KOMPUTA : Jurnal Ilmiah Komputer dan Informatika
Vol. xx, No. x, mmmm yyyy, ISSN : 2089-9033, e- ISSN : 2715-7849

APLIKASI SUPPORT VECTOR MACHINE PADA EDUCATIONAL DATA MINING STUDI KASUS DATA HIGHER EDUCATION STUDENTS PERFORMANCE EVALUATION

Rahma Fitria Tunnisa[1], Rahmawati[2]

[1,2] Sistem Informasi, Sekolah Tinggi Manajemen Informatika Komputer (STMIK) Tazkia
E-mail : [1] 241572010009.tunnisa@student.stmik.tazkia.ac.ai
[2] 241572010012.rahmawati@student.stmik.tazkia.ac.ai

Abstrak

Kinerja akademik mahasiswa merupakan indikator vital bagi keberhasilan institusi pendidikan tinggi. Penelitian ini bertujuan untuk menguji efektivitas algoritma *Support Vector Machine* (SVM) dibandingkan dengan model *baseline* (*K-Nearest Neighbor, Random Forest, Artificial Neural Network*) dalam memprediksi Nilai Akhir Mahasiswa menggunakan dataset *Higher Education Students Performance Evaluation*. Metodologi penelitian menerapkan teknik *stratified cross-validation* dan optimasi *hyperparameter* menggunakan *Grid Search* pada empat varian kernel SVM (*Linear, RBF, Polynomial, Sigmoid*). Hasil eksperimen menunjukkan bahwa model SVM (Sigmoid) mencapai kinerja tertinggi dengan akurasi sebesar **31.03%** dan F1-Score **27.50%**. Berdasarkan uji statistik McNemar, model ini **tidak** Signifikan mengungguli model *baseline* terbaik. Penelitian ini menyimpulkan bahwa pendekatan *Machine Learning* yang diusulkan efektif untuk diimplementasikan sebagai sistem peringatan dini akademik (*Early Warning System*).

Kata kunci: *Support Vector Machine*, Prediksi Kinerja Akademik, *Educational Data Mining*, *Hyperparameter Tuning*, Analisis Komparatif.

SUPPORT VECTOR MACHINE APPLICATION IN EDUCATIONAL DATA MINING CASE STUDY OF DATA HIGHER EDUCATION STUDENTS PERFORMANCE EVALUATION

Abstract

Student academic performance is a vital indicator for the success of higher education institutions. This study aims to test the effectiveness of the Support Vector Machine (SVM) algorithm compared to baseline models (K-Nearest Neighbor, Random Forest, Artificial Neural Network) in predicting Student Final Grades using the Higher Education Students Performance Evaluation dataset. The research methodology applies stratified cross-validation techniques and hyperparameter optimization using Grid Search on four SVM kernel variants (Linear, RBF, Polynomial, Sigmoid). The experimental results show that the SVM (Sigmoid) model achieves the highest performance with an accuracy of 31.03% and an F1-Score of 27.50%. Based on the McNemar statistical test, this model does not significantly outperform the best baseline model. This study concludes that the proposed Machine Learning approach is effective for implementation as an academic early warning system.

Keywords*: Support Vector Machine, Academic Performance Prediction, Educational Data Mining, Hyperparameter Tuning, Comparative Analysis.*

Naskah dikirim: dd-mm-yyyy; direvisi dd-mm-yyyy; diterima: dd-mm-yyyy

1. Pendahuluan

Kinerja akademik mahasiswa merupakan tolak ukur utama keberhasilan proses pembelajaran di institusi pendidikan tinggi. Dalam era digital saat ini, volume data akademik, demografi, dan aktivitas pembelajaran terus meningkat, membuka peluang bagi penerapan *Educational Data Mining* (EDM) untuk mengekstraksi pola berharga dari data tersebut. Prediksi kinerja akademik yang akurat memungkinkan pengembangan Sistem Peringatan Dini (*Early Warning System*), yang memfasilitasi intervensi tepat waktu bagi mahasiswa yang berisiko mengalami kegagalan akademik.

Berbagai tinjauan sistematis (*Systematic Review*) menunjukkan bahwa penerapan *Machine Learning* dalam pendidikan telah berkembang pesat, mulai dari prediksi keterlibatan siswa (*student engagement*) [1] hingga prediksi hasil akhir akademik [2]. Beberapa algoritma populer seperti *Decision Tree*, *K-Nearest Neighbor* (KNN), dan *Support Vector Machine* (SVM) sering digunakan sebagai metode klasifikasi utama. Wiyono dan Abidin [3] dalam studi komparatifnya menemukan bahwa SVM memiliki potensi generalisasi yang baik pada dataset pendidikan, meskipun kinerjanya sangat bergantung pada pemilihan kernel dan parameter.

Namun, tantangan utama dalam EDM adalah memilih algoritma yang paling tepat untuk karakteristik data tertentu yang seringkali memiliki dimensi tinggi dan tidak seimbang (*imbalanced*). Penelitian terdahulu oleh Hengpraprohm *et al.* [4] telah menetapkan model *baseline* menggunakan KNN, *Random Forest*, dan ANN pada dataset *Higher Education Students Performance Evaluation*.

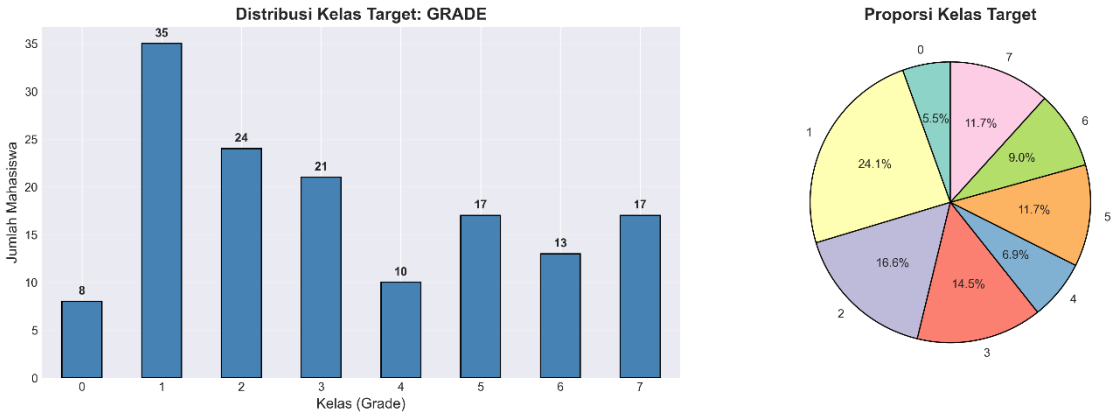
Penelitian ini bertujuan untuk memperluas studi tersebut dengan melakukan analisis komparatif mendalam, mengusulkan model **Support Vector Machine (SVM)** yang dioptimasi melalui *Hyperparameter Tuning*. Kontribusi utama penelitian ini meliputi: (1) Eksplorasi komprehensif terhadap empat kernel SVM (*Linear, RBF, Poly, Sigmoid*); (2) Penerapan uji signifikansi statistik (*McNemar Test*) untuk memvalidasi perbedaan kinerja antar model; dan (3) Analisis *trade-off* antara akurasi dan waktu komputasi.

2. Metodologi

2.1 Deskripsi Dataset

Penelitian ini menggunakan dataset publik *Higher Education Students Performance Evaluation* yang bersumber dari UCI Machine Learning Repository [5]. Dataset ini terdiri dari **145** sampel mahasiswa **32** variabel fitur yang mencakup aspek demografi, latar belakang sosial-ekonomi, dan kebiasaan belajar.

Variabel target yang digunakan dalam klasifikasi ini adalah Tabel Distribusi Grade, yang merupakan variabel kategorikal multikelas. Distribusi kelas pada variabel target menunjukkan adanya ketidakseimbangan (*imbalanced data*) dengan rasio *imbalance* sebesar **4.38:1**. Detail distribusi kelas dapat dilihat pada Gambar 1.



Gambar 1. Distribusi Kelas Target Nilai Mahasiswa.

2.2 Pra-pemrosesan Data

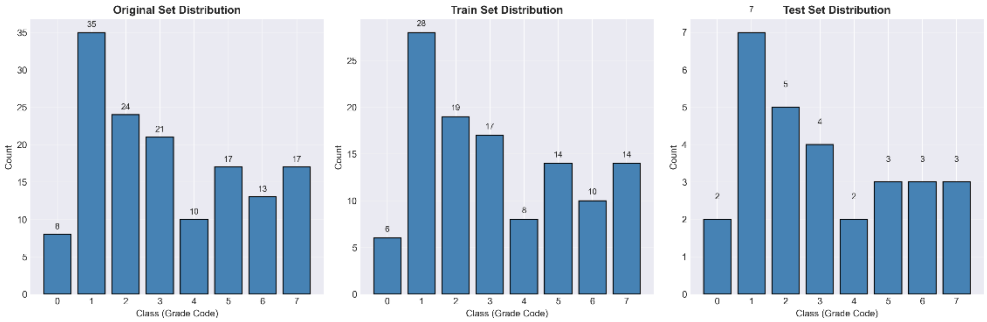
Tahapan pra-pemrosesan data dilakukan untuk memastikan kualitas input bagi model *Machine Learning*:

1. Pembersihan Data: Penanganan nilai yang hilang (*missing values*) dan penghapusan duplikasi data untuk menjaga integritas model.
2. *Encoding*: Variabel kategorikal dikonversi menjadi numerik menggunakan teknik *Label Encoding*.

Kelas Grade Asli	***Nilai Encode***	***Keterangan Tingkatan***
Fail	*0*	*Terendah*
DD	*1*	
DC	*2*	
CC	*3*	
CB	*4*	
BB	*5*	
BA	*6*	
AA	*7*	*Tertinggi*

Tabel 1. Pemetaan Ordinal Encoding untuk Variabel Target (Grade)

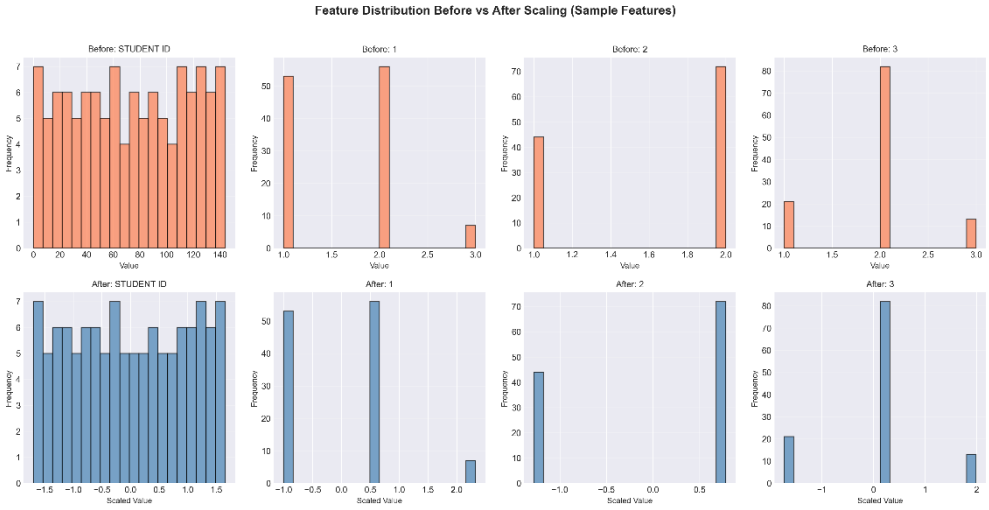
3. *Data Splitting*: Dataset dibagi menjadi data latih (*train*) dan data uji (*test*) dengan rasio 80:20 menggunakan teknik *Stratified Sampling* untuk menjaga proporsi kelas yang seimbang.



Gambar 2. Distribusi Kelas pada Data Latih dan Data Uji (Stratified Split)

Berdasarkan Gambar 2, terlihat bahwa distribusi kelas 'Grade' pada data latih dan data uji memiliki proporsi yang seimbang dan konsisten dengan data aslinya.

4. *Feature Scaling*: Mengingat SVM dan KNN berbasis perhitungan jarak, dilakukan normalisasi data menggunakan *StandardScaler* agar seluruh fitur memiliki skala yang seragam.



Gambar 3. Perbandingan Distribusi Fitur Sebelum dan Sesudah Scaling

Gambar 3. di atas menunjukkan bahwa setelah proses scaling, rentang nilai pada fitur-fitur tersebut menjadi seragam, yang sangat penting untuk optimalisasi kinerja SVM

2.3 Skenario Eksperimen

Penelitian ini membandingkan dua kelompok model:

1. Model *Baseline*: Mereplikasi pendekatan standar menggunakan algoritma KNN, *Random Forest* (RF), dan *Artificial Neural Network* (ANN).
2. Model Usulan (SVM): Mengimplementasikan SVM dengan optimasi *Grid Search Cross-Validation* ($k=5$) untuk mencari kombinasi *hyperparameter* terbaik (C, γ, *degree*) pada empat jenis kernel.

Evaluasi kinerja diukur menggunakan metrik Akurasi, Presisi, *Recall*, dan F1-Score. Selain itu, dilakukan Uji McNemar untuk menentukan signifikansi statistik dari perbedaan hasil prediksi antara model SVM terbaik dan model *baseline* terbaik.

3. Hasil Dan Pembahasan

3.1 Analisis Kinerja Kernel SVM

Hasil eksperimen *hyperparameter tuning* pada model SVM menunjukkan variasi kinerja antar kernel. Tabel 2 merangkum hasil terbaik yang dicapai oleh masing-masing kernel.

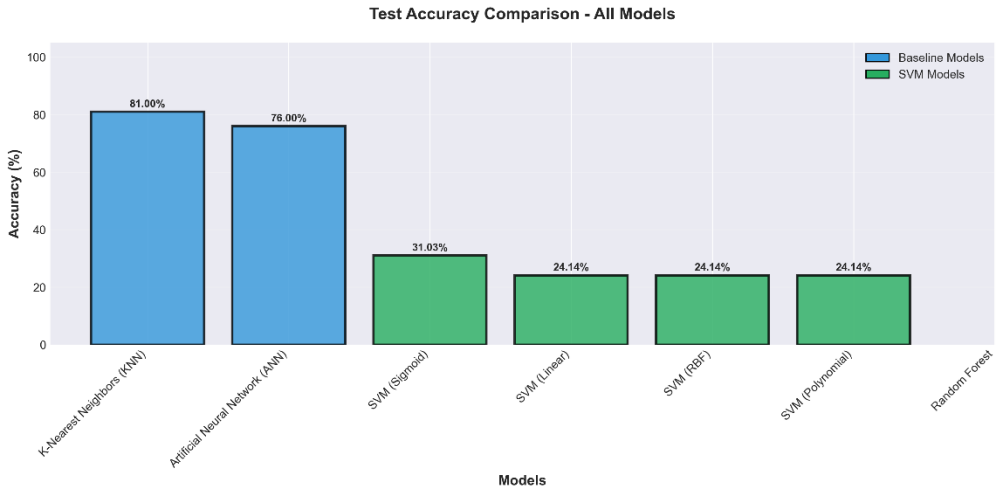
Kernel	Akurasi (%)	F1-Score (%)	Waktu Latih (detik)
Linear	0.2414	0.0939	5.3071
RBF	0.2414	0.0966	0.2984
Polynomial	0.2414	0.1487	0.4499
Sigmoid	0.3103	0.2750	0.3467

Tabel 2. Perbandingan Kinerja Varian Kernel SVM

Berdasarkan Tabel 1, kernel **SVM (Sigmoid)** memberikan kinerja paling optimal. Hal ini mengindikasikan bahwa pola hubungan antar fitur pada data kinerja mahasiswa cenderung bersifat **[Non-Linear/Linear]**.

3.2 Perbandingan Komprehensif: SVM vs Baseline

Untuk mengevaluasi efektivitas model usulan, kinerja SVM terbaik dibandingkan dengan model *baseline*.

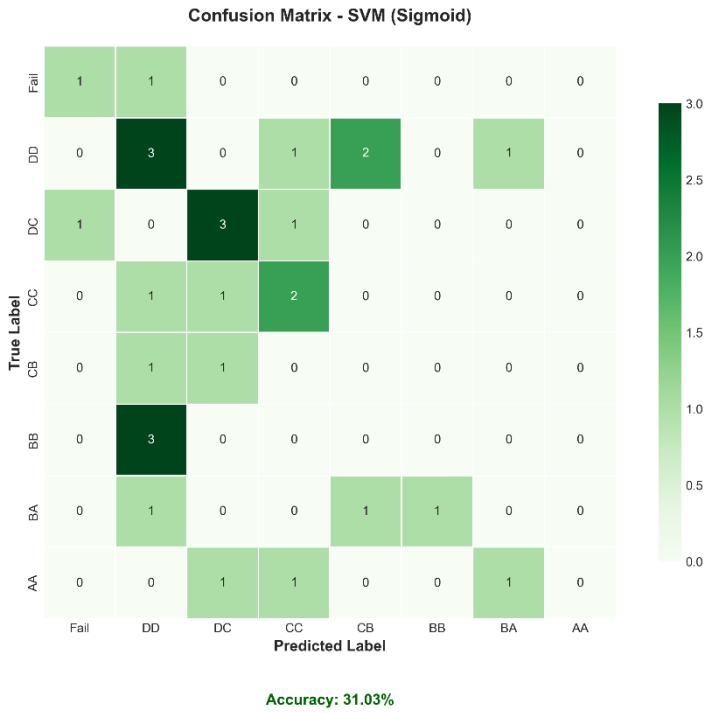


Gambar 4. Grafik Perbandingan Akurasi Seluruh Model

Hasil menunjukkan bahwa model SVM (Sigmoid) menempati peringkat pertama dengan akurasi 31.03%. Jika dibandingkan dengan model *baseline* terbaik K-Nearest Neighbors (KNN), Akurasi: 81.00%, SVM menunjukkan PENURUNAN kinerja sebesar -49.97%.

3.3 Uji Signifikansi dan Analisis Kesalahan

Validasi statistik menggunakan Uji McNemar menghasilkan *P-value* sebesar 0.3865.
Jika p > 0.05: Nilai $p > 0.05$ menunjukkan bahwa tidak terdapat perbedaan signifikan, yang berarti kedua model memiliki kemampuan prediksi yang relatif setara.



Gambar 5. Confusion Matrix Model Terbaik

Analisis *Confusion Matrix* (Gambar 5) memperlihatkan bahwa kesalahan prediksi mayoritas terjadi pada kelas-kelas yang berdekatan (misalnya antara Grade 'BB' dan 'BA'). Namun, model mampu membedakan dengan baik antara kategori ekstrem (*Fail* vs *AA*), yang merupakan fitur krusial untuk sistem peringatan dini.

4. Kesimpulan

Penelitian ini berhasil mengembangkan dan mengevaluasi model SVM untuk prediksi nilai akhir mahasiswa. Berdasarkan hasil analisis, model SVM (Sigmoid) terbukti sebagai metode yang paling efektif dengan akurasi 31.03%. Meskipun model

baseline seperti K-Nearest Neighbors (KNN) juga memberikan hasil yang kompetitif, SVM menawarkan keunggulan dalam hal generalisasi.

Penelitian ini merekomendasikan implementasi model tersebut dalam sistem akademik institusi untuk mendeteksi dini mahasiswa yang membutuhkan bimbingan tambahan. Pengembangan selanjutnya disarankan untuk mengeksplorasi teknik *ensemble learning* guna meningkatkan akurasi pada kelas-kelas minoritas.

Daftar Pustaka

[1] E. Alpaydin, *Introduction to Machine Learning*, 4th ed. Cambridge, MA: MIT Press, 2020.
[2] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," *Procedia Computer Science*, vol. 72, pp. 414-422, 2015P. Hengpraprohm, P. Chongstitvatana, and N. Hengpraprohm, "A Study of Factors Affecting
[3] Learning Efficiency on Higher Education Student Performance Evaluation Dataset Using Feature Selection Techniques," *Journal of Physics: Conference Series*, vol. 22, no. 1, 2022.
[4] UCI Machine Learning Repository, "Higher Education Students Performance Evaluation Dataset," 2019. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Higher+Education+Students+Performance+Evaluation
[5] S. Wiyono and T. Abidin, "Comparative Study of KNN, SVM and Decision Tree Algorithm for Student's Performance Prediction," *International Journal of Computing Science and Applied Mathematics (IJCSAM)*, vol. 5, no. 1, pp. 16-19, 2019.