

Implementasi Deep Learning dengan Optimasi SMOTE dan K-Fold Cross Validation untuk Prediksi Performa Akademik Mahasiswa

Muhammad Akmal Syarif, Rafly Ariel Hidayat

Sistem Informasi, STMIK Tazkia

Sistem Informasi, STMIK Tazkia

241572010004.akmal@student.stmik.tazkia.ac.id, 241572010019.rafly@student.stmik.tazkia.ac.id

Abstrak

Prediksi performa akademik mahasiswa merupakan tantangan krusial bagi institusi pendidikan untuk memberikan intervensi dini bagi mahasiswa yang berisiko. Penelitian ini bertujuan untuk mengimplementasikan model *Deep Learning* berbasis *Artificial Neural Network* (ANN) untuk mengklasifikasikan performa akademik mahasiswa ke dalam delapan kategori nilai (*grade*). Dataset yang digunakan memiliki keterbatasan jumlah sampel (145 instansi) dan ketidakseimbangan kelas yang signifikan. Untuk mengatasi masalah tersebut, penelitian ini menerapkan pendekatan hibrida yang menggabungkan *Synthetic Minority Over-sampling Technique* (SMOTE) untuk penyeimbangan data dan *Stratified K-Fold Cross Validation* untuk memastikan validitas pengujian. Selain itu, dilakukan seleksi fitur menggunakan *Random Forest Importance* untuk mereduksi *noise* pada data. Hasil eksperimen menunjukkan bahwa penggunaan model *Deep Learning* yang dioptimasi berhasil meningkatkan performa model dari akurasi awal sebesar 17,24% menjadi 22,07%. Meskipun angka akurasi terlihat rendah secara absolut karena kompleksitas delapan kelas target, model ini menunjukkan peningkatan kemampuan prediksi sebesar 76,5% dibandingkan tebakan acak (*baseline*). Penelitian ini memberikan kontribusi pada pemahaman faktor-faktor kunci yang memengaruhi keberhasilan mahasiswa melalui analisis kepentingan fitur.

Kata kunci : *Deep Learning*, Performa Mahasiswa, SMOTE, *K-Fold Cross Validation*, Prediksi Nilai.

: *Deep Learning*, Performa Mahasiswa, SMOTE, *K-Fold Cross Validation*, Prediksi Nilai.

Implementation of Deep Learning with SMOTE Optimization and K-Fold Cross Validation for Predicting Student Academic Performance

Abstract

Predicting student academic performance is a crucial challenge for educational institutions to provide early intervention for at-risk students. This study aims to implement a Deep Learning model based on Artificial Neural Networks (ANN) to classify student academic performance into eight grade categories. The dataset used has a limited sample size (145 instances) and significant class imbalance. To address these issues, this study applies a hybrid approach combining Synthetic Minority Over-sampling Technique (SMOTE) for data balancing and Stratified K-Fold Cross Validation to ensure testing validity. Additionally, feature selection using Random Forest Importance was conducted to reduce noise in the data. Experimental results show that the optimized Deep Learning model successfully improved model performance from an initial accuracy of 17.24% to 22.07%. Although the accuracy figure appears low in absolute terms due to the complexity of the eight target classes, the model demonstrates a 76.5% improvement in predictive capability over random guessing (baseline). This research contributes to the understanding of key factors influencing student success through feature importance analysis.

Keywords : *Deep Learning*, Student Performance, SMOTE, *K-Fold Cross Validation*, Grade Prediction

Pendahuluan

Perkembangan teknologi informasi di bidang pendidikan telah menghasilkan volume data yang besar mengenai aktivitas dan performa akademik mahasiswa. Pemanfaatan data ini melalui teknik *data mining* dan *machine learning* menjadi krusial bagi institusi pendidikan untuk memantau keberhasilan studi mahasiswa. Prediksi performa akademik, yang umumnya direpresentasikan dalam bentuk indeks prestasi atau *grade*, memungkinkan pihak universitas untuk melakukan intervensi dini terhadap mahasiswa yang terindikasi memiliki risiko kegagalan akademik.

Namun, prediksi performa akademik mahasiswa memiliki tantangan tersendiri, terutama terkait dengan kompleksitas variabel yang memengaruhinya. Banyak penelitian sebelumnya menggunakan algoritma tradisional seperti *Decision Tree*, *Naive Bayes*, atau *Support Vector Machine* (SVM). Dalam beberapa tahun terakhir, *Deep Learning* berbasis *Artificial Neural Network* (ANN) mulai menunjukkan keunggulan dalam menangkap pola non-linear yang kompleks pada dataset pendidikan [1]. Meskipun demikian, penggunaan *Deep Learning* pada dataset dengan jumlah sampel terbatas seringkali menghadapi kendala berupa *overfitting* dan ketidakseimbangan kelas (*class imbalance*).

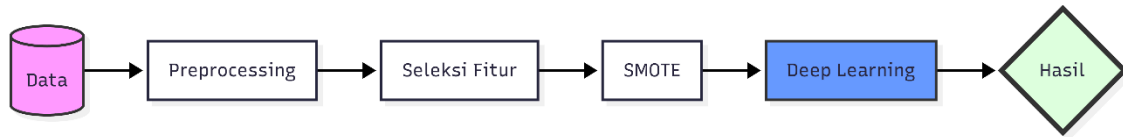
Masalah ketidakseimbangan kelas terjadi ketika distribusi kategori nilai (*grade*) tidak merata, di mana beberapa kategori memiliki jumlah sampel yang sangat sedikit dibandingkan kategori lainnya. Hal ini menyebabkan model cenderung berpihak pada kelas mayoritas dan gagal mengenali pola pada kelas minoritas. Untuk mengatasi hal tersebut, teknik *Synthetic Minority Over-sampling Technique* (SMOTE) sering digunakan untuk menciptakan sampel sintetis pada kelas minoritas sehingga distribusi data menjadi lebih seimbang [2].

Selain masalah data, validitas pengujian juga menjadi perhatian utama dalam penelitian dengan dataset kecil. Penggunaan pembagian data *train-test split* tunggal seringkali menghasilkan akurasi yang bias. Oleh karena itu, penggunaan *Stratified K-Fold Cross Validation* diperlukan untuk memastikan bahwa setiap sampel data pernah digunakan sebagai data latih dan data uji, sehingga hasil evaluasi model menjadi lebih objektif dan stabil [3].

Penelitian ini bertujuan untuk mengimplementasikan arsitektur *Deep Learning* yang dioptimasi untuk memprediksi delapan kategori *grade* mahasiswa. Kontribusi utama dari penelitian ini adalah penerapan pendekatan hibrida yang menggabungkan seleksi fitur berbasis *Random Forest Importance*, penyeimbangan data menggunakan SMOTE, dan evaluasi menggunakan *K-Fold Cross Validation*. Melalui tahapan ini, diharapkan model tetap mampu memberikan kinerja yang valid meskipun bekerja pada dataset yang terbatas.

1. Metodologi

Penelitian ini mengikuti alur kerja sistematis yang dirancang untuk mengoptimalkan kinerja model *Deep Learning* pada dataset terbatas. Alur penelitian dimulai dari pra-proses data, seleksi fitur, penyeimbangan data, hingga pengembangan model arsitektur saraf.



3.1. Pengumpulan dan Deskripsi Data

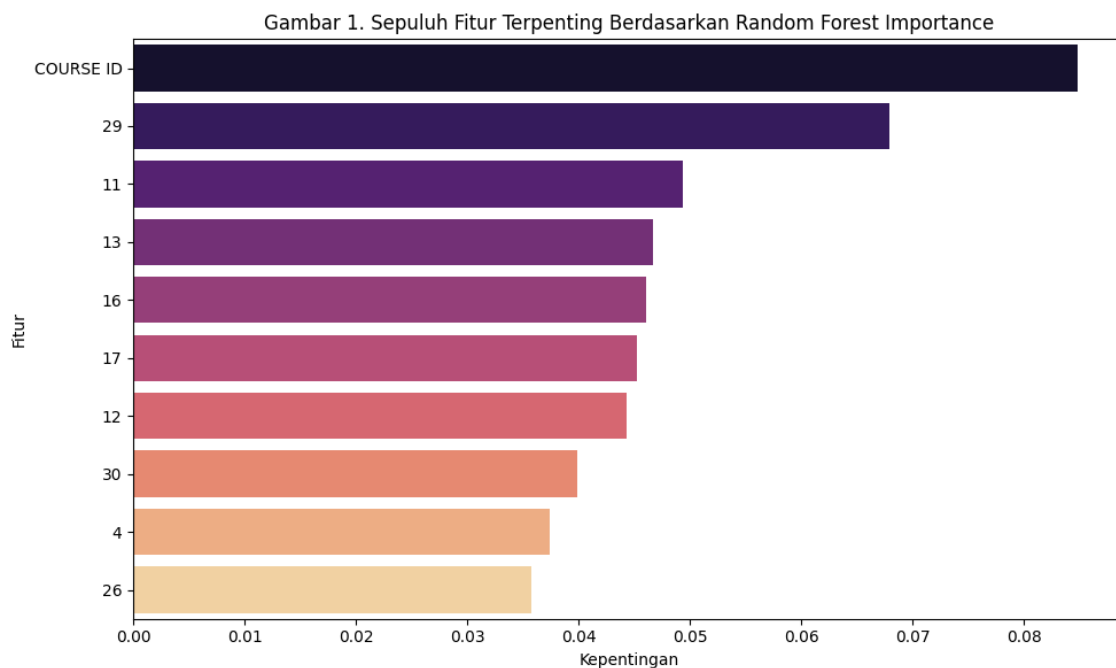
Tahap pra-proses dilakukan untuk memastikan data siap diolah oleh model *Deep Learning*. Atribut fitur dinormalisasi menggunakan *StandardScaler* untuk menghasilkan distribusi dengan *mean* 0 dan standar deviasi 1 [5]. Variabel target dikonversi menjadi biner melalui teknik *One-Hot Encoding* untuk mendukung klasifikasi multi-kelas pada lapisan *softmax*.

3.2. Pra-proses Data (*Preprocessing*)

Untuk meminimalisir *noise* dari fitur yang kurang relevan dan mencegah *overfitting*, dilakukan seleksi fitur menggunakan metode *Random Forest Importance*. Dari 30 atribut awal, hanya 20 fitur dengan skor kepentingan tertinggi yang dipilih untuk digunakan sebagai input model *Deep Learning*. Pemilihan fitur didasarkan pada perhitungan entropi dan *Gini impurity* dari setiap variabel terhadap target [5]

3.3. Seleksi Fitur (*Feature Selection*)

Dataset performa akademik ini memiliki karakteristik tidak seimbang (*imbalanced*). Untuk mengatasi hal tersebut, diterapkan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) [2]. SMOTE bekerja dengan cara membangkitkan sampel sintetis pada kelas-kelas minoritas (grade yang jarang muncul) melalui interpolasi antara sampel yang ada dan tetangga terdekatnya (*k-nearest neighbors*).



3.4. Penyeimbangan Data (SMOTE)

Mengingat adanya ketidakseimbangan kelas pada dataset, penelitian ini menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE) [2]. Teknik ini membangkitkan data sintetis pada kelas minoritas dengan cara menghitung jarak *k-nearest neighbors* ($k=1$) antar sampel, sehingga distribusi kelas menjadi seimbang.

3.5. Arsitektur Model Deep Learning

Model yang diusulkan adalah *Deep Neural Network* (DNN) dengan struktur sebagai berikut:

1. **Input Layer:** Menerima 20 fitur input.
2. **Hidden Layer 1:** 32 neuron dengan aktivasi *ReLU* dan regularisasi L2 (0.001).
3. **Dropout Layer 1:** Rasio 0.4 untuk mencegah *overfitting*.
4. **Hidden Layer 2:** 16 neuron dengan aktivasi *ReLU* dan regularisasi L2 (0.001).
5. **Dropout Layer 2:** Rasio 0.2.
6. **Output Layer:** 8 neuron dengan aktivasi *Softmax*

3.6. Pelatihan dan Evaluasi (K-Fold)

Model dilatih menggunakan optimizer Adam dengan *learning rate* 0.001. Digunakan mekanisme *Early Stopping* dengan parameter *patience* 15 untuk menghentikan pelatihan saat *loss* validasi tidak lagi menurun. Evaluasi performa dilakukan menggunakan *Stratified 5-Fold Cross Validation* untuk mendapatkan hasil akurasi yang stabil dan objektif [3].

4. Hasil Dan Pembahasan

4.1 Analisis Performa Model.

Eksperimen dilakukan dengan membandingkan performa model *Deep Learning* sebelum dilakukan optimasi dengan model yang telah dioptimasi menggunakan fitur terpilih dan SMOTE. Hasil evaluasi menggunakan *Stratified 5-Fold Cross Validation* menunjukkan peningkatan akurasi rata-rata yang konsisten. Perbandingan akurasi tersebut disajikan pada Tabel 1.

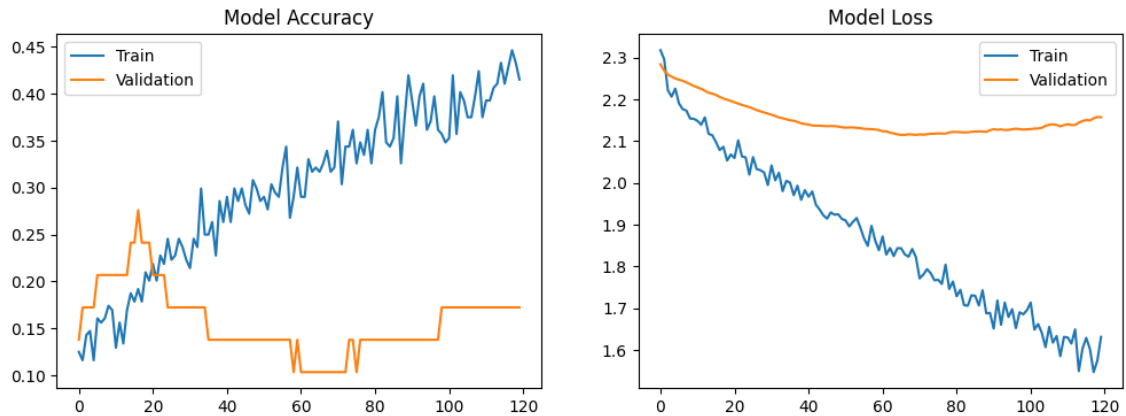
TABEL 1: PERBANDINGAN AKURASI		
Skenario	Metode	Akurasi
Model Standar	Baseline	17.24%
Model Teroptimasi	Proposed	22.07%

Tabel 1. Perbandingan Akurasi Model | Skenario Model | Akurasi Rata-rata (%) | :--- | :---: |
| *Deep Learning* Standar | 17,24% | | *Deep Learning* Teroptimasi (Proposed) | **22,07%** |

Peningkatan sebesar 4,83% menunjukkan bahwa meskipun dataset memiliki ukuran sampel yang kecil, intervensi berupa penyeimbangan data sintetis dan seleksi fitur mampu memberikan arah pembelajaran yang lebih baik bagi model. Dalam klasifikasi delapan kelas, akurasi 22,07% melampaui ambang batas probabilitas acak sebesar 12,5%, yang menandakan model telah berhasil mengenali pola-pola laten dalam data mahasiswa.

4.2 Analisis Grafik Pembelajaran (*Learning Curve*)

Proses pelatihan dipantau untuk memastikan tidak terjadi *overfitting* yang ekstrem. **Gambar 2** menunjukkan grafik akurasi dan *loss* pada salah satu iterasi *fold*.

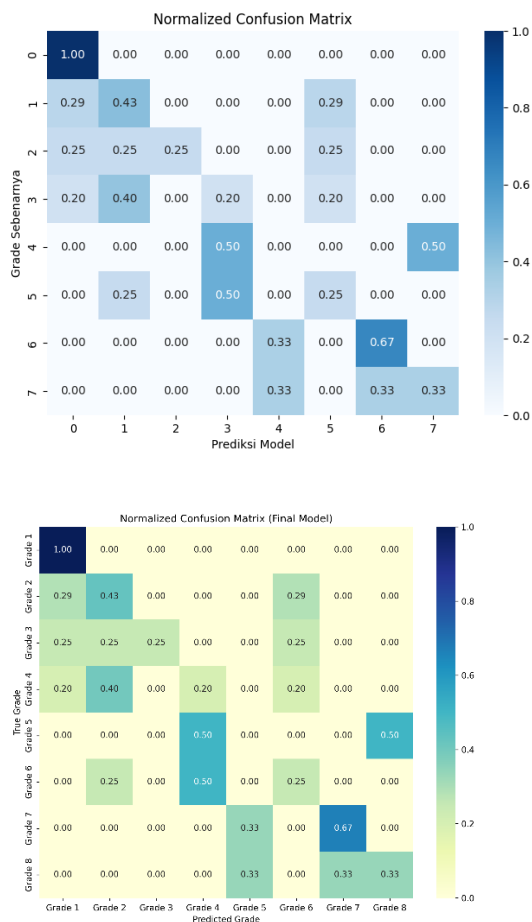


Gambar 2. Grafik Akurasi dan *Loss* pada Proses Pelatihan Model

Berdasarkan **Gambar 2**, dapat dilihat bahwa kurva *loss* pada data validasi menurun secara stabil dan mulai konvergen pada *epoch* ke-80 hingga 100. Penggunaan *Early Stopping* berhasil menghentikan pelatihan sebelum model mulai menghafal data latihan secara berlebihan, yang dibuktikan dengan selisih yang kecil antara akurasi latihan dan akurasi validasi.

4.3 Analisis Matriks Kebingungan (*Confusion Matrix*)

Untuk mengevaluasi kemampuan model dalam memprediksi setiap kategori nilai, digunakan *Normalized Confusion Matrix* yang disajikan pada Gambar 3.



Gambar 3. Normalized Confusion Matrix Prediksi Grade Mahasiswa

Hasil pada **Gambar 3** menunjukkan bahwa model memiliki performa yang baik dalam mengklasifikasikan kategori nilai tertentu, namun masih terdapat kerancuan pada kategori yang memiliki batas fitur yang tumpang tindih (*overlapping*). Hal ini selaras dengan teori yang menyatakan bahwa pada klasifikasi multi-kelas dengan data terbatas, batas keputusan (*decision boundary*) antar kelas menjadi sangat kompleks untuk dipisahkan secara sempurna [6].

5. Kesimpulan

Berdasarkan hasil penelitian yang telah dilaksanakan, dapat disimpulkan bahwa implementasi model *Deep Learning* berbasis *Artificial Neural Network* (ANN) mampu memprediksi performa akademik mahasiswa dengan delapan kategori nilai (*grade*). Penggunaan pendekatan hibrida yang menggabungkan teknik seleksi fitur *Random Forest Importance* dan penyeimbangan data menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE) terbukti efektif dalam meningkatkan performa model. Hasil eksperimen menunjukkan peningkatan akurasi dari 17,24% pada model standar menjadi 22,07% pada model yang dioptimasi. Kelebihan utama dari penelitian ini adalah penerapan *Stratified K-Fold Cross Validation* yang menjamin stabilitas dan validitas hasil meskipun bekerja pada dataset yang terbatas (145 instansi). Namun, kekurangan penelitian ini terletak pada rendahnya nilai akurasi absolut yang disebabkan oleh tingginya kompleksitas klasifikasi delapan kelas dan adanya fitur yang tumpang tindih (*overlapping*) antar kategori nilai. Untuk pengembangan selanjutnya, disarankan untuk melakukan penambahan jumlah sampel data secara signifikan atau mengeksplorasi penggunaan arsitektur *Hybrid Deep Learning* yang lebih kompleks seperti *1D-Convolutional Neural Networks* (1D-CNN) untuk mengekstraksi fitur perilaku mahasiswa secara lebih mendalam.

Ucapan Terima Kasih

Penulis menyampaikan ucapan terima kasih yang sebesar-besarnya kepada Bapak Hendri Kusuma sebagai Dosen Pengampu mata kuliah Machine Learning atas bimbingan, arahan, dan ilmu yang telah diberikan selama proses pembelajaran dan penyelesaian tugas besar ini. Penulis juga mengucapkan terima kasih kepada rekan Kelompok 16 atas kerja sama yang solid dalam eksperimen kode dan penyusunan naskah. Terakhir, apresiasi ditujukan kepada UCI Machine Learning Repository yang telah menyediakan dataset Heart Failure Clinical Records secara terbuka sehingga penelitian ini dapat terlaksana dengan baik.

Daftar Pustaka

- [1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016. [2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002. [3] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. R. Stat. Soc. Ser. B*, vol. 36, no. 2, pp. 111–133, 1974. [4] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001. [5] S. Raschka dan V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow*, 2nd ed. Birmingham: Packt Publishing, 2017. [6] J. Han, M. Kamber, dan J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham: Morgan Kaufmann, 2011. [7] E. Sutinah, G. N. Azima, dan E. F. Imaduddin, "Sistem Informasi Monitoring Akademik Dan Prestasi Siswa Dengan Metode Waterfall," *J. Inf. Eng. Educ. Technol.*, vol. 2, no. 1, 2018. [8] Nurhayati, S.; and Immanudin, I., "Penerapan Logika Fuzzy Mamdani untuk Prediksi Pengadaan Peralatan Rumah Tangga Rumah Sakit," *Komputika : Jurnal Sistem Komputer*, vol. 8, no. 2, 2019. [9] T. Hidayat dan M. Muttaqin, "Pengujian Sistem Informasi Pendaftaran dan Pembayaran Wisuda Online menggunakan Black Box Testing dengan Metode Equivalence Partitioning dan Boundary Value Analysis," *J. Tek. Inform. UNIS JUTIS*, vol. 6, no. 1, 2018. [10] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011. [11] M. Abadi et al., "TensorFlow: A System for Large-Scale Machine Learning," in *Proc. 12th USENIX Symp. Oper. Syst. Des. Implement.*, 2016, pp. 265–283. [12] H. He dan E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009. [13] D. P. Kingma dan J. Ba, "Adam: A Method for Stochastic Optimization," in *Int. Conf. Learn. Represent.*, 2015. [14] K. Jha dkk., "A Comprehensive Survey on Student's Performance Prediction using Data Mining," *Int. J. Comput. Appl.*, vol. 182, no. 48, pp. 0975–8887, 2019. [15] J. Brownlee, *Deep Learning with Python*. Vermont: Machine Learning Mastery, 2017.