

PREDIKSI MAHASISWA DROP-OUT DI UNIVERSITAS XYZ

Tubagus Ahmad Marzuqi^{*1}, Evelline Kristiani², Marcel³

^{1,2,3}Universitas Kristen Krida Wacana, Jakarta Barat

Email: ¹tubagus.marzuqi@ukrida.ac.id, ²evelline.kristiani@ukrida.ac.id, ³marcel@ukrida.ac.id

^{*}Penulis Korespondensi

(Naskah masuk: 31 Januari 2024, diterima untuk diterbitkan: 22 November 2024)

Abstrak

Akreditasi dan reputasi merupakan faktor krusial bagi setiap perguruan tinggi, termasuk Universitas XYZ. Salah satu hal yang dapat memengaruhi akreditasi adalah jumlah mahasiswa yang mengalami drop-out (DO). Untuk mencegah penurunan akreditasi dan reputasi akibat masalah tersebut, penelitian ini berfokus pada pengembangan model prediksi mahasiswa DO. Algoritma Random Forest, Gradient Boosting, dan Decision Tree digunakan untuk mengevaluasi seberapa akurat model klasifikasi dalam memprediksi potensi mahasiswa DO berdasarkan data akademik. Sebelum membangun model, digunakan metode SMOTE untuk menangani masalah ketidakseimbangan data. Hasil penelitian menunjukkan bahwa model prediksi menggunakan algoritma Random Forest berhasil mencapai akurasi sebesar 99,67%. Algoritma Gradient Boosting menghasilkan akurasi 99,21%, sementara Decision Tree mencapai akurasi sebesar 98,67%. Selain mengukur akurasi model, penelitian ini juga mengidentifikasi faktor-faktor utama yang berkontribusi terhadap prediksi mahasiswa DO. Faktor-faktor tersebut meliputi adanya tunggakan pembayaran, IPK rata-rata di bawah 2, jumlah mata kuliah yang diulang lebih dari satu kali, dan kegagalan dalam melakukan KRS lebih dari dua kali. Penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam bidang prediksi akademik, khususnya dalam upaya mengurangi tingkat mahasiswa drop-out (DO) di perguruan tinggi.

Kata kunci: *Prediksi, Deteksi, Mahasiswa DO, SMOTE, Decision Tree, Random Forest, dan Gradient Boosting*

PREDICTION OF STUDENT DROP-OUT AT XYZ UNIVERSITY

Abstract

Accreditation and reputation are critical factors for higher education institutions, including XYZ University. One factor that can negatively impact accreditation is the number of students who drop out (DO). To prevent a decline in accreditation and reputation due to this issue, this study aims to develop a predictive model for student dropouts. The Random Forest, Gradient Boosting, and Decision Tree algorithms were utilized to evaluate the accuracy of classification models in predicting potential dropouts using academic baseline data. Prior to model building, the SMOTE method was applied to address the issue of imbalanced data. The results indicate that the predictive model using the Random Forest algorithm achieved an accuracy of 99.67%. The Gradient Boosting algorithm yielded an accuracy of 99.21%, while the Decision Tree algorithm achieved 98.67% accuracy. In addition to determining model accuracy, this study also identified key factors contributing to the prediction of student dropouts. These factors include outstanding payment history, a GPA below 2.0, repeating courses more than once, and failing to register for courses (KRS) more than twice. This research is expected to make a significant contribution to the field of academic prediction, particularly in efforts to reduce the dropout (DO) rate among university students.

Keywords: *Prediction, Detection, Drop-out Students, SMOTE, Decision Tree, Random Forest, and Gradient Boosting*

1. PENDAHULUAN

Mahasiswa *drop-out* (DO) merupakan salah satu masalah yang cukup mengganggu perguruan tinggi di Indonesia (Mutrofin et al., 2019). Sebab kondisi tersebut akan berdampak pada penilaian akreditasi serta pemeringkatan klasterisasi universitas. Lebih lanjut lagi, dampak tersebut

seterusnya mempengaruhi reputasi instansi yang bersangkutan dihadapan pemangku kepentingan (stakeholders) internal dan eksternal (Perez, Castellanos and Correal, 2018).

Berkaitan dengan hal tersebut, maka dianggap penting untuk dapat melakukan pencegahan sebelum masalah terjadi. Namun, Universitas XYZ sebagai

subjek penelitian ini belum memiliki sistem untuk memprediksi mahasiswa yang berpotensi akan DO. Salah satu kasusnya ialah didapati beberapa kali ada mahasiswa yang sebenarnya sudah melewati masa studi tapi data status mahasiswanya masih disebutkan aktif.

Berdasarkan latar masalah, melalui penelitian ini akan dibuatkan sebuah model untuk prediksi mahasiswa DO secara otomatis. Model ini diharapkan bisa membantu universitas mengetahui lebih dini mahasiswa yang memiliki potensi terkena DO. Data yang digunakan adalah data informasi dasar akademik seperti Status Bayar, IPK, Jumlah mengulang Mata kuliah, Jumlah Tidak KRS, serta Asal Sekolah (SMA) Mahasiswa (Purnamasari, Rini and Sukemi, 2019; Samuel, Hutapea and Jonathan, 2019).

Algoritma *Random Forest & Gradient Boosting* yang akan digunakan pada penelitian ini. Karena penelitian terkait sebelumnya, hasil akurasi mencapai 92% untuk *Random Forest* & 93% untuk *Gradient Boosting* (Tenpipat and Akkarajitsakul, 2020) Serta satu algoritma *Decision Tree* karena cukup populer digunakan (Purnamasari, Rini and Sukemi, 2019; Samuel, Hutapea and Jonathan, 2019; Solichin, 2019; Tenpipat and Akkarajitsakul, 2020). Pada penelitian ini juga akan menggunakan SMOTE Berdasarkan penelitian (Samuel, Hutapea and Jonathan, 2019)(Sutarman et al., 2024) untuk menanggulangi data yang tidak seimbang. SMOTE digunakan karena telah terbukti untuk mengatasi masalah ketidakseimbangan pada sebuah label/kelas (Fernández et al., 2018).

Berdasarkan uraian yang dijelaskan, penulis merumuskan pertanyaan penelitian yaitu:

- Seberapa akurat model klasifikasi pada status terakhir mahasiswa *drop-out* (DO) dapat diprediksi menggunakan informasi dasar akademik?
- Berdasarkan pertanyaan penelitian pertama, faktor-faktor apa saja yang mempengaruhi prediksi mahasiswa *drop-out* (DO)?

Meskipun pembelajaran mesin telah banyak digunakan dalam prediksi mahasiswa *drop-out* (DO) pada penelitian-penelitian sebelumnya, kebaruan utama dari penelitian ini terletak pada penerapan metode SMOTE untuk menangani ketidakseimbangan distribusi label dalam dataset. Ketidakseimbangan ini sering kali menyebabkan bias pada model prediksi. Selain itu, penelitian ini menawarkan pendekatan analitis yang komprehensif dalam mengidentifikasi berbagai faktor penentu DO. Model prediktif yang dikembangkan memberikan institusi pendidikan tinggi untuk mengidentifikasi mahasiswa yang berisiko DO, sehingga memungkinkan penerapan intervensi dini yang lebih efektif.

Hasil Penelitian yang diharapkan dapat memberikan kontribusi yang signifikan dalam upaya mengurangi tingkat mahasiswa *drop-out* (DO) di perguruan tinggi. Hasil penelitian juga diharapkan membantu pihak universitas dalam mengambil langkah-langkah pencegahan yang lebih proaktif, seperti memberikan intervensi akademik atau bimbingan khusus kepada mahasiswa yang terdeteksi berisiko tinggi.

2. METODE PENELITIAN

Pada bagian ini menjelaskan mengenai metodologi penelitian yang terdiri dari desain penelitian, tahapan penelitian, dan metode yang akan digunakan.

2.1. DESAIN PENELITIAN

Jenis penelitian ini adalah penelitian kuantitatif. Karakteristik dari penelitian kuantitatif adalah mempertimbangkan hubungan antar variabel yang dapat diukur secara numerik (Saunders, Lewis and Thornhill, 2019). Selain itu, penelitian kuantitatif juga menggunakan paradigma positivisme sehingga menjanjikan pengetahuan yang tidak ambigu karena tidak dipengaruhi oleh interpretasi maupun hasil bias dari manusia.

Selain itu, penelitian ini juga merupakan *Supervised Learning* karena berkaitan dengan prediksi mahasiswa yang rentan DO. *Supervised Learning* adalah metode yang mencoba menemukan hubungan antara atribut input dan atribut target (atau biasa disebut label) (Shalev-Shwartz & Ben-David, 2014). Dalam *Supervised Learning*, dataset dilatih dengan training set, dan variabel input yang ada dalam dataset akan mempengaruhi nilai akurasi dari variabel prediksi (Wang et al., 2021).

2.2. TAHAPAN PENELITIAN

Berikut merupakan tahapan-tahapan pada penelitian ini:

- Langkah yang pertama adalah melakukan identifikasi masalah yang diangkat, yakni mahasiswa DO di Universitas XYZ. Kondisi saat ini, pengecekan mahasiswa yang rentan dan sudah melewati batas studi masih dilaksanakan manual sehingga memerlukan waktu lebih lama dan rentan terjadi kesalahan dalam pengecekan. Hasil dari analisis menunjukkan bahwa dibutuhkan sistem prediksi yang dapat secara otomatis mengantisipasi kondisi mahasiswa yang rentan DO
- Langkah selanjutnya dilakukan tinjauan pustaka untuk menentukan kerangka teoritis dalam penelitian ini.
- Melakukan penarikan data menggunakan *query* dan penulisan kode PHP untuk menarik serta mengolah data IPK, pembayaran karena

berbentuk presentase, data jumlah mengulang mata kuliah dan data jumlah tidak KRS serta label Lulus & DO.

- Data yang didapatkan sudah terstruktur, tetapi masih terdapat data yang kosong pada beberapa kolom serta terjadi kemungkinan terdapat data yang tidak seimbang. Maka perlu dilakukan pembersihan data terlebih dahulu. Metode yang dilakukan adalah *mean* untuk data kuantitatif dan modus untuk data kualitatif. Metode ini dipilih untuk memastikan kelengkapan dataset tanpa menghapus data yang berpotensi penting.
- Sebelum melakukan proses *classification model*, data akan dibagi menjadi dua untuk *training data* dan *testing data*. Pembagian tersebut akan menggunakan *10-fold cross validation*.

2.3. METODE & DATA YANG DIGUNAKAN

Penelitian ini menggunakan aplikasi *Google Collab* dan bahasa pemrograman *Python* sebagai *tools*. Algoritma yang digunakan adalah *Decision Tree*, *Random Forest*, dan *Gradient Boosting*. Data yang digunakan dalam penelitian ini mencakup variabel-variabel akademik dan administratif dari mahasiswa yang meliputi:

1. Status bayar (kualitatif) mencerminkan status keuangan mahasiswa, dengan tiga kategori yaitu hutang, deposit, dan lunas.
2. Indeks prestasi kumulatif (IPK) (kuantitatif) digunakan sebagai ukuran kinerja akademik mahasiswa.
3. Jumlah mengulang mata kuliah (kuantitatif) merupakan jumlah total mata kuliah yang diulang oleh mahasiswa.
4. Jumlah tidak KRS (kuantitatif) mengenai frekuensi mahasiswa tidak melakukan pengisian Kartu Rencana Studi (KRS) selama beberapa semester.
5. Asal sekolah (SMA) (kualitatif) atas mahasiswa, yang dikategorikan berdasarkan nama sekolah.
6. Label DO/Lulus (kualitatif) menunjukkan status akhir mahasiswa, apakah mereka telah *drop-out* (DO) atau lulus. Label ini berfungsi sebagai target variabel dalam model klasifikasi.

Sebelum membangun model prediksi, dilakukan beberapa langkah *preprocessing*. Langkah-langkah tersebut meliputi:

1. Penanganan data kualitatif seperti Status Bayar, Asal Sekolah dan Label DO/Lulus diolah dengan menggunakan metode label encoding dan one-hot encoding untuk mengonversi nilai kualitatif menjadi representasi numerik yang dapat diproses oleh algoritma pembelajaran mesin. Sebagai contoh, variabel Status Bayar diubah menjadi numerik: hutang = 0, deposit = 1, dan lunas = 2. Label DO/Lulus diubah menjadi numerik: DO = 1, Lulus = 0. Sementara untuk variabel Asal Sekolah, one-hot encoding diterapkan, di mana

setiap kategori sekolah diubah menjadi kolom biner terpisah.

2. Penanganan ketidakseimbangan pada data antara mahasiswa yang DO dan yang lulus, kami menggunakan metode SMOTE (*Synthetic Minority Over-sampling Technique*) untuk membuat distribusi data lebih seimbang.
3. Normalisasi data kuantitatif untuk memastikan bahwa semua fitur kuantitatif, seperti IPK, jumlah mengulang mata kuliah, dan jumlah tidak KRS, berada pada skala yang seragam, dilakukan normalisasi menggunakan metode min-max scaling. Hal ini penting untuk mencegah salah satu variabel mendominasi proses pembelajaran model karena skala yang berbeda.
4. Data yang tidak lengkap atau hilang diidentifikasi, dan dilakukan imputasi menggunakan metode *mean* untuk data kuantitatif dan modus untuk data kualitatif.

3. PENELITIAN TERKAIT

Terdapat beberapa penelitian sebelumnya yang memiliki keterkaitan dengan prediksi maupun deteksi mahasiswa DO. Penelitian "*Student Dropout Prediction: A KMUTT Case Study*" (Tenpipat and Akkarajitsakul, 2020) meneliti mengenai faktor-faktor yang mempengaruhi status pendidikan sarjana dan membuat model klasifikasi untuk memprediksi status pendidikan mahasiswa apakah akan DO atau tidak ketika mereka belajar di Universitas Teknologi Raja Mongkut Thonburi (KMUTT), Bangkok, Thailand. Secara khusus, data yang digunakan dalam analisis adalah 13.714 mahasiswa sarjana di 7 tahun akademik yaitu dari tahun akademik 2012-2013 hingga 2019-2020. Hasil penelitian menunjukkan bahwa akurasi prediksi model *Gradient Boosting*, *Decision Tree* dan *Random Forest* masing-masing adalah 93%, 92%, dan 92%.

Penelitian lainnya adalah "*Comparison of Decision Tree, Naïve Bayes and KNearest Neighbors for Predicting Thesis Graduation*" (Solichin, 2019). Penelitian ini untuk mengetahui faktor-faktor yang mempengaruhi kelulusan mahasiswa dalam skripsi di Universitas Budi Luhur (UBL), khususnya di Jurusan Teknik Informatika. Dataset diperoleh dari mahasiswa Jurusan Teknik Informatika yang mengambil skripsi pada tahun ajaran 2016/2017, dan 2017/2018. Algoritma yang digunakan adalah *Decision Tree*, *Naïve Bayes*, dan *k Nearest Neighbors* (kNN) dengan hasil *presicion* masing-masing adalah 75%, 73,16%, 80,32%.

Pada penelitian terkait lainnya "*Prediction of the Student Graduation's Level using C4.5 Decision Tree Algorithm*" (Purnamasari, Rini and Sukemi, 2019). Penelitian untuk memprediksi berdasarkan durasi kelulusan siswa untuk menyelesaikan studinya menggunakan algoritma *Decision Tree*. Data yang diambil untuk mendapatkan nilai prediksi tingkat kelulusan mahasiswa berasal dari sistem akademik

yang perlu diolah lebih lanjut. 69,79% didapatkan hasil akurasi pada penelitian ini.

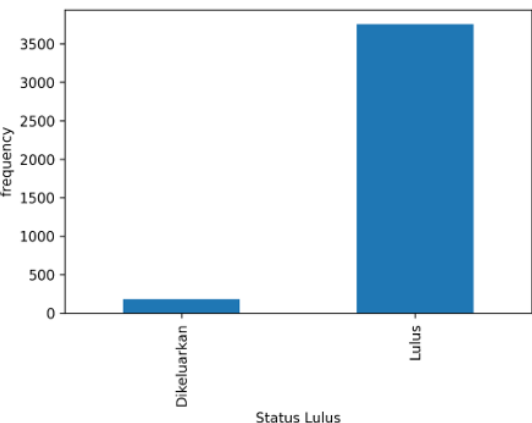
Selanjutnya, pada penelitian “Predicting the Timeliness of Student Graduation Using Decision Tree C4.5 Algorithm in Universitas Advent Indonesia” (Samuel, Hutapea and Jonathan, 2019). Tujuan penelitian ini memprediksi mahasiswa akan lulus tepat waktu atau tidak. Metode yang digunakan untuk prediksi adalah *Decision Tree C4.5*. Penelitian ini juga menggunakan SMOTE (*Synthetic Minority Oversampling Technique*). Hasilnya Penggunaan SMOTE meningkatkan nilai presisi dan recall.

Lewat studi terhadap penelitian-penelitian tersebut, ditentukan algoritma-algoritma yang akan digunakan. Berdasarkan popularitas, *Decision Tree* merupakan yang paling banyak digunakan. Namun untuk akurasi, satu penelitian yang memiliki hasil akurasi lebih dari 90% adalah Tenpipat & Akkarajitsakul tahun 2020 (Tenpipat & Akkarajitsakul, 2020). Sehingga penelitian ini akan menggunakan algoritma *Decision Tree*, *Random Forest* dan *Gradient Boosting* sebagai pembandingan hasil dari prediksi mahasiswa DO. Selain itu, untuk menanggulangi data yang tidak seimbang, penelitian juga akan menggunakan SMOTE (Samuel, Hutapea and Jonathan, 2019). Algoritma SMOTE dianggap sebagai standar *framework of learning* dari data yang tidak seimbang (Fernández et al., 2018).

4. ANALISIS DAN HASIL

Pada bab ini menjelaskan mengenai karakteristik data, pembentukan model dan melakukan *balancing data*. Pada bab ini juga dilakukan perbandingan algoritma untuk menentukan algoritma terbaik dalam mendeteksi mahasiswa DO.

4.1. KARAKTERISTIK DATASET



Gambar 1. Confussion Matrix dari Model *Gradient Boosting* Deteksi Mahasiswa DO

Total data yang didapatkan dengan kategori status lulus yang diambil dari angkatan 2011 sampai 2018 adalah berjumlah 3939 data. Jumlah data pada label “Lulus” adalah 3756 data, sedangkan pada label “Dikeluarkan”/DO berjumlah 183 data. Gambar 1 menunjukkan grafik pada dataset untuk deteksi

mahasiswa DO. Grafik menunjukkan ketidakseimbangan data yang cukup signifikan. Berdasarkan data tersebut akan menggunakan SMOTE untuk mengatasi ketidakseimbangan pada data.

4.2. HASIL MODEL PREDIKSI MAHASISWA

Penjelasan mengenai hasil model prediksi mahasiswa DO yang terdiri dari penggunaan SMOTE dan tanpa penggunaan SMOTE pada seluruh confusion matrix dan hasil performa. Kelas positif dianalogikan sebagai label “Dikeluarkan” dan kelas negatif dianalogikan sebagai label “Lulus” pada tabel *confussion matrix* yang ditampilkan.

Tabel 1. Confussion Matrix Dari Model *Decision Tree* Deteksi Mahasiswa DO

Decision Tree		No SMOTE	
		Predicted	
		Lulus	Dikeluarkan
Actual	Lulus	3734	22
	Dikeluarkan	18	165
SMOTE			
		Predicted	
		Lulus	Dikeluarkan
Actual	Lulus	3734	22
	Dikeluarkan	30	153

Tabel 1 menunjukkan confusion matrix dari model yang menggunakan algoritma *Decision Tree*. Hasilnya menunjukkan True Positif (TP) sebesar 165 data, True Negatif (TN) sebesar 3734 data, False Positif (FP) sebesar 22 data, dan False Negatif (FN) sebesar 18 data ketika tanpa menggunakan SMOTE. Setelah penggunaan SMOTE diterapkan, terdapat penurunan data pada TP sebesar 153 data. TN dan FP tidak ada perubahan data, sama seperti saat tidak menggunakan SMOTE. FN mengalami kenaikan hasil sebesar 30 data.

Tabel 2. Confussion Matrix Dari Model *Random Forest* Deteksi Mahasiswa DO

Random Forest		No SMOTE	
		Predicted	
		Lulus	Dikeluarkan
Actual	Lulus	3756	0
	Dikeluarkan	11	172
SMOTE			
		Predicted	
		Lulus	Dikeluarkan
Actual	Lulus	3755	1
	Dikeluarkan	12	171

Confussion matrix dari model yang menggunakan algoritma *Random Forest* ditunjukkan pada Tabel 2. Hasilnya menunjukkan True Positif (TP) sebesar 172 data, True Negatif (TN) sebesar 3756 data, False Positif (FP) sebesar 0 data, dan False Negatif (FN) sebesar 11 data ketika tanpa menggunakan SMOTE. Setelah penggunaan SMOTE diterapkan, terdapat penurunan data pada TP sebesar 171 data. FN dan FP mengalami kenaikan sebesar 12 data dan 1 data. TN mengalami penurunan hasil sebesar 3755 data.

Tabel 3. Confussion Matrix Dari Model *Gradient Boosting* Deteksi Mahasiswa DO

Gradient Boosting			
No SMOTE			
		<i>Predicted</i>	
		Lulus	Dikeluarkan
<i>Actual</i>	Lulus	3751	5
	Dikeluarkan	12	171
SMOTE			
		<i>Predicted</i>	
		Lulus	Dikeluarkan
<i>Actual</i>	Lulus	3735	21
	Dikeluarkan	10	173

Tabel 3 menunjukkan confusion matrix dari model yang menggunakan algoritma *Gradient Boosting*. Hasilnya menunjukkan True Positif (TP) sebesar 171 data, True Negatif (TN) sebesar 3751 data, False Positif (FP) sebesar 5 data, dan False Negatif (FN) sebesar 12 data ketika tanpa menggunakan SMOTE. Setelah penerapan SMOTE, terdapat kenaikan data pada TP sebesar 173 data dan FP sebesar 21 data. TN dan FN mengalami penurunan hasil sebesar 3735 data dan 10 data.

Tabel 4. Hasil Performa Dari Seluruh Model Deteksi Mahasiswa Do (No SMOTE)

No SMOTE				
	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>Average Precision</i>
<i>Decision Tree</i>	98.98%	88.23%	90.16%	80.01%
<i>Random Forest</i>	99.72%	100%	93.98%	98.30%
<i>Gradient Boosting</i>	99.56%	97.15%	93.44%	88.44%

Tabel 5. Hasil Performa Dari Seluruh Model Deteksi Mahasiswa Do (SMOTE)

SMOTE				
	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>Average Precision</i>
<i>Decision Tree</i>	98.67%	87.43%	83.60%	73.86%
<i>Random Forest</i>	99.67%	99.42%	93.44%	98.58%
<i>Gradient Boosting</i>	99.21%	89.17%	94.53%	90.99%

Hasil performa model dari seluruh algoritma yang digunakan untuk deteksi mahasiswa DO tanpa menggunakan SMOTE dan dengan menggunakan SMOTE dapat dilihat pada Tabel 4 & Tabel 5. Hasil

performa yang akan dijadikan acuan adalah dengan penggunaan SMOTE, berikut merupakan hasilnya:

- Hasil performa dari algoritma *Decision Tree* menunjukkan nilai *accuracy* 98,67%, *precision* 87,43%, *recall* 83,60%, dan *average precision* 73,86%.
- Algoritma *Random Forest* menunjukkan nilai *accuracy* 99,67%, *precision* 99,42%, *recall* 93,44%, dan *average precision* 98,58%.
- Algoritma *Gradient Boosting* menunjukkan nilai *accuracy* 99,21%, *precision* 89,17%, *recall* 94,53%, dan *average precision* 90,99%.

4.3. ANALISIS FAKTOR

Tabel 6 menunjukkan beberapa sampel dari hasil deteksi mahasiswa DO yang telah dilakukan. Pada sampel data dan juga keseluruhan data, hanya terdapat 1 mahasiswa yang terprediksi DO tetapi secara aktual lulus. Hal tersebut terjadi kemungkinan karena jumlah ulang mata kuliah lebih dari 0. Sampel selanjutnya adalah mahasiswa yang terprediksi lulus tetapi aktualnya DO. Kolom jumlah ulang mata kuliah pada sampel tersebut menunjukkan angka 0 serta status bayar lunas dan deposit. Kemungkinan terjadinya prediksi yang salah adalah pada kolom jumlah ulang mata kuliah. Berdasarkan hasil deteksi mahasiswa DO, faktor-faktor yang menyebabkan mahasiswa DO adalah status bayar karena terdapat mahasiswa yang hutang terprediksi DO, IPK dengan rata-rata di bawah 2 diprediksi DO, jumlah ulang mata kuliah di atas 1, tidak KRS di atas 2.

Tabel 6. Hasil Prediksi Mahasiswa DO Dengan Algoritma Terbaik

ID	Status Bayar	IPK	Ulang Matkul	Tidak Asal KRS Sekolah	Label	Prediksi
5943	Lunas	3.65	0	0 SMAN 1 KELAM PERMAI KALBAR	Lulus	Lulus
5745	Deposit	3.54	0	0 SMA TIARA KASIH	Lulus	Lulus
4622	Lunas	2.65	2	0 SMA MARSUD IRINI	Lulus	DO
7233	Lunas	1.7	0	3 SMA NEGERI 2 SENDAWAR - KALTIM	DO	Lulus
6033	Deposit	0.94	0	2 SMA KATOLIK GIOVANI KUPANG	DO	Lulus
4271	Lunas	1.68	15	0 SMA PRESIDEN	DO	DO
1457	Deposit	1.57	30	3 SMA NEGERI 1 LANGOWAN	DO	DO
4239	Hutang	1.38	12	0 SMA NEGERI 3 MAKALE	DO	DO

5. KESIMPULAN

Penelitian ini memiliki tujuan menganalisis model dalam membuat prediksi mahasiswa DO di Universitas XYZ. Hasil model tersebut diharapkan dapat menjadi pertimbangan untuk Universitas XYZ dapat memiliki sistem prediksi mahasiswa DO.

Berdasarkan hasil eksperimen yang dianalisis adalah sebagai berikut:

- Berdasarkan data yang digunakan untuk melakukan prediksi mahasiswa DO berhasil mencapai 99.67% pada *metric accuracy*, 99,42% pada *metric precision* dan *average precision* 98,58% dengan algoritma *Random Forest*.
- Faktor-faktor yang memengaruhi mahasiswa DO adalah status bayar karena terdapat mahasiswa yang hutang terprediksi DO, IPK dengan rata-rata di bawah 2 diprediksi DO, jumlah ulang mata kuliah di atas 1, tidak KRS di atas 2.

Penerapan SMOTE dalam penelitian ini secara efektif mengatasi masalah ketidakseimbangan representasi kasus drop-out (DO) dalam dataset akademik, sehingga meningkatkan performa model prediktif. Keunikan penelitian ini terletak pada penerapan metode SMOTE yang diintegrasikan dengan analisis mendalam terhadap faktor-faktor utama yang memengaruhi risiko DO mahasiswa, sehingga memberikan wawasan yang lebih komprehensif. Model prediktif yang dihasilkan memungkinkan institusi pendidikan tinggi untuk mengoptimalkan proses pengambilan keputusan terkait strategi dukungan dan retensi mahasiswa secara lebih efektif.

Saran untuk penelitian selanjutnya pada prediksi mahasiswa DO yang datanya sangat tidak seimbang mungkin dapat menggunakan teknik lain selain SMOTE sebagai pembanding untuk hasil akurasi.

DAFTAR PUSTAKA

- FERNÁNDEZ, A., GARCÍA, S., HERRERA, F., CHAWLA, N. V., 2018. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, .
- MUTROFIN, S., KHALIMI, A.M., KURNIAWAN, E., GINARDI, R.V.H., FATICAH, C., SARI, Y.A., 2019. Detection of Potentially Students Drop Out of College in Case of Missing Value Using C4.5. In: *2019 International Conference on Sustainable Engineering and Creative Computing (ICSECC)*. pp.349–354. <https://doi.org/10.1109/ICSECC.2019.8907014>.
- PEREZ, B., CASTELLANOS, C., CORREAL, D., 2018. Applying Data Mining Techniques to Predict Student Dropout: A Case Study. In: *2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI)*. pp.1–6. <https://doi.org/10.1109/ColCACI.2018.8484847>.
- PURNAMASARI, E., RINI, D.P., SUKEMI, 2019. Prediction of the Student Graduation's Level using C4.5 *Decision Tree* Algorithm. In: *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*. pp.192–195. <https://doi.org/10.1109/ICECOS47637.2019.8984493>.
- SAMUEL, Y.T., HUTAPEA, J.J., JONATHAN, B., 2019. Predicting the Timeliness of Student Graduation Using *Decision Tree* C4.5 Algorithm in Universitas Advent Indonesia. In: *2019 12th International Conference on Information & Communication Technology and System (ICTS)*. pp.276–280. <https://doi.org/10.1109/ICTS.2019.885094>.
- SAUNDERS, M.N.K., LEWIS, P., THORNHILL, A., 2019. *Research Methods for Business Students. Qualitative Market Research: An International Journal*, <https://doi.org/10.1108/qmr.2000.3.4.215.2>.
- SHALEV-SHWARTZ, S., BEN-DAVID, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.
- SOLICHIN, A., 2019. Comparison of *Decision Tree*, Naïve Bayes and K-Nearest Neighbors for Predicting Thesis Graduation. In: *2019 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*. pp.217–222. <https://doi.org/10.23919/EECSI48112.2019.8977081>.
- SUTARMAN, SIRINGORINGO, R., ARISANDI, D., KURNIAWAN, E., NABABAN, E. B. (2024). Model Klasifikasi Dengan Logistic Regression Dan Recursive Feature Elimination Pada Data Tidak Seimbang. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, 11(4), 735–742. <https://doi.org/10.25126/jtiik.1148198>
- TENPIPAT, W., AKKARAJITSAKUL, K., 2020. Student Dropout Prediction: A KMUTT Case Study. In: *2020 1st International Conference on Big Data Analytics and Practices, IBDAP 2020*. [online] Institute of Electrical and Electronics Engineers Inc. <https://doi.org/10.1109/IBDAP50342.2020.9245457>.
- WANG, L., ZHANG, Z., ZHANG, X., ZHOU, X., WANG, P., ZHENG, Y. (2021). Chapter One - A Deep-forest based approach for detecting fraudulent online transaction. In A. R. Hurson & S. Wu (Eds.), *AI and Cloud Computing* (Vol. 120, pp. 1–38). Elsevier. <https://doi.org/https://doi.org/10.1016/bs.adcom.2020.10.001>