PREDIKSI PENERIMA BEASISWA UNIVERSITAS MUHAMMADIYAH PRINGSEWU DENGAN MENGGUNAKAN ALGORITMA NAÏVE BAYES

[1]Bambang Triraharjo, [2]Roby Novianto, [3]Baskoro
[1]bambangtriraharjo@umpri.ac.id, [2]robynovianto@umpri.ac.id, [3]baskoro@umpri.ac.id

[1,2,3)]Universitas Muhammadiyah Pringsewu

Abstract: *Scholarship assignment is an operations management problem facing university administrators, which is usually resolved based on the administrator's personal experience. This research proposes an incentive method inspired by dynamic programming to replace the traditional decision-making process in scholarship assignments. The aim is to find the optimal scholarship assignment scheme with the highest equity while taking into account practical constraints and equity requirements. The methodology used in determining scholarship recipients at Pringsewu Muhammdiyah University uses the Naïve Bayes algorithm. The research results show that the Naïve Bayes algorithm with K-10 and K-Fold Cross Validation with k=10 has an accuracy of 95.01%. This shows that Naïve Bayes is an algorithm that can predict.*

Keywords*: Scholarship, Prediction, Datamining, Naïve Bayes*

Abstrak: Penugasan beasiswa adalah masalah manajemen operasi yang dihadapi administrator universitas, yang biasanya diselesaikan berdasarkan pengalaman pribadi administrator. Penelitian ini mengusulkan metode insentif yang terinspirasi oleh pemrograman dinamis untuk menggantikan proses pengambilan keputusan tradisional dalam penugasan beasiswa. Tujuannya adalah untuk menemukan skema penugasan beasiswa yang optimal dengan ekuitas tertinggi sambil memperhitungkan kendala praktis dan persyaratan ekuitas Metodologi yang digunakan dalam menentukan penerima beasiswa di Universitas Muhammdiyah Pringsewu dengan menggunakan algoritma Naïve Bayes. Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes dengan K-10 dan K-Fold Cross Validation dengan k=10 akurasi mencapai 95.01%. Hal ini menunjukkan bahwa Naïve Bayes merupakan algoritmadapat dalam memprediksi.

Kata Kunci: Beasiswa,Prediksi,Datamining,Naïve Bayes

I. PENDAHULUAN

Universitas Muhammadiyah Pringsewu (UMPRI) merupakan salah satu perguruan tinggi muhammadiyah yang ada di kabupaten pringsewu yang berdiri pada tahun 2019 merupakan penggabungan antara tiga sekolah tinggi muhammadiyah yaitu Sekolah Tinggi Keguruan dan Ilmu Pendidikan (STKIP) Muhammadiyah Pringsewu, Sekolah Tinggi Ilmu Kesehatan

(STIKes) Muhammadiyah Pringsewu dan Sekolah Tinggi Ilmu Ekonomi (STIE) Muhammadiyah Pringsewu. Dengan banyaknya jumlah mahasiswa dan untuk menjadikan mahasiwa berprestasi dalam bidang akademik maka universitas mempunyai program beasiswa.

Penghargaan beasiswa merupakan penghargaan yang diberikan kepada individu sarjana untuk melanjutkan pendidikan ke jenjang yang lebih tinggi. Itu reward yang diberikan dapat berupa akses khusus pada suatu institusi atau keuangan pendampingan. Pada dasarnya, beasiswa memberikan penghasilan bagi yang menerimanya. Biasanya, itu dalam bentuk dana yang dihabiskan untuk mahasiswa selama masa perkuliahan pada masa studi yang diinginkan. Pemerintah Provinsi selalu menawarkan program beasiswa ini setiap tahun. Sayangnya, beasiswa diberikan kepada para siswa subyektif sehingga banyak siswa yang memenuhi syarat tidak mendapatkan beasiswa dan sebaliknya(Pengembangan et al., 2015). Dalam hal ini, sangat penting dan sangat penting bagi administrator universitas untuk mengembangkan alat yang sistematis untuk memberikan beasiswa dengan cara yang efisien dan adil. mekanisme seperti itu memakan waktu dan energi bagi siswa karena mereka cenderung berfokus pada pengumpulan informasi pesaing mereka dan mengembangkan strategi aplikasi yang tepat dengan mengorbankan studi dan penelitian mereka, sehingga memberikan dampak negatif secara keseluruhan pada kinerja akademik mereka. Dalam hal ini, sangat penting dan sangat penting bagi administrator universitas untuk mengembangkan alat yang sistematis untuk memberikan beasiswa dengan cara yang efisien dan adil. mekanisme seperti itu memakan waktu dan energi bagi siswa karena mereka cenderung berfokus pada pengumpulan informasi pesaing mereka dan mengembangkan strategi aplikasi yang tepat dengan mengorbankan studi dan penelitian mereka, sehingga memberikan dampak negatif secara keseluruhan pada kinerja akademik mereka. Dalam hal ini, sangat penting dan sangat penting bagi administrator universitas untuk mengembangkan alat yang sistematis untuk memberikan beasiswa dengan cara yang efisien dan adil.

Penambangan data (DM) adalah ekstraksi dan pemrosesan informasi berharga dari gudang data besar. DM adalah bagian dari penambangan data. Langkah pertama dalam penambangan data (DM) adalah melihat data dengan berbagai cara dan menemukan informasi yang paling berharga dalam bentuk yang paling diringkas(Masters, 2018). Dalam strategi pemasaran, pendekatan DM sangat bermanfaat karena meminimalkan data yang tidak perlu dan menghemat sumber daya.

Mereka juga membantu menemukan pola perilaku konsumen dan praktis karena pengetahuannya yang sederhana. Terlepas dari hubungan yang jelas antara DM dan analisis data statistik/matematis, sebagian besar pendekatan yang digunakan dalam DM sejauh ini muncul dari subjek statistik(Bruce Ratner, 2017). Sebagai bagian dari penyelidikan kami, kami akan melihat beberapa model dan praktik pendidikan terbaru. Penambangan data berguna untuk mengekstraksi informasi dari kumpulan data yang besar. Ada banyak masalah data diselesaikan dengan menggunakan teknik penambangan data seperti asosiasi, prediksi, klasifikasi, dan pengelompokan. Untuk memecahkan beasiswa masalah penerima, klasifikasi dan cluster akan dilakukan oleh menggunakan teknik penambangan data(Galit Shmueli, 2018). Itu dilakukan dengan membandingkan dua Metode C4.5 dan K-Nearest Neighbors. Algoritma C4.5 membuat pohon keputusan berdasarkan konsep perolehan informasi, dengan setiap keputusan klasifikasi dikaitkan dengan klasifikasi target. Cara terbaik untuk menilai ketidakpastian adalah dengan menggunakan entropi.

Penelitian mengenai Penerapan Algoritma C4.5 Pada Asuransi dan Jasa Keuangan Menggunakan Metode Data Mining Setelah melakukan percobaan, didapat hasil akurasi tertinggi yaitu 96,25%(Xuanyuan et al., 2022), Selain itu Penerapan Algoritma Pohon Keputusan C4.5 untuk Mengevaluasi Pendidikan Musik Perguruan Tinggi Instruktur musik di universitas dan perguruan tinggi terus memperbarui metode pengajaran mereka dan memanfaatkan beberapa teknik untuk memberikan pengajaran mendalam di ruang kelas. Untuk memperluas antusiasme dan keterlibatan siswa sekaligus mengembangkan bakat kreatif musik mereka, sistem manajemen administrasi pendidikan informasi berbasis web telah banyak digunakan di banyak universitas dan perguruan tinggi.

Hasil penelitian menunjukkan bahwa enam model ML terutama digunakan: pohon keputusan (DT), jaringan syaraf tiruan (ANN), mesin vektor pendukung (SVM), tetangga terdekat K (KNN), regresi linier (LinR), dan Naive Bayes (NB)(Alsariera et al., 2022).Pengklasifikasi pembelajaran mesin seperti BPNN, RF, dan NB digunakan untuk mengklasifikasikan data kinerja akademik siswa. BPNN memiliki akurasi yang lebih baik untuk klasifikasi dan prediksi prestasi akademik mahasiswa. Untuk memprediksi siswa dengan keterlibatan rendah, kami menerapkan beberapa algoritme ML ke kumpulan data. Dengan menggunakan algoritme ini, model yang dilatih pertama kali diperoleh; kemudian, akurasi dan nilai kappa model dibandingkan. Hasilnya menunjukkan

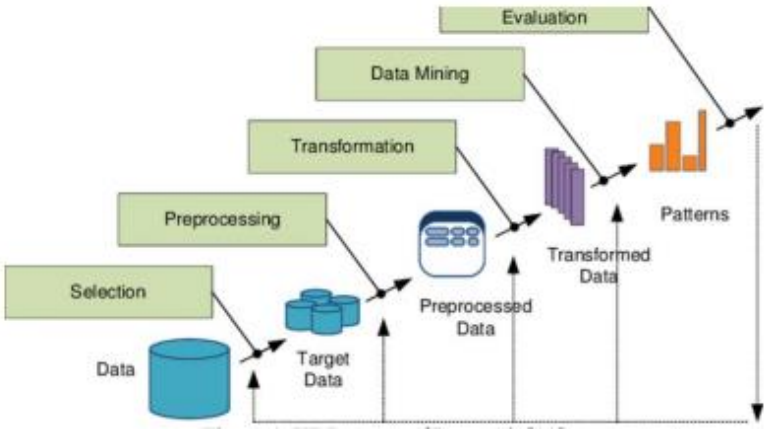
bahwa J48, pohon keputusan, JRIP, dan pengklasifikasi yang ditingkatkan gradien menunjukkan kinerja yang lebih baik dalam hal akurasi, nilai kappa, dan daya ingat dibandingkan dengan model yang diuji lainnya. Berdasarkan temuan tersebut, kami mengembangkan dashboard untuk memfasilitasi instruktur di OU. Model-model ini dapat dengan mudah dimasukkan ke dalam sistem VLE untuk membantu instruktur mengevaluasi keterlibatan siswa selama kursus VLE sehubungan dengan berbagai aktivitas dan materi dan untuk memberikan intervensi tambahan bagi siswa sebelum ujian akhir mereka(Hussain et al., 2018).

Penelitian ini melakukan analisis komparatif antara algoritma Naïve Bayes untuk prediksi penerimaan beasiswa. Pilihan penggunaan algoritma ini karena sangat populer dan banyak digunakan dalam praktek. Penelitian ini dilakukan untuk memberikan dukungan keputusan kepada perguruan tinggi, meminimalkan kesalahan yang dilakukan oleh penyedia beasiswa.

II. METODE

2.1. Alur Penelitian

Penelitian ini menggunakan Knowledge Discovery in Database (KDD)(Forsyth, 2018), tahapan model proses seperti yang ditunjukkan pada Gambar 1 Penjelasan dari setiap langkah penelitian seperti pada gambar 1 dibawah ini:



Gambar 1 Alur Penelitian

a. Pilihan
   Pada tahap ini dilakukan seleksi data terhadap data mahasiswa aktif dan mahasiswa yang telah mendaftarkan diri sebagai penerima beasiwa.
b. Preprocessing

Seluruh Mahasiswa Angkatan 2019 dan 2020 dari semua jurusan di Universitas Muhammadiyah Pringsewu digunakan sebagai data. Sebuah data pembersihan adalah dilakukan pada data tersebut untuk memeriksa nilai yang hilang, duplikasi data, atau data outlier.

c. Transformasi

Setelah dilakukan pembersihan data, tahap selanjutnya adalah transformasi data berdasarkan tipe data, dimana data tersebut adalah klasifikasi berdasarkan kategorinya.

d. Penambangan Data

Pada tahap ini, teknik penambangan data yang tepat dipilih. Untuk fungsi Prediksi, Naïve Bayes digunakan. Prediksi merupakan pembelajaran terawasi, jadi tahap ini termasuk dalam model pembelajaran yang diawasi.

e. Evaluasi

Tahapan ini dilakukan untuk mengevaluasi hasil prediksi Algoritma yang memiliki nilai relatif dengan klasifikasi data sebenarnya. Metode Confusion Matrix digunakan sebagai metode evaluasi. Penampilan nilai penilaian adalah akurasi dan error.

2.2. Naïve Bayes

Merupakan teknik klasifikasi berdasarkan Teorema Bayes dengan asumsi independensi antar prediktor. Secara sederhana, pengklasifikasi Naive Bayes mengasumsikan bahwa kehadiran fitur tertentu di suatu kelas tidak berhubungan dengan kehadiran fitur lainnya(Yizhou Sun, 2017).Pengklasifikasi Naïve Bayes adalah algoritme pembelajaran mesin terawasi populer yang digunakan untuk tugas klasifikasi seperti klasifikasi teks. Ini termasuk dalam keluarga algoritma pembelajaran generatif, yang berarti memodelkan distribusi masukan untuk kelas atau kategori tertentu. Pendekatan ini didasarkan pada asumsi bahwa fitur data masukan bersifat independen bersyarat berdasarkan kelasnya, sehingga memungkinkan algoritme membuat prediksi dengan cepat dan akurat.

Dalam statistik, Naïve Bayes adalah pengklasifikasi probabilistik sederhana yang menerapkan teorema Bayes. Teorema ini didasarkan pada kemungkinan suatu hipotesis, berdasarkan data dan beberapa pengetahuan sebelumnya. Pengklasifikasi Naive Bayes mengasumsikan bahwa semua fitur dalam data masukan tidak bergantung satu sama lain, yang seringkali tidak berlaku

dalam skenario dunia nyata. Namun, meskipun asumsi ini disederhanakan, pengklasifikasi Naive Bayes banyak digunakan karena efisiensi dan kinerjanya yang baik di banyak aplikasi dunia nyata(Berrar, 2019).

Selain itu, perlu dicatat bahwa pengklasifikasi Naive Bayes adalah salah satu model jaringan Bayesian yang paling sederhana, namun mereka dapat mencapai tingkat akurasi yang tinggi bila digabungkan dengan estimasi kepadatan kernel. Teknik ini melibatkan penggunaan fungsi kernel untuk memperkirakan fungsi kepadatan probabilitas dari data masukan, yang memungkinkan pengklasifikasi mening-katkan kinerjanya dalam skenario kompleks di mana distribusi data tidak terdefinisi dengan baik. Hasilnya, pengklasifikasi Naive Bayes menjadi alat yang ampuh dalam pembelajaran mesin, khususnya dalam klasifikasi teks, pemfilteran spam, dan analisis sentimen, dan lain-lain.

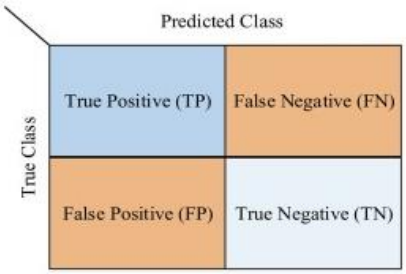
2.3. Evaluasi Kinerja

a. k-fold Cross Validation

   k-fold cross-validation adalah teknik untuk memvalidasi akurasi model yang dibangun pada kumpulan data tertentu, yang membagi kumpulan data menjadi dua bagian, yaitu data pelatihan dan data pengujian. Untuk masalah prediksi, model biasanya diberi dataset dari data yang diketahui untuk dilatih (dataset pelatihan) dan data yang tidak diketahui (atau data yang pertama kali muncul) untuk menguji model (disebut validasi). atau data uji)(Unpingco, 2016). Tujuan validasi silang adalah untuk menguji kemampuan model untuk memprediksi data baru yang tidak digunakan dalam evaluasinya, untuk menandai masalah seperti overfitting atau bias seleksi, dan untuk memberikan wawasan tentang bagaimana model menggeneralisasi data independen . set (yaitu dataset yang tidak diketahui, misalnya masalah).

b. Confusion matrix

   Confusion matrix atau Matriks kebingungan adalah ukuran yang sangat populer digunakan saat memecahkan masalah klasifikasi. Ini dapat diterapkan untuk klasifikasi biner serta untuk masalah klasifikasi multikelas(Caelen, 2017). Matriks ini digunakan untuk evaluasi kinerja metode yang digunakan setelah klasifikasi. Untuk klasifikasi biner, skema dari matriks konfusi terlihat pada Gambar 2



Gambar 2 Skema Confusion Matrix

Matriks konfusi merepresentasikan nilai TP yang diklasifikasikan dengan benar, nilai FP di kelas yang relevan saat seharusnya berada di kelas lain, dan nilai FN di kelas lain saat seharusnya berada di kelas yang relevan dan nilai TN yang diklasifikasikan dengan benar di kelas lain. Metrik kinerja yang paling sering digunakan untuk klasifikasi menurut nilai-nilai ini adalah akurasi (ACC), presisi (P), sensitivitas (Sn), spesifisitas (Sp), dan nilai skor- F. Perhitungan metrik kinerja ini menurut nilai-nilai dalam matriks kebingungan dibuat menurut Persamaan.

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

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

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

$$Sp = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \quad (4)$$

$$F - score = 2\mathrm{x}\frac{\mathrm{P\ x\ Sn}}{\mathrm{P+Sn}} \quad (5)$$

c. Kurva ROC

Kurva ROC adalah plot grafis yang mengilustrasikan kemampuan diagnostik sistem pengklasifikasi biner karena ambang diskriminasinya bervariasi. Metode ini awalnya dikembangkan untuk operator penerima radar militer mulai tahun 1941, yang kemudian memunculkan namanya. Kurva ROC dibuat dengan memplot true positive rate (TPR) terhadap false positive rate (FPR) pada berbagai pengaturan ambang batas. Tingkat positif sejati juga dikenal sebagai sensitivitas , daya ingat , atau probabilitas deteksi. Tingkat positif palsu juga dikenal sebagai probabilitas alarm palsu dan dapat dihitung sebagai (1 - spesifisitas). ROC juga dapat dianggap sebagai

sebidang kekuatan sebagai fungsi dari Kesalahan Tipe Idari aturan keputusan (ketika kinerja dihitung hanya dari sampel populasi, dapat dianggap sebagai estimator dari jumlah ini). Performance keakurasian AUC dapat diklasifikasikan menjadi beberapa kelompok yaitu(Moolayil, 2019):

1. 0.90 – 1.00 = *Exellent Classification*
2. 0.80 – 0.90 = *Good Classification*
3. 0.70 – 0.80 = *Fair Classification*
4. 0.60 – 0.70 = *Poor Classification*
5. 0.50 – 0.60 = *Failure Classification*

III. HASIL DAN PEMBAHASAN

3.1. Dataset

Pengolahan dataset yang dibagi menjadi dataset traning dan testing dengan jumlah 1102 record yang terdiri dari 18 atribut. Data tersebut bisa dilihat atau pada Gambar 3 dibawah ini.

Row No.	BEASISWA	Tahun	Nama Siswa	Status DTKS	Jenis Kelamin	Pekerjaan A...	Penghasila...	Status Ayah	Pekerjaan Ibu
1	1	2021	REYKA ANIS...	0	0	2	1250000	1	1
2	1	2021	AHMAD WAH...	1	1	4	1500000	1	1
3	1	2021	Nur Habibah ...	0	0	5	1250000	1	1
4	1	2021	LENI AGUSTIN	1	0	2	750000	1	2
5	1	2021	WAHYU FIRM...	0	1	5	1000000	1	1
6	1	2021	ROBBY KUR...	0	1	2	1250000	1	1
7	1	2021	Khusnul Khot...	1	0	2	750000	1	2
8	1	2021	RIDAYANI	1	0	2	1000000	1	1
9	0	2021	MUHAMMAD ...	0	1	4	5500000	1	1
10	0	2021	YOHANES W...	0	1	4	5500000	1	1
11	1	2021	NITA KURNIA...	0	0	3	750000	1	1
12	1	2021	Jesisca Arind...	1	0	4	1250000	1	1
13	0	2021	MERY	0	0	4	5500000	1	2

Gambar 3 Potongan Dataset

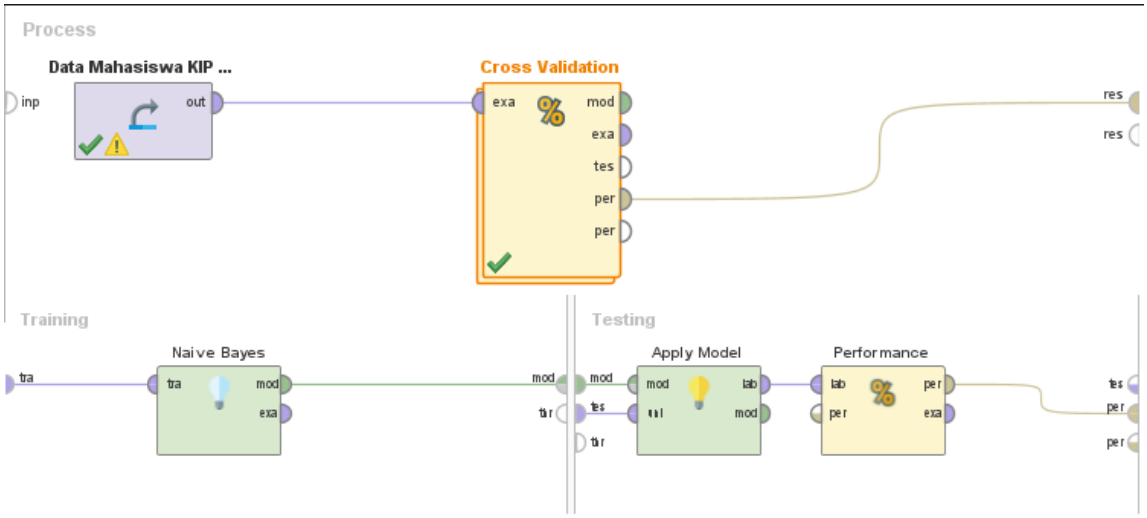
3.2. Pembahasan

3.2.1. Eksperimen dan Pengujian

Tahap implementasi dari algoritma ini adalah penulis melakukan pengujian. Dalam pengujian dari total 1102 data preprocessing penulis membagi data menjadi data latih (pelatihan) dan data uji (testing).

Dalam pengujian ini penulis menggunakan operator validasi silang. Operator validasi silang ininantinya akan memisahkan data dengan cara

membagi total data dari data training dan sisanya untuk menguji data. Penerapan data pada Rapid Miner digunakan untuk Prediksi Penerima Beasiswa menggunakan algoritma *Naïve Bayes* ditunjukan pada gambar 4 dibawah ini:



Gambar 4 Skema pengujian dengan rapid miner

3.2.2. Evaluasi dan Validasi Hasil

1. Hasil Pengujian Algoritma Naïve Bayes

   Pada tahap ini peneliti menggunakan metode algoritma Naïve Bayes untuk mengaplikasikan data yang telah mengalami proses preprocessing data atau pembersihan data pada aplikasi Rapidminer. Berdasarkan pengujian yang dilakukan menggunakan aplikasi Rapidminer didapatkan hasil yaitu: algoritma Naïve Bayes mencapai akurasi 95.01%

accuracy: 95.01% +/- 1.35% (micro average: 95.01%)

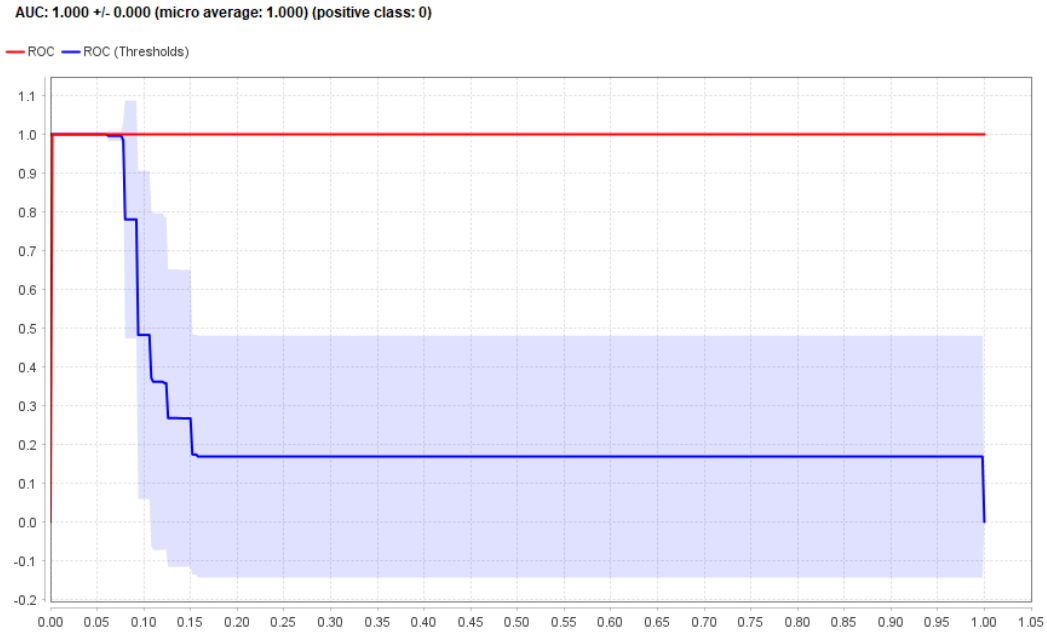
	true 1	true 0	class precision
pred. 1	590	0	100.00%
pred. 0	55	457	89.26%
class recall	91.47%	100.00%	

Gambar 5 Confusion Matrix C4.5

$$ACC = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} = \frac{590+457}{590+457+55+0} = \frac{1047}{1102} = 95{,}009\%$$

Selain Confusion Matrix untuk mengetahui kinerja dari pengujian ini

kami mengandalkan kurva ROC/AUC.(Area Under Curve) yang dihasilkan. Perbandingan hasil Kurva AUC menggunaakan Algoritma Naïve Bayes dapat kita lihat pada Gambar 2 dibawah ini.



Gambar 6 Hasil Kurva AUC

Berdasarkan klasifikasi tersebut dapat disimpulkan bahwa algoritma Naïve Bayes merupakan algoritma yang akurat untuk memprediksi karena nilai AUC termasuk dalam predikat Excellent Classification yaitu dengan nilai 0.90 – 1.00.

IV. SIMPULAN

Penelitian ini mengembangkan metode untuk mendapatkan skema penerimaan beasiswa yang optimal dengan pemerataan tertinggi bagi penyelenggara universitas. Metode tersebut dapat diterapkan karena memenuhi persyaratan pemerataan bahwa siswa yang berprestasi lebih baik harus menerima beasiswa yang sama atau lebih dari yang diterima oleh siswa yang kurang berprestasi; pemberian beasiswa meniadakan kebutuhan mahasiswa untuk mengajukan beasiswa tertentu secara manual, yang merupakan proses yang memakan waktu dan energi.Berdasarkan hasil penelitian dapat disimpulkan bahwa algoritma Naïve Bayes memiliki performansi yang lebih baik yaitu presisi 89,32%, akurasi 95,01% dan nilai recall 100%, dengan hasil AUC sebesar 1,000.

DAFTAR RUJUKAN

Alsariera, Y. A., Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A. A., & Ali, N. (2022). Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance. In Computational Intelligence and Neuroscience (Vol. 2022). Hindawi Limited. doi: 10.1155/2022/4151487

Berrar, D. (2019). Bayes' Theorem and Naive Bayes Classifier. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), Encyclopedia of Bioinformatics and Computational Biology (pp. 403–412). Oxford: Academic Press. doi: https://doi.org/10.1016/B978-0-12-809633-8.20473-1

Bruce Ratner. (2017). *Statistical and Machine-Learning Data Mining Techniques for Better Predictive Modeling and Analysis of Big Data Third Edition*.

Caelen, O. (2017). *A Bayesian Interpretation of the Confusion Matrix*.

Forsyth, D. (2018). *Probability and Statistics for Computer Science*.

Galit Shmueli, P. C. B. I. Y. N. R. P. K. C. L. Jr. (2018). *DATA MINING FOR BUSINESS ANALYTICS*.

Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. R. (2018). Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores. *Computational Intelligence and Neuroscience*, *2018*. doi: 10.1155/2018/6347186

Masters, T. (2018). Data Mining Algorithms in C++. In Data Mining Algorithms in C++. Apress. doi: 10.1007/978-1-4842-3315-3

Moolayil, J. (2019). Learn Keras for Deep Neural Networks. In Learn Keras for Deep Neural Networks. Apress. doi: 10.1007/978-1-4842-4240-7

Pengembangan, L., Informasi, T., & Komunikasi, D. (2015). *KAMUS ABREVIASI BAHASA INDONESIA*.

Unpingco, J. (2016). Python for probability, statistics, and machine learning. In Python for Probability, Statistics, and Machine Learning. Springer International Publishing. doi: 10.1007/978-3-319-30717-6

Xuanyuan, S., Xuanyuan, S., & Yue, Y. (2022). Application of C4.5 Algorithm in Insurance and Financial Services Using Data Mining Methods. *Mobile Information Systems*, *2022*. doi: 10.1155/2022/5670784

Yizhou Sun. (2017). *CS249: ADVANCED DATA MINING Classification Evaluation and Practical Issues*.