Terbit *online* pada laman : http://teknosi.fti.unand.ac.id/

Jurnal Nasional Teknologi dan Sistem Informasi

| ISSN (Print) 2460-3465 | ISSN (Online) 2476-8812 |

Artikel Penelitian

Seleksi Fitur pada Supervised Learning: Klasifikasi Prestasi Belajar Mahasiswa Saat dan Pasca Pandemi COVID-19

Akhas Rahmadeyan [a,b*], *Mustakim* [a,b]

[a]*Program Studi Sistem Informasi, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru 28293, Indonesia*
[b]*Puzzle Research Data Technology (Predatech), Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru 28293, Indonesia*

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi:13 Maret 2023
Revisi Akhir: 05 Mei 2023
Diterbitkan *Online*: 07 Mei 2023

KATA KUNCI

Klasifikasi,
Chi-Square,
Mutual Information,
Seleksi Fitur,
Prestasi Belajar

KORESPONDENSI

E-mail: 11950314479@students.uin-suska.ac.id*

A B S T R A C T

Dampak pandemi COVID-19 membuat lembaga pendidikan mengubah metode belajar menjadi pembelajaran jarak jauh secara daring. Banyak perguruan tinggi menyatakan keprihatinannya pada prestasi akademik mahasiswanya selama periode tersebut, namun disisi lain terdapat mahasiswa yang merasa puas dan nyaman. Di masa pasca pandemi terjadi transisi bertahap untuk kembali ke pembelajaran tatap muka. Ini dilakukan karena pembelajaran tatap muka dianggap lebih efektif dibandingkan pembelajaran daring. Untuk meningkatkan dan memantau kemajuan prestasi akademik mahasiswa demi menghasilkan lulusan yang berkualitas, maka diperlukan analisis terkait perilaku dan prestasi belajar mahaiswa, salah satunya dengan menggunakan teknik data mining. Penelitian ini bertujuan untuk menemukan pola dan faktor yang mempengaruhi prestasi akademik mahasiswa saat dan pasca pandemi COVID-19. Chi-Square dan Mutual Information diterapkan sebagai seleksi fitur untuk menentukan fitur paling berpengaruh pada data. Data dengan fitur optimal akan dilakukan klasifikasi dengan algoritma NBC, CART, RF, dan SVM. Berdasarkan hasil dan analisis yang dilakukan, dapat disimpulkan seleksi fitur efektif dalam meningkatkan kemampuan model dan mempercepat waktu komputasi. Pemodelan dengan 4 algoritma dan 2 metode seleksi fitur menghasilkan CART dengan Chi-Square pada 2 fitur sebagai model terbaik dengan akurasi 89,00%, precision 87,72%, recall 93,57% dan waktu komputasi 0,01814s. Dibandingkan tanpa seleksi fitur, performa CART dengan Chi-Square mengalami peningkatan akurasi sebesar 3% dan waktu komputasi 0,00629s. Chi-Square menjadi seleksi fitur yang efektif pada penelitian ini, yang mana Chi-Square unggul pada rara-rata recall dan waktu komputasi dibandingkan Mutual Information.

1. PENDAHULUAN

Pendidikan menjadi salah satu bidang yang paling terdampak oleh pandemi *COVID-19* [1]–[3]. Sebagai langkah darurat, lembaga pendidikan merespon dengan cepat untuk terus menjalankan fungsinya dengan mengubah metode belajar menjadi pembelajaran jarak jauh secara daring [4], [5]. Hal tersebut dilakukan untuk membatasi kontak fisik demi mengendalikan penyebaran infeksi *COVID-19* [2], [5]. Ini menjadi tantangan baru karena pembelajaran daring belum pernah dipikirkan, direncanakan, dan diinginkan sebelumnya [6], terutama di daerah yang belum berkembang [7]. Tercapainya pembelajaran daring bergantung pada kualitas belajar yang berdampak pada prestasi akademik [8]. Motivasi belajar dan keterampilan teknologi menjadi hal yang mendasar selama pembelajaran daring [9]. Tidak hanya itu, manajemen waktu, komunikasi, kemauan dan keinginan untuk belajar juga menjadi aspek yang penting [8], [10].

Banyak perguruan tinggi telah menyatakan keprihatinannya pada prestasi akademik mahasiswanya selama pembelajaran daring, terutama bagi daerah yang tidak memiliki akses internet [6], [11] serta mahasiswa dengan media belajar yang tidak mendukung [12]–[14]. Disisi lain, terdapat mahasiswa yang merasa puas dan nyaman dengan aktivitas dan lingkungan pembelajaran daring

https://doi.org/10.25077/TEKNOSI.v9i1.2023.21-32

[5], [15], [16]. Hal tersebut berdasarkan analisis prestasi mahasiswa selama *COVID-19*, ditemukan bahwa prestasi beberapa mahasiswa meningkat dibandingkan dengan tahun sebelumnya [16], [17]. Namun, ini tidak dapat dijadikan acuan karena tingkat kecurangan pembelajaran daring dinilai tinggi sehingga mempengaruhi evaluasi pembelajaran [4], [16].

Pada masa pasca pandemi *COVID-19*, terjadi transisi bertahap untuk kembali ke pembelajaran tatap muka dan pada beberapa perguruan tinggi masih tetap menerapkan pembelajaran daring dibeberapa pelajaran [18]. Transisi ini dilakukan karena pembelajaran tatap muka dianggap lebih efektif daripada pembelajaran daring [15]. Pembelajaran tatap muka mendorong keterlibatan belajar sehingga menghasilkan pembelajaran yang sukses, produktif dan bermakna dengan hasil yang lebih baik [19]. Metode pembelajaran secara tatap muka masih disukai dan dipercaya menjadi metode utama dalam meningkatkan keterampilan karena adanya interaksi secara langsung [15].

Dalam beberapa tahun terakhir, telah banyak penelitian yang mempelajari tentang analisis perilaku dan prestasi belajar mahaiswa [20], salah satunya dengan menggunakan teknik data mining [21]–[23]. Ini dilakukan dengan menganalisis dan mempelajari pola dari data gaya belajar, perilaku, ataupun karakteristik mahasiswa di masa lalu untuk memberikan gambaran prestasi belajar di masa mendatang [20], [23]. Dengan menggunakan algoritma klasifikasi data mining, hal tersebut dapat dicapai dengan sangat baik [21], [22]. Pengetahuan yang dihasilkan dapat dimanfaatkan oleh perguruan tinggi untuk menyusun strategi pembelajaran demi menghasilkan lulusan yang berkualitas [21]–[23].

Terdapat banyak algoritma klasifikasi pada data mining, beberapa yang digunakan pada penelitian ini diantaranya *Naïve Bayes Classifier* (NBC), *Classification and Regression Tree* (CART), *Random Forest* (RF), dan *Support Vector Machine* (SVM). NBC merupakan algoritma klasifikasi probabilistik sederhana berdasarkan teorema Bayes [24]–[26]. Berikutnya, CART adalah algoritma berbasis pohon keputusan [27] yang akan menghasilkan pohon klasifikasi jika target bertipe kategori dan menghasilkan pohon regresi jika target bertipe numerik atau kontinu [28], [29]. Kemudian, RF adalah algoritma klasifikasi dengan pendekatan *ensamble learning* berdasarkan *Decission Tree* [26], [30]–[32] yang menciptakan sejumlah pohon keputusan saat melakukan proses klasifikasi [26], [30], [33]. Sedangkan SVM merupakan algoritma pembelajaran mesin berbasis ruang vektor [26], [34] yang efektif dalam menangani kasus kumpulan data yang berdimensi tinggi [31].

Beberapa penelitian sebelumnya pernah menerapkan NBC, CART, RF dan, SVM untuk melakukan klasifikasi. Penelitian [35] menentuan rehabilitasi narkoba dengan algoritma NBC dan KNN menghasilkan NBC sebagai yang terbaik dengan akurasi 80,55%. Selanjutny, penelitian [27] memprediksi penyakit jantung dengan algoritma CART menghasilkan akurasi akurasi prediksi 87%. Kemudian, penelitian [32] memprediksi penyakit hepatitis menggunakan KNN, NBC SVM, MLP dan RF menghasilkan RF sebagai yang terbaik dengan akurasi 92,41%.Penelitian [36] membandingan NBC dan SVM untuk klasifikasi diabetes menghasilkan SVM sebagai algoritma terbaik dengan akurasi 82%.

Selain menerapkan beberapa algoritma klasifikasi, penelitian ini juga menerapkan metode seleksi fitur. Seleksi fitur merupakan langkah penting dalam pengenalan pola, *data mining*, dan *machine learning* [37]–[39]. Berbeda dengan teknik reduksi dimensi, seleksi fitur tidak menghasilkan kombinasi fitur yang baru [40]. Seleksi fitur membuang fitur yang tidak relevan dengan mengidentifikas fitur yang optimal tanpa mengubahnya [37], [39], [41], [42]. Teknik ini dinilai dapat menghindari *overfitting*, mempercepat komputasi, mengurangi kompleksitas, serta meningkatkan kemampuan model [37], [38], [41]–[44]. Pemilihan fitur lebih disukai pada beberapa kasus karena mempertahankan keaslian dari seluruh fitur saat mengurangi jumlah fitur [45]. Namun, menemukan fitur yang optimal tidaklah mudah karena dapat terjadi bias atau varians yang tinggi [37]. Untuk itu diperlukan metode seleksi fitur yang baik dan sesuai dengan data. Adapun beberapa metode seleksi fitur berbasis filter yang umum digunakan adalah *Chi-Square* dan *Mutual Information* [43].

Chi-square adalah metode statistik yang melakukan uji signifikansi data terhadap hubungan antara nilai suatu fitur dengan kelas target [46]–[49]. Sedangkan *Mutual Informaton* adalah bentuk umum dari koefisien korelasi yang digunakan untuk mengukur ketergantungan antar variabel [37], [43], [46]. Penerapan *chi-square* dan *mutual information* pernah dilakukan oleh beberapa penelitian sebelumnya. Penelitian [50] menerapkan seleksi fitur Chi-Square, Relief-F dan SU untuk memprediksi penyakit jantung menghasilkan Chi-Square dengan BayesNet sebagai yang terbaik dengan akurasi 85,00%. Selanjutnya, penelitian [51] mengklasifikasi penyakit hati berlemak menggunakan SMO, IBk, AdaBoostM1 dan BF-Tree dengan seleksi fitur MI menghasilkan SMO sebagai yang terbaik dengan akurasi 95,55% dan sensitivitas 97,77% pada 20 fitur. Penelitian [46] memprediksi kanker payudara menggunakan seleksi fitur Chi-Square dan MI pada random forest menghasilkan akurasi 84,70% dan AUC 0,9023 serta berhasil meningkatkan akurasi dibandingkan tanpa seleksi fitur.

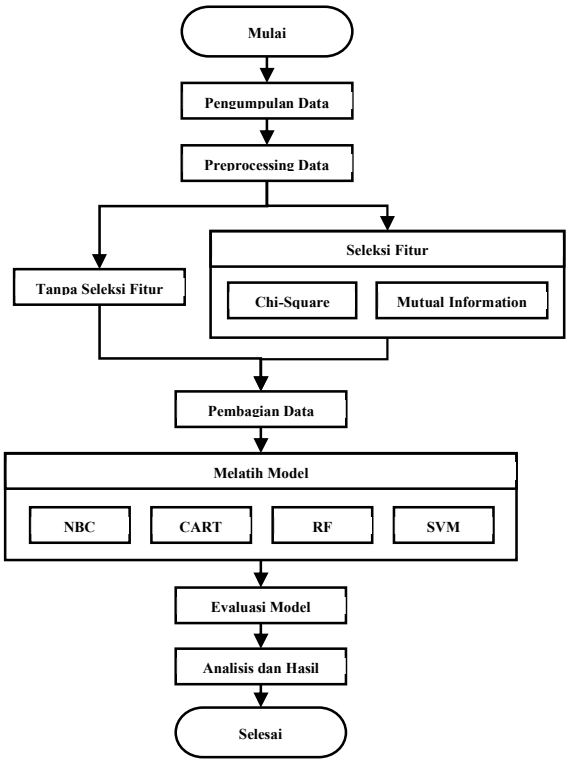
Pada penelitian ini akan menerapkan seleksi fitur untuk mengetahui dan menentukan fitur optimal pada data dengan *Chi-Square* dan *Mutual Information*. Data dengan fitur optimal akan dilakukan pemodelan *supervised learning* untuk dilakukan klasifikasi dengan algoritma NBC, CART, RF, dan SVM. Tujuan dari penelitian ini yaitu menemukan pola dan faktor yang mempengaruhi prestasi akademik mahasiswa saat dan pasca pandemi COVID-19. Selain itu penelitian ini juga bertujuan mengetahui metode seleksi fitur efektif dan mengetahui algoritma terbaik dalam mengklasifikasi data. Hasil dari penelitian ini diharapkan dapat bermanfaat bagi perguruan tinggi dalam menyusun strategi pembelajaran untuk meningkatkan dan memantau kemajuan prestasi akademik mahasiswa demi menghasilkan lulusan yang berkualitas.

2. METODE

Setiap tahapan pada penelitian ini gambarkan pada diagram alir dimulai dari tahap pengumpulan data hingga analisis dan hasil. Data yang digunakan bersumber dari kuesioner dengan target responden mahasiswa yang telah mengikuti pembelajaran daring saat pandemi dan luring setelah pandemi *COVID-19*. Penelitian ini menerapkan algoritma NBC, CART, RF dan SVM dengan seleksi fitur *chi-square* dan *mutual information* serta K-*Fold Cross Validation* sebagai teknik pembagian data untuk

 https://doi.org/10.25077/TEKNOSI.v9i1.2023.21-32

melakukan klasifikasi. Adapun penerapan seleksi fitur dan pemodelan algoritma dilakukan dengan menggunakan *Python* pada *Google Colabolatory*.



Gambar 1. Metodologi Penelitian

2.1. Pengumpulan Data

Pengumpulan data dilakukan dengan menyebarkan kuesioner kepada mahasiswa di beberapa Universitas secara *online* menggunakan *google form*. Kriteria yang ditetapkan sebagai responden yaitu mahasiswa aktif yang telah mengikuti pembelajaran daring saat pandemi dan pembelajaran luring setelah pandemi *COVID-19*.

2.2. Preprocessing Data

Preprocessing merupakan tahapan penting yang bertujuan untuk mengubah data menjadi format yang terstruktur sehingga siap untuk dianalisis [52]. Beberapa tahapan *preprocessing* yang dilakukan pada penelitian ini adalah menghapus baris dan kolom yang tidak relevan, seleksi data, dan melakukan *enconding* dengan mengubah nilai pada data yang bertipe kategorik menjadi representasi numerik.

2.3. Seleksi Fitur

2.3.1. Chi-Square

Chi-Square merupakan salah satu metode seleksi fitur berbasis filter [53], [54]. *Chi-square* adalah metode statistik yang melakukan uji signifikansi data terhadap hubungan antara nilai suatu fitur dengan kelas target [46]–[49]. Nilai *chi-square* yang tinggi menunjukkan suatu fitur memiliki hubungan yang signifikan dengan kelas target [50], [54]–[57]. *Chi-square* umumnya diterapkan pada data kategori ataupun campuran [37], [49], [58]. Berikut adalah persamaan dari *chi-square*:

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

2.3.2. Mutual Information

Mutual Information merupakan metode seleksi fitur berbasis filter [38], [54], [59], [60]. MI populer karena komputasinya yang cepat dan sederhana [54]. MI adalah bentuk umum dari koefisien korelasi yang digunakan untuk mengukur ketergantungan antar variabel [37], [43], [46]. Nilai MI merepresentasikan derajat korelasi antara fitur dan kelas [61]. Semakin tinggi nilai MI, semakin tinggi hubungan antara fitur dengan kelas target [61], [62]. Metode ini dapat diterapkan pada dependensi *linear* dan *non-linear* [60], dapat menangani kasus klasifikasi dan regresi, serta data yang tidak seimbang [44]. Berikut adalah persamaan dari *mutual information*:

$$I(X;Y) = H(X) - H(X|Y) \quad (2)$$

2.4. Pembagian Data

Teknik pembagian data yang digunakan pada penelitian ini adalah K-*Fold Cross Validation*. K-*Fold Cross Validation* merupakan teknik validasi untuk mengevaluasi algoritma *machine learning* pada data yang terbatas dengan membagi data secara acak menjadi K bagian [63], [64]. Setiap bagian dari data dipilih secara bergantian sebagai data latih dan data uji [63]. Kemudian hasil evaluasi diperoleh berdasarkan nilai rata-rata untuk memberikan perkiraan kinerja model [63]–[65]. Adapun pada penelitian ini menerapkan 10 K-*Fold* untuk pembagian data.

2.5. Melatih Model

2.5.1. Naïve Bayes Classifier

Naïve Bayes Classifier (NBC) merupakan algoritma klasifikasi berbasis probabilistik sederhana berdasarkan teorema Bayes [24]–[26]. Algoritma NBC menganggap setiap atribut tidak memiliki ketergantungan satu sama lain atau independen [24], [26], [36], [66]. NBC dapat diterapkan pada atribut numerik dan kategorikal [67]. Selain itu, NBC dinilai dapat bekerja dengan baik bahkan dalam skenario yang rumit [24]–[26].

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

2.5.2. Classification and Regression Tree

Classification and Regression Tree (CART) merupakan algoritma terbaru dari *Decision Tree* [27]. Algoritma CART dapat menangani fitur bertipe kategori dan kontinu [27], [68] serta kasus klasifikasi dan regresi [68], [69]. CART akan menghasilkan pohon klasifikasi jika target bertipe kategori dan akan menghasilkan pohon regresi jika target bertipe numerik atau kontinu [28], [29]. Algoritma CART menggunakan *gini index* untuk membangun pohon keputusan [29], [69], [70]. Setiap node mewakili fitur, cabang mewakili nilai dari fitur, dan daun mewakili target [71].

$$Gini\ (D) = 1 - \sum_{j=1}^{n} p_j^2 \quad (4)$$

2.5.3. *Random Forest*

Random Forest (RF) adalah algoritma klasifikasi dengan pendekatan *ensamble learning* yang dibangun berdasarkan *decission tree* [26], [30]–[32]. Algoritma ini menciptakan sejumlah pohon keputusan saat melakukan proses klasifikasi [26], [30], [33]. Hasil klasifikasi pada RF yaitu dengan mengambil suara mayoritas dari seluruh pohon keputusan yang dihasilkan [33], [52], [72]. Penerapan RF pada pembelajaran mesin sangat populer untuk kasus klasifikasi dan regresi [32]. RF telah terbukti menjadi algoritma klasifikasi yang andal, fleksibel dan efisien dalam berbagai literatur [30], [31].

2.5.4. *Support Vector Machine*

Support Vector Machine (SVM) merupakan algoritma pembelajaran mesin berbasis ruang vektor [26], [34]. SVM mencari hyperlane terbaik yang memisahkan data dengan margin terbesar [31], [32], [72], [73]. Algoritma ini populer untuk tujuan klasifikasi dan regresi [32]. SVM bekerja sangat baik untuk menyelesaikan masalah *linear* dan *non-linear* [73]. Jika kedua kelas dapat dipisahkan secara linier, model klasifikasi linier diterapkan. Jika tidak, maka model klasifikasi nonlinier yang akan diterapkan [57]. Selain itu, SVM efektif dalam menangani kasus kumpulan data yang berdimensi tinggi.

2.6. *Evaluasi Model*

Model yang efektif harus memiliki akurasi, *precision*, dan *recall* yang tinggi dengan waktu komputasi yang rendah [74]. Untuk itu model harus dievaluasi untuk mengetahui keefektifannya. Prosedur evaluasi model klasifikasi dapat dilakukan menggunakan *confussion matrix* [33].

Akurasi merupakan rasio jumlah total prediksi yang dianggap benar dengan jumlah total kejadian [66], [75].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+F} \quad (5)$$

Precision adalah tingkat ketepatan berdasarkan rasio prediksi positif yang benar terhadap jumlah total instance yang diprediksi di kelas positif [66].

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

Recall atau sensitivitas merupakan proporsi nilai pengamatan positif yang diprediksi dengan benar sebagai positif [66], [75].

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

2.7. *Analisis dan Hasil*

Tahapan ini menghasilkan informasi dan pengetahuan berdasarkan permasalahan yang diteliti, yaitu fitur yang paling berpengaruh, kemampuan algoritma *supervised learning* dalam melakukan klasifikasi dan metode seleksi fitur yang efektif pada data hasil belajar mahasiswa saat dan pasca pandemi *COVID-19*.

3. HASIL

3.1. *Pengumpulan Data*

Pengumpulan data dilakukan dengan menyebarkan kuesioner secara *online* dengan menggunakan *google form*. Terdapat 17 pertanyaan utama pada kuesioner dengan 16 pertanyaan akan digunakan sebagai fitur dan 1 pertanyaan sebagai kelas target. Untuk jumlah responden pada penelitian ini tidak dibatasi, karena semakin banyak data maka hasil klasifikasi akan semakin baik. Berikut adalah deskripsi dari pertanyaan pada kuesioner yang akan digunakan pada penelitian ini.

Tabel 1. Deskripsi Data

Fitur (Atribut)	Keterangan	Kategori
X0	Nama Lengkap	-
X1	Cara belajar saat pandemi	0. Mandiri 1. Diskusi 2. Mandiri & Diskusi
X2	Akses internet saat pandemi	0. Kurang Lancar 1. Lancar
X3	Waktu belajar saat pandemi	0. Jarang 1. Sering
X4	Media belajar saat pandemi	0. Kurang Bagus 1. Bagus
X5	Pemberian tugas saat pandemi	0. Jarang 1. Sering
X6	Pemahaman materi saat pandemi	0. Tidak Paham 1. Kurang Paham 2. Paham
X7	Lingkungan tempat tinggal saat pandemi	0. Tidak Kondusif 1. Kondusif
X8	Waktu tidur saat pandemi	0. Kurang 1. Cukup
X9	Cara belajar setelah pandemi	0. Mandiri, 1. Diskusi 2. Mandiri & Diskusi
X10	Akses ke kampus setelah pandemi	0. Dekat 1. Cukup Jauh 2. Jauh
X11	Waktu belajar setelah pandemi	0. Jarang 1. Sering
X12	Media belajar setelah pandemi	0. Kurang Bagus 1. Bagus
X13	Pemberian tugas setelah pandemi	0. Jarang 1. Sering
X14	Pemahaman materi setelah pandemi	0. Tidak Paham 1. Kurang Paham 2. Paham
X15	Lingkungan tempat tinggal setelah pandemi	0. Tidak Kondusif 1. Kondusif
X16	Waktu tidur setelah pandemi	0. Kurang 1. Cukup
Y	Indeks Prestasi saat pandemi dengan setelah pandemi	0. Menurun 1. Meningkat

Setelah menyebarkan kuesioner, didapatkan total sebanyak 123 responden yang telah mengisi kuesioner. Menurut data yang telah didapat, diketahui responden merupakan mahasiswa dari beberapa Universitas di Riau dan Sumatera. Berikut adalah data awal dari hasil kuesioner pada penelitian ini.

https://doi.org/10.25077/TEKNOSI.v9i1.2023.21-32

Tabel 2. Data Awal

X0	X1	X2	…	X16	Y
Abdul Alim	Mandiri	Lancar	…	Kurang	Meningkat
Adyah Widiarni	Mandiri & Diskusi	Kurang Lancar	…	Kurang	Meningkat
Afifah Pendri	Mandiri & Diskusi	Lancar	…	Kurang	Menurun
Agung Permana	Mandiri & Diskusi	Kurang Lancar	…	Cukup	Meningkat
Ahmad Miftah	Mandiri & Diskusi	Lancar	…	Kurang	Meningkat
…	…	…	…	…	…
Yogi Pranata	Mandiri	Lancar	…	Cukup	Meningkat

3.2. *Preprocessing Data*

Data yang telah terkumpul kemudian dilakukan *preprocessing* dengan menghapus baris dan kolom yang tidak relevan. Tahap selanjutnya adalah melakukan seleksi data dengan mengambil 100 dari 123 baris data. Hal ini bertujuan untuk menyeimbangkan jumlah kelas target pada data untuk mencegah model *overfitting*. Terakhir yaitu melakukan perubahan bentuk data dengan teknik *label encoder* agar data dapat diterima oleh model *machine learning*.

Tabel 3. Hasil *Preprocessing* Data

X1	X2	X3	X4	…	X16	Y
0	1	0	0	…	0	1
2	0	0	0	…	0	1
2	1	1	1	…	0	0
2	0	0	0	…	1	1
2	1	0	1	…	0	1
…	…	…	…	…	…	…
0	1	0	1	…	1	1

3.3. *Pembagian Data*

Penerapan K-*Fold Cross Validation* sebagai pembagian data dikarenakan teknik ini sangat baik dalam membagi data yang berukuran kecil, hal tersebut sesuai dengan ukuran data pada penelitian ini. Nilai umum yang digunakan pada K-*Fold Cross Validation* adalah 5 atau 10, karena nilai tersebut telah ditunjukkan secara empiris tidak mengalami bias dan varians yang tinggi. Untuk itu pada penelitian ini menggunakan 10 K-*Fold* sebagai teknik pembagian data.

3.4. *Seleksi Fitur dengan Chi-Square*

Implementasi seleksi fitur dengan *chi-square* menghasilkan beberapa fitur yang paling berpengaruh pada data, 3 diantaranya yaitu pemahaman materi setelah pandemi, waktu belajar setelah pandemi, dan pemahaman materi saat pandemi. Pada *chi-square*, terdapat beberapa cara yang sering digunakan untuk memilih fitur pada pemodelan yaitu dengan melalukan pengujian satu-persatu pada jumlah fitur [76] ataupun dengan menggunakan nilai *chi-score* dan p-*value* [74]. Semakin kecil nilai p-*value* menunjukkan semakin berpengaruh suatu fitur terhadap data, sedangkan *chi-score* adalah sebaliknya.

Umumnya pada penelitian yang menggunakan p-*value* untuk memilih fitur menetapkan nilai < 0,01 atau < 0,05 sebagai *threshold* [74]. Pada penelitian ini menetapkan p-*value* < 0,05 sebagai *threshold* dan juga akan melakukan pengujian satu-persatu berdasarkan persentase jumlah fitur (*cut-off*). Hal ini dilakukan agar dapat memilih kombinasi dan jumlah fitur yang paling optimal pada data sehingga menghasilkan model dengan kemampuan yang baik. Adapun nilai persentase *cut-off* yang digunakan yaitu 25%, 50% dan 75% dari jumlah fitur berdasarkan urutan *chi score* dan p-*value* teratas [76].

Tabel 4. Hasil Seleksi Fitur dengan *Chi-Square*

No	Fitur	Chi Score	P-Value
1	X14	8,582	0,003
2	X11	3,918	0,048
3	X6	0,924	0,336
4	X15	0,591	0,442
5	X2	0,399	0,528
6	X4	0,394	0,530
7	X13	0,323	0,570
8	X10	0,294	0,588
9	X9	0,216	0,642
10	X8	0,082	0,774
11	X12	0,072	0,788
12	X5	0,050	0,822
13	X7	0,029	0,865
14	X1	0,018	0,892
15	X3	0,018	0,895
16	X16	0,006	0,939

3.5. *Seleksi Fitur dengan Mutual Information*

Seleksi fitur dengan *mutual information* menghasilkan beberapa fitur yang paling berpengaruh pada data. Pada 3 teratas menghasilkan fitur yang tidak jauh berbeda dengan *chi-square* yaitu pemahaman materi setelah pandemi, cara belajar setelah pandemi, dan pemahaman materi saat pandemi. Beberapa cara umum yang digunakan untuk memilih fitur pada *mutual information* yaitu dengan melalukan pengujian satu-persatu pada setiap fitur dan dengan menggunakan nilai *mutual information score* (MI-*score*) [51]. Semakin tinggi nilai MI-*score* menunjukkan semakin berpengaruh suatu fitur terhadap data [77]. Pada penelitian ini akan menggunakan MI-*score* > 0,037 sebagai *threshold* dan juga akan melakukan pengujian satu-persatu berdasarkan persentase jumlah fitur (*cut-off*). Adapun nilai persentase yang digunakan sama seperti pada *chi-square* yaitu 25%, 50% dan 75% dari jumlah fitur berdasarkan urutan MI-*score* teratas [78].

3.6. *Melatih Model*

3.6.1. *Pemodelan Tanpa Seleksi Fitur*

Pemodelan algoritma dengan NBC, RF, CART, dan SVM tanpa seleksi fitur menghasilkan SVM sebagai model terbaik dengan akurasi 89,00%, *precision* 87,72%, dan *recall* 93,57%. Dalam hal komputasi, SVM berada pada urutan ketiga, dibawah CART dan NBC dengan waktu 0,03207s. Perbedaan kinerja yang dihasilkan

oleh model disebabkan oleh perbedaan karakteristik dari masing-masing algoritma dalam mengklasifikasikan data. Hasil pemodelan setiap algoritma tanpa seleksi fitur dapat dilihat pada tabel 6.

Tabel 5. Hasil Seleksi Fitur dengan *Mutual Information*

No	Fitur	MI-Score
1	X14	0,3364
2	X9	0,0373
3	X6	0,0370
4	X11	0,0357
5	X15	0,0226
6	X1	0,0112
7	X10	0,0082
8	X13	0,0048
9	X4	0,0047
10	X2	0,0038
11	X5	0,0021
12	X12	0,0017
13	X8	0,0008
14	X7	0,0003
15	X3	0,0001
16	X16	0,0001

Tabel 6. Evaluasi Model Tanpa Seleksi Fitur

Model	Accuracy (%)	Precision (%)	Recall (%)	Komputasi (s)
NBC	85,00%	88,20%	85,71%	0,02422s
RF	87,00%	87,58%	92,32%	0,32561s
CART	86,00%	87,17%	89,82%	0,02172s
SVM	89,00%	87,72%	93,57%	0,03207s

3.6.2. Pemodelan dengan Seleksi Fitur Chi-Square

Berdasarkan hasil pemodelan dengan seleksi fitur *chi-square*, diketahui bahwa jumlah fitur sangat mempengaruhi kemampuan algoritma, namun hal tersebut tidak terlalu berpengaruh pada SVM. Model SVM memiliki kemampuan yang sama walaupun dengan kombinasi dan jumlah fitur yang berbeda berdasarkan nilai akurasi, *precision* dan *recall*, namun memiliki pengaruh dan perbedaan dalam hal waktu komputasi. Jumlah fitur yang lebih sedikit membuat waktu komputasi jauh lebih cepat pada setiap algoritma yang diterapkan. Berikut adalah hasil pemodelan dengan uji coba jumlah dan kombinasi fitur pada *chi-square.*

Tabel 7. Evaluasi Model dengan Seleksi Fitur Chi-Square

Threshold (Jumlah Fitur)	Model	Accuracy (%)	Precision (%)	Recall (%)	Waktu Komputasi (S)
P-Value 0,05 (2 Fitur)	NBC	89,00%	87,72%	93,57%	0,01685s
	RF	89,00%	87,72%	93,57%	0,22561s
	CART	**89,00%**	**87,72%**	**93,57%**	**0,01543s**
	SVM	89,00%	87,72%	93,57%	0,02063s
25% (4 Fitur)	NBC	86,00%	88,20%	87,38%	0,01821s
	RF	89,00%	87,72%	92,14%	0,27847s
	CART	**89,00%**	**87,72%**	**93,57%**	**0,01615s**
	SVM	89,00%	87,72%	93,57%	0,02182s
50% (8 Fitur)	NBC	86,00%	88,20%	87,38%	0,02054s
	RF	86,00%	87,84%	87,14%	0,29126s
	CART	86,00%	87,17%	89,82%	0,01783s
	SVM	**89,00%**	**87,72%**	**93,57%**	**0,02842s**
75% (12 Fitur)	NBC	85,00%	88,20%	85,71%	0,02237
	RF	84,00%	87,20%	83,63%	0,30277
	CART	86,00%	87,17%	89,82%	0,01959
	SVM	**89,00%**	**87,72%**	**93,57%**	**0,02939**

Implementasi seleksi fitur dengan *chi-square* dengan uji coba jumlah dan kombinasi fitur menghasilkan beberapa model terbaik. Model dengan 2 dan 4 fitur menghasilkan CART sebagai yang terbaik. Model CART dengan 2 fitur memiliki akurasi 89,00%, *precision* 87,72%, *recall* 93,57% dan waktu komputasi 0,01543s, sedangkan pada 4 fitur memiliki akurasi 89,00%, *precision* 87,72%, *recall* 93,57% dan waktu komputasi 0,01615s. Selanjutnya pada 8 dan 12 fitur menghasilkan SVM sebagai yang terbaik. Model SVM dengan 8 fitur memiliki akurasi 89,00%, *precision* 87,72%, *recall* 93,57% dan waktu komputasi 0,02842s, sedangkan pada 8 fitur memiliki akurasi 89,00%, *precision* 87,72%, dan waktu komputasi 0,02939s.

3.6.3. Pemodelan dengan Seleksi Fitur Mutual Information

Penerapan *mutual information* sebagai seleksi fitur pada setiap algoritma menghasilkan kemampuan yang tidak berbeda jauh dengan *chi-square*. Beberapa perbedaan diantaranya adalah pada model NBC dengan 2 fitur terjadi penurunan performa, namun pada 4 fitur terjadi peningkatan. Kemudian pada model RF dengan 4 dan 8 fitur mengalami penurunan, sedangkan pada 12 fitur mengalami peningkatan. Selain itu, jumlah dan kombinasi fiur kembali tidak mempengaruhi performa SVM dalam hal akurasi, *precision* dan *recall* seperti pada *chi-square*, namun tetap memiliki pengaruh dan perbedaan terhadap waktu komputasi.

https://doi.org/10.25077/TEKNOSI.v9i1.2023.21-32

Hasil uji coba pemodelan dengan seleksi fitur *mutual information* dapat dilihat pada tabel 8.

Tabel 8. Evaluasi Model dengan Seleksi Fitur Mutual Information

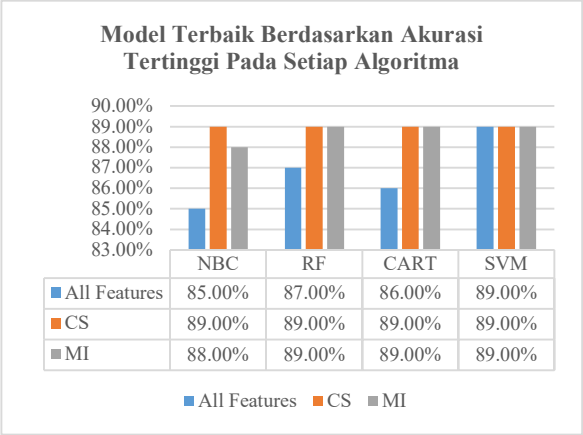
Threshold (Jumlah Fitur)	Model	Accuracy (%)	Precision (%)	Recall (%)	Waktu Komputasi (S)
MI Score 0,037 (2 Fitur)	NBC	88,00%	87,72%	90,24%	0,01708s
	RF	89,00%	87,72%	93,57%	0,27654s
	CART	**89,00%**	**87,72%**	**93,57%**	**0,01685s**
	SVM	89,00%	87,72%	93,57%	0,01987s
25% (4 Fitur)	NBC	87,00%	87,31%	90,48%	0,01804s
	RF	88,00%	87,72%	90,24%	0,28075s
	CART	**89,00%**	**87,72%**	**93,57%**	**0,01734s**
	SVM	89,00%	87,72%	93,57%	0,02157s
50% (8 Fitur)	NBC	86,00%	88,20%	87,38%	0,02110s
	RF	84,00%	87,17%	83,12%	0,29319s
	CART	86,00%	87,17%	89,82%	0,01829s
	SVM	**89,00%**	**87,72%**	**93,57%**	**0,02883s**
75% (12 Fitur)	NBC	85,00%	88,20%	85,71%	0,02221s
	RF	87,00%	88,99%	86,49%	0,31084s
	CART	86,00%	87,17%	89,82%	0,01956s
	SVM	**89,00%**	**87,72%**	**93,57%**	**0,03054s**

Dari hasil seleksi fitur dengan *mutual information* terdapat beberapa model terbaik berdasarkan kombinasi dan jumlah fitur yang ditetapkan. Pada 2 dan 4 fitur menghasilkan CART sebagai model yang terbaik. Model CART dengan 2 fitur memiliki akurasi 89,00%, *precision* 87,72%, *recall* 93,57% dan waktu komputasi 0,01685s, sedangkan pada 4 fitur memiliki akurasi 89,00%, *precision* 87,72%, *recall* 93,57% dan waktu komputasi 0,01734s. selanjutnya pada 8 dan 12 fitur menghasilkan SVM sebagai yang terbaik. Model SVM dengan 8 fitur memiliki akurasi 89,00%, *precision* 87,72%, *recall* 93,57% dan waktu komputasi 0,02883s. Sedangkan pada 12 fitur memiliki akurasi 89,00%, *precision* 87,72%, *recall* 93,57% dan waktu komputasi 0,03054s.

4. PEMBAHASAN

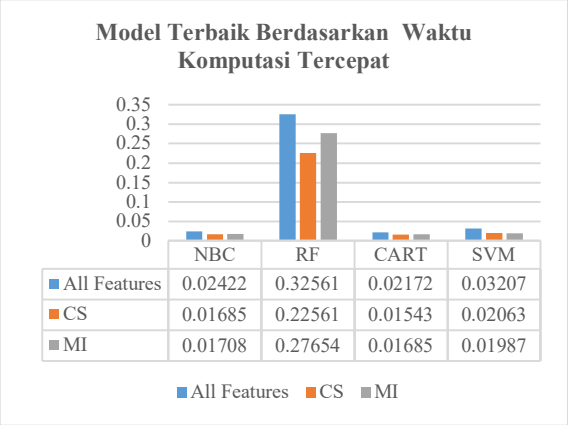
4.1. Menentukan Model Terbaik

Berdasarkan implementasi seleksi fitur yang digunakan pada penelitian ini menunjukkan bahwa kemampuan algoritma mengalami peningkatan dibandingkan tanpa seleksi fitur terutama dalam hal akurasi, namun tidak pada SVM. Model SVM dengan dan tanpa seleksi fitur menghasilkan akurasi yang sama yaitu 89,00%. Ini menunjukkan bahwa SVM memiliki kemampuan klasifikasi yang baik walaupun dengan data berdimensi besar. Peningkatan akurasi tertinggi terjadi pada model NBC dengan seleksi fitur *chi-square* yaitu sebesar 4%. Grafik perbandingan akurasi setiap algoritma dapat dilihat pada gambar 2.

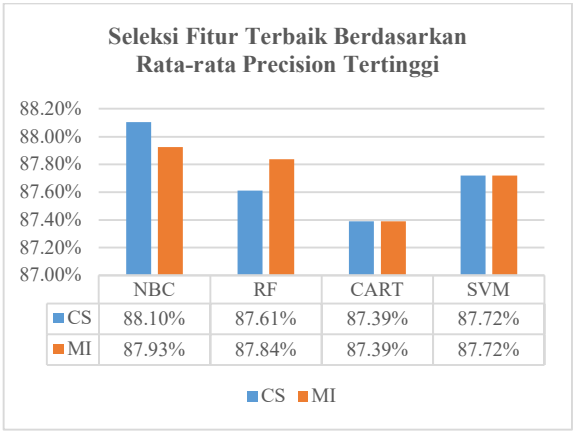


Gambar 2. Model Terbaik Berdasarkan Akurasi

Selain meningkatkan akurasi, penerapan seleksi fitur juga mempercepat waktu komputasi. Model CART dengan *chi-square* menjadi yang tercepat dalam hal rata-rata waktu komputasi dibandingkan model lainnya yaitu 0,01543s. Sedangkan Model RF memiliki waktu komputasi terlama dengan waktu komputasi tercepat yaitu 0,27654s.



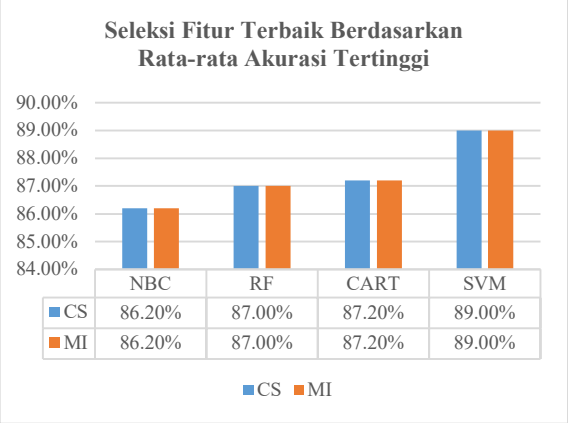
Gambar 3. Model Terbaik Berdasarkan Waktu Komputasi



Gambar 5. Seleksi Fitur Terbaik Berdasarkan Rata-rata Precision

4.2. Menentukan Seleksi Fitur yang Efektif

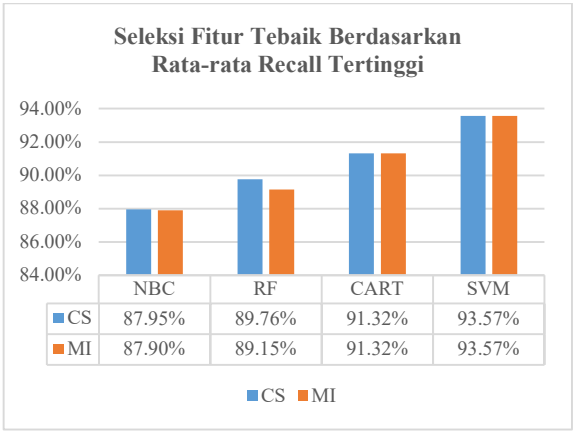
Untuk menentukan seleksi fitur yang efektif pada penelitian ini, perlu dilakukan perbandingan rata-rata performa model pada setiap metode. Berdasarkan rata-rata akurasi, *chi-square* dan *mutual information* memiliki nilai yang sama, yaitu pada NBC sebesar 86,20%, RF sebesar 87,00%, CART sebesar 87,20% dan SVM sebesar 89,00%. Hal ini menunjukkan bahwa kedua metode memiliki kemampuan dan efektifitas yang sama berdasarkan rata-rata akurasi.



Gambar 4. Seleksi Fitur Terbaik Berdasarkan Rata-rata Akurasi

Selanjutnya berdasarkan nilai *precision*, *chi-square* lebih baik dibandingkan *mutual information* terutama pada model NBC, namun pada RF adalah sebaliknya. Selain itu, kedua metode pada model CART dan SVM memiliki rata-rata yang sama. Dengan begitu, kedua metode dapat dikatakan efektif berdasarkan rata-rata *precision* tergantung dari algoritma yang digunakan.

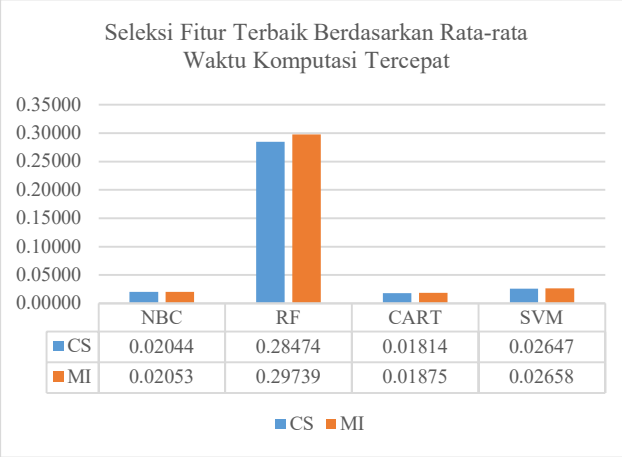
Kemudian pada rata-rata *recall*, *chi-square* sedikit lebih unggul daripada *mutual information* terutama pada model NBC dan RF. Sedangkan pada model CART dan SVM kedua metode sama baiknya. Maka dari itu, dapat dikatakan bahwa *chi-square* lebih unggul dibandingkan *mutual information* berdasarkan rata-rata *recall*.



Gambar 6. Seleksi Fitur Terbaik Berdasarkan Rata-rata Recall

Terakhir adalah membandingkan rata-rata waktu komputasi pada kedua metode. Berdasarkan hasil percobaan, *chi-square* unggul disetiap pemodelan. Hal ini mengindikasikan bahwa *chi-square* lebih baik dibanding *mutual information* dalam hal rata-rata waktu komputasi.

https://doi.org/10.25077/TEKNOSI.v9i1.2023.21-32



Gambar 7. Seleksi Fitur Terbaik Berdasarkan Rata-rata Waktu Komputasi

Dari perbandingan yang dilakukan berdasarkan rata-rata akurasi, *precision*, *recall* dan waktu komputasi, dapat disimpulkan *chi-square* adalah metode seleksi fitur yang efektif dibandingkan *mutual information* pada penelitian ini. Hal ini dikarenakan *chi-square* unggul dalam hal rata-rata *recall* dan waktu komputasi. Hasil perbandingan metode seleksi fitur *chi-square* dan *mutual information* dapat dilihat pada tabel 9.

Tabel 9. Hasil Perbandingan Metode Seleksi Fitur

Aspek Penilaian	Chi-Square				Mutual Information			
	NBC	RF	CART	SVM	NBC	RF	CART	SVM
Accuracy	-	-	-	-	-	-	-	-
Precision	V	-	-	-	-	V	-	-
Recall	V	V	-	-	-	-	-	-
Komputasi	V	V	V	V	-	-	-	-

5. KESIMPULAN

Berdasarkan hasil dan analisis dari penelitian yang dilakukan, dapat disimpulkan seleksi fitur efektif dalam meningkatkan kemampuan model dan mempercepat waktu komputasi. Implementasi seleksi fitur dengan *chi-square* menghasilkan 3 atribut yang paling berpengaruh pada data yaitu pemahaman materi setelah pandemi, waktu belajar setelah pandemi dan pemahaman materi setelah pandemi. Sedangkan 3 atribut paling berpengaruh pada seleksi fitur dengan *mutual information* yaitu pemahaman materi setelah pandemi, cara belajar setelah pandemi, dan pemahaman materi saat pandemi. Selain itu, berdasarkan pemodelan dengan 4 algoritma dan 2 metode seleksi fitur, menghasilkan CART dengan *chi-square* menggunakan 2 fitur sebagai model terbaik dengan akurasi 89,00%, *precision* 87,72%, *recall* 93,57% dan waktu komputasi 0,01814s. Dibandingkan tanpa seleksi fitur, performa CART dengan *chi-square* mengalami peningkatan akurasi sebesar 3% dan waktu komputasi 0,00629s. *Chi-square* menjadi seleksi fitur yang efektif pada penelitian ini, yang mana *chi-square* unggul pada rara-rata *recall* dan waktu komputasi dibandingkan *mutual information* pada pemodelan.

DAFTAR PUSTAKA

[1] A. Cahyadi, Hendryadi, S. Widyastuti, V. N. Mufidah, and Achmadi, "Emergency remote teaching evaluation of the higher education in Indonesia," *Heliyon*, vol. 7, no. 8, 2021, doi: 10.1016/j.heliyon.2021.e07788.

[2] N. Aisha and A. Ratra, "Online education amid COVID-19 pandemic and its opportunities, challenges and psychological impacts among students and teachers: a systematic review," *Asian Assoc. Open Univ. J.*, vol. 17, no. 3, pp. 242–260, 2022, doi: 10.1108/AAOUJ-03-2022-0028.

[3] W. Rahayu, M. D. K. Putra, Faturochman, Meiliasari, E. Sulaeman, and R. B. Koul, "Development and validation of Online Classroom Learning Environment Inventory (OCLEI): The case of Indonesia during the COVID-19 pandemic," *Learn. Environ. Res.*, vol. 25, no. 1, pp. 97–113, 2022, doi: 10.1007/s10984-021-09352-3.

[4] G. H. Mardini and O. A. Mah'd, "Distance learning as emergency remote teaching vs. traditional learning for accounting students during the COVID-19 pandemic: Cross-country evidence," *J. Account. Educ.*, vol. 61, p. 100814, 2022, doi: 10.1016/j.jaccedu.2022.100814.

[5] M. Al-Nasa'h, L. Al-Tarawneh, F. M. Abu Awwad, and I. Ahmad, "Estimating students' online learning satisfaction during COVID-19: A discriminant analysis," *Heliyon*, vol. 7, no. 12, p. e08544, 2021, doi: 10.1016/j.heliyon.2021.e08544.

[6] L. Anthonysamy and P. Singh, "The impact of satisfaction, and autonomous learning strategies use on scholastic achievement during Covid-19 confinement in Malaysia," *Heliyon*, vol. 9, no. 2, p. e12198, 2023, doi: 10.1016/j.heliyon.2022.e12198.

[7] S. Ali, Y. Hafeez, M. A. Abbas, M. Aqib, and A. Nawaz, "Enabling remote learning system for virtual personalized preferences during COVID-19 pandemic," *Multimed. Tools Appl.*, vol. 80, no. 24, pp. 33329–

33355, 2021, doi: 10.1007/s11042-021-11414-w.
[8] S. Adewale and M. B. Tahir, "Virtual learning environment factors as predictors of students' learning satisfaction during COVID-19 period in Nigeria," *Asian Assoc. Open Univ. J.*, vol. 17, no. 2, pp. 120–133, 2022, doi: 10.1108/AAOUJ-10-2021-0121.
[9] W. Wagiran, S. Suharjana, M. Nurtanto, and F. Mutohhari, "Determining the e-learning readiness of higher education students: A study during the COVID-19 pandemic," *Heliyon*, vol. 8, no. 10, p. e11160, 2022, doi: 10.1016/j.heliyon.2022.e11160.
[10] J. Melgaard, R. Monir, L. A. Lasrado, and A. Fagerstrøm, "Academic Procrastination and Online Learning during the COVID-19 Pandemic," *Procedia Comput. Sci.*, vol. 196, no. 2021, pp. 117–124, 2021, doi: 10.1016/j.procs.2021.11.080.
[11] A. Cahyadi, Hendryadi, S. Widyastuti, and Suryani, "COVID-19, emergency remote teaching evaluation: the case of Indonesia," *Educ. Inf. Technol.*, vol. 27, no. 2, pp. 2165–2179, 2022, doi: 10.1007/s10639-021-10680-3.
[12] F. Abdullah and S. Kauser, "Students' perspective on online learning during pandemic in higher education," *Qual. Quant.*, no. 0123456789, 2022, doi: 10.1007/s11135-022-01470-1.
[13] M. Kerres and J. Buchner, "Education after the Pandemic: What We Have (Not) Learned about Learning," *Educ. Sci.*, vol. 12, no. 5, 2022, doi: 10.3390/educsci12050315.
[14] S. I. N. W. Abdullah, K. Arokiyasamy, S. L. Goh, A. J. Culas, and N. M. A. Manaf, "University students' satisfaction and future outlook towards forced remote learning during a global pandemic," *Smart Learn. Environ.*, vol. 9, no. 1, 2022, doi: 10.1186/s40561-022-00197-8.
[15] Y. Zhu, G. Geng, L. Disney, and Z. Pan, *Changes in university students' behavioral intention to learn online throughout the COVID-19: Insights for online teaching in the post-pandemic era*, no. 0123456789. Springer US, 2022.
[16] C. Rapanta, L. Botturi, P. Goodyear, L. Guàrdia, and M. Koole, "Balancing Technology, Pedagogy and the New Normal: Post-pandemic Challenges for Higher Education," *Postdigital Sci. Educ.*, vol. 3, no. 3, pp. 715–742, 2021, doi: 10.1007/s42438-021-00249-1.
[17] A. Patricia Aguilera-Hermida, "College students' use and acceptance of emergency online learning due to COVID-19," *Int. J. Educ. Res. Open*, vol. 1, no. July, p. 100011, 2020, doi: 10.1016/j.ijedro.2020.100011.
[18] V. Ratten, "The post COVID-19 pandemic era: Changes in teaching and learning methods for management educators," *Int. J. Manag. Educ.*, vol. 21, no. 2, p. 100777, 2023, doi: 10.1016/j.ijme.2023.100777.
[19] A. Sharma and I. Alvi, "Evaluating pre and post COVID 19 learning: An empirical study of learners' perception in higher education," *Educ. Inf. Technol.*, vol. 26, no. 6, pp. 7015–7032, 2021, doi: 10.1007/s10639-021-10521-3.
[20] N. Yan and O. T. S. Au, "Online learning behavior analysis based on machine learning," *Asian Assoc. Open Univ. J.*, vol. 14, no. 2, pp. 97–106, 2019, doi: 10.1108/AAOUJ-08-2019-0029.
[21] A. Dinesh Kumar, R. Pandi Selvam, and V. Palanisamy, "Hybrid Classification Algorithms for Predicting Student Performance," in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, 2021, pp. 1074–1079, doi: 10.1109/ICAIS50930.2021.9395974.
[22] R. Patil, S. Salunke, M. Kalbhor, and R. Lomte, "Prediction System for Student Performance Using Data Mining Classification," in *2018 4th International Conference on Computing, Communication Control and Automation (ICCUBEA)*, 2018, pp. 1–4, doi: 10.1109/ICCUBEA.2018.8697770.
[23] C. C. Kiu, "Data Mining Analysis on Student's Academic Performance through Exploration of Student's Background and Social Activities," in *2018 4th International Conference on Advances in Computing, Communication and Automation (ICACCA)*, 2018, pp. 1–5, doi: 10.1109/ICACCAF.2018.8776809.
[24] A. Zamir *et al.*, "Phishing web site detection using diverse machine learning algorithms," *Electron. Libr.*, vol. 38, no. 1, pp. 65–80, 2020, doi: 10.1108/EL-05-2019-0118.
[25] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," *J. Supercomput.*, vol. 77, no. 5, pp. 5198–5219, 2021, doi: 10.1007/s11227-020-03481-x.
[26] K. Alpan and G. S. Ilgi, "Classification of Diabetes Dataset with Data Mining Techniques by Using WEKA Approach," in *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2020, pp. 1–7, doi: 10.1109/ISMSIT50672.2020.9254720.
[27] M. Ozcan and S. Peker, "A classification and regression tree algorithm for heart disease modeling and prediction," *Healthc. Anal.*, vol. 3, no. July 2022, p. 100130, 2023, doi: 10.1016/j.health.2022.100130.
[28] P. Subarkah, A. N. Ikhsan, and A. Setyanto, "The effect of the number of attributes on the selection of study program using classification and regression trees algorithms," in *2018 3rd International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2018, pp. 1–5, doi: 10.1109/ICITISEE.2018.8721030.
[29] S. N. Lathifah, F. Nhita, A. Aditsania, and D. Saepudin, "Rainfall forecasting using the classification and regression tree (CART) algorithm and adaptive synthetic sampling (study case: Bandung regency)," in *7th International Conference on Information and Communication Technology (ICoICT)*, 2019, pp. 1–5, doi: 10.1109/ICoICT.2019.8835308.
[30] S. İlkin, T. H. Gençtürk, F. Kaya Gülağız, H. Özcan, M. A. Altuncu, and S. Şahin, "hybSVM: Bacterial colony optimization algorithm based SVM for malignant melanoma detection," *Eng. Sci. Technol. an Int. J.*, vol. 24, no. 5, pp. 1059–1071, 2021, doi: 10.1016/j.jestch.2021.02.002.
[31] A. Mahmood and H. U. Khan, "Identification of critical factors for assessing the quality of restaurants using data mining approaches," *Electron. Libr.*, vol. 37, no. 6, pp. 952–969, 2019, doi: 10.1108/EL-12-2018-0241.
[32] M. J. Nayeem, S. Rana, F. Alam, and M. A. Rahman, "Prediction of Hepatitis Disease Using K-Nearest Neighbors, Naive Bayes, Support Vector Machine, Multi-Layer Perceptron and Random Forest," in *2021 International Conference on Information and Communication Technology for Sustainable Development, (ICICT4SD)*, 2021, pp. 280–284, doi: 10.1109/ICICT4SD50815.2021.9397013.
[33] G. Ramaswami, T. Susnjak, A. Mathrani, J. Lim, and P. Garcia, "Using educational data mining techniques to increase the prediction accuracy of student academic performance," *Inf. Learn. Sci.*, vol. 120, no. 7–8, pp. 451–467, 2019, doi: 10.1108/ILS-03-2019-0017.
[34] V. Chang, V. R. Bhavani, A. Q. Xu, and M. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms," *Healthc.*

Anal., vol. 2, p. 100016, 2022, doi: 10.1016/j.health.2022.100016.

[35] Islamiyah, A. N. Afiyah, N. Dengen, and M. Taruk, "Comparison Performance of C4.5 , Naïve Bayes and K-Nearest Neighbor in Determination Drug Rehabilitation," in *2019 5th International Conference on Science in Information Technology (ICSITech)*, 2019, pp. 112–117, doi: 10.1109/ICSITech46713.2019.8987455.

[36] R. S. Raj, D. S. Sanjay, M. Kusuma, and S. Sampath, "Comparison of Support Vector Machine and Naïve Bayes Classifiers for Predicting Diabetes," in *2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing and Communication Engineering (ICATIECE)*, 2019, pp. 41–45, doi: 10.1109/ICATIECE45860.2019.9063792.

[37] D. H. Jeong, B. K. Jeong, N. Leslie, C. Kamhoua, and S.-Y. Ji, "Designing a supervised feature selection technique for mixed attribute data analysis," *Mach. Learn. with Appl.*, vol. 10, no. March, p. 100431, 2022, doi: 10.1016/j.mlwa.2022.100431.

[38] H. Zhou, X. Wang, and Y. Zhang, "Feature selection based on weighted conditional mutual information," *Appl. Comput. Informatics*, 2020, doi: 10.1016/j.aci.2019.12.003.

[39] H. Zhou, X. Wang, and R. Zhu, "Feature selection based on mutual information with correlation coefficient," *Appl. Intell.*, vol. 52, no. 5, pp. 5457–5474, 2022, doi: 10.1007/s10489-021-02524-x.

[40] F. Macedo, R. Valadas, E. Carrasquinha, M. R. Oliveira, and A. Pacheco, "Feature selection using Decomposed Mutual Information Maximization," *Neurocomputing*, vol. 513, pp. 215–232, 2022, doi: 10.1016/j.neucom.2022.09.101.

[41] J. Gonzalez-Lopez, S. Ventura, and A. Cano, "Distributed multi-label feature selection using individual mutual information measures," *Knowledge-Based Syst.*, vol. 188, 2020, doi: 10.1016/j.knosys.2019.105052.

[42] K. Gajowniczek, J. Wu, S. Gupta, and C. Bajaj, "HOFS: Higher order mutual information approximation for feature selection in R," *SoftwareX*, vol. 19, p. 101148, 2022, doi: 10.1016/j.softx.2022.101148.

[43] H. Chauhan, K. Modi, and S. Shrivastava, "Development of a classifier with analysis of feature selection methods for COVID-19 diagnosis," *World J. Eng.*, vol. 19, no. 1, pp. 49–57, 2022, doi: 10.1108/WJE-10-2020-0537.

[44] F. Souza, C. Premebida, and R. Araújo, "High-order conditional mutual information maximization for dealing with high-order dependencies in feature selection," *Pattern Recognit.*, vol. 131, p. 108895, 2022, doi: 10.1016/j.patcog.2022.108895.

[45] T. Bhadra, S. Mallik, N. Hasan, and Z. Zhao, "Comparison of five supervised feature selection algorithms leading to top features and gene signatures from multi-omics data in cancer," *BMC Bioinformatics*, vol. 23, pp. 1–18, 2022, doi: 10.1186/s12859-022-04678-y.

[46] S. Williamson, K. Vijayakumar, and V. J. Kadam, "Predicting breast cancer biopsy outcomes from BI-RADS findings using random forests with chi-square and MI features," *Multimed. Tools Appl.*, vol. 81, no. 26, pp. 36869–36889, 2022, doi: 10.1007/s11042-021-11114-5.

[47] N. Peker and C. Kubat, "Application of Chi-square discretization algorithms to ensemble classification methods," *Expert Syst. Appl.*, vol. 185, no. June 2020, p. 115540, 2021, doi: 10.1016/j.eswa.2021.115540.

[48] H. M. Abdelaal, B. R. Elemary, and H. A. Youness, "Classification of Hadith According to Its Content Based on Supervised Learning Algorithms," *IEEE Access*, vol. 7, pp. 152379–152387, 2019, doi: 10.1109/ACCESS.2019.2948159.

[49] I. Sumaiya Thaseen and C. Aswani Kumar, "Intrusion detection model using fusion of chi-square feature selection and multi class SVM," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 29, no. 4, pp. 462–472, 2017, doi: 10.1016/j.jksuci.2015.12.004.

[50] R. Spencer, F. Thabtah, N. Abdelhamid, and M. Thompson, "Exploring feature selection and classification methods for predicting heart disease," *Digit. Heal.*, vol. 6, pp. 1–10, 2020, doi: 10.1177/2055207620914777.

[51] V. Sharma and K. C. Juglan, "Automated Classification of Fatty and Normal Liver Ultrasound Images Based on Mutual Information Feature Selection," *IRBM*, vol. 39, no. 5, pp. 313–323, 2018, doi: 10.1016/j.irbm.2018.09.006.

[52] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *Int. J. Cogn. Comput. Eng.*, vol. 2, pp. 40–46, 2021, doi: 10.1016/j.ijcce.2021.01.001.

[53] W. BinSaeedan and S. Alramlawi, "CS-BPSO: Hybrid feature selection based on chi-square and binary PSO algorithm for Arabic email authorship analysis," *Knowledge-Based Syst.*, vol. 227, p. 107224, 2021, doi: 10.1016/j.knosys.2021.107224.

[54] D. Effrosynidis and A. Arampatzis, "An evaluation of feature selection methods for environmental data," *Ecol. Inform.*, vol. 61, no. January, p. 101224, 2021, doi: 10.1016/j.ecoinf.2021.101224.

[55] G. Dimic, D. Rancic, N. Macek, P. Spalevic, and V. Drasute, "Improving the prediction accuracy in blended learning environment using synthetic minority oversampling technique," *Inf. Discov. Deliv.*, vol. 47, no. 2, pp. 76–83, 2019, doi: 10.1108/IDD-08-2018-0036.

[56] A. Madasu and S. Elango, "Efficient feature selection techniques for sentiment analysis," *Multimed. Tools Appl.*, vol. 79, no. 9–10, pp. 6313–6335, 2020, doi: 10.1007/s11042-019-08409-z.

[57] S. Mochammad, Y. J. Kang, Y. Noh, S. Park, and B. Ahn, "Stable Hybrid Feature Selection Method for Compressor Fault Diagnosis," *IEEE Access*, vol. 9, pp. 97415–97429, 2021, doi: 10.1109/ACCESS.2021.3092884.

[58] A. Kaur, K. Guleria, and N. K. Trivedi, "Feature Selection in Machine Learning: Methods and Comparison," in *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 2021, pp. 789–795, doi: 10.1109/ICACITE51222.2021.9404623.

[59] H. Utama, "Sentiment analysis in airline tweets using mutual information for feature selection," in *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (CITISEE)*, 2019, pp. 295–300, doi: 10.1109/ICITISEE48480.2019.9003903.

[60] M. Alduailij, Q. W. Khan, M. Tahir, M. Sardaraz, M. Alduailij, and F. Malik, "Machine-Learning-Based DDoS Attack Detection Using Mutual Information and Random Forest Feature Importance Method," *Symmetry (Basel).*, vol. 14, no. 6, pp. 1–15, 2022, doi: 10.3390/sym14061095.

[61] K. Wang, W. Mao, W. Feng, and H. Wang, "Research on spam filtering technology based on new mutual information feature selection algorithm," *J. Phys. Conf. Ser.*, vol. 1673, no. 1, 2020, doi: 10.1088/1742-

6596/1673/1/012028.
[62] M. Malekipirbazari, V. Aksakalli, W. Shafqat, and A. Eberhard, "Performance comparison of feature selection and extraction methods with random instance selection," *Expert Syst. Appl.*, vol. 179, no. April, p. 115072, 2021, doi: 10.1016/j.eswa.2021.115072.
[63] T. H. Kerbaa, A. Mezache, and H. Oudira, "Model Selection of Sea Clutter Using Cross Validation Method," in *Procedia Computer Science*, 2019, vol. 158, pp. 394–400, doi: 10.1016/j.procs.2019.09.067.
[64] O. Karal, "Performance comparison of different kernel functions in SVM for different k value in k-fold cross-validation," in *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2020, pp. 1–5, doi: 10.1109/ASYU50717.2020.9259880.
[65] S. Saud, B. Jamil, Y. Upadhyay, and K. Irshad, "Performance improvement of empirical models for estimation of global solar radiation in India: A k-fold cross-validation approach," *Sustain. Energy Technol. Assessments*, vol. 40, p. 100768, 2020, doi: 10.1016/j.seta.2020.100768.
[66] T. T. S. Nguyen and P. M. T. Do, "Classification optimization for training a large dataset with Naïve Bayes," *J. Comb. Optim.*, vol. 40, no. 1, pp. 141–169, 2020, doi: 10.1007/s10878-020-00578-0.
[67] P. Kamal and S. Ahuja, "An ensemble-based model for prediction of academic performance of students in undergrad professional course," *J. Eng. Des. Technol.*, vol. 17, no. 4, pp. 769–781, 2019, doi: 10.1108/JEDT-11-2018-0204.
[68] H. F. Zhou, J. W. Zhang, Y. Q. Zhou, X. J. Guo, and Y. M. Ma, "A feature selection algorithm of decision tree based on feature weight," *Expert Syst. Appl.*, vol. 164, no. February 2020, p. 113842, 2021, doi: 10.1016/j.eswa.2020.113842.
[69] E. Z. Aziza, L. Mohamed El Amine, M. Mohamed, and B. Abdelhafid, "Decision tree CART algorithm for diabetic retinopathy classification," in *2019 6th International Conference on Image and Signal Processing and their Applications (ISPA)*, 2019, pp. 1–5, doi: 10.1109/ISPA48434.2019.8966905.
[70] E. Pekel Özmen and T. Özcan, "Diagnosis of diabetes mellitus using artificial neural network and classification and regression tree optimized with genetic algorithm," *J. Forecast.*, vol. 39, no. 4, pp. 661–670, 2020, doi: 10.1002/for.2652.
[71] S. Cano-Ortiz, P. Pascual-Muñoz, and D. Castro-Fresno, "Machine learning algorithms for monitoring pavement performance," *Autom. Constr.*, vol. 139, no. September 2021, 2022, doi: 10.1016/j.autcon.2022.104309.
[72] N. Rust-Nguyen, S. Sharma, and M. Stamp, "Darknet traffic classification and adversarial attacks using machine learning," *Comput. Secur.*, vol. 127, p. 103098, 2023, doi: 10.1016/j.cose.2023.103098.
[73] K. Iqbal and M. S. Khan, "Email classification analysis using machine learning techniques," *Appl. Comput. Informatics*, 2022, doi: 10.1108/ACI-01-2022-0012.
[74] L. A. C. Ahakonye, C. I. Nwakanma, J. M. Lee, and D. S. Kim, "SCADA intrusion detection scheme exploiting the fusion of modified decision tree and Chi-square feature selection," *Internet of Things (Netherlands)*, vol. 21, no. August 2022, p. 100676, 2023, doi: 10.1016/j.iot.2022.100676.
[75] F. Itoo, Meenakshi, and S. Singh, "Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection," *Int. J. Inf. Technol.*, vol. 13, no. 4, pp. 1503–1511, 2021, doi: 10.1007/s41870-020-00430-y.
[76] F. Thabtah, F. Kamalov, S. Hammoud, and S. R. Shahamiri, "Least Loss: A simplified filter method for feature selection," *Inf. Sci. (Ny).*, vol. 534, pp. 1–15, 2020, doi: 10.1016/j.ins.2020.05.017.
[77] M. Jansi Rani and D. Devaraj, "Two-Stage Hybrid Gene Selection Using Mutual Information and Genetic Algorithm for Cancer Data Classification," *J. Med. Syst.*, vol. 43, no. 8, 2019, doi: 10.1007/s10916-019-1372-8.
[78] M. S. Al-Batah, M. Alzyoud, R. Alazaidah, M. Toubat, H. Alzoubi, and A. Olaiyat, "Early Prediction of Cervical Cancer Using Machine Learning Techniques," *Jordanian J. Comput. Inf. Technol.*, vol. 08, no. 04, pp. 357–369, 2022, doi: 10.5455/jjcit.71-1661691447.