Terbit online pada laman web jurnal: **http://journal.isas.or.id/index.php/JACOST**

JOURNAL OF APPLIED COMPUTER SCIENCE AND TECHNOLOGY (JACOST)

Vol. 5 No. 1 (2024) 109 – 119 | ISSN: 2723-1453 (Media Online)

Klasifikasi Metode Data Mining untuk Prediksi Kelulusan Tepat Waktu Mahasiswa dengan Algoritma Naïve Bayes, Random Forest, Support Vector Machine (SVM) dan Artificial Neural Nerwork (ANN)

Satrio Junaidi[1], Rani Valicia Anggela[2], Delsi Kariman[3]

[1,2,3]Sains Data, Fakultas Sains dan Teknologi, Universitas PGRI Sumatera Barat, Padang, Indonesia

[1]satriojunaidi@upgrisba.ac.id, [2]ranivalicia@gmail.com, [3]delsik@upgrisba.ac.id

Abstract

Timely graduation of students is essential for determining the quality of college. Universities must know the percentage of students' ability to complete their studies on time. So, to deal with this problem, data mining classification is carried out to predict student graduation on time to find patterns for student on-time graduation predictions. This research can yield new information to help colleges anticipate student graduations that are not on time. The method used is a classification data mining method with 4 algorithms: naïve Bayes, random forest, support vector machine (SVM), and artificial neural network (ANN). The attributes used are gender, parental income, length of guidance, working student status or not, semester 1 to semester 8 grades, and GPA. This study used Python 3 programming language on jupyter notebooks in Anaconda to process datasets. The distribution of datasets is divided by 70% for training data and 30% for testing data. The results of this study were obtained with the best algorithm accuracy in the support vector machine (SVM) algorithm is 0.94. Based on the results of this study, the accuracy is good for predicting student graduation on time.

Keywords: data classification mining, naïve Bayes, random forest, SVM, ANN

Abstrak

Kelulusan mahasiswa tepat waktu sangat penting untuk yang menentukan kualitas perguruan tinggi. Perguruan tinggi harus mengetahui persentasi kemampuan mahasiswa untuk menyelesaikan studi tepat waktu. Maka, untuk menghadapi masalah tersebut dilakukan klasifikasi data mining untuk memprediksi kelulusan tepat waktu mahasiswa dengan tujuan mencari pola untuk prediksi kelulusan tepat waktu mahasiswa. Penelitian ini dapat menghasilkan informasi baru membantu perguruan tinggi mengantisipasi kelulusan mahasiswa yang tidak tepat waktu. Metode yang digunakan adalah metode data mining klasifikasi dengan 4 algoritma yaitu *naïve bayes*, *random forest*, *support vector machine* (SVM) dan *artificial neural network* (ANN). Atribut yang digunakan adalah jenis kelamin, penghasilan orang tua, lama bimbingan, status mahasiswa bekerja atau tidak, nilai semester 1 sampai semester 8, dan IPK. Penelitian ini menggunakan bahasa pemrograman python 3 pada jupyter notebook di anaconda untuk memproses dataset. Distribusi dataset dibagi 70% untuk data training dan 30% untuk data testing. Hasil penelitian ini didapatkan dengan akurasi algoritma terbaik adalah algoritma *support vector machine* (SVM) adalah 0.94. Berdasarkan hasil penelitian ini akurasi tersebut baik untuk memprediksi kelulusan tepat waktu mahasiswa.

Kata kunci: klasifikasi data mining, *naïve bayes*, random forest, SVM, ANN

1. Pendahuluan

Lulus tepat waktu adalah keinginan seluruh mahasiswa, tidak hanya itu lulus tepat waktu adalah keuntungan bagi kedua belah pihak, yaitu mahasiswa dan instansi pendidikan [1]. Kelulusan mahasiswa bisa diprediksi menggunakan suatu sistem. Namun, beberapa perguruan tinggi belum memiliki sistem untuk memprediksi keterlambatan kelulusan mahasiswa, sehingga perguruan tinggi belum bisa melakukan pencegahan akan hal tersebut. Lulus tepat waktu merupakan hal yang sangat penting karena tingkat kelulusan sebagai dasar efektifnya suatu fakultas di suatu perguruan tinggi. Jika terjadi penurunan tingkat kelulusan maka akan menjadi suatu permasalahan yang akan mempengaruhi akreditasi di sebuah perguruan tinggi. Dalam mengatasi permasalahan tersebut Perguruan Tinggi perlu mengantisipasi dampak buruk yang akan terjadi pada kelulusan mahasiswa nantinya. Banyak faktor yang memengaruhi tidak lulus tepat waktunya mahasiswa selain IPK yaitu lama bimbingan, penghasilan orang tua, status mahasiswa bekerja atau tidak, nilai setiap semester dan IPK. Dataset didapatkan dari admin setiap program studi yang ada di fakultas sains dan teknologi, data yang diminta adalah nama mahasiswa, NPM, IP semester 1 sampai semester akan di wisuda dan IPK. Sedangkan data penghasilan orang tua, status mahasiswa bekerja atau tidak selama kuliah, lama bimbingan

didapatkan dari mahasiswa bersangkutan dengan cara menyebarkan angket mengisi link form drive yang dibagikan.

Adapun beberapa kajian terdahulu terkait data mining menggunakan metode naïve bayes pada penelitian yang berjudul Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode *Naive Bayes* di Program Studi Teknik Informatika UHAMKA, Model dengan hasil terbaik yaitu model ke-3 dengan tingkat akurasi sebesar 80.19%, recall 80.26%, precision 92.75% dan F-Measure 86.05% yang nantinya akan digunakan untuk implementasi pada aplikasi prediksi kelulusan mahasiswa [2]. Penelitian yang berjudul Optimasi Model Prediksi Kelulusan Mahasiswa Menggunakan Algoritma Naïve Bayes, hasil dari penelitian ini mendapatkan nilai akurasi untuk metode naïve bayes sebesar 70% dan akurasi untuk model prediksi dengan parameter sosial sebesar 85% dengan selisih akurasi 10% [3].

Penelitian yang berjudul Prediksi Kelulusan Mahasiswa menggunakan Algoritma *Naive Bayes* (Studi Kasus 5 PTS di Banda Aceh), Hasil penelitian didapatkan bahwa algoritma data mining untuk prediksi kelulusan berdasarkan atribut ketepatan lulus yang dipilih mengungkapkan bahwa tingkat prediksi seragam dengan algoritma yang digunakan yaitu Naïve Bayes, akurasi prediksi sebesar 84%. Atribut data yang ditemukan memiliki signifikan dipengaruhi proses klasifikasi adalah IPK dan Lama Studi. Hasil yang diperoleh bahwa mahasiswa yang lulus sebesar 60% yaitu mahasiswa yang berpendidikan di ASM Nusantara dan AMIK Indonesia, sedangkan pada STIES Banda Aceh dan Universitas Serambi Mekkah sebesar prediksi lulus sebesar 52%. Hal lain berbeda dengan STIA Iskandar Thani dimana prediksi lulus hanya sebesar 48% dan tidak lulus tepat waktu 52%. Hasil prediksi ini bisa mengungkapkan dan menjadi sebuah rekomendasi bagi calon mahasiswa atau pihak akademik agar meningkatkan kuantitas lulusan dan meningkatkan kepercayaan mahasiswa terhadap Perguruan Tinggi [4].

Penelitian yang berjudul Prediksi Kelulusan Mahasiswa Fakultas Teknik Universitas Bina Darma Menggunakan Algoritma *Naïve Bayes*, hasil akurasi penelitian dengan menggunakan perhitungan *confusion matrix multiclass* bahwa nilai akurasi dari hasil prediksi menggunakan algoritma *naïve bayes* yaitu 78 mahasiswa yang akan lulus di semester 8 dengan akurasi sebesar 98%, lulus di semester 9 sebanyak 24 mahasiswa dengan akurasi 96%, lulus di semester 10 sebanyak 3 mahasiswa dengan akurasi 100%, dan lulus disemester 12 sebanyak 15 mahasiswa dengan akurasi 98%, Kemudian dari nilai akurasi secara keseluruhan untuk prediksi kelulusan mahasiswa menggunakan algoritma *naïve bayes* sebesar 95,33%. Berdasarkan hasil dari penelitian menggunakan algoritma *naïve bayes* maka akurasi tersebut sudah cukup untuk menentukan prediksi kelulusan mahasiswa [5].

Penelitian yang berjudul Sistem Prediksi Lama Studi Kuliah Menggunakan Metode Naive Bayes, Hasil Penelitian dalam Uji sistem telah memiliki tingkat akurasi 80%, recall 50% dan presisi 100%.[5]

Penelitian yang berjudul Algoritma Decision Tree C.45 dalam Analisa kelulusan tepat waktu mahasiswa Program Studi Manajemen Informatika UMPP, Dari analisa yang dilakukan didapatkan bahwa tahun masuk mahasiswa menjadi faktor variabel utama dalam kelulusan mahasiswa dengan didapatkan akurasi algoritma decision tree sebesar 73,48%. Dengan Tahun mahasiswa menjadi point utama dikarenakan kondisi pemenuhan SDM dan fasilitas dalam menangani mahasiswa untuk segera menyelesaikan Tugas Akhirnya [6].

Penelitian yang berjudul Implementasi Data Mining Menggunakan Metode *Naive Bayes* Dengan Feature Selection Untuk Prediksi Kelulusan Mahasiswa Tepat Waktu, hasil penelitian ini menghasilkan tingkat akurasi 81% dengan presisi sebesar 83,563% dan recall 88,41%. Metode yang digunakan termasuk dalam Klasifikasi Baik dan akan menjadi acuan pihak manajemen perguruan tinggi, untuk mengatasi masalah yang mungkin timbul dalam penurunan kualitas pendidikan (misalnya penurunan rasio dosen dengan mahasiswa) [7].

Penelitian yang berjudul Metode Data Mining *Assosiation Rule* dengan Algoritma *FP-Growth* untuk Mengetahui Kelulusan tepat Waktu Mahasiswa, Data yang diolah adalah data akademik mahasiswa yang dilihat dari IPK dan Total SKS, kemudian untuk mengetahui jumlah kelulusan mahasiwa tepat waktu menggunakan metode Data Mining *Association Rule* dengan algoritma *FP-Growth* dan juga menggunakan aplikasi WEKA (Weikato Environment Knowledge and Analiysis). Hasil didapatkan bahwa *rule* terbaik dijadikan patokan untuk mengetahui mahasiswa lulus tepat waktu dengan *support* 14,8 % dan confidance 93,7 % . hasil support dan *confidance* terendah diprediksi mahasiswa tidak lulus tepat waktu [8]

Penelitian yang berjudul Perbandingan Metode Data Mining Untuk Prediksi Nilai Dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4,5, Naïve Bayes, KNN Dan SVM dengan hasil diperoleh bahwa algoritma Naïve Bayes merupakan algoritma terbaik untuk memprediksi kelulusan mahasiswa yang tepat waktu dan IPK $\geqslant$ 3 dengan nilai accuracy (76,79%), error (23,17%) , dan AUC (0,850) [9]. Penelitian yang berjudul Pola Prediksi Kelulusan Siswa Madrasah Aliyah Swasta dengan Support Vector Machine dan Random Forest dengan hasil algoritma SVM dan RF sangat baik dalam memprediksi kelulusan siswa MAS terlihat dari akurasi yang sangat tinggi yaitu SVM(98,98%) dan RF(99,49%) [10]. Penelitian berjudul Algoritma Support Vector Machine Untuk Memprediksi Nilai ujian Nasional dengan hasil Root

DOI : https://doi.org/10.52158/jacost.v5i1.489

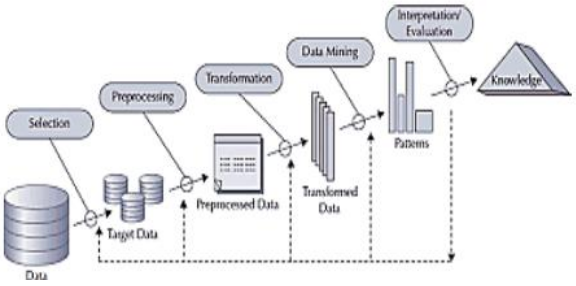
mean squared Error (RMSE) terbaik untuk Bahasa Indonesia adalah 0.713 +/- 0.173, Bahasa Inggris sebesar 0.586 +/- 0.066, dan Matematika sebesar 0.882 +/- 0.188 [11]. Penelitian berjudul Klasifikasi Penyakit Daun Pada Tanaman Jagung Menggunakan Algoritma Support Vector Machine, K-Nearest Neighbors dan Multilayer Perceptron, Hasil yang didapatkan menunjukan bahwa algoritma Multilayer Perceptron menghasilkan nilai terbaik dengan accuracy, precision dan recall masing-masing sebesar 97.4% [12]. Analisis Perbandingan Metode Regresi Linier, Random Forest Regressiondan Gradient Boosted Trees Regression Method untuk Prediksi Harga Rumah telah didapat hasil perbandingan prediksi dengan error rate lebih kecil pada metode Random Forest Regression. Evaluasi prediksi dilakukan dengan melihat hasil error pada RMSE setiap model, yaitu pada Random Forest Regression didapat nilai error 0.440, Linear Regression didapat nilai errorterkecil di range 0.515 dan pada Gradient Boosted Trees Regression sebesar 0.508 [13].

Berdasarkan latar belakang tersebut penelitian ini bertujuan untuk memanfaatkan klasifikasi data mining untuk memprediksi kelulusan tepat waktu mahasiswa dengan atribut yang digunakan adalah jenis kelamin, penghasilan orang tua, lama bimbingan, status mahasiswa bekerja atau tidak, nilai semester 1 sampai semester 8, dan IPK menggunakan algoritma yaitu *naïve bayes*, *random forest*, *support vector machine* (SVM) dan *artificial neural network* (ANN) dengan bahasa pemrograman python dengan jupyter notebook pada ananconda. Selanjutnya penelitian ini bertujuan untuk membandingkan dari keempat algoritma yang digunakan yang memberikan hasil paling baik atau hasil yang paling mendekati hasil sebenarnya.

2. Metode Penelitian

Metode pengumpulan data dilakukan menggunakan data primer karena data yang digunakan merupakan data yang diperoleh secara langsung dari Staff UPT TI Universitas PGRI Sumatera Barat, admin masing-masing prodi serta mahasiswa masing-masing program studi yang ada di fakultas sains dan teknologi Universitas PGRI Sumatera Barat peneliti selain itu juga menggunakan studi pustaka dan literatur yang berkaitan dengan tema penelitian. Data-data yang sudah dikumpulkan kemudian disimpan dalam bentuk excel agar dapat diolah menggunakan data mining.

Data Mining merupakan sebuah proses ekstraksi dari data yang banyak untuk dijadikan pengetahuan dan informasi yang dibutuhkan. Data mining melibatkan banyak data dan dibantu oleh Teknologi *Artificial Intelligence*, Statistik, matematika dan s learning. Selain itu data mining dapat membantu dalam mengambil keputusan berdasarkan data yang telah diproses menjadi informasi. Data mining sering juga disebut dengan *Knowledge Discovery in Database* atau disingkat KDD, tahappan KDD dapat di lihat pada Gambar 1[14].



Gambar 1. Tahapan KDD

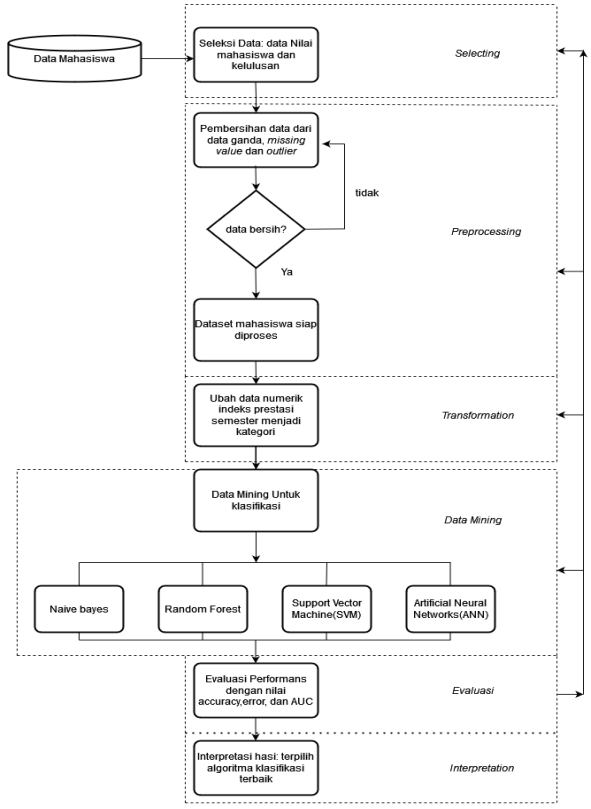
Berikut penjelasan tiap tahap penelitian pada Gambar 1.

2.1. Selection

Pada tahap ini dilakuakan seleksi data lulusan yang terdiri dari variabel-variabel prediktor dan satu target variabel. Variabel target yaitu klasifikasi lulusan yang lulus tepat waktu yaitu 4 tahun atau kurang dan memiliki nilai IPK minimal 3,00 sedangkan variabel-variabel prediktor yaitu, jenis kelamin, indeks prestasi semester 1 sampai semester 8, penghasilan orang tua, status mahasiswa bekerja atau tidak.

2.2. Preprocessing

Data yang diambil sesuai dengan banyaknya lulusan di Fakultas sains dan teknologi. Dari data yang diambil dilakukan pembersihan data apabila terdapat data yang hilang, data ganda atau bersifat outlier. Berikut pada Gambar 2 tahapan penelitian berdasarkan KDD.



Gambar 2. Tahapan Penelitian berdasakan KDD

DOI : https://doi.org/10.52158/jacost.v5i1.489

2.3. Transformation

Setelah data bersih dari kesalahan, selanjutnya dilakukan transformasi pada data sesuai dengan jenis data pada tahapan transformasi dimana jenis akan dikelompokkan menjadi data yang bersifat kategori seperti Tabel 1.

Tabel 1. Atribut Data Mahasiswa

Atribut	Tipe	Keterangan
Jenis Kelamin	Teks	Mahasiswa Perempuan atau Laki--Laki
Penghasilan Orang Tua	Numerik	Kategori Rendah dan Tinggi
Lama Bimbingan	Numerik	Hitungan dalam Bulan
Status Mahasiswa	Teks	Mahasiswa/Bekerja
IPS 1	Numerik	IP Semester 1
IPS 2	Numerik	IP Semester 2
IPS 3	Numerik	IP Semester 3
IPS 4	Numerik	IP Semester 4
IPS 5	Numerik	IP Semester 5
IPS 6	Numerik	IP Semester 6
IPS 7	Numerik	IP Semester 7
IPS 8	Numerik	IP Semester 8
IPK	Numerik	Indek Prestasi Komulatif
Keterangan	Teks	Terlambat/Tepat Waktu

2.4. Data Mining

Pada tahap ini dilakukan pemilihan teknik data mining yang sesuai, untuk fungsi klasifikasi digunakan algoritma diantaranya *naïve bayes*, *Random Forest*, *Support Vector Machine* (SVM), atau *Artificial Neural Networks* (ANN). Klasifikasi adalah metode *supervised learning* yang mengkategorikan informasi ke dalam kelompok yang telah ditentukan. Tujuannya untuk membuat model yang dapat mengkategorikan populasi data yang besar[15].

2.4.1. Naïve Bayes

Naïve bayes merupakan metode *machine learning* yang dapat digunakan untuk membedakan objek yang berbeda berdasarkan fitur tertentu. Algoritma ini mudah dibuat dan sangat berguna pada dataset yang besar. *Naïve bayes* merupakan teknik prediksi berbasis *probabilistic* sederhana yang berdasar pada penerapan teorema bayes (aturan bayes) dengan asumsi independensi (ketidaktergantungan) yang kuat (naif). Dengan kata lain, dalam *naïve bayes* model yang digunakan adalah model fitur independent.

$$P(X|H) = \frac{P(X|H)P(H)}{P(H)} \quad \text{..... (1)}$$

Penjelasan dari persamaan (1) adalah sebagai berikut: P(H|E): Probabilitas akhir bersyarat (conditional probability) suatu hipotesis H terjadi jika diberikan bukti (evidence) E terjadi. P(E|H): Probabilitas sebuah bukti E terjadi tanpa memandang hipotesis Atau bukti yang lain

2.4.2. *Random Forest*

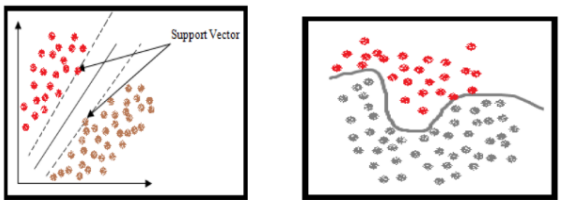
Random Forest (RF) adalah pengklasifikasi yang membawa konsep pohon keputusan lebih jauh dengan menghasilkan sejumlah besar pohon keputusan. Pendekatan ini pertama-tama mengambil sampel data secara acak dan mengidentifikasi serangkaian fitur membangunkan setiap pohon keputusan. Kemudian pohon keputusan ini ditentukan kesalahannya (tingkat kesalahan model) dan kemudian membandingkan kumpulan pohon keputusan untuk menemukan kumpulan variabel gabungan yang menghasilkan metode klasifikasi terkuat, di antara algoritma pengklasifikasi saat ini, RF memiliki akurasi yang sangat baik. Untuk memperkirakan data yang hilang, RF memiliki metode yang efektif dan ketika sebagian besar data hilang, akurasinya tetap terjaga.

2.4.3. *Support Vector Machine* (SVM)

Support vector Machine adalah jenis metode klasifikasi *supervised*. Algoritma SVM melibatkan dua jenis versi seperti versi linier dan non linier. Versi pertama yaitu versi linier, hyperplanes atau himpunan hyperplanes digunakan untuk pemisahan kelas. Sebuah hyperplane diwakili oleh Persamaan 2.

$$WX + b = 0 \quad \text{..... (2)}$$

Pada versi kedua yaitu versi non linier, kelas bukanlah partisi sehingga tidak ada garis lurus yang memisahkan kelas. Dengan bantuan vektor dan margin, SVM menemukan hyperplane. Untuk teks klasifikasi, SVh5M adalah metode yang paling akurat. Metode ini juga banyak digunakan dalam klasifikasi untuk analisis sentimen. SVM dapat digunakan untuk mempelajari pengklasifikasi radial basis functional (RBF), polinomial, dan multi-layer perceptron (MLP) seperti Gambar 3.



Gambar 3. Linear dan Non Linear SVM

2.4.4. *Artificial Neural Networks* (ANN).

Istilah jaringan saraf adalah sirkuit yang digunakan untuk menyusun sejumlah besar elemen pemrosesan yang disebut Neuron. Oleh karena itu, jaringan saraf sangatlah kompleks. Ide utama dari jaringan saraf adalah untuk memperoleh atribut dari kombinasi linier dari data masukan, dan kemudian memodelkan keluaran sebagai fungsi non linier dari atribut tersebut. Hasilnya adalah salah satu bentuk sistem pembelajaran yang paling populer dan efektif. Jaringan saraf diwakili oleh diagram jaringan yang terdiri dari node yang dihubungkan oleh tautan terarah. Node diselaraskan dalam lapisan dan struktur jaringan saraf yang paling banyak digunakan. melibatkan 3 lapisan: lapisan masukan, lapisan tersembunyi dan lapisan keluaran. ANN adalah metode yang paling banyak diterapkan dalam meramalkan konsumsi energi gedung.

DOI : https://doi.org/10.52158/jacost.v5i1.489

2.5. Evaluation

Tahap ini digunkan untuk evaluasi hasil-hasil prediksi yang dihasilnya oleh keempat algoritma dan dipilih metode algoritma yang menghasilkan nilai mendekati klasifikasi data sebenarnya. Evaluasi dilakukan dengan menggunakan metode Confusion Matrix dan kurva ROC (Receiver Operating Characteristic).

3. Hasil dan Pembahasan

Data Selection

Sumber data mentah yang digunakan dalam penelitian ini adalah data mahasiswa fakultas sains dan teknologi universitas PGRI Sumatera Barat wisuda tahun 2023. Data mahasiswa terdiri dari nama, jenis kelamin, nilai dan status kelulusan di dapat dari Kepala UPT TI, admin masing-masing program studi. Sedangkan data status mehasiswa bekerja atau tidak selama kuliha, lama bimbingan, dan penghasilan orang tua di dapat dari mahasiswa tersebut.

Preprocessing Data

Berikut data mentah mahasiswa di Fakultas Sains dan Teknologi Universitas PGRI Sumatera Barat. Dataset berisikan atribut data mentah yang di dapat dari Fakultas Sains dan Teknologi. Data Mahasiswa yang diperoleh seperti Jenis Kelamin (JK), Gaji Orang Tua (G), Status Mahasiswa sedang bekerja atau tidak (SM), Lama Bimbingan (LB), Indek Prestasi (IP) semseter 1 sampai 8, IPK dan Status mahasiswa Terlambat (T) atau Tepat Waktu (TW). Untuk lebih jelas seperti tabel 2 di bawah ini.

Tabel 2. Atribut data mentah mahasiswa

JK	G	SM	LB	1	2	3	4	5	6	7	8	IPK	S
P	<2,5	B	3	3,2	3	2,8	2,9	2,9	3,2	0	4	2,7	T
P	<2,5	B	8	3,3	3	3	3,7	3,5	3,7	3,7	0	2,9	T
…	…..	….	….	….	….	….	….	….	….	….	….	….	….
P	<2,5	M	2	3,6	3,6	3,8	3,8	2,9	3,8	3,7	0	3,6	TW
L	<2,5	B	7	3,2	3	2,6	3,6	3,1	3,3	2,8	3	3,07	

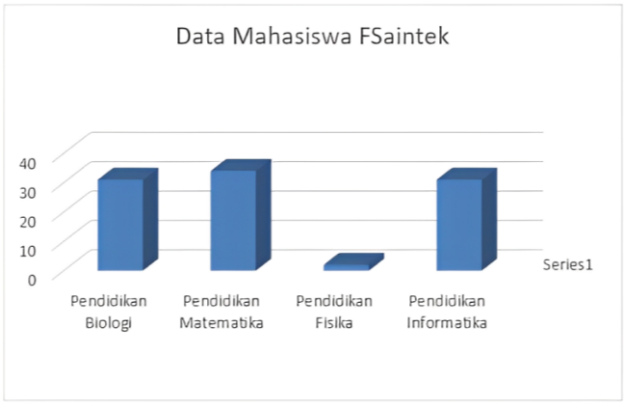
Transformation

Selanjutnya data tersebut di jadikan data numerik, data yang akan di olah seperti terlihat pada Tabel 3.

Tabel 3. Data numerik mahasiswa

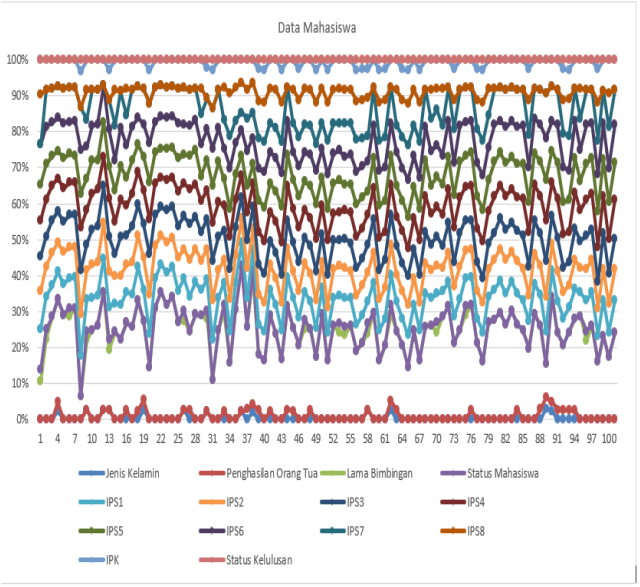
JK	G	SM	LB	1	2	3	4	5	6	7	8	IPK	S
0	0	1	3	3,2	3	2,8	2,9	2,9	3,2	0	4	2,7	0
0	0	1	8	3,3	3	3	3,7	3,5	3,7	3,7	0	2,9	0
…	…..	….	….	….	….	….	….	….	….	….	….	….	….
0	0	0	2	3,6	3,6	3,8	3,8	2,9	3,8	3,7	0	3,6	1
0	0	0	7	3,2	3	2,6	3,6	3,1	3,3	2,8	3	3,07	0

Berikut merupakan grafik distribusi data berdasarkan attribut program studi di fakultas sains dan teknologi Universitas PGRI Sumatera Barat dan jenis kelas pada tiap mahasiswa, persebaran data terdiri dari data sebelum di proses cleaning kemudian data setelah proses cleaning data, pada Gambar 4



Gambar 4. Grafik persebaran data

Selanjutnya dapat kita lihat grafik data setelah dilakukan proses cleaning dan transformasi data dapat dilihat pada Gambar 5 grafik persebaran data berdasarkan program studi.



Gambar 5. Grafik persebaran data setelah cleaning

Pada gambar 5 merupakan gambar yang berisikan data persebaran pada dataset, maka proses selanjutnya adalah klasifikasi dengan model Naïve Bayes menggunakan bahasa pemrograman Python 3 dengan tools jupyter notebook pada anaconda.

Data Mining

dataset yang digunakan pada penelitian ini berjumlah 101 mahasiswa yang berasal dari data calon wisudawan Fakultas Sains dan Teknologi Universitas PGRI Sumatera Barat, setelah itu dataset dibagi menjadi data training dan data testing. Data training berjumlah 70% dan data testing berjumlah 30% dari keseluruhan data. .

Proses Klasifikasi Naïve Bayes

Proses klasifikasi random forest, yang pertama; pemanggilan library import numpy as np, import pandas as pd, from sklearn.model_selection import

DOI : https://doi.org/10.52158/jacost.v5i1.489

train_test_split, from sklearn.naive_bayes import GaussianNB dan from sklearn.metrics import accuracy_score,classification_report, confusion_matrix, kedua; import dataset, ketiga; pisahkan fitur dan label. Ketiga; bgai data menjadi data training dan testing, keempat; membangun model naïve bayes, kelima; latih model pada data pelatihan, keenam; lakukan prediksi pada data pengujian, keenam; evaluasi performa model pada tahap ini akan dilakukan uji data dan evaluasi model dari *source code* sebelumnya terhadap evaluasi model dengan *confussion matrix*. data training, data *confusion matrix* dan hasil dari *precision, recall, dan F1-Scor*.

Selanjutnya mencetak hasil prediksi kelulusan mahasiswa dengan Hasil 0 Terlambat dan Hasil 1 Tepat Waktu. Pada hasil pengujian dilakukan proses klasifikasi terhadap sebaran data testing. Pengujian dilakukan dengan melakukan klasifikasi terhadap data testing Hasil *confussion matrix* dari tiap sebaran data dapat dilihat pada Tabel 4.

Setelah nilai-nilai pada *Confussion Matrix* diketahui, maka proses selanjutnya adalah mengetahui atau menghitung dari nilai *precission, accuracy, recall* dan *F1-Score*. Precission merupakan ketepatan nilai antara permintaan pengguna pada respon system, sedangkan accuracy merupakan perbandingan antara informasi benar yang dijawab sistem dengan keseluruhan data, lalu recall merupakan ketepatan antara informasi yang sama dengan informasi yang pernah dipanggil sebelumnya, terakhir adalah *F1-Score* yaitu merupakan perbandingan rata-rata pada *precission dan recall* yang dibobotkan. Berikut hasil dari *source code* untuk melakukan pengujian model yang dapat dilihat pada Tabel 5.

Tabel 4. Hasil Persebaran *Confussion Matrix Algoritma Naïve Bayes*

		Tepat Waktu	Terlambat
Actual Value	Tepat Waktu	25	4
	Terlambat	0	2

Tabel 5. Pengujian Model

	precision	recall	f1-score	support
0	1.00	0.86	0.93	29
1	0.33	1.00	0.50	2
Accuracy			0.87	31
Macro avg	0.67	0.93	0.71	31
Weightad avg	0.96	0.87	0.90	31

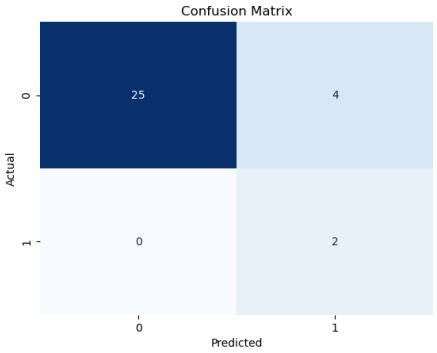
Dari Tabel 5 dapat dilihat bahwa hasil dari evaluasi model yang dapat dilihat nilai *precision* dan *recall* pada setiap kelas dapat terlihat kemampuan pemrosesan sistem dalam mencari tingkat ketepatan kelulusan tepat waktu terhadap mahasiswa antara Tepat Waktu yang memiliki nilai "33%" dan Terlambat "100%". Tingkat keberhasilan dari pemrosesan sistem dalam memperoleh kembali informasi label Tepat Waktu adalah "100%", untuk label Terlambat adalah "86%". Berikut ini gambar 6 yang merupakan hasil dari nilai *accuracy* dari algoritma *Naïve Bayes* yang sudah dijalankan.

```
# Evaluasi performa model
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy}')
```

```
Accuracy: 0.8709677419354839
```

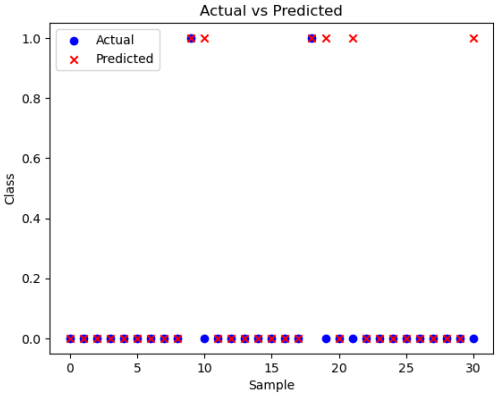
Gambar 6. Hasil akurasi naïve bayes

Gambar 7 adalah visualisasi untuk confusion matrix



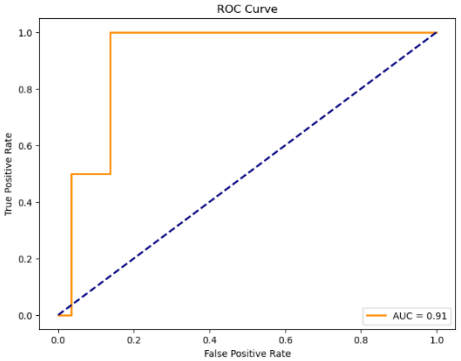
Gambar 7. Confusion Matrix

Gambar 8 adalah visualisasi hasil prediksi



Gambar 8. Visualisasi hasil prediksi

Menampilkan kurva ROC



Gambar 9. Kurva ROC

Proses Klasifikasi Random Forest

Proses klasifikasi random forest, yang pertama; pemanggilan library from sklearn.model_selection

DOI : https://doi.org/10.52158/jacost.v5i1.489

import train_test_split, from sklearn.ensemble import RandomForestClassifier, from sklearn.metrics import accuracy_score, classification_report dan import pandas as pd, kedua; import dataset, ketiga; pisahkan fitur dan label. Ketiga; bgai data menjadi data training dan testing, keempat; inisialisasi model random forest, kelima; latih model pada data pelatihan, keenam; lakukan prediksi pada data pengujian, keenam; evaluasi kinerja model.

Pengujian dilakukan dengan melakukan klasifikasi terhadap data testing Hasil *confussion matrix* dari tiap sebaran data dapat dilihat pada Tabel 6.

Tabel 6. Hasil Persebaran *Confussion Matrix Algoritma Naïve Bayes*

		Tepat Waktu	Terlambat
	Tepat Waktu	18	2
Actual Value	Terlambat	3	8

Setelah nilai-nilai pada *Confussion Matrix* diketahui, maka proses selanjutnya adalah mengetahui atau menghitung dari nilai *precission*, *accuracy*, *recall* dan *F1-Score*. Precission merupakan ketepatan nilai antara permintaan pengguna pada respon system, sedangkan accuracy merupakan perbandingan antara informasi benar yang dijawab sistem dengan keseluruhan data, lalu recall merupakan ketepatan antara informasi yang sama dengan informasi yang pernah dipanggil sebelumnya, terakhir adalah *F1-Score* yaitu merupakan perbandingan rata-rata pada *precission dan recall* yang dibobotkan. Berikut hasil dari *source code* untuk melakukan pengujian model yang dapat dilihat pada Tabel 7.

Tabel 7. Pengujian Model

	precision	recall	f1-score	support
0	0.86	0.90	0.88	20
1	0.80	0.73	0.76	11
Accuracy			0.84	101
Macro avg	0.83	0.81	0.82	101
Weightad avg	0.84	0.84	0.84	101

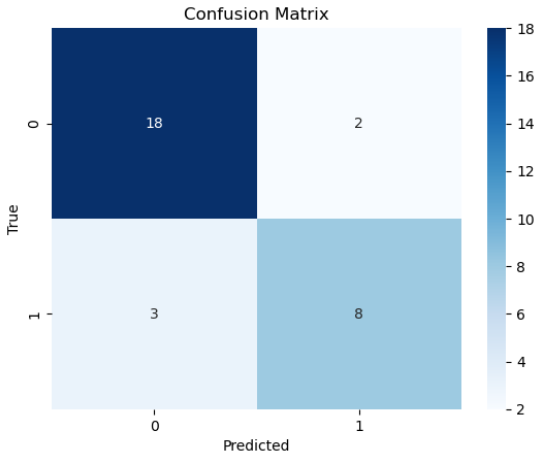
Dari Tabel 7 dapat dilihat bahwa hasil dari evaluasi model yang dapat dilihat nilai *precision* dan *recall* pada setiap kelas dapat terlihat kemampuan pemrosesan sistem dalam mencari tingkat ketepatan kelulusan tepat waktu terhadap mahasiswa antara Tepat Waktu yang memiliki nilai "80%" dan Terlambat "86%". Tingkat keberhasilan dari pemrosesan sistem dalam memperoleh kembali informasi label Tepat Waktu adalah "73%", untuk label Terlambat adalah "90%". Berikut ini gambar 10 yang merupakan hasil dari nilai *accuracy* dari algoritma *Naïve Bayes* yang sudah dijalankan.

```
print(f'Accuracy: {accuracy}')
print('Classification Report:\n', report)
```

```
Accuracy: 0.8387096774193549
```

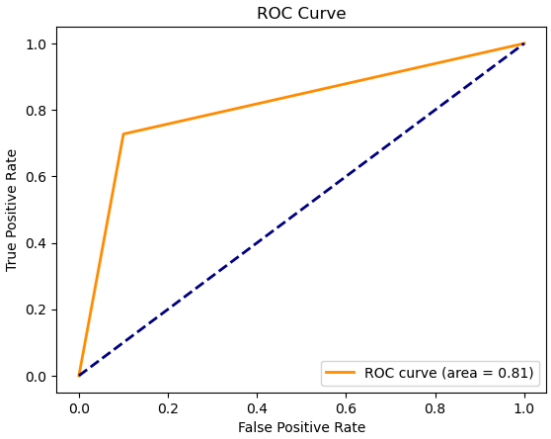
Gambar 10. Hasil Akurasi Random Forest

Gambar 11 adalah visualisasi hasil prediksi random forest confusion matrix



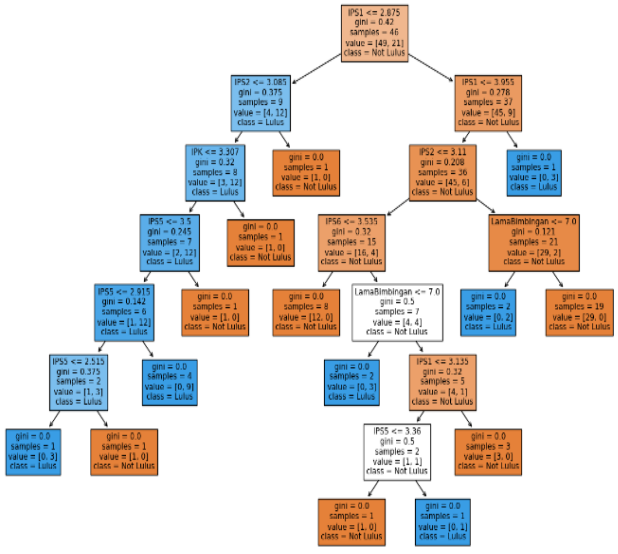
Gambar 11. Confusion Matrix Random Forest

Gambar 12 adalah Receiver Operating Characteristic (ROC) Curve.



Gambar 12. Receiver Operating Characteristic (ROC) Curve

Gambar 13 adalah visualisasi Pohon Keputusan.



Gambar 13. Visualisasi Pohon Keputusan

DOI : https://doi.org/10.52158/jacost.v5i1.489

Proses Klasifikasi Support Vector Machine (SVM)

Proses klasifikasi random forest, yang pertama; pemanggilan library from sklearn.model_selection import train_test_split, from sklearn.svm import SVC, from sklearn.metrics import accuracy_score, classification_report import pandas as pd kedua; import dataset, ketiga; pisahkan fitur dan label. Ketiga; bgai data menjadi data training dan testing, keempat; inisialisasi model SVM, kelima; latih model pada data pelatihan, keenam; lakukan prediksi pada data pengujian, keenam; evaluasi kinerja model.

Pengujian dilakukan dengan melakukan klasifikasi terhadap data mengetahui atau menghitung dari nilai *precission, accuracy, recall* dan *F1-Score.* Precission merupakan ketepatan nilai antara permintaan pengguna pada respon system, sedangkan accuracy merupakan perbandingan antara informasi benar yang dijawab sistem dengan keseluruhan data, lalu recall merupakan ketepatan antara informasi yang sama dengan informasi yang pernah dipanggil sebelumnya, terakhir adalah *F1-Score* yaitu merupakan perbandingan rata-rata pada *precission dan recall* yang dibobotkan. Berikut hasil dari *source code* untuk melakukan pengujian model yang dapat dilihat pada Tabel 8.

Tabel 8. Pengujian Model

	precision	recall	f1-score	support
0	1.00	0.90	0.95	20
1	0.85	1.00	0.92	11
Accuracy			0.94	101
Macro avg	0.92	0.95	0.93	101
Weightad avg	0.95	0.84	0.94	101

Dari tabel dapat dilihat bahwa hasil dari evaluasi model yang dapat dilihat nilai *precision* dan *recall* pada setiap kelas dapat terlihat kemampuan pemrosesan sistem dalam mencari tingkat ketepatan kelulusan tepat waktu terhadap mahasiswa antara Tepat Waktu yang memiliki nilai "85%" dan Terlambat "100%". Tingkat keberhasilan dari pemrosesan sistem dalam memperoleh kembali informasi label Tepat Waktu adalah "100%", untuk label Terlambat adalah "90%". Berikut ini gambar 14 yang merupakan hasil dari nilai *accuracy* dari algoritma *Naïve Bayes* yang sudah dijalankan.

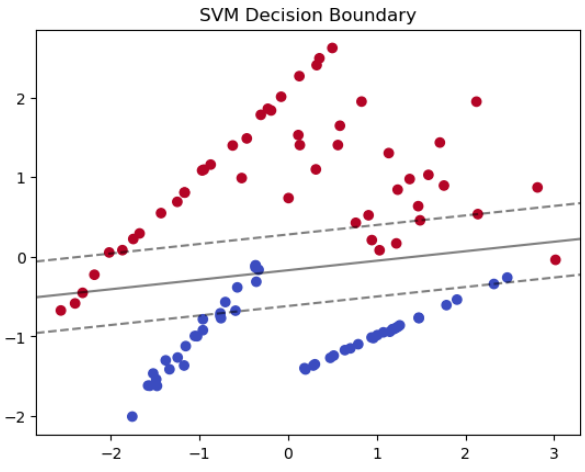
```
: # Evaluasi kinerja model
accuracy = accuracy_score(y_test, predictions)
report = classification_report(y_test, predictions)

print(f'Accuracy: {accuracy}')
print('Classification Report:\n', report)
```

```
Accuracy: 0.9354838709677419
```

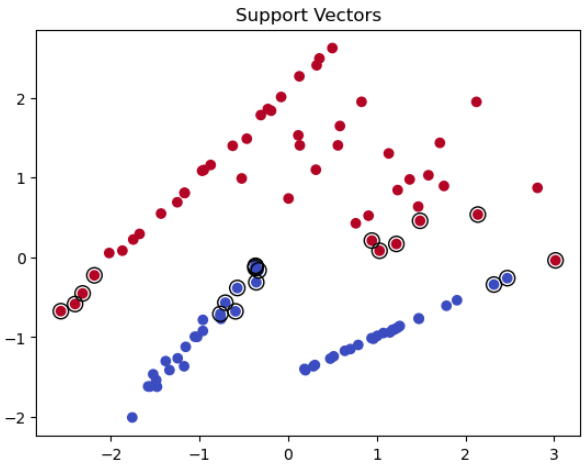
Gambar 14. Hasil Akurasi SVM

Visualisasi data batas keputusan



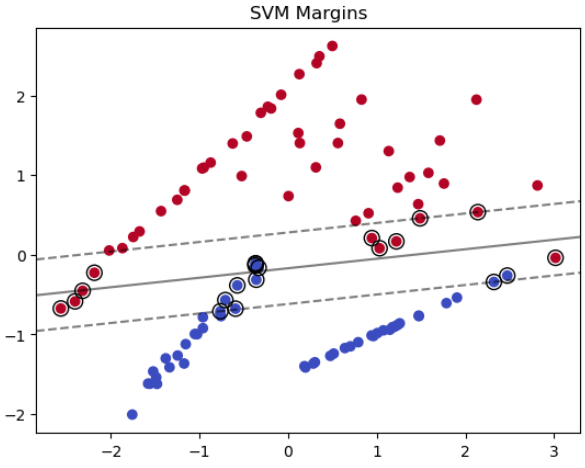
Gambar 15. Decision Boundary SVM

Menampilkan support vectors



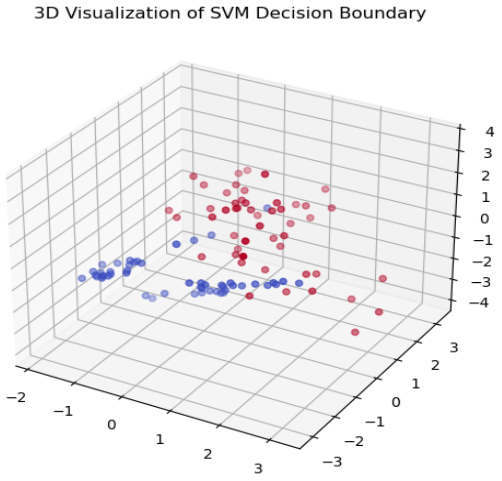
Gambar 16. Support Vectors

Menampilkan SVM Margins



DOI : https://doi.org/10.52158/jacost.v5i1.489

Visualisasi 3D



Gambar 18. Visualisasi 3D

Proses Klasifikasi Artificial Neural Network (ANN)

Proses klasifikasi random forest, yang pertama; pemanggilan library import numpy as np, import pandas as pd, from sklearn.model_selection import train_test_split, from sklearn.preprocessing import StandardScaler, from sklearn.metrics import accuracy_score, from keras.models import Sequential, from keras.layers import Dense, kedua; import dataset, ketiga; pisahkan fitur dan label. Ketiga; bgai data menjadi data training dan testing, keempat; normalisasi data dan membangun model neural network, kelima; latih model pada data pelatihan, keenam; lakukan prediksi pada data pengujian, keenam; evaluasi kinerja model.

Pengujian dilakukan dengan melakukan klasifikasi terhadap data testing Hasil *confussion matrix* dari tiap sebaran data dapat dilihat pada Tabel 4.

Tabel 9. Hasil Persebaran *Confussion Matrix Algoritma Naïve Bayes*

		Tepat Waktu	Terlambat
Actual Value	Tepat Waktu	29	71
	Terlambat	27	73

Setelah nilai-nilai pada *Confussion Matrix* diketahui, maka proses selanjutnya adalah mengetahui atau menghitung dari nilai *precission, accuracy, recall* dan *F1-Score*. Precission merupakan ketepatan nilai antara permintaan pengguna pada respon system, sedangkan accuracy merupakan perbandingan antara informasi benar yang dijawab sistem dengan keseluruhan data, lalu recall merupakan ketepatan antara informasi yang sama dengan informasi yang pernah dipanggil sebelumnya, terakhir adalah *F1-Score* yaitu merupakan perbandingan rata-rata pada *precission dan recall* yang dibobotkan. Berikut hasil dari *source code* untuk melakukan pengujian model yang dapat dilihat pada tabel 10 berikut ini:

Tabel 10. Pengujian Model

	precision	recall	f1-score	support
0	0.51	0.73	0.60	15
1	0.56	0.69	0.62	16
Accuracy			0.64	31

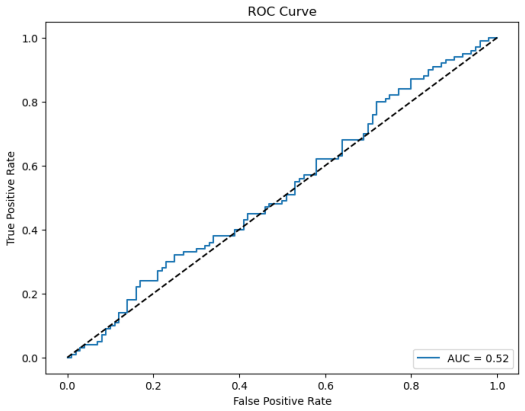
Dari tabel dapat dilihat bahwa hasil dari evaluasi model yang dapat dilihat nilai *precision* dan *recall* pada setiap kelas dapat terlihat kemampuan pemrosesan sistem dalam mencari tingkat ketepatan kelulusan tepat waktu terhadap mahasiswa antara Tepat Waktu yang memiliki nilai "56%" dan Terlambat "51%". Tingkat keberhasilan dari pemrosesan sistem dalam memperoleh kembali informasi label Tepat Waktu adalah "69%", untuk label Terlambat adalah "73%". Berikut ini gambar 19 yang merupakan hasil dari nilai *accuracy* dari algoritma *Naïve Bayes* yang sudah dijalankan.

```
print("Accuracy:", accuracy_score(y_test, y_pred))
print("Precision:", precision_score(y_test, y_pred))
print("Recall:", recall_score(y_test, y_pred))
print("F1 Score:", f1_score(y_test, y_pred))
```

```
Accuracy: 0.6451612903225806
```

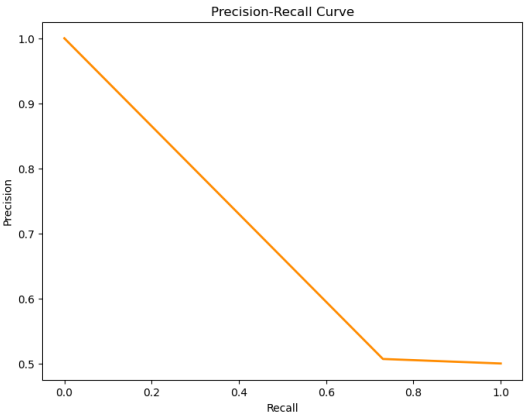
Gambar 19. Hasil Akurasi ANN

Tampilan Kurva ROC (Receiver Operating Characteristic)



Gambar 20. Kurva ROC

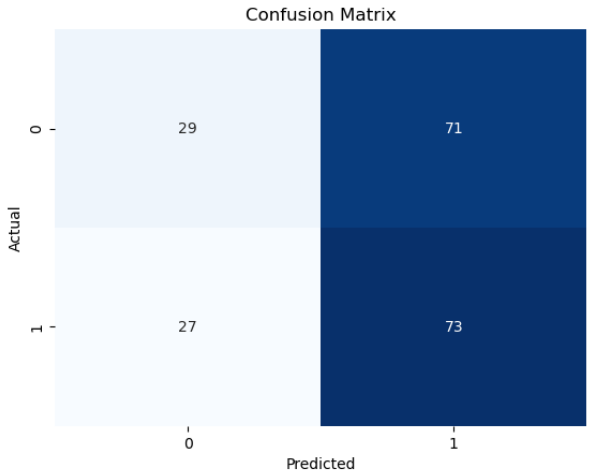
Kurva Precision-Recall:



Gambar 21. Kurva Precision-Recall

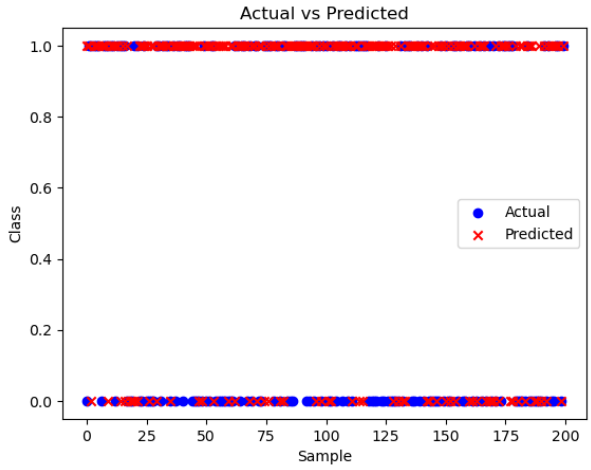
DOI : https://doi.org/10.52158/jacost.v5i1.489

Confusion Matrix



Gambar 22. Confusion Matrix ROC ANN

Grafik Perbandingan Hasil Prediksi dan Nilai Sebenarnya



Gambar 23. Hasil Prediksi

4. Kesimpulan

Hasil klasifikasi tersebut di dapatkan, pada *naïve bayes* memiliki tingkat Akurasi sebesar 0.87. Untuk nilai *percision* Kelas Tepat Waktu memiliki nilai "0.33" sedangkan Terlambat sebesar "1.00". Untuk nilai *recall* Kelas Tepat Waktu memiliki nilai "1.00" sedangkang kelas Terlambat sebesar "0.86". pada *random forest* memiliki tingkat Akurasi sebesar 0.84. Untuk nilai *percision* Kelas Tepat Waktu memiliki nilai "0.80" sedangkan Terlambat sebesar "0.86". Untuk nilai *recall* Kelas Tepat Waktu memiliki nilai "0.73" sedangkang kelas Terlambat sebesar "0.90". Pada *support vector machine* (SVM) memiliki tingkat Akurasi sebesar 0.94. Untuk nilai *percision* Kelas Tepat Waktu memiliki nilai "0.85" sedangkan Terlambat sebesar "1.00". Untuk nilai *recall* Kelas Tepat Waktu memiliki nilai "1.00" sedangkang kelas Terlambat sebesar "0.90"dan sedangkan pada *artificial neural network* (ANN) memiliki tingkat Akurasi sebesar 0.65. Untuk nilai *percision* Kelas Tepat Waktu memiliki nilai "0.56" sedangkan Terlambat sebesar "0.51". Untuk nilai *recall* Kelas Tepat Waktu memiliki nilai "0.69" sedangkang kelas Terlambat sebesar "0.73"

Dari hasil perbandingan klasifikasi tersebut di dapat bahwa algoritma *support vector machine* (SVM) memiliki tingkat Akurasi sebesar 0.94 maka memiliki nilai paling baik di antara algoritma yang lain. Maka dapat disimpulkan bahwa dengan data mining dapat di peroleh informasi kelulusan mahasiswa pada database mahasiswa, Dari klasisfikasi data mining terdapat algoritma-algoritma yang mamu memrediksi tingkat keluusan yang diharakan, dari hasil evaluasi di dapatkan bahwa algoritma *support vector machine* (SVM) adalah yang paling baik untuk memprediksi kelulusan mahasiswa karena memiliki akurasi tertinggi dari algoritma lainnya. Saran penelitian selanjutnya agar bisa menggunakan data yang lebih banyak, mencari atribut yang lebih berpengaruh serta menggunakan metode lain untuk prediksi kelulusan tepat waktu mahasiswa.

Daftar Rujukan

[1] U. R. Habibah and A. Solichin, "Prediksi Kelulusan Mahasiswa Dengan Metode Naïve Bayes dan Artificial Neural Network: Studi Kasus Fakultas Teknik UNIS Tangerang," *Fakt. Exacta*, vol. 15, no. 1, pp. 73–83, 2022.

[2] D. Anugrah Putra and M. Kamayani, "Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Naive Bayes di Program Studi Teknik Informatika UHAMKA," *Pros. Semin. Nas. Teknoka*, vol. 5, no. 2502, pp. 34–40, 2020, doi: 10.22236/teknoka.v5i.331.

[3] Hartatik, "Optimasi Model Prediksi Kelulusan Mahasiswa Menggunakan Algoritma Naïve Bayes," *IJAI (Indonesian J. Appl. Informatics)*, vol. 5, 2020.

[4] M. Munawir and T. Iqbal, "Prediksi Kelulusan Mahasiswa menggunakan Algoritma Naive Bayes (Studi Kasus 5 PTS di Banda Aceh)," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 3, no. 2, p. 59, 2019, doi: 10.35870/jtik.v3i2.77.

[5] R. Sepriansyah, S. D. Purnamasari, K. R. N. Wardani, and N. Halim, "Prediksi Kelulusan Mahasiswa Fakultas Teknik Universitas Bina Darma Menggunakan Algoritma Naïve Bayes," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 8, no. 1, pp. 313–322, 2023, doi: 10.29100/jipi.v8i1.3459.

[6] A. Fatkhudin, M. Y. Febrianto, F. A. Artanto, M. W. N. Hadinata, and R. Fahlevi, "Algoritma Decision Tree C.45 Dalam Analisa Kelulusan Mahasiswa Program Studi Manajemen Informatika Umpp," *J. Ilm. ILMU Komput.*, vol. 8, no. 2, pp. 83–86, 2022, doi: 10.35329/jiik.v8i2.240.

[7] R. H. Sukarna and Y. Ansori, "Implementasi Data Mining Menggunakan Metode Naive Bayes Dengan Feature Selection Untuk Prediksi Kelulusan Mahasiswa Tepat Waktu," *J. Ilm. Sains dan Teknol.*, vol. 6, no. 1, pp. 50–61, 2022, doi: 10.47080/saintek.v6i1.1467.

[8] T. M. Satrio Junaidi, "Metode Data Mining Association Rule Dengan Algoritma Fp-Growth Untuk Mengetahui Kelulusan Tepat Waktu Mahasiswa (Studi Kasus Stkip Pgri Sumatera Barat)," *J. Edik Inform.*, vol. 7, no. 1, pp. 9–18, 2018.

[9] S. Widaningsih, "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: 10.36787/jti.v13i1.78.

[10] A. Darmawan, I. Yudhisari, A. Anwari, and M. Makruf, "Pola Prediksi Kelulusan Siswa Madrasah Aliyah Swasta dengan Support Vector Machine dan Random Forest," *J. Minfo Polgan*, vol. 12, no. 1, pp. 387–400, 2023, doi:

DOI : https://doi.org/10.52158/jacost.v5i1.489

10.33395/jmp.v12i1.12388.

[11] E. Rizky and H. Himawan, "Algoritma Support Vector Machine Untuk Memprediksi Nilai Ujian Nasional," *J. Teknol. Inf.*, vol. 11, pp. 172–184, 2015, [Online]. Available: http://research.pps.dinus.ac.id

[12] J. Kusuma, Rubianto, R. Rosnelly, Hartono, and B. H. Hayadi, "Klasifikasi Penyakit Daun Pada Tanaman Jagung Menggunakan Algoritma Support Vector Machine, K-Nearest Neighbors dan Multilayer Perceptron," *J. Appl. Compt. Sci. Technol.*, vol. 4, no. 1, pp. 1–6, 2023, doi: 10.52158/jacost.v4i1.484.

[13] E. Fitri, "Analisis Perbandingan Metode Regresi Linier, Random Forest Regression dan Gradient Boosted Trees Regression Method untuk Prediksi Harga Rumah," *J. Appl. Comput. Sci. Technol.*, vol. 4, no. 1, pp. 58–64, 2023, doi: 10.52158/jacost.v4i1.491.

[14] D. Gellysa Urva, Isa Albanna, Muchamad Sobri Sungkar, S. R. I Made Agus Oka Gunawan, Iwan Adhicandra, H. Rifky Lana Rahardian, H. Rahmadya Trias Handayanto, Anak Agung Gede Bagus Ariana, and S. J. Prima Dina Atika, *Penerapan Data Mining Di Berbagai Bidang (Konsep, Metode, dan Studi Kasus)*. Jambi: SonPedia, 2023.

[15] A. W. Zunan Setiawan, Muhammad Fajar, Arif Mudi Priyatno, Anggi Yhurinda Perdana Putri, Mediana Aryuni, Siti Yuliyanti, Harya Widiputra, Budanis Dwi Meilani, Rohmat Nur Ibrahim, Rezania Agramanisti Azdy, Satrio Junaidi, *Buku Ajar Data Mining*. Jambi: SonPedia, 2023.

DOI : https://doi.org/10.52158/jacost.v5i1.489