Komparasi Metode Seleksi Fitur Dalam Prediksi Keterlambatan Pembayaran Biaya Kuliah

Taghfirul Azhima Yoga Siswa [1*], Renaldi Panji Wibowo [2]

[1] Program Studi Teknik Informatika, Universitas Muhammadiyah Kalimantan Timur, Samarinda, Kalimantan Timur
[2] Program Studi Informatika, Universitas Mulawarman, Samarinda, Kalimantan Timur
Email: [1*] tay758@umkt.ac.id, [2] renaldipanji@student.unmul.ac.id

(Naskah masuk: 15 Feb 2023, direvisi: 2 Mar 2023, diterima: 7 Mar 2023)

Abstrak

Penelitian *data mining* pada keterlambatan pembayaran SPP telah banyak dilakukan namun mayoritas penelitian memiliki *dataset* yang berdimensi rendah. Hal ini dapat menjadi bahan kajian bagi para peneliti selanjutnya dikarenakan penelitian terkait *dataset* keterlambatan biaya SPP yang berdimensi tinggi hanya mendapatkan akurasi dibawah 60%. Ditambah lagi penelitian klasifikasi *data mining* yang menguji hubungan antar atribut-atribut yang digunakan pada pemodelan terhadap label data relatif masih minim. Penelitian ini bertujuan untuk menganalisis peningkatan akurasi algoritma klasifikasi yakni *K-Nearest Neighbor, Naive Bayes, C4.5, Random forest*, dan *Logistic Regression* dalam memprediksi keterlambatan biaya kuliah yang dioptimasi dengan beberapa perbandingan algoritma seleksi fitur diantaranya *Mutual Information, Forward Selection, Backward,* dan *Recursive Elimination*. Data yang digunakan adalah data pembayaran SPP mahasiswa dari tahun 2019 - 2021 dengan teknik pembagian data menggunakan metode *5-fold cross validation*. Hasil dari penelitian ini ditemukan bahwa algoritma *Backward Elimination* memberikan peningkatan akurasi tertinggi dengan nilai rata-rata 0,52%, sedangkan algoritma klasifikasi yang memiliki akurasi tertinggi terdapat pada *random forest* dan C4.5 dengan nilai akurasi sebesar 62,6%, *precision* 65%, *recall* 63% dan *f1-score* 61%.

Kata Kunci: *Data mining*, Klasifikasi, *Mutual Information, Forward Selection, Backward Elimination, Recursive Elimination*.

Comparison of Feature Selection Methods in Predicting Late Payment of Tuition Fees

Abstract

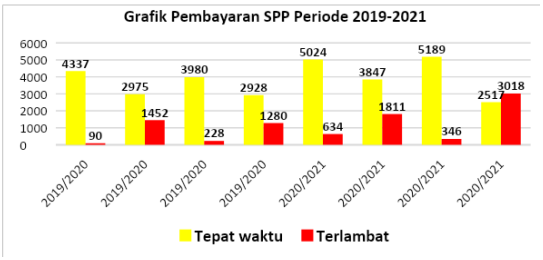
Data mining research on late payment of tuition fees has been carried out a lot, but the majority of studies have low-dimensional datasets. This can be material for study for future researchers because research related to the high-dimensional tuition fee delay dataset only obtains an accuracy of below 60%. In addition, data mining classification research that examines the relationship between the attributes used in modeling and relatively minimal data labels. This study aims to analyze the increase in the accuracy of the classification algorithm, namely K-Nearest Neighbor, Naive Bayes, C4.5, Random forest, and Logistic Regression in predicting delays in tuition fees which are optimized with several comparisons of feature selection algorithms including Mutual Information, Forward Selection, Backward, and Recursive Elimination. The data used is student tuition payment data from 2019 - 2021 with the data sharing technique using the 5-fold cross validation method. The results of this study found that the Backward Elimination algorithm provides the highest increase in accuracy with an average value of 0.52%, while the classification algorithm that has the highest accuracy is in random forest and C4.5 with an accuracy value of 62.6%, 65% precision, 63% recall and 61% f1-score.

Keywords: *Data mining, Classification, Mutual Information, Forward Selection, Backward Elimination, Recursive Elimination.*

I. PENDAHULUAN

Pasca pandemi Covid-19 jelas berdampak pada sektor ekonomi dunia yang mengakibatkan banyak perusahaan mengalami penurunan *income*. Hal ini menyebabkan perusahaan mencari cara agar dapat bertahan, salah satunya dengan melakukan pemutusan hubungan kerja (PHK), pengurangan gaji, penghilangan fasilitas karyawan dan melakukan perubahan pada struktur organisasi [1]. Tidak heran hal ini juga berimbas pada pembayaran biaya kuliah di perguruan tinggi khususnya Universitas Muhammadiyah Kalimantan Timur (UMKT).

Universitas Muhammadiyah Kalimantan Timur (UMKT) menjadi kampus swasta milik Muhammadiyah sebagai dedikasi amal usaha yang beroperasi di dunia pendidikan. Dalam meningkatkan kualitas pembelajaran, UMKT terus berupaya melakukan perkembangan seperti perbaikan sarana prasarana, pembangunan infrastruktur, dan penambahan jumlah dosen. Sumber dana terbesar yang digunakan adalah keuangan mahasiswa yaitu salah satunya melalui pembayaran Sumbangan Pembangunan Pendidikan (SPP). Jika terjadi keterlambatan pembayaran biaya kuliah tentu dapat merugikan pihak UMKT yang berpengaruh pada biaya operasional. Sedangkan bagi mahasiswa itu sendiri dapat mengganggu proses perkuliahan seperti tidak dapat mengambil Kartu Rencana Studi (KRS) dan tidak bisa mencetak kartu ujian yang akan digunakan sebagai syarat untuk mengikuti ujian akhir semester.



Gambar 1. Pembayaran SPP Periode Tahun 2019-2021
(Sumber: Bagian Keuangan UMKT, 2022)

Gambar 1 menunjukkan grafik mahasiswa dalam melakukan pembayaran SPP pada periode tahun 2019 – 2021 mengalami kenaikan dan penurunan. Namun pada periode akhir 2021 justru mengalami kenaikan keterlambatan pembayaran SPP yang sangat drastis dengan jumlah 3.018 dari total 5.533 mahasiswa. Berdasarkan fenomena tersebut perlu adanya analisis tentang prediksi keterlambatan, salah satunya dengan pendekatan *data mining*. *Data mining* merupakan aktivitas yang berkaitan dengan pengumpulan data, pemakaian data historis untuk menemukan pengetahuan, informasi, keteraturan, pola atau hubungan dalam data yang berukuran besar [2]. Salah satu metode *data mining* yang sering digunakan adalah klasifikasi.

Penelitian *data mining* terkait dengan prediksi keterlambatan SPP telah banyak dilakukan diantaranya, Abdullah dkk. [3] Prediksi Keterlambatan Pembayaran SPP Sekolah dengan menggunakan metode K-NN pada 236 data siswa dari angkatan 2017/2018 dengan tingkat akurasi sebesar 86%. Ginting dkk. [4] melakukan prediksi pada keterlambatan pembayaran sumbangan pembangunan pendidikan menggunakan metode C4.5 dengan hasil tingkat akurasi sebesar 73%, Firman [5] menggunakan algoritma *random forest* dalam memprediksi keterlambatan pembayaran UKT dengan hasil akurasi sebesar 58,70%, sedangkan Apandi dkk [6] menganalisis kemungkinan keterlambatan pembayaran SPP pada Politeknik TEDC Bandung dengan mendapatkan tingkat akurasi sebesar 75%.

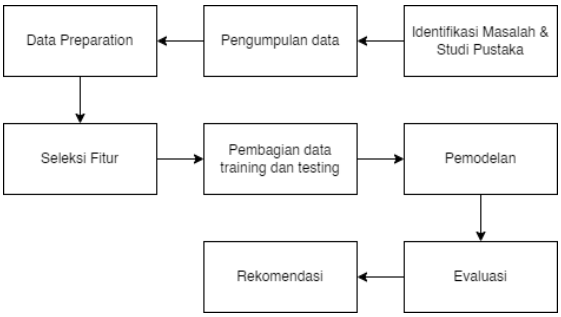
Berdasarkan studi literatur yang telah dilakukan sebelumnya, hampir di semua penelitian memiliki *dataset* yang berdimensi rendah, kecuali pada penelitian Firman [7]. Hal ini dapat menjadi bahan kajian bagi para peneliti dikarenakan penelitian terkait *dataset* keterlambatan biaya spp yang berdimensi tinggi hanya mendapatkan akurasi dibawah 60%. Ditambah lagi dengan minimnya penelitian yang menguji hubungan atribut-atribut yang digunakan pada pemodelan terhadap label data yang digunakan. Hal ini akan menganggu efektifitas dan kinerja dari algoritma klasifikasi, sebab kemungkinan adanya atribut yang kurang relevan untuk digunakan dalam melakukan analisis prediksi. Oleh karena itu perlu diterapkannya seleksi fitur sebelum melakukan analisis seperti yang dilakukan oleh Rohmayani [8] memprediksi keterlambatan pembayaran biaya kuliah dengan menggunakan metode *Naive Bayes* dan dukungan seleksi fitur *particle swarm optimization* yang dihasilkan peningkatan akurasi sebesar 13,2%. Penelitian lain yang dilakukan oleh Muqorobin dkk [9] melakukan prediksi keterlambatan pembayaran biaya pendidikan sekolah menggunakan metode *Naive Bayes* dengan seleksi fitur *information gain*, diperoleh bahwa terjadi peningkatan akurasi sebesar 10% dibanding tanpa menggunakan seleksi fitur. Hal ini memperkuat bukti bahwa seleksi fitur dapat meningkatkan akurasi algoritma klasifikasi walaupun pada penelitian sebelumnya memiliki *dataset* yang berdimensi rendah [8], [9]. Penelitian ini akan mengkomparasi beberapa algoritma seleksi fitur dengan tujuan mencari peningkatan nilai akurasi tertinggi yang diterapkan pada data keterlambatan pembayaran biaya kuliah, khususnya dengan *dataset* yang berdimensi tinggi. Hasil seleksi fitur tersebut akan diterapkan ke dalam beberapa algoritma klasifikasi seperti *algoritma K-NN*, *Naïve Bayes*, *C4.5*, *Logistic Regression*, dan *Random forest* untuk melihat peningkatan nilai akurasi yang dihasilkan dari masing-masing algoritma seleksi fitur yang dikomparasikan.

II. METODE PENELITIAN

Metode penelitian yang digunakan dalam penelitian ini adalah metode eksperimen. Metode penelitian eksperimen 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*) [10]. Dalam penelitian ini dilakukan beberapa eksperimen terhadap komparasi algoritma seleksi fitur *seperti Mutual Information, Forward Selection, Backward Elimination, dan Recursive Feature Elimination*

DOI: 10.34148/teknika.v12i1.601

untuk peningkatan akurasi algoritma klasifikasi pada kasus keterlambatan pembayaran SPP mahasiswa di UMKT. Ada beberapa tahapan yang dilakukan dalam mencapai tujuan penelitian yang dapat dilihat pada Gambar 2 berikut:



Gambar 2. Alur Tahapan Penelitian

1. Mengidentifikasi masalah dan mencari pemecahan masalah melalui studi pustaka.
2. Pengumpulan data diambil dari Bagian Administrasi Akademik dan Biro Keuangan UMKT.
3. *Data preparation* dilakukan agar data yang digunakan memiliki kualitas yang baik.
4. Seleksi fitur yang digunakan yakni, *Mutual Information, Forward Selection, Backward Elimination, dan Recursive Feature Elimination*.
5. Dalam pembagian data *training* dan data *testing* menggunakan 10-*Fold Cross Validation*.
6. Pemodelan yang digunakan adalah C4.5, *Naïve Bayes, Random Forest*, dan *Logistic Regression*.
7. Teknik evaluasi yang digunakan adalah *confusion matrix*.
8. Rekomendasi yang diberikan seperti fitur-fitur apa saja yang sangat berpengaruh dalam prediksi dan metode seleksi fitur dengan performa terbaik.

A. Algoritma C4.5

Algoritma C4.5 adalah algoritma yang dipergunakan dalam membentuk pohon keputusan berdasarkan kriteria pembentuk keputusan [11]. Algoritma C4.5 mirip sebuah pohon dimana terdapat *node internal* (bukan daun) yang mendeskripsikan atribut-atribut, setiap cabang menggambarkan hasil dari atribut yang diuji, dan setiap daun menggambarkan kelas [12].

Secara umum algoritma C4.5 dalam membangun pohon keputusan sebagai berikut:

1. Pilih atribut sebagai akar

Untuk memilih atribut sebagai akar, didasarkan pada nilai gain tertinggi dari atribut-atribut yang ada. Untuk menghitung gain digunakan, rumus seperti pada persamaan berikut.

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

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

Sementara itu, perhitungan nilai entropi dapat dilihat pada persamaan berikut.

$$Entropy\ (S) = \sum_{i=1}^{n} -pi * \log_2 p \quad (2)$$

Dimana:
S : Himpunan kasus
A : Fitur
n : jumlah partisi S
pi : proporsi dari S_i terhadap S

2. Buat cabang untuk tiap-tiap nilai.
3. Bagi kasus dalam cabang
4. Ulangi proses setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

B. Algoritma *Naïve Bayes*

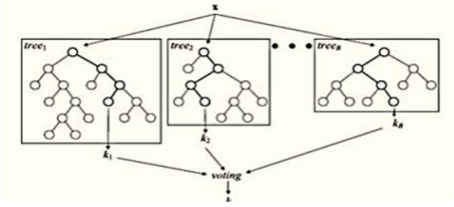
Algoritma *Naive Bayes* merupakan algoritma klasifikasi yang menganut teorema *Bayesian* pada statistika [13]. Algoritma *Naive Bayes* biasa digunakan dalam memprediksi probabilitas keanggotaan suatu kelas. Dalam melakukan perhitungan nilai probabilitas P(H|X), Teorema *Bayesian* menggunakan probabilitas P(X), P(H), dan P(X|H) sebagai berikut [13].

$$P(X) = \frac{P(X|H) * P(H)}{P(X)} \quad (3)$$

Dimana:
X : *data testing* yang kelasnya belum diketahui.
H : hipotesis data X yang kelasnya lebih spesifik.
P(H) : peluang dari hipotesa H.
P(X) : *predictor prior* yang merupakan probabilitas X.
P(X|H) :*likelihood* yang merupakan probabilitas hipotesis X berdasarkan kondisi H.

C. Algoritma *Random Forest*

Random forest adalah salah satu algoritma dari *machine learning* untuk mengembangkan *decision tree*. *Random forest* dapat dianggap sebagai kombinasi dari beberapa buah *decision tree* [14]. *Random forest* merupakan salah satu bentuk yang berasal dari metode *ensemble* yang bertujuan untuk meningkatkan akurasi klasifikasi data dari sebuah pemilah tunggal yang tidak stabil melalui kombinasi dari banyak jenis metode yang sama sebagai proses *majority voting* untuk menghasilkan prediksi tentang klasifikasi akhir [15].



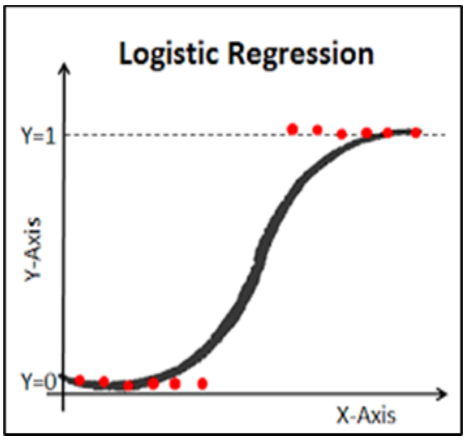
Gambar 3. Gambaran *Random Forest* [15]

Langkah-langkah *random forest* sebagai berikut :

1. Melakukan pengambilan contoh acak yang berukuran n dengan perbaikan pada gugus data. Tahapan ini merupakan tahapan dari *bootstrap*.
2. Melakukan pengambilan sampel data secara acak (*random*) dengan kemungkinan pengambilan sampel data yang sama (tahapan ini disebut tahapan *bootstrap*).
3. Ulangi langkah 1 dan 2 sebanyak k kali, sehingga terbentuk sebuah hutan yang terdiri atas k pohon.

D. Algoritma *Logistic Regression*

Logistic Regression adalah algoritma yang dapat memisahkan *dataset* menjadi dua partisi yang disebut dengan *binary classification* menggunakan metode prediksi probabilitas. Output yang dihasilkan oleh algoritma *Logistic Regression* bersifat kategori dan kualitatif [14].



Gambar 4. Pola Kurva Dalam *Logistic Regression* [16]

Pada Gambar 4 di atas menunjukkan jika kurva munuju ke positif, nilai y (*output*) maka akan diprediksi menjadi 1, jika kurva menuju ke negatif, nilai y (*output*) maka diprediksi menjadi 0. Jika dirumuskan:

$$p \geq 0.5, class = 1 \quad (4)$$

$$p < 0.5, class = 0 \quad (5)$$

Apabila jumlah variabel tidak dibatasi, Persamaan *Logistic Regression* dinyatakan dengan rumus sebagai berikut:

$$P(Y = 1) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \quad (6)$$

Atau

$$logit(p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad (7)$$

Dimana:

Ln : Logaritma natural
β_0 : Konstanta
β_1 : Koefisien masing-masing variabel
p : Probabilitas logistik

E. Seleksi Fitur *Mutual Information*

Mutual Information (MI) adalah salah satu metode seleksi fitur yang sudah sering digunakan sebelum melakukan pemodelan. Metode ini melakukan perhitungan dengan mengukur jumlah informasi yang terdapat pada fitur dan mengetahui serta menetapkan fitur tersebut sebagai nilai tertinggi [17]. Berdasarkan hasil pengukuran tersebut dapat diketahui fitur-fitur yang memiliki pengaruh dalam melakukan proses klasifikasi yang tepat. Perhitungan yang terdapat pada metode *Mutual Information* memiliki rumus yang ditunjukkan pada persamaan berikut:

$$I(X,Y) = \sum_{y\in Y}\sum_{x\in X} p(X,Y)^{(x,y)\log(\frac{p(X,Y)^{(x,y)}}{pX^{(x)}pY^{(y)}}} \quad (8)$$

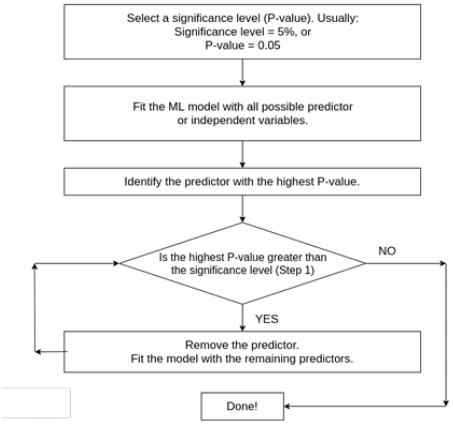
F. Seleksi Fitur *Forward Selection*

Menurut Saleh dalam [18] *Forward Selection* adalah metode seleksi fitur yang menyeleksi atribut berdasarkan kofisien korelasi dan meregresikan semua atribut bebas hingga diperoleh persamaan yang sempurna. *Forward Selection* dimulai dengan pemilihan atribut yang kosong serta dalam setiap putaran ditambahkan setiap atribut yang tidak terpakai. Untuk setiap atribut yang ditambahkan diperkirakan menggunakan kinerja operator batin, misalkan *cross validation*. Dimana hanya atribut yang memiliki kinerja tinggi untuk digunakan dalam *modelling*. Untuk prosedur *Forward Selection* dapat dirumuskan sebagai berikut.

1. Menentukan model awal $\hat{y} = b_0$
2. Memasukkan variabel respon dengan setiap variabel berprediktor, misalnya $X_1, X_2, \ldots X_n$ yang terkait $\hat{y} = b_0 + b_1 X_1$
3. Uji F terhadap peubah pertama yang dipilh. Jika $F_{hitung} < F_{tabel}$ maka peubah terpilih dibuang proses dihentikan. Apabila $F_{hitung} < F_{tabel}$ maka peubah terpilih memiliki pengaruh nyata terhadap peubah terkait *y*, sehingga layak untuk diperhitungkan didalam model.
4. Masukan peubah bebas terpilih (yang paling signifikan) ke dalam model. Misalkan X_2, sehingga membentuk suatu model $\hat{y} = b_0 + b_1 X_1 + b_2 X_2$
5. Uji F, jika $F_{hitung} < F_{tabel}$ maka proses dihentikan dan model terbaik adalah model berikutnya.

G. Seleksi Fitur *Backward Elimination*

Backward Elimination adalah salah satu metode seleksi fitur yang memiliki fungsi untuk mengoptimalkan performa suatu model dengan sistem kerja pemilihan mundur. Pemilihan atribut/fitur dilakukan dengan cara kedepan yakni menguji semua atribut/fitur kemudian menghapus atribut-atribut yang dianggap tidak signifikan [19]. Selain itu metode ini memiliki beberapa keunggulan seperti peningkatan waktu pelatihan, penurunan kompleksitas dan peningkatan kinerja dan akurasi. Oleh karena itu, metode ini sangat berguna untuk memilih fitur-fitur yang ingin digunakan sebelum dilakukan pemodelan.



Gambar 5. Proses *Backward Elimination*

Pertama, semua atribut yang akan diuji dalam model regresi, dengan tingkat signifikansi 0,05. Ketika nilai p dari suatu atribut lebih besar dari tingkat signifikansi (nilai P > 0,05). Langkah ini diulangi sampai semua atribut menjadi signifikan (nilai P < 0,05). Terakhir, model ini dilengkapi dengan serangkaian atribut baru [14].

H. Seleksi Fitur *Recursive Elimination*

Recursive Elimination adalah metode seleksi fitur yang bekerja dengan cara mengurangi fitur yang tidak saling berhubungan dan dilakukan secara terus-menerus secara rekursif hingga fitur terbaik diperoleh untuk digunakan dalam membangun model. Fitur-fitur yang memiliki hubungan tinggi tersebut digunakan dalam membangun model dan dihitung nilai akurasinya. Fitur diurutkan secara relatif sesuai urutan eliminasi. Perangkingan fitur yang dipilih akan membuat model yang cocok dengan semua atribut.

Atribut dengan nilai p-*value* tertinggi akan dipilih sedangkan jika nilai p-*value* lebih besar dari tingkatan signifikansi maka akan ditolak. Dapat diingat bahwa, model ini dibangun dengan nilai atribut atau fitur yang tersisa.

Atribut akan dihapus berulang kali sampai model mendapatkan akurasi yang optimal. Dalam melakukan perangkingan fitur, metode seleksi fitur *Recursive Elimination* menggunakan nilai koefisien dari atribut, dimana semakin tinggi nilai koefisien maka semakin baik perangkingannya dan semakin besar kemungkinan untuk dipilih [20].

Untuk menentukan perangkingan pada setiap atribut di dalam data, diperlukan rumus untuk menghitung bobot *w*. Berikut rumus dalam menghitung bobot fitur [21]:

$$w = \sum_{i=1}^{k} a_i \, y_i x_i \tag{9}$$

I. *Cross Validation*

Menurut Written & Frank [22], *cross validation* adalah bentuk sederhana dari teknik statistik. 10-*fold cross validation* merupakan jumlah standar dalam memprediksi tingkat *erorr* pada data. Metode *k-fold cross validation* melakukan generalisasi dengan membagi data kedalam *k* bagian berukuran sama. Selama proses berlangsung, salah satu dari bagian dipilih sebagai data uji, dan sisanya digunakan untuk data latih. Langkah ini di ulangi sebanyak nilai *k* sehingga setiap bagian digunakan untuk data uji tepat satu kali. Metode *k-fold cross validation* menetapkan *k* = N, yang berdasar pada ukuran *dataset*. Pendekatan ini memiliki kelebihan yakni dalam penggunaan data sebanyak mungkin dalam proses *training* agar model yang dibangun memiliki performa yang optimal. Data *test* secara efektif mencakup keseluruhan *dataset*. Pada teknik ini memiliki kekurangan yakni banyaknya komputasi untuk mengulangi prosedur sebanyak N kali sehingga menyebabkan proses *modelling* memakan waktu lebih lama dari teknik lainnya. *K-fold cross validation* juga merupakan salah satu teknik yang biasa digunakan dalam mengevaluasi keakuratan model [23]. Ilustrasi proses pada 10-*fold cross validation* dapat dilihat pada Gambar 6.

Gambar 6. Ilustrasi 10-*Fold Cross Validation* [23]

J. *Confusion Matrix*

Confusion matrix merupakan suatu teknik yang dapat digunakan untuk mengetahui seberapa akurat model klasifikasi menggunakan tabel *confusion matrix* [14]. *Confusion matrix* memberikan ringkasan dari semua hasil prediksi yang dihasilkan dengan membandingkan antara hasil prediksi dan hasil yang diharapkan [24].

Confusion matrix memberikan keputusan yang diperoleh selama pelatihan dan pengujian, dan *confusion matrix* juga memberikan penilaian kinerja klasifikasi berdasarkan apakah objek itu benar atau salah [25].

Tabel 1. Tabel *Confusion Matrix* [26]

Class		***Actual***	
		TRUE	***FALSE***
Prediction	*TRUE*	*True Positive (TP)*	*False Positive (FP)*
	FALSE	*False Negative (FN)*	*True Positive (TP)*

Berikut cara mengukur *confusion matrix* [26]:

a. *Accuracy* (Akurasi)
Mengukur akurasi model. Rumusnya jumlah prediksi benar dibagi dengan total seluruh populasi.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{10}$$

b. *Precision* (Ketepatan)
Mengukur jumlah data yang sukses diprediksi positif, dibandingkan dengan seluruh data yang diprediksi positif, yang kenyataannya benar dan tidak benar.

$$precision = \frac{TP}{TP+FP} \tag{11}$$

c. *Sensitivity (Recall)*
Mengukur banyaknya data yang sukses saat diprediksi sebagai positif dibandingkan dengan seluruh data yang pada kenyataannya positif.

$$sensitivity = \frac{TP}{TP+FN} \quad (12)$$

III. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Penelitian ini menggunakan data sekunder yang bersumber dari Bagian Administrasi Akademik dan Bagian Biro Keuangan UMKT 2019 - 2021. Data yang digunakan adalah data profil mahasiswa yang didapatkan pada Bagian Administrasi Akademik sedangkan data keterlambatan pembayaran biaya kuliah didapatkan pada bagian keuangan UMKT. Data-data tersebut meliputi NIM, nama mahasiswa, jenis kelamin, fakultas, program studi, angkatan, pendidikan ibu, penghasilan ibu, pendidikan ayah, penghasilan ayah, dan status pembayaran biaya kuliah. Data awal yang didapatkan dari Bagian Biro Keuangan sebanyak 8.833 *record* data mahasiswa yang terlambat melakukan pembayaran SPP dan 30.811 *record* data mahasiswa yang tepat waktu melakukan pembayaran SPP dengan total 39.644 *record* di tahun 2019 - 2021.

Tabel 2. Gambaran Data yang Didapatkan dari Bagian Administrasi Akademik

No	NIM	Nama Mahasiswa	Fakultas	Program Studi	Angkatan	Gender	Penghasilan Ayah	Penghasilan Ibu	Pendidikan Ayah	Pendidikan Ibu
1	1811102411002	ADAM MUH. AGUSSALIM	Ilmu Keperawatan	Keperawatan	2018	L	Kurang dari Rp. 500,000	Rp. 2,000,000 - Rp. 4,999,999	SMA	S1
…	…	…	…	…	…	…	…	…	…	…
5658	17111024430021	ROBBY HADI CAHYONO	Sains Dan Teknologi	Teknik Sipil	2017	L	Rp. 1,000,000 - Rp. 1,999,999	Kurang dari Rp. 500,000	SMP	SD

Tabel 3. Gambaran Data yang Didapatkan dari Bagian Biro Keuangan

No	NIM	Nama Mahasiswa	Terlambat
1	1811102411002	ADAM MUH. AGUSSALIM	Tidak
…	…	…	…
3451	17111024430021	ROBBY HADI CAHYONO	Tidak

B. Data *Preparation*

Pada tahapan ini, *dataset* akan melalui proses data *selection*, data *transformation*, dan data *cleaning*. Semua proses tersebut dilakukan agar *dataset* yang digunakan memiliki kualitas yang baik.

1. *Data Selection*

Proses pemilihan data yang dilakukan pada *dataset* pembayaran biaya SPP di UMKT dilakukan untuk memilih atribut yang sesuai. Proses pemilihan data dibatasi pada rentang tahun 2019 hingga 2021. Atribut yang digunakan dalam penelitian ini adalah fakultas, program studi, angkatan, gender, penghasilan ayah, penghasilan ibu, pendidikan ayah, pendidikan ibu dan status pembayaran SPP yang dijadikan kelas target dalam penelitian ini, sedangkan atribut-atribut yang dihapus karena tidak digunakan dalam pemodelan adalah NIM dan nama mahasiswa.

Tabel 3. Gambaran Data yang Telah Diseleksi

No	Fakultas	Program Studi	Angkatan	Gender	Penghasilan Ayah	Penghasilan Ibu	Pendidikan Ayah	Pendidikan Ibu	Status Pembayaran SPP
1	Ilmu Keperawatan	Keperawatan	2018	L	Kurang dari Rp. 500,000	Rp. 2,000,000 - Rp. 4,999,999	SMA	S1	Tepat
…	…	…	…	…	…	…	…	…	
39644	Sains Dan Teknologi	Teknik Sipil	2017	L	Rp. 1,000,000 - Rp. 1,999,999	Kurang dari Rp. 500,000	SMP	SD	Tepat

DOI: 10.34148/teknika.v12i1.601

2. *Data Transformation*

Proses transformasi data yang dilakukan adalah mengubah nilai dari atribut-atribut yang bersifat kategorikal menjadi numerik, hal ini dilakukan karena pada penerapan menggunakan *library sklearn* hanya bisa menerima nilai atribut numerik. Beberapa atribut yang ditransformasi datanya meliputi fakultas, program studi, angkatan, gender, penghasilan ayah, penghasilan ibu, pendidikan ayah dan pendidikan ibu. Contoh data yang telah dilakukan transformasi dapat dilihat pada Tabel 5.

Tabel 4. Gambaran Data yang Telah Ditransformasi

No	Fakultas	Program Studi	Angkatan	Gender	Penghasilan Ayah	Penghasilan Ibu	Pendidikan Ayah	Pendidikan Ibu	Status Pembayaran SPP
1	3	3	1	0	0	3	4	8	Tepat
…	…	…	…	…	…	…	…	…	
39644	7	12	0	0	2	0	2	3	Tepat

3. *Data Cleaning*

Pembersihan data yang dilakukan pada *dataset* keterlambatan biaya kuliah adalah menghapus nilai data yang bernilai 0 untuk atribut yang bernilai numerik dan memiliki nilai #N/A (*no value is available*) atau tidak memiliki nilai. Dalam proses ini total data yang dibersihkan adalah 10.099 *record* sehingga data yang dimiliki setelah dilakukan proses *cleaning* yakni 29.545 *record*.

C. Pembagian Data

Pembagian data *training* dan data *testing* dalam penelitian ini menggunakan teknik *cross validation*, dimana nilai *k-fold* yang digunakan adalah *k*=10.

D. Seleksi Fitur dan Pemodelan

Pada tahap ini, ada dua tahapan yang dilakukan yakni penerapan algoritma seleksi fitur terhadap *dataset* dan hasil dari seleksi fitur tersebut diterapkan ke dalam pemodelan algoritma klasifikasi. Pada tahap seleksi fitur dilakukan perangkingan terhadap atribut-atribut yang ada, mulai dari atribut yang memiliki pengaruh terbesar terhadap hasil prediksi diberikan peringkat 1 sampai atribut yang memiliki pengaruh terkecil atau justru tidak memiliki pengaruh terhadap prediksi diberikan peringkat 8. Proses perankingan atribut tersebut menggunakan bantuan *library sklearn*. Pada tahap pemodelan dilakukan beberapa percobaan penggunaan atribut yang mengacu pada hasil perankingan atribut yang dihasilkan dari penerapan seleksi fitur, kemudian hasil tersebut akan dikomparasikan untuk menemukan algoritma seleksi fitur dengan peningkatan akurasi tertinggi.

1. *Mutual Information*

Pada seleksi fitur menggunakan algoritma *Mutual Information* (MI) ditemukan bahwa atribut angkatan memiliki pengaruh tertinggi terhadap hasil prediksi sedangkan atribut pendidikan ibu memiliki tingkat pengaruh paling rendah dibandingkan atribut-atribut lainnya. Berikut hasil perangkingan atribut yang dihasilkan oleh algoritma MI:

Tabel 5. Hasil Perangkingan Atribut Oleh Algoritma MI.

Atribut	Nilai MI	Ranking
Angkatan	0,0478	1
Program Studi	0,0267	2
Fakultas	0,0250	3
Gender	0,0111	4
Penghasilan Ibu	0,0084	5
Pendidikan Ayah	0,0052	6
Penghasilan Ayah	0	7
Pendidikan Ibu	0	8

Hasil penerapan seleksi fitur MI diperoleh bahwa algoritma *naïve bayes* mendapatkan peningkatan akurasi tertinggi sebesar 1,2% dengan penggunaan 7 atribut dibandingkan dengan pemodelan tanpa seleksi fitur sedangkan *algoritma K-NN* sebaliknya justru menurunkan akurasi dibandingkan algoritma klasifikasi lainnya yakni sebesar -0,6%.

Tabel 6. Hasil Peningkatan Akurasi di Setiap Algoritma Klasifikasi Dengan MI

No	Algoritma Klasifikasi	Jumlah atribut	Peningkatan Akurasi Tertinggi
1.	K-NN	7	-0,6%
2.	C4.5	2	0,7%
3.	*Naive Bayes*	7	1,2%
4.	*Random forest*	2	0,1%
5.	*Logistic Regression*	3	0,2%
Rata-Rata Hasil Peningkatan Akurasi Tertinggi			**0,32%**

2. *Forward Selection*

Proses perankingan pada Tabel 8 didapatkan melalui urutan atribut yang dihasilkan *Forward Selection* menggunakan *library sklearn*, dimana atribut yang memiliki pengaruh terbesar mendapatkan rangking 1 dan begitu seterusnya hingga atribut dengan pengaruh terkecil. Berikut hasil perangkingan atribut yang dihasilkan oleh algoritma *Forward Selection*:

Tabel 8. Hasil Perangkingan Atribut Oleh Algoritma FS

Atribut	Ranking
Angkatan	1
Program Studi	2
Penghasilan Ibu	3
Fakultas	4
Pendidikan Ibu	5
Pendidikan Ayah	6
Gender	7
Penghasilan Ayah	8

Pada Tabel 8 ditemukan bahwa atribut angkatan memiliki pengaruh tertinggi terhadap hasil prediksi sedangkan atribut penghasilan ayah memiliki tingkat pengaruh paling rendah dibandingkan atribut-atribut lainnya. Hasil penerapan seleksi fitur *Forward Selection* dihasilkan bahwa algoritma C4.5 mendapatkan peningkatan akurasi tertinggi sebesar 0,7% dengan penggunaan 7 atribut, sedangkan algoritma K-NN mendapatkan peningkatan akurasi terendah dibandingkan algoritma klasifikasi lainnya yakni sebesar -0,9%. Adapun hasil peningkatan akurasi tertinggi di setiap algoritma klasifikasi dapat dilihat pada Tabel 9 berikut.

Tabel 9. Hasil Peningkatan Akurasi di Setiap Algoritma Klasifikasi Dengan *Forward Selection*.

No	Algoritma Klasifikasi	Jumlah atribut	Peningkatan Akurasi Tertinggi
1.	K-NN	6	-0,9%
2.	C4.5	7	0,7%
3.	*Naive Bayes*	2	0,2%
4.	*Random forest*	2	0,1%
5.	*Logistic Regression*	5	0,2%
Rata-Rata Hasil Peningkatan Akurasi Tertinggi			**0,02%**

3. *Backward Elimination*

Pada seleksi fitur *Backward Elimination* (BE) proses perankingan juga hampir sama dengan *forward selecion*, namun urutan atribut pada BE dimulai dari atribut yang memiliki pengaruh paling kecil. Berikut hasil perangkingan atribut yang dihasilkan oleh algoritma *Backward Elimination*:

Tabel 10. Hasil Perangkingan Atribut Oleh Algoritma *Backward Elimination*.

Atribut	Ranking
Angkatan	1
Program Studi	2
Gender	3
Pendidikan Ayah	4
Pendidikan Ibu	5
Penghasilan Ayah	6
Fakultas	7
Penghasilan Ibu	8

Pada Tabel 10 diketahui bahwa atribut angkatan memiliki pengaruh tertinggi terhadap hasil prediksi sedangkan atribut penghasilan ibu memiliki tingkat pengaruh paling rendah dibandingkan atribut-atribut lainnya. Hasil penerapan seleksi fitur *Backward Elimination* didapatkan hasil bahwa algoritma *naïve bayes* mendapatkan peningkatan akurasi tertinggi sebesar 2,1%, sedangkan algoritma K-NN mendapatkan peningkatan akurasi terendah dibandingkan algoritma klasifikasi lainnya yakni sebesar 0,3%. Hasil peningkatan akurasi tertinggi di setiap algoritma klasifikasi dapat dilihat pada tabel berikut.

Tabel 11. Hasil Peningkatan Akurasi di Setiap Algoritma Klasifikasi Dengan *Backward Elimination*.

No	Algoritma Klasifikasi	Jumlah atribut	Peningkatan Akurasi Tertinggi
1.	K-NN	7	-0,3%
2.	C4.5	2	0,7%
3.	*Naive Bayes*	2	2,1%
4.	*Random forest*	2	0,1%
5.	*Logistic Regression*	6	0%
Rata-Rata Hasil Peningkatan Akurasi Tertinggi			**0,52%**

4. *Recursive Elimination*

Pada seleksi fitur menggunakan algoritma *Recursive Elimination* (RE) proses perankingan atribut dilakukan dengan menghitung bobot pada setiap atribut yang dapat dilihat pada persamaan 9, dimana atribut dengan bobot tertinggi, mendapatkan ranking tertinggi. Berikut hasil perangkingan atribut yang dihasilkan oleh algoritma *Recursive Elimination*:

Tabel 12. Hasil Perangkingan Atribut oleh Algoritma *Recursive Elimination*.

Atribut	Ranking
Angkatan	1
Program Studi	2
Penghasilan Ibu	3
Fakultas	4
Pendidikan Ibu	5
Pendidikan Ayah	6
Gender	7
Penghasilan Ayah	8

Pada Tabel 12 ditemukan bahwa atribut angkatan memiliki pengaruh tertinggi terhadap hasil prediksi sedangkan atribut gender memiliki tingkat pengaruh paling rendah dibandingkan atribut-atribut lainnya. Hasil seleksi fitur menggunakan algoritma *Recursive Elimination* terhadap beberapa algoritma klasifikasi diperoleh bahwa algoritma C4.5 mendapatkan peningkatan akurasi tertinggi sebesar 0,7% dengan penggunaan 2 atribut dibandingkan dengan pemodelan tanpa seleksi fitur. Sedangkan algoritma K-NN mendapatkan peningkatan akurasi terendah dibandingkan algoritma klasifikasi lainnya yakni sebesar -0,8%. Hasil peningkatan akurasi tertinggi di setiap algoritma klasifikasi dapat dilihat pada tabel berikut.

Tabel 13. Hasil Peningkatan Akurasi di Setiap Algoritma Klasifikasi Dengan *Recursive Elimination*

No	Algoritma Klasifikasi	Jumlah atribut	Peningkatan Akurasi Tertinggi
1.	K-NN	7	-0,8%
2.	C4.5	2	0,7%
3.	*Naive Bayes*	7	0,4%
4.	*Random forest*	2	0,1%
5.	*Logistic Regression*	3	0,1%
Rata-Rata Hasil Peningkatan Akurasi Tertinggi			**0,1%**

E. Evaluasi

Pada tahap ini dilakukan komparasi hasil seleksi fitur yang diterapkan pada data keterlambatan biaya kuliah di UMKT. Hasil yang dikomparasikan adalah nilai rata-rata dari hasil peningkatan akurasi pada setiap algoritma seleksi fitur, dimana perangkingan atribut yang dihasilkan melalui seleksi fitur diterapkan kedalam algoritma klasifikasi untuk menguji akurasi ataupun kinerja dari algoritma seleksi fitur tersebut. Berikut hasil komparasi dari pengujian algoritma seleksi fitur.

Tabel 14. Hasil Peningkatan Akurasi Tertinggi di Setiap Algoritma Klasifikasi.

No	Algoritma Seleksi Fitur	Algoritma Klasifikasi	Peningkatan Akurasi Tertinggi	Rata-Rata Peningkatan Akurasi Tertinggi
1.	*Mutual Information*	K-NN	-0,6%	**0,32%**
		C4.5	0,7%	
		Naive Bayes	1,2%	
		Random forest	0,1%	
		Logistic Regression	0,2%	
2.	*Forward Selection*	K-NN	-0,9%	**0,02%**
		C4.5	0,7%	
		Naive Bayes	0,1%	
		Random forest	0,1%	
		Logistic Regression	0%	
3.	*Backward Elimination*	K-NN	-0,3%	**0,52%**
		C4.5	0,7%	
		Naive Bayes	2,1%	
		Random forest	0,1%	
		Logistic Regression	0%	
4.	*Recursive Elimination*	K-NN	-0,8%	**0,1%**
		C4.5	0,7%	
		Naive Bayes	0,4%	
		Random forest	0,1%	
		Logistic Regression	0,1%	

Hasil komparasi algoritma seleksi fitur yang digunakan pada data keterlambatan biaya kuliah diperoleh bahwa algoritma *Backward Elimination* memiliki peningkatan akurasi tertinggi dengan nilai rata-rata akurasi tertinggi sebesar 0,52%. Berdasarkan Tabel 14 juga ditemukan bahwasanya penerapan seleksi fitur menggunakan *backward elimintation* pada algoritma *naïve bayes* mendapatkan peningkatan akurasi tertinggi dibandingkan dengan algoritma klasifikasi lainnya yakni sebesar 2,1%, sedangkan penerapan seleksi fitur pada algoritma K-NN justru menurunkan akurasi pemodelan disetiap percobaan yang dilakukan pada penelitian ini, dengan nilai penurunan akurasi terendah mencapai 6,6%. Terakhir, jika dilihat dari performa algoritma klasifikasi, ditemukan bahwa performa terbaik didapatkan oleh algoritma *random forest* dan c4.5 dengan nilai akurasi sebesar 62,6%, *precision* 65%, *recall* 63% dan *f1-score* 61%.

IV. KESIMPULAN

Dari hasil pengujian yang dilakukan pada beberapa algoritma seleksi fitur yang diterapkan pada data keterlambatan biaya kuliah di UMKT dengan penggunaan teknik 10-*fold cross validation* dan uji coba penggunaan atribut berdasarkan perangkingan atribut, diperoleh bahwa algoritma seleksi fitur *Backward Elimination* menghasilkan peningkatan akurasi tertinggi dengan nilai rata-rata sebesar 0,52%, sedangkan penerapan *Backward Elimination* pada algoritma *naïve bayes* mendapatkan peningkatan akurasi tertinggi dibandingkan algoritma klasifikasi lainnya yakni mencapai 2,1% dengan penggunaan 2 atribut. Dari hasil tersebut dapat disimpulkan bahwa algoritma seleksi fitur *Backward Elimination* lebih baik daripada algoritma seleksi fitur lainnya. Sedangkan hasil dari perangkingan atribut yang dilakukan dalam penerapan seleksi fitur dihasilkan atribut yang memiliki pengaruh paling besar dalam klasifikasi keterlambatan pembayaran biaya kuliah adalah atribut angkatan.

REFERENSI

[1] Ropidin and S. Riyanto, "Dampak Pemutusan Hubungan Kerja Pada Perusahaan Farmasi Terkait Covid-19 Di Indonesia," *J. Syntax Transform.*, vol. 1,

no. 5, pp. 167–174, 2020, doi: 10.46799/jst.v1i5.63.
[2] E. Buulolo, *Data Mining Untuk Perguruan Tinggi*. Sleman: Deepublish, 2020.
[3] R. W. Abdullah, Kusrini, and E. T. Luthfii, "Prediksi Keterlambatan Pembayaran Spp Sekolah Dengan Metode K-Nearest Neighbor (Studi Kasus Smk Al-Islam Surakarta)," *J. Inf. Interaktif*, vol. 4, no. 3, pp. 160–164, 2019.
[4] V. S. Ginting, K. Kusrini, and E. T. Luthfi, "Penerapan Algoritma C4.5 Dalam Memprediksi Keterlambatan Pembayaran Uang Sekolah Menggunakan Python," *JurTI (Jurnal Teknol. Informasi)*, 2020, [Online]. Available: http://jurnal.una.ac.id/index.php/jurti/article/view/1101.
[5] F. Firman, *Pemodelan Klasifikasi Keterlambatan Pembayaran UKT Mahasiswa IPB dengan Random Forest dan AdaBoost*. 202.124.205.241, 2021.
[6] T. H. Apandi, R. B. Maulana, R. Piarna, and D. Vernanda, "Menganalisis Kemungkinan Keterlambatan Pembayaran Spp Dengan Algoritma C4. 5 (Studi Kasus Politeknik Tedc Bandung)," *J. Techno Nusa Mandiri*, vol. 16, no. 2, pp. 93–98, 2019, [Online]. Available: https://core.ac.uk/download/pdf/229771622.pdf.
[7] F. Firman, *Pemodelan Klasifikasi Keterlambatan Pembayaran UKT Mahasiswa IPB dengan Random Forest dan AdaBoost*. 202.124.205.241, 2021.
[8] D. Rohmayani, "Analysis of Student Tuition Fee Pay Delay Prediction Using Naive Bayes Algorithm With Particle Swarm Optimation Optimazation (Case Study : Politeknik Tedc Bandung)," *J. Teknol. Inf. dan Pendidik.*, vol. 13, no. 2, pp. 1–8, 2020, doi: 10.24036/tip.v13i2.317.
[9] M. Muqorobin, K. Kusrini, and E. T. Luthfi, "Optimasi Metode Naive Bayes Dengan Feature Selection Information Gain Untuk Prediksi Keterlambatan Pembayaran Spp Sekolah," *J. Ilm. SINUS*, 2019, [Online]. Available: https://p3m.sinus.ac.id/jurnal/index.php/e-jurnal_SINUS/article/view/378.
[10] R. Rukminingsih, G. Adnan, and M. A. Latief, *Metode Penelitian Pendidikan (Kuantitatif, Kualitatif & Penelitian Tindakan Kelas)*. Yogyakarta: Erhaka Utama, 2020.
[11] D. Nofriansyah, *Konsep Data Mining Vs Sistem Pendukung Keputusan*. Yogyakarta: Deepublish, 2015.
[12] A. Rohman, V. Suhartono, and C. Supriyanto, "Penerapan Agoritma C4.5 Berbasis Adaboost Untuk Prediksi Penyakit Jantung," *J. Teknol. Inf.*, vol. 13, pp. 13–19, 2017.
[13] J. Suntoro, *Data mining : algoritma dan implementasi dengan pemrograman PHP*. Jakarta: Elex Media Komputindo, 2019.
[14] R. Primartha, *Algoritma Machine Learning*. Informatika, 2021.
[15] M. I. Putra, "Sistem Rekomendasi Kelayakan Kredit Menggunakan Metode Random Forest pada BRI Kantor Cabang Pelaihari," Universitas Islam Negeri Sunan Ampel Surabaya, 2019.
[16] R. Tyasnurita and A. Y. M. Pamungkas, "Deteksi Diabetik Retinopati menggunakan Regresi Logistik," *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 130–135, 2020, doi: 10.33096/ilkom.v12i2.578.130-135.
[17] J. Ling, I. P. E. N. Kencana, and T. B. Oka, "Analisis Sentimen Menggunakan Metode Naive Bayes Classifier Dengan Seleksi Fitur Chi Square," *E-Jurnal Mat.*, vol. 3, no. 3, pp. 92–9, 2014, doi: 10.24843/mtk.2014.v03.i03.p070.
[18] M. Hasan, "Prediksi Tingkat Kelancaran Pembayaran Kredit Bank Menggunakan Algoritma Naïve Bayes Berbasis Forward Selection," *Ilk. J. Ilm.*, vol. 9, no. 3, pp. 317–324, 2017, doi: 10.33096/ilkom.v9i3.163.317-324.
[19] A. Bode, "K-Nearest Neighbor Dengan Feature Selection Menggunakan Backward Elimination Untuk Prediksi Harga Komoditi Kopi Arabika," *Ilk. J. Ilm.*, vol. 9, no. 2, pp. 188–195, 2017, doi: 10.33096/ilkom.v9i2.139.188-195.
[20] A. R. I. Pratama, S. A. Latipah, and B. N. Sari, "Optimasi Klasifikasi Curah Hujan Menggunakan Support Vector Machine (SVM) Dan Recursive Feature Elimination (RFE)," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 07, no. 02, pp. 314–324, 2022, [Online]. Available: http://jurnal.stkippgritulungagung.ac.id/index.php/jipi/article/view/2675%0Ahttps://jurnal.stkippgritulungagung.ac.id/index.php/jipi/article/download/2675/1166.
[21] J. C. Jeong, "Enhanced Recursive Feature Elimination," in *Proceedings - 6th International Conference on Machine Learning and Applications, ICMLA 2007*, 2007, no. January 2008, pp. 429–435, doi: 10.1109/ICMLA.2007.35.
[22] I. H. Written and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco: Morgan Kaufmann, 2005.
[23] R. Ramadhan, "Perbandingan Produk Smartphone Berdasarkan Klasifikasi Komentar Website Menggunakan Metode Naive Bayes Classification," Universitas Islam Negeri Sultan Syarif Kasim Riau, 2017.
[24] I. D. Id, *Machine Learning: Teori, Studi Kasus dan Implementasi Menggunakan Python*. Riau: UR Press, 2021.
[25] J. Indriyanto, *Algoritma K-Nearrest Neighbor Untuk Prediksi Nasabah Asuransi*. Pekalongan: PT Nasya Expanding Management, 2021.
[26] D. Kurniawan, *Pengenalan Machine Learning dengan Python*. Jakarta: Elex Media Komputindo, 2020.