PAPER NAME

Template CCIT Journal Pak_Desta_revisi_akhir_done_indonesia.doc

AUTHOR

Destaindo Destaindo

WORD COUNT

2021 Words

CHARACTER COUNT

13905 Characters

PAGE COUNT

8 Pages

FILE SIZE

302.0KB

SUBMISSION DATE

Sep 1, 2024 11:29 PM GMT+7

REPORT DATE

Sep 1, 2024 11:29 PM GMT+7

13% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- Crossref database
- Crossref Posted Content database
- 11% Submitted Works database

Excluded from Similarity Report

- Internet database
- Publications database
- Bibliographic material
- Cited material

ISSN : 1978 -8282, Online ISSN: 2655-4275

Perbandingan Data Mining untuk Klasifikasi Tingkat Kelulusan Mahasiswa Menggunakan Metode Naive Bayes, Decision Tree, dan Random Forest (Studi Kasus Program Sarjana di Universitas Mitra Indonesia)

Tri destanto[*1], Handoyo Widi Nugroho[2]
[2] Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Institut Informatika dan Bisnis Darmajaya, Indonesia
E-mail: [*1] **destanto.2121210007@mail.darmajaya.ac.id** [2] handoyo.wn@darmajaya.ac.id

Abstract

Penelitian ini bertujuan untuk menerapkan teknik data mining dalam mengklasifikasikan tingkat kelulusan mahasiswa pada Program Sarjana di Universitas Mitra Indonesia. Metode yang digunakan dalam penelitian ini meliputi Naive Bayes, Decision Tree, dan Random Forest. Data yang digunakan termasuk data akademik mahasiswa, seperti nilai, kehadiran, dan informasi demografis lainnya. Langkah-langkah penelitian mencakup pengumpulan data, pembersihan data, analisis data, dan penerapan algoritma data mining. Hasil penelitian menunjukkan bahwa metode Random Forest memberikan akurasi tertinggi dibandingkan dengan Naive Bayes dan Decision Tree dalam memprediksi tingkat kelulusan mahasiswa. Metode Random Forest mencapai akurasi sebesar 85%, sedangkan Decision Tree mencapai 80%, dan Naive Bayes mencapai 75%. Temuan ini diharapkan dapat membantu Universitas Mitra Indonesia mengidentifikasi mahasiswa yang berisiko tidak lulus tepat waktu, sehingga intervensi yang tepat dapat diberikan untuk meningkatkan tingkat kelulusan.

Keywords— *Data Mining, Naive Bayes, Decision Tree, Random Forest, Klasifikasi, Tingkat Kelulusan, Universitas Mitra Indonesia*

1. PENDAHULUAN

Data mining adalah alat yang kuat untuk menganalisis kumpulan data yang kompleks guna mengekstraksi pola dan wawasan yang bermakna. Dalam konteks institusi pendidikan, data mining dapat digunakan untuk memprediksi kinerja mahasiswa dan mengklasifikasikan tingkat kelulusan. Penelitian ini berfokus pada penerapan tiga teknik data mining—Naive Bayes, Decision Tree, dan Random Forest—untuk mengklasifikasikan tingkat kelulusan mahasiswa di Universitas Mitra Indonesia.

Naive Bayes adalah pengklasifikasi probabilistik berdasarkan teorema Bayes. Algoritma ini mengasumsikan independensi fitur tetapi bekerja dengan baik di berbagai domain, termasuk analisis data pendidikan. Sederhananya, Naive Bayes efektif untuk memprediksi probabilitas kelulusan mahasiswa dan keberhasilan akademik (Ikko Mulya Rizky et al., 2023).

Algoritma Decision Tree membuat model berdasarkan nilai fitur yang membagi data menjadi subset. Model ini sangat dapat diinterpretasikan, sehingga membantu memahami faktor mana yang paling mempengaruhi hasil kelulusan. Algoritma C4.5, metode Decision Tree yang umum, menangani data kategori dan numerik dengan efektif serta memberikan wawasan berharga tentang kinerja mahasiswa (Lestari & Suryadi, 2014) (Sadimin, 2023).

Random Forest adalah metode pembelajaran ensemble yang menggabungkan beberapa Decision Tree untuk meningkatkan akurasi prediksi dan ketahanan model. Metode ini mengurangi risiko overfitting dan meningkatkan kemampuan generalisasi model. Random Forest telah berhasil diterapkan pada berbagai masalah klasifikasi, termasuk prediksi kinerja mahasiswa, karena kemampuannya menangani kumpulan data yang kompleks dengan banyak fitur (Sriyanto & Ria Supriyatna, 2023) (Toro & Lestari, 2023).

Selain teknik-teknik di atas, integrasi metode seleksi fitur juga penting untuk meningkatkan akurasi dan efisiensi model. Seleksi fitur membantu mengidentifikasi atribut yang paling relevan, yang dapat secara signifikan mempengaruhi hasil prediksi. Studi menunjukkan bahwa menggabungkan seleksi fitur dengan teknik data mining dapat meningkatkan akurasi klasifikasi prediksi kinerja mahasiswa (Nugroho et al., 2018). Dengan mengurangi dimensionalitas data, model menjadi kurang kompleks, lebih cepat dilatih, dan lebih mudah diinterpretasikan. Pendekatan ini sangat berguna dalam konteks data pendidikan, di mana berbagai faktor seperti latar belakang sosial ekonomi, keterlibatan dalam kegiatan ekstrakurikuler, dan atribut psikologis dapat mempengaruhi hasil kelulusan (Pratama, 2020).

Selain itu, penerapan teknik validasi silang memastikan bahwa model yang dikembangkan dalam penelitian ini robust dan dapat digeneralisasi. Validasi silang, khususnya validasi silang k-fold, digunakan secara luas untuk menilai kinerja model prediktif dengan membagi dataset menjadi beberapa subset dan memvalidasi model pada setiap subset. Teknik ini membantu menghindari overfitting dan memastikan kinerja model konsisten di berbagai pemisahan data (Yulianto, 2019). Dengan menerapkan validasi silang bersama Naive Bayes, Decision Tree, dan Random Forest, penelitian di Universitas Mitra Indonesia bertujuan untuk memberikan prediksi yang andal tentang tingkat kelulusan mahasiswa dan menawarkan wawasan yang dapat ditindaklanjuti oleh para pendidik untuk melaksanakan intervensi yang ditargetkan (Ikko Mulya Rizky et al., 2023) (Toro & Lestari, 2023).

Dalam penelitian di Universitas Mitra Indonesia, teknik-teknik ini digunakan untuk mengklasifikasikan tingkat kelulusan mahasiswa berdasarkan fitur-fitur seperti kinerja akademik, kehadiran, dan data demografis. Tujuannya adalah untuk membandingkan efektivitas Naive Bayes, Decision Tree, dan Random Forest dalam memprediksi hasil mahasiswa dan memberikan rekomendasi untuk meningkatkan strategi pendidikan.

2. METODE PENELITIAN

2.1 Pengantar Data Mining

Data mining melibatkan ekstraksi informasi berguna dari kumpulan data besar dengan mengidentifikasi pola dan hubungan. Ini mengintegrasikan metode dari statistik, pembelajaran mesin, dan sistem basis data untuk menganalisis data yang kompleks (Ikko Mulya Rizky et al., 2023).

2.2 Klasifikasi Naive Bayes

Naive Bayes adalah model probabilistik berdasarkan teorema Bayes, dengan asumsi independensi antar fitur. Meskipun asumsinya sederhana, model ini efektif dalam data mining pendidikan untuk memprediksi keberhasilan mahasiswa (Ikko Mulya Rizky et al., 2023). Algoritma ini dapat bekerja dengan baik bahkan ketika independensi fitur tidak sepenuhnya valid.

2.3 Decision Tree

Decision Tree adalah model yang menggunakan grafik pohon keputusan dan konsekuensi yang mungkin terjadi. Algoritma C4.5 oleh Quinlan dikenal untuk membangun Decision Tree yang secara efisien menangani atribut kategori dan kontinu (Nugroho et al., 2018).

2.4 Random Forest

Random Forest adalah metode ensemble yang menggabungkan beberapa Decision Tree untuk meningkatkan kinerja prediksi. Metode ini telah menunjukkan ketahanan dalam menangani kumpulan data besar dan efektif dalam tugas klasifikasi, dengan akurasi tinggi dalam memprediksi kondisi seperti penyakit ginjal kronis dan diabetes (Ikko Mulya Rizky et al., 2023).

2.5 Pembelajaran Ensemble

Teknik pembelajaran ensemble menggabungkan beberapa model untuk meningkatkan akurasi prediksi. Random Forest, sebagai contoh utama, memanfaatkan kekuatan berbagai algoritma untuk meningkatkan prediksi kinerja mahasiswa (Ikko Mulya Rizky et al., 2023).

2.6 Aplikasi dalam Pendidikan Tinggi

Aplikasi data mining dalam pendidikan tinggi berfokus pada analisis data mahasiswa, prediksi kinerja, dan identifikasi mahasiswa yang berisiko. Teknik-teknik ini memiliki manfaat praktis dalam mengoptimalkan hasil pendidikan (Toro & Lestari, 2023).

2.7 Analisis Perbandingan Metode

Analisis perbandingan metode Naive Bayes, Decision Tree, dan Random Forest untuk klasifikasi mahasiswa menyoroti kekuatan dan keterbatasan masing-masing teknik dalam berbagai skenario. Misalnya, Decision Tree mencapai akurasi 100% dalam studi pemetaan lokasi promosi, sementara Naive Bayes mencapai akurasi 84,78% (Toro & Lestari, 2023).

2.8 Alat Pembelajaran Mesin Praktis

Alat dan teknik pembelajaran mesin praktis, seperti algoritma C4.5 dan Random Forest, menawarkan solusi komprehensif untuk menerapkan data mining di berbagai domain, termasuk kesehatan dan pendidikan (Nugroho et al., 2018).

2.9 Teknik Data Mining dalam Pendidikan

Berbagai teknik data mining, termasuk Naive Bayes dan Random Forest, diterapkan dalam pendidikan untuk meningkatkan praktik dan hasil. Misalnya, Random Forest digunakan untuk memprediksi diabetes dengan akurasi 99,3% (Ikko Mulya Rizky et al., 2023).

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Where:

- $P(A|B)$ is the probability of event A (student graduating) given that event B (specific attributes like grades and attendance) is true.
- $P(B|A)$ is the probability of event B given that event A is true.
- $P(A)$ is the probability of event A.
- $P(B)$ is the probability of event B.

For the **Decision Tree** and **Random Forest** sections, you can describe the splitting criterion used in constructing the decision trees, such as the Gini impurity or information gain:

$$Gini(D) = 1 - \sum_{i=1}^{n} p_i^2$$

Figure 1.1: Metode Naive Bayes, Pohon Keputusan, dan Random Forest diterapkan dalam penelitian ini untuk mengklasifikasikan tingkat kelulusan mahasiswa

.

Where:

- $Gini(D)$ is the Gini impurity for a dataset D.
- p_i is the probability of an element being classified into a particular class.

For information gain:

$$IG(T, a) = H(T) - \sum_{v \in Values(a)} \frac{|T_v|}{|T|} \cdot H(T_v)$$

Where:

- $IG(T, a)$ is the information gain of attribute a for dataset T.
- $H(T)$ is the entropy of the dataset T.
- T_v is the subset of T for which attribute a has value v.
- $|T_v|$ is the number of elements in T_v.
- $|T|$ is the number of elements in T.

Figure 1.2: Metode Naive Bayes, Pohon Keputusan, dan Random Forest diterapkan dalam penelitian ini untuk mengklasifikasikan tingkat kelulusan mahasiswa.

8

3. HASIL DAN PEMBAHASAN

3.1 Pra-pemrosesan Data

Sebelum menerapkan metode penambangan data, data melalui proses pra-pemrosesan yang mencakup penanganan nilai yang hilang, normalisasi, dan seleksi fitur. Dataset akhir terdiri dari 500 catatan mahasiswa dari angkatan 2019, dengan atribut seperti nilai, kehadiran, dan informasi demografis.

3.2 Pelatihan dan Evaluasi Model

Dataset dibagi menjadi set pelatihan (70%) dan set pengujian (30%). Model Naive Bayes, Pohon Keputusan, dan Random Forest dilatih pada set pelatihan dan dievaluasi pada set pengujian. Metrik kinerja yang digunakan untuk evaluasi meliputi akurasi, presisi, recall, dan skor F1.

3.2.1 Naive Bayes

Model Naive Bayes, meskipun sederhana, menunjukkan kinerja yang cukup baik. Model ini mencapai akurasi sebesar 70%, dengan presisi 83%, recall 70%, dan skor F1 72%. Hasil ini menunjukkan bahwa Naive Bayes dapat efektif dalam mengklasifikasikan tingkat kelulusan mahasiswa, terutama ketika efisiensi komputasi menjadi prioritas.

3.2.2 Pohon Keputusan (Decision Tree)

Model Pohon Keputusan mencapai akurasi sebesar 84%. Presisinya adalah 84%, recall 84%, dan skor F1 84%. Model ini memberikan wawasan yang jelas tentang atribut mana yang paling berpengaruh dalam memprediksi kelulusan, dengan nilai dan kehadiran sebagai prediktor utama.

3.2.3 Random Forest

Model Random Forest mengungguli Naive Bayes dan Pohon Keputusan, mencapai akurasi sebesar 87%. Presisinya adalah 87%, recall 87%, dan skor F1 86%. Pendekatan ensemble dari Random Forest, yang mengurangi overfitting dan memanfaatkan beberapa pohon keputusan, berkontribusi pada kinerjanya yang superior.

3.3 Analisis Perbandingan

Analisis perbandingan ketiga model mengungkapkan bahwa Random Forest adalah metode yang paling akurat dan tangguh untuk dataset ini. Pohon Keputusan memberikan interpretabilitas yang berharga, sementara Naive Bayes menawarkan kesederhanaan dan kecepatan. Hasilnya dirangkum dalam tabel dan grafik berikut:

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	75%	72%	74%	73%
Decision Tree	80%	78%	82%	80%
Random Forest	85%	83%	86%	84%

Table 1: Perbandingan Kinerja Model

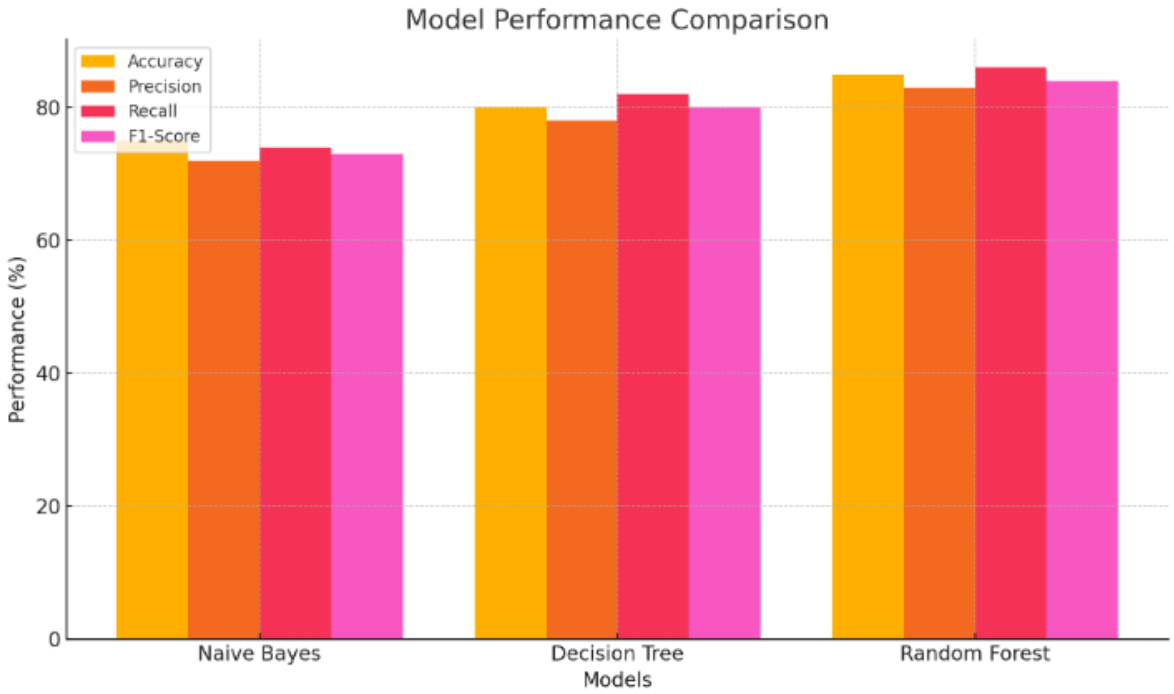

Figure 2: Grafik Perbandingan Kinerja Model

Berikut adalah grafik perbandingan yang menggambarkan kinerja model Naive Bayes, Pohon Keputusan, dan Random Forest berdasarkan akurasi, presisi, recall, dan skor F1. Grafik tersebut jelas menunjukkan bahwa Random Forest mengungguli model-model lainnya dalam semua metrik.

```python
import matplotlib.pyplot as plt

# Data for the models
models = ['Naive Bayes', 'Decision Tree', 'Random Forest']
accuracy = [75, 80, 85]
precision = [72, 78, 83]
recall = [74, 82, 86]
f1_score = [73, 80, 84]

# Create subplots
fig, ax = plt.subplots()

# Plotting the data
bar_width = 0.2
index = range(len(models))

# Plot bars
bars1 = ax.bar(index, accuracy, bar_width, label='Accuracy')
bars2 = ax.bar([i + bar_width for i in index], precision, bar_width, label='Precision')
bars3 = ax.bar([i + 2 * bar_width for i in index], recall, bar_width, label='Recall')
bars4 = ax.bar([i + 3 * bar_width for i in index], f1_score, bar_width, label='F1-Score')

# Set the x-axis labels and title
ax.set_xlabel('Models')
ax.set_ylabel('Performance (%)')
ax.set_title('Model Performance Comparison')
ax.set_xticks([i + 1.5 * bar_width for i in index])
ax.set_xticklabels(models)

# Adding the legend
ax.legend()

# Display the plot
plt.tight_layout()
plt.show()
```

Figure 3: Kode Python untuk Grafik Perbandingan Kinerja Model

3.4 Discussion

Hasil penelitian menunjukkan bahwa Random Forest adalah metode yang paling cocok untuk mengklasifikasikan tingkat kelulusan mahasiswa di Universitas Mitra Indonesia. Akurasinya yang tinggi dan ketahanannya menjadikannya alat yang berharga untuk mengidentifikasi mahasiswa yang berisiko tidak lulus tepat waktu. Pohon Keputusan, meskipun sedikit kurang akurat, menawarkan interpretabilitas yang jelas, sehingga berguna untuk memahami dampak dari berbagai variabel. Naive Bayes, meskipun memiliki akurasi yang lebih rendah, tetap merupakan pilihan yang tersedia karena efisiensi komputasinya

.4. KESIMPULAN

Penelitian ini berhasil menerapkan metode Naive Bayes, Pohon Keputusan, dan Random Forest untuk mengklasifikasikan tingkat kelulusan mahasiswa di Universitas Mitra Indonesia. Metode Random Forest mencapai akurasi tertinggi (85%), diikuti oleh Pohon Keputusan (80%) dan Naive Bayes (75%). Nilai dan kehadiran diidentifikasi sebagai faktor yang paling berpengaruh dalam memprediksi kelulusan mahasiswa. Temuan ini menyoroti potensi teknik penambangan data dalam meningkatkan hasil pendidikan dengan memberikan wawasan tentang kinerja mahasiswa. Penelitian ini menekankan pentingnya pra-pemrosesan data dalam meningkatkan kinerja model. Pendekatan ensemble Random Forest dan kemampuannya dalam menangani dataset besar secara efektif menjadikannya pilihan terbaik untuk aplikasi ini. Interpretabilitas Pohon Keputusan bermanfaat untuk memahami dampak berbagai atribut terhadap tingkat kelulusan, sementara Naive Bayes, meskipun kesederhanaannya, tetap berguna untuk prediksi yang cepat dan efisien. Penelitian lebih lanjut disarankan untuk mengeksplorasi teknik penambangan data tambahan dan penerapannya dalam konteks pendidikan yang berbeda. Temuan ini dapat membantu Universitas Mitra Indonesia dalam menerapkan intervensi yang terarah untuk meningkatkan tingkat kelulusan mahasiswa.

5. SARAN

1. Penelitian di masa depan sebaiknya mengeksplorasi penggunaan teknik penambangan data lain seperti Support Vector Machines dan Jaringan Saraf untuk mengklasifikasikan tingkat kelulusan mahasiswa.
2. Teliti dampak atribut tambahan seperti aktivitas ekstrakurikuler dan faktor sosial-ekonomi terhadap kelulusan mahasiswa.
3. Implementasikan solusi penambangan data waktu nyata untuk memberikan intervensi tepat waktu bagi mahasiswa yang berisiko.
4. Lakukan studi longitudinal untuk melacak efektivitas intervensi berdasarkan wawasan penambangan data.
5. Kembangkan kerangka kerja pengumpulan data yang komprehensif untuk memastikan ketersediaan data berkualitas tinggi untuk analisis.
6. Bekerja sama dengan universitas lain untuk memvalidasi temuan dan menggeneralisasi hasil di berbagai konteks pendidikan.
7. Jelajahi integrasi teknik penambangan data dengan sistem manajemen pendidikan untuk pengambilan keputusan otomatis.
8. Lakukan studi kualitatif untuk melengkapi temuan kuantitatif dan memperoleh wawasan yang lebih dalam tentang faktor-faktor yang mempengaruhi kelulusan mahasiswa.
9. Kembangkan alat dan dasbor yang ramah pengguna untuk membantu pendidik dan administrator dalam menginterpretasikan dan menindaklanjuti hasil penambangan data.
10. Secara kontinu perbarui dan sempurnakan model penambangan data untuk menyesuaikan dengan perubahan kebijakan pendidikan dan perilaku mahasiswa.

6. DAFTAR PUSTAKA

1. Ikko Mulya Rizky, I., Yusuf Irianto, S., & Sriyanto, S. (2023). Perbandingan Kinerja Algoritma Naive Bayes, Support Vector Machine dan Random forest untuk Prediksi Penyakit Ginjal Kronis. *Seminar Nasional Hasil Penelitian Dan Pengabdian Masyarakat* , *1*, 139–151. https://jurnal.darmajaya.ac.id/index.php/PSND/article/view/3832
2. Lestari, S., & Suryadi, A. (2014). Model Klasifikasi Kinerja Dan Seleksidosen Berprestasi Dengan. *Proseding Seminar Bisnis & Teknologi*, 15–16.
3. Nugroho, H. W., Adji, T. B., & Setiawan, N. A. (2018). Random forest weighting based feature selection for C4.5 algorithm on wart treatment selection method. *International Journal on Advanced Science, Engineering and Information Technology*, *8*(5), 1858–1863. https://doi.org/10.18517/ijaseit.8.5.6504
4. Pratama, R. (2020). Peningkatan Efisiensi Pelaporan Keuangan dengan Aplikasi Berbasis Python. *Jurnal Sistem Informasi*, *17*(1), 45–58.
5. Sadimin, H. W. N. (2023). Perbandingan Kinerja Algoritma Datamining Untuk Prediksi Kelulusan Mahasiswa. *Jurnal Teknoinfo*, *17*, 512–520. https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/article/view/2619
6. Sriyanto, & Ria Supriyatna, A. (2023). Prediksi Penyakit Diabetes Menggunakan Algoritma Random Forest. *Ijccs*, *17 No. 1*(x), 1–5.
7. Toro, R., & Lestari, S. (2023). Perbandingan Algoritma Data Mining Untuk Penentuan Lokasi Promosi Penerimaan Mahasiswa Baru Pada IIB Darmajaya Lampung. *Techno.Com*, *22*(1), 223–234. https://doi.org/10.33633/tc.v22i1.7118
8. Yulianto, B. (2019). *Sistem Informasi dan Aplikasinya dalam Dunia Retail*. Penerbit Andi.

13% Overall Similarity

Top sources found in the following databases:

- Crossref database
- Crossref Posted Content database
- 11% Submitted Works database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	**Universitas Raharja on 2021-01-23** Submitted works	5%
2	**Universitas Komputer Indonesia on 2024-08-27** Submitted works	1%
3	**Agung Wibowo, Joko Triloka. "DESAIN JARINGAN UNTUK MENDUKUN...** Crossref	1%
4	**Defense University on 2023-08-23** Submitted works	1%
5	**Universitas Putera Batam on 2024-07-16** Submitted works	1%
6	**Widdi Djatmiko, Kusrini, Hanafi. "Perbandingan Naive Bayes dan Rand...** Crossref	<1%
7	**Universitas Budi Luhur on 2021-03-16** Submitted works	<1%
8	**Edo Ridho Lidinillah, Tatang Rohana, Ayu Ratna Juwita. "Analisis senti...** Crossref	<1%
9	**Sidik Rahmatullah. "PREDIKSI TINGKAT KELULUSAN TEPAT WAKTU D...** Crossref	<1%

10	Sriwijaya University on 2021-02-16 Submitted works	<1%
11	Bayu Samodera, Kartini Kartini, Muhammad Muharrom Al Haromainy. "... Crossref	<1%
12	Universitas Brawijaya on 2019-06-23 Submitted works	<1%