e-ISSN 2597-4963 dan p-ISSN 1858-4853 DOI : https://dx.doi.org/10.30872/jim.v18i1.11330

Komparasi Optimasi Chi-Square, CFS, Information Gain Dan ANOVA Dalam Evaluasi Peningkatan Akurasi Algoritma Klasifikasi Data Performa Akademik Mahasiswa

Taghfirul Azhima Yoga Siswa
Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Muhammadiyah Kalimantan Timur
Jl. Ir. H. Juanda No.15, Sidodadi, Kec. Samarinda Ulu, Kota Samarinda, Kalimantan Timur, 75124
E-Mail : tay758@umkt.ac.id

ABSTRAK

Telah banyak penelitian implementasi data mining pada perfoma akademik mahasiswa yang dilakukan untuk mencari kinerja terbaik dari algoritma klasifikasi, namun penelitian yang menguji hubungan atribut-atribut dengan dimensi data yang tinggi pada pemodelan terhadap label data yang digunakan masih rendah. Penelitian ini bertujuan untuk mengkomparasi peningkatan akurasi algoritma klasifikasi yakni Naive Bayes, C4.5, Random Forest, dan Logistic Regression yang telah dioptimasi dengan beberapa algoritma seleksi fitur seperti Chi-Square, CFS, Information Gain dan ANOVA. Dataset yang digunakan berjumlah 2663 record, dengan membagi data menggunakan metode 5-*fold cross validation* kemudian dilakukan evaluasi kinerja algoritma menggunakan *confusion matrix*. Hasil penelitian yang diperoleh adalah optimasi *Chi-square* memiliki nilai tertinggi dalam meningkatkan akurasi pemodelan algoritma klasifikasi, dengan rata-rata peningkatan akurasi sebesar 2.45%. Sementara, hasil perbandingan algoritma klasifikasi dalam menangani data prediksi performa mahasiswa menghasilkan algoritma *Random Forest* sebagai algoritma klasifikasi tertinggi dengan persentase *accuracy* sebesar 94.5%, *precision* 95%, *recall* 94*, f1-score* 94%.

Kata Kunci – klasifikasi, *c4.5, naïve bayes, random forest, logistic regression*, optimasi, *chisquare, cfs, information gain, anova*.

1. PENDAHULUAN

Pelaksanaan pembelajaran daring dilakukan melalui *platform Learning Management System* (LMS) untuk penyediaan materi, penugasan, dan penilaian, yang kombinasikan dengan *platform* tatap muka online seperti *zoom*, *google meet*, *microsoft teams* dan lain-lain. Disisi lain survey menunjukan bahwa 73,2% peserta didik merasa terbebani dalam melakukan pembelajaran jarak jauh (*e-learning*) (KPAI, 2021). Hal ini dapat menjadi sebuah ancaman bagi institusi pendidikan dalam upaya untuk menjaga kualitas pembelajaran untuk mendapatkan lulusan yang berkualitas. Sejalan dengan hal itu Universitas Muhammadiyah Kalimantan Timur (UMKT) menggunakan *platform* LMS *OpenLearning* untuk menunjang modul pembelajaran, penugasan, serta penilaian mahasiswa.

Dalam *platform OpenLearning*, juga terdapat data rekam jejak mahasiswa yang dapat dimanfaatkan sebagai indikator dalam melakukan prediksi terhadap performa perkuliahan daring dengan tambahan data pendukung yang didapatkan dari bagian akademik UMKT. Untuk mendukung UMKT dalam menjaga mutu pendidikan, perlu dilakukan pendekatan data analitik dengan menggunakan metode klasifikasi data mining untuk memprediksi kinerja mahasiswa dalam pembelajaran daring. *Data mining* menurut Pramudiono (2006) dalam buku (Nofriansyah & Nurcahyo, 2015) adalah analisis otomatis dari data yang berjumlah besar atau kompleks dengan tujuan untuk menemukan pola atau kecenderungan yang penting yang biasanya tidak disadari keberadaannya. Informasi atau pengetahuan yang dihasilkan oleh *data mining* dapat digunakan untuk meningkatkan proses pengambilan keputusan (Santoso & Umam, 2018).

Penelitian terkait prediksi menggunakan metode *data mining* dalam dunia pendidikan telah banyak dilakukan oleh peneliti-peneliti sebelumnya. Penelitian yang dilakukan oleh Sihombing et al. (2021) tentang Analisis Keberhasilan Pembelajaran Daring pada Masa Pandemi *Covid-19* menggunakan Algoritma C4.5 dan *Naïve Bayes*, didapatkan hasil bahwa tingkat akurasi *Naïve Bayes* lebih unggul dibanding C4.5, dimana akurasi *Naïve Bayes* sebesar 99% sedangkan C4.5 sebesar 98%. Abubakar & Ahmad (2017) melakukan Prediksi kinerja peserta didik dengan membandingkan beberapa metode klasifikasi, didapatkan *Random Forest* memperoleh akurasi sebesar 76.9% lebih baik dari KNN dan *Decision Tree* dengan akurasi sebesar 69.2% dan 61.5%.. Annisa & Sasongko (2020) melakukan penelitian prediksi nilai akademik siswa dengan menggunakan algoritma *Naive Bayes*. Pada penelitian ini diperoleh hasil *accuracy* 96,24%, *precision* 95,76%, dan *recall* 100%. Hastuti (2012) pada penelitiannya melakukan komparasi terhadap 4 algoritma data mining, yaitu *Logistic Regression*, *Decision Tree*, *Naive Bayes*, dan *Neural Network* dengan hasil algoritma *Decision Tree* memiliki nilai akurasi tertinggi yaitu 95,29%, *Neural Network* 94,56%, *Naïve Bayes* 93,47% dan *Logistic Regression* 81,64%.

Penelitian sebelumnya pada umumnya hanya berfokus pada implementasi dan evaluasi kinerja algoritma data mining, namun penelitian yang menguji hubungan atribut-atribut dengan dimensi data yang tinggi pada pemodelan terhadap label data yang digunakan masih rendah. Dalam penelitian ini,

e-ISSN 2597-4963 dan p-ISSN 1858-4853 DOI : https://dx.doi.org/10.30872/jim.v18i1.11330

fokus permasalahan yang akan diteliti adalah pada tahapan seleksi fitur pada persiapan data (*data preparation*) yang merupakan salah satu tahapan terpenting dalam proses klasifikasi data. Proses *data preparation* haruslah sangat diperhatikan sebelum terbentuknya pemodelan klasifikasi, khususnya pada tahapan seleksi fitur. Menurut Jassim & Abdulwahid (2021), perbedaan keakuratan dan ketidaktepatan model algoritma sangat bergantung pada tahapan persiapan data yang dilakukan sebelumnya. Pada prosesnya, tingkat korelasi dalam fitur / atribut seringkali menghasilkan nilai yang sama dalam identifikasinya, dimana apabila proses seleksi fitur dilakukan dengan menggunakan metode tradisional dapat menghasilkan ketidakstabilan yang akan mengurangi tingkat kepercayaan fitur yang terpilih (Khaire & Dhanalakshmi, 2022). Dari banyaknya seleksi fitur yang ada, pada penelitian ini akan dilakukan komparasi antara algoritma seleksi fitur Chi-square, CFS, Information Gain, ANOVA untuk mencari algoritma seleksi fitur terbaik yang mana dapat meningkatkan performa algoritma klasifikasi dalam menangani prediksi performa mahasiswa UMKT dalam pembelajaran daring berbasis LMS *OpenLearning*.

2. TINJAUAN PUSTAKA

A. Performa Akademik Mahasiswa

Performa akademik mahasiswa merupakan sebuah tolak ukur bagi dosen dalam proses evaluasi tingkat keaktifan mahasiswa dalam mengikuti pembelajaran. Menurut Caballero et. al. (2007), performa akademik meliputi tujuan, pencapaian, dan objektif yang telah ditetapkan pada sebuah program mata kuliah yang diikuti oleh mahasiswa. Mahasiswa adalah komponen yang dimiliki perguruan tinggi yang memiliki kewajiban untuk menuntut ilmu (Warasto, 2016). Torres dan Rodríguez dalam Willcox (2011) mendefinisikan performa akademik sebagai tingkat pengetahuan yang ditunjukkan mahasiswa dalam suatu bidang atau mata pelajaran yang umumnya diukur dengan menggunakan nilai rata-rata.

Berdasarkan dari beberapa penelitian yang dipaparkan, penulis menyimpulkan bahwa performa akademik mahasiswa adalah sebuah tujuan, capaian, dan objektif yang harus diraih oleh mahasiswa dalam mengikuti pembelajaran. Serta, performa akademik dapat diukur dengan nilai rata-rata yang diraih oleh mahasiswa dalam suatu bidang atau mata pelajaran.

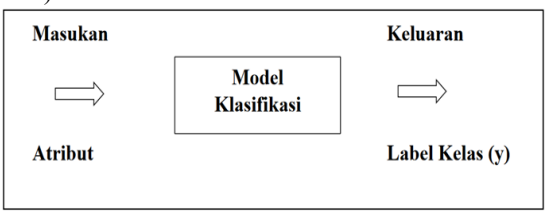
B. Data Mining

Menurut Hermawati (2013), data mining adalah proses yang melakukan pekerjaan satu atau lebih teknik machine learning atau pembelajaran komputer yang dapat digunakan untuk menganalisis dan mengekstraksi pengetahuan (knowledge) yang dilakukan secara otomatis. Sementara pengertian data mining menurut Mardi (2017) adalah sebuah proses yang memanfaatkan teknik statistik, matematika, kecerdasan buatan, dan machine learning atau pembelajaran mesin untuk proses ekstraksi dan identifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database yang besar.

Berdasarkan definisi-definisi yang telah dijabarkan di atas, dapat disimpulkan bahwa data mining merupakan proses melakukan pengambilan data secara besar untuk ekstraksi dan identifikasi informasi bermanfaat yang terdapat pada sekumpulan data dengan menggunakan teknik-teknik pembelajaran komputer atau machine learning. Data mining memiliki beberapa teknik, yaitu klasterisasi (*clustering*), regresi (*regression*), klasifikasi (*classification)*, dan kaidah asosiasi (*association rule)* (Hermawati, 2013).

C. Klasifikasi

Menurut Hermawati (2013), Klasifikasi merupakan proses pembelajaran suatu fungsi tujuan (target) f yang pada prosesnya memetakan dari setiap himpunan atribut x ke satu dari label kelas yang sebelumnya telah dilakukan proses pendefinisian. Klasifikasi adalah suatu cara untuk mengelompokkan benda berdasarkan ciri-ciri yang dimiliki oleh suatu objek klasifikasi, yang dalam prosesnya klasifikasi dapat dilakukan dengan berbagai cara baik secara manual maupun dengan bantuan teknologi (Wibawa, 2018).



Gambar 1. Konsep Klasifikasi

D. Algoritma C4.5

Menurut Berry & Linoff (2018) dalam buku (Nofriansyah & Nurcahyo, 2015) Algoritma C4.5 merupakan salah satu solusi pemecahan kasus yang sering digunakan dalam pemecahan masalah pada teknik klasifikasi. Keluaran dari algoritma C4.5 berupa sebuah *decision tree* layaknya teknik klasifikasi lain. Menurut Nofriansyah & Nurcahyo (2015) Untuk penyelesaian kasus di dalam algoritma C4.5 terdapat 2 elemen beberapa elemen yaitu entropy dan gain. *Entropy*(S) merupakan jumlah bit yang diperkirakan dibutuhkan untuk dapat mengekstrak suatu kelas (+ atau -) dari sejumlah data acak pada ruang sampel S. *Entropy* dapat dikatakan sebagai kebutuhan bit untuk menyatakan suatu kelas untuk digunakan dalam mengekstrak suatu kelas. *Entropy* digunakan untuk mengukur ketidakaslian S. Adapun rumus untuk mencari nilai *entropy*.

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

Keterangan :
S = himpunan kasus
A = atribut
n = jumlah partisi S
p^i = proporsi dari S^i terhadap S

e-ISSN 2597-4963 dan p-ISSN 1858-4853 DOI : https://dx.doi.org/10.30872/jim.v18i1.11330

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 :

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

Dimana :
A = atribut
S = sampel
n = jumlah partisi himpunan atribut a
|Si| = jumlah sampel pada partisi ke-i
|S| = jumlah sampel dalam S

E. Algoritma Naïve Bayes

Algoritma *Naïve Bayes* merupakan metode yang menggunakan probability untuk membuat model prediksi klasifikasi yang memanfaatkan data mengenai kejadian masa lampau, model ini juga dapat menghitung probabilitas suatu kejadian dan dapat berubah jika ada informasi pendukung tambahan yang disediakan (Kurniawan, 2020) .Algoritma *Naïve Bayes* dapat dirumuskan sebagai berikut (Nofriansyah & Nurcahyo, 2015):

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

Keterangan:
X = Merupakan data testing yang kelasnya belum diketahui
H = Merupakan hipotesis data X yang kelasnya lebih spesifik
P(H) = Disebut juga dengan *prior probability* yang merupakan probabilitas hipotesis
P(X) = Disebut dengan *predictor prior* yang merupakan probabilitas X
P(X|H) = Disebut juga dengan *likelihood* yang merupakan probabilitas hipotesis X berdasarkan kondisi H.

F. Algoritma Random Forest

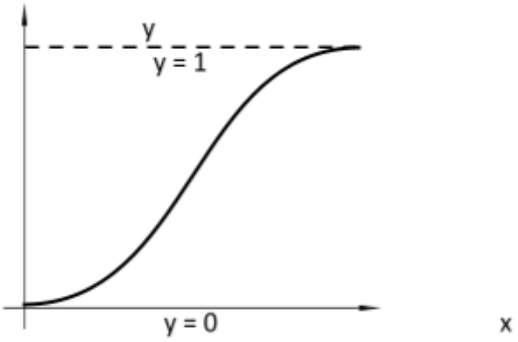
Random Forest merupakan metode hasil pengembangan dari algoritma *Classification And Regression Tree* (CART) yang pada penerapannya menggunakan metode *bootstrap aggregating* (*bagging*) dan *random feature selection* (Breiman, 2001). Menurut Jonathan (2021), *Random Forest* memiliki sebuah mekanisme internal yang menyediakan estimasi dari proses generalization error-nya sendiri, atau yang biasa disebut dengan *out-of-bag* (OOB) *error estimate*. Perumusan untuk RF yang terdiri dari N trees dinyatakan sebagai berikut (Liparas, 2014):

$$l_{(y)} = agrmax_c(\sum_{n=1}^{n} h_n(y) = c)$$

Dimana variabel *I* adalah fungsi indikator dan hn merupakan *tree* ke-n dari *RF*.

G. Algoritma Logistic Regression

Logistic Regression merupakan algoritma yang dapat memisahkan dataset menjadi dua bagian yang disebut dengan *binary classification* menggunakan metode prediksi probabilitas. *Logistic Regression* menghasilkan output yang bersifat kualitatif dan kategori (Primartha, 2021).



Gambar 2. *Logistic Regression*
Sumber : Kurniawan (2020)

Grafik ini membagi dataset menjadi dua class = 1 dan class = 0 tepat di tengah, yaitu saat Y = 0.5. *Class* merupakan prediksi probabilitas (p atau P) yang dirumuskan:

$p \geq 0.5$, *class*=1
$p < 0.5$, *class*=0

Probabilitas regresi logistik sebagai berikut:

$$p = \frac{e^{\beta_0+\beta_1+\varepsilon_i}}{1 + e^{\beta_0+\beta_1 X+\varepsilon_i}}$$

Atau

$$p = \frac{1}{1 + e^{-(\beta_0+\beta_1 X+\varepsilon_i)}}$$

Dimana:
P = Probabilitas
e = Fungsi exponen

H. Chi-square

Pengujian Chi-square menurut Usman & Akbar (2000) biasa dipakai untuk mengidentifikasi keterkaitan antara dua variabel nominal, yang diikuti dengan pengukuran kekuatan hubungan antar dua variabel. Chi-square feature selection merupakan metode yang juga umum digunakan untuk menyeleksi fitur dengan cara perangkingan tiap fitur berdasarkan hasil dari penyeleksian fitur dari nilai terbesar ke nilai terkecil. (Ling. et al. 2014). Berikut adalah rumus perhitungan Chi-square (Sugiyono, 2010)

$$X^2 = \sum_{i=1}^{k} \frac{(f_0 - f_2)^2}{f_h}$$

Keterangan:
x^2 = Chi-kuadrat
f_0 = Frekuensi yang diamati
f_h = Frekuensi yang diharapkan

I. Correlation-based Feature Selection (CFS)

Menurut Hall (1999), *Correlation-based Feature Selection* (CFS) adalah algoritma filterisasi yang melakukan perangkingan subset fitur berdasarkan evaluasi heuristik berbasis korelasi. Bias yang terjadi

e-ISSN 2597-4963 dan p-ISSN 1858-4853 DOI : https://dx.doi.org/10.30872/jim.v18i1.11330

pada fungsi evaluasi adalah pada subset yang memuat fitur yang sangat berhubungan dengan class dan tidak berhubungan satu sama lain. Fitur yang tidak relevan harus diindahkan karena fitur tersebut akan memiliki nilai korelasi yang rendah terhadap class. Berikut adalah rumus perhitungan CFS:

$$M_S = \frac{k\underline{r_{cf}}}{\sqrt{k + k(k-1)r_{ff}}}$$

Dimana M_S merupakan "*merit*" heuristik dari subset fitur *S* yang memuat fitur k, r_{cf} adalah mean dari korelasi fitur terhadap kelas (f ∈ S), dan r_{ff} merupakan fitur rata-rata/fitur interkorelasi.

J. Information Gain

Menurut Jensen & Shen (2008), *Information Gain* adalah reduksi yang diharapkan dalam *entropy* (keacakan atau mengukur ketidakteraturan informasi yang sedang diproses dalam Pembelajaran Mesin) yang dihasilkan dari mempartisi objek kumpulan data menurut fitur tertentu.. Berikut adalah rumus perhitungan *Information Gain*:

$$IG(S,A) = Entropy(S) - \sum_{c \in values(A)} |Sv||S|Entropy(Sv)$$

Keterangan:
S = Himpunan kasus
A = Atribut
$|S_i|$ = Jumlah kasus pada partisi ke-*i*
$|S|$ = Jumlah kasus dalam *S*

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

Keterangan:
S = Himpunan Kasus
c = Jumlah Partisi

K. Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) merupakan pendekatan statistik yang dikenal untuk membandingkan beberapa nilai independen rata-rata (Johnson & Synovec, 2002). Pendekatan ANOVA menggunakan pemeringkatan fitur dengan menghitung rasio dari varian antara dan dalam grup (Lin & Ding, 2011).

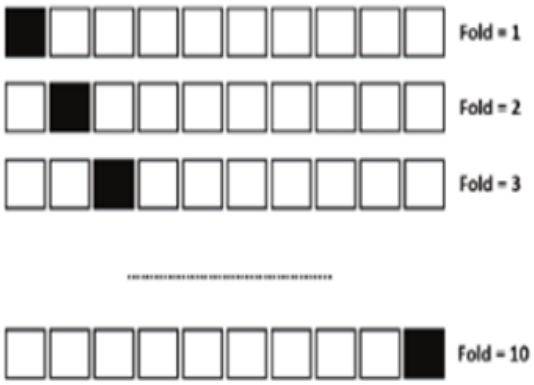
Rasio menunjukkan seberapa kuat fitur λ tersebut terkait dengan variabel grup. Persamaan berikut digunakan untuk menghitung nilai rasio λ th g-gap dipeptida dalam dua set data benchmark:

$$F(\lambda) = \frac{S_B^2(\lambda)}{S_W^2(\lambda)}$$

Dimana:
$S_R^2(\lambda)$ dan $S_R^2(\lambda)$ = Sampel varian antar grup (juga dikenal sebagai *Mean Square Between*, MSB), dan dalam grup (juga dikenal sebagai *Mean Square Within*, MSW), secara masing-masing

L. Cross Validation

Menurut Menurut Nurhayati dkk. (2014) *cross validation* adalah metode untuk memperkirakan kesalahan prediksi untuk evaluasi kinerja model. Dalam *cross validation* dikenal sebagai estimasi rotasi, dengan membagi data menjadi himpunan bagian k dengan ukuran yang hampir sama, model dalam klasifikasi dilatih dan diuji sebanyak *k*. Di setiap pengulangan, salah satu himpunan bagian akan digunakan sebagai data penguji dan sub kelompok data *k* lainnya berfungsi sebagai data pelatihan.



Gambar 3. Ilustrasi 10-*Fold Cross Validation*

M. Confusion Matrix

Confusion matrix merupakan matrik berukuran N x N yang dapat digunakan untuk permasalahan klasifikasi, dimana N merupakan jumlah kelas yang akan diprediksi (Daqiqil *Id, 2021)*. Tahapan ini menerapkan aturan Gambar 4 dengan melakukan perhitungan menggunakan 3 keluaran yaitu *precision*, *recall* dan *f-score* yang diperkenalkan oleh Baeza-Yates & Ribeiro-Neto pada tahun 1999.

Class	Actual = Yes	Actual =No
Predicted = Yes	TP	FP
Predicted = No	FN	TN

Gambar 4. *Confusion Matrix*

Pada confusion matrix terdapat beberapa istilah yang digunakan pada kasus klasifikasi yaitu (Pulungan, 2019):

a. *True Positive* (TP) *confusion matrix* data positif yang terdeteksi benar
b. *False Positive* (FP) *confusion matrix* data negatif namun terdeteksi dengan benar
c. *False Negative* (FN) *confusion matrix* data positif yang terdeteksi sebagai data negatif
d. *True Negative* (TN) *confusion matrix* data negatif yang terdeteksi benar.

Beberapa perhitungan kinerja klasifikasi dapat dijelaskan pada *confusion matrix*. Berikut adalah beberapa perhitungan kinerja klasifikasi (Sokolova & Lapalme, 2009):

1. *Accuracy*

Accuracy adalah nilai efektivitas keseluruhan dari proses klasifikasi

$$accuracy = \frac{TP + TN}{n}$$

e-ISSN 2597-4963 dan p-ISSN 1858-4853 DOI : https://dx.doi.org/10.30872/jim.v18i1.11330

2. Precision

Precision merupakan tingkat ketepatan kelas label data dengan label positif yang diberikan oleh pengklasifikasi

$$precision = \frac{TP}{TP + FP}$$

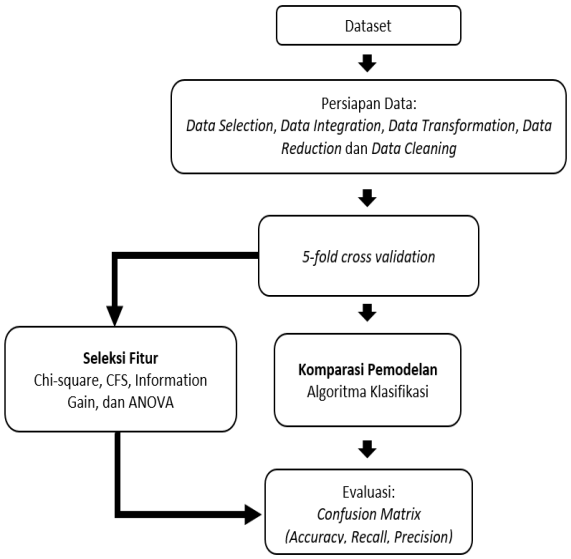
3. Recall

Recall merupakan tingkat efektivitas untuk mengidentifikasi label positif pada proses klasifikasi.

$$recall = \frac{TP}{TP + FN}$$

3. METODE PENELITIAN

Metode penelitian yang digunakan dalam penelitian ini adalah adalah metode eksperimen. Metode penelitian eksperimen menurut Rukminingsih (2020) merupakan salah satu penelitian kuantitatif dimana peneliti memanipulasi satu atau lebih variabel bebas (*independent variable*), mengontrol variabel lain yang relevan, dan mengamati efek dari manipulasi pada variabel terikat (*dependent variable*). Sebuah eksperimen dengan sengaja dan sistematis memperkenalkan perubahan dan kemudian mengamati konsekuensi dari perubahan itu. Hanya masalah penelitian yang memungkinkan peneliti untuk memanipulasi kondisi yang tepat untuk penelitian eksperimental. Dalam penelitian ini dilakukan beberapa eksperimen terhadap komparasi algoritma seleksi fitur seperti optimasi *Chi-square*, *Correlation-based Feature Selection* (CFS), *Information Gain*, dan *Analysis Of Variance* (ANOVA) untuk peningkatan akurasi algoritma klasifikasi pada kasus performa akademik mahasiswa dalam perkuliahan online berbasis LMS.
Dalam penelitian ini ada beberapa tahapan yang dilakukan dalam mencapai tujuan penelitian. Adapun tahapan-tahapan pada penelitian ini dapat dilihat pada gambar 5 berikut:



Gambar 5. Alur Tahapan Penelitian

4. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Data yang digunakan pada penelitian ini adalah data nilai akhir mahasiswa yang diperoleh dari BAA UMKT dan *platform OpenLearning*. Data yang digunakan adalah data mata kuliah Kewarganegaraan tahun akademik 2020/2021 dan 2021/2022 semester ganjil seluruh program studi. Data yang diperoleh pada BAA UMKT dan *platform OpenLearning* menghasilkan record sebanyak 2663 mahasiswa.

Data nilai akhir mahasiswa yang diperoleh dari bagian administrasi akademik UMKT didapatkan sebanyak 5 atribut yang meliputi nomor induk mahasiswa (nim), nama, nilai akhir, bobot, dan simbol. Data yang diperoleh dari bagian administrasi akademik UMKT dapat dilihat pada tabel 1.

Tabel 1. Data Yang Diperoleh Dari BAA UMKT

No	NIM	Nama	Program Studi	Nilai Akhir	Bobot	Simbol
1	1911102416071	Ade Nur Fitriani	416	79.9	3.5	AB
2	1811102416004	Ade Rino Setiawan	416	79.9	3.5	AB
3	1911102416059	Affina Aulia Firda	416	79.9	3.5	AB
…	…	…	…	…	…	…
2663	2011102432109	Tomy Julian	432	82	4	A

Sedangkan data rekam jejak mahasiswa yang diperoleh dari platform *OpenLearning* didapatkan sebanyak 13 atribut yang meliputi *Learner name, Learner email, Enrolment ID, Institution Membership ID, Enrolment date, Completion date, Time spent on course, Progress, % Course completed, Certificate ID, Comments, Kudos, Enrolment cost,* Tugas, UTS, dan *Quiz*. Pada atribut Tugas, data merupakan perolehan dari nilai rata-rata tugas 1 sampai dengan tugas 10 yang telah dikerjakan oleh mahasiswa.

B. Persiapan Data

1) Seleksi dan Integrasi Data

Seluruh data yang diperoleh dari BAA UMKT dan *platform OpenLearning* hanya beberapa atribut saja yang akan digunakan. Proses seleksi dan integrasi tahap pertama adalah dengan menyeleksi atribut yang tidak diperlukan. Atribut yang dihapus pada data BAA UMKT adalah Nama dan Bobot. Sementara atribut yang dihapus pada data *OpenLearning* adalah *Profile name, Learner name, Learner email, Enrolment ID, Institution Membership ID, Enrolment date, Completion date, Progres,* dan *Certificate ID*. Berikut adalah hasil seleksi dan integrasi data.

Tabel 2. Hasil Seleksi dan Integrasi Data

No	*Time spent on course*	*Course Completed*	Program Studi	*Comments*	*Kudos*	Tugas	*Quiz*	UTS	Simbol
1	4 Hrs 56 Mins	69.55	416	32	3	54.1	0	0	AB
2	5 Hrs 45 Mins	80.91	416	27	4	90	0	57	AB
3	17 Hrs 47 Mins	81.36	416	24	1	79.6	0	57	AB
…	…	…	…	…	…	…	…	…	…
2663	4 Hrs 42 Mins	95	432	17	0	72.4	84	0	A

e-ISSN 2597-4963 dan p-ISSN 1858-4853 DOI : https://dx.doi.org/10.30872/jim.v18i1.11330

2) Transformasi Data

Tahapan ini dilakukan untuk merubah isi data atau tipe data sebelum masuk pada pemodelan. Atribut yang akan ditransformasi adalah *time spent on course* dan simbol. Atribut *time spent on course* merupakan lama waktu mahasiswa menyelesaikan kursus yang tertera waktu berupa jam dan menit, hal ini akan ditransformasi menjadi menit secara keseluruhan. Pada atribut *Time spent on course*, data lamanya mahasiswa berada pada *course* atau mata kuliah Kewarganegaraan memuat dua jenis tipe data dalam *record*-nya, yaitu *integer* dan *string*. Hal ini dapat membuat pemodelan tidak dapat memproses *dataset* kedalam tahapan prediksi. Pada atribut ini, nilai akan diubah menjadi akumulasi dari lamanya mahasiswa berada di *course* menjadi hitungan menit. Adapun perhitungan transformasi atribut *time spent on course* adalah sebagai berikut.

Transformasi atribut *time spent on course*
1 jam = 60 menit

time spent on course = 4 *Hrs* (jam) 56 *Mins* (menit)
= 4 x 60 + 56
= 240 + 56
= 296

Berikut adalah hasil transformasi pada atribut *time spent on course*

Tabel 3. Atribut *Time Spent On Course* Setelah Di Transformasi

Time spent on course (Sebelum di Transformasi)	*Time spent on course* (Setelah di Transformasi)
4 Hrs 56 Mins	296
5 Hrs 45 Mins	345
17 Hrs 47 Mins	1067
7 Hrs 44 Mins	467
5 Hrs 45 Mins	345
:	:
4 Hrs 42 Mins	282

Kemudian atribut kedua yang ditransformasi adalah simbol, simbol ini akan digunakan sebagai kelas target pada pemodelan. Kelas target yang awalnya memiliki 8 kriteria yaitu A, AB, B, BC, D, E dan T akan diubah menjadi 2 kelas target. Perubahan yang dilakukan yaitu nilai A, AB, B diubah menjadi BAIK. Sedangkan nilai BC, C, D, E dan T diubah menjadi BURUK. Pengubahan kelas target BAIK dan BURUK dilakukan berdasarkan ketentuan standar kelulusan mata kuliah dasar umum UMKT yaitu dengan nilai akhir minimal B. Adapun rentang penilaian nilai akhir mengacu pada norma Penilaian Acuan Patokan (PAP) program akademik UMKT dengan uraian pada tabel 4.

Tabel 4. Norma Penilaian Berdasarkan PAP Program Akademik UMKT

No	Huruf	Angka	Nilai Akhir	Predikat
1	A	4	≥ 80	Sangat Baik
2	AB	3,5	75 - < 80	
3	B	3	70 - < 75	Baik
4	BC	2,5	65 - < 70	
5	C	2	65 - < 65	Cukup
6	D	1	50 - < 60	Kurang
7	E	0	< 50	Gagal

Tabel 5 adalah hasil transformasi atribut Simbol.

Tabel 5. Atribut Simbol Setelah Di Transformasi

Simbol (Sebelum di Transformasi)	Simbol (Setelah di Transformasi)
A	BAIK
AB	BAIK
B	BAIK
BC	BURUK
C	BURUK
D	BURUK
E	BURUK
T	BURUK

Hasil transformasi atribut *time spent on course* dan simbol dapat dilihat pada tabel 6 dibawah ini.

Tabel 6. Hasil Data Setelah Di Transformasi

No	*Time spent on course*	*Course Completed*	*Comments*	*Kudos*	Program Studi	Tugas	*Quiz*	UTS	Simbol
1	296	70	32	3	416	54.1	0	0	BURUK
2	345	81	27	4	416	90	57	0	BURUK
3	1067	81	24	1	416	79.6	57	0	BURUK
…	…	…	…	…	…	…	…	…	…
2663	282	95	17	0	432	72.4	0	84	BAIK

3) Reduksi dan Pembersihan Data

Tahapan ini dilakukan untuk menghindari dataset yang tidak seimbang dengan mengurangi data mayoritas. Data mayoritas merupakan data kelas BAIK yang berjumlah 2663 sehingga dikurangi sejumlah data minoritas BURUK yang berjumlah 127 data. Jadi total data kelas baik berjumlah 127 yang diambil secara random. Setelah itu dilakukan tahapan *data cleaning* atau pembersihan data terhadap data yang tidak konsisten. Dimana pada kasus yang ditemukan, mahasiswa memiliki nilai tugas, *quiz* dan uts yang buruk diklasifikasikan dengan baik. Oleh karena itu pada tahapan reduksi ini dilakukan proses pembersihan data yang tidak konsisten. Dataset yang telah melalui tahapan *data reduction* dan *data cleaning* menghasilkan sebanyak 254 yang tertera pada tabel 7.

Tabel 7. Data Setelah Proses Reduksi

No	*Time spent on course*	*Course Completed*	*Comments*	*Kudos*	Program Studi	Tugas	*Quiz*	UTS	Simbol
1	62	49	18	0	413	22	0	62	BURUK
2	321	81	15	0	431	58.4	54	32	BURUK
3	85	77	20	0	413	19	0	30	BURUK
…	…	…	…	…	…	…	…	…	…
254	606	97	8	0	415	70.9	87	78	BAIK

C. Pemodelan

Proses perbandingan optimasi algoritma pemodelan klasifikasi diawali dengan melakukan seleksi fitur terhadap dataset yang telah melalui proses persiapan data. Metode seleksi fitur yang dilakukan meliputi *Chi-square*, CFS, *Information Gain*, dan ANOVA. Tahapan seleksi fitur dimulai dengan melakukan perangkingan pengaruh masing-masing fitur terhadap label dengan ketentuan tingkat pengaruh tertinggi diletakan pada peringkat 1, sementara tingkat pengaruh terendah diletakkan pada peringkat 8. Setelah masing-masing ranking seleksi fitur diperoleh, barulah data dimasukan pada pemodelan klasifikasi yang meliputi C4.5, *Naive Bayes*, *Random*

e-ISSN 2597-4963 dan p-ISSN 1858-4853 DOI : https://dx.doi.org/10.30872/jim.v18i1.11330

Forest, dan *Logistic Regression*. Adapun teknik pembagian *data training* dan *data testing* untuk menguji pemodelan menggunakan pengujian 5-*fold cross validation*.

1) **Optimasi Chi-square**

Tabel 8. Perangkingan Fitur *Chi-square*

No	Atribut	Rangking
1	*time_spent_on_course*	1
2	*quiz*	2
3	tugas	3
4	uts	4
5	*course_completed*	5
6	*kudos*	6
7	*comments*	7
8	program_studi	8

Berdasarkan hasil penerapan seleksi fitur menggunakan *Chi-square* pada beberapa algoritma klasifikasi dengan beberapa skema perangkingan, didapatkan hasil peningkatan akurasi tertinggi pada algoritma *Naive Bayes*. Peningkatan akurasi pada pemodelan *Naive Bayes* berada pada 3.5% lebih tinggi dibandingkan dengan pemodelan tanpa seleksi fitur. Pada pemodelan klasifikasi yang dilakukan terhadap data performa akademik mahasiswa juga menunjukan bahwa algoritma random *Random Forest* memperoleh tingkat akurasi paling tinggi dibandingkan dengan algoritma klasifikasi lainnya, yakni sebesar 94.5%. Berikut adalah hasil peningkatan akurasi algoritma klasifikasi menggunakan *Chi-square*.

Tabel 9. Hasil Perbandingan Peningkatan Akurasi Algoritma Klasifikasi

No	Algoritma	Jumlah Atribut	Peningkatan Akurasi
1	C4.5	5	2.4%
2	*Naive Bayes*	4	3.5%
3	*Random Forest*	5	1.2%
4	*Logistic Regression*	3	2.7%
Rata-rata Hasil Peningkatan Akurasi Tertinggi			**2.45%**

2) **Optimasi CFS**

Tabel 10. Perangkingan Fitur CFS

No	Atribut	Rangking
1	*quiz*	1
2	tugas	2
3	*time_spent_on_course*	3
4	uts	4
5	*course_completed*	5
6	*kudos*	6
7	program_studi	7
8	*comments*	8

Berdasarkan hasil penerapan seleksi fitur menggunakan CFS pada beberapa algoritma klasifikasi dengan beberapa skema perangkingan, didapatkan hasil peningkatan akurasi tertinggi pada algoritma *Naive Bayes*. Peningkatan akurasi pada pemodelan *Naive Bayes* berada pada 3.5% lebih tinggi dibandingkan dengan pemodelan tanpa seleksi fitur. Pada pemodelan klasifikasi yang dilakukan terhadap data performa akademik mahasiswa juga menunjukan bahwa algoritma *Random Forest* memperoleh tingkat akurasi paling tinggi dibandingkan dengan algoritma klasifikasi lainnya, yakni sebesar 94.1%. Berikut adalah hasil peningkatan akurasi algoritma klasifikasi menggunakan CFS.

Tabel 11. Hasil Perbandingan Peningkatan Akurasi Algoritma Klasifikasi

No	Algoritma	Jumlah Atribut	Peningkatan Akurasi
1	C4.5	3	0.8%
2	*Naive Bayes*	4	3.5%
3	*Random Forest*	6	1.2%
4	*Logistic Regression*	3	2.7%
Rata-rata Hasil Peningkatan Akurasi Tertinggi			**2.05%**

3) **Optimasi Information Gain**

Tabel 12. Perangkingan Fitur *Information Gain*

No	Atribut	Rangking
1	*quiz*	1
2	tugas	2
3	*course_completed*	3
4	*time_spent_on_course*	4
5	uts	5
6	*comments*	6
7	program_studi	7
8	*kudos*	8

Berdasarkan hasil penerapan seleksi fitur menggunakan *Information Gain* pada beberapa algoritma klasifikasi dengan beberapa skema perangkingan, didapatkan hasil peningkatan akurasi tertinggi pada algoritma *Naive Bayes*. Peningkatan akurasi pada pemodelan *Naive Bayes* berada pada 3.5% lebih tinggi dibandingkan dengan pemodelan tanpa seleksi fitur. Pada pemodelan klasifikasi yang dilakukan terhadap data performa akademik mahasiswa juga menunjukan bahwa algoritma *Random Forest* memperoleh tingkat akurasi paling tinggi dibandingkan dengan algoritma klasifikasi lainnya, yakni sebesar 94.1%. Berikut adalah hasil peningkatan akurasi algoritma klasifikasi menggunakan *Information Gain*.

Tabel 13. Hasil Perbandingan Peningkatan Akurasi Algoritma Klasifikasi

No	Algoritma	Jumlah Atribut	Peningkatan Akurasi
1	C4.5	2	0.4%
2	*Naive Bayes*	5	3.5%
3	*Random Forest*	4	0.8%
4	*Logistic Regression*	2	2.3%
Rata-rata Hasil Peningkatan Akurasi Tertinggi			**1.75%**

4) **Optimasi ANOVA**

Tabel 14. Perangkingan Fitur ANOVA

No	Atribut	Rangking
1	*quiz*	1
2	tugas	2
3	*course_completed*	3
4	*time_spent_on_course*	4
5	uts	5
6	*comments*	6
7	program_studi	7
8	*kudos*	8

e-ISSN 2597-4963 dan p-ISSN 1858-4853 DOI : https://dx.doi.org/10.30872/jim.v18i1.11330

Berdasarkan hasil penerapan seleksi fitur menggunakan ANOVA pada beberapa algoritma klasifikasi dengan beberapa skema perangkingan, didapatkan hasil peningkatan akurasi tertinggi pada algoritma *Logistic Regression*. Peningkatan akurasi pada pemodelan *Logistic Regression* berada pada 2.7% lebih tinggi dibandingkan dengan pemodelan tanpa seleksi fitur. Pada pemodelan klasifikasi yang dilakukan terhadap data performa akademik mahasiswa juga menunjukan bahwa algoritma *Random Forest* memperoleh tingkat akurasi paling tinggi dibandingkan dengan algoritma klasifikasi lainnya, yakni sebesar 94.1%. Berikut adalah hasil peningkatan akurasi algoritma klasifikasi menggunakan *ANOVA.*

Tabel 15. Hasil Perbandingan Peningkatan Akurasi Algoritma Klasifikasi

No	Algoritma	Jumlah Atribut	Peningkatan Akurasi
1	C4.5	5	2.4%
2	*Naive Bayes*	4	1.9%
3	*Random Forest*	3	0.8%
4	*Logistic Regression*	3	2.7%
Rata-rata Hasil Peningkatan Akurasi Tertinggi			**1.95%**

D. Evaluasi

Pada penelitian ini, pembagian data 5-*fold cross validation* digunakan pada pengujian pemodelan algoritma klasifikasi dengan optimasi *Chi-square*, CFS, *Information Gain*, dan ANOVA. Hasil pengujian pada masing-masing pemodelan menggunakan algoritma optimasi menunjukkan bahwa algoritma *Chi-square* menghasilkan rata-rata peningkatan akurasi algoritma tertinggi dibandingkan dengan algoritma optimasi lainnya, yakni sebesar 2.45%. Berikut adalah perbandingan peningkatan akurasi algoritma klasifikasi menggunakan *Chi-square*, CFS, *Information Gain*, dan ANOVA.

Tabel 16. Perbandingan Peningkatan Akurasi Algoritma Klasifikasi

No	Algoritma Optimasi	Rata-Rata Peningkatan Akurasi
1	Chi-square	2.45%
2	CFS	2.05%
3	Information Gain	1.75%
4	ANOVA	1.95%

Adapun hasil pengujian algoritma klasifikasi dengan menggunakan optimasi *Chi-square*, CFS, *Information Gain*, dan ANOVA menunjukkan bahwa algoritma klasifikasi *Random Forest* memiliki tingkat akurasi klasifikasi tertinggi pada data performa akademik mahasiswa, yakni sebesar 94.5% dibandingkan dengan algoritma klasifikasi lainnya. Berikut adalah perbandingan hasil akurasi algoritma klasifikasi C4.5, *Naive Bayes*, *Random Forest*, dan *Logistic Regression*.

Tabel 17. Perbandingan Hasil Akurasi Algoritma Klasifikasi

Algoritma Klasifikasi	Chi-square	CFS	Information Gain	ANOVA
C4.5	90.2%	89.9%	89.4%	90.6%
Naive Bayes	93.7%	93.7%	93.7%	92.1%
Random Forest	94.5%	94.1%	94.1%	94.1%
Logistic Regression	92.5%	92.5%	92.1%	92.5%

5. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan maka dapat ditarik kesimpulan:

a) Prediksi performa akademik mahasiswa Universitas Muhammadiyah Kalimantan Timur dalam pembelajaran daring menggunakan *OpenLearning* dapat dilakukan dengan atribut *time spent on course*, *course completed*, *tugas*, *quiz*, dan uts.
b) Hasil komparasi pemodelan algoritma klasifikasi yang meliputi C4.5*, Naïve Bayes*, *Random Forest* dan *Logistic Regression* dengan optimasi pemodelan menggunakan algoritma *Chi-square,* CFS, *Information Gain,* dan ANOVA menghasilkan algoritma *Chi-square* sebagai algoritma optimasi tertinggi, dengan rata-rata peningkatan akurasi sebesar 2.45% terhadap algoritma klasifikasi.
c) Perbandingan algoritma klasifikasi C4.5*, Naïve Bayes*, *Random Forest* dan *Logistic Regression* menghasilkan algoritma *Random Forest* sebagai algoritma klasifikasi dengan akurasi tertinggi dalam menangani data prediksi performa akademik mahasiswa, dengan persentase *accuracy* sebesar 94.5%, *precision* 95%, *recall* 94, *f1-score* 94%.

Adapun saran dari penulis untuk penelitian ini adalah:

a) Untuk menghasilkan akurasi yang lebih baik dapat menambahkan atribut terhadap atribut yang telah diteliti, dengan menambahkan nilai tugas-tugas tambahan, absensi kehadiran dan nilai ujian akhir semester.
b) Dapat menggunakan algoritma klasifikasi dan menerapkan algoritma optimasi lainnya dalam upaya meningkatkan *accuracy, precision, recall* dan *f1-score* pemodelan untuk memprediksi performa akademik mahasiswa dalam pembelajaran *online* maupun *offline*.

6. DAFTAR PUSTAKA

Abubakar, Y., & Ahmad, N. B. H. (2017). Prediction of Students Performance in ELearning Environment Using Random Forest. International Journal of Innovative Computing, 7(2).

Annisa, R., & Sasongko, A. (2020). Prediksi Nilai Akademik Mahasiswa Menggunakan Algoritma Naïve Bayes. Jurnal Sains & Teknologi. Vol. 9 (1)

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. Springer.

Daqiqil Id, I. (2021). Machine Learning: Teori, Studi Kasus dan Implementasi Menggunakan Python. Riau: UR PRESS

Hall, M. A. (1999). Correlation-based feature selection for machine learning

Hermawati, F.A. (2013). Data Mining. Yogyakarta: Penerbit Andi

Hastuti, K. (2012). Analisis komparasi algoritma klasifikasi data mining untuk prediksi mahasiswa non aktif. Semantik, 2(1).

e-ISSN 2597-4963 dan p-ISSN 1858-4853 DOI : https://dx.doi.org/10.30872/jim.v18i1.11330

Jensen, R., & Shen, Q. (2008). Computational Intelligence and Feature Selection - Rough and Fuzzy Approaches. IEEE Press series on computational intelligence.Johnson, K. J., & Synovec, R. E. (2002). Pattern recognition of jet fuels: comprehensive GC× GC with ANOVA-based feature selection and principal component analysis. Chemometrics and Intelligent Laboratory Systems, 60(1-2), 225-237.

Jassim, M. A., & Abdulwahid, S. N. (2021). Data Mining preparation: Process, Techniques and Major Issues in Data Analysis. In IOP Conference Series: Materials Science and Engineering (Vol. 1090, No. 1, p. 012053). IOP Publishing.

Khaire, U. M., & Dhanalakshmi, R. (2022). Stability of feature selection algorithm: A review. Journal of King Saud University-Computer and Information Sciences, 34(4), 1060-1073.

Kurniawan, D., Pengenalan Machine Learning Python. Jakarta: PT Alex Media Komputindo, 2020.

KPAI. 2021. Survei Pelaksanaan Pembelajaran Jarak Jauh (PJJ) dan Sistem Penilaian Jarak Jauh Berbasis Pengaduan KPAI [pdf] Komisi Perlindungan Anak Indonesia. Tersedia di: https://bankdata.kpai.go.id/files /2021/02/ Paparan-Survei-PJJ-KPAI-29042020_Final-update.pdf [Diakses 10 Februari 2022]

Lin, H., & Ding, H. (2011). Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. Journal of theoretical biology, 269(1), 64-69.

Ling, J., Kencana, I. P. E. N., & Oka, T. B. (2014). Analisis Sentimen Menggunakan Metode Naïve Bayes Classifier Dengan Seleksi Fitur Chi Square. E-Jurnal Matematika, 3(3), 92-99.

Liparas, D., Ha, Cohen-Kerner, Y., Moumtzidou, A., Vrochidis, S., & Kompatsiaris, I. (2014). News articles classification using random forests and weighted multimodal features. In Information Retrieval Facility Conference (pp. 63-75). Springer, Cham.

Nurhayati, Soekarno, I., Hadihardaja, I. K., & Cahyono, M. (2015). IEEE. A study of hold-out and k-fold cross validation for accuracy of groundwater modelling in tidal lowland reclamation using extreme learning machine. 10.1109/TIME-E.2014.7011623.

Nofriansyah, D., & Nurcahyo, G. (2015). Algoritma Data Mining dan Pengujian. Yogyakarta: DEEPUBLISH.

Pulungan, A. F. (2019). Analisis Kinerja Bray Curtis Distance, Canberra Distance dan Euclidean Distance pada Algoritma K-Nearest Neighbour. Tersedia di: https://repositori.usu.ac.id/handle/123456789/15051 [Diakses 28 Januari 2022]

Sugiyono. (2010). Statistika untuk Penelitian. Bandung: CV Alfabeta.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information processing & management, 45(4), 427-437.

Usman, H., & Akbar, S. R. Purnomo (2000), Pengantar Statistika, PT. Bumi Aksara, Jakarta

Primartha, R. (2021). Algoritma Machine Learning. Bandung: Informatika Bandung.

Santoso, B., & Umam, A. (2018). Data Mining dan Big Data Analytics. Yogyakarta: Penebar Media Pustaka.

Sihombing, I. A., Hartama, D., Parlina, I., Gunawan, I., & Kirana, I. O. (2021). Analisis Keberhasilan Pembelajaran Daring pada Masa Pandemi Covid-19 menggunakan Algoritma C4 . 5 dan Naive Bayes. Jurnal Komputer dan Informatika, 3(November), 89–96.

Wibawa, A. P. (2018). Metode-metode Klasifikasi. In Prosiding SAKTI (Seminar Ilmu Komputer dan Teknologi Informasi) (Vol. 3, No. 1, pp. 134-138).

Willcox, M. del R. (2011). Factores de riesgo y protección para el rendimiento académico: Un estudio descriptivo en estudiantes de Psicología de una universidad privada. Ibero-American Journal of Education, 55(1), 1-9. Recuperado de http://www.rieoei.org/deloslectores/3878Wilcox.pdf