

NASKAH PUBLIKASI (*MANUSCRIPTS*)

**OPTIMASI ALGORITMA *DECISION TREE* C4.5 MENGGUNAKAN
METODE CORRELATION-BASED FEATRUE SELECTION PADA
KLASIFIKASI NILAI MAHASISWA**

***OPTIMIZATION OF DECISION TREE C4.5 ALGORITHM USING
CORRELATION-BASED FEATURE SELECTION METHOD IN STUDENT
GRADE CLASSIFICATION***

Kamirul Hakim, Wawan Joko Pranoto, Faldi



DISUSUN OLEH :

KAMIRUL HAKIM

1911102441124

**PROGRAM STUDI S1 TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS MUHAMMADIYAH KALIMANTAN TIMUR
SAMARINDA
2023**

Naskah Publikasi (*Manuscripts*)

**Optimasi Algoritma *Decision Tree* C4.5 menggunakan Metode
Correlation-Based Feature Selection pada Klasifikasi Nilai
Mahasiswa**

***Optimization of Decision Tree C4.5 Algorithm using Correlation-
Based Feature Selection method in Student Grade Classification***

Kamirul Hakim, Wawan Joko Pranoto, Faldi



Disusun Oleh :

Kamirul Hakim

1911102441124

**PROGRAM STUDI S1 TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS MUHAMMADIYAH KALIMANTAN TIMUR
SAMARINDA
2023**

HALAMAN PENGESAHAN

**OPTIMASI ALGORITMA DECISION TREE C4.5 MENGGUNAKAN
METODE CORRELATION-BASED FEATURE SELECTION PADA
KLASIFIKASI NILAI MAHASISWA
STUDI KASUS: UNIVERSITAS MUHAMMADIYAH KALIMANTAN
TIMUR**

NASKAH PUBLIKASI

DISUSUN OLEH :

KAMIRUL HAKIM

1911102441124

Dosen Pembimbing

Wawan Joko Pranoto, S.Kom., M.TI
NIDN. 1102057701

Penguji

Faldi, S.Kom., M.TI
NIDN. 1121079101

Dekan

Prof. Dr. Sarjito, MT., Ph.D
NIDN. 0610116204

Ketua Program Studi

Asslia Johar Latipah, M.Cs
NIDN. 1124098902

OPTIMASI ALGORITMA *DECISION TREE* C4.5 MENGGUNAKAN METODE *CORRELATION-BASED FEATURE SELECTION* PADA KLASIFIKASI NILAI MAHASISWA

Kamirul Hakim¹, Wawan Joko Pranoto², Faldi³

^{1,2,3}Program Studi Teknik Informatika, Fakultas Sains Dan Teknologi, Universitas Muhammadiyah Kalimantan Timur

e-mail: ¹khamirul.hakim@gmail.com

ABSTRAK

Universitas X menerapkan sistem pembelajaran hybrid (daring dan luring) karena keterbatasan ruangan dan jumlah mahasiswa yang besar. Namun, pembelajaran jarak jauh memiliki dampak negatif terhadap hasil belajar mahasiswa. Data juga menunjukkan penurunan nilai rata-rata pada Mata Kuliah Dasar Umum Bahasa Arab di Universitas X, yang mungkin disebabkan oleh pembelajaran hybrid yang tidak efektif dan tantangan lainnya. Penelitian ini bertujuan untuk menganalisis faktor-faktor yang berpengaruh dalam klasifikasi nilai Bahasa Arab mahasiswa. Hasil penelitian menunjukkan bahwa Fitur-fitur yang berpengaruh signifikan dalam pengambilan keputusan adalah "Comments", "Tugas_2", "UTS", dan "Tugas_1", sementara jenis kelamin tidak memiliki pengaruh signifikan. Evaluasi model menggunakan validasi silang dengan rata-rata k-fold menunjukkan bahwa sebelum dilakukan seleksi fitur menggunakan metode CFS, akurasi mencapai 88%. Namun, setelah seleksi fitur menggunakan CSF, akurasi meningkat menjadi 92%. Hasil ini menunjukkan bahwa seleksi fitur dengan menggunakan metode CFS dapat signifikan meningkatkan kinerja klasifikasi dalam penilaian nilai Bahasa Arab mahasiswa.

Kata Kunci: Akurasi, Optimasi, Algoritma C4.5, CFS.

1. PENDAHULUAN

Mengukur hasil belajar sangatlah penting bagi mahasiswa. Hasil belajar mencakup perubahan perilaku seseorang setelah mengikuti pembelajaran. Untuk itu, mahasiswa harus bisa mengevaluasi perubahan perilaku siswa sebagai indikator dari hasil belajar yang telah dicapai. Dengan cara ini, evaluasi hasil belajar menjadi sangat relevan bagi mahasiswa karena dapat memberikan pengukuran yang akurat dan dapat dipercaya terhadap hasil belajar siswa [1]. Universitas X masih menerapkan sistem pembelajaran secara hybrid (daring dan luring) dengan membagi sesi berdasarkan nim ganjil dan genap, hal ini dilakukan karena banyaknya jumlah mahasiswa dan ruangan yang minim. Tidak menutup kemungkinan kegiatan pembelajaran seperti ini dapat mempengaruhi hasil pembelajaran mahasiswa. Pada pelaksanaan kegiatan pembelajaran jarak jauh terdapat survey yang menunjukkan bahwa 79,9% interaksi antara pengajar dan peserta didik tidak dilakukan, melainkan hanya pemberian tugas semata [2]. Berdasarkan data yang di peroleh untuk memperkuat penelitian ini, bahwa Mata Kuliah Dasar Umum (MKDU) Bahasa Arab pada angkatan tahun 2020 hingga 2021 di Universitas X mengalami penurunan.

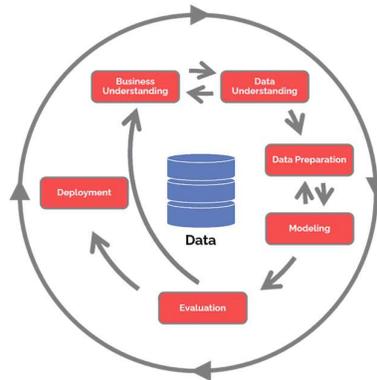
Pada tahun 2017, nilai rata-rata mahasiswa dalam mata kuliah ini adalah 57.77. Nilai ini mungkin mencerminkan kurangnya pengalaman atau pemahaman dalam mempelajari bahasa Arab bagi sebagian besar mahasiswa. Pada tahun 2018, terjadi peningkatan yang signifikan dalam nilai rata-rata mata kuliah Dasar Umum Bahasa Arab, yaitu menjadi 76.61. Ini mungkin terjadi karena adanya perubahan pendekatan pengajaran atau peningkatan kualitas dosen pengajar. Para mahasiswa pada tahun ini telah menunjukkan kemajuan yang baik dalam memahami dan menerapkan bahasa Arab. Pada tahun 2019, peningkatan kualitas pengajaran dan pendekatan pembelajaran yang terus ditingkatkan menghasilkan peningkatan nilai rata-rata yang signifikan menjadi 82.40. Nilai ini menunjukkan bahwa mahasiswa semakin memahami konsep dan aturan dasar bahasa Arab. Pada tahun 2020, nilai rata-rata mata kuliah Dasar Umum Bahasa Arab mengalami peningkatan lagi menjadi 84.80. Namun, pada tahun 2021, terjadi penurunan nilai rata-rata menjadi 74.18. Hal ini mungkin disebabkan oleh pembelajaran secara hybrid yang kurang maksimal dan berbagai tantangan lain yang dihadapi oleh mahasiswa.

Dalam permasalahan ini Universitas Muhammadiyah Kalimantan Timur harus memprioritaskan penanganan penurunan hasil belajar mahasiswa pada tahun 2020-2021 untuk memastikan kualitas pendidikan dan lulusan berkualitas. Solusi yang dapat diambil antara lain melakukan analisis faktor penyebab, meningkatkan kualitas dan motivasi pembelajaran, serta melakukan monitoring dan evaluasi terhadap kinerja mahasiswa. Evaluasi pembelajaran berfungsi untuk pengembangan program, perencanaan dan pengembangan kurikulum, serta untuk akreditasi program kelembagaan [3]. Untuk menjaga kualitas sistem pembelajaran hybrid, klasifikasi kinerja mahasiswa dalam perkuliahan hybrid perlu dilakukan di Universitas X.

2. METODE PENELITIAN

2.1. Alur Penelitian

Alur penelitian menggunakan tahap CRISP-DM dapat dilihat pada gambar 1.



Gambar 1. Alur Penelitian

2.2. Kinerja Mahasiswa

Mahasiswa yang mengikuti pendidikan di universitas atau institusi pendidikan tinggi memiliki performa akademis. Performa akademis dianggap sebagai salah satu indikator kualitas mahasiswa. Prestasi belajar mahasiswa didasarkan pada aktivitas perkuliahan yang mereka lakukan selama menempuh studi [4].

2.3. Data Mining

Data Mining merupakan sekumpulan data menjadi informasi yang memiliki potensi secara implisit (tidak nyata/ jelas) yang sebelumnya tidak diketahui, dengan menggunakan pranti otomatis atau semi otomatis, dari sejumlah besar data yang bertujuan untuk menemukan pola yang memiliki arti [5]

2.4. Klasifikasi

Klasifikasi merupakan pengelompokan berdasarkan hubungan antara variabel target. Contohnya pengelompokan dampak gempa bumi yaitu rusak berat, rusak berat dan tsunami, atau tidak terdampak [5]

2.5. Algoritma *Decision Tree* C4.5

Dalam Algoritma C4.5, terdapat beberapa elemen yang dikenal untuk menyelesaikan kasus, yaitu Entropi dan Gain. Entropi (S) dapat diartikan sebagai jumlah bit yang dibutuhkan untuk mengekstraksi kelas (+ atau -) dari sejumlah data acak dalam ruang sampel S. Entropi dapat dianggap sebagai ukuran kebutuhan bit untuk menyatakan suatu kelas, sehingga semakin kecil nilai entropi, semakin sedikit bit yang diperlukan untuk mengekstraksi kelas tersebut. Oleh karena itu, nilai entropi yang rendah lebih diutamakan dalam penggunaannya dalam mengekstraksi suatu kelas[6].

$$Entropy(S) = \sum_{i=1}^n - p_i * \log_2 p$$

Keterangan :

S : himpunan kasus

A : atribut

n : jumlah partisi

S : proporsi dari terhadap S

Gain (S,A) merupakan perolehan informasi dari atribut A relatif terhadap output data S. Perolehan informasi didapat dari output data atau variabel dependen S yang dikelompokkan berdasarkan atribut, dinotasikan dengan gain (S,A). Adapun rumus untuk mencari nilai gain yaitu :

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|s_i|}{|S|} * Entropy(s_i)$$

S = Himpunan kasus

A = Atribut

n = Jumlah atribut

|Si| = Jumlah partisi ke -i

|S| = jumlah kasus dalam S

2.6. Data Preprocessing

Data Preprocessing merupakan salah satu tahapan dalam melakukan mining data. Sebelum menuju ke tahap pemrosesan. Data mentah akan diolah terlebih dahulu. Data Preprocessing atau praproses data biasanya dilakukan melalui cara eliminasi data yang tidak sesuai. Selain itu dalam proses ini data akan diubah dalam bentuk yang akan lebih dipahami oleh sistem [7].

2.7. Confusion Matrix

Confusion Matrix adalah matrik yang berukuran N x N dimana N adalah jumlah kelas yang di prediksi. Jadi matrik ini cocok digunakan untuk permasalahan klasifikasi. *Confusion Matrix* menyajikan ringkasan semua hasil prediksi yang dihasilkan dengan membandingkan antara hasil prediksi dan hasil yang diharapkan [8].

Tabel 1 Perhitungan Confusion Matrix

	Positif	Negatif
Positif	TP	FP
Negatif	FN	TN

Pada confusion matrix tersebut ada 4 kolom dengan beberapa istilah yaitu :

1. TP (True Positif) berisi jumlah data points diberi label Yes yang sebenarnya bernilai Yes.
2. TN (True Negatif) berisi jumlah data points diberi label No yang sebenarnya bernilai No.
3. FP (False Positif) berisi jumlah data points diberi label No yang sebenarnya bernilai No.
4. FN (False Negatif) berisi jumlah data points diberi label No yang sebenarnya bernilai Yes.

Menghitung confusion matrix dapat menggunakan rumus sebagai berikut :

1. Accuracy (Akurasi)

Mengukur akurasi dengan menggunakan rumus dengan jumlah prediksi yang benar dibagi dengan total seluruh populasi.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

3. HASIL DAN PEMBAHASAN

3.1. Business Understanding (Pemahaman Bisnis)

Penelitian ini bertujuan untuk memprediksi kelulusan mahasiswa dalam mata kuliah Bahasa Arab di Universitas X. Data nilai mahasiswa diperoleh dari Mata Kuliah Dasar Umum (MKDU) dan Bagian Administrasi Akademik (BAA) dengan total 943 data mahasiswa dari angkatan 2020 dan 2021. Dalam tahap Business Understanding, peneliti mengidentifikasi variabel target sebagai kelulusan mahasiswa yang dibagi menjadi kategori "Lulus" dan "Tidak Lulus".

3.2. Data Understanding (Pemahaman Data)

Pada penelitian ini, data mengenai nilai Bahasa Arab dikumpulkan dari dua sumber yaitu Bagian Administrasi Akademik (BAA) dan Open Learning dari Mata Kuliah Dasar Umum (MKDU). Data nilai akhir

mahasiswa yang didapatkan dari BAA terdiri dari lima atribut. Atribut-atribut tersebut mencakup jenis kelamin, nomor induk mahasiswa (NIM), nama, nilai akhir, bobot, dan simbol dapat dilihat pada tabel 2.

Tabel 2 Data Atribut Dari BAA

No	Atribut	Keterangan
1	NIM	Nomor Induk Mahasiswa
2	Nama Mahasiswa	Nama Mahasiswa
3	Nilai Akhir	Nilai Akhir Mahasiswa
4	Bobot	Bobot nilai berdasarkan standar penilaian
5	Simbol	Skala penilaian

Tabel 3 Data Atribut Openlearning

No	Atribut	Keterangan
1	Profile name	Id mahasiswa pada sistem OpenLearning
2	Learner name	Nama mahasiswa
3	Learner email	Email mahasiswa
4	Enrolment ID	Id pendaftaran OpenLearning
5	Institution Mambership ID	Id anggota institusi
6	Enrolment date	Tanggal daftar
7	Completion date	Tanggal menyelesaikan mata kuliah
8	Time Spent on course	Lama waktu mahasiswa berada di mata kuliah
9	Progress	Persentase kemajuan mahasiswa
10	% Course completed	Persentase kemajuan mahasiswa menyelesaikan mata kuliah
11	Certificate ID	Id sertifikat
12	Comments	Banyaknya komentar mahasiswa selama perkuliahan
13	Kudos	Penghargaan
14	Enrolment cost	Biaya pendaftaran mata kuliah
15	Tugas	Nilai rata-rata tugas 1 – 10
16	Quis	Nilai quiz
17	UTS	Nilai Ujian Tengah Semester

3.3. Data Preparation (Pengolahan Data)

a. *Data Selection Dan Data Integration*

Langkah ini dimulai setelah data berhasil dikumpulkan dan dilakukan pengumpulan informasi yang relevan. Pada tahap ini, fokus diberikan pada pemilihan atribut yang akan digunakan untuk meningkatkan akurasi dan efisiensi dari algoritma yang akan diterapkan. Proses seleksi dan integrasi atribut dapat dibagi menjadi dua tahap. Tahap pertama adalah menghapus atribut yang tidak diperlukan seperti data BAA yaitu, Nama dan Bobot. Sedangkan atribut data OpenLearning yang dihapus adalah Profile name, Learner name, Learner email, Enrolment ID, Institution Membership ID, Enrolment date, Completion date, Progress, dan Certificate ID. Tahap seleksi atribut bertujuan untuk memilih atribut-atribut yang paling relevan dan memberikan kontribusi signifikan terhadap hasil akhir. Berikut adalah hasil seleksi dan integrasi tahap pertama.

Tabel 4 Data Seleksi Pertama

Jenis_Kelamin	Time_spent_on_course	Progress	Course_completed	Comments	Tugas_1	Tugas_2	UTS	Simbol
P	4 Hrs 45 Mins	98,37	98,37	8	71	73	74	A
P	5 Mins	59,80	59,80	5	61	77	75	AB
L	38 Mins	98,37	98,37	7	71	81	70	A
L	2 Hrs 27 Mins	98,37	98,37	26	60	75	74	A

Pada tahap kedua melakukan seleksi fitur menggunakan metode *Correlation-Based Feature Selection* CFS untuk mengurangi dimensi atribut dan meningkatkan akurasi serta efisiensi algoritma yang akan diterapkan. Melalui CFS, akan dianalisis korelasi antara atribut-atribut dengan label target. Proses ini akan mengidentifikasi atribut-atribut yang memiliki pengaruh signifikan terhadap label target. Analisis korelasi ini membantu memilih atribut-atribut yang paling relevan dan memiliki pengaruh yang signifikan untuk digunakan dalam algoritma berikutnya. Berikut adalah hasil seleksi fitur menggunakan CFS dapat dilihat pada gambar 2.

```

5 Fitur Terbaik:
Fitur: Tugas_1, Korelasi: 0.025497913030360454
Fitur: Tugas_2, Korelasi: -0.002379170298651742
Fitur: Jenis_Kelamin, Korelasi: -0.008797269205660428
Fitur: UTS, Korelasi: -0.013837522603261904
Fitur: Comments, Korelasi: -0.053969757624819044
    
```

Gambar 2 Hasil Seleksi fitur CFS

b. *Data Transformation*

Tahap transformasi digunakan untuk mengubah format dan tipe data sebelum dimasukkan ke dalam pemodelan. Dalam konteks ini, data yang mengalami konversi adalah atribut “Time spent on course”, “Jenis Kelamin”, dan “Simbol”. Pada tahap ini, dilakukan pengubahan format dan tipe data untuk memastikan konsistensi dan kecocokan dengan kebutuhan pemodelan yang akan dilakukan. Berikut hasil transformasi atribut jenis kelamin dan simbol dapat dilihat pada tabel 5, tabel 6 dan tabel 7.

Tabel 5 Hasil Transformasi Data Atribut Openlearning

Time spent on course	Time spent on course
4 Hrs 45 Mins	285
5 Mins	5
:	:
2 Hrs 27 Mins	147

Tabel 6 Hasil Transformasi Atribut Jenis Kelamin

Jenis Kelamin	Jenis Kelamin
Perempuan	1
Laki-Laki	2

Tabel 7 Hasil Transformasi Atribut Simbol

Simbol	Simbol	Simbol
A	Lulus	1
AB	Lulus	1
B	Lulus	1
BC	Lulus	1
D	Tidak Lulus	0
E	Tidak Lulus	0
T	Tidak Lulus	0

3.4. Permodelan Algoritma *Decision Tree C4.5*

Untuk menguji algoritma C4.5 menggunakan 10-fold cross validation, langkah pertama adalah mengimpor dataset menggunakan library pandas dengan menggunakan alias "pd". Selanjutnya, dataset CSV dibaca menggunakan fungsi `pd.read_csv`. Dataset yang digunakan sudah melalui tahap persiapan yang meliputi pembersihan data, transformasi atribut jenis kelamin dari "Perempuan" dan "Laki-Laki" menjadi "1" dan "2", transformasi atribut waktu yang dihabiskan pada kursus menjadi menit, serta transformasi atribut kelas target dari simbol "Lulus" dan "Tidak Lulus" menjadi "1" dan "0". Selain itu, jika diperlukan, dilakukan. Dataset ini terdiri dari 943 records dan sudah siap untuk diuji menggunakan algoritma C4.5

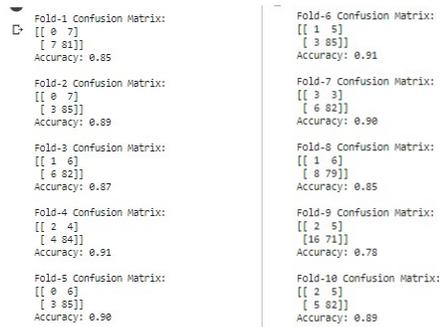
- a. Dataset Sebelum diseleksi fitur menggunakan CFS

Tabel 8 Dataset

Jenis_Kelamin	Time_spent_on_course	Progress	Course_completed	Comments	Tugas_1	Tugas_2	UTS	Simbol
1	463	98,37	98,37	8	71	73	74	1
1	129	59,8	59,8	5	61	77	75	1
1	135	98,69	98,69	9	63	85	72	1
2	173	98,37	98,37	7	71	81	70	1
2	165	98,37	98,37	26	60	75	74	1
:	:	:	:	:	:	:	:	:
1	947	98,69	98,69	4	71	75	73	1
1	268	64,38	64,38	14	76	73	75	0
1	403	94,12	94,12	106	74	64	72	1
1	222	98,69	98,69	13	73	71	70	1
1	65	98,69	98,69	10	74	73	69	1

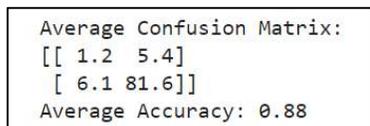
b. Tabel Confusion Matrix 10 K Fold Cross Validation

Berikut adalah hasil confusion matrix dari setiap fold dalam pengujian algoritma C4.5 dapat dilihat pada gambar 3.



Gambar 3 Tabel Confusion Matrix Setiap Fold

Dalam penelitian ini, confusion matrix yang disajikan menggunakan angka-angka tertentu untuk mewakili kelas "Lulus" dan "Tidak lulus". Dalam ini, angka 1 digunakan untuk mewakili kelas "Lulus" dan angka 0 digunakan untuk mewakili kelas "Tidak lulus". Hasil rata-rata dari confusion matrix dapat dilihat pada gambar 4 yang disajikan di bawah ini.

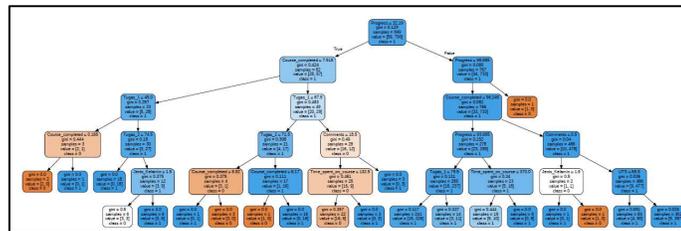


Gambar 4 Rata-Rata Convusion Matrix

Dalam confusion matrix pada gambar 4.3, terdapat empat nilai yang mewakili hasil dari klasifikasi yang dilakukan oleh suatu model :

- 1) Nilai 12 menunjukkan jumlah True Positive (TP), yaitu data yang benar-benar diklasifikasikan sebagai "Lulus".
- 2) Nilai 61 menunjukkan jumlah False Negative (FN), yaitu data yang seharusnya diklasifikasikan sebagai "Lulus" tetapi salah diklasifikasikan sebagai "Tidak Lulus".
- 3) Nilai 54 menunjukkan jumlah False Positive (FP), yaitu data yang seharusnya diklasifikasikan sebagai "Tidak Lulus" tetapi salah diklasifikasikan sebagai "Lulus".
- 4) Nilai 816 menunjukkan jumlah True Negative (TN), yaitu data yang benar-benar diklasifikasikan sebagai "Tidak Lulus".

c. Hasil Pohon Keputusan



Gambar 5 Pohon Keputusan

Berikut adalah hasil pohon keputusan dari model algoritma C4.5 :

- 1) Atribut "Progress" adalah atribut pertama yang digunakan dalam pemilihan keputusan. Jika nilai "Course_completed" kurang dari atau sama dengan 32.19, model akan mempertimbangkan atribut "Progress"; jika tidak, model akan mempertimbangkan atribut "Course_completed" yang lain.
- 2) Jika nilai "Progress" kurang dari atau sama dengan 7.515, model akan mempertimbangkan atribut "Tugas_1"; jika tidak, model akan mempertimbangkan atribut "Tugas_1" yang lain.

- 3) Jika nilai "Tugas_1" kurang dari atau sama dengan 45.0, keputusan yang diambil adalah kelas 0. Jika nilai "Tugas_1" lebih besar dari 45.0, keputusan yang diambil adalah kelas 1.
 - 4) Jika nilai "Tugas_1" lebih kecil dari atau sama dengan 67.5, model akan mempertimbangkan atribut "Tugas_1" yang lain. Jika tidak, model akan mempertimbangkan atribut "Tugas_1" yang lain.
 - 5) Jika nilai "Tugas_1" lebih kecil dari atau sama dengan 67.5, keputusan yang diambil adalah kelas 1. Jika tidak, keputusan yang diambil adalah kelas 0.
 - 6) Jika nilai "Course_completed" kurang dari atau sama dengan 98.855, model akan mempertimbangkan atribut "Course_completed" yang lain. Jika tidak, tidak ada simpul anak yang terhubung langsung.
 - 7) Jika nilai "Course_completed" kurang dari atau sama dengan 96.245, model akan mempertimbangkan atribut "Course_completed" yang lain. Jika tidak, model akan mempertimbangkan atribut "Course_completed" yang lain.
 - 8) Jika nilai "Course_completed" kurang dari atau sama dengan 96.245, keputusan yang diambil adalah kelas 1. Jika tidak, keputusan yang diambil adalah kelas 1.
 - 9) Jika nilai "Course_completed" lebih kecil dari atau sama dengan 96.245, keputusan yang diambil adalah kelas 1. Jika tidak, keputusan yang diambil adalah kelas 0.
- d. Dataset Setelah diseleksi fitur menggunakan CFS

Tabel 9 Dataset Seleksi Fitur CFS

Jenis_Kelamin	Comments	Tugas_1	Tugas_2	UTS	Simbol
1	8	71	73	74	1
1	5	61	77	75	1
1	9	63	85	72	1
2	7	71	81	70	1
2	26	60	75	74	1
:	:	:	:	:	:
1	4	71	75	73	1
1	14	76	73	75	0
1	106	74	64	72	1
1	13	73	71	70	1
1	10	74	73	69	1

e. Tabel Confusion Matrix 10 K Fold Cross Validation

Berikut adalah hasil confusion matrix dari setiap fold dalam pengujian algoritma C4.5 dapat dilihat pada gambar 6.

```

Fold-1 Confusion Matrix:
[[ 0  7]
 [ 0 88]]
Accuracy: 0.93

Fold-2 Confusion Matrix:
[[ 0  7]
 [ 0 88]]
Accuracy: 0.93

Fold-3 Confusion Matrix:
[[ 0  7]
 [ 0 88]]
Accuracy: 0.93

Fold-4 Confusion Matrix:
[[ 0  6]
 [ 0 88]]
Accuracy: 0.94

Fold-5 Confusion Matrix:
[[ 0  6]
 [ 1 87]]
Accuracy: 0.93

Fold-6 Confusion Matrix:
[[ 0  6]
 [ 0 88]]
Accuracy: 0.94

Fold-7 Confusion Matrix:
[[ 0  6]
 [ 0 88]]
Accuracy: 0.94

Fold-8 Confusion Matrix:
[[ 0  7]
 [ 0 87]]
Accuracy: 0.93

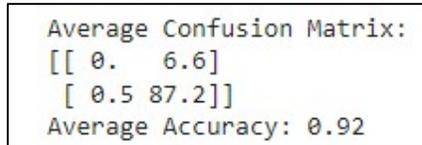
Fold-9 Confusion Matrix:
[[ 0  7]
 [ 1 86]]
Accuracy: 0.91

Fold-10 Confusion Matrix:
[[ 0  7]
 [ 3 84]]
Accuracy: 0.89

```

Gambar 6 Confusion Matrix 10 Fold CFS

Hasil rata-rata dari confusion matrix setelah dilakukan seleksi fitur menggunakan CFS dapat dilihat pada gambar 7.

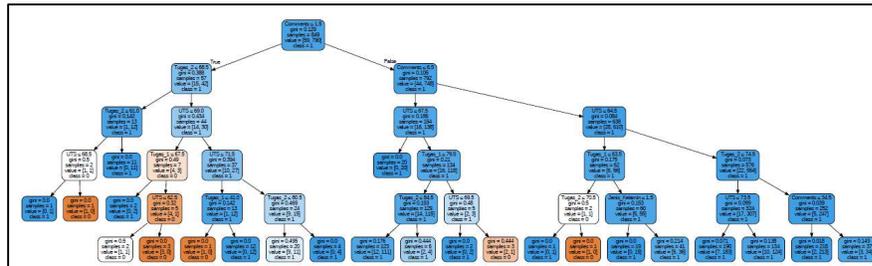


Gambar 7 Rata-Rata Confusion Matrix CFS

Dalam confusion matrix pada gambar 7, terdapat empat nilai yang mewakili hasil dari klasifikasi yang dilakukan oleh suatu model :

- 1) Nilai 0 menunjukkan jumlah True Positive (TP), yaitu data yang benar-benar diklasifikasikan sebagai "Lulus".
- 2) Nilai 5 menunjukkan jumlah False Negative (FN), yaitu data yang seharusnya diklasifikasikan sebagai "Lulus" tetapi salah diklasifikasikan sebagai "Tidak Lulus".
- 3) Nilai 66 menunjukkan jumlah False Positive (FP), yaitu data yang seharusnya diklasifikasikan sebagai "Tidak Lulus" tetapi salah diklasifikasikan sebagai "Lulus".
- 4) Nilai 872 menunjukkan jumlah True Negative (TN), yaitu data yang benar-benar diklasifikasikan sebagai "Tidak Lulus".

f. Hasil Pohon Keputusan



Hasil pohon keputusan setelah dilakukan seleksi fitur menggunakan CFS sebagai berikut :

- 1) Atribut "Comments" memiliki pengaruh yang signifikan dalam membuat keputusan. Jika jumlah komentar kurang dari atau sama dengan 1.5, keputusan yang diambil cenderung untuk kelas 1. Jika jumlah komentar lebih dari 1.5, keputusan yang diambil lebih bervariasi tergantung pada atribut-atribut lain yang dinilai.
- 2) Atribut "Tugas_2" juga memiliki pengaruh yang signifikan dalam membuat keputusan. Nilai tugas kedua yang lebih rendah atau sama dengan 68.5 atau 61.0 cenderung berhubungan dengan keputusan untuk kelas 1.
- 3) Atribut "UTS" juga mempengaruhi keputusan. Jika nilai UTS kurang dari atau sama dengan 68.5, keputusan cenderung untuk kelas 0. Jika nilai UTS lebih dari 68.5 atau 69.0, keputusan cenderung untuk kelas 1.
- 4) Atribut "Tugas_1" juga memiliki pengaruh dalam membuat keputusan, terutama pada percabangan yang melibatkan nilai Tugas_1. Nilai-nilai tertentu dari Tugas_1 dapat mengarahkan keputusan ke kelas 0 atau kelas 1.
- 5) Atribut "Jenis_Kelamin" tidak terlihat mempengaruhi keputusan dalam pohon keputusan ini, karena tidak ada percabangan yang melibatkan atribut tersebut.
- 6) Penggunaan pohon keputusan dapat membantu memahami hubungan antara atribut-atribut yang diuji dan keputusan yang diambil. Pohon keputusan dapat digunakan sebagai alat untuk memprediksi kelas target berdasarkan nilai-nilai atribut yang diberikan.
- 7) Pohon keputusan ini memiliki gini impurity yang rendah di setiap simpul, menunjukkan bahwa pohon keputusan ini efektif dalam membagi data menjadi kelompok yang homogen berdasarkan nilai-nilai atribut.

4. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan maka dapat ditarik kesimpulan sebagai berikut :

- 1) Penelitian ini mengoptimalkan Algoritma C4.5 untuk memprediksi nilai Bahasa Arab Mahasiswa tahun ajaran 2020/2021. Dengan menggunakan metode Correlation-Based Feature Selection (CFS) dan cross validation 10 K Fold, akurasi Algoritma C4.5 meningkat dari 88% menjadi 92%.
- 2) Fitur-fitur yang berpengaruh signifikan dalam pengambilan keputusan adalah "Comments", "Tugas_2", "UTS", dan "Tugas_1", sementara jenis kelamin tidak memiliki pengaruh signifikan.
- 3) Model yang dihasilkan memiliki akurasi 92% dan dapat diandalkan dalam memprediksi dan mengklasifikasikan nilai Bahasa Arab Mahasiswa secara akurat. Penelitian ini membuktikan bahwa optimasi

menggunakan CFS meningkatkan efisiensi dan kehandalan Algoritma C4.5 dalam evaluasi dan pengambilan keputusan terkait nilai Bahasa Arab Mahasiswa.

DAFTAR PUSTAKA

- [1] Zulkifli Mantondang, Ely Djulia, Sriadhi, and Janner Simarmata, *Evaluasi Hasil Belajar*, 1st ed. Medan : Yayasan Kita Menulis, 2019.
- [2] KPAI, “Survei Pelaksanaan Pembelajaran Jarak Jauh (PJJ) dan Sistem Penilaian Jarak Jauh Berbasis Pengaduan KPAI,” *Komisi Perlindungan Anak Indonesia (KPAI)*, Feb. 08, 2021. <https://bankdata.kpai.go.id/Infografis/Survei-Pelaksanaan-Pembelajaran-Jarak-Jauh-Pjj-Dan-Sistem-Penilaian-Jarak-Jauh-Berbasis-Pengaduan-Kpai>. (accessed Jul. 18, 2023).
- [3] I. Magdalena, H. N. Fauzi, and R. Putri, “PENTINGNYA EVALUASI DALAM PEMBELAJARAN DAN AKIBAT MEMANIPULASINYA,” 2020. [Online]. Available: <https://ejournal.stitpn.ac.id/index.php/bintang>
- [4] Khalida Kumalasari, “ANALISIS KINERJA AKADEMIK MAHASISWA PROGRAM STUDI AKUNTANSI FAKULTAS EKONOMI UNIVERSITAS ISLAM INDONESIA: PENGARUH INTEGRASI AKADEMIK DAN SOSIAL,” Universitas Islam Indonesia, Yogyakarta, 2018.
- [5] Lailil Muflikhah, Dian Eka Ratnawati, and Rekyan Regasari Mardi Putri, *DATA MINING*, Pertama. Malang: UB Press, 2018.
- [6] Dicky Nofriansyah and Gunadi Widi Nurcahyo, *Data Mining Dan Pengujiannya*, Pertama. Yogyakarta : CV BUDI UTAMA, 2015.
- [7] Supto, Ajeng Sekar Kirana, and Rr Nurul Rahmanita, “Teknik pre-processing dan classification dalam data science,” *Student corner*, Aug. 26, 2022. <https://mie.binus.ac.id/2022/08/26/teknik-pre-processing-dan-classification-dalam-data-science/> (accessed Mar. 20, 2023).
- [8] Ibnu Daqiqil Id, *Mechine Learning Teori, Studi Kasus, dan Implementasi Menggunakan Python*, Pertama. Pekanbaru : UR PRESS, 2021.



SURAT KETERANGAN ARTIKEL PUBLIKASI

Assalamu'alaikum Warahmatullahi wabarakatuh

Saya yang bertanda tangan dibawah ini:

Nama : Wawan Joko Pranoto, S.Kom., M.TI
NIDN : 1102057701
Nama : Kamirul Hakim
NIM : 1911102441124
Fakultas : Sains Dan Teknologi
Program Studi : S1 Teknik Informatika

Manyatakan bahwa artikel ilmiah yang berjudul "Optimasi Algoritma *Decision Tree C4.5* menggunakan *Metode Correlation-Based Feature Selection* pada klasifikasi nilai mahasiswa" telah di submit pada jurnal *Dinamika Informatika* pada tahun 2023.

<https://www.unisbank.ac.id/ojs/index.php/fti2>

Demikian surat keterangan ini dibuat untuk dapat dipergunakan sebagaimana mestinya.

Wassalamu'alaikum Warahmatullahi wabarakatuh

Mahasiswa

Kamirul Hakim

Samarinda, Rabu, 26 Juli 2023
Dosen Pembimbing

Wawan Joko Pranoto, S.Kom., M.TI
NIDN. 1102057701