Jurnal Komputer dan Informatika
Vol 17 No 1, April 2022: hlm 23-29

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

PERBANDINGAN KINERJA METODE KLASIFIKASI UNTUK MEMPREDIKSI PUTUS SEKOLAH DAN KEBERHASILAN AKADEMIK SISWA

William
Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara
Jl. Letjen S.Parman No.1, Jakarta Barat, DKI Jakarta, Indonesia 11410
e-mail: *william.535210013@stu.untar.ac.id*

ABSTRAK

Putus sekolah dan keberhasilan akademik siswa merupakan dua hal yang penting dalam pendidikan. Penelitian ini membandingkan kinerja metode klasifikasi untuk memprediksi putus sekolah dan keberhasilan akademik siswa. Metode klasifikasi yang digunakan adalah Random Forest Classifier, AdaBoost, Decision Tree, Logistic Regression, dan XGBoost. Dataset yang digunakan berasal dari perguruan tinggi yang memiliki 4424 sampel dengan 36 fitur dan 3 kelas. Hasil penelitian menunjukkan bahwa metode Random Forest Classifier memiliki kinerja terbaik dengan akurasi 76%, diikuti oleh XGBoost 76%, AdaBoost 74%, Logistic Regression 74%, dan Decision Tree 71%. Oleh karena itu, metode Random Forest Classifier dapat digunakan untuk memprediksi putus sekolah dan keberhasilan akademik siswa dengan lebih akurat. Namun, perlu dicatat bahwa meskipun semua metode klasifikasi yang digunakan dalam penelitian ini telah mengalami perbaikan kinerja melalui penggunaan teknik ADASYN dan penyetelan parameter, mereka masih menghadapi tantangan dalam mengidentifikasi dengan akurat kasus-kasus dalam salah satu kelas minoritas. Oleh karena itu, langkah selanjutnya yang perlu diambil adalah melakukan penelitian lebih lanjut untuk mengoptimalkan parameter dengan lebih cermat dan juga mempertimbangkan pendekatan lain yang dapat lebih lanjut meningkatkan kinerja model, seperti mempertimbangkan penambahan informasi tambahan yang mungkin ada dalam dataset.

Kata kunci: *Random Forest, Logistic Regression,* Algoritma *Boosting*, Prediksi Keberhasilan Akademik, Putus Sekolah.

ABSTRACT

Dropping out of school and students' academic success are two important things in education. This study compares the performance of classification methods for predicting school dropout and student academic success. The classification methods used are Random Forest Classifier, AdaBoost, Decision Tree, Logistic Regression, and XGBoost. The dataset used comes from universities which has 4424 samples with 36 features and 3 classes. The research results show that the Random Forest Classifier method has the best performance with an accuracy of 76%, followed by XGBoost 76%, AdaBoost 74%, Logistic Regression 74%, and Decision Tree 71%. Therefore, the Random Forest Classifier method can be used to predict school dropout and student academic success more accurately However, it should be noted that although all classification methods used in this study have experienced performance improvements through the use of ADASYN techniques and parameter tuning, they still face challenges in accurately identifying cases within one of the minority classes. Therefore, the next step that needs to be taken is to carry out further research to optimize parameters more carefully and also consider other approaches that can further improve model performance, such as considering the addition of additional information that may be present in the dataset.

Keywords: *Random Forest, Logistic Regression, Boosting Algorithm, Predict Academic Success, Dropout*.

1. PENDAHULUAN

Putus sekolah dan keberhasilan akademik siswa merupakan dua hal penting yang perlu diperhatikan dalam sistem pendidikan. Putus sekolah dapat berdampak negatif pada individu, keluarga, dan masyarakat. Sementara itu, keberhasilan akademik siswa merupakan indikator penting untuk keberhasilan pendidikan. Penyebab putus sekolah di Indonesia beragam, antara lain faktor ekonomi, faktor keluarga, faktor lingkungan, dan faktor individu. Faktor ekonomi merupakan faktor yang paling dominan. Faktor keluarga meliputi kurangnya perhatian orang tua, kekerasan dalam

rumah tangga, dan konflik keluarga. Faktor lingkungan meliputi kondisi lingkungan yang tidak mendukung pendidikan, seperti tingginya angka kriminalitas dan narkoba. Faktor individu meliputi motivasi belajar yang rendah, kurangnya minat belajar, dan gangguan belajar. Keberhasilan akademik siswa juga dipengaruhi oleh berbagai faktor, antara lain faktor internal dan faktor eksternal. Faktor internal meliputi kemampuan kognitif, motivasi belajar, dan minat belajar.

Faktor eksternal meliputi dukungan keluarga, dukungan guru, dan lingkungan belajar yang kondusif. Prediksi putus sekolah dan keberhasilan akademik siswa dapat dilakukan dengan menggunakan metode klasifikasi. Metode klasifikasi adalah metode yang digunakan untuk mengelompokkan data ke dalam dua atau lebih kelas. Metode klasifikasi yang digunakan untuk memprediksi putus sekolah dan keberhasilan akademik siswa antara lain metode *decision tree*, metode *logistic regression*, metode *boosting*, dan metode *random forest*.

Penelitian ini bertujuan untuk membandingkan kinerja metode klasifikasi untuk memprediksi putus sekolah dan keberhasilan akademik siswa. Penelitian ini menggunakan data dari *Polytechnic Institute of Portalegre* (IPP), Portugal. Data yang digunakan meliputi data siswa yang berisi variabel yang menggambarkan karakteristik demografi, sosial ekonomi, dan akademik siswa. Karakteristik demografi meliputi usia, jenis kelamin, status perkawinan, kewarganegaraan, kode alamat, dan kebutuhan khusus. Karakteristik sosial ekonomi meliputi pekerjaan pelajar, tempat tinggal orang tua, profesi orang tua, situasi pekerjaan orang tua, pelajar hibah, dan hutang siswa. Karakteristik akademik meliputi nilai masuk, tahun retensi di sekolah menengah atas, urutan pilihan mata pelajaran yang terdaftar, dan jenis mata pelajaran di sekolah menengah atas. Hasil penelitian ini diharapkan dapat memberikan informasi tentang metode klasifikasi yang paling efektif untuk memprediksi putus sekolah dan keberhasilan akademik siswa. Informasi ini dapat digunakan untuk mengembangkan program intervensi untuk mencegah putus sekolah dan meningkatkan keberhasilan akademik siswa.

2. METODE PENELITIAN

Bagian dibawah ini menyajikan data yang digunakan, metode yang digunakan untuk mengatasi ketidakseimbangan data, dan menggunakannya untuk membangun dan mengevaluasi model klasifikasi.

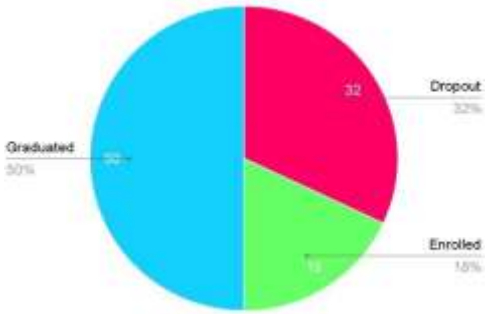
3.1 Dataset

Dalam studi ini, kami menggunakan data dari institusi yang terkait dengan mahasiswa yang mendaftar di program sarjana *Polytechnic Institute of Portalegre*, Portugal. Data ini mencakup informasi tentang mahasiswa yang terdaftar dari tahun akademik 2008/2009 hingga 2018/2019 dan berasal dari berbagai program sarjana seperti teknologi, layanan sosial, manajemen, jurnalisme, keperawatan, pendidikan, agronomi, desain. Data ini mencakup variabel-variabel yang berkaitan dengan faktor demografis, faktor sosial-ekonomi, serta jalur akademik mahasiswa [1].

Target data diklasifikasikan sebagai *Dropout*, *Enrolled*, dan *Graduated*, tergantung pada waktu yang diperlukan mahasiswa untuk lulus. *Dropout* berarti bahwa mahasiswa tidak berhasil lulus dalam rentang waktu yang ditentukan, *Graduated* berarti bahwa mahasiswa memerlukan waktu tambahan hingga tiga tahun untuk lulus, *Enrolled* berarti bahwa mahasiswa memerlukan lebih dari tiga tahun tambahan untuk lulus atau bahkan tidak lulus.

Yang memberikan tiga tingkat risiko: mahasiswa ‘risiko rendah' dengan probabilitas tinggi untuk lulus, mahasiswa 'risiko menengah/sedang', dengan probabilitas yang di mana jika diambil tindakan oleh institusi dapat berkontribusi pada keberhasilan untuk lulus, dan mahasiswa 'risiko tinggi', yang memiliki probabilitas tinggi untuk gagal. Pada Gambar 1, distribusi data di antara tiga kategori ini tidak seimbang, dengan dua kelas minoritas, yaitu *Dropout* dan *Enrolled*. *Dropout* mengisi 32% dari total data, dan *Enrolled* mengisi 18% dari total data, sedangkan kelas mayoritas, yaitu *Graduated* mengisi 50% dari total data.

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)



Gambar 1. Alur Penelitian

3.2 *Sampling Data Imbalanced*

Dalam penelitian ini, kami mengatasi masalah data yang tidak seimbang dengan menggunakan metode ADASYN (*Adaptive Synthetic Sampling*) yang telah terintegrasi dalam pustaka *imblearn* [2]. Data tidak seimbang adalah kondisi di mana terdapat perbedaan yang signifikan dalam jumlah data antara kelas mayoritas dan kelas minoritas, yang dapat menyebabkan bias dalam model pembelajaran mesin terhadap kelas mayoritas. Metode ADASYN bekerja dengan menciptakan data sintetik untuk kelas minoritas berdasarkan data terdekat dari kelas minoritas tersebut [3].

Keunikan metode ini adalah bahwa ia mengambil keseimbangan antara jumlah data sintetik yang harus dibuat dan tingkat kesulitan kelas minoritas dalam diklasifikasikan oleh model. Dengan menerapkan ADASYN, kami berhasil meningkatkan akurasi model pembelajaran mesin kami pada data yang tidak seimbang, karena metode ini memungkinkan model untuk lebih efektif belajar dari data minoritas yang sulit diklasifikasikan.

3.3 Algoritma Klasifikasi

Dalam studi ini menggunakan *scikit-learn* untuk algoritma *machine learning*. *Scikit-learn* adalah *library machine learning python* yang menyediakan berbagai algoritma klasifikasi [4].

Logistic regression adalah analisis statistik yang menggambarkan dan memperkirakan hubungan antara satu variabel dependen dan satu atau lebih variabel independen [5]. *Logistic regression* adalah metode analisis data yang digunakan untuk memprediksi probabilitas suatu peristiwa terjadi yang didasarkan pada model statistik yang menghubungkan variabel dependen (biasanya biner) dengan satu atau lebih variabel independen [6]. Algoritma ini bekerja dengan cara memodelkan fungsi logistik pada data, yang memetakan variabel input ke variabel output [7].

Decision Tree adalah jenis pengklasifikasi yang digunakan dalam pembelajaran mesin dan penambangan data yang memprediksi nilai variabel target berdasarkan beberapa variabel masukan [8]. *Decision tree* adalah model pembelajaran mesin yang digunakan untuk memprediksi kategori atau nilai variabel target berdasarkan nilai variabel input yang digambarkan sebagai struktur pohon, di mana setiap node mewakili keputusan, dan setiap cabang mewakili hasil dari keputusan tersebut [9]. Ini adalah model keputusan seperti pohon dan kemungkinan konsekuensinya, termasuk hasil kejadian yang tidak disengaja, biaya sumber daya, dan utilitas [10]. Algoritma ini dibangun dengan membagi kumpulan data menjadi sub-kumpulan yang lebih kecil secara rekursif berdasarkan nilai salah satu variabel masukan, dengan tujuan memaksimalkan homogenitas variabel target dalam setiap subset [11]. Pohon yang dihasilkan dapat digunakan untuk melakukan prediksi terhadap data baru dengan mengikuti jalur dari node akar ke node daun yang sesuai dengan nilai prediksi [12].

Random Forest adalah algoritma pembelajaran mesin yang menggunakan kumpulan *decision tree* untuk meningkatkan akurasi dan ketahanan model. Ini adalah jenis algoritma *bagging* yang menggabungkan beberapa *decision tree* untuk mengurangi *overfitting* dan meningkatkan kinerja generalisasi [13]. *Random forest* adalah metode ensemble untuk klasifikasi, regresi, dan tugas lainnya

yang beroperasi dengan membangun sejumlah besar pohon keputusan pada waktu pelatihan dan mengeluarkan kelas yang merupakan modus dari kelas (klasifikasi) atau rata-rata prediksi (regresi) dari pohon individu [14]. Dalam hutan acak, setiap pohon keputusan dilatih berdasarkan subset acak dari data pelatihan dan subset acak fitur. Keacakan ini membantu mengurangi korelasi antar pepohonan dan meningkatkan keanekaragaman *ensemble* [15].

Adaptive Boosting, juga dikenal sebagai *AdaBoost*, adalah algoritma pembelajaran mesin yang menggabungkan beberapa pengklasifikasi lemah untuk membuat pengklasifikasi yang kuat. Algoritma ini bekerja dengan melatih serangkaian pengklasifikasi lemah secara berulang pada kumpulan data yang sama, dengan setiap pengklasifikasi berikutnya lebih menekankan pada sampel yang salah diklasifikasikan oleh pengklasifikasi sebelumnya. Pengklasifikasi terakhir adalah jumlah tertimbang dari pengklasifikasi lemah, dengan bobot ditentukan oleh keakuratannya [16].

Extreme Gradient Boosting atau yang dikenal juga dengan nama *XGBoost* adalah metode pembelajaran mesin yang digunakan untuk masalah klasifikasi dan regresi. Ini adalah metode pembelajaran *ensemble* yang menggabungkan beberapa pohon keputusan untuk membuat prediksi [17]. *XGBoost* merupakan versi perbaikan dari metode *Gradient Boosting*, yaitu algoritma peningkatan yang menggabungkan pembelajar yang lemah untuk menciptakan pembelajar yang kuat [18]. *XGBoost* menggunakan algoritma penurunan gradien untuk meminimalkan fungsi kerugian dan meningkatkan akurasi model [19]. *XGBoost* juga memiliki beberapa parameter yang dapat disesuaikan untuk meningkatkan performa model, seperti *learning rate*, jumlah pohon keputusan, dan kedalaman pohon keputusan [20].

3.4 Skema Eksperimen

Data dibagi menjadi dua set, yaitu set pelatihan (80%) dan set pengujian (20%). Kemudian, setiap algoritma klasifikasi menggunakan *cross validation* 10-*fold* untuk menghindari *overfitting*. Ini berarti bahwa set data pelatihan dibagi menjadi 10 blok, dan pelatihan setiap algoritma klasifikasi dilakukan dengan 9 blok, sementara satu blok sisanya digunakan untuk tujuan validasi.

Proses ini diulang 10 kali, sekali untuk setiap blok, sehingga memungkinkan maksimalnya jumlah pengamatan yang digunakan untuk validasi sambil menghindari *overfitting*. Skor estimator validasi silang rata-rata terbaik dipilih. Metodologi ini juga mencakup prosedur untuk memastikan bahwa setiap kelas diwakili dengan baik dalam setiap lipatan. Kemudian, kinerja keseluruhan dari setiap model yang terpilih dievaluasi dengan set pengujian. Karena target data yang tidak seimbang, akurasi bukanlah metrik yang paling sesuai untuk kinerja model, karena itu adalah metrik keseluruhan yang mungkin menghasilkan nilai tinggi berdasarkan kinerja yang baik hanya untuk kelas mayoritas.

Dalam penelitian ini, kami menggunakan metrik *F1*, yang mempertimbangkan *trade-off* antara presisi dan *recall*. Skor F1 dihitung untuk setiap kelas, dan skor *F1* rata-rata untuk tiga kelas juga dihitung. Ini adalah metrik yang digunakan untuk penyetelan *hyperparameter*, seperti yang akan dijelaskan selanjutnya. Untuk model yang dioptimalkan, akurasi juga dihitung sebagai metrik keseluruhan [21].

Semua model telah melalui proses penyetelan *hyperparameter*, yang merupakan langkah penting dalam meningkatkan kinerja mereka. Salah satu teknik yang digunakan untuk menyesuaikan *hyperparameter* adalah dengan melakukan pencarian grid. Ini adalah pendekatan yang sangat komprehensif di mana berbagai konfigurasi parameter diuji secara sistematis, dan yang terbaik dipilih berdasarkan hasil validasi silang. Dalam hal ini, kami memanfaatkan metode *Grid Search* CV yang disediakan oleh pustaka *Scikit-learn*, dengan metrik *F1-score* dan akurasi sebagai fokus utama untuk memaksimalkan kinerja model.

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

3. HASIL DAN PEMBAHASAN

Pada bagian ini merupakan hasil pengujian kinerja metode-metode klasifikasi yang telah dilatih setelah diterapkannya ADASYN dan *hyperparameter*. Tabel 1 menggambarkan hasil evaluasi klasifikasi untuk beberapa metode klasifikasi standar yang digunakan dalam penelitian ini, yaitu *Logistic Regression, Decision Tree, dan Random Forest*. Evaluasi dilakukan dengan menggunakan metrik *F1-score*, yang mengukur presisi dan *recall* dari setiap kelas, yaitu "*Dropout*," "*Enrolled*," dan "*Graduated*," serta rata-rata *F1-score* dan akurasi keseluruhan.

Tabel 1. Evaluasi Klasifikasi Metode Klasifikasi Standar

	Logistic Regression	***Decision Tree***	***Random Forest***
F1-score Dropout	0.76	0.72	0.78
F1-score Enrolled	0.51	0.46	0.50
F1-score Graduated	0.83	0.80	0.84
Rata-rata *F1-score*	0.70	0.66	0.71
Akurasi	0.74	0.71	0.76

Hasil yang tercatat dalam tabel ini memberikan gambaran tentang seberapa baik setiap metode klasifikasi mengatasi perbedaan dalam kinerja antara kelas-kelas tersebut. Dalam konteks ini, metrik *F1-score* digunakan untuk mengukur kualitas prediksi model terhadap setiap kelas. Dapat dilihat bahwa metode *Random Forest* memiliki *F1-score* tertinggi untuk kelas "*Dropout*" 0.78 dan "*Graduated*" 0.84, menunjukkan kemampuan yang lebih baik dalam mengklasifikasikan mahasiswa dalam kategori ini. Logistic Regression juga mencapai F1-score yang baik, terutama untuk kelas "*Graduated*" 0.83, sementara *Decision Tree* menunjukkan performa yang sedikit lebih rendah dalam hal ini.

Rata-rata *F1-score* dihitung untuk memberikan gambaran keseluruhan tentang performa metode klasifikasi. Metode *Random Forest* mencapai rata-rata *F1-score* tertinggi 0.71, diikuti oleh *Logistic Regression* 0.70, dan *Decision Tree* 0.66. Selain metrik *F1-score*, akurasi juga dicatat untuk mewakili tingkat kebenaran total prediksi model. Metode *Random Forest* memiliki akurasi tertinggi 0.76 dan *F1-score* tertinggi 0.71, yang mengindikasikan bahwa model ini lebih tepat dalam mengklasifikasikan data secara keseluruhan. Tabel 2 menampilkan hasil evaluasi klasifikasi untuk *Adaptive Boost* dan *Extreme Gradient Boost*. Evaluasi ini dilakukan dengan menggunakan metrik *F1-score* untuk tiga kategori yang berbeda, yaitu "*Dropout*," "*Enrolled*," dan "*Graduated*."

Tabel 2. Evaluasi Klasifikasi Metode Klasifikasi *Boosting*

	Adaptive Boosting	***XGBoost***
F1-score Dropout	0.76	0.78
F1-score Enrolled	0.48	0.46
F1-score Graduated	0.82	0.84
Rata-rata *F1-score*	0.69	0.69
Akurasi	0.74	0.76

4. KESIMPULAN

Dalam penelitian ini, kami telah melakukan evaluasi kinerja beberapa metode klasifikasi yang berbeda, yaitu *Adaptive Boost, Extreme Gradient Boost, Logistic Regression, Decision Tree, dan Random Forest*, dalam konteks penanganan data yang tidak seimbang menggunakan teknik ADASYN dan *Grid Search* CV dengan *10-fold cross-validation*.

Dari hasil evaluasi, kami dapat menyimpulkan bahwa metode klasifikasi *Logistic Regression* dan *Random Forest* secara konsisten menunjukkan kinerja yang lebih baik dibandingkan dengan metode klasifikasi lainnya. Khususnya, ketika kita mempertimbangkan nilai rata-rata *F1-score Logistic Regression* dan *Random Forest* mencapai nilai tertinggi, yaitu 0.70 dan 0,71 secara berurutan, sementara metode lainnya memiliki rata-rata *F1-score* yang lebih rendah. Hal ini menunjukkan bahwa *Logistic Regression* dan *Random Forest* dapat mengatasi masalah ketidakseimbangan kelas dengan baik, menghasilkan prediksi yang lebih baik untuk kelas minoritas (*Dropout* dan *Enrolled*).

Namun, meskipun terdapat peningkatan yang signifikan dalam kinerja, kita perlu mencatat bahwa metode klasifikasi ini masih memiliki ruang untuk perbaikan. Penggunaan *Grid Search* CV untuk mencari parameter optimal adalah langkah yang positif dalam meningkatkan kinerja model, tetapi mungkin masih ada kemungkinan untuk mengoptimalkan parameter lebih lanjut atau menggunakan metode klasifikasi yang lebih canggih.

DAFTAR PUSAKA

[1] D. and M. J. and B. L. M. T. and R. V. Martins Mónica V. and Tolledo, 2021, Early Prediction of student's Performance in Higher Education: A Case Study, Trends and Applications in Information Systems and Technologies, vol. 1, hal. 166–175.

[2] G. Lemaître, F. Nogueira, dan C. K. Aridas, 2017, Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning, Journal of Machine Learning Research, vol. 18, no. 17, hal. 1–5, [Daring]. Tersedia pada: http://jmlr.org/papers/v18/16- 365.html

[3] Haibo He, Yang Bai, E. A. Garcia, dan Shutao Li, 2008, ADASYN: Adaptive synthetic sampling approach for imbalanced learning, dalam 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), IEEE, hal. 1322–1328. doi: 10.1109/IJCNN.2008.4633969.

[4] F. Pedregosa dkk., 2011, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, vol. 12, hal. 2825–2830.

[5] K. Lu, Logistic Regression in Biomedical Study, 2022, 2022 International Conference on Biotechnology, Life Science and Medical Engineering (BLSME 2022), [Daring]. Tersedia pada: https://api.semanticscholar.org/CorpusID:248935866

[6] J. Friedman, T. Hastie, dan R. Tibshirani, 2000,Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors), The Annals of Statistics, vol. 28, no. 2, doi: 10.1214/aos/1016218223.

[7] M. Zanchak, V. Vysotska, dan S. Albota, 2021, The Sarcasm Detection in News Headlines Based on Machine Learning Technology, dalam 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT), IEEE, hal. 131–137. doi: 10.1109/CSIT52700.2021.9648710.

[8] Y. Yue, L. Jia, H. Zhai, M. Kong, dan M. Li, 2020, CFS-DT: a Combined Feature Selection and Decision Tree based Method for Octane Number Prediction, dalam 2020 4th Annual International Conference on Data Science and Business Analytics (ICDSBA), IEEE, hal. 100–103. doi: 10.1109/ICDSBA51020.2020.00033.

[9] J. R. Quinlan, 1986, Induction of decision trees, Mach Learn, vol. 1, no. 1, hal. 81–106, doi: 10.1007/BF00116251.

[10] E. Momeni, M. R. Sahebi, dan A. Mohammadzadeh, 2020, Classification Of High-Resolution Satellite Images Using Fuzzy Logics Into Decision Tree, Malaysian Journal of Geosciences, vol. 4, no. 1, hal. 07–12, doi: 10.26480/mjg.01.2020.07.12.

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

[11] L. Wang dan Y. Zhang, 2020, Clustering Reduction Method Analysis of Rough Set and Decision Tree based on Weight Matrix Analysis, IOP Conf Ser Mater Sci Eng, vol. 750, no. 1, hal. 012205, doi: 10.1088/1757-899X/750/1/012205.
[12] N. Nakaryakova, S. Rusakov, dan O. Rusakova, 2020, Prediction Of The Risk Group (By Academic Performance) Among First Course Students By Using The Decision Tree Method, Applied Mathematics and Control Sciences, no. 4, hal. 121–136, doi: 10.15593/2499-9873/2020.4.08.
[13] S. Abdullah dan G. Prasetyo, 2020, Easy Ensemmble with Random Forest To Handle Imbalanced Data In Classification, Journal of Fundamental Mathematics and Applications (JFMA), vol. 3, no. 1, hal. 39–46, doi: 10.14710/jfma.v3i1.7415.
[14] L. Breiman, Random Forests, Mach Learn, 2001, vol. 45, no. 1, hal. 5–32, doi: 10.1023/A:1010933404324.
[15] C. Han dan H. Jia, 2022, Multi-Modal Representation Learning with Self-Adaptive Thresholds for Commodity Verification, dalam China Conference on Knowledge Graph and Semantic Computing.
[16] Z. Zheng dan Y. Yang, 2021, Adaptive Boosting for Domain Adaptation: Toward Robust Predictions in Scene Segmentation, IEEE Transactions on Image Processing, vol. 31, hal. 5371– 5382.
[17] N. A. Akbar, A. Sunyoto, M. Rudyanto Arief, dan W. Caesarendra, 2020, Improvement of decision tree classifier accuracy for healthcare insurance fraud prediction by using Extreme Gradient Boosting algorithm, dalam 2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), IEEE, hal. 110–114.
[18] M. Alqahtani, H. Mathkour, dan M. M. Ben Ismail, 2020, IoT Botnet Attack Detection Based on Optimized Extreme Gradient Boosting and Feature Selection, Sensors, vol. 20, no. 21, hal. 6336, doi: 10.3390/s20216336.
[19] Z. Yan dan H. Wen, 2020, Electricity Theft Detection Base on Extreme Gradient Boosting in AMI, dalam 2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), IEEE, hal. 1–6.
[20] T. Chen dan C. Guestrin, 2016, XGBoost, dalam Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM, hal. 785–794.
[21] D. Johannßen, C. Biemann, S. Remus, T. Baumann, dan D. Scheffer, 2020, GermEval 2020 Task 1 on the Classification and Regression of Cognitive and Motivational Style from Text: Companion Paper.