JIIC: JURNAL INTELEK INSAN CENDIKIA
https://jicnusantara.com/index.php/jiic
Vol : 3 No: 1, Januari 2026
E-ISSN : 3047-7824

Prediksi Kelulusan Tepat Waktu Berdasarkan Keaktifan Organisasi dan Pengembangan Diri Menggunakan *Algoritma Support Vector Machine*

Predicting On-Time Graduation Based on Organizational Activity and Self-Development Using the Support Vector Machine Algorithm

Rangga Saputra[1], Hasbi Firmansyah[2]
Universitas Pancasakti Tegal
Email: ranggasptra007@gmail.com[1], hasbifirmansyah@upstegal.ac.id[2]

Article Info

Article history :
Received : 28-12-2025
Revised : 30-12-2025
Accepted : 02-01-2026
Pulished : 04-01-2026

Abstract

Timely graduation is a vital indicator in the accreditation assessment of higher education institutions. Often, student involvement in non-academic activities such as organizations and communication forums is considered a hindering factor for graduation, even though these aspects contribute to soft skill development. This study aims to classify student graduation status (fast or late) by utilizing the Support Vector Machine (SVM) algorithm. The data used consists of 50 student samples with attributes including GPA, self-development training intensity, achievements, participation in communication forums, and organizational activities. Data processing was carried out using RapidMiner tools with the Cross Validation method. The test results showed an accuracy rate of 66.00%, with a Recall value for the "Fast Graduation" class reaching 100%. However, the model showed a bias tendency towards the majority class, indicating the need for handling imbalanced data in future research.

***Keywords**: Data Mining, Support Vector Machine, Graduation Prediction*

Abstrak

Ketepatan waktu kelulusan mahasiswa merupakan indikator vital dalam penilaian akreditasi perguruan tinggi. Seringkali, keterlibatan mahasiswa dalam kegiatan non-akademik seperti organisasi dan forum komunikasi dianggap sebagai faktor penghambat kelulusan, padahal aspek tersebut juga berkontribusi pada pengembangan *soft skill*. Penelitian ini bertujuan untuk mengklasifikasikan status kelulusan mahasiswa (lulus cepat atau terlambat) dengan memanfaatkan algoritma *Support Vector Machine* (SVM). Data yang digunakan terdiri dari 50 sampel mahasiswa dengan atribut meliputi IPK, intensitas pelatihan pengembangan diri, prestasi, keaktifan dalam forum komunikasi, dan kegiatan organisasi. Pengolahan data dilakukan menggunakan *tools* RapidMiner dengan metode validasi *Cross Validation*. Hasil pengujian menunjukkan tingkat akurasi sebesar 66,00%, dengan nilai *Recall* untuk kelas "Lulus Cepat" mencapai 100%. Namun, model menunjukkan kecenderungan bias terhadap kelas mayoritas, yang mengindikasikan perlunya penanganan ketidakseimbangan data (*imbalanced data*) pada penelitian selanjutnya.

Kata Kunci: *Data Mining, Support Vector Machine,* Prediksi Kelulusan

PENDAHULUAN

Perguruan tinggi dituntut untuk tidak hanya menghasilkan lulusan dengan kuantitas besar, tetapi juga menjaga kualitas dan ketepatan waktu studi. Lama masa studi mahasiswa menjadi salah satu poin krusial dalam borang akreditasi institusi maupun program studi(Tinggi & TERAKREDITASI, 2018). Fenomena yang sering terjadi di lingkungan akademik adalah adanya dikotomi antara pencapaian akademik (IPK) dan keterlibatan dalam kegiatan kemahasiswaan.

Banyak mahasiswa yang aktif dalam organisasi, forum komunikasi, atau pelatihan pengembangan diri menghadapi stigma bahwa kegiatan tersebut akan menghambat penyelesaian skripsi atau tugas akhir. Namun, di sisi lain, dunia industri saat ini menuntut lulusan yang memiliki *soft skill* mumpuni yang justru didapatkan dari kegiatan di luar kelas tersebut (Fujita, 2006). Oleh karena itu, diperlukan sebuah metode analitis untuk melihat pola hubungan antara aktivitas non-akademik tersebut dengan peluang kelulusan tepat waktu.

Educational Data Mining (EDM) hadir sebagai solusi untuk mengekstraksi pengetahuan dari data akademik guna membantu pengambilan keputusan strategis(Romero & Ventura, 2010). Berbagai algoritma klasifikasi telah diterapkan dalam EDM, seperti *Decision Tree*, *Naive Bayes*, dan *K-Nearest Neighbor*. Namun, penelitian ini memilih algoritma *Support Vector Machine* (SVM). SVM dikenal memiliki performa yang tangguh dalam menangani data dengan dimensi yang variatif dan mampu memisahkan kelas data dengan *margin* yang optimal, bahkan pada dataset dengan jumlah sampel terbatas(Hearst et al., 1998).

Penelitian ini berfokus pada prediksi status kelulusan (Lulus Cepat atau Terlambat) dengan memasukkan variabel yang jarang digabungkan secara bersamaan, yaitu intensitas pelatihan pengembangan diri dan keaktifan forum komunikasi, selain variabel standar seperti IPK dan Organisasi. Tujuannya adalah untuk membuktikan secara empiris apakah keaktifan non-akademik benar-benar menjadi prediktor kuat terhadap keterlambatan atau justru sebaliknya, serta mengukur performa SVM dalam mengklasifikasikan pola tersebut menggunakan bantuan *tools* RapidMiner.

METODE PENELITIAN

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari repositori data publik Kaggle. Dataset tersebut berfokus pada rekam jejak akademik dan non-akademik mahasiswa, yang mencerminkan berbagai aspek aktivitas selama masa studi. Atribut yang digunakan meliputi Indeks Prestasi Kumulatif (IPK), frekuensi keikutsertaan dalam pelatihan pengembangan diri atau soft skill, jumlah prestasi atau penghargaan yang diraih, tingkat keaktifan mahasiswa dalam forum komunikasi perkuliahan, serta tingkat keterlibatan dalam kegiatan organisasi kemahasiswaan. Selain itu, terdapat atribut target berupa label biner “Lulus Cepat”, di mana nilai 1 merepresentasikan mahasiswa yang lulus tepat waktu atau lebih cepat, sedangkan nilai 0 merepresentasikan mahasiswa yang mengalami keterlambatan kelulusan.

Sebelum dilakukan proses pemodelan, data mentah terlebih dahulu melalui tahapan preprocessing guna memastikan kualitas dan kesiapan data sebagai input model. Tahapan ini mencakup pengecekan terhadap nilai yang hilang (missing values) serta proses normalisasi data. Normalisasi diperlukan karena algoritma Support Vector Machine (SVM) merupakan metode berbasis perhitungan jarak (distance-based), sehingga perbedaan skala antar atribut—seperti IPK yang berada pada skala 4,0 dan atribut lain yang dapat bernilai puluhan—berpotensi menimbulkan bias dalam proses pembelajaran. Oleh karena itu, seluruh atribut numerik dinormalisasi agar berada pada rentang skala yang setara(Han et al., 2012).

Metode klasifikasi yang digunakan dalam penelitian ini adalah Support Vector Machine (SVM), yang bekerja dengan prinsip pencarian hyperplane optimal untuk memisahkan data ke dalam dua kelas yang berbeda. Dalam konteks penelitian ini, SVM digunakan untuk memisahkan mahasiswa yang termasuk kategori “Lulus Cepat” dan “Terlambat”. Untuk menangani

kemungkinan data yang tidak dapat dipisahkan secara linear, fungsi kernel diterapkan guna memetakan data ke ruang berdimensi lebih tinggi. Implementasi algoritma SVM dalam penelitian ini dilakukan menggunakan perangkat lunak RapidMiner dengan pengaturan standar(Anwar, 2025).

Evaluasi kinerja model dilakukan menggunakan Confusion Matrix untuk mengukur kemampuan prediksi model secara menyeluruh. Metrik evaluasi yang menjadi fokus utama adalah Accuracy, yang menunjukkan tingkat ketepatan prediksi model, serta Recall, yang mengukur kemampuan model dalam mengidentifikasi mahasiswa yang benar-benar lulus cepat. Pemilihan metrik Recall menjadi penting karena kesalahan dalam mendeteksi mahasiswa yang berpotensi lulus cepat atau sebaliknya dapat memiliki implikasi yang signifikan dalam pengambilan keputusan manajerial di lingkungan pendidikan(Hoffait & Schyns, 2017).

HASIL DAN PEMBAHASAN

Tahap inti dari penelitian ini melibatkan pengujian model klasifikasi menggunakan algoritma *Support Vector Machine* (SVM). SVM dipilih karena reputasinya dalam menangani ruang fitur berdimensi tinggi dan kemampuannya untuk menemukan *hyperplane* pemisah yang optimal. Namun, hasil pengujian pada dataset ini mengungkapkan anomali kinerja yang memerlukan analisis kritis mendalam, terutama terkait dengan fenomena ketidakseimbangan kelas (*class imbalance*).

Proses eksperimen dilakukan menggunakan perangkat lunak RapidMiner Studio. Dataset diimpor dan diproses sesuai tahapan metodologi.

1. Analisis Pola Data

Berdasarkan eksplorasi awal pada data *lulus.csv*, ditemukan variasi menarik. Beberapa mahasiswa dengan IPK di atas 3.50 ternyata tidak selalu masuk dalam kategori lulus cepat jika aktivitas organisasinya sangat rendah atau nol. Sebaliknya, terdapat pola dimana mahasiswa dengan IPK moderat (3.00 - 3.25) namun memiliki skor tinggi pada "Forum Komunikasi Kuliah" dan "Pelatihan Pengembangan Diri" justru masuk kategori lulus cepat. Ini mengindikasikan bahwa kemampuan komunikasi dan manajemen diri berkontribusi pada efisiensi penyelesaian studi.

2. Analisis Statistik Matriks Kebingungan*(Confusion Matrix)*

Pengujian dilakukan pada total 500 sampel data uji, dengan distribusi kelas aktual yang menunjukkan ketidakseimbangan signifikan:

a. Kelas Positif (Lulus Cepat): 330 sampel (66% dari populasi)
b. Kelas Negatif (Terlambat): 170 sampel (34% dari populasi)

Hasil prediksi model dirangkum dalam *Confusion Matrix* berikut:

Klasifikasi	Prediksi: Lulus Cepat	Prediksi: Terlambat	Total Aktual
Aktual: Lulus Cepat	330 (True Positive - TP)	0 (False Negative - FN)	330
Aktual: Terlambat	170 (False Positive - FP)	0 (True Negative - TN)	170
Total Prediksi	500	0	500

Dari tabel di atas, metrik evaluasi kinerja model dihitung sebagai berikut:

a. Akurasi Global:

$$Accuracy = \frac{TP + TN}{Total\ Sampel} = \frac{330 + 0}{500} = 66.00\%$$

b. Recall (Sensitivitas) pada Kelas Lulus Cepat:

$$Recall = \frac{TP}{TP + FN} = \frac{330}{330 + 0} = 100.00\%$$

c. Precision pada Kelas Lulus Cepat:

$$Precisoon = \frac{TP}{TP + FP} = \frac{330}{330 + 170} = 66.00\%$$

d. Specificity (Tingkat True Negative) pada Kelas Terlambat:

$$Specificity = \frac{TN}{TN + FP} = \frac{0}{0 + 170} = 0.00\%$$

3. Ilusi metrik: Menafsirkan Recall 100% dan Akurasi 66%

Angka statistik di atas menyajikan gambaran yang menipu (*misleading*) jika tidak dianalisis dengan konteks distribusi data. Poin pembahasan yang menyebutkan "nilai Recall sebesar 100% adalah poin yang sangat menarik" perlu dikoreksi dengan pemahaman bahwa pencapaian ini bersifat trivial dan artifisial.

Nilai Recall 100% memang menunjukkan bahwa *tidak ada* mahasiswa yang lulus cepat yang terlewat oleh sistem (Zero False Negative). Secara teoritis, ini berarti algoritma sangat sensitif terhadap kelas positif. Namun, sensitivitas ini dicapai dengan biaya yang tidak masuk akal: model memprediksi seluruh data, tanpa terkecuali, sebagai kelas mayoritas ("Lulus Cepat"). Ketika penyebut dalam rumus Recall ($TP + FN$) hanya berisi TP karena $FN=0$ akibat prediksi sapu jagat (*blanket prediction*), maka nilai Recall akan selalu sempurna.

Nilai Akurasi 66% juga merupakan representasi dari apa yang disebut sebagai Paradoks Akurasi (*Accuracy Paradox*). Dalam dataset tidak seimbang, sebuah model naif yang hanya menebak kelas mayoritas sepanjang waktu akan selalu menghasilkan akurasi yang setara dengan proporsi kelas mayoritas tersebut. Di sini, 66% bukanlah ukuran kecerdasan model dalam membedakan pola, melainkan hanya cerminan statistik dari distribusi populasi dataset (330 dari 500). Model ini, pada dasarnya, tidak memiliki kemampuan diskriminatif sama sekali (*zero discriminatory power*).

4. Kegagalan Fatal pada Spesifitas dan Dampaknya

Kegagalan terbesar model ini terletak pada nilai Specificity 0%. Model gagal total mendeteksi satu pun mahasiswa yang "Terlambat". Seluruh 170 mahasiswa yang berisiko terlambat diklasifikasikan secara salah sebagai "Lulus Cepat" (*False Positive*). Dalam konteks sistem peringatan dini akademik, kesalahan tipe ini (False Positive dalam konteks prediksi kelulusan, atau False Negative dalam konteks deteksi risiko) adalah yang paling berbahaya dan mahal. Jika sistem ini diimplementasikan:

a. Institusi akan memiliki rasa aman palsu bahwa semua mahasiswa baik-baik saja.
b. 170 mahasiswa yang sebenarnya membutuhkan intervensi bimbingan akademik intensif akan terabaikan.
c. Peluang untuk mencegah *dropout* atau keterlambatan akan hilang sepenuhnya. Prioritas dalam prediksi risiko pendidikan seharusnya adalah meminimalkan kegagalan deteksi risiko, yang berarti metrik *Recall* pada kelas minoritas ("Terlambat") seharusnya menjadi prioritas utama, bukan *Recall* pada kelas mayoritas.

5. Pembahasan: Mekanis Kegagalan SVM pada Data Tidak Seimbang dan Ber-Noise

Mengapa algoritma sekuat SVM, yang dirancang untuk menemukan margin optimal, gagal total dalam kasus ini? Jawabannya terletak pada prinsip matematis dasar SVM dan interaksinya dengan karakteristik intrinsik data: ketidakseimbangan (*imbalance*) dan tumpang tindih (*overlap*) fitur. Pada dataset lulus.csv, rasio ketidakseimbangan adalah sekitar 2:1. Meskipun ini bukan ketidakseimbangan ekstrem (seperti 1:1000 pada deteksi penipuan), ini sudah cukup untuk membiaskan SVM standar. Karena jumlah sampel kelas "Lulus Cepat" (330) jauh lebih banyak daripada kelas "Terlambat" (170), kontribusi kumulatif dari variabel *slack* (ξ) kelas mayoritas terhadap fungsi kerugian global jauh lebih besar.

Untuk meminimalkan total *loss*, SVM menemukan bahwa strategi yang paling "murah" secara komputasi adalah dengan menggeser *hyperplane* pemisah sejauh mungkin ke arah kelas minoritas, atau bahkan menempatkannya sedemikian rupa sehingga seluruh ruang fitur dianggap sebagai wilayah kelas mayoritas.[23] Dengan melakukan ini, SVM menerima kesalahan pada 170 sampel minoritas (sebagai biaya yang dapat ditoleransi) demi memastikan 330 sampel mayoritas diklasifikasikan dengan benar dan margin dimaksimalkan. Tanpa mekanisme pembobotan biaya (*cost-weighting*) yang memberikan penalti lebih berat untuk kesalahan pada kelas minoritas, SVM secara alami akan berlaku bias terhadap kelas mayoritas.

Nilai akurasi sebesar 66% menunjukkan performa yang moderat. Namun, poin yang sangat menarik adalah nilai *Recall* sebesar 100%. Artinya, algoritma SVM berhasil mengenali seluruh mahasiswa yang benar-benar lulus cepat tanpa ada yang terlewat (tidak ada *False Negative*).

Akan tetapi, model mengalami tantangan dalam memprediksi kelas "Terlambat". Model cenderung memprediksi semua data sebagai "Lulus Cepat" (terlihat dari kolom Prediksi Terlambat yang bernilai 0). Hal ini mengindikasikan adanya isu *Imbalanced Data* (ketidakseimbangan data) yang parah pada dataset, dimana jumlah sampel "Lulus Cepat" jauh mendominasi dibandingkan sampel yang terlambat (Chawla et al., 2002).

Meskipun bias terhadap kelas mayoritas, hasil ini tetap memberikan wawasan bahwa atribut *Keaktifan Organisasi* dan *Pengembangan Diri* memiliki pola yang konsisten pada mahasiswa yang lulus cepat, sehingga SVM dengan mudah mengelompokkannya ke dalam kelas positif. Penelitian oleh Sasa dkk. (Moonallika et al., 2020) juga menyebutkan bahwa atribut perilaku seringkali memiliki *noise* yang membuat batas pemisah (*hyperplane*) sulit ditentukan secara presisi tanpa kernel yang sangat kompleks atau penyeimbangan data.

KESIMPULAN

Penerapan algoritma *Support Vector Machine* (SVM) untuk memprediksi kelulusan mahasiswa berdasarkan atribut akademik dan non-akademik menghasilkan akurasi sebesar 66,00%. Penelitian ini menyimpulkan bahwa aktivitas organisasi dan pengembangan diri tidak selalu menjadi penghambat kelulusan; justru pola data menunjukkan kontribusi positif. Keunggulan model ini terletak pada sensitisitasnya yang tinggi (*Recall* 100%) dalam mendeteksi mahasiswa yang berpotensi lulus tepat waktu. Namun, kelemahannya terletak pada ketidakmampuan model membedakan mahasiswa yang terlambat lulus karena bias data mayoritas.

SARAN

Untuk penelitian selanjutnya, sangat disarankan menerapkan teknik *resampling* seperti SMOTE (*Synthetic Minority Over-sampling Technique*) untuk mengatasi ketidakseimbangan data agar prediksi kelas minoritas (terlambat lulus) menjadi lebih akurat. Selain itu, penambahan variabel eksternal seperti status ekonomi atau latar belakang sekolah asal dapat memperkaya fitur prediksi.

DAFTAR PUSTAKA

Anwar, N. R. (2025). *ANALISIS ALGORITMA SUPPORT VECTOR MACHINE UNTUK KLASIFIKASI OBAT TERHADAP PASIEN PRECISION MEDICINE THROUGH SUPPORT VECTOR MACHINES: ANALYZING PATIENT DATA FOR IMPROVED DRUG CLASSIFICATION*. Universitas Pelita Bangsa.

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Fujita, K. (2006). The effects of extracurricular activities on the academic performance of junior high students. *Undergraduate Research Journal for the Human Sciences*, *5*(1), 1–16.

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and. *Techniques, Waltham: Morgan Kaufmann Publishers*.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications*, *13*(4), 18–28.

Hoffait, A.-S., & Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems*, *101*, 1–11.

Moonallika, P. S. C., Fredlina, K. Q., & Sudiatmika, I. B. K. (2020). Penerapan Data Mining Untuk Memprediksi Kelulusan Mahasiswa Menggunakan Algoritma Naive Bayes Classifier (Studi Kasus STMIK Primakara). *Progresif: Jurnal Ilmiah Komputer*, *16*(1), 47–56.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618.

Tinggi, B. A. N. P., & TERAKREDITASI, S. A. D. A. N. P. (2018). *Badan Akreditasi Nasional Perguruan Tinggi*.