

PREDIKSI KETERLAMBATAN MAHASISWA DALAM MEMBAYAR BIAYA KULIAH MENGGUNAKAN ALGORITMA C4.5 (STUDI KASUS : UNIVERSITAS MUHAMMADIYAH SUKABUMI)

Restu Raina Rahmah, Iwan Rizal Setiawan, Fathia Frazna Az-Zahra

Teknik Informatika, Universitas Muhammadiyah Sukabumi

Jl. R. Syamsudin, S.H. No. 50, Cikole, Kec. Cikole, Kota Sukabumi, Jawa Barat 43113

resturr012@ummi.ac.id

ABSTRAK

Biaya kuliah merupakan sumber pendanaan utama yang penting bagi peningkatan mutu pendidikan dan pembangunan infrastruktur. Keterlambatan pembayaran dapat mengganggu alur keuangan institusi, mempengaruhi proses belajar mengajar, dan menghambat pengembangan infrastruktur. Untuk mengatasi masalah ini, diperlukan prediksi keterlambatan pembayaran mahasiswa. Algoritma C4.5, salah satu algoritma machine learning, digunakan untuk membangun model prediktif yang mampu mengidentifikasi mahasiswa yang kemungkinan terlambat membayar dengan akurasi yang baik. Evaluasi model dilakukan menggunakan confusion matrix dan k-fold cross validation. Confusion matrix menilai kinerja model melalui akurasi, presisi, dan recall, sementara k-fold cross validation digunakan untuk memastikan validitas hasil dan mengurangi overfitting. Hasil penelitian menunjukkan bahwa model C4.5 mencapai akurasi sebesar 87%, dengan presisi dan recall untuk kedua kelas masing-masing sekitar 86% dan 88%. Evaluasi menggunakan k-fold cross validation dengan 5 lipatan menunjukkan rata-rata akurasi 87%, menegaskan performa model yang stabil dan cukup baik dalam memprediksi keterlambatan pembayaran mahasiswa.

Kata kunci : *Prediksi keterlambatan pembayaran, Machine Learning, C4.5, Confusion Matrix, K-fold Cross Validation*

1. PENDAHULUAN

Biaya kuliah merupakan sumber dana utama yang tidak hanya mendukung operasional sehari-hari di institusi pendidikan tinggi, tetapi juga menjadi tulang punggung bagi pembangunan infrastruktur dan peningkatan mutu pendidikan [1].

Setiap penundaan dalam penerimaan biaya kuliah memiliki dampak langsung pada kemampuan universitas untuk memenuhi kewajiban finansialnya, seperti membayar gaji staf, memelihara dan meningkatkan fasilitas kampus, serta mendukung program-program akademik dan penelitian [2].

Universitas Muhammadiyah Sukabumi, yang merupakan universitas terbesar di Sukabumi dengan 7 fakultas dan 24 program studi, pengelolaan arus kas menjadi tantangan tersendiri. Keterlambatan pembayaran oleh mahasiswa, jika tidak diantisipasi dan dikelola dengan baik, dapat menimbulkan gangguan serius pada berbagai aspek operasional universitas. Misalnya, ketidaktepatan dalam pembayaran gaji staf atau keterlambatan dalam penyelesaian proyek-proyek infrastruktur dapat menciptakan ketidakpuasan di kalangan karyawan dan mahasiswa, yang pada gilirannya dapat menurunkan citra universitas.

Keterlambatan ini tidak hanya berdampak pada sisi finansial tetapi juga pada kualitas layanan pendidikan yang diberikan. Keterlambatan dalam pembelian peralatan laboratorium, pemeliharaan perpustakaan, atau pengadaan fasilitas teknologi informasi dapat membatasi akses mahasiswa terhadap sumber daya yang dibutuhkan untuk belajar secara

efektif. Hal ini, dalam jangka panjang, dapat memengaruhi reputasi akademik universitas dan mengurangi daya tariknya di mata calon mahasiswa baru [3].

Identifikasi pola keterlambatan pembayaran menjadi krusial dalam upaya universitas untuk memitigasi risiko ini. Namun, mengingat kompleksitas faktor-faktor yang mempengaruhi keterlambatan, mulai dari kondisi ekonomi mahasiswa, kebijakan pembayaran, hingga faktor-faktor pribadi seperti situasi keluarga, pendekatan manual tidak lagi memadai. Oleh karena itu, penggunaan data mining, khususnya dengan algoritma C4.5, dapat memberikan solusi berbasis data untuk masalah ini.

Algoritma C4.5, yang dikenal dengan kemampuannya dalam klasifikasi dan prediksi, dapat diterapkan pada data historis pembayaran mahasiswa untuk membangun model yang mampu memprediksi kemungkinan terjadinya keterlambatan pembayaran. Dengan menganalisis atribut-atribut seperti profil mahasiswa, riwayat pembayaran, program studi, dan kondisi ekonomi, algoritma ini dapat membantu universitas dalam mengidentifikasi kelompok mahasiswa yang berisiko tinggi mengalami keterlambatan. Hasil analisis ini memungkinkan pihak universitas untuk mengambil langkah proaktif, seperti pemberian peringatan dini atau penyesuaian kebijakan pembayaran, guna mengurangi tingkat keterlambatan dan meningkatkan efisiensi pengelolaan keuangan [4].

Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi nyata bagi Universitas Muhammadiyah Sukabumi dalam menghadapi

tantangan keterlambatan pembayaran biaya kuliah, sekaligus memberikan manfaat bagi institusi pendidikan tinggi lainnya yang menghadapi masalah serupa.

2. TINJAUAN PUSTAKA

2.1. Biaya Pendidikan/ SPP

Biaya pendidikan meliputi pengeluaran uang, barang, dan tenaga pendidik. Sumbernya berasal dari pemerintah, masyarakat, dan orangtua, dan digunakan untuk menyelenggarakan sekolah baik secara langsung maupun tidak langsung. SPP adalah sumbangan yang dibayar siswa untuk membantu membiayai operasional sekolah dan memastikan kelancaran kegiatan belajar mengajar [5].

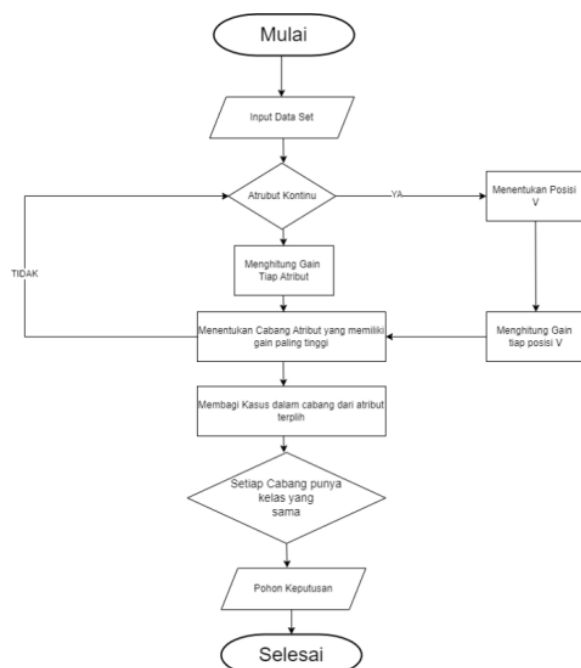
2.2. Machine Learning

Machine learning merupakan cabang dari ilmu kecerdasan buatan yang menjadikan system komputer untuk belajar dan melakukan tugas secara otomatis menggunakan algoritma tertentu. Dalam prosesnya, *machine learning* memiliki 3 kategori, yaitu *Supervised Learning*, *Unsupervised Learning* dan *Reinforcement Learning* [6].

2.3. Prediksi

Prediksi adalah proses memperkirakan nilai masa yang akan datang berdasarkan data masa lampau dan informasi saat ini. Digunakan untuk memprediksi apa yang akan terjadi, membantu perencanaan dan pengambilan Keputusan [7].

2.4. Algoritma C4.5



Gambar 1. Flowchart Algoritma C4.5

Algoritma C4.5 adalah metode klasifikasi dan prediksi yang membangun decision tree dari data pelatihan. Sebagai pengembangan dari ID3, C4.5

dapat menangani *missing values*, data kontinu dan melakukan *pruning*. Algoritma ini menggunakan *training samples* untuk membangun *tree* dan *samples* sebagai parameter dalam klasifikasi, menyediakan struktur hierarki untuk prediksi [8].

Berikut ini tahapan dari algoritma C4.5:

- Memilih atribut sebagai akar
- Membuang cabang untuk tiap-tiap nilai
- Membagi kasus dalam cabang
- Mengulangi proses setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.
- Memilih atribut sebagai akar yang didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada.

2.5. CRISP-DM

CRISP-DM (*Cross Industry Standard Process for Data Mining*) adalah standar untuk proses data mining yang memastikan setiap tahap dilakukan secara terstruktur dan efisien. Pemilihan algoritma yang tepat sangat mempengaruhi kinerja metode data *mining* [9]. CRISP-DM digunakan secara luas dalam bisnis dan penelitian untuk pemecahan masalah. Berikut adalah keenam tahapan dari metode CRISP-DM, yaitu:

- Business understanding* proses untuk mengenali dan mengidentifikasi tujuan bisnis, penilaian situasi dan penetapan tujuan data *mining* [9].
- Data understanding* secara awal data dikumpulkan, dideskripsikan, dikaji dan divalidasi kualitasnya [9].
- Data Preparation* mendeskripsikan *dataset*, kemudian dilakukan pemilihan data, dihubungkan serta dibangun. Proses ini juga mencakup pemilihan tabel, *record* dan atribut data, termasuk pembersihan dan transformasi data [9].
- Modeling* dilakukan pemilihan berbagai teknik pemodelan dan diimplementasikan pada *dataset* yang sudah disiapkan, serta pembangunan model. Dalam prosesnya menerapkan metode pembagian presentase tes dengan membagi data set menjadi dua bagian, data *training* dan data *testing* yang akan diuji pada model yang telah dibangun [9].
- Evaluation* bertujuan untuk menilai keefektifan dan kualitas model sebelum digunakan, serta memastikan apakah model dapat memenuhi tujuan yang ditetapkan pada tahap pemahaman bisnis awal [9].
- Deployment* hasil dari pengetahuan yang telah diperoleh akan dipresentasikan dalam format tertentu agar dapat digunakan oleh pengguna akhir [9].

2.6. Confusion Matrix

Confusion Matrix adalah metode untuk mengevaluasi akurasi model dengan menghasilkan nilai akurasi, presisi, dan *recall*. Akurasi mengukur persentase data yang diklasifikasikan dengan benar, presisi adalah proporsi kasus yang diprediksi positif dan benar-benar positif, dan *recall* adalah proporsi kasus positif yang diprediksi dengan benar [10].

Adapun rumus dari perhitungan akurasi, presisi dan *recall* yaitu sebagai berikut:

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Keterangan:

True Positive (TP): Persentase model dalam memprediksi kelas positif dengan benar.

True Negative (TN): Persentase model dalam memprediksi kelas negatif dengan benar.

False Positive (FP): Persentase model dalam memprediksi kelas positif, tetapi seharusnya negatif.

False Negative (FN): Persentase model dalam memprediksi kelas negatif, tetapi seharusnya positif.

2.7. K-fold Cross Validation

K-fold cross validation adalah metode evaluasi algoritma yang membagi data menjadi *k* bagian, melatih dan menguji model *k* kali. Metode ini memberikan estimasi akurasi yang lebih akurat dan mengurangi kemungkinan *overfitting* dengan mengintegrasikan hasil dari setiap lipatan [11].

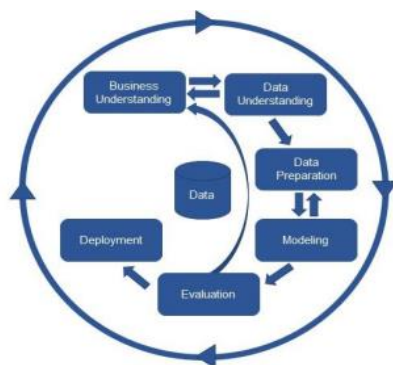
2.8. Google Colaboratory

Google Colaboratory (Google Colab) adalah alat berbasis cloud gratis untuk penelitian, mirip dengan Jupyter Notebook namun dapat diakses secara online. Mendukung berbagai library yang diperlukan untuk pengembangan *Artificial Intelligence* (AI), Colab menyediakan lingkungan notebook interaktif tanpa server, yang memungkinkan pembuatan, berbagi, dan menjalankan kode [12].

2.9. Flask

Flask adalah *micro framework* berbasis Python yang tidak memiliki lapisan abstraksi database, namun mendukung ekstensi untuk menambah fungsionalitas, seperti autentikasi, pemetaan relasional objek, dan validasi formulir. Ekstensi ini sering diperbarui lebih cepat dibandingkan framework flask itu sendiri [13].

3. METODE PENELITIAN



Gambar 2. Tahapan Penelitian CRISP-DM

Metode penelitian ini bertujuan memberikan pemahaman tentang masalah yang akan diteliti, termasuk prosedur pengumpulan data dari observasi pada lembaga. Tahap ini juga mencakup kebutuhan untuk menyusun jadwal penelitian. Alur penelitian mengikuti metode CRISP-DM (*Cross Industry Standard Process for Data Mining*).

3.1. Business Understanding

Tahap ini fokus pada pemahaman detail kebutuhan penelitian dalam lingkup bisnis atau unit penelitian. Pemahaman tersebut dirumuskan menjadi rencana awal data mining untuk mencapai tujuan, yaitu menggunakan algoritma C4.5 untuk memprediksi tingkat keterlambatan pembayaran biaya kuliah mahasiswa di Universitas Muhammadiyah Sukabumi.

3.2. Data Understanding

Tahap ini dilakukan pengumpulan data dengan dilakukan observasi dan wawancara ke biro keuangan Universitas Muhammadiyah Sukabumi. Kemudian, mengumpulkan data primer yang diperoleh dari UPT SIM Universitas Muhammadiyah Sukabumi. *Dataset* meliputi data riwayat pembayaran mahasiswa, serta data pekerjaan dan penghasilan orang tua, dari tahun 2020/2021 hingga 2023/2024.

3.3. Data Preparation

Setelah memahami data, peneliti menyusun *dataset* dengan cara menentukan atribut data yang akan menjadi kriteria dan mempersiapkannya untuk proses pemodelan *data mining*. Pada *data preparation* ini mencakup *data cleaning*, *data selection* dan *data transformation* agar cocok dengan pemodelan yang dilakukan.

a. Data cleaning

Pada tahap ini data yang memuat nilai *null*, duplikasi data atau data yang tidak lengkap akan dibersihkan dengan menghapus atribut tersebut, yang kemudian tidak akan digunakan pada tahap selanjutnya.

b. Data selection

Proses penyeleksian fitur yaitu untuk menentukan variabel atau atribut mana yang akan digunakan sebagai input untuk proses pemodelan klasifikasi dalam memprediksi menggunakan algoritma C4.5. Terdapat fitur-fitur yang berperan penting mencakup nama lengkap, nama prodi, nama fakultas, semester, tahun akademik, jumlah bayar, tanggal bayar, pekerjaan ayah, jumlah penghasilan, kolom lunas dan kolom terlambat bayar.

c. Data transformation

Proses dalam mengubah format atau struktur data menjadi format yang diinginkan, sesuai dengan kebutuhan dan ketentuan algoritma yang digunakan dalam penelitian ini.

3.4. Modeling

Tahap pemodelan merupakan tahap yang langsung melibatkan proses *data mining*, seperti pemilihan algoritma, pendekatan *data mining* dan penentuan nilai parameter yang optimal. Teknik dalam pemodelan *data mining* yang diterapkan dalam penelitian ini yaitu klasifikasi prediksi menggunakan metode *decision tree* C4.5. Penerapan metode klasifikasi C4.5 bertujuan untuk memprediksi tingkat keterlambatan mahasiswa dalam membayar biaya kuliah.

3.5. Evaluation

Tahap evaluasi merupakan tahap yang difokuskan pada model yang dihasilkan algoritma C4.5 untuk mengukur performa pada model yang telah dibuat apakah sesuai dengan tujuan awal yang telah ditetapkan. Tujuan dalam tahapan ini untuk mencapai tingkat akurasi yang stabil, sebagai bukti keberhasilan dari penelitian yang dilakukan. Adapun evaluasi dan uji validitas data pada penelitian ini menggunakan confusion matrix dan k-fold cross validation.

3.6. Deployment

Pada tahap terakhir yaitu deployment, peneliti menerapkan model prediksi dengan merancang antarmuka berbasis web untuk memprediksi tingkat keterlambatan pembayaran biaya kuliah oleh mahasiswa. Dalam prosesnya, peneliti menggunakan teks editor yaitu Visual Studio Code dan Flask untuk membuat sistem berbasis web serta menggabungkan hasil dari model yang sudah dibuat ke dalam sistem tersebut.

4. HASIL DAN PEMBAHASAN

4.1. Business Understanding

Pemahaman bisnis merupakan tahap pertama dalam metode CRISP-DM. Memahami masalah bisnis sangat penting untuk menentukan data apa yang diperlukan dan bagaimana data tersebut harus diproses. Fokus utama pada studi kasus ini adalah memprediksi apakah seorang mahasiswa akan terlambat dalam membayar biaya kuliah. Keterlambatan ini bisa disebabkan oleh beberapa faktor, seperti program studi yang diambil mahasiswa, pekerjaan serta jumlah penghasilan orang tua dan waktu pembayaran terkait semester yang berjalan.

Universitas Muhammadiyah Sukabumi berusaha untuk memahami faktor-faktor yang berpengaruh terhadap keterlambatan tersebut sehingga dapat menerapkan pola-pola pencegahan, seperti memberikan peringatan dini kepada mahasiswa yang kemungkinan besar akan terlambat membayar, guna mengurangi risiko jumlah mahasiswa yang terlambat membayar biaya kuliah. Dengan memanfaatkan algoritma C4.5, lembaga dapat membangun model klasifikasi yang akurat untuk memprediksi keterlambatan pembayaran biaya kuliah.

4.2. Data Understanding

Tahap selanjutnya setelah memahami kebutuhan bisnis, yaitu mengumpulkan dan memahami data. Dataset yang digunakan dalam penelitian ini diperoleh dari UPT SIM UMMI yang dikumpulkan dari Sistem Informasi dan Akademik (SIK) Universitas Muhammadiyah Sukabumi. Terdapat dua file csv yang digunakan, file pertama berisi data pembayaran mahasiswa, sedangkan file yang kedua berisi data pekerjaan dan penghasilan orang tua. Dataset yang digunakan merupakan data dari tahun akademik 2020/2021 sampai dengan 2023/2024 yang berjumlah 41.802 data.

Kedua file csv dataset digabungkan berdasarkan beberapa kolom kunci, yaitu *nama_lengkap*, *nama_prodi* dan *nama_fakultas*. Pada tahap ini menghasilkan dataset gabungan yang lengkap.

	nama_lengkap	nama_prodi	nama_fakultas	jumlah_bayar	tanggal_bayar	tahun_akademik	semester	nama_pekerjaan_orang_tua	jumlah_penghasilan
0	DARMAWAN EFENDI	ILMU HUKUM	FAKULTAS HUKUM	200000	2019-02-07 13:45:43.12889	2019/2019	1	Tidak Bersewa	< 3.000.000
1	DARMAWAN EFENDI	ILMU HUKUM	FAKULTAS HUKUM	150000	2020-02-20 09:32:54	2019/2020	1	Tidak Bersewa	< 3.000.000
2	DARMAWAN EFENDI	ILMU HUKUM	FAKULTAS HUKUM	975000	2020-03-26 16:33:03	2019/2020	1	Tidak Bersewa	< 3.000.000
3	DARMAWAN EFENDI	ILMU HUKUM	FAKULTAS HUKUM	722000	2020-06-16 08:26:53.08696	2019/2020	1	Tidak Bersewa	< 3.000.000
4	DARMAWAN EFENDI	ILMU HUKUM	FAKULTAS HUKUM	150000	2020-09-26 11:08:05	2020/2021	1	Tidak Bersewa	< 3.000.000

Gambar 3. Penggabungan Dua Dataset

	nama_lengkap	nama_prodi	nama_fakultas	jumlah_bayar	tanggal_bayar	tahun_akademik	semester	nama_pekerjaan_orang_tua	jumlah_penghasilan
40201	ZEN HARNO MUBAROK	PENDIDIKAN JERAMAH KESEHATAN DAN KEWASIHAN	FAKULTAS KESEHATAN DAN KEWASIHAN	175000	2024-03-26 15:50:19	2023/2024	1	Guru/Dosen Swasta	< 3.000.000
40202	MUHAMMAD FATHAN MULLANA MANULIS	TEKNIK INFORMATIKA	FAKULTAS SAINS DAN TEKNOLOGI	486000	2024-03-26 14:32:37	2023/2024	1	Pegawai Swasta	5.000.000 - 10.000.000
40203	TIBAL AR ROBBI	TEKNIK INFORMATIKA	FAKULTAS SAINS DAN TEKNOLOGI	398000	2024-03-26 10:23:20	2023/2024	1	Tidak Bersewa	3.000.000 - 5.000.000
40204	TIBAL AR ROBBI	TEKNIK INFORMATIKA	FAKULTAS SAINS DAN TEKNOLOGI	1000000	2024-03-26 10:23:21	2023/2024	1	Tidak Bersewa	3.000.000 - 5.000.000
40205	GENI SEPTIADI	ADMINISTRASI BISNIS	FAKULTAS ILMU SOSIAL	400000	2024-04-01 11:13:37	2023/2024	2	Pegawai Swasta	< 3.000.000

Gambar 4. Penggabungan Dua Dataset

Data ini akan diproses menggunakan algoritma C4.5. Algoritma C4.5 merupakan algoritma pengambilan keputusan yang digunakan untuk membentuk pohon keputusan dari data yang tersedia. Pohon keputusan tersebut berguna untuk melakukan klasifikasi atau prediksi berdasarkan aturan-aturan yang dihasilkan dari analisis data. Dengan menerapkan algoritma C4.5, akan ditemukan pola atau aturan yang menjelaskan faktor-faktor yang mempengaruhi seorang mahasiswa terlambat dalam membayar biaya kuliah. Sebagai contoh, “Jika mahasiswa yang membayar biaya kuliahnya tidak tepat waktu dari waktu yang telah ditentukan, maka kemungkinan besar mahasiswa tersebut akan masuk ke kategori yang terlambat membayar”.

4.3. Data cleaning

Tahap ini merupakan proses pembersihan data dari *missing values* dan duplikasi, yang dapat mempengaruhi hasil analisis jika tidak ditangani dengan tepat. Hal pertama yang dilakukan yaitu menghitung dan mengecek apakah terdapat *missing values* dan duplikasi pada data tersebut.

Dalam mencari *missing values*, nilai yang dihasilkan adalah 0, yang berarti tidak terdapat nilai ‘Nan’. Tetapi, terdapat 159 duplikasi data.

```

0
nama_lengkap 0
nama_prodi 0
nama_fakultas 0
jumlah_bayar 0
tanggal_bayar 0
tahun_akademik 0
semester 0
nama_pekerjaan_ayah 0
jumlah_penghasilan 0

dtype: int64

```

Gambar 5. Mencari Missing Values

```

159

```

Gambar 6. Menghitung duplikasi data

```

df_merged
df_merged.drop_duplicates()
df_merged.duplicated().sum()

```

Kode program di atas digunakan untuk menghapus semua baris yang duplikat pada *dataset* serta memastikan bahwa tidak ada baris yang muncul lebih dari satu kali.

```

0

```

Gambar 7. Menghapus Duplikasi

4.4. Data selection

Setelah tahap data cleaning selesai, tahap berikutnya yaitu data selection, yang mencakup pemilihan dan pengelompokan data yang tepat untuk proses yang lebih lanjut. Data dikelompokkan ('groupby') berdasarkan kolom kunci. Proses pengelompokan ini dilanjutkan dengan akumulasi data pada kolom-kolom lain, seperti menjumlahkan 'jumlah_bayar', mengambil nilai maksimal 'tanggal_bayar' dan memilih nilai pertama dari 'nama_pekerjaan_ayah' serta 'jumlah_penghasilan'. Pengelompokan dan akumulasi data ini bertujuan untuk mencari total pembayaran per-tahun akademik tiap mahasiswa. Dalam proses data selection, terdapat sebagian kolom-kolom yang tidak tepat atau tidak diperlukan untuk analisis lebih lanjut dihapus dari dataset.

```

data=
df_grouped.drop(['tanggal_bayar', 'nama_lengkap', 'tahun_akademik', 'jumlah_bayar', 'lunas'], axis=1)

```

```

nama_prodi      nama_fakultas  semester  nama_pekerjaan_ayah  jumlah_penghasilan  terlambat_bayar
0      PENDOKAN JASMANI KESEHATAN DAN REKREASI      FAKULTAS KESUKURAN DAN ILMU PENDOKAN      1      Wirawanda  3.000.000 - 5.000.000      iya
1      PENDOKAN JASMANI KESEHATAN DAN REKREASI      FAKULTAS KESUKURAN DAN ILMU PENDOKAN      1      Wirawanda  3.000.000 - 5.000.000      iya
2      PENDOKAN JASMANI KESEHATAN DAN REKREASI      FAKULTAS KESUKURAN DAN ILMU PENDOKAN      1      Wirawanda  3.000.000 - 5.000.000      iya
3      PENDOKAN JASMANI KESEHATAN DAN REKREASI      FAKULTAS KESUKURAN DAN ILMU PENDOKAN      1      Wirawanda  3.000.000 - 5.000.000      iya
4      AGRIBISNIS      FAKULTAS PERTANIAN      1      Wirawanda  + 3.000.000      iya
11772      TEKNIK SIPIL      FAKULTAS SAINS DAN TEKNOLOGI      1      Pegawai Swasta  3.000.000 - 5.000.000      iya
11774      SASTRA INGGRIS      FAKULTAS ILMU SOSIAL      1      Tidak Ditemui      + 3.000.000      iya
11776      SASTRA INGGRIS      FAKULTAS ILMU SOSIAL      1      Tidak Ditemui      + 3.000.000      iya
11778      SASTRA INGGRIS      FAKULTAS ILMU SOSIAL      1      Tidak Ditemui      + 3.000.000      iya
11777      SASTRA INGGRIS      FAKULTAS ILMU SOSIAL      2      Tidak Ditemui      + 3.000.000      tidak
11778 rows x 6 columns

```

Gambar 8. Menghapus Kolom-kolom

Pada tahap ini, kolom-kolom di atas dihapus dari *dataset* 'df_grouped', karena kolom-kolom tersebut tidak diperlukan untuk proses pemodelan. Hal ini dilakukan agar memastikan bahwa fitur yang penting yang disertakan dalam analisis, sehingga model dapat fokus pada variabel-variabel yang benar-benar mempengaruhi hasil prediksi.

4.5. Data transformation

Setelah mengakumulasi total pembayaran per-tahun akademik dari tiap mahasiswa pada tahap *data selection*, maka dibuatlah tabel lunas yang didapatkan dari jumlah_bayar dan patokan biaya per-semester dari tiap program studi. Sehingga ketika jumlah bayar sudah melewati jumlah pembayaran biaya kuliah program studinya, maka dinyatakan lunas.

```

nama_lengkap      nama_prodi      nama_fakultas  tahun_akademik  semester  jumlah_bayar  tanggal_bayar  nama_pekerjaan_ayah  jumlah_penghasilan  lunas
0      ALIYAH FAUZI      PENDOKAN JASMANI KESEHATAN DAN REKREASI      FAKULTAS KESUKURAN DAN ILMU PENDOKAN      2023/2024      1      870000      2023-06-22 19:40:48      Wirawanda  3.000.000 - 5.000.000      sudah
1      ALIYAH FAUZI      PENDOKAN JASMANI KESEHATAN DAN REKREASI      FAKULTAS KESUKURAN DAN ILMU PENDOKAN      2023/2024      1      870000      2023-06-19 13:43:59      Wirawanda  3.000.000 - 5.000.000      sudah
2      ALIYAH FAUZI      PENDOKAN JASMANI KESEHATAN DAN REKREASI      FAKULTAS KESUKURAN DAN ILMU PENDOKAN      2023/2024      1      870000      2023-07-10 14:17:46      Wirawanda  3.000.000 - 5.000.000      sudah
3      ALIYAH FAUZI      PENDOKAN JASMANI KESEHATAN DAN REKREASI      FAKULTAS KESUKURAN DAN ILMU PENDOKAN      2023/2024      1      870000      2023-07-23 08:52:29      Wirawanda  3.000.000 - 5.000.000      sudah
4      ALIYAH FAUZI      PENDOKAN JASMANI KESEHATAN DAN REKREASI      FAKULTAS KESUKURAN DAN ILMU PENDOKAN      2023/2024      1      870000      2023-07-23 08:52:29      Wirawanda  3.000.000 - 5.000.000      sudah
11773      ZYLA RAMBA ALYIA      PENDOKAN JASMANI KESEHATAN DAN REKREASI      FAKULTAS KESUKURAN DAN ILMU PENDOKAN      2023/2024      1      700000      2023-05-17 09:28:51      Pegawai Swasta  3.000.000 - 5.000.000      sudah
11774      ZYLA RAMBA ALYIA      PENDOKAN JASMANI KESEHATAN DAN REKREASI      FAKULTAS KESUKURAN DAN ILMU PENDOKAN      2023/2024      1      800000      2023-06-17 01:50:54      Total Ditemui      + 3.000.000      sudah
11776      ZYLA RAMBA ALYIA      PENDOKAN JASMANI KESEHATAN DAN REKREASI      FAKULTAS KESUKURAN DAN ILMU PENDOKAN      2023/2024      1      800000      2023-06-17 01:50:54      Total Ditemui      + 3.000.000      sudah
11778      ZYLA RAMBA ALYIA      PENDOKAN JASMANI KESEHATAN DAN REKREASI      FAKULTAS KESUKURAN DAN ILMU PENDOKAN      2023/2024      1      200000      2023-12-29 01:50:57      Total Ditemui      + 3.000.000      sudah
11777      ZYLA RAMBA ALYIA      PENDOKAN JASMANI KESEHATAN DAN REKREASI      FAKULTAS KESUKURAN DAN ILMU PENDOKAN      2023/2024      2      200000      2023-05-19 01:50:57      Total Ditemui      + 3.000.000      sudah
11778 rows x 10 columns

```

Gambar 9. Membuat Tabel Lunas

Selanjutnya membuat tabel terlambat bayar, untuk mengetahui informasi mengenai status keterlambatan pembayaran setiap mahasiswa. Tabel ini berfungsi sebagai label atau target yang akan digunakan dalam model prediksi. Tabel terlambat bayar berasal dari bulan bayar, semester dan tabel lunas. Rentang tanggal pembayaran untuk setiap tahun akademik dan semester diambil dari masa perbaikan atau perpanjangan pengisian KRS (Kartu Rencana Studi).

Selanjutnya, untuk memeriksa keterlambatan pembayaran, periksa apakah tahun akademik dan semester yang diberikan ada dalam 'tanggal_bayar_dict'. Kemudian, mengambil tanggal mulai dan tanggal akhir pembayaran. Jika 'tanggal_bayar' berada dalam rentang yang ditentukan ('start_date' hingga 'end_date'), maka akan mengembalikan 'tidak' yang berarti tidak terlambat jika status pembayaran sudah lunas. Jika tidak lunas, maka akan mengembalikan 'iya' yang berarti terlambat. Jika salah satu dari kondisi sebelumnya tidak terpenuhi (contohnya, tahun akademik atau semester tidak ditemukan, atau pembayaran dilakukan di luar periode yang ditentukan), fungsi akan mengembalikan 'iya', yang menunjukkan bahwa mahasiswa dianggap terlambat membayar.

```

if      tahun_akademik      in
tanggal_bayar_dict:
        if      semester      in
tanggal_bayar_dict[tahun akademik]
:

```

```

start_date, end_date =
tanggal_bayar dict[tahun akademik]
[semester]

if start_date <=
tanggal_bayar <= end_date:
    return 'tidak' if
lunas == 'sudah' else 'iya'
return 'iya'

```

Gambar 10. Membuat Tabel terlambat Bayar

```

terlambat_bayar_iya =
len(df_grouped[df_grouped['terlambat_bayar'] == 'iya'])
terlambat_bayar_tidak =
len(df_grouped[df_grouped['terlambat_bayar'] == 'tidak'])
print(f'data terlambat bayar iya
sebanyak {terlambat_bayar_iya} dan
data terlambat bayar tidak
sebanyak {terlambat_bayar_tidak}')

```

data terlambat bayar iya sebanyak 11633 dan data terlambat bayar tidak sebanyak 145

Gambar 11. Jumlah Data Terlambat dan Tidak Terlambat Bayar

Kode program di atas bertujuan untuk mengetahui jumlah data terlambat_bayar iya dan tidak. Terdapat 11.633 data yang terlambat bayar dan 145 data yang tidak terlambat bayar.

Mencakup konversi data dari format asli ke format lain yang lebih sesuai untuk proses pemodelan. Transformasi data dilakukan dengan mengonversi kolom-kolom kategorikal menjadi nilai numerik menggunakan 'LabelEncoder'.

Gambar 12. Mencetak Kategori Asli dan Nilai Numerik

Pada tahap ini juga dilakukan pembagian data menjadi data latih dan data uji dengan rasio 80:20, menunjukkan bahwa 20% dari data akan digunakan sebagai data uji, sedangkan 80% sisanya digunakan untuk melatih model.

data traing sebanyak 18612 dan data test sebanyak 4654

Gambar 13. Data Training dan Data Testing

4.6. Modeling

Tahap berikutnya yaitu *modeling* atau pemodelan. Pada tahap ini fokus pada pembuatan dan pelatihan model dengan data yang sudah disiapkan. Hal ini mencakup pemilihan algoritma, penyesuaian parameter model dan proses pelatihan tersebut. Pada penelitian ini, berfokus pada penggunaan *Decision Tree Classifier* dengan teknik pencarian grid (GridSearchCV) untuk mengidentifikasi kombinasi parameter terbaik yang menghasilkan kinerja model yang optimal.

'GridSearchCV' digunakan untuk melakukan pencarian grid, yaitu teknik pencarian *brute-force* untuk mengidentifikasi kombinasi terbaik dari hyperparameter yang telah didefinisikan dalam 'param_grid'. Kemudian dilakukan evaluasi performa model untuk setiap kombinasi tersebut menggunakan *cross-validation* ('cv=5'). Hal ini,, *cross-validation* dibagi menjadi 5 lapis (*folds*), dimana *dataset* pelatihan dibagi menjadi 5 bagian, lalu setiap bagian secara bergiliran digunakan sebagai set validasi, sementara sisanya digunakan untuk melatih model. Sementara 'scoring=accuracy' menunjukkan bahwa metrik yang digunakan untuk mengevaluasi performa model adalah akurasi.

```

grid_search =
GridSearchCV(estimator=model,
param_grid=param_grid, cv=5,
scoring='accuracy')

grid_search.fit(X_train,
y_train)

```

Setelah parameter terbaik telah diidentifikasi, model *decision tree* dilatih ulang menggunakan kombinasi parameter tersebut ('best_estimator_'). Kemudian model ini dilatih kembali pada seluruh *dataset* pelatihan untuk memastikan bahwa model telah sepenuhnya dioptimalkan sebelum digunakan pada data pengujian.

```

best_model =
grid_search.best_estimator_
best_model.fit(X_train, y_train)

```

Gambar 14. Parameter Terbaik dan Skor Validasi Terbaik

4.7. Evaluation

Tahap selanjutnya yaitu *evaluation* atau evaluasi. Evaluasi model merupakan tahap dimana menilai kinerja model berdasarkan hasil prediksi terhadap uji model yang sudah dilatih, kemudian dievaluasi menggunakan metrik-metrik seperti akurasi, presisi dan *recall*. Selain itu, laporan klasifikasi dan *confusion matrix* digunakan untuk memberikan *insight* lebih lanjut mengenai performa model, termasuk bagaimana

model menangani ketidakseimbangan data. *K-fold cross-validation* juga dilakukan untuk memastikan bahwa model memiliki kinerja yang konsisten pada berbagai subset data.

Metrik Evaluasi:
 Akurasi: 0.870042973785991
 Presisi: 0.8643130582730753
 Recall: 0.8766177739430544

Laporan Klasifikasi:

	precision	recall	f1-score	support
Iya	0.88	0.86	0.87	2336
Tidak	0.86	0.88	0.87	2318
accuracy			0.87	4654
macro avg	0.87	0.87	0.87	4654
weighted avg	0.87	0.87	0.87	4654

Gambar 15. Akurasi, Presisi dan Recall

Gambar menunjukkan nilai akurasi sebesar 0.87 atau 87%. Hal ini diartikan model *decision tree* yang digunakan dalam proses klasifikasi mampu mengklasifikasikan data uji dengan benar sebanyak 87% dari total data uji yang digunakan. Presisi menunjukkan hasil 0.86 dan 0.88, yang artinya model cukup baik. *Recall* menunjukkan bahwa model juga baik dengan nilai 0.86 dan 0.88. Secara keseluruhan pada laporan klasifikasi, model menunjukkan performa yang solid, dengan keseimbangan yang baik antara presisi dan *recall* untuk kedua kelas, yaitu kelas 'iya' dan kelas 'tidak', serta akurasi keseluruhan yang tinggi.

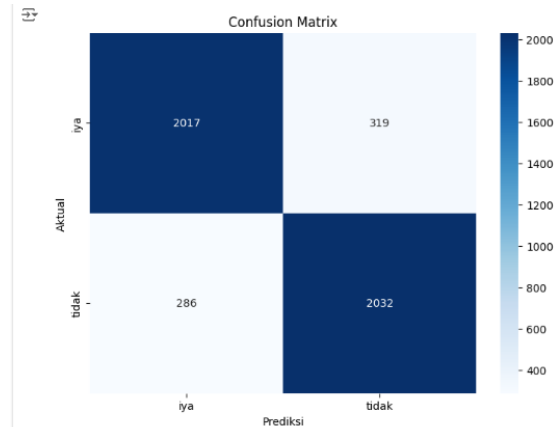
Hasil K-Fold Cross Validation (Accuracy):

Lipatan 1:	0.8753693258125168
Lipatan 2:	0.8654311039484287
Lipatan 3:	0.8747984954325632
Lipatan 4:	0.8656636217087588
Lipatan 5:	0.8651262761955938
Rata-rata akurasi:	0.8692777646195722

Gambar 16. Hasil *K-fold Cross Validation*

Gambar di atas menunjukkan hasil penerapan *k-fold cross-validation* dengan 5 lipatan pada model *decision tree*. Teknik ini digunakan untuk mengevaluasi performa model secara lebih akurat dengan membagi data pelatihan menjadi 5 bagian, lalu melatih dan menguji model pada masing-masing bagian secara bergiliran. Hasil yang ditampilkan menunjukkan akurasi untuk setiap lipatan, yang berkisar antara 0.86 hingga 0.87. Rata-rata akurasi dari kelima lipatan ini adalah 0.869 atau sekitar 87%. Hal ini menandakan bahwa model memiliki performa yang konsisten dan tidak *overfitting*, sehingga dapat diandalkan dalam memprediksi data baru.

Terakhir, membuat dan memvisualisasikan *confusion matrix* dari hasil prediksi model. *Confusion matrix* adalah tabel yang digunakan untuk mengevaluasi kinerja model klasifikasi dengan membandingkan hasil prediksi model terhadap data aktual.



Gambar 17. *Confusion Matrix*

Gambar di atas menunjukkan jumlah kasus *true* dan *false* yang diprediksi oleh model untuk dua kelas ('iya' dan 'tidak').

- True Positives (TP):** 2017 (memprediksi 2017 sampel sebagai 'iya' terlambat dan aktual 'iya' terlambat. Hal ini menunjukkan prediksi positif yang benar).
- True Negatives (TN):** 2032 (memprediksi 2032 sampel sebagai 'tidak' terlambat dan aktual 'tidak' terlambat. Hal ini menunjukkan prediksi negatif yang benar).
- False Positives (FP):** 319 (memprediksi 319 sampel sebagai 'iya' terlambat tetapi aktual 'tidak' terlambat. Hal ini menunjukkan prediksi positif yang salah).
- False Negatives (FN):** 286 (memprediksi 286 sampel sebagai 'tidak' terlambat tetapi aktual 'iya' terlambat. Hal ini menunjukkan prediksi negatif yang salah).

4.8. Deployment

Tahap terakhir dalam metode CRISP-DM adalah *deployment*. Tahap ini merupakan tahap dimana model telah dievaluasi dan dianggap cukup baik dibangun dan diterapkan ke dalam sistem berbasis *web* untuk digunakan pada data baru. Model yang sudah dilatih disimpan menggunakan *pickle* dan dapat digunakan kembali untuk memprediksi apakah mahasiswa baru yang belum muncul di data pelatihan akan terlambat membayar atau tidak.

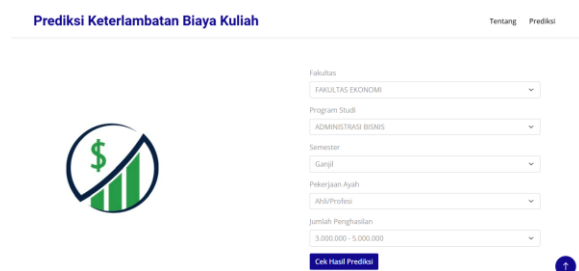
Website ini dibuat untuk memudahkan pengguna dalam menginput data yang memengaruhi keterlambatan mahasiswa dalam membayar biaya kuliah. Setelah data dimasukkan, *website* ini mengirimkan informasi tersebut ke model *machine learning* yang berjalan di server. Model tersebut kemudian memproses data dan memberikan prediksi keterlambatan mahasiswa dalam pembayaran (terlambat atau tidak terlambat) secara *real-time*.

Adapun tampilan *website*-nya sebagai berikut:



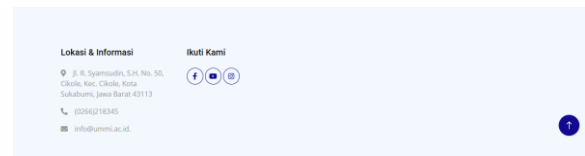
Gambar 18. Tampilan Halaman Tentang Website

Halaman utama menampilkan penerapan model prediksi keterlambatan mahasiswa dalam membayar biaya kuliah yang terintegrasi di situs web “Keuangan UMMI – Prediksi Terlambat”. Pada halaman ini juga ditampilkan penjelasan terkait lembaga penelitian dan website yang dibangun.



Gambar 19. Tampilan Halaman Prediksi

Melalui halaman ini, pengguna atau staf keuangan Universitas Muhammadiyah Sukabumi dapat memprediksi keterlambatan mahasiswa yang terlambat dalam membayar biaya kuliah secara mudah berdasarkan data input fakultas, program studi, semester, pekerjaan ayah dan jumlah penghasilan. Fitur ini dirancang untuk membantu biro keuangan Universitas Muhammadiyah Sukabumi dalam mengambil langkah untuk mencegah tingginya tingkat keterlambatan pembayaran yang terjadi.



Gambar 20. Tampilan Halaman Footer

Pada halaman ini ditampilkan lokasi dan informasi berupa nomor telepon serta email lembaga. Selain itu, terdapat tiga media sosial lembaga yaitu *facebook*, *instagram* dan *youtube*.

4.9. Pengujian Fungsionalitas

Pengujian fungsionalitas adalah proses untuk menilai apakah sebuah situs web yang dibangun sudah siap digunakan atau masih memerlukan perbaikan. Berikut ini adalah hasil pengujian terhadap Website Prediksi Keterlambatan Mahasiswa dalam Membayar Biaya Kuliah yang telah dilakukan:

Tabel 1. Pengujian Fungsionalitas

Pengujian Halaman Tentang Website				
No	Pengujian Fungsi	Deskripsi	Output	Hasil Pengujian
1	Halaman Tentang Website	Mengak-ses halaman Tentang Website dengan menekan tombol “Tentang”	Jika menekan tombol “Tentang” dengan benar, maka akan beralih ke halaman Tentang Website Prediksi Keterlambatan Pembayaran Biaya Kuliah	SUKSES
Pengujian Halaman Prediksi Keterlambatan				
1	Halaman Prediksi	Mengak-ses halaman prediksi dengan menekan tombol “Mulai Prediksi”.	Jika mengisi form-form untuk mengecek hasil prediksi dengan benar, maka akan keluar hasil prediksinya yaitu, “terlambat” atau “tidak terlambat”	SUKSES
Pengujian Halaman Footer				
1	Halaman Footer	Mengak-ses halaman footer dengan meng-scroll halaman sampai bawah.	Jika meng-scroll halaman sampai ke bawah dengan benar, maka terdapat halaman footer yang berisi lokasi dan social media lembaga.	SUKSES

5. KESIMPULAN DAN SARAN

Kesimpulan dari penelitian ini menunjukkan bahwa Algoritma C4.5 efektif dalam memprediksi keterlambatan pembayaran biaya kuliah dengan akurasi 87%, menggunakan variabel seperti fakultas, program studi, dan penghasilan orang tua, serta berhasil diimplementasikan dalam bentuk website. Untuk meningkatkan kinerja model ini, disarankan agar penelitian lebih lanjut menambahkan variabel data seperti jumlah SKS, status keaktifan mahasiswa, dan jenis bantuan keuangan, serta membandingkan hasil dengan algoritma dan metode data mining

lainnya. Selain itu, pengembangan fitur website seperti dashboard statistik interaktif diusulkan untuk memperkaya tampilan dan fungsionalitas, sementara evaluasi berkelanjutan diperlukan untuk memastikan relevansi model seiring perubahan data dan kondisi ekonomi.

DAFTAR PUSTAKA

- [1] F. L. Handayani *et al.*, “Analisis Penggunaan Teknologi pada Pembelajaran IPA di Sekolah Dasar,” *J. Teknol. Pendidik.*, vol. 1, no. 2, p. 7, 2023, doi: <https://doi.org/10.47134/jtp.v1i2.105>.

- [2] T. A. Y. Siswa and R. P. Wibowo, "Komparasi Metode Seleksi Fitur Dalam Prediksi Keterlambatan Pembayaran Biaya Kuliah," *Teknika*, vol. 12, no. 1, pp. 73–82, 2023, doi: 10.34148/teknika.v12i1.601.
- [3] L. Iryani, "Penerapan Datamining Menentukan Minat Baca Mahasiswa di Perpustakaan Universitas Bina Darma Palembang Menggunakan Metode Clustering," *J. Inf. Technol. Comput. Sci.*, vol. 3, no. 1, pp. 82–89, 2020.
- [4] L. Y. L. Gaol, M. Safii, and D. Suhendro, "Prediksi Kelulusan Mahasiswa Stikom Tunas Bangsa Prodi Sistem Informasi dengan Menggunakan Algoritma C4.5," *BRAHMANA J. Penerapan Kecerdasan Buatan*, vol. 2, no. 2, pp. 97–106, 2021.
- [5] S. Khoiridah, D. Ferriswara, F. Augustinah, and I. Devy, *Merdeka Berpikir: Catatan Harian Pandemi Covid-19*. Surabaya: Unitomo Press, 2020.
- [6] A. D. Sidik and A. Ansawarman, "Prediksi Jumlah Kendaraan Bermotor Menggunakan Machine Learning," *Formosa J. Multidiscip. Res.*, vol. 1, no. 3, pp. 558–568, 2022, doi: <https://doi.org/10.55927/fjmr.v1i3.745>.
- [7] S. Agustinus and C. Lesmana, "Perbandingan Analisis Perkuatan Struktur Pelat dengan Metode Elemen Hingga," *J. Tek. Sipil*, vol. 15, no. 1, pp. 1–25, 2019, doi: <https://doi.org/10.28932/jts.v15i1.1852>.
- [8] S. Widianingsih, "Perbandingan Metode Data Mining Untuk Prediksi Nilai dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4.5, Naïve Bayes, KNN, dan SVM," *J. Tekno Insentif*, vol. 13, no. 1, pp. 16–25, 2019, doi: <https://doi.org/10.36787/jti.v13i1.78>.
- [9] M. A. Hasanah, S. Soim, and A. S. Handayani, "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," vol. 5, no. 2, 2021.
- [10] V. S. Ginting, Kusri, and E. T. Luthfi, "Keterlambatan Pembayaran Uang Sekolah Menggunakan Python," vol. 4, no. 1, 2020.
- [11] A. Purwanto, "Jurnal Teknoinfo," *Anal. PERBANDINGAN KINERJA Algoritma C4.5 DAN Algoritma K-NEAREST NEIGHBORS UNTUK KLASIFIKASI PENERIMA BEASISWA*, vol. 17, pp. 48–58, 2023, [Online]. Available: <https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/article/view/336/329>.
- [12] E. Supriyadi *et al.*, "Pendampingan Guru Matematika di Cianjur dalam Penggunaan Bahasa Pemrograman Python dengan Google Collaboratory," *DIMASTEK J. Pengabd. Kpd. Masy. Berbas. Teknol.*, vol. 3, no. 2, pp. 39–44, 2023.
- [13] D. Renita, E. D. Wahyuni, and S. F. A. Wati, "Implementasi Data Mining Pada Sistem Informasi Menggunakan Fp-Growth," *J. Ilm. Sist. Inf. dan Ilmu Komput.*, vol. 3, no. 3, pp. 1–20, 2023, doi: <https://doi.org/10.55606/juisik.v3i3.620>.