JNATIA Volume 2, Nomor 3, Mei 2024 p-ISSN: 2986-3929
Jurnal Nasional Teknologi Informasi dan Aplikasinya e-ISSN: 3032-1948

Memprediksi Kelulusan Mahasiswa: Graduate dan Dropout dengan Support Vector Machine dan GridSearchCV

Ni Putu Eka Marita Anggarini[a1], Agus Muliantara[a2]

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Udayana
Jalan Raya Kampus UNUD, Bukit Jimbaran, Kuta Selatan, Badung, Bali, Indonesia
[1]anggarini.2208561032@student.unud.ac.id
[2]muliantara@unud.ac.id

Abstract

In today's educational landscape, having a model to predict whether a student will graduate or drop out based on their academic statistics is highly beneficial. Such a model allows for early assessment of academic success. Human calculations alone can be time-consuming and often lack accuracy, hence the introduction of machine learning models to address this issue. This research utilizes a dataset comprising undergraduate students from various majors in higher education institutions. The data were collected while the students were still enrolled, with their grades from the first year serving as a key feature. The response variable in the dataset is labeled as either 'dropout' or 'graduate'. We employ Support Vector Machines (SVM) with GridSearchCV optimization to build the predictive model. The goal of this model is to predict a student's academic success as early as their first-year statistics are available. If a student is predicted to drop out, targeted interventions can be provided to help them overcome challenges, ultimately aiming to improve graduation rates.

Keywords: *siswa, akademik, dropout, graduate, SVM, hyperparameter tuning, klasifikasi, prediksi, machine learning, GridSearchCV*

1. Pendahuluan

Mengetahui apakah seorang siswa akan drop out atau lulus sangat penting sejak dini, bahkan sejak tahun pertama mereka. Hal ini penting agar siswa bisa memahami kondisi mereka dan mendapatkan dukungan yang diperlukan untuk membantu mereka dalam akademis, sehingga pada akhirnya mereka semua dapat lulus. Prediksi dan klasifikasi awal tingkat kinerja siswa memberikan peringatan dini dan memberikan resep untuk meningkatkan kinerja siswa yang kurang baik serta untuk pengaturan manajerial lainnya [1]. Di era revolusi informasi, analisis basis data dalam lingkungan pendidikan seperti analitik pembelajaran, analitik prediktif, penambangan data pendidikan, dan teknik *machine learning* telah menjadi bidang penelitian yang populer [2]. *Machine Learning* adalah proses mempelajari serangkaian aturan dari contoh-contoh atau lebih umum lagi, menciptakan sebuah pengklasifikasi yang dapat digunakan untuk menggeneralisasi dari contoh-contoh baru. Pembuatan pengklasifikasi adalah proses dua langkah. Pada langkah pertama, model pengklasifikasi dibangun menggunakan set data yang diberikan. Langkah ini disebut *training*. Pada langkah ini, aturan-aturan klasifikasi dibuat. Langkah kedua disebut *testing*, yang menentukan akurasi aturan-aturan klasifikasi [3]. *Machine learning* digunakan untuk memprediksi, mengklasifikasikan kinerja siswa, dan menganalisis perilaku belajar mereka guna memantau kemajuan mereka di bidang akademik. Namun, yang menjadi tantangan adalah menemukan algoritma optimal yang dapat menghasilkan hasil yang memuaskan [1]. Algoritma *machine learning* seperti *naïve Bayes, logistic regression, artificial neural network, decision trees, random forest, support vector machine, k-nearest neighbor,* dan lainnya, banyak digunakan untuk menganalisis dan memprediksi kinerja akademik [2]. Dalam penelitian ini, diusulkan penggunaan model pembelajaran mesin untuk memprediksi keberhasilan akademik siswa, yang akan mengeliminasi kebutuhan perhitungan manual. Penggunaan model pembelajaran mesin tidak

hanya menghemat waktu tetapi juga meningkatkan akurasi dibandingkan perhitungan manual. Model ini akan menggunakan teknologi terkini dalam pembelajaran mesin dan teknik optimasi, khususnya model Support Vector Machine (SVM) dan algoritma tuning hyperparameter GridSearchCV untuk mencapai akurasi tertinggi. Dengan memanfaatkan dataset dari University of California, Irvine, model pembelajaran mesin dilatih untuk memprediksi apakah seorang siswa akan drop out atau lulus.

2. Metode Penelitian

2.1 Dataset

Untuk melatih model pembelajaran mesin agar dapat memprediksi apakah seseorang akan lulus, putus sekolah, atau masih terdaftar sebagai mahasiswa, terlebih dahulu dibutuhkan dataset yang berisi catatan informasi pendidikan atau atribut pribadi mahasiswa dan status mereka, apakah lulus, putus sekolah, atau masih terdaftar. Dataset yang digunakan dalam penelitian ini diperoleh dari Kaggle.com, sebuah repositori sumber terbuka untuk dataset pembelajaran mesin, yang berasal dari University of California, Irvine, terkait dengan mahasiswa yang terdaftar dalam berbagai program sarjana. Dataset ini mencakup informasi yang diketahui saat pendaftaran mahasiswa (jalur akademik, demografi, dan faktor sosial-ekonomi) dan kinerja akademik mahasiswa (GDP) pada akhir tahun pertama kuliah. Data ini memiliki tiga kategori sebagai target: putus sekolah, terdaftar, dan lulus. Dengan 4424 instance, dataset ini mencakup 36 fitur. Namun dalam penelitian ini, label "*Enrolled*" pada variabel target dihapus, sehingga hanya 3630 baris yang akan digunakan, dan hanya akan ada dua kategori sebagai target: putus sekolah dan lulus. Fitur target adalah variabel respon yang dihitung menggunakan variabel lainnya. Deskripsi masing-masing fitur dan variabel target disajikan dalam Tabel 1.

Tabel 1. Karakteristik Basis Data

Nama Variabel	Tipe Variable	Deskripsi	Nilai
Marital Status	*Feature*	*Status of marriages/relationship*	*Any value from 1-6 (6 unique marital statuses)*
Application mode	*Feature*	*Mode of application (e.g., change of course, transfer, etc.)*	*Some value from 1-57 (17 unique application modes)*
Application order	*Feature*	*Order of application (e.g., first choice, last choice, etc.)*	*Any value from 0-9 (10 unique application orders)*
Course	*Feature*	*Type of courses*	*Some value from 33-9991 (17 unique courses)*
Daytime/evening attendance	*Feature*	*Attendance time (e.g., evening, daytime)*	*0,1*
Previous qualification	*Feature*	*Previous education level*	*Some value from 1-43 (17 unique education levels)*
Previous qualification (grade)	*Feature*	*Grade of previous qualification*	*Any values between 0-200*
Nacionality	*Feature*	*Nationality*	*Some value from 1-109 (21 unique nationalities)*
Mother's qualification	*Feature*	*Education level*	*Some value from 1-44 (29 unique education levels)*
Father's qualification	*Feature*	*Education level*	*Some value from 1-44 (29 unique education levels)*

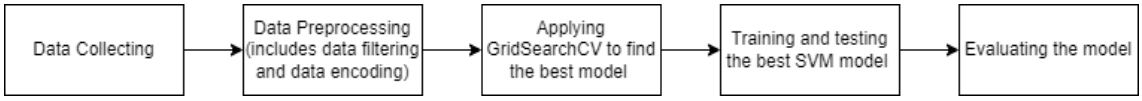
p-ISSN: 2986-3929
e-ISSN: 3032-1948

Nama Variabel	Tipe Variable	Deskripsi	Nilai
Mother's occupation	*Feature*	*Types of job*	*Some values from 0-194 (32 unique jobs)*
Father's occupation	*Feature*	*Type of job*	*Some values from 0-195 (46 unique jobs)*
Admission grade	*Feature*	*Admission grade*	*Any values between 95-190*
Displaced	*Feature*	*1 – yes 0 – no if displaced*	0,1
Educational special needs	*Feature*	*1 – yes 0 – no if needs educational special needs*	0,1
Debtor	*Feature*	*1 – yes 0 – no if a debtor*	0,1
Tuition fees up to date	*Feature*	*1 – yes 0 – no if tuition fees are up to date*	0,1
Gender	*Feature*	*1 – male 0 – female*	0,1
Scholarship holder	*Feature*	*1 – yes 0 – no if a scholarship holder*	0,1
Age at enrollment	*Feature*	*Age of student at enrollment*	*Any values from 17-70*
International	*Feature*	*1 – yes 0 – no if an international student*	0,1
Curricular units 1st sem (credited)	*Feature*	*Number of curricular units credited in the 1st semester*	*Any values from 0-20*
Curricular units 1st sem (enrolled)	*Feature*	*Number of curricular units enrolled in the 1st semester*	*Any values from 0-26*
Curricular units 1st sem (evaluations)	*Feature*	*Number of evaluations to curricular units in the 1st semester*	*Any values from 0-45*
Curricular units 1st sem (approved)	*Feature*	*Number of curricular units approved in the 1st semester*	*Any values from 0-26*
Curricular units 1st sem (grade)	*Feature*	*Grade average in the 1st semester*	*Any values from 0-18.875*
Curricular units 1st sem (without evaluations)	*Feature*	*Number of curricular units without evalutions in the 1st semester*	*Any values from 0-12*
Curricular units 2nd sem (credited)	*Feature*	*Number of curricular units credited in the 2nd semester*	*Any values from 0-19*
Curricular units 2nd sem (enrolled)	*Feature*	*Number of curricular units enrolled in the 2nd semester*	*Any values from 0-23*
Curricular units 2nd sem (evaluations)	*Feature*	*Number of evaluations to curricular units in the 2nd semester*	*Any values from 0-33*
Curricular units 2nd sem (approved)	*Feature*	*Number of curricular units approved in the 2nd semester*	*Any values from 0-20*
Curricular units 2nd sem (grade)	*Feature*	*Grade average in the 2nd semester*	*any values from 0-18.6*

Nama Variabel	Tipe Variable	Deskripsi	Nilai
Curricular units 2nd sem (without evaluations)	*Feature*	*Number of curricular units without evalutions in the 1st semester*	*Any values from 0-12*
Unemployment rate	*Feature*	*Unemployment rate (%)*	*Any values from 7.6 to 16.2*
Inflation rate	*Feature*	*Inflation rate (%)*	*Any values from (-0.8) to 3.7*
GDP	*Feature*	*GDP*	*Any values from (-4.06) to 3.51*
Target	*Target*	*Dropout, Graduate*	0,1

2.2 *Support Vector Machine*

Semua prosedur dalam penelitian ini dilakukan menggunakan bahasa pemrograman Python, lingkungan pengembangan Google Colaboratory, dan pustaka pembelajaran mesin sklearn. Untuk melatih model pembelajaran mesin, diperlukan set pelatihan dan set pengujian. Oleh karena itu, dataset dibagi secara programatis. Pembagian data ini dilakukan secara acak. Konsep utama dari SVM adalah untuk mendapatkan Hyperplane Pemisah Optimal (OSH) antara sampel positif dan negatif. Ini dapat dilakukan dengan memaksimalkan margin antara dua hyperplane paralel. Dengan menemukan hyperplane ini, SVM kemudian dapat memprediksi klasifikasi sampel yang tidak berlabel dengan menentukan di sisi mana dari hyperplane pemisah sampel tersebut berada. SVM mampu menangani berbagai jenis masalah klasifikasi seperti masalah klasifikasi linier dan non-linier [4]. Linear SVM digunakan dengan data yang dapat dipisahkan secara linier, artinya data tidak memerlukan transformasi untuk memisahkan kelas yang berbeda. Jika data tidak dapat dipisahkan secara linier (*non-linear data*), SVM menyelesaikannya dengan membuat variabel baru menggunakan kernel. Kernel adalah fungsi matematis yang digunakan untuk memetakan titik data input asli ke ruang fitur berdimensi tinggi, sehingga *hyperplane* dapat dengan mudah ditemukan meskipun titik data tidak dapat dipisahkan secara linier dalam ruang input asli. Beberapa fungsi kernel yang umum digunakan adalah linear, polynomial, radial basis function (RBF), dan sigmoid. Dalam pengaplikasiannya, pertama, data akan dibagi menjadi X_train dan y_train dengan variabel 'Target' bertindak sebagai label model kami. Kemudian, data dibagi menjadi 80% sebagai data pelatihan dan 20% sebagai data pengujian. Sebelum membuat model dan memasukkan data ke dalamnya, metode *hyperparameter tuning* dengan *GridSearchCV* diterapkan untuk mengoptimalkan model agar mencapai akurasi tertinggi yang mungkin. Berikut adalah gambaran dari metode yang akan digunakan.



Gambar 1. Metode Aplikasi SVM dan GridSearchCV

2.3 *GridSearchCV*

GridSearchCV (Grid Search Cross-Validation) adalah sebuah teknik yang digunakan untuk mengatur hyperparameter dan pemilihan model [5]. *GridSearchCV* mengotomatiskan proses untuk menemukan set parameter yang optimal dan menghindari overfitting [6]. Metode ini secara sistematis menguji semua kombinasi yang mungkin dari *hyperparameter* yang disediakan untuk menemukan set yang memberikan kinerja terbaik untuk model yang diberikan. Proses ini melibatkan pembagian dataset menjadi set pelatihan dan validasi beberapa kali menggunakan *cross-validation*, yang memastikan setiap kombinasi parameter dievaluasi secara adil dan mengurangi risiko *overfitting*. Untuk setiap kombinasi, model dilatih pada set pelatihan dan dievaluasi pada set validasi. Metrik kinerja dari setiap lipatan *cross-validation* dirata-ratakan untuk menentukan efektivitas keseluruhan dari set parameter tersebut. Kombinasi yang menghasilkan

p-ISSN: 2986-3929
e-ISSN: 3032-1948

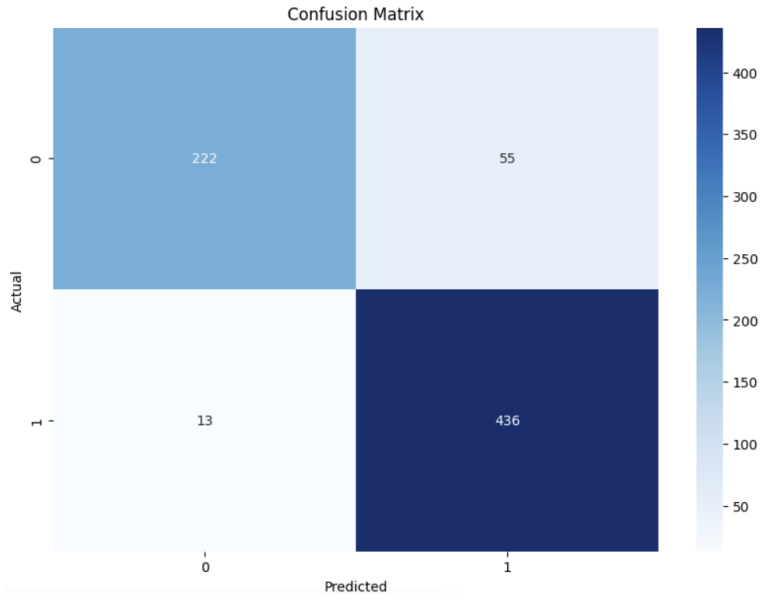
kinerja rata-rata tertinggi dipilih sebagai set *hyperparameter* optimal. Metode ini menyediakan pendekatan komprehensif untuk penyetelan *hyperparameter*, meningkatkan kinerja prediktif dan kemampuan generalisasi model. Dalam pengaplikasiannya di penelitian ini, parameter yang digunakan meliputi parameter C, degree, dan kernel dengan berbagai nilai. Ide utamanya adalah model akan dibuat menggunakan semua nilai berbeda untuk setiap parameter, dan yang memiliki akurasi terbaik akan dipilih sebagai parameter terbaik dan estimator terbaik yang akan digunakan untuk model akhir.

3. Hasil dan Diskusi

Setelah menguji berbagai kombinasi *hyperparameter* menggunakan metode *GridSearchCV*, kami menemukan bahwa parameter terbaik untuk model *Support Vector Machine* (SVM) adalah C: 100, degree: 1, dan kerne: poly. Kombinasi ini menghasilkan estimator terbaik yaitu SVC (C=100, degree=1, kernel=poly). Langkah selanjutnya adalah menerapkan model ini pada data pelatihan (X_train dan y_train). Model dilatih untuk mengenali pola dan hubungan dalam data pelatihan dengan menggunakan parameter optimal yang telah ditemukan. Setelah pelatihan selesai, model kemudian diuji pada data pengujian untuk mengevaluasi kinerjanya. Hasilnya menunjukkan bahwa model tersebut mencapai tingkat akurasi sebesar 0.90633608815427 pada data pengujian, yang menunjukkan bahwa model tersebut mampu memprediksi dengan tingkat ketepatan yang tinggi. Berikut adalah laporan klasifikasi yang merinci kinerja model pada data pengujian, yang meliputi metrik seperti precision, recall, f1-score, dan support untuk masing-masing kelas target:

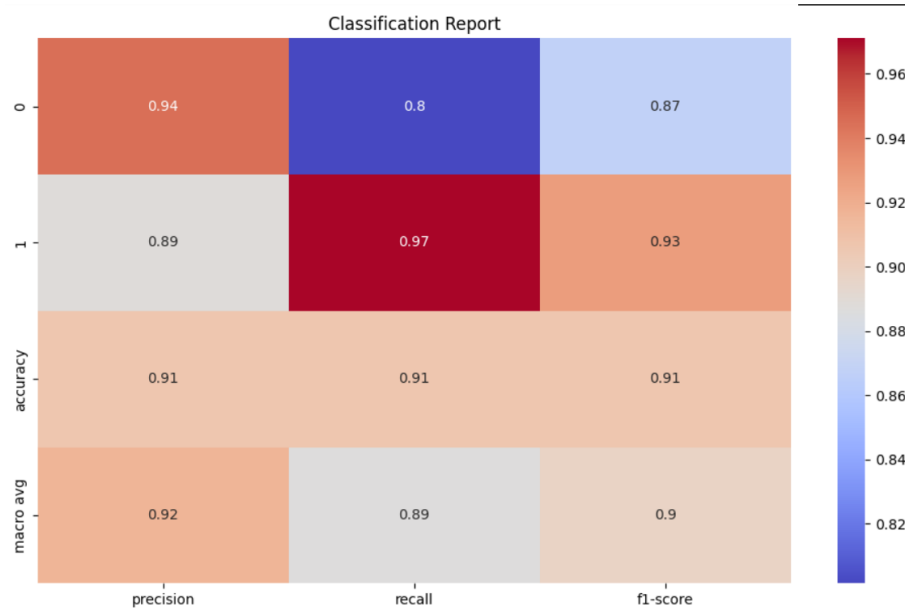
Tabel 2. *Classification Report*

	Precision	*Recall*	*F1-Score*	*Support*
0 *(Dropout)*	0.94	0.8	0.87	277
1 *(Graduate)*	0.89	0.97	0.93	449
Accuracy			0.91	726
Macro Avg	0.92	0.89	0.9	726
Weighted Avg	0.91	0.91	0.9	726



Gambar 2. *Confusion Matrix*

p-ISSN: 2986-3929
e-ISSN: 3032-1948



Gambar 3. *Classification Report*

4. Kesimpulan

Algoritma SVM dapat memprediksi kesuksesan akademik mahasiswa dengan baik menggunakan penyetelan *hyperparameter* dengan *GridSearchCV*. Model terbaik dari algoritma SVM menggunakan parameter C: 100, degree: 1, dan kernel: poly yang diperoleh dengan menggunakan metode *GridSearchCV*. Model ini mencapai akurasi hampir 91% yang akan baik untuk memprediksi keberhasilan akademik mahasiswa berdasarkan statistik tahun pertama mereka untuk menentukan apakah mereka akan lulus atau putus sekolah. Ini akan bermanfaat karena mahasiswa dengan kinerja akademik yang kurang baik dapat mendapatkan akomodasi sehingga mereka dapat berprestasi lebih baik dan tidak putus sekolah pada akhir tahun universitas.

Daftar Pustaka

[1] S. Phauk and O. Takeo, “Hybrid Machine Learning Algorithms for Predicting Academic Performance,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no.1, pp. 32-41, 2020.

[2] S. Slater, S. Joksimovic, V. Kovanovic, R.s Baker, and D. Gasevic, "Tools for Educational Data Mining: A Review", *Journal of Educational and Behavioral Statistics*, Vol. 42, No. 1, 2016, pp. 88-106.

[3] W. Slamet and A. Taufiq, “Comparative Study of Machine Learning KNN, SVM, and Decision Tree Algorithm to Predict Student’s Performance”, *International Journal of Research - Granthaalayah*, vol. 7, no.1, pp. 190-196, 2019.

[4] K. Motaz, A. Tarek, and S. Ghada, "A Comparison among Support Vector Machine and other Machine Learning Classification Algorithms," *International Journal of Computer Science*, vol. 3, no.5, pp. 25-45, 2015.

[5] T. Yan, S.-L. Shen, A. Zhou, and X. Chen, “Prediction of geological characteristics from shield operational parameters by integrating grid search and K-fold cross validation into stacking classification algorithm, "*Journal of Rock Mechanics and Geotechnical Engineering*, vol. 14, no. 4, pp. 1292–1303, Aug. 2022: https://doi.org/10.1016/j.jrmge.2022.03.002.

[6] A. Talal, “Using Artificial Neural Networks with GridSearchCV for Predicting Indoor Temperature in a Smart Home”, *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13437-13443, 2024.