Klasifikasi Lama Studi dan Predikat Kelulusan Mahasiswa menggunakan Metode Naïve Bayes

I Nyoman Rudy Hendrawan[1], I Made Arya Budhi Saputra[2], Gusti Ayu Putu Cahya Dewi[3], I Gede Surya Adi Pranata[4], Ni Luh Nyoman Wedasari[5]
Program Studi Sistem Informasi
Institut Teknologi dan Bisnis STIKOM Bali
Denpasar, Indonesia
e-mail: [1]rudyhendrawan@stikom-bali.ac.id, [2]aryabudhi@stikom-bali.ac.id, [3]cahyadewi908@gmail.com, [4]spranata55@gmail.com, [5]weda@stikom-bali.ac.id

Diajukan: 27 Agustus 2021; Direvisi: 27 September 2021; Diterima: 30 September 2021

Abstrak

Ketepatan waktu studi mahasiswa adalah hal yang penting dalam perguruan tinggi. Ketepatan waktu mahasiswa dalam menyelesaikan studi menjadi salah satu penunjang penilaian kualitas perguruan tinggi. Hal ini tentu saja berpengaruh terhadap mutu kelulusan mahasiswa dan predikat kelulusan pada mahasiswa itu sendiri terutama pada saat proses akreditasi. Oleh karena itu, pada penelitian ini diklasifikasikan lama studi dan predikat kelulusan mahasiswa dengan tujuan untuk membantu pihak program studi dan fakultas dalam menganalisis luaran pembelajaran. Metode klasifikasi yang diterapkan pada penelitian ini adalah Naïve Bayes. Data yang digunakan adalah data mahasiswa Institut Teknologi dan Bisnis STIKOM Bali tahun 2008 sampai dengan tahun 2016 dengan total jumlah data sebanyak 5.081. Atribut dataset yang digunakan untuk mengklasifikasikan Lama Studi dan Predikat Kelulusan adalah Jenis Kelamin, Prodi, Konsentrasi, Tahun Masuk, dan Tahun Lulus. Hasil eksperimen menunjukkan bahwa akurasi tes classifier untuk klasifikasi lama studi sebesar 0,74 dan untuk akurasi tes klasifikasi predikat kelulusan sebesar 0,61 pada kelompok data Program Studi Sistem Komputer. Kemudian untuk kelompok data Program Studi Sistem Informasi akurasi tes klasifikasi lama studi sebesar 0,73 dan untuk akurasi klasifikasi predikat kelulusan sebesar 0,67.

Kata kunci: *Klasifikasi, Lama studi, Predikat kelulusan, Naïve Bayes.*

Abstract

Length of study is important at the university level. The punctuality of students in completing their studies is one of the supporting factors for evaluating the quality of higher education. This of course affects the quality of student graduation and the graduation predicate on the students themselves, especially during the accreditation process. Therefore, in this study, study period and the graduate honor of students were classified in regard to assisting faculty in analyzing the learning outcomes. Institut Teknologi dan Bisnis STIKOM Bali students' data is used from 2008 to 2016 with a total of 5.081 data. The dataset attributes that used to classify study period dan graduation honor are gender, concentration, and year of admission. The experimental results show that the classifier test accuracy of study period is 0.74 and the classification test accuracy of the graduate honor is 0.61 in the data group of the Computer Systems study program. Then for the data group of the Information Systems study program, the classification test accuracy of study period is 0.73, and for the classification test accuracy of the graduate honor is 0.67.

Keywords*: Classification, Study period, Graduate honor, Naïve Bayes.*

1. Pendahuluan

Institut Teknologi dan Bisnis STIKOM Bali adalah salah satu perguruan tinggi swasta di Bali yang memiliki dua program studi yaitu program studi sarjana Sistem Informasi dan Sistem Komputer. Lama masa studi yang ditempuh oleh mahasiswa merupakan salah satu standar yang termasuk ke dalam standar penilaian pada Standar Nasional Pendidikan Tinggi atau SN-DIKTI, masa studi untuk program sarjana maksimal tujuh tahun akademik dengan beban belajar mahasiswa paling sedikit 144 SKS. Mahasiswa harus berjuang melewati lebih dari 100 SKS dengan ketentuan IPK minimal 2.00 [1].

Seperti dalam buku Pedoman Pendidikan Tahun 2014-2015 (STIKOM Bali) evaluasi hasil studi pada akhir jenjang studi Strata 1 (S-1) menyebutkan bahwa mahasiswa yang mencapai Indeks Predikat Kumulatif (IPK) minimal 2,00, tidak ada nilai E. Artinya bahwa mahasiswa bisa menempuh perkuliahan hanya dengan 3,5 tahun bila mencapai syarat yang telah ditentukan. Salah satu permasalahan yang sering terjadi di perguruan tinggi khususnya Institut Teknologi dan Bisnis STIKOM Bali yaitu ketidakseimbangan antara mahasiswa yang masuk dan lulus. Mahasiswa yang masuk dalam jumlah banyak, namun jumlah yang lulus tepat waktu jauh lebih sedikit dari pada mahasiswa yang masuk ke Institut Teknologi dan Bisnis STIKOM Bali yang akan berpengaruh terhadap mutu mahasiswa dan predikat kelulusan pada mahasiswa itu sendiri. Data ini dapat dijadikan sebuah informasi yang berharga dalam pengambilan keputusan dengan menganalisis informasi yang ada, maka untuk membantu dalam menemukan informasi berharga itu diperlukan teknik *data mining*. Pada dasarnya *data mining* berhubungan erat dengan analisa data dan penggunaan perangkat lunak untuk mencari pola dan kesamaan dalam pengumpulan data [2]. Beberapa teknik atau algoritma dalam *data mining*, salah satunya adalah Naïve Bayes. Kelebihan dari algoritma Naïve Bayes yaitu relatif mudah untuk diimplementasikan karena tidak menggunakan optimasi numerik, perhitungan matriks dan lainnya, efisien dalam pelatihannya dan penggunaannya, bisa menggunakan data *binary* atau *polinom* karena diasumsikan independen maka memungkinkan metode ini diimplementasikan dengan berbagai macam *dataset* [2].

Penelitian oleh [3], menggunakan metode Naïve Bayes untuk memprediksi peluang kelulusan mahasiswa baru pada suatu perguruan tinggi. Hasil evaluasi klasifikasi menunjukkan bahwa Naïve Bayes menghasilkan keakuratan prediksi sebesar 93,6%. Penelitian lainnya oleh [4], menggunakan metode yang sama untuk mengklasifikasikan kelulusan mahasiswa di Universitas Dian Nuswantoro. Serupa dengan penelitian oleh [4], penelitian oleh [5]-[9] juga menggunakan metode Naïve Bayes untuk memprediksi kelulusan mahasiswa. Berdasarkan penelitian-penelitian tersebut didapatkan bahwa Naïve Bayes memiliki kemampuan yang cukup baik dalam memprediksi kelulusan mahasiswa dengan tingkat akurasi dari 68% hingga 80%.

Pada penelitian ini dilakukan suatu analisa dan prediksi untuk menentukan lama masa studi dan predikat kelulusan yang didapat dari data mahasiswa Institut Teknologi dan Bisnis STIKOM Bali menggunakan metode Naïve Bayes. Untuk data yang digunakan sebagai keperluan penelitian adalah data mahasiswa dari angkatan 2008-2016, menggunakan parameter antara lain jenis kelamin, program studi, konsentrasi, tahun masuk, dan tahun lulus. Penelitian yang dilakukan diharapkan membantu pihak program studi dan fakultas dalam menganalisis luaran pembelajaran.

2. Metode Penelitian

2.1. Tipe Atribut

Metode yang digunakan dalam penelitian ini adalah metode penelitian eksperimen, yang ditunjukkan pada Gambar 1. Data yang digunakan adalah *dataset* mahasiswa Institut Teknologi dan Bisnis STIKOM Bali dari tahun angkatan 2008 sampai dengan tahun angkatan 2016, total jumlah *dataset* ini sebanyak 5.081 data mahasiswa. Jumlah keseluruhan kolom pada *dataset* ini sebanyak sepuluh kolom di antaranya, Nomor Induk Mahasiswa (NIM), nama mahasiswa, jenis kelamin, program studi, konsentrasi, tahun masuk, tahun lulus, lama studi (dalam tahun), predikat kelulusan, dan keterangan kelulusan. Namun, untuk melakukan klasifikasi atribut NIM, nama mahasiswa, dan program studi tidak diikutsertakan, sedangkan kolom predikat kelulusan dan keterangan kelulusan digunakan sebagai label data.

Atribut *dataset* yang digunakan untuk mengklasifikasikan Lama Studi dan Predikat Kelulusan adalah Jenis Kelamin, Prodi, Konsentrasi, Tahun Masuk, dan Tahun Lulus. Kelas Lama Studi dikategorikan berdasarkan tahun lulus yang ditempuh untuk mencapai kelulusan yaitu, Tepat Waktu, jika lama studi delapan semester atau kurang dari delapan semester. Kemudian dikategorikan Terlambat, jika lama studi lebih dari delapan semester. Kelas predikat kelulusan ditentukan berdasarkan IPK yang dibagi menjadi tiga bagian yaitu, Memuaskan, jika IPK antara 2,00 dan 2,75; Sangat Memuaskan, jika IPK antara 2,76 dan 3,50; dan *Cumlaude*, jika IPK antara 3,51 dan 4,0. Tabel 1 menunjukkan detail atribut *dataset* mahasiswa yang digunakan untuk proses klasifikasi.

Tabel 1. Atribut data mahasiswa.

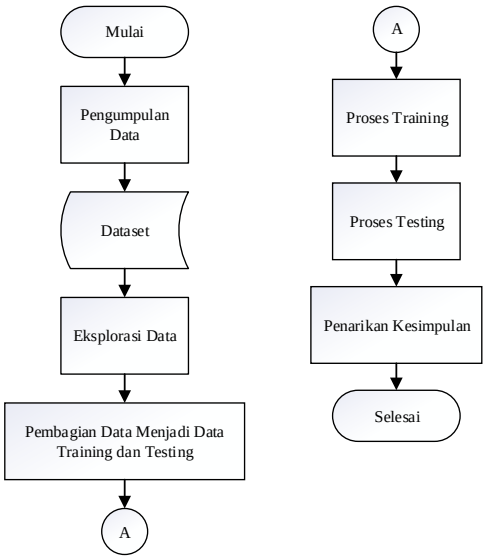
Atribut	Tipe Atribut	Keterangan
Jenis Kelamin	Kategori	Jenis kelamin mahasiswa
Prodi	Kategori	Program Studi mahasiswa
Konsentrasi	Kategori	Konsentrasi mahasiswa
Tahun Masuk	Kategori	Tahun masuk mahasiswa
Tahun Lulus	Kategori	Tahun lulus mahasiswa

2.2. Eksplorasi Data

Berikut adalah contoh *dataset* yang diambil secara acak, Tabel 2 adalah *dataset* yang digunakan untuk klasifikasi Lama Studi dan Tabel 3 digunakan untuk klasifikasi Predikat Kelulusan. Proses klasifikasi pada penelitian ini menggunakan bantuan perangkat lunak Orange. Gambar 2 menunjukkan alur kerja perangkat lunak Orange dalam proses klasifikasi hingga proses evaluasi.

2.3. Pembagian Data

Proses klasifikasi dibagi menjadi dua kategori berdasarkan program studi mahasiswa, yaitu Program Studi Sistem Komputer dan Sistem Informasi, sehingga jumlah data untuk masing-masing program studi adalah sebanyak 2.813 dan 2.246 baris data. Setiap kelompok *dataset* tersebut dibagi menjadi *train set* dan *test set* dengan proporsi sebesar 80% dan 20% secara berturut-turut. Berikut Tabel 4 rincian pembagian data *train set* dan *test set*.



Gambar 1. *Flowchart* alur penelitian.

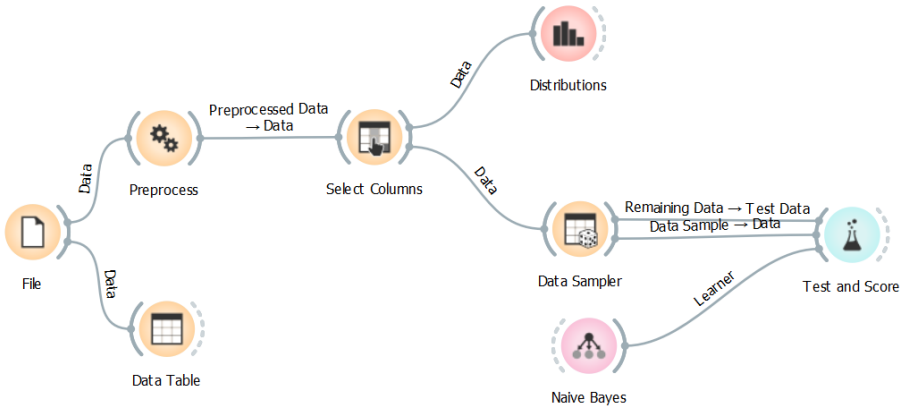
Pada *dataset* kelompok Program Studi Sistem Komputer dan Sistem Informasi terdapat *missing value* sebanyak 486 dan 2 secara berturut-turut pada atribut data Konsentrasi. Untuk mengatasi permasalahan pada jumlah *missing value* yang cukup banyak (sebesar 17% data) di *dataset* kelompok Program Studi Sistem Komputer dilakukan dengan cara mengganti nilai tersebut dengan nilai *random* [10], sedangkan pada Program Studi Sistem Informasi dilakukan penghapusan *record* data dikarenakan jumlahnya jauh lebih sedikit [10].

Tabel 2. Contoh *dataset* mahasiswa kelas lama studi.

Jenis Kelamin	Prodi	Konsentrasi	Tahun Masuk	Tahun Lulus	Lama Studi
L	SK	Networking and Cyber Security	2010	2015/2016	Terlambat
P	SI	Enterprise System	2010	2015/2016	Terlambat
L	SK	Robotic and Embedded System	2010	2013/2014	Tepat

Tabel 3. Contoh *dataset* mahasiswa kelas predikat kelulusan.

Jenis Kelamin	Prodi	Konsentrasi	Tahun Masuk	Tahun Lulus	Predikat kelulusan
P	SI	Intelligence System	2010	2013/2014	Cumlaude
L	SI	Intelligence System	2010	2014/2015	Sangat Memuaskan
L	SK	Robotic and Embeded System	2010	2014/2015	Sangat Memuaskan



Gambar 2. Alur kerja perangkat lunak Orange.

Tabel 4. Rincian pembagian data.

Kelompok Data	Jumlah Train Set	Jumlah Test Set
	80%	20%
Program Studi Sistem Komputer	2250	563
Program Studi Sistem Informasi	1796	449

Berdasarkan alur kerja (Gambar 2) di atas terdapat *widget* yang digunakan antara lain dapat dilihat pada Tabel 5 di bawah ini:

Tabel 5. *Widget* pada Orange.

Widget	Keterangan
File	Digunakan untuk menentukan atribut yang digunakan pada label lama studi dan predikat kelulusan.
Preprocess	Digunakan untuk melakukan proses *preprocessing*.
Data Table	Digunakan untuk menampilkan data dalam tabel.
Select Column	Digunakan untuk menentukan target atau kelas data.
Distributions	Digunakan untuk menampilkan visualisasi data dalam bentuk histogram.
Data Sampler	Digunakan untuk mengatur proporsi pada *train set* dan *test set*.
Test and Score	Digunakan untuk menampilkan hasil pengujian berdasarkan parameter *Area Under the Curve*, *Classifier Accuracy (CA)*, *F1 Score*, *Precission*, dan *Recall*
Naïve Bayes	Metode Naïve bayes yang diimplementasikan
ROC Analysis	Digunakan untuk menampilkan kurva Receiver Operating Characteristic (ROC)

Pengujian dilakukan berdasarkan metrik *Area Under the Curve* (AUC), Akurasi (CA), F1, *Precision* dan *Recall* melalui perangkat lunak *Orange*. Berikut penjelasan mengenai metrik pengujian yang digunakan:

1. AUC adalah suatu daerah di bawah ROC. ROC merupakan kurva yang dihasilkan dari tarik ulur antara sensitivitas dan spesifisitas pada berbagai titik potong. Semakin besar *area under curve* maka semakin baik variabel yang diteliti dalam memprediksi kejadian.
2. Akurasi (CA) merupakan ukuran kinerja yang paling intuitif dan ini hanya rasio pengamatan yang diprediksi dengan benar terhadap total pengamatan.
3. F1 adalah penggabungan *Recall* dan *Precission* ke satu metrik kinerja. Skor F1 adalah rata-rata dari *Precission* and *Recall*. Oleh karena itu, skor ini memperhitungkan prediksi dengan nilai positif dan prediksi dengan nilai negatif.
4. *Precission* merupakan rasio pengamatan positif yang diprediksi dengan benar dengan total pengamatan positif yang diprediksi.
5. *Recall* menunjukkan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif.

3. Hasil dan Pembahasan

3.1. Hasil

Pada bagian ini disajikan hasil klasifikasi kelas Lama Studi pada kategori data Program Studi Sistem Komputer dan Sistem Informasi. Hasil klasifikasi ini (Tabel 6 dan Tabel 7) berdasarkan lima metrik

pengukuran yang telah dijelaskan pada bagian sebelumnya, setiap tabel di bawah ini adalah hasil pengukuran metrik pada *train set* dan *test set*.

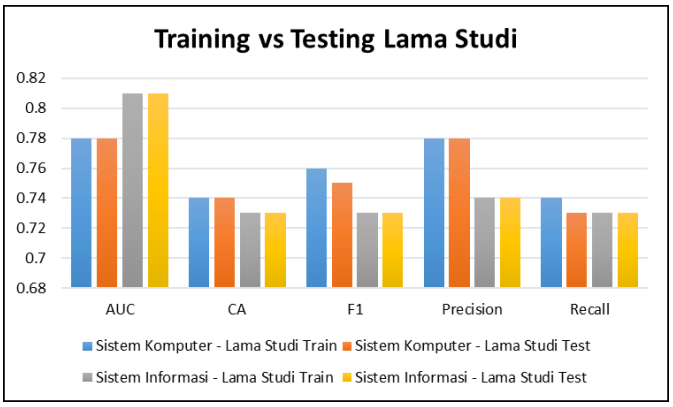
Tabel 6. Hasil pengukuran pada kategori Program Studi Sistem Komputer.

Kelas	Data	AUC	CA	F1	Precision	Recall
Lama Studi	Train	0,78	0,74	0,76	0,78	0,74
	Test	0,78	0,74	0,75	0,78	0,73
Predikat Kelulusan	Train	0,65	0,64	0,57	0,52	0,64
	Test	0,62	0,61	0,54	0,49	0,61

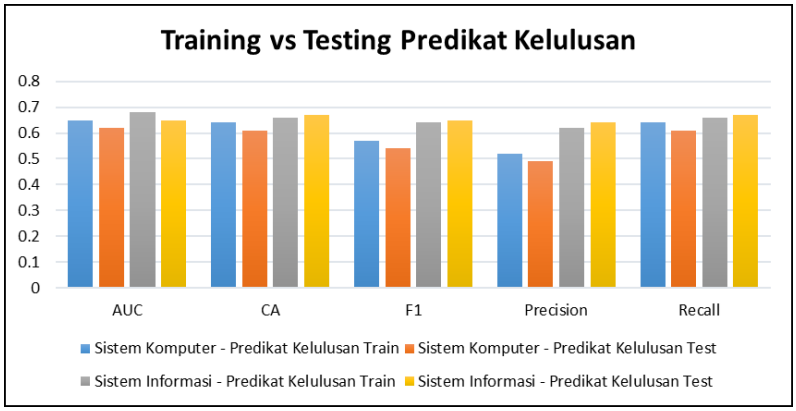
Tabel 7. Hasil pengukuran pada kategori Program Studi Sistem Informasi.

Kelas	Data	AUC	CA	F1	Precision	Recall
Lama Studi	Train	0,81	0,73	0,73	0,74	0,73
	Test	0,81	0,73	0,73	0,74	0,73
Predikat Kelulusan	Train	0,68	0,66	0,64	0,62	0,66
	Test	0,65	0,67	0,65	0,64	0,67

Pada Gambar 3, terlihat grafik perbandingan hasil *training* dan *testing* di semua metrik pengukuran. Tabel 6 dan Tabel 7 juga memperjelas hasil pengukuran tersebut di mana hasil pengukuran *training* dan *testing* tidak jauh berbeda, bahkan jika dilihat secara keseluruhan hasil *training* dan *testing* cenderung menunjukkan hasil yang sama. Namun, jika dilihat dari perbandingan antar kelas pada masing-masing kelompok data, terdapat perbedaan pada hasil pengukuran. Berdasarkan kelima metrik pengukuran, kelas Lama Studi memiliki hasil pengukuran yang lebih baik daripada kelas Predikat Kelulusan. Hal ini dapat dilihat pada Tabel 6 dan Tabel 7, di mana nilai metrik pada kelas Predikat Kelulusan tidak ada yang mencapai nilai 0,7.



Gambar 3. Grafik hasil *training* dan *testing* pada kelas lama studi.

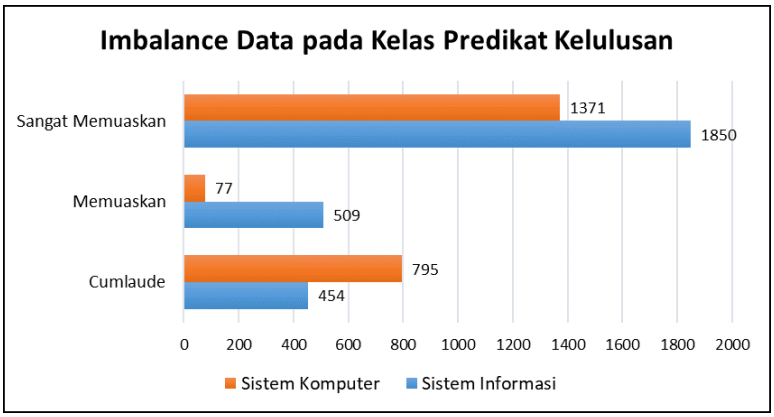


Gambar 4. Grafik hasil *training* dan *testing* pada kelas lama studi.

3.2. Pembahasan

Seperti yang telah dijelaskan pada bagian sebelumnya di mana kinerja *classifier* Naïve Bayes tidak cukup baik pada kelas Predikat Kelulusan baik itu dari kelompok data Program Studi Sistem Komputer ataupun Program Studi Sistem Informasi. Berdasarkan studi literatur yang dilakukan [11][12][13], didapatkan bahwa salah satu penyebab dari permasalahan ini adalah adanya ketidakseimbangan (pada bagian selanjutnya digunakan istilah *imbalanced*) data pada nilai kelas yang diklasifikasikan, yaitu Predikat Kelulusan. Gambar 5 memperjelas permasalahan *imbalanced* ini, di mana ada ketidakseimbangan antara nilai *Cumlaude*, Memuaskan, dan Sangat Memuaskan. Jumlah data dengan nilai Sangat Memuaskan jauh lebih banyak daripada nilai Memuaskan dan juga *Cumlaude*.

Imbalanced pada data menyebabkan adanya kelas mayoritas dan juga minoritas. Hal ini tentu saja bermasalah karena *classifier* akan cenderung lebih banyak belajar pada pola data mayoritas daripada yang minoritas, sehingga keputusan yang dihasilkan oleh *classifier* akan menjadi bias, sehingga metrik CA tidak dapat lagi digunakan sebagai satu-satunya metrik yang digunakan sebagai penentu kinerja *classifier* [14]. Berdasarkan [12], terdapat beberapa teknik pendekatan yang dapat dilakukan untuk mengatasi permasalahan *imbalanced* data seperti pada penelitian ini, pertama berdasarkan pendekatan dekomposisi, yaitu mengubah permasalahan *multi-class* menjadi *binary-class*. Kedua, pendekatan *ad-hoc*, yaitu pendekatan untuk mengatasi masalah *imbalanced* mulai dari proses *preprocessing* sampai dengan proses klasifikasinya. Namun untuk kedua teknik pendekatan tersebut di luar dari lingkup penelitian ini sehingga tidak dapat dijelaskan lebih lanjut dengan hasil penelitian.



Gambar 5. *Imbalanced* data pada kelas predikat kelulusan.

4. Kesimpulan

Berdasarkan hasil penelitian maka didapat kesimpulan sebagai berikut:

1. Perbandingan hasil *training* dan *testing* di semua metrik pengukuran. menunjukkan bahwa hasil pengukuran *training* dan *testing* tidak jauh berbeda, bahkan jika dilihat secara keseluruhan hasil *training* dan *testing* cenderung menunjukkan hasil yang sama.
2. Jika dilihat dari perbandingan antar kelas pada masing-masing kelompok data, terdapat perbedaan pada hasil pengukuran. Berdasarkan kelima metrik pengukuran, kelas Lama Studi memiliki hasil pengukuran yang lebih baik daripada kelas Predikat Kelulusan, di mana nilai metrik pada kelas Predikat Kelulusan tidak ada yang mencapai nilai 0,7.
3. Kinerja *classifier* Naïve Bayes tidak cukup baik pada kelas Predikat Kelulusan baik itu dari kelompok data Program Studi Sistem Komputer ataupun Program Studi Sistem Informasi, hal ini dikarenakan adanya *imbalanced* data.
4. *Imbalanced* pada data menyebabkan adanya kelas mayoritas dan juga minoritas, hal ini menyebabkan *classifier* cenderung lebih banyak belajar pada pola data mayoritas daripada yang minoritas, sehingga keputusan yang dihasilkan oleh *classifier* akan menjadi bias.

Daftar Pustaka

[1] BAPPEDA, "Data Koperasi dan UKM," *Dataku*, 2020. [Online]. Available: bappeda.jogjaprov.go.id/dataku/data_dasar?id_skpd=18.

[2] T. T. Maskoen and D. Purnama, "Area Under the Curve dan Akurasi Cystatin C untuk Diagnosis Acute Kidney Injury pada Pasien Politrauma," *Maj. Kedokt. Bandung*, vol. 50, no. 4, pp. 259–264, Dec. 2018.

[3] S. Syarli and A. Muin, "Metode Naive Bayes Untuk Prediksi Kelulusan (Studi Kasus: Data Mahasiswa Baru Perguruan Tinggi)," *J. Ilm. Ilmu Komput.*, vol. 2, no. 1, pp. 22–26, Apr. 2016.
[4] N. Y. Septian, "Data Mining Menggunakan Algoritma Naïve Bayes Untuk Klasifikasi Kelulusan Mahasiswa Universitas Dian Nuswantoro," *J. Semant. 2013*, pp. 1–11, 2009.
[5] S. Salmu and A. Solichin, "Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naïve Bayes : Studi Kasus UIN Syarif Hidayatullah Jakarta," in *Seminar Nasional Multidisiplin Ilmu (SENMI) 2017*, 2017, no. April, pp. 701–709.
[6] 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, Apr. 2019.
[7] P. A. Lizsara, S. Oyama, and S. Wardani, "Implementasi Data Mining Menggunakan Metode Naïve Bayes Untuk Memprediksi Ketepatan Waktu Tingkat Kelulusan Mahasiswa (Study Kasus: Program Studi Informatika Universitas PGRI Yogyakarta)," *Seri Pros. Semin. Nas. Din. Inform.*, vol. 4, no. 1, pp. 34–37, Apr. 2020.
[8] S. Rahmatullah, "Prediksi Tingkat Kelulusan Tepat Waktu dengan Metode Naïve Bayes dan K-Nearest Neighbor," *J. Inf. dan Komput.*, vol. 7, no. 1, pp. 7–16, Apr. 2019.
[9] P. S. C. Moonallika, K. Q. Fredlina, and I. B. K. Sudiatmika, "Penerapan Data Mining Untuk Memprediksi Kelulusan Mahasiswa Menggunakan Algoritma Naive Bayes Classifier (Studi Kasus STMIK Primakara)," *J. Ilm. Komput.*, vol. 6, no. 1, pp. 47–56, Feb. 2020.
[10] L. A. Hunt, "Missing data imputation and its effect on the accuracy of classification," in *Studies in Classification, Data Analysis, and Knowledge Organization*, 2017, no. 195089, pp. 3–14.
[11] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Springer International Publishing, 2018.
[12] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, "Imbalanced Classification with Multiple Classes," in *Learning from Imbalanced Data Sets*, Springer International Publishing, 2018, pp. 197–226.
[13] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, "Foundations on Imbalanced Classification," in *Learning from Imbalanced Data Sets*, Springer International Publishing, 2018, pp. 19–46.
[14] S. Fotouhi, S. Asadi, and M. W. Kattan, "A comprehensive data level analysis for cancer diagnosis on imbalanced data," *J. Biomed. Inform.*, vol. 90, Feb. 2019.