Terakreditasi SINTA Peringkat 2
Surat Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti No. 10/E/KPT/2019
masa berlaku mulai Vol. 1 No. 1 tahun 2017 s.d Vol. 5 No. 3 tahun 2021

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 4 No. 1 (2020) 95 - 101 ISSN Media Elektronik: 2580-0760

Educational Data Mining untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritme Naïve Bayes Classifier

Edi Sutoyo[1], Ahmad Almaarif[2]
[1,2]Program Studi Sistem Informasi, Fakultas Rekayasa Industri, Universitas Telkom
[1]edisutoyo@telkomuniversity.ac.id, [2]ahmadalmaarif@telkomuniversity.ac.id

Abstract

Students are one of the main components in the world of education and are expected to develop academic and non-academic quality as long as they are still students. The quality can be seen from the academic achievements, which are evidence of the efforts made by students. Student academic achievement is evaluated at the end of each semester to determine the learning outcomes that have been achieved. If a student cannot meet certain academic criteria that are stated by fulfilling the requirements to continue his studies, the student may have the potential to not graduate on time or even Drop Out (DO). The high number of students who do not graduate on time or DO in higher education institutions can be minimized by detecting students who are at risk in the early stages of education and is supported by making policies that can direct students to complete their education. Also, if the time for completion of student studies can be predicted then the handling of students will be more effective. One technique for making predictions that can be used is data mining techniques. Therefore, in this study, the Naive Bayes Classifier (NBC) algorithm will be used to predict student graduation at Telkom University. The dataset was obtained from the Information Systems Directorate (SISFO), Telkom University which contained 4000 instance data. The results of this study prove that NBC was successfully implemented to predict student graduation. Prediction of the graduation of these students is able to produce an accuracy of 73,725%, precision 0.742, recall 0.736 and F-measure of 0.735.

Keywords: Data Mining, Classification, Naive Bayes Classifier, Student Graduation

Abstrak

Mahasiswa merupakan salah satu komponen utama pada dunia pendidikan dan diharapkan dapat mengembangkan kualitas akademik maupun non-akademik selama masih berstatus menjadi mahasiswa. Kualitas tersebut dapat dilihat dari prestasi akademik yang diraih, yang merupakan bukti dari usaha yang dilakukan oleh mahasiswa. Prestasi akademik mahasiswa dievaluasi pada akhir setiap semester untuk menentukan hasil belajar yang telah dicapai. Jika mahasiswa tidak dapat memenuhi kriteria akademik tertentu yang dinyatakan dengan memenuhi syarat untuk melanjutkan studi, mahasiswa tersebut dapat berpotensi untuk tidak lulus tepat waktu atau bahkan *Drop Out* (DO). Tingginya jumlah mahasiswa yang tidak lulus tepat waktu atau DO di institusi pendidikan tinggi dapat diminimalkan dengan melakukan deteksi mahasiswa yang berisiko pada tahap awal pendidikan dan ditunjang dengan membuat kebijakan yang dapat mengarahkan mahasiswa agar dapat menyelesaikan pendidikannya. Selain itu, jika waktu penyelesaian studi mahasiswa dapat diprediksikan maka penanganan mahasiswa akan lebih efektif. Salah satu teknik melakukan prediksi yang dapat digunakan adalah dengan teknik data mining. Oleh karena itu, pada penelitian ini Algoritme Naive Bayes Classifier (NBC) akan digunakan untuk melakukan prediksi kelulusan mahasiswa di Universitas Telkom. *Dataset* diperoleh dari Direktorat Sistem Informasi (SISFO), Universitas Telkom yang berisi 4000 data *instance*. Hasil penelitian ini membuktikan bahwa NBC berhasil diimplementasikan untuk memprediksi kelulusan mahasiswa. Prediksi kelulusan mahasiswa tersebut mampu menghasilkan *accuracy* sebesar 73.725%, *precision* 0.742, *recall* 0.736 dan F-*measure* sebesar 0.735.

Kata kunci: Data Mining, Klasifikasi, Naive Bayes Classifier, Kelulusan Mahasiswa

© 2020 Jurnal RESTI

1. Pendahuluan

Lembaga pendidikan tinggi dituntut untuk memberikan pendidikan yang berkualitas bagi mahasiswa sehingga mereka dapat menghasilkan sumber daya manusia yang berpengetahuan, cakap, kreatif, dan kompetitif. Setiap tahun akademik baru, universitas menyelenggarakan

Diterima Redaksi : 16-12-2019 | Selesai Revisi : 06-02-2020 | Diterbitkan Online : 08-02-2020

proses penerimaan mahasiswa baru. Dilihat dari jumlah peminat dari setiap tahun akademik pendidikan tinggi dapat memperhatikan berbagai faktor yang mempengaruhi manajemen kapasitas mahasiswa, salah satunya adalah keakuratan masa studi mahasiswa mengikuti waktu yang ditentukan. Dalam sistem pendidikan tinggi, mahasiswa adalah aset penting untuk lembaga pendidikan, sehingga perlu dipertimbangkan tingkat kelulusan mahasiswa agar tepat waktu. Persentase naik turunnya kemampuan mahasiswa untuk menyelesaikan studi tepat waktu adalah salah satu elemen penilaian akreditasi universitas. Untuk alasan ini, ada kebutuhan untuk dilakukan pemantauan dan evaluasi kecenderungan bagi mahasiswa untuk lulus tepat waktu atau tidak.

Secara umum di Indonesia, evaluasi kegiatan akademik mahasiswa dinyatakan dalam satuan kredit semester (SKS). SKS adalah sistem penyediaan pendidikan menggunakan satuan kredit semester dan menggunakan satuan waktu semester yang terdiri dari 2 semester dalam satu tahun akademik. Mahasiswa dapat dinyatakan lulus setelah menyelesaikan sejumlah sks tertentu. Misalnya, mahasiswa Strata 1 (S1) harus menyelesaikan 144-160 kredit, dan program Diploma 3 (D3) membutuhkan 110-120 SKS. Indeks Prestasi Mahasiswa (IPK) adalah angka yang menunjukkan pencapaian kumulatif atau kemajuan belajar mahasiswa, mulai dari semester pertama hingga semester terakhir yang telah diambil. Indeks Prestasi Semester (IPS) dan IPK digunakan sebagai kriteria untuk memberikan evaluasi akademik dan evaluasi studi di akhir program [1]. Mahasiswa diperbolehkan untuk mengambil beban studi kurang dari jumlah minimum beban per semester, tetapi tidak diizinkan untuk mengambil beban studi lebih besar dari jumlah maksimum yang telah ditentukan.

Perguruan tinggi perlu mendeteksi perilaku mahasiswa sehingga faktor-faktor yang dapat menyebabkan kegagalan mahasiswa dapat diidentifikasi sehingga dapat ditentukan tingkat kelulusan sesuai dengan masa studi, termasuk kemampuan akademik yang rendah, usia masuk perguruan tinggi, indeks prestasi atau faktor-faktor lainnya [2], [3]. Di setiap tahun ajaran baru, sering terjadi tidak seimbangnya mahasiswa baru yang masuk dengan mahasiswa yang lulus. Selain itu, jumlah besar penerimaan mahasiswa kadang tidak seimbang dengan jumlah mahasiswa yang lulus tepat waktu mengikuti ketentuan 4 tahun atau 8 semester. Akibatnya, ada akumulasi mahasiswa yang signifikan di setiap periode kelulusan sehingga proses pendidikan tidak berjalan optimal.

Sebagai pendidikan tinggi, mengetahui tingkat kelulusan tepat waktu mahasiswa sangat penting. Tujuannya adalah untuk mengetahui kinerja mahasiswa lebih awal, dapat merencanakan program dan langkah strategis sehingga di masa depan dapat meningkatkan tingkat kelulusan mahasiswa tepat waktu [4]. Tingkat kelulusan tepat waktu dapat ditingkatkan dengan meningkatkan kualitas pembelajaran dan layanan akademik bagi mahasiswa. Selain itu, jika waktu penyelesaian studi mahasiswa dapat diprediksi, maka penanganan mahasiswa akan lebih efektif. Salah satu teknik untuk membuat prediksi yang dapat digunakan adalah teknik *data mining*. Penambangan data berdasarkan data pendidikan di perguruan tinggi dapat meningkatkan kualitas pembelajaran mahasiswa [5].

Data mining atau penambangan data adalah serangkaian proses untuk mendapatkan pengetahuan atau pola dari kumpulan data [6]. Penambangan data akan memecahkan masalah dengan menganalisis data yang sudah ada dalam database. Penambangan data sering juga disebut *Knowledge Discovery in Databases* (KDD) yaitu suatu aktivitas yang meliputi pengumpulan, penggunaan data historis untuk menemukan pola reguler, pola hubungan dalam kumpulan data yang besar [7]. Hasil dari penambangan data dapat digunakan untuk meningkatkan pengambilan keputusan di masa depan. Sejauh ini, terdapat banyak penelitian yang berkaitan dengan penambangan data dalam tahap pengembangan teori maupun yang telah diimplementasikan pada permasalahan dunia nyata, seperti pada area *clustering*, *association rules*, *classification*, dan *conflict analysis* [8]–[15]. Penerapan unsupervised learning juga telah dilakukan untuk clustering data history dari log transaksi pengguna internet di website microsoft dan MSNBC dengan menggunakan kaidah pada soft set theory dan parameter reduction [8], [9], [16], [17]. Sedangkan penerapan supervised learning juga telah dilakukan untuk klasifikasi batuan berjenis igneous [10], klasifikasi dan predikti kualitas air sungai [11], klasifikasi data citra pada bidang medis [12], seleksi atribut pada *dataset* yang berpengaruh secara signifikan pada proses pengambilan keputusan [13], prediksi penggunaan minyak dalam kurun waktu tertentu [14], dan peramalan kejadian penyakit demam berdarah [15].

Lebih jauh lagi, *data mining* yang berfokus pada data dari pendidikan disebut *Education Data Mining* (EDM) [18]. EDM adalah disiplin ilmu yang muncul terkait dengan pengembangan metode untuk memperoleh informasi tersembunyi dari kumpulan data yang berasal dari data pendidikan. Informasi tersembunyi ini kemudian dievaluasi sehingga dapat digunakan untuk mengatasi masalah di dunia pendidikan seperti *e-learning*, mengukur kinerja mahasiswa, penambangan jaringan sosial, memprediksi kegagalan studi, sistem bimbingan cerdas, penambangan teks, dan sebagainya [18]. Karena pada kenyataannya, berbagai data di universitas terus bertambah karena meningkatnya jumlah kegiatan akademik dan non-akademik di dunia pendidikan.

Salah satu algoritme klasifikasi yang dapat digunakan untuk melakukan proses klasifikasi dan prediksi dalam EDM adalah Naïve Bayes Classifier (NBC) [19]. NBC

telah banyak digunakan untuk berbagai permasalahan klasifikasi dan prediksi, serta telah terbukti bahwa NBC dapat menggapai akurasi lebih tinggi dibandingkan algoritme klasifikasi lainnya [20]–[24]. Mccue [20] melakukan penelitian membandingkan Support Vector Machine (SVM) dengan Naïve Bayes dalam hal akurasi untuk klasifikasi spam. Hasil yang didapatkan, Naïve Bayes mendapatkan hasil yang lebih baik yaitu dengan capaian akurasi 97.8% dibandingkan dengan SVM yang mencapai akurasi 96%. Selanjutnya penelitian dari Huang, Hsu dan Lin [21] yang melakukan perbandingan Naïve Bayes, SVM dan C4.5 Decision Tree untuk memprediksi *Chronic Fatigue Syndrome* (CFS). Hasil menunjukkan bahwa dengan menggunakan 10-*fold cross-validation*, Naïve Bayes mampu mengungguli algoritme SVM dan C4.5 dalam hal *Area Under Curve (AUC), Sensitivity* dan *Specificity*. Penelitian yang dilakukan oleh Aninditya, Hasibuan, dan Sutoyo [22] menunjukkan bahwa Naïve Bayes yang mampu menghasilkan akurasi sebesar 83% untuk klasifikasi *text mining* soal ujian yang disesuaikan dengan level kognitif pada *taxonomy bloom*. Begitu juga Shah dan Jivani [23] menggunakan Decision Tree, Naïve Bayes, dan K-Nearest Neighbor untuk memprediksi kanker payudara. Penelitian ini juga menunjukkan bahwa Naïve Bayes lebih unggul dibandingkan dengan algoritme lainnya dalam hal akurasi dan waktu yang digunakan untuk eksekusi algoritme. Hämäläinen dan Vinni [24] juga melakukan penelitian untuk memprediksi dan mengklasifikasi keberhasilan siswa dalam belajar dengan menggunakan data yang terbatas, yaitu 100 baris data. Perbandingan akurasi untuk algoritme Linear Regression (LR), SVM dan Naive Bayes dilakukan, hasil menunjukkan bahwa Naive Bayes mampu unggul dengan akurasi 80% dibandingkan algoritme lainnya. Oleh karena itu pada penelitian ini, algoritme Naïve Bayes Classifier akan digunakan untuk klasifikasi kelulusan mahasiswa Universitas Telkom.

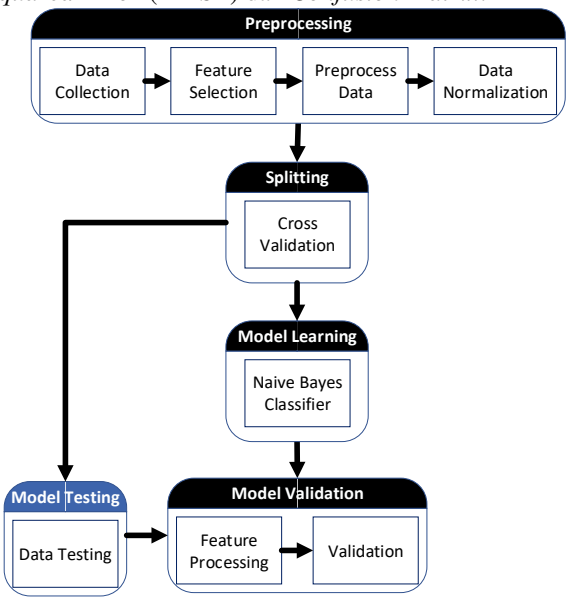
Kelanjutan dari makalah ini disusun sebagai berikut. Pada Bagian 2 menjelaskan tentang metode penelitian, teori Naive Bayes Classifier, dan metriks pengukuran performa algoritme, dan penjelasan mengenai *dataset* yang digunakan beserta *metadata*-nya. Selanjutnya, hasil dan pembahasan penelitian akan diuraikan pada Bagian 3, yang berisi performansi Naïve Bayes Classifier dan detail pembahasannya. Akhirnya, kesimpulan dari penelitian ini disajikan pada Bagian 4.

2. Metode Penelitian

Penelitian ini merupakan penelitian berbasis eksperimental. Pada bagian ini menjelaskan tentang metode penelitian, teori Naïve Bayes Classifier, dan metriks pengukuran performa algoritme seperti *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE), dan *confusion matrix*.

2.1. Metodologi Penelitian

Gambar 1 menjelaskan metodologi penelitian yang digunakan. Terdapat 5 tahap utama, yaitu *preprocessing*, *splitting dataset*, *model learning*, *model testing* dan *model validation*. Pada tahap *preprocessing*, terdapat 4 tahap yaitu data *collection, feature selection, preprocess* data, dan data *normalization*. Pada *model learning*, Naïve Bayes Classifier akan digunakan untuk proses *learning*. Sedangkan untuk *model testing*, k-*Fold Cross-Validation* akan digunakan untuk proses membagi antara data *training set* yang digunakan untuk *model learning* dan *testing set* yang digunakan untuk *model testing*. Pada tahap *model validation* akan digunakan *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE) dan *Confusion Matrix*.



Gambar 1. Metodologi Penelitian

2.2. Naïve Bayes Classifier (NBC)

Naïve bayes merupakan pengklasifikasian sederhana yang menghitung probabilitas dengan menjumlahkan frekuensi dan nilai dari data yang ada. Naïve Bayes berdasar kepada teorema bayes yang digunakan untuk menghitung probabilitas dari tiap kelas dengan asumsi kelas satu dengan yang lain independen (tidak saling tergantung). Definisi lain yaitu Naïve Bayes merupakan metode untuk memprediksi peluang dimasa depan berdasarkan dengan pengalaman sebelumnya. Naïve Bayes memiliki tingkat akurasi dan kecepatan yang lebih tinggi saat di aplikasikan kedalam suatu database dengan nilai yang besar [19].

Bentuk Umum dari teorema Bayes adalah [19], [25]:

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

Keterangan;

X : Data dengan class yang belum diketahui
H : Hipotesis data X merupakan suatu class spesifik
$P(H|X)$: Probabilitas hipotesis H berdasarkan kondisi X (posteriori probabilitas)

$P(H)$: Probabilitas hipotesis H (prior probabilitas)
$P(X|H)$: Probabilitas X berdasarkan kondisi pada hipotesis H
$P(X)$: Probabilitas X

Naïve Bayes merupakan penyerdehanaan dari metode bayes. Teorema Bayes disederhanakan menjadi:

$$P(H|X) = P(H)P(X) \quad (2)$$

Bayes Rules digunakan untuk menghitung posterior dan probabilitas dari data sebelumnya. Hasil akhirnya akan memberi informasi prior dan posterior untuk menghasilkan probabilitas menggunakan Bayes.

2.3. *Mean Absolute Error (MAE)*

Mean Absolute Error (MAE) mengukur besarnya nilai rata-rata kesalahan dalam serangkaian prediksi, tanpa mempertimbangkan arahnya. MAE adalah rata-rata di atas sampel uji perbedaan absolut antara prediksi dan pengamatan aktual di mana semua perbedaan individu memiliki bobot yang sama.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right| \quad (3)$$

2.4. *Root Mean Squared Error (RMSE)*

Root Mean Squared Error (RMSE) adalah aturan penilaian kuadrat yang juga mengukur besarnya rata-rata kesalahan. RMSE adalah akar kuadrat dari rata-rata perbedaan kuadrat antara prediksi dan observasi aktual.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2} \quad (4)$$

2.5. *Confusion Matrix*

Confusion matrix adalah tabel yang digunakan untuk menggambarkan kinerja model klasifikasi (atau "*classifier*") pada set data uji yang nilai sebenarnya diketahui. Ini memungkinkan identifikasi yang mudah dari kebingungan antar *class*, misal satu kelas umumnya salah diberi label sebagai yang lain. Jumlah prediksi yang benar dan salah dirangkum dengan nilai-nilai hitung dan dipecah oleh masing-masing *class*. *Confusion matrix* menunjukkan cara-cara ketika klasifikasi bingung menentukan *class*-nya dalam membuat prediksi. Ini memberi informasi detail tidak hanya tentang kesalahan yang dibuat oleh classifier tetapi lebih penting lagi jenis kesalahan yang dibuat.

Confusion matrix memvisualisasikan akurasi *classifier* dengan membandingkan kelas aktual dan prediksi. Biner *classifier* memprediksi semua *instance* data dari *dataset* uji sebagai positif atau negatif. Klasifikasi ini menghasilkan empat hasil yaitu *True Positive* (TP), *False Positive* (FP), *True Negative* (TN*), False Negative* (FN). TP menghasilkan nilai yang diprediksi dengan benar diprediksi sebagai positif *actual*, FP menghasilkan nilai yang diprediksi salah memprediksi positif sebenarnya. misal, nilai-nilai negatif diprediksi sebagai positif. Sedangkan FN menghasilkan nilai positif tetapi diprediksi negatif dan TN menghasilkan nilai yang diprediksi dengan tepat diprediksi sebagai negatif aktual. Pada Gambar 2 berikut menjelaskan 4 (empat) hasil dari klasifikasi *confusion matrix*.

Confusion Matrix		*Modeled Values:* x_m	
		True	*False*
Actual Values: x	*True*	*TP*	*FN (Type II error)*
	False	*FP (Type I error)*	*TN*

Gambar 2. Empat pembagian *class* pada *Confusion Matrix*

Dari *confusion matrix* tersebut dapat dibuat metriks pengukuran untuk mendapatkan nilai *Accuracy*, *Precision, Recall*, dan F-*Measure* [26].

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

$$Precision = (TP)/(TP + FP) \quad (6)$$

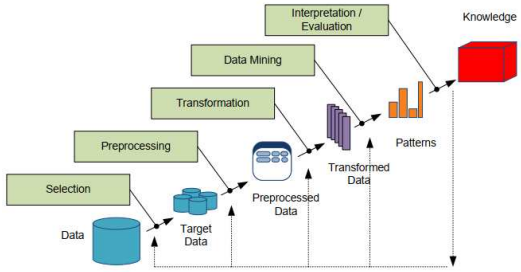
$$Recall = (TP)/(TP + FN) \quad (7)$$

$$F - Measure = 2\frac{(Recall * Precision)}{(Recall + Precision)} \quad (8)$$

2.6. Dataset

Dataset diperoleh dari Direktorat Sistem Informasi (SISFO), Universitas Telkom dari mahasiswa dengan Program Studi Sistem Informasi pada tahun wisuda 2017/2018. Dataset ini berisi 4000 data *instances* disertai enam (6) atribut numerik dan satu (1) label output. Atribut yang digunakan adalah Nilai Indeks Prestasi Semester (IPS) pada semester pertama, Nilai IPS semester kedua, Nilai IPS semester ketiga, Nilai IPS semester keempat, Nilai IPS semester keenam dan ditambah dengan *class* label kelulusan (Tepat Waktu (TW) / Tidak Tepat Waktu (TTW) sebagai label *output*. Pada *class* label TW dan TTW, masing-masing terdapat 2000 data *instances*. Pada proses *feature selection*, atribut yang tidak digunakan seperti nama mahasiswa, nomor telepon, alamat email, nama orang tua, pekerjaan orang tua, alamat lengkap telah dihapus.

Pada penelitian ini, agar dataset cocok dengan model algoritma yang digunakan, proses pengolahan dari *raw data* menjadi dataset mengikuti langkah-langkah di *Knowledge Discovery in Databases* (KDD). Proses KDD ini adalah proses yang interaktif dan berulang, melibatkan banyak langkah, termasuk *preprocessing* data (pembersihan data) dan *postprocessing* (evaluasi hasil). Proses KDD tersebut dapat ditunjukkan pada Gambar 3.



Gambar 3. Proses *Knowledge Discovery in Databases* (KDD) [6]

Data yang telah dikumpulkan kemudian direkapitulasi dan diperoleh analisis deskriptif dari masing-masing variabel. Hasil analisis dapat ditunjukkan pada Tabel I.

Tabel 1. Sampel Dataset yang Digunakan

No	IPS1	IPS2	IPS3	IPS4	IPS5	IPS6	Label
1	3.14	3.20	3.10	3.10	3.20	3.20	TW
2	3.00	2.70	2.50	2.80	3.00	2.80	TTW
3	3.00	3.10	3.20	3.00	3.00	3.30	TW
4	3.20	3.25	3.00	3.20	3.30	3.40	TW
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3996	2.90	2.70	2.55	2.80	3.00	2.80	TTW
3997	3.00	2.20	2.50	2.80	2.75	2.80	TTW
3998	2.90	3.20	2.90	3.20	2.25	3.20	TTW
3999	3.20	3.20	3.48	3.48	3.75	3.75	TW
4000	3.78	3.48	3.75	3.50	3.48	3.10	TW

Pada proses analisis, *dataset* tersebut diubah menjadi bentuk kategorisasi data agar dapat diimplementasikan pada algoritme NBC. Pada proses ini, peneliti akan mengklasifikasikan dan mengategorikan masing-masing nilai pada masing-masing atribut. Dalam mengategorikan nilai, digunakan penghitungan dengan Norma Absolut Skala 5. Adapun rumus penghitungan tersebut ditampilkan pada Tabel 2.

Tabel 2. Konversi *Dataset* Menjadi Bentuk Kategorisasi

No	Rentang *Output*	Kategori	Nilai Kategori
1	$x > \mu + 1.5\sigma$	A	5
2	$\mu + 1.5\sigma < x < \mu + 0.5\sigma$	B	4
3	$\mu + 0.5\sigma < x < \mu - 0.5\sigma$	C	3
4	$\mu - 0.5\sigma < x < \mu - 1.5\sigma$	D	2
5	$x < \mu - 1.5\sigma$	E	1

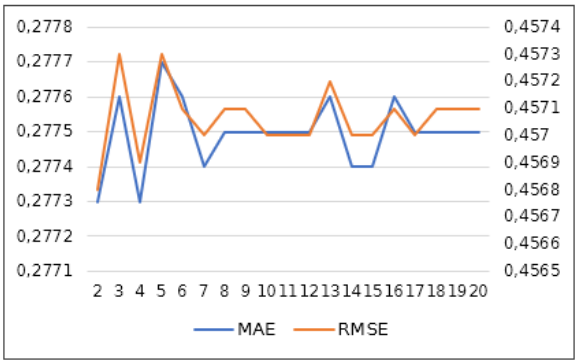
Konversi *output* target aktual ke dalam bentuk representasi huruf dan nilai kategori didasarkan pada distribusi *output* target. Hasil dari sampel *dataset* yang digunakan setelah dikonversi menjadi bentuk kategorisasi dapat ditampilkan pada Tabel 3.

Tabel 3. Hasil Kategorisasi dari Sampel Dataset

No	IPS1	IPS2	IPS3	IPS4	IPS5	IPS6	Label
1	2	3	2	2	3	3	TW
2	2	2	2	2	3	2	TTW
3	2	2	3	2	2	3	TW
4	3	3	2	3	3	3	TW
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3996	2	2	2	2	3	2	TTW
3997	2	1	2	2	2	2	TTW
3998	2	3	2	3	1	3	TTW
3999	3	3	4	4	5	5	TW
4000	5	4	5	4	4	4	TW

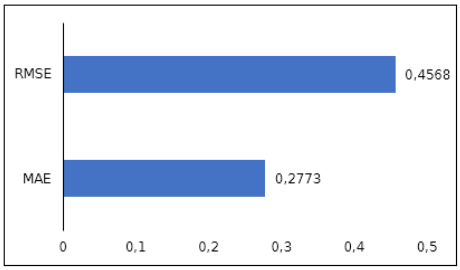
3. Hasil dan Pembahasan

Eksperimen telah dilakukan menggunakan bahasa pemrograman Python yang dijalankan pada komputer dengan spesifikasi Prosesor Intel Core i5 7th *Generation*, memori RAM sebesar 8GB, kapasitas *hard disk* 128GB (SSD)+1TB dan Windows 10 sebagai sistem operasi. Dataset dilakukan partisi menjadi dua bagian menggunakan k-*Fold Cross-Validation* untuk menentukan jumlah data *training set* dan *testing set*, sedangkan k yang dipakai adalah rentang antara 2 sampai dengan 20. Hasil eksperimen tersebut menghasilkan nilai MAE dan RMSE yang beragam, seperti yang ditampilkan pada Gambar 4.

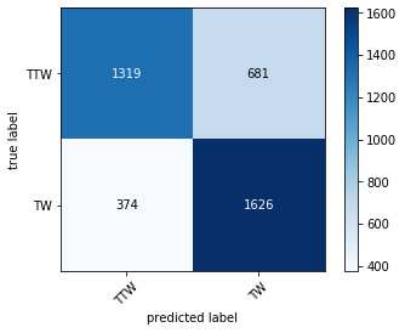


Gambar 4. Hasil Pengukuran MAE dan RMSE Menggunakan Cross-Validation dengan Rentang nilai k=2-20

Berdasarkan pada Gambar 4, dapat diketahui bahwa nilai MAE dan RMSE terkecil adalah ketika nilai k=4. Sehingga hasil eksperimennya didapatkan nilai MAE dan RMSE adalah 0.2773 dan 0.4569, seperti yang ditampilkan pada Gambar 5.



Gambar 5. Hasil Pengukuran MAE dan RMSE Menggunakan *Cross-Validation* dengan Nilai k=4



Gambar 6. Hasil Klasifikasi *Confusion Matrix*

Agar hasil dapat diklasifikasikan lebih spesifik, maka pada penelitian ini juga digunakan *confusion matrix*. Hasil klasifikasi *confusion matrix* digambarkan pada Gambar 6. Dari total 2000 data yang mempunyai *class* TTW, Naïve Bayes Classifier (NBC) berhasil memprediksi 1319 data, yaitu yang masuk pada hasil *True Positive* (TP), dan pada *class* TW, NBC telah berhasil memprediksi sejumlah 1626 data yaitu yang masuk pada *True Negative* (TN). Sedangkan pada *Type 1 Error* atau FP terdapat 681 data dan pada *Type 2 Error* atau FN terdapat 374 data.

Tabel 4. Hasil Eksperimen Menggunakan Naïve Bayes Classifier

Metriks Ukur	Hasil
Accuracy	73.725%
Precision	0.742
Recall	0.736
F-*Measure*	0.735

Seperti yang ditampilkan pada Tabel 4, dari hasil *confusion matrix* tersebut dapat didapatkan hasil untuk *accuracy*, *precision*, *recall* dan F-*measure*. Pada eksperimen ini NBC mampu menghasilkan *accuracy* sebesar 73.725% dengan *precision* 0.742, *recall* 0.736 dan F-*measure* sebesar 0.735. Hasil tersebut menunjukkan bahwa NBC dapat digunakan sebagai model untuk prediksi kelulusan mahasiswa, walaupun akurasi yang didapatkan masih belum memuaskan. Oleh karena itu, penelitian lebih lanjut dimasa depan perlu dilakukan guna memperbaiki tingkat akurasinya.

4. Kesimpulan

Pada penelitian ini pengimplementasian Naïve Bayes Classifier (NBC) untuk prediksi kelulusan mahasiswa telah dilakukan. Penelitian ini menggunakan dataset yang diperoleh dari Direktorat Sistem Informasi (SISFO), Universitas Telkom yang berisi 4000 data *instances* dengan enam (6) atribut, seperti Nilai Indeks Prestasi Semester (IPS) pada semester pertama, Nilai IPS semester kedua, Nilai IPS semester ketiga, Nilai IPS semester keempat, Nilai IPS semester keenam dan ditambah dengan *class* label kelulusan Tepat Waktu (TW) / Tidak Tepat Waktu (TTW) sebagai label *output*. Hasil penelitian menunjukkan bahwa NBC berhasil digunakan untuk mengklasifikasikan kelulusan mahasiswa dengan *accuracy* sebesar 73.725% dengan *precision* 0.742, *recall* 0.736 dan F-*measure* sebesar 0.735. Untuk alasan ini, pendidikan tinggi diharapkan dapat merancang program strategis dan memberikan perlakuan khusus kepada siswa yang berisiko pada tahap awal pendidikan untuk meningkatkan prestasi akademik sehingga mahasiswa yang tidak lulus tepat waktu atau putus sekolah dapat diminimalkan. Rencana kerja masa depan akan berfokus pada evaluasi teknik yang ada dengan tujuan mengembangkan teknik alternatif untuk meningkatkan akurasi, serta menambahkan atribut lain yang berkaitan dengan latar belakang mahasiswa seperti jalur masuk, nilai rapor, nilai ujian nasional (UN), nilai ujian sekolah (US) dan atribut lainnya yang dapat meningkatkan akurasi hasil prediksinya.

Daftar Rujukan

[1] M. R. T. dan P. Tinggi, *Peraturan Menteri Ristek Dan Dikti No 44 Tahun 2015 Standar Nasional Pendidikan Tinggi*. 2015.

[2] P. A. Murtaugh, L. D. Burns, and J. Schuster, "Predicting the retention of university students," *Res. High. Educ.*, vol. 40, no. 3, pp. 355–371, 1999.

[3] C. Márquez-Vera, C. Romero Morales, and S. Ventura Soto, "Predicting school failure and dropout by using data mining techniques," *Rev. Iberoam. Tecnol. del Aprendiz.*, vol. 8, no. 1, pp. 7–14, 2013.

[4] Z. Ibrahim and D. Rusli, "Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression," in *21st Annual SAS Malaysia Forum, 5th September*, 2007.

[5] C. Márquez-Vera, A. Cano, C. Romero, and S. Ventura, "Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data," *Appl. Intell.*, vol. 38, no. 3, pp. 315–330, 2013.

[6] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[7] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[8] E. Sutoyo, I. T. R. Yanto, R. R. Saedudin, and T. Herawan, "A soft set-based co-occurrence for clustering web user transactions," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 15, no. 3, 2017.

[9] E. Sutoyo, I. T. R. Yanto, Y. Saadi, H. Chiroma, S. Hamid, and T. Herawan, "A Framework for Clustering of Web Users Transaction Based on Soft Set Theory," in *Springer*, 2019, pp. 307–314.

[10] I. T. R. Yanto, E. Sutoyo, A. Apriani, and O. Verdiansyah, "Fuzzy Soft Set for Rock Igneous Clasification," in *2018 International Symposium on Advanced Intelligent Informatics (SAIN)*, 2018, pp. 199–203.

[11] E. Sutoyo, R. R. Saedudin, I. T. R. Yanto, and A. Apriani, "Application of adaptive neuro-fuzzy inference system and chicken swarm optimization for classifying river water quality," in *Electrical, Electronics and Information Engineering (ICEEIE), 2017 5th International Conference on*, 2017, pp. 118–122.

[12] M.-L. Antonie, O. R. Zaiane, and A. Coman, "Application of data mining techniques for medical image classification," in *Proceedings of the Second International Conference on Multimedia Data Mining*, 2001, pp. 94–101.

[13] R. R. Saedudin, E. Sutoyo, S. Kasim, H. Mahdin, and I. T. R. Yanto, "Attribute selection on student performance dataset using maximum dependency attribute," in *Electrical, Electronics and Information Engineering (ICEEIE), 2017 5th International Conference on*, 2017, pp. 176–179.

[14] H. Chiroma *et al.*, "An intelligent modeling of oil consumption," *Adv. Intell. Syst. Comput.*, vol. 320, 2015.

[15] A. R. Muhajir, E. Sutoyo, and I. Darmawan, "Forecasting Model Penyakit Demam Berdarah Dengue Di Provinsi DKI Jakarta Menggunakan Algoritma Regresi Linier Untuk Mengetahui Kecenderungan Nilai Variabel Prediktor Terhadap Peningkatan Kasus," *Fountain Informatics J.*, vol. 4, no. 2, pp. 33–40, Nov. 2019.

[16] M. A. T. Mohammed, W. M. W. Mohd, R. A. Arshah, M. Mungad, E. Sutoyo, and H. Chiroma, "Analysis of Parameterization Value Reduction of Soft Sets and Its Algorithm," *Int. J. Softw. Eng. Comput. Syst.*, vol. 2, no. 1, pp. 51–57, 2016.

[17] M. A. T. Mohammed, W. M. W. Mohd, R. A. Arshah, M. Mungad, E. Sutoyo, and H. Chiroma, "Hybrid Framework Parameterization Reduction Combination in Soft Set," in *The Second International Conference on Advanced Data and Information Engineering (DaEng-2015)*, 2019, vol. 520, pp. 233–243.

[18] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 135–146, 2007.

[19] P. Domingos and M. Pazzani, "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss," *Mach. Learn.*, vol. 29, no. 2–3, pp. 103–130, 1997.

[20] R. Mccue, "A Comparison of the Accuracy of Support Vector Machine and Nave Bayes Algorithms In Spam Classification," p. 17, 2009.

[21] L. C. Huang, S. Y. Hsu, and E. Lin, "A comparison of classification methods for predicting chronic fatigue syndrome based on genetic data," *J. Transl. Med.*, vol. 7, p. 81, 2009.

[22] A. Aninditya, M. A. Hasibuan, and E. Sutoyo, "Text Mining Approach Using TF-IDF and Naive Bayes for Classification of Exam Questions Based on Cognitive Level of Bloom's Taxonomy," in *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTAIS)*, 2019.

[23] C. Shah and A. G. Jivani, "Comparison of data mining classification algorithms for breast cancer prediction," in *2013 4th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2013*, 2013, pp. 4–7.

[24] W. Hämäläinen and M. Vinni, "Comparison of machine learning methods for intelligent tutoring systems," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4053 LNCS, pp. 525–534, 2006.

[25] I. Rish, "An empirical study of the naive Bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 2001, vol. 3, no. 22, pp. 41–46.

[26] D. L. Olson and D. Delen, *Advanced data mining techniques*. Springer Berlin Heidelberg, 2008.