JURNAL TEKNOINFO
Volume 17, Nomor 1, Januari 2023, Page 236-243
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

ANALISA PERBANDINGAN KINERJA ALGORITMA C4.5 DAN ALGORITMA K-NEAREST NEIGHBORS UNTUK KLASIFIKASI PENERIMA BEASISWA

Agung Purwanto[1)], Handoyo Widi Nugroho[2)]

[1,2]Fakultas Ilmu Komputer, IIB Darmajaya
[1,2]Jl. ZA. Pagar Alam No.93, Gedong Meneng, Kec. Rajabasa, Kota Bandar Lampung, Lampung
Email : [1]purwanto.agung@gmail.com, [2]handoyo.wn@darmajaya.ac.id

Abstract

Scholarship assignment is an operations management problem faced by university administrators, which is usually solved based on the administrator's personal experience. This study proposes an incentive method inspired by dynamic programming to replace the traditional decision-making process in awarding scholarships. The aim is to find the optimal scholarship assignment scheme with the highest payout while taking into account practical constraints and equity requirements. The methodology used in determining scholarship recipients at Muhammdiyah Pringsewu University is to compare the stages of the C.45 Algorithm and the K-Nearest Neighbors Algorithm. From some sample data recipient candidates from the K-Nearest Neighbors algorithm have better performance, namely 98.08% precision, 98.30% accuracy and 98.00% recall value, with an AUC result of 1,000 while the C4.5 algorithm. reached 97.23% with a precision value of 94.43%, a recall value of 100.00% and an AUC result of 0.956.

Keywords: *Scholarship, Classification, C4.5, K-Nearest Neighbors*

Abstrak

Penugasan beasiswa adalah masalah manajemen operasi yang dihadapi administrator universitas, yang biasanya diselesaikan berdasarkan pengalaman pribadi administrator. Penelitian ini mengusulkan metode insentif yang terinspirasi oleh pemrograman dinamis untuk menggantikan proses pengambilan keputusan tradisional dalam penugasan beasiswa. Tujuannya adalah untuk menemukan skema penugasan beasiswa yang optimal dengan ekuitas tertinggi sambil memperhitungkan kendala praktis dan persyaratan ekuitas Metodologi yang digunakan dalam menentukan penerima beasiswa di Universitas Muhammdiyah Pringsewu adalah dengan membandingkan tahapan Algoritma C.45 dan Algoritma K-Nearest Neighbors. Dari beberapa data sampel calon penerima dari jurusan algoritma K-Nearest Neighbors memiliki performansi yang lebih baik yaitu presisi 98,08%, akurasi 98,30% dan nilai recall 98,00%, dengan hasil AUC sebesar 1,000 sedangkan C4,5 algoritma. mencapai 97,23% dengan nilai precision 94.43%, nilai recall 100,00% dan hasil AUC 0,956.

Kata Kunci: Beasiswa,Klasifikasi,C4.5, K-Nearest Neighbors

1. PENDAHULUAN

Penghargaan beasiswa merupakan penghargaan yang diberikan kepada individu sarjana untuk melanjutkan pendidikan ke jenjang yang lebih tinggi. Itu reward yang diberikan dapat berupa akses khusus pada suatu institusi atau keuangan pendampingan. Pada dasarnya, beasiswa memberikan penghasilan bagi yang menerimanya. Biasanya, itu dalam bentuk dana yang dihabiskan untuk mahasiswa selama masa perkuliahan pada masa studi yang diinginkan. Pemerintah Provinsi selalu menawarkan program beasiswa ini setiap tahun. Sayangnya, beasiswa diberikan kepada para siswa subyektif sehingga banyak siswa yang memenuhi syarat tidak mendapatkan beasiswa dan sebaliknya. [1]. Dalam hal ini, sangat penting dan sangat penting bagi administrator universitas untuk mengembangkan alat yang sistematis untuk memberikan beasiswa dengan cara yang efisien dan adil. mekanisme seperti itu memakan waktu dan energi bagi siswa karena mereka cenderung berfokus pada pengumpulan informasi pesaing mereka dan mengembangkan strategi aplikasi yang tepat dengan mengorbankan studi dan penelitian mereka, sehingga memberikan dampak negatif secara keseluruhan pada kinerja akademik mereka. Dalam hal ini, sangat penting dan sangat penting bagi administrator universitas untuk mengembangkan alat yang sistematis untuk memberikan beasiswa dengan cara yang efisien dan adil. mekanisme seperti itu memakan waktu dan energi bagi siswa karena mereka cenderung berfokus pada pengumpulan informasi pesaing mereka dan mengembangkan strategi aplikasi yang tepat dengan mengorbankan studi dan penelitian mereka, sehingga memberikan dampak negatif secara keseluruhan pada kinerja akademik mereka. Dalam hal ini, sangat penting dan sangat penting bagi administrator universitas untuk mengembangkan alat yang sistematis untuk memberikan beasiswa dengan cara yang efisien dan adil.

JURNAL TEKNOINFO
Volume 17, Nomor 1, Januari 2023, Page 236-243
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

Penambangan data (DM) adalah ekstraksi dan pemrosesan informasi berharga dari gudang data besar. DM adalah bagian dari penambangan data. Langkah pertama dalam penambangan data (DM) adalah melihat data dengan berbagai cara dan menemukan informasi yang paling berharga dalam bentuk yang paling diringkas [2]. Dalam strategi pemasaran, pendekatan DM sangat bermanfaat karena meminimalkan data yang tidak perlu dan menghemat sumber daya. Mereka juga membantu menemukan pola perilaku konsumen dan praktis karena pengetahuannya yang sederhana. Terlepas dari hubungan yang jelas antara DM dan analisis data statistik/matematis, sebagian besar pendekatan yang digunakan dalam DM sejauh ini muncul dari subjek statistik[3]. Sebagai bagian dari penyelidikan kami, kami akan melihat beberapa model dan praktik pendidikan terbaru. Penambangan data berguna untuk mengekstraksi informasi dari kumpulan data yang besar. Ada banyak masalah data diselesaikan dengan menggunakan teknik penambangan data seperti asosiasi, prediksi, klasifikasi, dan pengelompokan. Untuk memecahkan beasiswa masalah penerima, klasifikasi dan cluster akan dilakukan oleh menggunakan teknik penambangan data[4]. Itu dilakukan dengan membandingkan dua Metode C4.5 dan K-Nearest Neighbors. Algoritma C4.5 membuat pohon keputusan berdasarkan konsep perolehan informasi, dengan setiap keputusan klasifikasi dikaitkan dengan klasifikasi target. Cara terbaik untuk menilai ketidakpastian adalah dengan menggunakan entropi.

Penelitian mengenai Penerapan Algoritma C4.5 Pada Asuransi dan Jasa Keuangan Menggunakan Metode Data Mining Setelah melakukan percobaan, didapat hasil akurasi tertinggi yaitu 96,25%[5], Selain itu Penerapan Algoritma Pohon Keputusan C4.5 untuk Mengevaluasi Pendidikan Musik Perguruan Tinggi Instruktur musik di universitas dan perguruan tinggi terus memperbarui metode pengajaran mereka dan memanfaatkan beberapa teknik untuk memberikan pengajaran mendalam di ruang kelas. Untuk memperluas antusiasme dan keterlibatan siswa sekaligus mengembangkan bakat kreatif musik mereka, sistem manajemen administrasi pendidikan informasi berbasis web telah banyak digunakan di banyak universitas dan perguruan tinggi. Akhirnya, ini mengidentifikasi atribut pengambilan keputusan yang mempengaruhi evaluasi guru[6]. Kinerja siswa sangat penting untuk keberhasilan perguruan tinggi. Terutama, prestasi akademik adalah salah satu metrik yang digunakan dalam pemeringkatan universitas berkualitas tinggi.Studi terkait yang diterbitkan antara 2015 dan 2021 diidentifikasi melalui pencarian sistematis dari berbagai database online. Tiga puluh sembilan studi dipilih dan dievaluasi.

Hasil penelitian menunjukkan bahwa enam model ML terutama digunakan: pohon keputusan (DT), jaringan syaraf tiruan (ANN), mesin vektor pendukung (SVM), tetangga terdekat K (KNN), regresi linier (LinR), dan Naive Bayes (NB)[7].Pengklasifikasi pembelajaran mesin seperti BPNN, RF, dan NB digunakan untuk mengklasifikasikan data kinerja akademik siswa. BPNN memiliki akurasi yang lebih baik untuk klasifikasi dan prediksi prestasi akademik mahasiswa[8]. Untuk memprediksi siswa dengan keterlibatan rendah, kami menerapkan beberapa algoritme ML ke kumpulan data. Dengan menggunakan algoritme ini, model yang dilatih pertama kali diperoleh; kemudian, akurasi dan nilai kappa model dibandingkan. Hasilnya menunjukkan bahwa J48, pohon keputusan, JRIP, dan pengklasifikasi yang ditingkatkan gradien menunjukkan kinerja yang lebih baik dalam hal akurasi, nilai kappa, dan daya ingat dibandingkan dengan model yang diuji lainnya. Berdasarkan temuan tersebut, kami mengembangkan dashboard untuk memfasilitasi instruktur di OU. Model-model ini dapat dengan mudah dimasukkan ke dalam sistem VLE untuk membantu instruktur mengevaluasi keterlibatan siswa selama kursus VLE sehubungan dengan berbagai aktivitas dan materi dan untuk memberikan intervensi tambahan bagi siswa sebelum ujian akhir mereka[9]. K-Nearest Neighbors atau biasa disingkat dengan KNN adalah metode nonparametrik untuk mengklasifikasikan data baru yang kelasnya belum diketahui dan pilihlah data sebanyak k yang terdekat terletak dari data baru. Umumnya, k ditentukan sebagai ganjil angka untuk menghindari munculnya jumlah jarak yang sama dalam proses klasifikasi. Metode ini memberikan lebih banyak pendekatan fleksibel dengan bentuk eksplisit untuk f(k). Metode ini seringkali lebih kompleks untuk dipahami dan ditafsirkan. Untuk beberapa dari pengamatan pada prediktor data, metode parametrik bekerja lebih baik[10]. Pengklasifikasi KNN adalah salah satu lingkungan yang paling populer pengklasifikasi dalam pengenalan pola. Jika dibandingkan dengan Bayes algoritma dan perhitungan jarak Euclidean lainnya, K-Nearest algoritma tetangga memiliki efisiensi dan kinerja yang lebih baik.

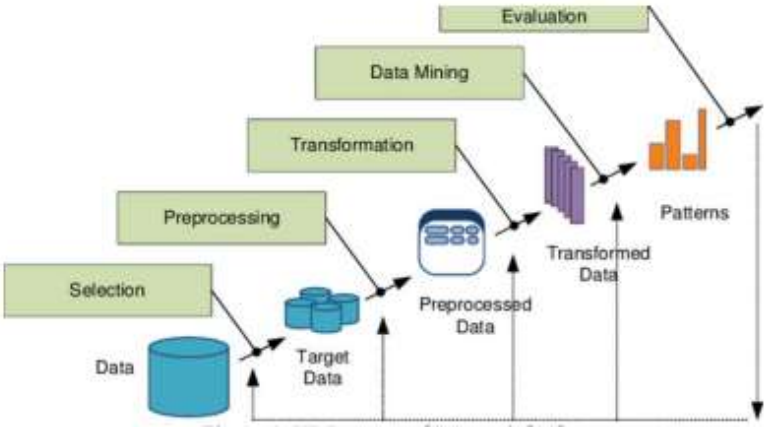
Penelitian ini melakukan analisis komparatif antara algoritma C4.5 dan k-nearest neighbor untuk klasifikasi penerimaan siswa. Pilihan penggunaan kedua algoritma ini sangat populer dan banyak digunakan dalam praktek. Penelitian ini dilakukan untuk memberikan dukungan keputusan kepada perguruan tinggi, meminimalkan kesalahan yang dilakukan oleh penyedia beasiswa, dan mengetahui algoritma mana yang memiliki nilai akurasi paling tinggi dalam mengklasifikasikan penerima beasiswa.

2. METODE PENELITIAN

2.1 Alur Penelitian

Penelitian ini menggunakan Knowledge Discovery in Database (KDD)[11], tahapan model proses seperti yang ditunjukkan pada Gambar 1 Penjelasan dari setiap langkah penelitian seperti pada gambar 1 dibawah ini:

JURNAL TEKNOINFO
Volume 17, Nomor 1, Januari 2023, Page 236-243
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index



Gambar 1 Alur Penelitian

a. Pilihan
Pada tahap ini dilakukan seleksi data terhadap data mahasiswa aktif dan mahasiswa yang telah mendaftarkan diri sebagai penerima beasiwa.
b. Preprocessing
Seluruh Mahasiswa Angkatan 2020 dan 2021 dari semua jurusan di Universitas Muhammadiyah Pringsewu digunakan sebagai data. Sebuah data pembersihan adalah dilakukan pada data tersebut untuk memeriksa nilai yang hilang, duplikasi data, atau data outlier.
c. Transformasi
Setelah dilakukan pembersihan data, tahap selanjutnya adalah transformasi data berdasarkan tipe data, dimana data tersebut adalah klasifikasi berdasarkan kategorinya.
d. Penambangan Data
Pada tahap ini, teknik penambangan data yang tepat dipilih. Untuk fungsi klasifikasi, C4.5 dan *K-Nearest Neighbor* digunakan. Klasifikasi merupakan pembelajaran terawasi, jadi tahap ini termasuk dalam model pembelajaran yang diawasi.
e. Evaluasi
Tahapan ini dilakukan untuk mengevaluasi hasil prediksi Algoritma yang memiliki nilai relatif dengan klasifikasi data sebenarnya. Metode Confusion Matrix digunakan sebagai metode evaluasi. Penampilan nilai penilaian adalah akurasi dan error.

2.2 *Decision tree (C4.5)*

C4.5 adalah kumpulan algoritma untuk teknik klasifikasi dalam pembelajaran mesin dan penambangan data. Tujuannya adalah pembelajaran terawasi, di mana setiap tupel dalam kumpulan data dapat dijelaskan oleh sekumpulan nilai atribut, dan setiap tupel milik salah satu dari banyak kelas yang berbeda dan tidak kompatibel. Tujuan C4.5 adalah mempelajari pemetaan dari nilai atribut ke kategori yang dapat digunakan untuk mengkategorikan item yang tidak diketahui ke dalam kategori baru. J. Rossi Quinlan menyarankan C4.5 berdasarkan ID3. Sebuah pohon keputusan dibangun menggunakan algoritma ID3. Sebuah pohon keputusan adalah struktur pohon yang seperti flowchart, dengan setiap node internal (node nonleaf) yang mewakili tes pada atribut, setiap cabang mewakili hasil tes, dan setiap node daun memegang label kelas. Setelah membuat pohon keputusan untuk tupel yang tidak menyediakan label klasifikasi, kami memilih jalur dari simpul akar ke simpul daun, dan jalur tersebut menyimpan informasi prediksi tupel. Pohon keputusan memiliki keuntungan karena tidak memerlukan informasi domain atau konfigurasi parameter, menjadikannya ideal untuk penggalian informasi eksplorasi.

Algoritma C4.5 didasarkan pada ID3 yang ditambahkan ke atribut kontinyu, nilai atribut, dan pemrosesan informasi, dengan membangkitkan pohon untuk membangun pohon keputusan pemangkasan dalam dua tahap. Pada setiap atribut dengan perhitungan informasi algoritma C4.5, kita dapat mengetahui Rasio Gain laju perolehan informasi. Akhirnya, dipilih dengan tingkat perolehan informasi tertinggi dari atribut uji set yang diberikan untuk mengatur cabang. Menurut nilai atribut uji menggunakan algoritma rekursif, dapatkan pohon keputusan awal. Rumus komputasi terkait algoritma C4.5 sebagai berikut[12]. Pertama, nilai ekspektasi yang diperlukan untuk klasifikasi sampel diberikan sebagai berikut:Tentukan akar pohon dengan menghitung nilai gain tertinggi dari setiap atribut atau nilai indeks entropi terendah. Sebelumnya, nilai indeks entropi dihitung menggunakan rumus:

$$Entropy\ (i) = \sum_{j=1}^{m} \quad f(i,j).\ 2f[(i,j)] \tag{1}$$

a. Nilai gain dengan rumus:

JURNAL TEKNOINFO
Volume 17, Nomor 1, Januari 2023, Page 236-243
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

$$gain = -\sum_{i=1}^{p} \quad . IE(i) \tag{2}$$

b. Untuk menghitung gain ratio perlu diketahui suatu term baru yang disebut Split Information dengan rumus:

$$SplitInformation = -\sum_{t=1}^{c} \frac{S1}{S} log2 \frac{S1}{S} \tag{3}$$

c. Selanjutnya menghitung gain ratio

$$Gainratio(S, A) = \frac{\text{Gain(S.A)}}{\text{SplitInformation (S,A)}} \tag{4}$$

d. Ulangi langkah 2 sampai semua record telah terpecah. Proses pemisahan pohon keputusan berakhir ketika:
   1) Semua tupel dalam catatan simpul m adalah kelas yang sama.
   2) Atribut dalam dataset tidak dibagi lagi.
   3) Cabang kosong tidak memiliki catatan

2.3 *K-Nearest Neighbors*

Algoritma KNN adalah algoritma klasifikasi atau regresi nonparametrik dalam bidang pengenalan pola. Pada subbagian ini, kami secara singkat memperkenalkan algoritma klasifikasi KNN. Asumsikan ada beberapa data pelatihan yang memiliki beberapa atribut dan label. Selanjutnya ada kelompok data testing yang hanya memiliki beberapa atribut saja tanpa label[13]. Tujuan dari algoritma klasifikasi KNN adalah untuk mendapatkan label dari data pengujian. Proses spesifiknya adalah sebagai berikut.

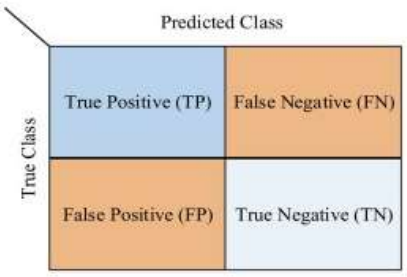
- (x_i, y_i), $i = 1, 2, \ldots, N$
- x_i adalah data pelatihan dalam R^n
- y_i adalah kelas yang sesuai dari data x_i, dan $y_i \{c_j, j\ 1, 2, \ldots M\}$
- $dist\ (x - x_i) = || x - x_i ||$

Algoritma KNN klasik dapat digunakan untuk prediksi numerik, yaitu mengembalikan nilai prediksi sebenarnya menurut tuple yang tidak diketahui[12]. Strategi pemasaran harga harus memahami dasar objektif penetapan harga; melakukan riset pasar; menganalisis secara objektif kebijakan nasional, lingkungan pasar, nilai proyek pembangunan itu sendiri, hubungan penawaran dan permintaan pasar real estat, situasi persaingan pasar, psikologi pembeli rumah, dan faktor lainnya; pilih metode penetapan harga ilmiah; dan mengadopsi strategi yang sesuai.

2.4 Evaluasi Kinerja

a. k-fold Cross Validation
   k-fold cross-validation adalah teknik untuk memvalidasi akurasi model yang dibangun pada kumpulan data tertentu, yang membagi kumpulan data menjadi dua bagian, yaitu data pelatihan dan data pengujian. Untuk masalah prediksi, model biasanya diberi dataset dari data yang diketahui untuk dilatih (dataset pelatihan) dan data yang tidak diketahui (atau data yang pertama kali muncul) untuk menguji model (disebut validasi). atau data uji)[14]. Tujuan validasi silang adalah untuk menguji kemampuan model untuk memprediksi data baru yang tidak digunakan dalam evaluasinya, untuk menandai masalah seperti overfitting atau bias seleksi, dan untuk memberikan wawasan tentang bagaimana model menggeneralisasi data independen . set (yaitu dataset yang tidak diketahui, misalnya masalah).

b. Confusion matrix
   Confusion matrix atau Matriks kebingungan adalah ukuran yang sangat populer digunakan saat memecahkan masalah klasifikasi. Ini dapat diterapkan untuk klasifikasi biner serta untuk masalah klasifikasi multikelas[15]. Matriks ini digunakan untuk evaluasi kinerja metode yang digunakan setelah klasifikasi. Untuk klasifikasi biner, skema dari matriks konfusi terlihat pada Gambar 2



Gambar 2 Skema Confusion Matrix

Matriks konfusi merepresentasikan nilai TP yang diklasifikasikan dengan benar, nilai FP di kelas yang relevan saat seharusnya berada di kelas lain, dan nilai FN di kelas lain saat seharusnya berada di kelas yang relevan dan nilai TN yang diklasifikasikan dengan benar di kelas lain. Metrik kinerja yang paling sering digunakan untuk klasifikasi menurut nilai-nilai ini adalah akurasi (ACC), presisi (P), sensitivitas (Sn), spesifisitas (Sp), dan

JURNAL TEKNOINFO
Volume 17, Nomor 1, Januari 2023, Page 236-243
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

nilai skor- F[15] . Perhitungan metrik kinerja ini menurut nilai-nilai dalam matriks kebingungan dibuat menurut Persamaan.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$P = \frac{TP}{TP+FP} \quad (7)$$

$$Sn = \frac{TP}{TP+FN} \quad (8)$$

$$Sp = \frac{TN}{TN+FP} \quad (9)$$

$$F - score = 2x\frac{P \times Sn}{P+Sn} \quad (10)$$

c. Kurva ROC

Kurva ROC adalah plot grafis yang mengilustrasikan kemampuan diagnostik sistem pengklasifikasi biner karena ambang diskriminasinya bervariasi. Metode ini awalnya dikembangkan untuk operator penerima radar militer mulai tahun 1941, yang kemudian memunculkan namanya. Kurva ROC dibuat dengan memplot true positive rate (TPR) terhadap false positive rate (FPR) pada berbagai pengaturan ambang batas. Tingkat positif sejati juga dikenal sebagai sensitivitas , daya ingat , atau probabilitas deteksi. Tingkat positif palsu juga dikenal sebagai probabilitas alarm palsu dan dapat dihitung sebagai (1 - spesifisitas)[15]. ROC juga dapat dianggap sebagai sebidang kekuatan sebagai fungsi dari Kesalahan Tipe Idari aturan keputusan (ketika kinerja dihitung hanya dari sampel populasi, dapat dianggap sebagai estimator dari jumlah ini). Performance keakurasian AUC dapat diklasifikasikan menjadi beberapa kelompok yaitu[17]:

1. 0.90 – 1.00 = *Exellent Classification*
2. 0.80 – 0.90 = *Good Classification*
3. 0.70 – 0.80 = *Fair Classification*
4. 0.60 – 0.70 = *Poor Classification*
5. 0.50 – 0.60 = *Failure Classification*

3. HASIL DAN PEMBAHASAN

3.1 Dataset

Pengolahan dataset yang dibagi menjadi dataset traning dan testing dengan jumlah 940 record yang terdiri dari 18 atribut. Data tersebut bisa dilihat atau pada Gambar 3 dibawah ini.

ber Lis...	Luas Tanah	Luas Bang...	Sumber Air	MCK	Prestasi	SNMPN	BEASISWA
	integer	integer	polynominal	polynominal	polynominal	polynominal	binominal
dan Genset	50	25	Sumur	Kepemilikan Sen...	Tidak Ada	-	DITERIMA
Ada	0	0	Tidak Ada	Tidak Ada	Tidak Ada	-	DITERIMA
	200	100	Sumur	Kepemilikan Sen...	Ada	-	DITERIMA
Ada	0	0	Tidak Ada	Tidak Ada	Ada	-	DITERIMA
	25	25	Sumur	Kepemilikan Sen...	Tidak Ada	-	DITERIMA
	100	50	Sumur	Kepemilikan Sen...	Tidak Ada	-	DITERIMA
	0	0	Tidak Ada	Tidak Ada	Tidak Ada	-	DITERIMA
Ada	0	0	Tidak Ada	Tidak Ada	Ada	-	DITERIMA
	220	100	Sumur	Kepemilikan Sen...	Tidak Ada	?	TIDAK DITERIMA
	220	100	Sumur	Kepemilikan Sen...	Tidak Ada	?	TIDAK DITERIMA
	99	50	Sungai/Mata Air	Kepemilikan Sen...	Ada	-	DITERIMA
	99	50	Sungai/Mata Air	Kepemilikan Sen...	Tidak Ada	-	DITERIMA
	220	100	Sumur	Kepemilikan Sen...	Tidak Ada	?	TIDAK DITERIMA
	0	0	Tidak Ada	Tidak Ada	Tidak Ada	-	DITERIMA
	220	100	Sumur	Kepemilikan Sen...	Tidak Ada	?	TIDAK DITERIMA
Ada	0	0	Tidak Ada	Tidak Ada	Ada	-	DITERIMA
	100	80	Sumur	Kepemilikan Sen...	Tidak Ada	?	TIDAK DITERIMA
	200	100	Sumur	Kepemilikan Sen...	Tidak Ada	-	DITERIMA

Gambar 3 Potongan Dataset

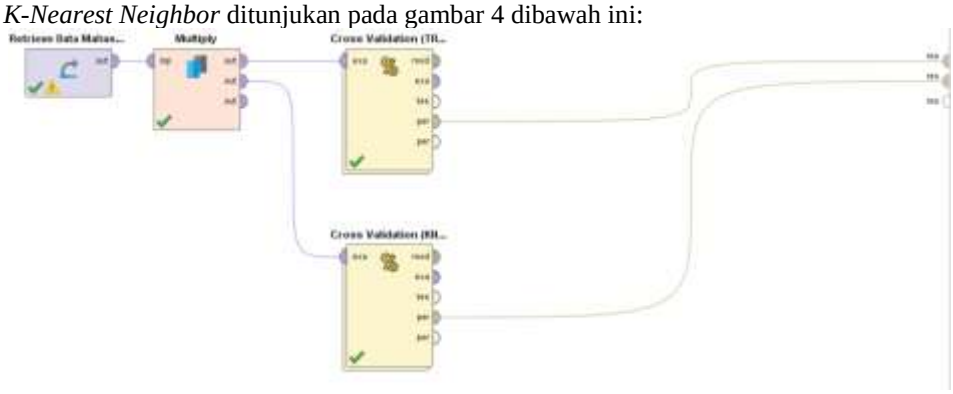
3.2 Pembahasan

3.2.1. Eksperimen dan Pengujian

Tahap implementasi dari algoritma ini adalah penulis melakukan pengujian. Dalam pengujian dari total 940 data preprocessing penulis membagi data menjadi data latih (pelatihan) dan data uji (testing).

JURNAL TEKNOINFO
Volume 17, Nomor 1, Januari 2023, Page 236-243
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

Dalam pengujian ini penulis menggunakan operator validasi silang. Operator validasi silang ininantinya akan memisahkan data dengan cara membagi total data dari data training dan sisanya untuk menguji data. Penerapan data pada Rapid Miner digunakan untuk Klasifikasi Penerima Beasiswa menggunakan algoritma *Decision tree* dan algoritma *K-Nearest Neighbor* ditunjukan pada gambar 4 dibawah ini:



Gambar 4 Skema Pengujian Dengan Rapid Miner

3.2.2. Evaluasi dan Validasi Hasil

1. Hasil Pengujian Algoritma C4.5

Pada tahap ini peneliti menggunakan metode algoritma C4.5 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 C4.5 mencapai akurasi 97.23%

accuracy: 97.23% +/- 3.66% (micro average: 97.23%)

	true DITERIMA	true TIDAK DITERIMA	class precision
pred. DITERIMA	514	0	100.00%
pred. TIDAK DITERIMA	26	400	93.90%
class recall	95.19%	100.00%	

Gambar 5 Confusion Matrix C4.5

$$ACC = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} = \frac{514+400}{514+400+26+0} = \frac{914}{940} = 97{,}23\%$$

2. Hasil Pengujian Algoritma KNN

Pada tahap ini peneliti menggunakan metode algoritma *K-Nearest Neighbor* 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 *K-Nearest Neighbor* mencapai hasil akurasi 98,72%.

accuracy: 98.30% +/- 1.96% (micro average: 98.30%)

	true DITERIMA	true TIDAK DITERIMA	class precision
pred. DITERIMA	532	8	98.52%
pred. TIDAK DITERIMA	8	392	98.00%
class recall	98.52%	98.00%	

Gambar 5 Confusion Matrix KNN

$$ACC = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} = \frac{532+392}{532+392+8+8} = \frac{924}{940} = 98{,}30\%$$

Hasil pengujian ini diantaranya adalah mendapatkan nilai akurasi yang merupakan ketepatan antara nilai derajat yang diukur dengan nilai sebenarnya, Precission yang merupakan pemilihan oleh system dokumen teks yang relevan dari semua dokumen teks, dan Recall merupakan proporsi jumlah dokumen teks relevan yang ada pada koleksi, Perbandingan hasil pengujian menggunaakan Algoritma C4.5 dan *K-Nearest Neighbor* dapat kita lihat pada Tabel 1 dibawah ini.

Tabel 1 Perbandingan Kinerja

Model	Akurasi	Presisi	Recall
K-Nearest Neighbor	98.30%	98.08%	98.00%
C4.5	97.23%	94.43%	100.00%

JURNAL TEKNOINFO
Volume 17, Nomor 1, Januari 2023, Page 236-243
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

Selain Confusion Matrix untuk mengetahui kinerja dari pengujian ini kami mengandalkan kurva ROC/AUC.(Area Under Curve) yang dihasilkan. Perbandingan hasil Kurva AUC menggunaakan Algoritma C4.5 dan *K-Nearest Neighbor* dapat kita lihat pada Tabel 2 dibawah ini.

Tabel 2 Perbandingan hasil AUC

Model	AUC
K-Nearest Neighbor	1.000
C4.5	0.726

Berdasarkan klasifikasi tersebut dapat disimpulkan bahwa algoritma C4.5 dan *K-Nearest Neighbor* merupakan algoritma yang akurat untuk memprediksi karena nilai AUC termasuk dalam predikat Excellent Classification yaitu dengan nilai 0.90 – 1.00.

3. KESIMPULAN

Penelitian ini mengembangkan metode untuk mendapatkan skema penerimaan beasiswa yang optimal dengan pemerataan tertinggi bagi penyelenggara universitas. 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 K-Nearest Neighbors memiliki performansi yang lebih baik yaitu presisi 98,08%, akurasi 98,30% dan nilai recall 98,00%, dengan hasil AUC sebesar 1,000 sedangkan C4,5 algoritma. mencapai 97,23% dengan nilai precision 94.43%, nilai recall 100,00% dan hasil AUC 0,956.

UCAPAN TERIMA KASIH

Terima kasih keluarga yang tak pernah henti memberikan support, teman-temen seperjuangan MTI angkatan IIB Darmajaya.

DAFTAR PUSTAKA

[1] L. Pengembangan, T. Informasi, and D. Komunikasi, *KAMUS ABREVIASI BAHASA INDONESIA*. 2015.
[2] T. Masters, *Data Mining Algorithms in C++*. Apress, 2018. doi: 10.1007/978-1-4842-3315-3.
[3] Bruce Ratner, "Statistical and Machine-Learning Data Mining Techniques for Better Predictive Modeling and Analysis of Big Data Third Edition," 2017.
[4] P. C. B. I. Y. N. R. P. K. C. L. Jr. Galit Shmueli, "DATA MINING FOR BUSINESS ANALYTICS," 2018.
[5] S. Xuanyuan, S. Xuanyuan, and Y. Yue, "Application of C4.5 Algorithm in Insurance and Financial Services Using Data Mining Methods," *Mobile Information Systems*, vol. 2022, 2022, doi: 10.1155/2022/5670784.
[6] J. Wang, "Application of C4.5 Decision Tree Algorithm for Evaluating the College Music Education," *Mobile Information Systems*, vol. 2022, 2022, doi: 10.1155/2022/7442352.
[7] Y. A. Alsariera, Y. Baashar, G. Alkawsi, A. Mustafa, A. A. Alkahtani, and N. Ali, "Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance," *Computational Intelligence and Neuroscience*, vol. 2022. Hindawi Limited, 2022. doi: 10.1155/2022/4151487.
[8] M. Kamal *et al.*, "Metaheuristics Method for Classification and Prediction of Student Performance Using Machine Learning Predictors," *Math Probl Eng*, vol. 2022, 2022, doi: 10.1155/2022/2581951.
[9] M. Hussain, W. Zhu, W. Zhang, and S. M. R. Abidi, "Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores," *Comput Intell Neurosci*, vol. 2018, 2018, doi: 10.1155/2018/6347186.
[10] Parteek Bhatia, "Data Mining and Data Warehousing," 2019.
[11] D. Forsyth, "Probability and Statistics for Computer Science," 2018.
[12] P. V. Ngoc, C. V. T. Ngoc, T. V. T. Ngoc, and D. N. Duy, "A C4.5 algorithm for english emotional classification," *Evolving Systems*, vol. 10, no. 3, pp. 425–451, Sep. 2019, doi: 10.1007/s12530-017-9180-1.
[13] Andi Ainun Khaerunnisyah Qodrat, "Perbandingan Algoritma Naïve Bayes Dan K-Nearest Neighbor Untuk Sistem Kelayakan Kredit Pada Nasabah (Studi Kasus: PT. Armada Finance Cabang Makassar)," 2017.
[14] J. Unpingco, *Python for probability, statistics, and machine learning*. Springer International Publishing, 2016. doi: 10.1007/978-3-319-30717-6.
[15] O. Caelen, "A Bayesian Interpretation of the Confusion Matrix," 2017.

[16] D. J. H. Wojtek J. Krzanowski, "ROC Curves for Continuous Data," 2009.
[17] J. Moolayil, *Learn Keras for Deep Neural Networks*. Apress, 2019. doi: 10.1007/978-1-4842-4240-7.