

Perbandingan Algoritma C4.5 dan k-Means Dalam Memprediksi Kategori Indeks Prestasi Mahasiswa

Gabriella Amelia Prasetyo¹, R. Gunawan Santosa², Antonius Rachmat Chrismanto³

Program Studi Informatika, Universitas Kristen Duta Wacana Yogyakarta

Jl. Dr. Wahidin Sudirohusodo No.5-25, Yogyakarta

¹gabriella.amelia@ti.ukdw.ac.id

²gunawan@staff.ukdw.ac.id

³anton@ti.ukdw.ac.id

Abstract— The purpose of this research is to compare the accuracy between C4.5 and k-Means in predicting the FTI UKDW students' grade point average for the first semester. The dataset is taken from the data of FTI UKDW student from class of 2008 to 2016 (training data) and class of 2017 (test data). Achievement admission and regular admission use different number of features. Achievement admission only uses category, status, location, and ICE level as features, while regular admission also uses numeric, verbal, spatial, and analogy score from the admission test. The accuracy score is calculated by using cross-tabulation. C4.5 algorithm highest accuracy score is 77,45%, while k-Means highest is 60,78%. Achievement admission scenarios average accuracy score is 55,27% while regular admission scenarios average accuracy score is 38,95%.

Intisari— Penelitian ini dilakukan untuk membandingkan akurasi prediksi algoritma C4.5 dan k-Means dalam memprediksi nilai indeks prestasi semester 1 mahasiswa FTI UKDW. Data yang digunakan adalah data mahasiswa FTI UKDW angkatan 2008-2016 sebagai data latih, dan angkatan 2017 sebagai data uji. Atribut yang digunakan akan dibedakan berdasarkan jalur prestasi dan non-prestasi. Jalur prestasi menggunakan atribut kategori, status, lokasi, dan level ICE, sedangkan jalur non-prestasi menggunakan atribut kategori, status, lokasi, level ICE, numerik, verbal, spasial, analogi. Akurasi akan dihitung menggunakan *cross-tabulation*. Algoritma C4.5 mendapatkan hasil terbaik sebesar 77,45% dan algoritma k-Means mendapatkan hasil terbaik sebesar 60,78%. Skenario dengan jalur prestasi mendapatkan hasil rata-rata akurasi sebesar 55,27% dan skenario dengan jalur non-prestasi mendapatkan hasil rata-rata akurasi sebesar 38,95%.

Kata Kunci— prediksi indeks prestasi mahasiswa, akurasi data mining, C4.5, k-Means

I. PENDAHULUAN

Setiap tahunnya, Universitas Kristen Duta Wacana (UKDW) menerima sejumlah mahasiswa baru. Jalur penerimaan mahasiswa baru dibagi menjadi dua, yaitu: jalur prestasi dan jalur non-prestasi. Jalur prestasi adalah jalur penerimaan mahasiswa baru berdasarkan nilai rapor calon mahasiswa, sedangkan jalur non-prestasi adalah jalur penerimaan mahasiswa baru berdasarkan hasil tes seleksi [1]. Data yang didapatkan dari jalur prestasi adalah kategori sekolah (SMA atau SMK), status sekolah (Negeri atau Swasta) dan lokasi sekolah (Jawa atau luar Jawa), sedangkan data yang didapatkan dari jalur non-prestasi adalah data yang

sama dengan jalur prestasi, ditambah dengan kemampuan Spasial, Verbal, Numerik dan Analogi. Selain itu, kemampuan bahasa Inggris mahasiswa juga menjadi pertimbangan penerimaan mahasiswa, dimana calon mahasiswa diklasifikasikan ke dalam empat level kemampuan bahasa Inggris, yaitu: level 1, 2, 3 atau ESP (*English for Special Purpose*) [2].

Data penerimaan mahasiswa yang didapatkan oleh UKDW mengandung banyak informasi. Dari data tersebut dapat dilakukan pencarian pola dengan metode *data mining* atau penambangan data. Hasil dari penambangan data adalah informasi mengenai pola dari suatu data dan dapat digunakan untuk memprediksi data yang akan datang. Dengan kata lain, data penerimaan mahasiswa yang sudah ada dapat digunakan untuk memprediksi data mahasiswa yang akan datang. Salah satu implementasi dari penambangan data ini adalah untuk memprediksi indeks prestasi mahasiswa baru pada semester 1. Dari data mahasiswa yang sudah ada akan dicari pola yang menarik untuk memprediksi nilai indeks prestasi mahasiswa dengan data pribadi sesuai dengan jalur penerimaannya (jalur prestasi atau non-prestasi). Hasil prediksi ini dapat dimanfaatkan sebagai acuan baru dalam penerimaan mahasiswa, juga membantu pengajar memberi perhatian lebih kepada mahasiswa baru dengan nilai indeks prestasi tertentu.

Penelitian ini merupakan lanjutan dari penelitian [2], [3], [4], [5], dimana perbedaannya terletak pada algoritma dan data mahasiswa yang digunakan. Penelitian terdahulu menggunakan data mahasiswa angkatan 2008-2014 sebagai data latih dan angkatan 2015 sebagai data uji, sedangkan pada penelitian ini menggunakan data mahasiswa angkatan 2008-2016 sebagai data latih dan angkatan 2017 sebagai data uji.

Tujuan dari penelitian ini adalah menerapkan dan membandingkan akurasi algoritma C4.5 dan k-Means *clustering* dalam memprediksi kategori IPK mahasiswa FTI UKDW.

Penelitian mengenai implementasi algoritma dalam memprediksi indeks prestasi mahasiswa di FTI UKDW sudah pernah dilakukan sebelumnya oleh Santosa dan Chrismanto [2] & [4], Alverina [3], dan Sari [5]. Keempat penelitian tersebut menggunakan data mahasiswa FTI UKDW angkatan 2008-2015. Santosa dan Chrismanto [2] & [4] menggunakan regresi logistik, Alverina [3] membandingkan algoritma C4.5 dan CART, sedangkan Sari

[5] membandingkan algoritma k-Nearest Neighbor (KNN) dan Naïve Bayes Classifier.

Penelitian ini dan penelitian Alverina [3] sama-sama menggunakan algoritma C4.5, tetapi berbeda dalam ukuran untuk memilih atributnya. Alverina [3] menggunakan *Gain* sebagai ukuran dalam memilih atribut pada algoritma C4.5, sedangkan penelitian ini menggunakan *Gain Ratio* [6]. Hasil dari ketiga penelitian tersebut adalah algoritma C4.5 dan CART memiliki akurasi terbaik untuk jalur prestasi yaitu 86.86%, dan algoritma C4.5 memiliki akurasi terbaik kedua untuk jalur non-prestasi yaitu 61.45%.

Hssina, Merbouha, Ezzikouri, dan Erritali [7] melakukan penelitian tentang membandingkan algoritma *decision tree* ID3 dan C4.5. Hasil dari penelitian tersebut menunjukkan bahwa algoritma C4.5 mendapatkan hasil akurasi dan waktu eksekusi yang lebih baik daripada algoritma ID3. Algoritma C4.5 menghasilkan rata-rata akurasi sebesar 87,94% dan rata-rata waktu eksekusi sebesar 0,13383 detik, sedangkan algoritma ID3 menghasilkan rata-rata akurasi sebesar 84,94% dan rata-rata waktu eksekusi sebesar 0,308 detik.

Oyelade, Oladipupo, dan Obagbuwa [8] melakukan penelitian mengenai penggunaan algoritma k-Means *clustering* untuk memprediksi performa akademik murid di sekolah. Hasil dari penelitiannya menyatakan bahwa algoritma k-Means *clustering* memberikan hasil yang bagus untuk memantau performa akademik murid. Hasil penelitian ini juga bermanfaat bagi pengajar untuk mengetahui kandidat murid yang harus diberi perhatian agar performa akademiknya meningkat setiap semester.

Ramadhani [9] dalam penelitiannya menggunakan algoritma k-Means *clustering* untuk menentukan strategi promosi sebuah universitas. Atribut yang digunakan adalah kota asal, program studi, dan IPK mahasiswa. Untuk data yang berjenis nominal, seperti kota asal dan program studi, dilakukan proses inialisasi dengan mengurutkan data dari frekuensi paling besar ke paling kecil. Penelitian tersebut menghasilkan tiga *cluster*, dimana masing-masing *cluster* memiliki kota asal, program studi, dan rata-rata IPK yang berbeda. Hasil dari *clustering* ini bermanfaat bagi admisi dalam menyusun strategi promosi dengan melihat program studi yang paling banyak diminati pada setiap wilayah dan potensi akademik berdasarkan rata-rata IPK yang ada pada masing-masing *cluster*.

Berdasarkan tinjauan terhadap penelitian-penelitian di atas, penulis akan melakukan penelitian untuk membandingkan algoritma C4.5 dan k-Means *clustering* dalam memprediksi kategori indeks prestasi mahasiswa. Penelitian-penelitian di atas digunakan sebagai acuan bagi peneliti dalam melakukan penelitian ini.

II. LANDASAN TEORI

Ada beberapa landasan teori yang digunakan dalam penelitian ini, yaitu Data Mining, Decision Tree, Clustering, Algoritma C4.5, Algoritma k-Means Clustering, dan Cross-tabulation.

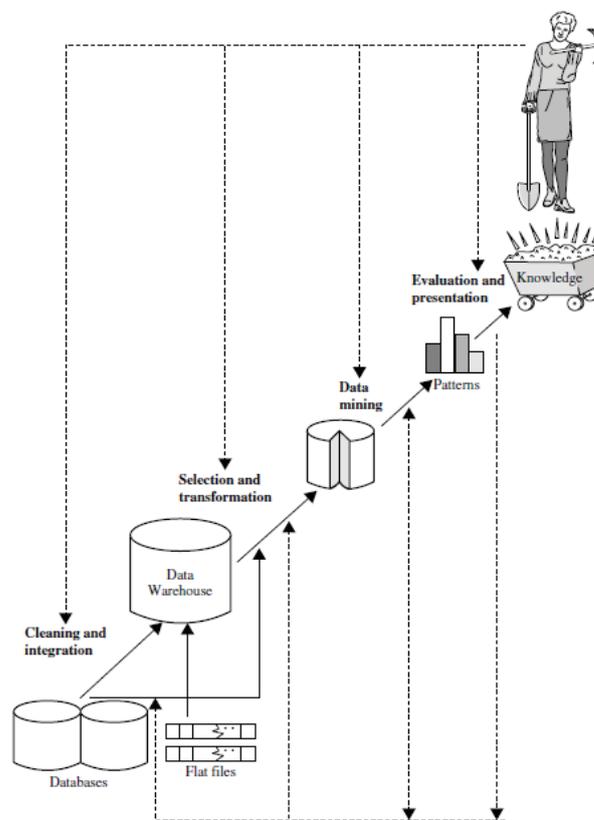
A. Data Mining

Data mining adalah proses pencarian pola yang menarik dan pengetahuan yang didapatkan dari data dalam jumlah yang besar [6]. *Data mining* memiliki banyak istilah, antara lain: penambahan pengetahuan dari data, ekstraksi

pengetahuan, analisis data atau pola, arkeologi data, dan pengerukan data.

Data mining sering disebut sebagai bagian dari proses *Knowledge Discovery from Data* (KDD), namun ada juga yang menyebut bahwa data mining adalah KDD itu sendiri. Gambar 1 menunjukkan proses dari KDD menurut [6], berikut ini adalah penjelasannya:

1. *Data cleaning*, untuk menghilangkan *noise* dan data yang tidak konsisten.
2. *Data integration*, proses dimana berbagai sumber data bisa saja dikombinasikan.
3. *Data selection*, proses dimana data yang relevan dengan *task* yang dianalisis akan dimasukkan dalam basis data.
4. *Data transformation*, proses dimana data ditransformasi atau dikonsolidasi ke dalam bentuk yang lebih sesuai untuk ditambang.
5. *Data mining*, proses yang penting dimana metode tertentu diimplementasikan untuk menemukan pola dalam data.
6. *Pattern evaluation*, untuk mengidentifikasi pola yang menarik yang merepresentasikan pengetahuan.
7. *Knowledge presentation*, proses dimana visualisasi dan representasi pengetahuan digunakan untuk menampilkan pengetahuan yang didapatkan kepada pengguna.

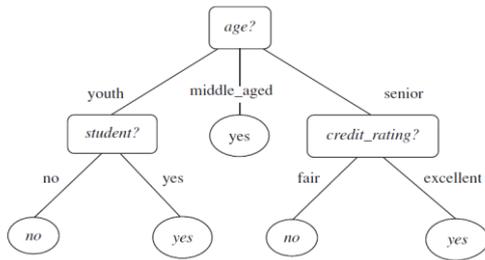


Gambar 1. Data mining sebagai bagian dari proses KDD.

B. Decision Tree

Decision tree atau pohon klasifikasi adalah salah satu cara menghasilkan aturan klasifikasi dengan menggunakan struktur seperti pohon [10]. *Node* yang berada di paling atas disebut *root*. Sedangkan *node* yang berada di akhir 8 cabang

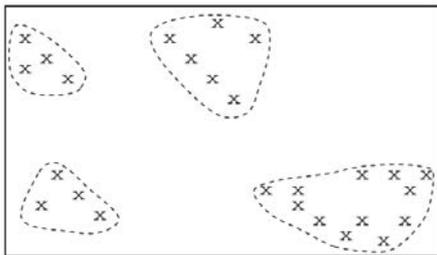
adalah *leaf node* [6]. Gambar 2 menunjukkan contoh kasus penggunaan *decision tree* dalam memprediksi apakah seseorang akan membeli komputer atau tidak berdasarkan atribut umur, status pendidikan, dan *credit rating*.



Gambar 2. *Decision tree* untuk pembelian komputer.

C. Clustering

Clustering adalah pengelompokkan objek-objek yang mirip satu dengan yang lainnya dan berbeda dengan objek-objek yang berada di *cluster* lain [10]. Metode *clustering* dapat dibagi ke dalam beberapa kategori, yaitu metode *partitioning*, metode *hierarchical*, metode *density-based*, metode *grid-based*, metode *model-based* [6]. Gambar 3 menunjukkan contoh *clustering* yang dilakukan terhadap sejumlah data ke dalam 4 *cluster*.



Gambar 3. *Clustering* pada data ke dalam 4 *cluster*

D. Algoritma C4.5

Algoritma C4.5 adalah algoritma *decision tree* yang dikembangkan dari algoritma ID3 (Iterative Dichotomiser). Algoritma C4.5 dan ID3 ditemukan oleh J. Ross Quinlan, seorang peneliti di bidang *machine learning* [6]. *Decision tree* dibangun dengan menggunakan *Gain Ratio* sebagai ukuran untuk menentukan atribut mana yang menjadi *node* yang dipilih.

Persamaan (1) digunakan untuk menghitung nilai *Entropy*.

$$Info(D) = - \sum_{i=1}^m p_i \log_2(p_i) \quad (1)$$

Keterangan:

- D : data yang digunakan
- m : jumlah partisi
- pi : probabilitas dari i terhadap D

Persamaan (2) digunakan untuk menghitung nilai *Entropy* terhadap atribut.

$$Info_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} \times Info(D_j). \quad (2)$$

Keterangan:

- D : data yang digunakan

- A : atribut yang mempartisi D
- v : jumlah partisi
- $\frac{|D_j|}{|D|}$: probabilitas dari partisi j terhadap D

Persamaan (3) digunakan untuk menghitung *Information Gain*.

$$Gain(A) = Info(D) - Info_A(D) \quad (3)$$

Keterangan:

- D : data yang digunakan
- A : atribut yang mempartisi D

Persamaan (4) digunakan untuk melakukan normalisasi terhadap *Information Gain* dengan menggunakan *Split Information*.

$$SplitInfo_A(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \times \log_2 \left(\frac{|D_j|}{|D|} \right) \quad (4)$$

Keterangan:

- D : data yang digunakan
- A : atribut yang mempartisi D
- v : jumlah partisi
- $\frac{|D_j|}{|D|}$: probabilitas dari partisi j terhadap D

Persamaan (5) digunakan untuk menghitung *Gain Ratio*, dimana *Gain Ratio* dipakai sebagai ukuran untuk menentukan atribut mana yang dipilih.

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)} \quad (5)$$

Keterangan:

- D : data yang digunakan
- A : atribut yang mempartisi D

Gain Ratio adalah pengembangan dari *Information Gain* yang semula bias terhadap atribut yang bernilai banyak. Meskipun begitu, *Gain Ratio* memiliki kekurangan, yaitu cenderung membentuk cabang yang tidak seimbang dimana satu partisi lebih kecil dibanding yang lainnya [6].

E. Algoritma k-Means Clustering

Algoritma k-Means *clustering* merupakan salah satu algoritma *clustering* yang menggunakan metode *partitioning* [6]. Di dalam algoritma k-Means *clustering*, pusat dari tiap *cluster* diwakili oleh rata-rata nilai dari setiap objek yang ada di dalam *cluster* tersebut. Alur algoritma k-Means *clustering* adalah sebagai berikut:

Input:

- k : jumlah *cluster*
- D : sebuah data set yang memiliki objek sejumlah n

Output:

- *Cluster* sejumlah k dengan objeknya masing-masing

Metode:

1. Memilih objek sebanyak k dari D untuk dijadikan *centroid* awal
2. Menempatkan setiap objek ke dalam *cluster* dimana objek paling mirip dengan *centroid* pada *cluster* tersebut
3. Menghitung rata-rata nilai objek yang ada di setiap *cluster* dan menjadikannya nilai *centroid* yang baru

- Mengulangi langkah nomor 2 dan 3 sampai tidak terjadi perubahan

Algoritma k-Means memiliki beberapa variasi, dimana perbedaan dapat terletak pada inisialisasi *centroid* awal, penghitungan kemiripan, dan strategi dalam menghitung nilai rata-rata sebuah cluster [6]. Dalam penghitungan kemiripan antar dua objek dapat dilakukan dengan menghitung jarak dari objek satu dengan yang lainnya, salah satunya adalah dengan menggunakan *Euclidian distance* yang terdapat pada (6). Jarak yang pendek menandakan bahwa objek yang satu dengan yang lainnya memiliki kemiripan, begitu juga sebaliknya, jarak yang panjang menandakan bahwa objek-objek tersebut tidak mirip.

$$Jarak = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (6)$$

Keterangan:

- x1 : nilai atribut pertama objek pertama
- x2 : nilai atribut pertama objek kedua
- y1 : nilai atribut kedua objek pertama
- y2 : nilai atribut kedua objek kedua

Algoritma ini memiliki kekurangan, yaitu sensitif terhadap *outlier* karena objek tersebut letaknya jauh dari data mayoritas, sehingga dapat mengubah nilai rata-rata *cluster* ke dalam nilai yang sangat berbeda. Hal ini dapat menyebabkan kesalahan dalam penempatan objek dalam *cluster* [6].

F. Cross-tabulation

Cross-tabulation atau tabulasi silang adalah metode yang mentabulasikan beberapa variabel ke dalam suatu matriks yang hasilnya disajikan dalam suatu tabel dengan variabel yang disusun dalam baris dan kolom. Tabel 1 menunjukkan *cross-tabulation* antara data aktual dengan data prediksi.

TABEL 1
CROSS-TABULATION ANTARA DATA AKTUAL DENGAN DATA PREDIKSI

Aktual	Prediksi			Total
	Positif	Negatif		
Positif	P ₁₁	P ₁₂	P ₁₊	
Negatif	P ₂₁	P ₂₂	P ₂₊	
Total	P ₊₁	P ₊₂	n	

Keterangan:

- P₁₁ : data aktual positif dan diprediksi positif
- P₁₂ : data aktual positif dan diprediksi negatif
- P₂₁ : data aktual negatif dan diprediksi positif
- P₂₂ : data aktual negatif dan diprediksi negatif
- P₁₊ : P₁₁ + P₁₂
- P₂₊ : P₂₁ + P₂₂
- P₊₁ : P₁₁ + P₂₁
- P₊₂ : P₁₂ + P₂₂
- n : jumlah total data

Akurasi dari *cross-tabulation* dapat dihitung dengan menjumlahkan data yang aktualnya sama dengan hasil prediksinya. Persamaan (7) menunjukkan cara menghitung akurasi dari *cross-tabulation* [11].

$$Akurasi (\%) = \frac{P_{11} + P_{22}}{n} \quad (7)$$

Keterangan:

- P₁₁ : data aktual positif dan diprediksi positif
- P₂₂ : data aktual negatif dan diprediksi negatif
- n : jumlah total data

III. METODOLOGI PENELITIAN

A. Sumber Data

Data set yang digunakan adalah data mahasiswa FTI UKDW angkatan 2008-2017 sebanyak 2428. Data set dibagi menjadi 2, yaitu data mahasiswa angkatan 2008-2016 sebagai data latih, dan data mahasiswa angkatan 2017 sebagai data uji. Data set akan dibagi berdasarkan jalur penerimaannya, yaitu data mahasiswa jalur prestasi, dan data mahasiswa jalur non-prestasi. Tabel 2 menunjukkan jumlah mahasiswa FTI UKDW angkatan 2008-2017.

Data mahasiswa jalur prestasi memiliki variabel NIM, kategori sekolah (SMA atau SMK), status sekolah (Negeri atau Swasta), lokasi sekolah (Jawa atau luar Jawa), dan level kemampuan bahasa Inggris (level 1, 2, 3, atau ESP). Sedangkan untuk data mahasiswa jalur non-prestasi memiliki variabel NIM, kategori sekolah (SMA atau SMK), status sekolah (Negeri atau Swasta), lokasi sekolah (Jawa atau luar Jawa), level kemampuan bahasa Inggris (level 1, 2, 3, atau ESP), kemampuan Spatial, Verbal, Numerik, dan Analogi.

TABEL 2
JUMLAH MAHASISWA FTI UKDW ANGKATAN 2008-2017

No	Angkatan	Jalur Prestasi	Jalur Non-prestasi	Total
1	2008	63	305	368
2	2009	11	249	260
3	2010	55	209	264
4	2011	144	107	251
5	2012	125	119	244
6	2013	125	80	205
7	2014	90	81	171
8	2015	193	61	254
9	2016	99	111	210
10	2017	102	99	201
Total		1007	1421	2428

Gambar 4 menunjukkan rancangan basis data yang akan digunakan pada penelitian ini. Data mahasiswa akan disimpan ke dalam tabel 'Data Mahasiswa'. Tabel tersebut terdiri dari nim, prodi, kategori, status, lokasi, level, numerik, verbal, spasial, analogi, jalur, angkatan, ips1, nprodi, nkategori, nstatus, nlokasi, nlevel, dan njalur dengan atribut nim sebagai primary key. Atribut nprodi, nkategori, nstatus, nlokasi, nlevel, dan njalur, merupakan atribut numerik dari atribut prodi, kategori, status, lokasi, level, dan jalur.

DATA_MAHASISWA	
PK	nim
	prodi
	kategori
	status
	lokasi
	level
	numerik
	verbal
	spasial
	analogi
	jalur
	angkatan
	ips1
	nprodi
	nkategori
	nstatus
	nlokasi
	nlevel
	njalur

Gambar 4. Rancangan basis data.

B. Rancangan Pengujian

Pengujian akan dilakukan pada data mahasiswa FTI UKDW tahun 2017 sejumlah 201. Pengujian akan dilakukan ke dalam 24 skenario yang dapat dilihat pada Tabel 3. Skenario menggunakan jalur prestasi sejumlah 8 skenario (skenario 1-8), sedangkan skenario menggunakan jalur non-prestasi sejumlah 16 skenario (skenario 9-24). Skenario menggunakan algoritma C4.5 sejumlah 12 skenario (skenario 1-4, dan 9-16), dan skenario menggunakan algoritma k-Means sejumlah 12 skenario (skenario 5-8, dan 17-24). Skenario menggunakan data seimbang sejumlah 12 skenario (skenario 1, 2, 5, 6, 9-12, dan 17-20), dan skenario menggunakan data tidak seimbang sejumlah 12 skenario (skenario 3, 4, 7, 8, 13-16, dan 21-24). Skenario menggunakan binning sejumlah 8 skenario (skenario 9, 10, 13, 14, 17, 18, 21, 22), dan skenario tidak menggunakan binning sejumlah 16 skenario (skenario 1-8, 11, 12, 15, 16, 19, 20, 23, 24). Skenario menggunakan 2 kategori IPS sejumlah 12 skenario (skenario 1, 3, 5, 7, 9, 11), dan skenario menggunakan 3 kategori IPS sejumlah 12 skenario (skenario 2, 4, 6, 8, 10, 12).

TABEL 3
SKENARIO PENGUJIAN SISTEM

Nama Skenario	Jalur	Algoritma	Seimbang	Binning	Kategori IPS
Skenario 1	Prestasi	C4.5	Ya	Tidak	2
Skenario 2	Prestasi	C4.5	Ya	Tidak	3
Skenario 3	Prestasi	C4.5	Tidak	Tidak	2
Skenario 4	Prestasi	C4.5	Tidak	Tidak	3
Skenario 5	Prestasi	k-Means	Ya	Tidak	2
Skenario 6	Prestasi	k-Means	Ya	Tidak	3
Skenario 7	Prestasi	k-Means	Tidak	Tidak	2
Skenario 8	Prestasi	k-Means	Tidak	Tidak	3
Skenario 9	Non-prestasi	C4.5	Ya	Ya	2

Skenario 10	Non-prestasi	C4.5	Ya	Ya	3
Skenario 11	Non-prestasi	C4.5	Ya	Tidak	2
Skenario 12	Non-prestasi	C4.5	Ya	Tidak	3
Skenario 13	Non-prestasi	C4.5	Tidak	Ya	2
Skenario 14	Non-prestasi	C4.5	Tidak	Ya	3
Skenario 15	Non-prestasi	C4.5	Tidak	Tidak	2
Skenario 16	Non-prestasi	C4.5	Tidak	Tidak	3
Skenario 17	Non-prestasi	k-Means	Ya	Ya	2
Skenario 18	Non-prestasi	k-Means	Ya	Ya	3
Skenario 19	Non-prestasi	k-Means	Ya	Tidak	2
Skenario 20	Non-prestasi	k-Means	Ya	Tidak	3
Skenario 21	Non-prestasi	k-Means	Tidak	Ya	2
Skenario 22	Non-prestasi	k-Means	Tidak	Ya	3
Skenario 23	Non-prestasi	k-Means	Tidak	Tidak	2
Skenario 24	Non-prestasi	k-Means	Tidak	Tidak	3

Binning hanya akan digunakan untuk skenario dengan jalur non-prestasi, karena binning dilakukan pada atribut numerik, sementara jalur prestasi tidak memiliki atribut numerik. Atribut yang akan dilakukan binning adalah atribut numerik, verbal, spasial, dan analogi, yang memiliki rentang nilai antara 0 sampai 200. Tabel 4 menunjukkan nama bin beserta rentang nilainya.

TABEL 4
BINNING ATRIBUT NUMERIK

Nama Bin	Atribut Numerik
A	0-40
B	41-80
C	81-120
D	121-160
E	161-200

Skenario dengan kategori IPS sejumlah 2 kategori dapat dilihat pembagiannya di Tabel 5. Kategori IPS dibagi menjadi kategori “Rendah”, dan kategori “Tinggi”. Skenario dengan kategori IPS sejumlah 3 kategori dapat dilihat pembagiannya di Tabel 6. Kategori IPS dibagi menjadi kategori “Rendah”, “Sedang”, dan “Tinggi”. Pengelompokan kategori IPS “Rendah” didasarkan pada nilai minimum IPK mahasiswa FTI UKDW untuk lulus, yaitu 2,25.

TABEL 5
PEMBAGIAN 2 KATEGORI IPS

Nama Kategori	Nilai IPS
Rendah	0-2,25
Tinggi	2,26-4

TABEL 6
PEMBAGIAN 3 KATEGORI IPS

Nama Kategori	Nilai IPS
Rendah	0-2,25
Sedang	2,26-3
Tinggi	3,01-4

Dari skenario pengujian pada Tabel 3, diharapkan hasil dari pengujian berupa akurasi prediksi setiap skenario. Akurasi akan dihitung dengan menggunakan Persamaan (7), dimana nilai akurasi semakin mendekati nilai 100% semakin baik hasilnya.

IV. HASIL DAN ANALISIS

Setelah dilakukan pengujian menggunakan skenario yang telah dibentuk, penghitungan hasil pengujian dinyatakan dalam persentase akurasi prediksi. Rata-rata dan nilai akurasi untuk skenario 1 sampai 24 dapat dilihat pada Tabel 7. Nilai akurasi tertinggi terdapat pada skenario 3 sebesar 77,45%, sedangkan nilai akurasi terendah terdapat pada skenario 24 sebesar 31,31%. Rata-rata nilai akurasi setiap skenario adalah 44,39%.

TABEL 7
HASIL AKURASI SETIAP SKENARIO

Nama Skenario	Algoritma	Seimbang	Binnig	Kategori IPS	Akurasi
Skenario 1	C4.5	Ya	Tidak	2	65,69 %
Skenario 2	C4.5	Ya	Tidak	3	44,12 %
Skenario 3	C4.5	Tidak	Tidak	2	77,45 %
Skenario 4	C4.5	Tidak	Tidak	3	48,04 %
Skenario 5	k-Means	Ya	Tidak	2	60,78 %
Skenario 6	k-Means	Ya	Tidak	3	47,06 %
Skenario 7	k-Means	Tidak	Tidak	2	60,78 %
Skenario 8	k-Means	Tidak	Tidak	3	38,24 %
Skenario 9	C4.5	Ya	Ya	2	46,46 %
Skenario 10	C4.5	Ya	Ya	3	38,38 %
Skenario 11	C4.5	Ya	Tidak	2	39,39 %
Skenario 12	C4.5	Ya	Tidak	3	33,33 %
Skenario 13	C4.5	Tidak	Ya	2	46,46 %
Skenario 14	C4.5	Tidak	Ya	3	35,35 %
Skenario 15	C4.5	Tidak	Tidak	2	45,45 %
Skenario 16	C4.5	Tidak	Tidak	3	32,32 %
Skenario 17	k-Means	Ya	Ya	2	42,42 %
Skenario 18	k-Means	Ya	Ya	3	35,35 %

Skenario 19	k-Means	Ya	Tidak	2	43,43 %
Skenario 20	k-Means	Ya	Tidak	3	34,34 %
Skenario 21	k-Means	Tidak	Ya	2	41,41 %
Skenario 22	k-Means	Tidak	Ya	3	34,34 %
Skenario 23	k-Means	Tidak	Tidak	2	43,43 %
Skenario 24	k-Means	Tidak	Tidak	3	31,31 %
Rata-rata					44,39 %

Rata-rata dan nilai akurasi untuk skenario menggunakan algoritma C4.5 dapat dilihat pada Tabel 8. Nilai akurasi tertinggi terdapat pada skenario 3 sebesar 77,45%, sedangkan nilai akurasi terendah terdapat pada skenario 16 sebesar 32,32%. Rata-rata nilai akurasi setiap skenario adalah 45,28%.

TABEL 8
AKURASI SKENARIO DENGAN ALGORITMA C4.5 DAN RATA-RATANYA

Nama Skenario	Algoritma	Seimbang	Binnig	Kategori IPS	Akurasi
Skenario 1	C4.5	Ya	Tidak	2	65,69 %
Skenario 2	C4.5	Ya	Tidak	3	44,12 %
Skenario 3	C4.5	Tidak	Tidak	2	77,45 %
Skenario 4	C4.5	Tidak	Tidak	3	48,04 %
Skenario 9	C4.5	Ya	Ya	2	46,46 %
Skenario 10	C4.5	Ya	Ya	3	38,38 %
Skenario 11	C4.5	Ya	Tidak	2	39,39 %
Skenario 12	C4.5	Ya	Tidak	3	33,33 %
Skenario 13	C4.5	Tidak	Ya	2	46,46 %
Skenario 14	C4.5	Tidak	Ya	3	35,35 %
Skenario 15	C4.5	Tidak	Tidak	2	45,45 %
Skenario 16	C4.5	Tidak	Tidak	3	32,32 %
Rata-rata					46,04 %

Rata-rata dan nilai akurasi untuk skenario menggunakan algoritma k-Means dapat dilihat pada Tabel 9. Nilai akurasi tertinggi terdapat pada skenario 5 dan 7 sebesar 60,78%, sedangkan nilai akurasi terendah terdapat pada skenario 20 dan 24 sebesar 31,31%. Rata-rata nilai akurasi setiap skenario adalah 42,74%.

TABEL 9
AKURASI SKENARIO DENGAN ALGORITMA K-MEANS DAN RATA-RATANYA

Nama Skenario	Algoritma	Seimbang	Binnig	Kategori IPS	Akurasi
Skenario 5	k-Means	Ya	Tidak	2	60,78 %
Skenario 6	k-Means	Ya	Tidak	3	47,06 %
Skenario 7	k-Means	Tidak	Tidak	2	60,78 %
Skenario 8	k-Means	Tidak	Tidak	3	38,24 %
Skenario 17	k-Means	Ya	Ya	2	42,42 %
Skenario 18	k-Means	Ya	Ya	3	35,35 %
Skenario 19	k-Means	Ya	Tidak	2	43,43 %
Skenario 20	k-Means	Ya	Tidak	3	34,34 %
Skenario 21	k-Means	Tidak	Ya	2	41,41 %
Skenario 22	k-Means	Tidak	Ya	3	34,34 %
Skenario 23	k-Means	Tidak	Tidak	2	43,43 %
Skenario 24	k-Means	Tidak	Tidak	3	31,31 %
Rata-rata					42,74 %

Rata-rata dan nilai akurasi untuk skenario jalur prestasi dapat dilihat pada Tabel 10. Nilai akurasi tertinggi terdapat pada skenario 3 sebesar 77,45%, sedangkan nilai akurasi terendah terdapat pada skenario 8 sebesar 38,24%. Rata-rata nilai akurasi setiap skenario adalah 55,27%.

TABEL 10
AKURASI SKENARIO JALUR PRESTASI DAN RATA-RATANYA

Nama Skenario	Algoritma	Seimbang	Binnig	Kategori IPS	Akurasi
Skenario 1	C4.5	Ya	Tidak	2	65,69 %
Skenario 2	C4.5	Ya	Tidak	3	44,12 %
Skenario 3	C4.5	Tidak	Tidak	2	77,45 %
Skenario 4	C4.5	Tidak	Tidak	3	48,04 %
Skenario 5	k-Means	Ya	Tidak	2	60,78 %
Skenario 6	k-Means	Ya	Tidak	3	47,06 %
Skenario 7	k-Means	Tidak	Tidak	2	60,78 %
Skenario 8	k-Means	Tidak	Tidak	3	38,24 %
Rata-rata					55,27 %

Rata-rata dan nilai akurasi untuk skenario jalur non-prestasi dapat dilihat pada Tabel 11. Nilai akurasi tertinggi terdapat pada skenario 9 dan 13 sebesar 46,46%, sedangkan nilai akurasi terendah terdapat pada skenario 24 sebesar

31,31%. Rata-rata nilai akurasi setiap skenario adalah 38,95%.

TABEL 11
AKURASI SKENARIO JALUR NON-PRESTASI DAN RATA-RATANYA

Nama Skenario	Algoritma	Seimbang	Binnig	Kategori IPS	Akurasi
Skenario 9	C4.5	Ya	Ya	2	46,46 %
Skenario 10	C4.5	Ya	Ya	3	38,38 %
Skenario 11	C4.5	Ya	Tidak	2	39,39 %
Skenario 12	C4.5	Ya	Tidak	3	33,33 %
Skenario 13	C4.5	Tidak	Ya	2	46,46 %
Skenario 14	C4.5	Tidak	Ya	3	35,35 %
Skenario 15	C4.5	Tidak	Tidak	2	45,45 %
Skenario 16	C4.5	Tidak	Tidak	3	32,32 %
Skenario 17	k-Means	Ya	Ya	2	42,42 %
Skenario 18	k-Means	Ya	Ya	3	35,35 %
Skenario 19	k-Means	Ya	Tidak	2	43,43 %
Skenario 20	k-Means	Ya	Tidak	3	34,34 %
Skenario 21	k-Means	Tidak	Ya	2	41,41 %
Skenario 22	k-Means	Tidak	Ya	3	34,34 %
Skenario 23	k-Means	Tidak	Tidak	2	43,43 %
Skenario 24	k-Means	Tidak	Tidak	3	31,31 %
Rata-rata					38,95 %

Rata-rata dan nilai akurasi untuk skenario data latih seimbang dapat dilihat pada Tabel 12. Nilai akurasi tertinggi terdapat pada skenario 1 sebesar 65,69%, sedangkan nilai akurasi terendah terdapat pada skenario 12 sebesar 33,33%. Rata-rata nilai akurasi setiap skenario adalah 44,23%.

TABEL 12
AKURASI SKENARIO DATA LATIH SEIMBANG DAN RATA-RATANYA

Nama Skenario	Algoritma	Seimbang	Binnig	Kategori IPS	Akurasi
Skenario 1	C4.5	Ya	Tidak	2	65,69 %
Skenario 2	C4.5	Ya	Tidak	3	44,12 %
Skenario 5	k-Means	Ya	Tidak	2	60,78 %
Skenario 6	k-Means	Ya	Tidak	3	47,06 %
Skenario 9	C4.5	Ya	Ya	2	46,46 %
Skenario 10	C4.5	Ya	Ya	3	38,38 %
Skenario 11	C4.5	Ya	Tidak	2	39,39 %

Skenario 12	C4.5	Ya	Tidak	3	33,33 %
Skenario 17	k-Means	Ya	Ya	2	42,42 %
Skenario 18	k-Means	Ya	Ya	3	35,35 %
Skenario 19	k-Means	Ya	Tidak	2	43,43 %
Skenario 20	k-Means	Ya	Tidak	3	34,34 %
Rata-rata					44,23 %

Rata-rata dan nilai akurasi untuk skenario data latih tidak seimbang dapat dilihat pada Tabel 13. Nilai akurasi tertinggi terdapat pada skenario 3 sebesar 77,45%, sedangkan nilai akurasi terendah terdapat pada skenario 24 sebesar 31,31%. Rata-rata nilai akurasi setiap skenario adalah 44,55%.

TABEL 13
AKURASI SKENARIO DATA LATIH TIDAK SEIMBANG DAN RATA-RATANYA

Nama Skenario	Algoritma	Seimbang	Binning	Kategori IPS	Akurasi
Skenario 3	C4.5	Tidak	Tidak	2	77,45 %
Skenario 4	C4.5	Tidak	Tidak	3	48,04 %
Skenario 7	k-Means	Tidak	Tidak	2	60,78 %
Skenario 8	k-Means	Tidak	Tidak	3	38,24 %
Skenario 13	C4.5	Tidak	Ya	2	46,46 %
Skenario 14	C4.5	Tidak	Ya	3	35,35 %
Skenario 15	C4.5	Tidak	Tidak	2	45,45 %
Skenario 16	C4.5	Tidak	Tidak	3	32,32 %
Skenario 21	k-Means	Tidak	Ya	2	41,41 %
Skenario 22	k-Means	Tidak	Ya	3	34,34 %
Skenario 23	k-Means	Tidak	Tidak	2	43,43 %
Skenario 24	k-Means	Tidak	Tidak	3	31,31 %
Rata-rata					44,55 %

Rata-rata dan nilai akurasi untuk skenario menggunakan *binning* dapat dilihat pada Tabel 14. Nilai akurasi tertinggi terdapat pada skenario 9 dan 13 sebesar 46,46%, sedangkan nilai akurasi terendah terdapat pada skenario 22 sebesar 34,34%. Rata-rata nilai akurasi setiap skenario adalah 40,02%.

TABEL 14
AKURASI SKENARIO DENGAN BINNING DAN RATA-RATANYA

Nama Skenario	Algoritma	Seimbang	Binning	Kategori IPS	Akurasi
Skenario 9	C4.5	Ya	Ya	2	46,46 %

Skenario 10	C4.5	Ya	Ya	3	38,38 %
Skenario 13	C4.5	Tidak	Ya	2	46,46 %
Skenario 14	C4.5	Tidak	Ya	3	35,35 %
Skenario 17	k-Means	Ya	Ya	2	42,42 %
Skenario 18	k-Means	Ya	Ya	3	35,35 %
Skenario 21	k-Means	Tidak	Ya	2	41,41 %
Skenario 22	k-Means	Tidak	Ya	3	34,34 %
Rata-rata					40,02 %

Rata-rata dan nilai akurasi untuk skenario tidak menggunakan *binning* dapat dilihat pada Tabel 15. Nilai akurasi tertinggi terdapat pada skenario 3 sebesar 77,45%, sedangkan nilai akurasi terendah terdapat pada skenario 24 sebesar 31,31%. Rata-rata nilai akurasi setiap skenario adalah 46,57%.

TABEL 15
AKURASI SKENARIO TIDAK BINNING DAN RATA-RATANYA

Nama Skenario	Algoritma	Seimbang	Binning	Kategori IPS	Akurasi
Skenario 1	C4.5	Ya	Tidak	2	65,69 %
Skenario 2	C4.5	Ya	Tidak	3	44,12 %
Skenario 3	C4.5	Tidak	Tidak	2	77,45 %
Skenario 4	C4.5	Tidak	Tidak	3	48,04 %
Skenario 5	k-Means	Ya	Tidak	2	60,78 %
Skenario 6	k-Means	Ya	Tidak	3	47,06 %
Skenario 7	k-Means	Tidak	Tidak	2	60,78 %
Skenario 8	k-Means	Tidak	Tidak	3	38,24 %
Skenario 11	C4.5	Ya	Tidak	2	39,39 %
Skenario 12	C4.5	Ya	Tidak	3	33,33 %
Skenario 15	C4.5	Tidak	Tidak	2	45,45 %
Skenario 16	C4.5	Tidak	Tidak	3	32,32 %
Skenario 19	k-Means	Ya	Tidak	2	43,43 %
Skenario 20	k-Means	Ya	Tidak	3	34,34 %
Skenario 23	k-Means	Tidak	Tidak	2	43,43 %
Skenario 24	k-Means	Tidak	Tidak	3	31,31 %
Rata-rata					46,57 %

Rata-rata dan nilai akurasi untuk skenario dengan 2 kategori IPS dapat dilihat pada Tabel 16. Nilai akurasi tertinggi terdapat pada skenario 3 sebesar 77,45%,

sedangkan nilai akurasi terendah terdapat pada skenario 11 sebesar 39,39%. Rata-rata nilai akurasi setiap skenario adalah 51,1%.

TABEL 16
AKURASI SKENARIO DENGAN 2 KATEGORI IPS DAN RATA-RATANYA

Nama Skenario	Algoritma	Seimbang	Binning	Kategori IPS	Akurasi
Skenario 1	C4.5	Ya	Tidak	2	65,69 %
Skenario 3	C4.5	Tidak	Tidak	2	77,45 %
Skenario 5	k-Means	Ya	Tidak	2	60,78 %
Skenario 7	k-Means	Tidak	Tidak	2	60,78 %
Skenario 9	C4.5	Ya	Ya	2	46,46 %
Skenario 11	C4.5	Ya	Tidak	2	39,39 %
Skenario 13	C4.5	Tidak	Ya	2	46,46 %
Skenario 15	C4.5	Tidak	Tidak	2	45,45 %
Skenario 17	k-Means	Ya	Ya	2	42,42 %
Skenario 19	k-Means	Ya	Tidak	2	43,43 %
Skenario 21	k-Means	Tidak	Ya	2	41,41 %
Skenario 23	k-Means	Tidak	Tidak	2	43,43 %
Rata-rata					51,10 %

Rata-rata dan nilai akurasi untuk skenario dengan 3 kategori IPS dapat dilihat pada Tabel 17. Nilai akurasi tertinggi terdapat pada skenario 4 sebesar 48,04%, sedangkan nilai akurasi terendah terdapat pada skenario 24 sebesar 31,31%. Rata-rata nilai akurasi setiap skenario adalah 37,68%.

TABEL 17
AKURASI SKENARIO DENGAN 2 KATEGORI IPS DAN RATA-RATANYA

Nama Skenario	Algoritma	Seimbang	Binning	Kategori IPS	Akurasi
Skenario 2	C4.5	Ya	Tidak	3	44,12 %
Skenario 4	C4.5	Tidak	Tidak	3	48,04 %
Skenario 6	k-Means	Ya	Tidak	3	47,06 %
Skenario 8	k-Means	Tidak	Tidak	3	38,24 %
Skenario 10	C4.5	Ya	Ya	3	38,38 %
Skenario 12	C4.5	Ya	Tidak	3	33,33 %
Skenario 14	C4.5	Tidak	Ya	3	35,35 %
Skenario 16	C4.5	Tidak	Tidak	3	32,32 %
Skenario 18	k-Means	Ya	Ya	3	35,35 %
Skenario 20	k-Means	Ya	Tidak	3	34,34 %

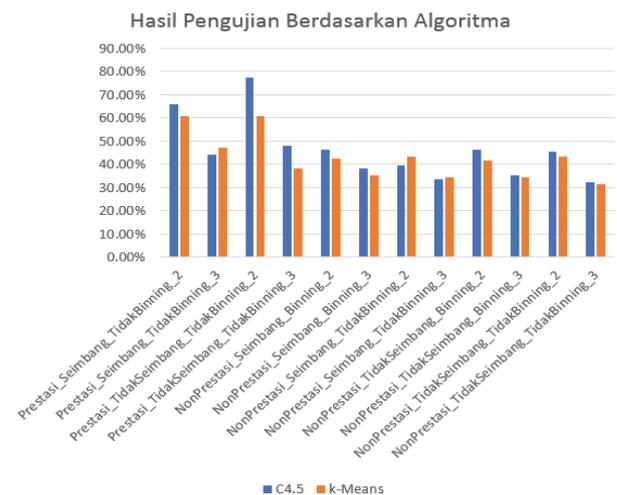
Skenario 22	k-Means	Tidak	Ya	3	34,34 %
Skenario 24	k-Means	Tidak	Tidak	3	31,31 %
Rata-rata					37,68 %

Grafik hasil pengujian berdasarkan skenario dapat dilihat pada Gambar 5. Secara keseluruhan, skenario 3 memiliki hasil terbaik, dimana skenario 3 merupakan skenario dengan jalur prestasi, algoritma C4.5, data tidak seimbang, tidak binning, dan kategori IPS 3. Skenario 3 mendapatkan hasil akurasi sebesar 77,45%. Skenario dengan hasil terendah dimiliki oleh skenario 24, dimana skenario 24 merupakan skenario dengan jalur non-prestasi, algoritma k-Means, data tidak seimbang, tidak binning, dan kategori IPS 3. Skenario 24 mendapatkan hasil akurasi sebesar 31,31%.



Gambar 5. Grafik hasil pengujian berdasarkan skenario.

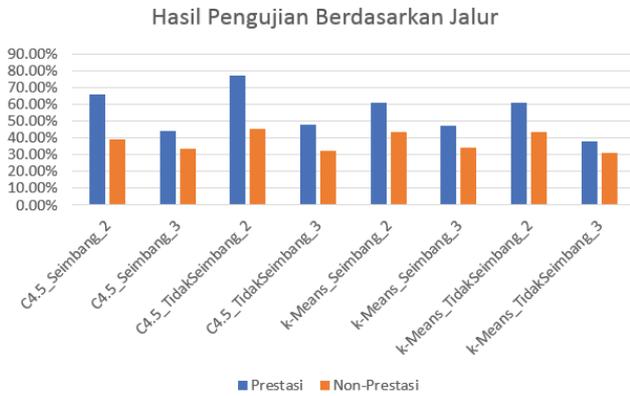
Grafik hasil pengujian berdasarkan algoritma dapat dilihat pada Gambar 6. Rata-rata nilai akurasi dengan menggunakan algoritma C4.5 lebih tinggi daripada algoritma k-Means. Secara keseluruhan dengan melihat skenario dengan atribut yang sama, algoritma C4.5 memiliki hasil yang lebih baik jika dibandingkan dengan algoritma k-Means.



Gambar 6. Grafik hasil pengujian berdasarkan algoritma.

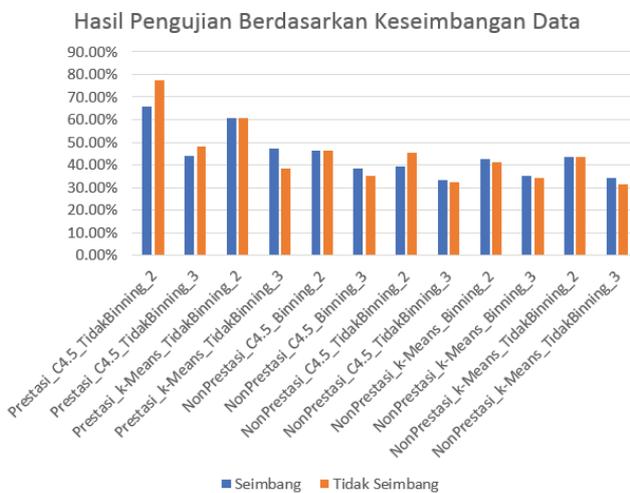
Grafik hasil pengujian berdasarkan jalur penerimaan dapat dilihat pada Gambar 7. Data pada grafik tersebut merupakan data uji yang tidak menggunakan binning dan dibagi berdasarkan jalur penerimaannya. Hasil pengujian dengan menggunakan jalur prestasi memiliki hasil yang lebih baik dibandingkan dengan jalur non-prestasi. Hal ini dapat dipengaruhi oleh perbedaan atribut yang dimiliki kedua jalur

tersebut, dimana jalur prestasi memiliki atribut kategori, status, lokasi, dan level ICE, sementara jalur non-prestasi memiliki atribut kategori, status, lokasi, level ICE, nilai numerik, verbal, spasial, dan analogi. Atribut nilai numerik, verbal, spasial, dan analogi membuat hasil akurasi prediksi menurun. Hal ini bisa saja karena adanya faktor eksternal yang membuat nilai-nilai tersebut menjadi tidak relevan dengan nilai IPS.



Gambar 7. Grafik hasil pengujian berdasarkan jalur.

Grafik hasil pengujian berdasarkan keseimbangan data dapat dilihat pada Gambar 8. Secara keseluruhan, skenario dengan data seimbang dan data tidak seimbang tidak memiliki perbedaan yang terlalu jauh. Rata-rata hasil prediksi data seimbang dan data tidak seimbang tidak berbeda jauh.



Gambar 8. Grafik hasil pengujian berdasarkan keseimbangan data.

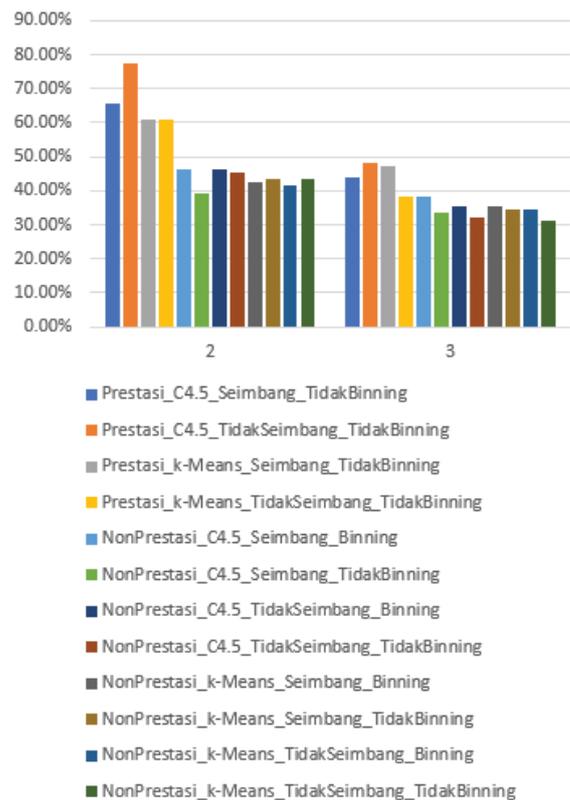
Grafik hasil pengujian berdasarkan *binning* dapat dilihat pada Gambar 9. Data pada grafik tersebut merupakan data uji yang jalur non-prestasi dan dibagi berdasarkan penggunaan *binning*-nya. Skenario dengan menggunakan *binning* memiliki hasil yang lebih baik jika dibandingkan dengan tidak menggunakan *binning*. Hal ini dapat dilihat jelas saat menggunakan skenario dengan algoritma C4.5, data seimbang, dan kategori 2. Secara keseluruhan, *binning* dengan 2 kategori IPS memiliki hasil yang lebih baik dari *binning* dengan 3 kategori IPS.



Gambar 9. Grafik hasil pengujian berdasarkan *binning*.

Grafik hasil pengujian berdasarkan jumlah kategori IPS dapat dilihat pada Gambar 10. Dari grafik tersebut, dapat terlihat jelas bahwa skenario dengan menggunakan 2 kategori IPS memiliki hasil yang lebih baik daripada 3 kategori IPS. Hasil terbaik dapat dilihat pada skenario dengan jalur prestasi, algoritma C4.5, tidak seimbang, dan tidak *binning*. Semakin sedikit jumlah kategori IPS yang digunakan akan memberikan hasil prediksi yang semakin baik.

Hasil Pengujian Berdasarkan Kategori



Gambar 10. Grafik hasil pengujian berdasarkan kategori IPS.

Tabel 18 menunjukkan perbandingan persentase nilai akurasi terbaik algoritma C4.5 (*Gain Ratio*), k-Means dibandingkan dengan C4.5 (*Gain*), CART, Regresi Logistik, K-Nearest Neighbor, dan Naïve Bayes Classifier. Jika dibandingkan dengan hasil dari algoritma C4.5 (*Gain*) dan CART milik Alverina [3], algoritma Regresi Logistik milik Santosa dan Chrismanto [2] & [4], K-Nearest Neighbor dan

Naïve Bayes Classifier milik Sari [5], hasil terbaik dimiliki oleh Alverina [3]. Algoritma C4.5 (*Gain Ratio*) berada di urutan ketiga algoritma terbaik untuk memprediksi jalur prestasi, dan urutan kelima algoritma terbaik untuk memprediksi jalur non-prestasi. Algoritma k-Means berada di urutan keenam algoritma terbaik untuk memprediksi jalur prestasi, dan urutan keenam algoritma terbaik untuk memprediksi jalur non-prestasi.

TABEL 18
PERBANDINGAN PERSENTASE NILAI AKURASI TERBAIK ALGORITMA C4.5 (GAIN RATIO), K-MEANS, C4.5 (GAIN), CART, REGRESI LOGISTIK, K-NEAREST NEIGHBOR, DAN NAÏVE BAYES CLASSIFIER

Algoritma	Jalur Prestasi	Jalur Non-Prestasi
C4.5 (<i>Gain Ratio</i>)	77,45%	46,46%
k-Means	60,78%	43,43%
C4.5 (<i>Gain</i>) [3]	86,86%	61,54%
CART [3]	86,86%	63,16%
Regresi Logistik [2], [4]	72%	60,7%
K-Nearest Neighbor [5]	61%	52%
Naïve Bayes Classifier [5]	55%	41%

V. KESIMPULAN

Kesimpulan yang didapat dari penelitian ini adalah:

1. Pada skenario jalur prestasi dengan data latihan sebanyak 905 data dan data uji sebanyak 102 data, algoritma C4.5 mendapat hasil terbaik sebesar 77,45%, sedangkan algoritma k-Means mendapat hasil terbaik sebesar 60,78%.
2. Pada skenario jalur non-prestasi dengan data latihan sebanyak 1322 data dan data uji sebanyak 99 data, algoritma C4.5 mendapat hasil terbaik sebesar 46,46%, sedangkan algoritma k-Means mendapat hasil terbaik sebesar 43,43%.
3. Berdasarkan skenario dengan jalur penerimaannya, algoritma C4.5 mendapatkan hasil yang lebih baik dibandingkan dengan algoritma k-Means.
4. Skenario dengan jalur prestasi memiliki hasil yang lebih baik dibandingkan dengan jalur non-prestasi, dimana rata-rata akurasi skenario jalur prestasi adalah 55,27% sedangkan rata-rata akurasi skenario jalur non-prestasi adalah 38,95%.
5. Skenario dengan 2 kategori IPS memiliki hasil yang lebih baik dibandingkan dengan 3 kategori IPS.

6. Angkatan 2017 FTI UKDW (Data Uji) menggunakan kurikulum yang berbeda dari angkatan-angkatan sebelumnya, sehingga mempengaruhi hasil prediksi.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Tuhan Yang Maha Esa sehingga penulis dapat menyelesaikan penelitian ini. Penulis juga mengucapkan terima kasih kepada Bapak R. Gunawan Santosa dan Bapak Antonius Rachmat Chrismanto yang telah membimbing dan mengarahkan penelitian ini sehingga penelitian ini dapat selesai. Penulis juga berterima kasih kepada teman-teman program studi Informatika UKDW angkatan 2015 yang selalu memberi semangat dan mendukung penulis.

DAFTAR PUSTAKA

- [1] Andi, "Pendaftaran Program Sarjana", *UKDW*, 11-Sep-2019. [Online]. Available: <https://www.ukdw.ac.id/pendaftaran-program-sarjana/>
- [2] R. G. Santosa and A. R. Chrismanto, "Perbandingan Akurasi Model Regresi Logistik untuk Prediksi Kategori IP Mahasiswa Jalur Prestasi dengan Non Jalur Prestasi," *Jurnal Teknik & Ilmu Komputer*, vol. 7, no. 25, pp. 107-121, Januari 2018.
- [3] D. Alverina, A. R. Chrismanto, and R. G. Santosa, "Perbandingan Algoritma C4.5 Dan Cart Dalam Memprediksi Kategori Indeks Prestasi Mahasiswa," *Jurnal Teknologi dan Sistem Komputer*, vol. 6, no. 2, April 2018. doi: 10.14710/jtsiskom.6.2.2018.76-83, [Online].
- [4] R. G. Santosa and A. R. Chrismanto, "Logistic Regression Model for Predicting First Semester Students GPA Category Based on High School Academic Achievement.," *Journal of Arts, Science & Commerce*, vol. VIII, no. 2(1), pp. 58-66, April 2017.
- [5] V. H. A. Sari, "Perbandingan Algoritma K-nearest Neighbor Dan Naïve Bayes Classifier Dalam Memprediksi Kategori Indeks Prestasi Mahasiswa". [Online]. Available: <http://sinta.ukdw.ac.id>
- [6] J. Han, M. Kamber, and J. Pei, *Data mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann, 2012.
- [7] B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali, "A comparative study of decision tree ID3 and C4.5," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 2. doi: 10.14569/specialissue.2014.040203. [Online].
- [8] O. J. Oyelade, O. O. Oladipupo, and I. C. Obagbuwa, "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance," *International Journal of Computer Science and Information Security*, vol. 7, pp. 292-295, 2010. [Online].
- [9] R. D. Ramadhani. "Data Mining Menggunakan Algoritma K-Means Clustering Untuk Menentukan Strategi Promosi Universitas Dian Nuswantoro". [Online]. Available: <http://eprints.dinus.ac.id>
- [10] M. Bramer, *Principles of Data Mining* (2nd ed.). London: Springer-Verlag London, 2013
- [11] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms* (2nd ed.). John Wiley & Sons, 2011.