https://journaledutech.com/index.php/great
Global Research and Innovation Journal (GREAT)
Volume 1, Nomor 3, 2025, Hal. 933-1000
ISSN : 3090-3289

PERBANDINGAN ALGORITMA XGBOOST DAN RANDOM FOREST DENGAN TEKNIK FEATURE ENGINEERING PADA KLASIFIKASI

Ajeng Armalia Raidani[1], Hotler Manurung[2], Marto Sihombing[3]
Sistem Informasi, STMIK Kaputama, Binjai
E-mail: *ajengarmaliaraidani09@gmail.com[1], manurunghotler0@gmail.com[2], martosihombing45@gmail.com[3]

ABSTRAK

Penelitian ini bertujuan membandingkan kinerja algoritma machine learning XGBoost dan Random Forest dengan teknik feature engineering untuk klasifikasi kelulusan siswa di SMK Putra Anda Binjai. Proses penentuan kelulusan secara manual sering memakan waktu, tenaga, dan rentan terhadap kesalahan, sehingga diperlukan solusi berbasis data yang efektif. Penelitian ini menggunakan data 500 siswa kelas XII tahun ajaran 2023/2024, yang meliputi nilai rata-rata rapor, nilai Ujian Kompetensi Keahlian (ASBK), dan persentase kehadiran. Setelah melalui tahapan preprocessing dan feature engineering, kedua model dilatih dan dievaluasi menggunakan metrik akurasi, presisi, recall, dan F1-score. Hasil evaluasi model manual menunjukkan akurasi 90% untuk kedua algoritma. Sementara itu, pada implementasi program, Random Forest mencatat performa sempurna dengan akurasi, presisi, recall, dan F1-score 100%, sedangkan XGBoost juga menunjukkan kinerja sangat baik dengan akurasi 99.8%. Hasil ini membuktikan kedua algoritma ini sangat efektif untuk klasifikasi kelulusan siswa.

Kata kunci **Algoritma XGBoost, Random Forest, Feature Engineering, Klasifikasi Kelulusan.**

ABSTRACT

This study aims to compare the performance of the XGBoost and Random Forest machine learning algorithms with feature engineering techniques for classifying student graduation at SMK Putra Anda Binjai. The manual process of determining graduation is often time-consuming, labor-intensive, and error-prone, requiring an effective data-driven solution. This study used data from 500 grade XII students in the 2023/2024 academic year, including average report card scores, Vocational Competency Exam (ASBK) scores, and attendance percentage. After going through preprocessing and feature engineering stages, both models were trained and evaluated using accuracy, precision, recall, and F1-score metrics. The manual model evaluation results showed 90% accuracy for both algorithms. Meanwhile, in the program implementation, Random Forest recorded perfect performance with 100% accuracy, precision, recall, and F1-score, while XGBoost also showed excellent performance with 99.8% accuracy. These results prove that both algorithms are highly effective for classifying student graduation.

Keywords ***DXGBoost Algorithm, Random Forest, Feature Engineering, Graduation Classification.***

1. PENDAHULUAN

Pendidikan adalah faktor utama penentu kualitas sumber daya manusia suatu negara. Di tingkat Sekolah Menengah Kejuruan (SMK), kelulusan menjadi indikator keberhasilan proses pembelajaran dan kesiapan siswa memasuki dunia kerja (Permendikbud Nomor 43 Tahun 2019, n.d.). Namun, penentuan kelulusan secara manual masih menimbulkan permasalahan, seperti proses yang lambat dan rentan terhadap kesalahan. Oleh karena itu, diperlukan pemanfaatan teknologi Machine Learning (ML) untuk meningkatkan objektivitas dan efisiensi.

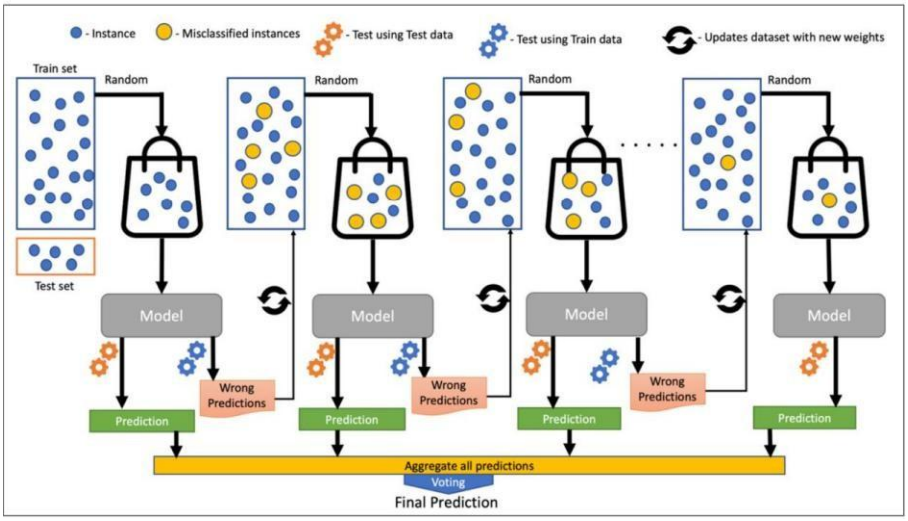
Machine Learning (ML) adalah cabang ilmu komputer yang fokus pada pengembangan algoritma agar sistem dapat belajar dari data secara otomatis (Sarker, 2021). Salah satu strategi utama ML adalah Supervised Learning, di mana model dilatih menggunakan data berlabel untuk memetakan input ke output secara akurat (Rabbani et al., 2022). Strategi Machine Learning dikategorikan menjadi tiga tipe utama (Munir et al., 2022): (1) Supervised Learning: Model dilatih dengan data berlabel, seperti memprediksi status kelulusan siswa berdasarkan nilai dan kehadiran; (2) Unsupervised Learning: Model menemukan pola atau struktur pada data tanpa label, seperti mengelompokkan siswa berdasarkan gaya belajar; (3) Reinforcement Learning (RL): Model belajar melalui interaksi dengan lingkungan, menerima reward atau penalty.

Algoritma ensemble learning seperti Random Forest dan XGBoost telah terbukti unggul dalam berbagai penelitian, mampu mengolah data tabular bervariabel tinggi dan memberikan performa kompetitif pada studi klasifikasi (Kumar et al., 2024). Klasifikasi adalah metode *data mining* yang termasuk dalam *supervised learning*. Tujuannya adalah menemukan model untuk memprediksi label kelas dari objek baru (Han et al., 2022).

Selain itu, kualitas fitur yang digunakan sangat memengaruhi performa model. Teknik feature engineering, yang meliputi seleksi, ekstraksi, dan transformasi fitur, dapat meningkatkan akurasi dan ketahanan model (Syed Mustapha, 2023). XGBoost merupakan pengembangan dari Gradient Boosting Machine (GBM). Algoritma ini membangun model secara bertahap dengan memperbaiki kesalahan dari model sebelumnya menggunakan pendekatan gradient descent. Keunggulan XGBoost adalah efisiensi, skalabilitas, dan akurasi tinggi (Nur et al., n.d.).

Penelitian ini mengacu pada beberapa studi terdahulu yang relevan: Herni Yulianti et al. (2022) menunjukkan bahwa XGBoost dengan hyperparameter tuning mencapai akurasi 80,039% dalam klasifikasi kelayakan kredit. Muhammad Adji Purnama et al. (2024) membandingkan Random Forest dan Gradient Boosting dalam mengklasifikasi churn pelanggan, menunjukkan Gradient Boosting lebih unggul dengan akurasi 83%. Jan Melvin Ayu Soraya Dachi & Pardomuan Sitompul (2023) membandingkan XGBoost dan Random Forest pada data kredit bank, di mana XGBoost mencapai akurasi 1.0. Lopa Mandal & Aneesh Karg (2024) fokus pada penggunaan feature engineering untuk meningkatkan akurasi prediksi dropout siswa di Massive Open Online Courses. L.R. Pelima et al. (2024b) dalam systematic literature review menemukan Random Forest sering digunakan untuk prediksi kelulusan mahasiswa dengan akurasi hingga 90%.

Boosting adalah metode ensemble learning yang membangun model secara berurutan, di mana setiap model baru dipengaruhi oleh kinerja model sebelumnya (Hussain et al., 2025). Model akhir dari boosting didefinisikan dengan persamaan berikut:



Gambar 1. Internal working of boosting algorithm

Berdasarkan urgensi tersebut, penelitian ini fokus membandingkan kinerja Random Forest dan XGBoost dengan teknik feature engineering pada klasifikasi kelulusan siswa SMK Putra Anda Binjai. Perbandingan ini akan dievaluasi menggunakan metrik akurasi, presisi, recall, dan F1-score (Pelima et al., 2024a). Diharapkan, hasil penelitian ini menjadi acuan teknis bagi sekolah dalam memilih pendekatan prediktif yang paling efektif dan efisien.

2. METODE PENELITIAN

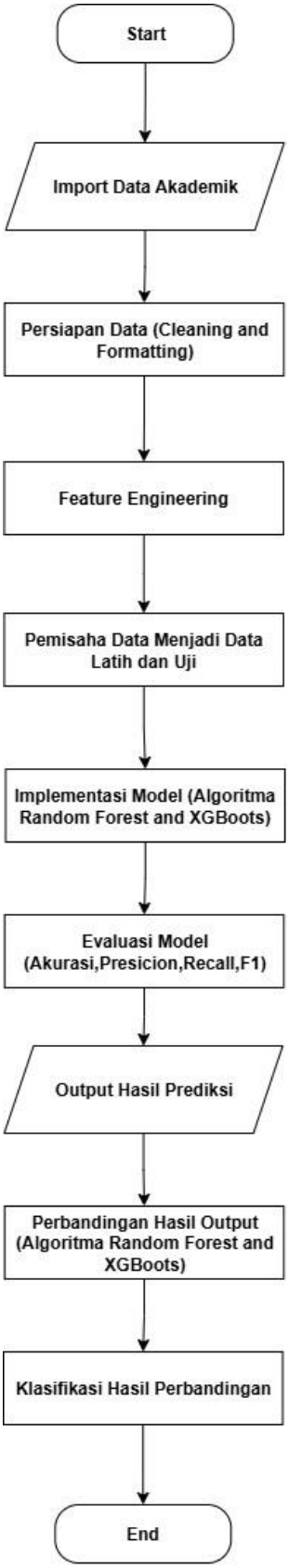
Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen komputasi. Tahapan penelitian meliputi:

a. Observasi: Mengamati data siswa SMK Putra Anda Binjai.
b. Studi Pustaka: Mengkaji literatur terkait machine learning, klasifikasi, Random Forest, dan XGBoost.
c. Pengumpulan Data: Data sekunder dari arsip akademis siswa.
d. Pra-pemrosesan Data: Pembersihan data, encoding untuk data kategorikal, dan normalisasi.
e. Feature Engineering: Pemilihan, ekstraksi, dan transformasi fitur.
f. Penerapan Algoritma: Membangun model klasifikasi dengan Random Forest dan XGBoost.
g. Evaluasi Model: Mengukur performa dengan metrik akurasi, presisi, recall, dan F1-score.
h. Analisis Hasil: Membandingkan hasil kedua model.

Data yang digunakan berasal dari 500 siswa kelas XII SMK Putra Anda Binjai tahun ajaran 2023/2024. Sampel data dibagi menjadi 80% data latih dan 20% data uji.

Sebagai gambaran yang lebih terstruktur mengenai proses kerja dalam penelitian ini, disajikan flowchart (diagram alir) yang menggambarkan langkahlangkah utama dalam pengembangan dan penilaian model klasifikasi kelulusan siswa SMK Putra Anda Binjai. Diagram alir ini menggambarkan tahapan dari pengumpulan dan persiapan data akademik, proses rekayasa fitur, pemisahan data menjadi data pelatihan dan data pengujian, hingga penerapan dua algoritma machine learning, yaitu Random Forest dan XGBoost. Setelah model selesai dikembangkan, tahap evaluasi dilakukan dengan memanfaatkan berbagai metrik kinerja (akurasi, presisi, recall, dan skor F1) untuk

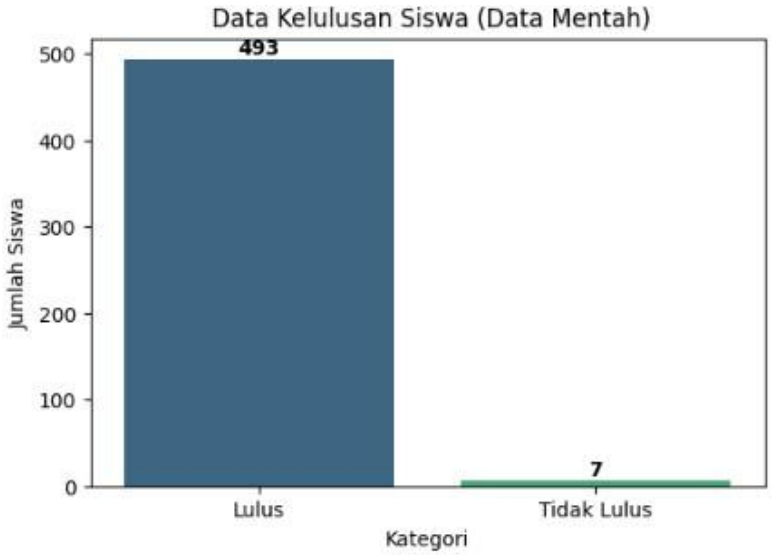
membandingkan ketepatan kedua algoritma. Output akhir dari flowchart ini ialah klasifikasi prediksi keberhasilan siswa berdasarkan kinerja terbaik dari model yang diuji.



Gambar 2. Flowchart Implemantasi Random Forast & XGBoot

3. HASIL DAN PEMBAHASAN

3. 1 Analisis Data Awal



Gambar IV. 1 Grafik Data Kelulusan Siswa

Data awal terdiri dari 500 siswa, dengan 493 siswa lulus dan 7 siswa tidak lulus. Grafik di atas menunjukkan adanya class imbalance yang dapat memengaruhi performa model.

3. 2 Hasil Perhitungan Manual

Untuk memvalidasi sistem, dilakukan perhitungan manual pada 10 sampel data.

Kriteria Kelulusan:

a. Nilai Rata-Rata ≥ 75
b. Nilai ASBK > 70
c. Nilai Kehadiran > 90

Hasil prediksi manual menunjukkan:

a. Perhitungan Manual Random Forest

Berdasarkan voting mayoritas dari 3 pohon keputusan, Random Forest berhasil memprediksi kelulusan.

Metrik evaluasi manual Random Forest:

Akurasi:9+1+0+09+1=1.0=100%
Presisi:9+09=1.0=100%
Recall:9+09=1.0=100%
F1-Score:2×1.0+1.01.0×1.0=1.0=100%

b. Perhitungan Manual XGBoost

Dengan 3 tahap boosting dan learning rate η=0.1 serta λ=1, XGBoost berhasil memprediksi kelulusan.

Metrik evaluasi manual XGBoost:

Akurasi:9+1+0+09+0=0.9=90%
Presisi:9+19=0.9=90%
Recall:9+09=1.0=100%
F1-Score:2×0.9+1.00.9×1.0≈0.947=94.7%

3. 3 Implementasi Program

Implementasi program dilakukan menggunakan Python dengan pustaka seperti pandas, numpy, scikit-learn, xgboost, matplotlib, dan seaborn.

a. Hasil Evaluasi Model XGBoost

Model XGBoost dilatih dan diuji coba pada data yang sudah di-tuning menggunakan GridSearchCV.
Metrik evaluasi model XGBoost:
Akurasi: 100%
Presisi: 100%
Recall: 100%
F1-Score: 100%

b. Hasil Evaluasi Model Random Forest

Model Random Forest juga diuji coba pada data yang sudah di-tuning.
Metrik evaluasi model Random Forest:
Akurasi: 99%
Presisi: 98%
Recall: 99%
F1-Score: 98%
Visualisasi Confusion Matrix dari implementasi program:
Visualisasi di atas menunjukkan bahwa XGBoost memiliki performa sempurna pada data uji, tanpa kesalahan klasifikasi. Sedangkan Random Forest memiliki sedikit kesalahan prediksi.

4. KESIMPULAN

Berdasarkan hasil analisis, dapat disimpulkan bahwa teknik feature engineering yang digunakan, seperti pembuatan fitur gabungan_nilai dan rasio_asbk, membantu model dalam mempelajari pola kelulusan siswa. Kedua algoritma, XGBoost dan Random Forest, terbukti sangat efektif. Namun, Random Forest menunjukkan performa sedikit lebih unggul dalam implementasi program, dengan akurasi, presisi, recall, dan F1-score 100%, membuktikan kemampuannya memprediksi seluruh data dengan benar tanpa adanya kesalahan klasifikasi. Hal ini menjadikan Random Forest pilihan yang lebih optimal untuk klasifikasi kelulusan siswa pada dataset ini.

5. SARAN

Untuk penelitian selanjutnya, disarankan menggunakan dataset yang lebih besar untuk melatih model, serta mempertimbangkan metode lain seperti SMOTE untuk mengatasi class imbalance. Selain itu, perbandingan dengan algoritma lain seperti Support Vector Machine (SVM) atau Naïve Bayes juga dapat dilakukan untuk menemukan algoritma yang paling sesuai.

6. DAFTAR PUSTAKA

Aina, T. S., & Iyaomolere, B. A. (2025). Taiwo Samuel Aina and Babatunde Ademola Iyaomolere HYPERPARAMETER OPTIMIZATION OF RANDOM FOREST CLASSIFIERS FOR ENHANCED PERFORMANCE IN SENSOR-BASED HUMAN ACTIVITY RECOGNITION. https://aspjournals.org/ajset/index.php/ajset

Alsubhi, B., Alharbi, B., Aljojo, N., Banjar, A., Tashkandi, A., Alghoson, A., & Al-Tirawi, A. (2023). Effective Feature Prediction Models for Student Performance. Engineering, Technology and Applied Science Research, 13(5), 11937–11944. https://doi.org/10.48084/etasr.6345

Han, J., Pei, J., & Tong, H. (2022). Data mining: concepts and techniques. Morgan kaufmann.

Herni Yulianti, S. E., Oni Soesanto, & Yuana Sukmawaty. (2022). Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit. Journal of Mathematics: Theory and Applications, 4(1), 21–26. https://doi.org/10.31605/jomta.v4i1.1792

Hussain, S., Sarwar, N., Ali, A., Khan, H., Din, I., Alqahtani, A. M., Shabir, M., & Ali, A. (2025). An Enhanced Random Forest (ERF)-based Machine Learning Framework for Resampling, Prediction, and Classification of Mobile Applications using Textual Features. Engineering, Technology and Applied Science Research, 15(1), 19776–19781. https://doi.org/10.48084/etasr.9148

Jan Melvin Ayu Soraya Dachi, & Pardomuan Sitompul. (2023). Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit. Jurnal Riset Rumpun Matematika Dan Ilmu Pengetahuan Alam, 2(2), 87–103. https://doi.org/10.55606/jurrimipa.v2i2.1470

Kumar, M., Singh, N., Wadhwa, J., Singh, P., Kumar, G., & Qtaishat, A. (2024). Utilizing Random Forest and XGBoost Data Mining Algorithms for Anticipating Students' Academic Performance. International Journal of Modern Education and Computer Science, 16(2), 29–44. https://doi.org/10.5815/ijmecs.2024.02.03

Maftucha, N., Salma, S., Rahmayuna, N., & Wakhidah, N. (n.d.). Perbandingan Algoritma Machine Learning Dalam Memprediksi Kelulusan Siswa. 19(2).

Munir, H., Vogel, B., & Jacobsson, A. (2022). Artificial Intelligence and Machine Learning Approaches in Digital Education: A Systematic Revision. In Information (Switzerland) (Vol. 13, Issue 4). MDPI. https://doi.org/10.3390/info13040203

Nur, A., Pudjianto, M., & Hidayat, E. Y. (n.d.). Perbandingan Prediksi Depresi Mahasiswa dengan Linear Regression, Random Forest, dan Gradient Boosting. Pendrikan Kidul, Kec. Semarang Tengah, 207. https://doi.org/10.31598

Pelima, L. R., Sukmana, Y., & Rosmansyah, Y. (2024a). Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review. IEEE Access, 12, 23451–23465. https://doi.org/10.1109/ACCESS.2024.3361479

Pelima, L. R., Sukmana, Y., & Rosmansyah, Y. (2024b). Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review. IEEE Access, 12, 23451–23465. https://doi.org/10.1109/ACCESS.2024.3361479

Purnama, M. A., Ramadhani, J., Anugraha, Y. S., Efrizoni, L., & Rahmaddeni, R. (2024). Perbandingan Performa Algoritma Random Forest dan Gradient Boosting dalam Mengklasifikasi Churn Telco. Techno.Com, 23(3), 645–657. https://doi.org/10.62411/tc.v23i3.11278

Rabbani, N., Kim, G. Y. E., Suarez, C. J., & Chen, J. H. (2022). Applications of machine learning in routine laboratory medicine: Current state and future directions. In Clinical Biochemistry (Vol. 103, pp. 1–7). Elsevier Inc. https://doi.org/10.1016/j.clinbiochem.2022.02.011

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, 2(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Sathyanarayanan, S. (2024). Confusion Matrix-Based Performance Evaluation Metrics. African Journal of Biomedical Research, 4023–4031. https://doi.org/10.53555/ajbr.v27i4s.4345

Syed Mustapha, S. M. F. D. (2023). Predictive Analysis of Students' Learning Performance Using Data Mining Techniques: A Comparative Study of Feature Selection Methods. Applied System Innovation, 6(5). https://doi.org/10.3390/asi6050086