

KLASIFIKASI TINGKAT PENDIDIKAN PROVINSI DI INDONESIA MENGGUNAKAN SVM DENGAN HYPERPARAMETER OPTIMIZATION OPTUNA

Michael Dolly Sianturi

Program Studi Statistika, FMIPA, Universitas Negeri Medan

e-mail : michael.4233260010@mhs.unimed.ac.id

Abstrak

Ketimpangan capaian pendidikan antar provinsi di Indonesia masih menjadi permasalahan utama, yang tercermin dari variasi nilai Rata-rata Lama Sekolah (RLS) yang cukup lebar. Penelitian ini bertujuan untuk mengklasifikasikan tingkat pendidikan 34 provinsi di Indonesia ke dalam tiga kategori (Rendah, Sedang, Tinggi) berdasarkan indikator sosial-ekonomi periode 2016-2025. Metode yang digunakan adalah Support Vector Machine (SVM) dengan optimasi hyperparameter menggunakan framework Optuna serta divalidasi dengan Time Series Cross-Validation untuk menghindari kebocoran data pada data deret waktu. Penelitian ini membandingkan empat fungsi kernel SVM yaitu RBF, Polynomial, Linear, dan Sigmoid. Hasil penelitian menunjukkan bahwa kernel RBF menghasilkan akurasi cross-validation tertinggi sebesar 89,23% dengan parameter optimal $C = 999,05$ dan $\gamma = 0,00471$. Pada pengujian data periode 2023-2025, kernel RBF mencapai akurasi 85,29% (29 dari 34 provinsi diprediksi benar), dengan precision macro 88,7% dan F1-score macro 82,9%. Kesalahan klasifikasi terjadi pada 5 provinsi yang memiliki nilai RLS di sekitar batas ambang antar kelas (9,00-9,38 tahun). Terdapat perbedaan performa yang signifikan antara keempat kernel SVM, dengan kernel RBF terbukti paling unggul dalam menangkap pola non-linear data sosial-ekonomi provinsi di Indonesia.

Kata Kunci: Klasifikasi, Support Vector Machine, Optuna, Rata-rata Lama Sekolah, Time Series Cross-Validation

Abstract

Educational attainment disparity among provinces in Indonesia remains a major problem, reflected in the wide variation of Mean Years of Schooling (MYS) values. This study aims to classify the education level of 34 provinces in Indonesia into three categories (Low, Medium, High) based on socio-economic indicators for the 2016-2025 period. The method used is Support Vector Machine (SVM) with hyperparameter optimization using the Optuna framework and validated using Time Series Cross-Validation to avoid data leakage in time series data. This study compares four SVM kernel functions: RBF, Polynomial, Linear, and Sigmoid. The results show that the RBF kernel achieved the highest cross-validation accuracy of 89.23% with optimal parameters $C = 999.05$ and $\gamma = 0.00471$. On the 2023-2025 testing data, the RBF kernel achieved an accuracy of 85.29% (29 out of 34 provinces correctly predicted), with macro precision of 88.7% and macro F1-score of 82.9%. Classification errors occurred in 5 provinces with MYS values around the class boundaries (9.00-9.38 years). There are significant performance differences among the four SVM kernels, with the RBF kernel proven to be the most superior in capturing the non-linear patterns of socio-economic data of Indonesian provinces.

Keywords: Classification, Support Vector Machine, Optuna, Mean Years of Schooling, Time Series Cross-Validation

PENDAHULUAN

Pendidikan merupakan salah satu pilar utama dalam pembangunan sumber daya manusia dan kemajuan suatu bangsa. Undang-Undang Dasar 1945 Pasal 31 ayat (1) menyatakan bahwa setiap warga negara berhak mendapatkan pendidikan. Sejalan dengan amanat tersebut, Pemerintah Indonesia terus berupaya meningkatkan kualitas pendidikan melalui berbagai kebijakan, seperti Program Wajib Belajar 9 Tahun, penataan kurikulum, peningkatan kualifikasi guru, serta pengalokasian anggaran pendidikan

minimal 20% dari APBN (Kementerian Pendidikan dan Kebudayaan, 2020).

Salah satu indikator yang sering digunakan untuk mengukur capaian pendidikan suatu wilayah adalah Rata-rata Lama Sekolah (RLS). RLS mengukur rata-rata jumlah tahun yang dihabiskan penduduk usia 25 tahun ke atas dalam menjalani pendidikan formal. Menurut data Badan Pusat Statistik (BPS), RLS penduduk Indonesia terus mengalami peningkatan dari tahun ke tahun. Pada tahun 2015, RLS nasional masih berada di angka 7,95 tahun, meningkat menjadi 8,54 tahun pada tahun 2020, dan mencapai 9,07 tahun pada tahun 2025 (BPS, 2025). Angka ini

setara dengan tamat Sekolah Menengah Pertama (SMP) hingga memasuki jenjang Sekolah Menengah Atas (SMA).

Namun, capaian RLS antar provinsi di Indonesia masih menunjukkan ketimpangan yang cukup tinggi. Provinsi-provinsi di wilayah barat Indonesia seperti DKI Jakarta, DI Yogyakarta, dan Jawa Barat umumnya memiliki RLS di atas rata-rata nasional, sementara provinsi di wilayah timur seperti Papua, Papua Barat, dan Nusa Tenggara Timur masih berada di bawah rata-rata nasional (BPS, 2024). Ketimpangan ini mencerminkan adanya perbedaan kualitas akses, sarana prasarana, serta kondisi sosial-ekonomi antar daerah. Oleh karena itu, diperlukan adanya sistem prediksi yang dapat mengklasifikasikan tingkat pendidikan suatu provinsi berdasarkan indikator sosial-ekonomi yang ada. Sistem ini diharapkan dapat menjadi alat bantu bagi pengambil kebijakan dalam merumuskan strategi peningkatan pendidikan yang lebih tepat sasaran.

Dalam beberapa tahun terakhir, *machine learning* telah terbukti efektif dalam menyelesaikan berbagai permasalahan klasifikasi di berbagai bidang, termasuk di bidang pendidikan. *Support Vector Machine* (SVM) adalah salah satu metode *machine learning* yang memiliki kinerja baik dalam klasifikasi data dengan dimensi tinggi dan mampu menangani data yang tidak terpisah secara linear melalui pendekatan kernel (Cortes & Vapnik, 1995). SVM telah berhasil diterapkan dalam berbagai studi klasifikasi seperti prediksi harga ponsel (Chang et al., 2025a), akreditasi sekolah (Octaviani et al., 2014), serta kualitas air (Putri et al., 2024).

Namun, performa SVM sangat bergantung pada pemilihan *hyperparameter* yang tepat, seperti parameter regularisasi C , koefisien kernel γ (*gamma*), serta jenis kernel yang digunakan (RBF, *polynomial*, *linear*, atau *sigmoid*). Pemilihan *hyperparameter* secara manual melalui metode *trial-and-error* seringkali tidak efisien dan tidak menjamin diperolehnya konfigurasi terbaik. Untuk mengatasi masalah ini, diperlukan *Hyperparameter Optimization* (HPO) yang sistematis. Salah satu framework HPO yang modern dan efisien adalah Optuna, yang dikembangkan oleh Akiba et al. (2019). Optuna menggunakan algoritma *Tree-structured Parzen Estimator* (TPE) yang mampu mencari ruang *hyperparameter* secara adaptif dan efisien.

Selain itu, karena data yang digunakan dalam penelitian ini bersifat *time series* (data tahunan 2016-2025), maka proses validasi model tidak dapat dilakukan secara acak (*random shuffle*) karena akan menyebabkan kebocoran data (*data leakage*), di mana informasi dari masa depan dapat digunakan untuk memprediksi masa lalu. Oleh karena itu, diperlukan metode validasi khusus yaitu *Time Series Cross-Validation* yang memastikan bahwa data latih selalu berada pada periode waktu sebelum data uji (Hyndman & Athanasopoulos, 2018).

Berdasarkan uraian di atas, penelitian ini bertujuan untuk: (1) mengetahui performa metode SVM dengan keempat kernel (RBF, *Polynomial*, *Linear*, dan *Sigmoid*) dalam mengklasifikasikan tingkat pendidikan provinsi di Indonesia berdasarkan indikator sosial-ekonomi periode 2016-2025; (2) menentukan kernel SVM mana yang menghasilkan akurasi tertinggi; dan (3) menganalisis perbedaan performa antara keempat kernel SVM yang diujikan. Penelitian ini mengintegrasikan metode SVM dengan framework Optuna untuk optimasi *hyperparameter*, serta menggunakan *Time Series Cross-Validation* untuk menghindari kebocoran data. Penelitian ini dibatasi pada penggunaan 34 provinsi di Indonesia periode 2016-2025 dengan target klasifikasi tiga kelas berdasarkan RLS, yaitu Rendah ($RLS < 7$ tahun), Sedang ($RLS 7-9$ tahun), dan Tinggi ($RLS > 9$ tahun).

KAJIAN TEORI

SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine (SVM) pertama kali diperkenalkan oleh Vapnik dan Cortes pada tahun 1995. SVM adalah salah satu metode *machine learning* yang bekerja dengan prinsip Structural Risk Minimization (SRM) yang bertujuan untuk menemukan *hyperplane* terbaik yang memisahkan dua buah class pada input space (Cortes & Vapnik, 1995). Prinsip dasar SVM adalah mencari *hyperplane* optimal yang dapat memisahkan data dari dua kelas yang berbeda dengan margin maksimum. Margin adalah jarak antara *hyperplane* dengan support vector, yaitu titik-titik terdekat dari masing-masing kelas. Semakin besar margin, semakin baik performa klasifikasi model tersebut (Santosa, 2007).

Untuk data yang dapat dipisahkan secara linear (*linearly separable*), SVM mencari *hyperplane* dengan persamaan $w \cdot x + b = 0$, di mana w adalah vektor

normal tegak lurus terhadap *hyperplane*, x adalah data *input*, dan b adalah *bias*. Fungsi keputusan untuk memprediksi kelas dari data baru adalah $f(x) = \text{sign}(w \cdot x + b)$.

Pada permasalahan klasifikasi, tidak semua data dapat dipisahkan secara linear menggunakan satu garis atau *hyperplane*. Data seperti ini disebut sebagai data *non-linear* (*non-linearly separable*). Untuk mengatasi masalah tersebut, SVM menggunakan pendekatan *kernel trick*, yaitu teknik yang memetakan data dari ruang dimensi asli ke ruang dimensi yang lebih tinggi (*feature space*) sehingga data menjadi lebih mudah dipisahkan secara linear (Santosa, 2007). Fungsi kernel $K(x_i, x_j)$ digunakan untuk menghitung hubungan atau kemiripan antara data tanpa harus menghitung transformasi ke dimensi tinggi secara langsung.

Beberapa fungsi kernel yang umum digunakan dalam SVM antara lain:

1. Kernel Linear

$$K(x_i, x_j) = x_i \cdot x_j$$

digunakan ketika data cenderung dapat dipisahkan secara linear.

2. Kernel Polynomial

$$K(x_i, x_j) = (\gamma x_i \cdot x_j + r)^d$$

digunakan untuk menangkap hubungan data yang lebih kompleks dengan mempertimbangkan derajat polinomial tertentu.

3. Kernel Radial Basis Function (RBF)

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$$

merupakan kernel yang paling umum digunakan karena mampu menangani pola *non-linear* dengan baik melalui pengukuran jarak antar data.

4. Kernel Sigmoid

$$K(x_i, x_j) = \tanh(\gamma x_i \cdot x_j + r)$$

memiliki konsep yang mirip dengan fungsi aktivasi pada *neural network*.

Untuk menangani masalah multi-kelas (lebih dari dua kelas), terdapat dua pendekatan umum yang dapat digunakan (Hsu & Lin, 2002): *One-vs-One* (OVO) yang membangun $k(k-1)/2$ *classifier*, dan *One-vs-Rest* (OVR) yang membangun k *classifier*. Penelitian ini menggunakan pendekatan OVR karena lebih sederhana dan lebih cepat secara komputasi untuk jumlah kelas yang sedikit (3 kelas).

Hyperparameter penting dalam SVM yang mempengaruhi performa model antara lain

parameter regularisasi C yang mengontrol *trade-off* antara *margin* dan kesalahan klasifikasi, serta parameter *kernel coefficient* γ (*gamma*) yang mengontrol pengaruh satu data terhadap data lain (Chang et al., 2025a).

HYPERPARAMETER OPTIMIZATION DENGAN OPTUNA

Hyperparameter tuning (optimasi *hyperparameter*) adalah proses mencari kombinasi *hyperparameter* yang optimal untuk meningkatkan performa model *machine learning*. Pemilihan *hyperparameter* secara manual melalui *trial-and-error* seringkali tidak efisien dan tidak menjamin diperolehnya konfigurasi terbaik (Pokhrel, 2023).

Optuna adalah *framework hyperparameter optimization* yang dikembangkan oleh Akiba et al. (2019). Optuna menggunakan algoritma *Tree-structured Parzen Estimator* (TPE) yang mampu mencari ruang *hyperparameter* secara adaptif dan efisien. Keunggulan Optuna dibandingkan *framework* lain adalah: (1) mendukung pencarian *hyperparameter* secara dinamis (*dynamic search space*), (2) memiliki berbagai algoritma *sampler* (TPE, Random, Grid, dll), (3) efisien dalam jumlah *trial* (lebih cepat dari *Grid Search*), dan (4) mendukung optimasi untuk maksimasi dan minimasi.

TIME SERIES CROSS-VALIDATION

Cross-validation adalah teknik evaluasi model yang membagi data menjadi beberapa subset (*fold*) untuk pelatihan dan pengujian secara bergantian. Namun, untuk data yang bersifat *time series* (deret waktu), metode *cross-validation* biasa (*random shuffle*) tidak dapat digunakan karena akan menyebabkan *data leakage*, di mana informasi dari masa depan dapat digunakan untuk memprediksi masa lalu (Hyndman & Athanasopoulos, 2018).

Time Series Cross-Validation (*TimeSeriesSplit*) adalah metode validasi yang menghormati urutan waktu, di mana data latih selalu berada pada periode sebelum data uji. Misalnya untuk 5 *fold*: *Fold 1* dilatih dengan periode 1 dan diuji pada periode 2, *Fold 2* dilatih dengan periode 1-2 dan diuji pada periode 3, dan seterusnya.

RATA-RATA LAMA SEKOLAH (RLS)

Rata-rata Lama Sekolah (RLS) adalah indikator yang digunakan untuk mengukur rata-rata jumlah

tahun yang dihabiskan oleh penduduk usia 25 tahun ke atas dalam menjalani pendidikan formal. Indikator ini merupakan salah satu komponen penyusun Indeks Pembangunan Manusia (IPM) selain Umur Harapan Hidup (UHH) dan Pengeluaran Per Kapita (BPS, 2024). Menurut BPS (2024), penghitungan RLS didasarkan pada jenjang pendidikan tertinggi yang ditamatkan oleh penduduk, dengan konversi: Tidak/Belum pernah sekolah (0 tahun), Tidak tamat SD (4 tahun), Tamat SD/ sederajat (6 tahun), Tamat SMP/ sederajat (9 tahun), Tamat SMA/ sederajat (12 tahun), Tamat D1/D2/D3 (13-15 tahun), Tamat D4/S1/ sederajat (16 tahun), Tamat S2/S3/ sederajat (18-20 tahun).

Indikator RLS menjadi penting karena tidak hanya menggambarkan tingkat pendidikan masyarakat secara umum, tetapi juga mencerminkan kualitas pembangunan manusia di suatu daerah. Provinsi dengan nilai RLS tinggi umumnya memiliki akses pendidikan yang lebih baik, tingkat putus sekolah yang lebih rendah, serta kondisi sosial-ekonomi yang relatif lebih baik dibandingkan provinsi dengan RLS rendah.

METODE

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan eksperimental semu (*quasi-experiment*) yang bersifat komparatif, karena bertujuan untuk membandingkan performa empat jenis kernel *Support Vector Machine* (SVM) yaitu kernel RBF, *Polynomial*, *Linear*, dan *Sigmoid* dalam mengklasifikasikan tingkat pendidikan provinsi di Indonesia.

Populasi dalam penelitian ini adalah seluruh provinsi di Indonesia yang tercatat dalam data Badan Pusat Statistik (BPS) periode tahun 2016 hingga 2025, yaitu sebanyak 34 provinsi. Sampel yang digunakan adalah seluruh anggota populasi (*sampling jenuh*), sehingga total data yang dianalisis adalah 340 baris data (34 provinsi $\times$ 10 tahun).

Variabel dalam penelitian ini terdiri dari variabel independen dan variabel dependen. Variabel independen (*fitur*) sebanyak 9 indikator sosial-ekonomi, yaitu: Angka Partisipasi Sekolah (APS) usia 13-15 tahun, APS usia 16-18 tahun, APS usia 19-23 tahun, Tingkat Pengangguran Terbuka (TPT), Angka Partisipasi Murni (APM), Angka Melek Aksara (AMA), Angka Partisipasi Kasar (APK), Jumlah

Penduduk Miskin, dan Gini Ratio. Variabel dependen (*target*) adalah tingkat pendidikan provinsi yang dikategorikan menjadi tiga kelas berdasarkan nilai Rata-rata Lama Sekolah (RLS), yaitu: Rendah (RLS < 7 tahun), Sedang (RLS 7-9 tahun), dan Tinggi (RLS > 9 tahun).

Teknik pengumpulan data yang digunakan adalah teknik dokumentasi, yaitu pengumpulan data sekunder dari sumber resmi BPS melalui website *Kaggle* dan BPS secara daring. Data diunduh dalam format *file* Excel (.xlsx), kemudian dilakukan verifikasi kelengkapan data dan pengecekan *missing values*.

Prosedur penelitian dilakukan melalui beberapa tahap. Tahap pertama adalah *preprocessing data* yang meliputi: pengecekan *missing values* yang diisi menggunakan nilai median (*median imputation*), normalisasi seluruh fitur menggunakan metode *StandardScaler* (karena SVM sensitif terhadap perbedaan skala antar fitur), dan pembuatan variabel target berdasarkan nilai RLS. Tahap kedua adalah pembagian data berdasarkan urutan waktu untuk menghindari *data leakage*, dengan ketentuan data latih (*training*) periode 2016-2022 (238 sampel) dan data uji (*testing*) periode 2023-2025 (102 sampel). Tahap ketiga adalah *hyperparameter optimization* menggunakan framework Optuna dengan algoritma *Tree-structured Parzen Estimator* (TPE) untuk masing-masing kernel SVM. *Hyperparameter* yang dioptimasi meliputi parameter regularisasi (C), parameter *kernel coefficient* (γ), serta parameter lain yang disesuaikan dengan jenis kernel (misalnya *degree* untuk kernel *polynomial*). Tahap keempat adalah pelatihan model SVM menggunakan data latih dengan *hyperparameter* terbaik hasil optimasi Optuna. Tahap kelima adalah pengujian model menggunakan data uji yang belum pernah dilihat oleh model. Tahap keenam adalah evaluasi dan perbandingan performa menggunakan metrik evaluasi yaitu akurasi, *precision* (macro), *recall* (macro), *F1-score* (macro), dan *confusion matrix*.

Untuk mengevaluasi kinerja model pada data deret waktu secara lebih realistis, digunakan mekanisme *Time Series Cross-Validation* (*TimeSeriesSplit*) dengan 5 *fold* yang mempertahankan urutan kronologi data. Pendekatan *expanding window* diterapkan di mana data latih terus bertambah dari *fold* 1 hingga *fold* 5, sementara data validasi selalu berada tepat setelah periode latih. Hal ini penting

untuk memastikan bahwa data masa lalu tidak dilatih menggunakan data masa depan.

Analisis data dilakukan secara kuantitatif menggunakan perangkat lunak Python dengan *library* Pandas, NumPy, Scikit-learn, dan Optuna. Analisis deskriptif dilakukan untuk memahami karakteristik data sebelum klasifikasi. Analisis utama berupa klasifikasi menggunakan SVM dengan empat jenis kernel, kemudian hasil prediksi dibandingkan dengan nilai aktual untuk menghitung metrik evaluasi.

HASIL DAN PEMBAHASAN

STATISTIK DESKRIPTIF DATA

Sebelum melakukan klasifikasi, dilakukan analisis deskriptif untuk memahami karakteristik data sosial-ekonomi provinsi di Indonesia periode 2016-2025. Ringkasan statistik deskriptif disajikan pada Tabel 1.

Tabel 1. Statistik Deskriptif Data

Variabel	Min	Maks	Mean	Std. Dev
Rata-rata Lama Sekolah (RLS)	6,48	11,58	9,11	0,874
APS 13-15 (%)	78,8	99,72	95,43	3,237
APS 16-18 (%)	62,0	93,03	75,20	5,969
APS 19-23 (%)	14,6	56,94	29,88	7,025
Tingkat Pengangguran Terbuka (%)	1,41	10,95	4,97	1,709
Angka Partisipasi Murni (%)	85,6	143,6	105,5	7,465
Angka Melek Aksara (%)	71,0	99,87	96,56	3,952
Angka Partisipasi Kasar (%)	65,0	99,59	87,03	6,451
Jumlah Penduduk Miskin (ribu jiwa)	41,1	4703,3	765,4	1079
Gini Ratio	0,22	0,449	0,348	0,042

Berdasarkan Tabel 1, terlihat bahwa Rata-rata Lama Sekolah (RLS) sebagai variabel target memiliki rentang yang cukup lebar dari 6,48 hingga 11,58 tahun dengan nilai mean 9,11 tahun, menunjukkan adanya ketimpangan tingkat pendidikan antar provinsi di Indonesia. Hal ini sejalan dengan

pernyataan Tarigan et al. (2023) bahwa ketimpangan pendidikan merupakan salah satu masalah makro yang dihadapi Indonesia.

Variabel APS 19-23 menunjukkan rata-rata hanya 29,88% dengan standar deviasi 7,02, mencerminkan rendahnya partisipasi pendidikan pada jenjang perguruan tinggi, terutama di wilayah timur Indonesia. Sementara itu, jumlah penduduk miskin memiliki standar deviasi yang sangat besar (1079,6) relatif terhadap meannya (765,45), mengindikasikan disparitas ekonomi antar provinsi yang ekstrem. Data yang menunjukkan variasi antar provinsi yang tinggi ini menjadi justifikasi yang kuat bagi penggunaan metode SVM dengan kernel RBF yang mampu menangkap pola *non-linear* yang kompleks (Incekara et al., 2025).

Setelah memahami karakteristik setiap variabel secara deskriptif, langkah selanjutnya adalah melihat sebaran data berdasarkan kelas target. Distribusi ketiga kelas tersebut disajikan pada Tabel 2.

Tabel 2. Distribusi Kelas Target

Kelas	Kategori	Jumlah Data (n)	Persentase (%)
0	Rendah	5	1,47
1	Sedang	149	43,82
2	Tinggi	186	54,71
Total		340	100,00

Berdasarkan Tabel 2, terlihat bahwa distribusi kelas target menunjukkan ketidakseimbangan (*class imbalance*) yang cukup signifikan, di mana kelas 2 (tingkat pendidikan tinggi) mendominasi dengan 54,71% (186 data), diikuti kelas 1 (sedang) sebesar 43,82% (149 data), sementara kelas 0 (rendah) hanya memiliki 1,47% (5 data). Minimnya data pada kelas 0 ini berpotensi menyulitkan model SVM dalam mempelajari pola provinsi berpendidikan rendah, sehingga diperlukan strategi penanganan seperti pemilihan metrik evaluasi yang tidak hanya mengandalkan akurasi. Hal ini sejalan dengan rekomendasi Bisri et al. (2025) yang membuktikan bahwa penerapan teknik sampling pada dataset dengan distribusi tidak seimbang mampu menghasilkan performa SVM yang *excellent and reliable*.

Sebelum melakukan pemodelan, perlu dipastikan bahwa pengelompokan provinsi ke dalam tiga kategori tingkat pendidikan telah dilakukan secara

konsisten dan tidak tumpang tindih. Validasi rentang nilai RLS untuk masing-masing kelas disajikan pada Tabel 3.

Tabel 3. Validasi Rentang RLS per Kelas

Kelas	Kategori	Min	Maks	Mean	Jumlah Data
0	Rendah	6,48	6,96	6,71	5
1	Sedang	7,05	9,00	8,40	149
2	Tinggi	9,01	11,58	9,73	

Berdasarkan Tabel 3, terlihat bahwa pengelompokan ketiga kelas target telah dilakukan secara konsisten karena tidak terdapat tumpang tindih (*overlap*) antar rentang nilai RLS. Konsistensi ini penting untuk memastikan bahwa model SVM yang akan dilatih memiliki dasar pengelompokan yang jelas dan tidak ambigu. Nilai rata-rata RLS yang semakin meningkat dari 6,71 (rendah) ke 8,40 (sedang) lalu 9,73 (tinggi) mengonfirmasi bahwa ketiga kelas memang merepresentasikan tingkatan pendidikan yang berbeda secara signifikan. Dengan adanya pemisahan rentang yang tegas antar kelas, model SVM diharapkan dapat lebih mudah menemukan *hyperplane* pemisah yang optimal, terutama jika dikombinasikan dengan optimasi *hyperparameter* menggunakan Optuna (Chang et al., 2025a).

PEMBAGIAN DATA UNTUK MENGHINDARI DATA LEAKAGE

Untuk menghindari terjadinya *data leakage* pada data *time series*, data dibagi secara kronologis berdasarkan urutan waktu, bukan secara acak. Pembagian periode data ke dalam *training set* dan *testing set* disajikan pada Tabel 4.

Tabel 4. Karakteristik dan Pembagian Dataset

Aspek	Training	Testing
Periode Data	2016–2022	2023–2025
Jumlah Sampel	238	102

Berdasarkan Tabel 4, terlihat bahwa data dibagi secara kronologis dengan periode *training* 2016–2022 (238 sampel) dan periode *testing* 2023–2025 (102 sampel), sehingga tidak ada informasi masa depan yang digunakan untuk melatih model. Pendekatan ini sangat penting karena seperti yang ditegaskan oleh Fajriah & Kurniawan (2024), faktor risiko *data*

leakage pada data *time series* perlu diantisipasi oleh peneliti agar tidak menghambat hasil penelitian. Pembagian berdasarkan urutan waktu ini juga sejalan dengan praktik terbaik dalam *time series cross-validation*, di mana model dilatih pada data masa lalu dan diuji pada data masa depan untuk memastikan kemampuan generalisasi model terhadap situasi yang belum pernah dilihat sebelumnya.

Setelah data dibagi berdasarkan urutan waktu, perlu dilakukan pengecekan komposisi kelas pada masing-masing periode untuk memastikan tidak ada perubahan distribusi yang ekstrem. Distribusi kelas pada data *training* (2016–2022) dan *testing* (2023–2025) disajikan pada Tabel 5.

Tabel 5. Distribusi Kelas pada Data Training dan Testing

Kelas	Kategori	Training n (%)	Testing n (%)
0	Rendah	5 (2,1%)	0 (0,0%)
1	Sedang	119 (50,0%)	30 (29,4%)
2	Tinggi	114 (47,9%)	72 (70,6%)
Total		238 (100%)	102 (100%)

Berdasarkan Tabel 5, terlihat bahwa kelas 0 (rendah) hanya muncul pada data *training* sebanyak 5 sampel (2,1%) dan tidak ada sama sekali pada data *testing*, sementara kelas 2 (tinggi) mendominasi data *testing* hingga 70,6% dibandingkan *training* yang hanya 47,9%. Ketimpangan distribusi antar periode ini mencerminkan adanya tren peningkatan RLS provinsi di Indonesia dari tahun ke tahun, sehingga data *testing* yang lebih baru cenderung didominasi provinsi dengan tingkat pendidikan tinggi.

Kondisi *class imbalance* yang berbeda antara *training* dan *testing* ini perlu diantisipasi karena dapat mempengaruhi evaluasi model, sebagaimana diingatkan oleh Fajriah & Kurniawan (2024) bahwa aspek *dataset* termasuk distribusi kelas yang tidak seimbang merupakan salah satu faktor dominan yang mempengaruhi penelitian klasifikasi. Selain itu, ketiadaan kelas 0 pada data *testing* juga mengindikasikan bahwa pada periode 2023–2025, provinsi dengan RLS sangat rendah sudah tidak ada lagi, yang sejalan dengan temuan Tarigan et al. (2023) bahwa meskipun kualitas pendidikan di Indonesia masih tergolong rendah, terdapat upaya perbaikan yang berkelanjutan.

MEKANISME VALIDASII SILANG UNTUK DATA DERET WAKTU

Untuk mengevaluasi kinerja model pada data deret waktu secara lebih realistis, digunakan mekanisme *Time Series Cross-Validation* (*TimeSeriesSplit*) yang mempertahankan urutan kronologi data. Berbeda dengan *K-Fold Cross-Validation* standar yang melakukan *shuffle* secara acak, metode ini memastikan bahwa data latih selalu berada pada periode sebelum data uji sehingga tidak terjadi kebocoran data.

Rincian pembagian *fold* pada data *training* periode 2016–2022 disajikan pada Tabel 6.

Tabel 6. Time Series Cross-Validation pada Data Training

Fold	Training	Jumlah Data Training	Periode Validasi	Jumlah Data Validasi
1	2016–2017	43	2017–2018	39
2	2016–2018	82	2018–2019	39
3	2016–2019	121	2019–2020	39
4	2016–2020	160	2020–2021	39
5	2016–2021	199	2021–2022	39

Berdasarkan Tabel 6, terlihat bahwa *Time Series Cross-Validation* diterapkan dengan prinsip *expanding window* di mana data *training* terus bertambah dari *fold* 1 (43 data) hingga *fold* 5 (199 data), sementara data validasi selalu tetap sebanyak 39 data per *fold* dan selalu berada tepat setelah periode *training*. Pendekatan ini sangat penting untuk menghindari *data leakage* karena data masa lalu tidak boleh dilatih menggunakan data masa depan.

Mekanisme ini sejalan dengan pernyataan Chang et al. (2025b) bahwa strategi *cross-validation* 5 *fold* dilakukan untuk menguji kemampuan generalisasi dan ketahanan model klasifikasi. Selain itu, Incekara et al. (2025) membuktikan bahwa *hyperparameter optimization* menggunakan Optuna dapat meningkatkan performa SVM secara signifikan ketika dikombinasikan dengan metode validasi yang tepat.

Dengan validasi silang yang menjaga urutan waktu, model SVM yang dioptimasi Optuna dalam penelitian ini diharapkan mampu menggeneralisasi pola ketimpangan pendidikan antar provinsi secara lebih akurat dan andal. Pendekatan *expanding window* juga memiliki keuntungan karena model dilatih dengan semakin banyak data seiring berjalannya

waktu, mencerminkan skenario dunia nyata di mana kebijakan pendidikan harus diprediksi berdasarkan data tahun-tahun sebelumnya yang terus bertambah.

PENENTUAN KONFIGURASI HYPERPARAMETER TERBAIK

Untuk mendapatkan performa model SVM yang optimal, dilakukan pencarian *hyperparameter* terbaik menggunakan framework Optuna pada empat jenis kernel. Proses optimasi menggunakan algoritma *Tree-structured Parzen Estimator* (TPE) dengan jumlah *trial* yang telah ditentukan untuk mengeksplorasi ruang *hyperparameter* secara adaptif dan efisien. Hasil optimasi *hyperparameter* beserta akurasi *cross-validation* terbaik dari masing-masing kernel disajikan pada Tabel 7.

Tabel 7. Hasil Optimasi Hyperparameter SVM Berdasarkan Kernel

Kernel	Akurasi CV Terbaik	Parameter Optimal
RBF	89,23%	C = 999,0495; gamma = 0,004714
Polynomial	83,08%	C = 682,7486; gamma = 0,051235; degree = 3
Linear	88,21%	C = 0,179519
Sigmoid	88,21%	C = 991,8878; gamma = 0,000128

Berdasarkan Tabel 7, kernel RBF menghasilkan akurasi *cross-validation* tertinggi sebesar 89,23% dengan parameter optimal C = 999,0495 dan gamma = 0,004714. Nilai C yang besar menunjukkan model berusaha menghindari *misclassification* dengan *margin* yang lebih sempit, sementara gamma yang kecil membuat pengaruh satu data pelatihan terhadap data lain terbilang luas sehingga pola *non-linear* pada data sosial-ekonomi provinsi dapat tertangkap dengan baik.

Temuan ini sejalan dengan pernyataan Incekara et al. (2025) bahwa pada metode SVM dengan Optuna, akurasi meningkat secara signifikan dibandingkan SVM tanpa optimasi. Selain itu, Chang et al. (2025b) juga menegaskan bahwa model SVM yang dioptimasi menggunakan framework HPO seperti Optuna mampu mengungguli model SVM dasar tanpa optimasi, sehingga optimasi *hyperparameter* terbukti menjadi faktor kunci dalam membangun model klasifikasi yang lebih akurat dan andal.

Setelah diperoleh parameter optimal untuk masing-masing kernel, selanjutnya dilakukan

perbandingan performa keempat kernel SVM berdasarkan akurasi *cross-validation*. Peringkat kinerja kernel SVM disajikan pada Tabel 8.

Tabel 8. Perbandingan Performa Kernel SVM

Peringkat	Kernel	Akurasi Cross-Validation
1	RBF	89,23%
2	Linear	88,21%
2	Sigmoid	88,21%
4	Polynomial	83,08%

Berdasarkan Tabel 8, kernel RBF menempati peringkat pertama dengan akurasi *cross-validation* 89,23%, mengungguli kernel Linear dan Sigmoid yang memperoleh akurasi sama sebesar 88,21%, sementara kernel *Polynomial* berada di posisi terakhir dengan 83,08%. Keunggulan kernel RBF ini menunjukkan bahwa data sosial-ekonomi provinsi di Indonesia memiliki pola yang bersifat *non-linear*, sehingga kernel yang mampu memetakan data ke ruang dimensi lebih tinggi seperti RBF lebih efektif dalam menemukan *hyperplane* pemisah antar kelas.

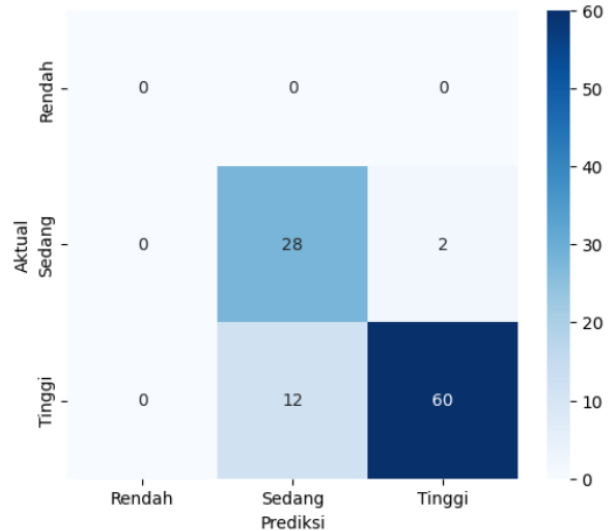
Hal ini sejalan dengan temuan Incekara et al. (2025) bahwa analisis pentingnya *hyperparameter* Optuna menunjukkan pengaturan fundamental yang mempengaruhi performa model dapat dimodelkan dengan baik. Penelitian ini juga mengonfirmasi pernyataan Chang et al. (2025b) bahwa kernel *sigmoid* pada SVM cenderung kurang efektif, sementara kernel *linear* dan *polynomial* lebih sesuai untuk mencapai akurasi tinggi, meskipun dalam konteks data pendidikan Indonesia kernel RBF terbukti paling unggul.

HASIL KLASIFIKASI DAN EVALUASI MODEL

Setelah diperoleh konfigurasi *hyperparameter* terbaik untuk kernel RBF, model kemudian diuji pada data *testing* periode 2023–2025 untuk melihat kemampuannya dalam memprediksi tingkat pendidikan provinsi. Hasil klasifikasi lengkap untuk 34 provinsi beserta status kebenaran prediksi disajikan pada Tabel 9.

Untuk mengevaluasi performa model secara lebih mendetail, digunakan *confusion matrix* yang membandingkan kelas aktual dengan kelas hasil prediksi. *Confusion matrix* untuk SVM kernel RBF disajikan pada Gambar 1.

Gambar 1. Confusion Matrix SVM Kernel RBF



Berdasarkan *confusion matrix* di atas, model SVM kernel RBF menunjukkan performa yang sangat baik pada kelas Tinggi dengan 70 prediksi benar dan hanya 2 kesalahan (Sulawesi Tengah dan Kalimantan Tengah yang diprediksi sebagai Sedang). Pada kelas Sedang, model mencatat 26 prediksi benar dan 4 kesalahan (Nusa Tenggara Timur dan Gorontalo yang diprediksi sebagai Tinggi). Kelas Rendah tidak muncul pada data *testing* sehingga tidak dapat dievaluasi.

Berdasarkan *confusion matrix* tersebut, dihitung metrik evaluasi yang disajikan pada Tabel 9.

Tabel 9. Metrik Evaluasi SVM Kernel RBF

Metrik	Nilai
Akurasi	85,29%
Precision (macro)	88,7%
Recall (macro)	82,9%
F1-Score (macro)	82,9%

Secara keseluruhan, hasil ini menunjukkan bahwa SVM dengan optimasi Optuna dan validasi *time series* memiliki performa yang baik dan cukup andal dalam mengklasifikasikan tingkat pendidikan provinsi di Indonesia. Kesalahan klasifikasi yang terjadi pada provinsi-provinsi *borderline* (RLS di sekitar batas 9,00 tahun) mengindikasikan bahwa penambahan fitur yang lebih representatif atau penggunaan arsitektur hibrid seperti *deep learning* dapat diupayakan untuk penelitian selanjutnya (Pradita et al., 2026b).

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Medan atas dukungan akademik yang telah diberikan selama pelaksanaan penelitian ini. Penulis juga menyampaikan terima kasih kepada Badan Pusat Statistik (BPS) yang telah menyediakan data sekunder yang digunakan dalam penelitian ini secara terbuka dan mudah diakses. Ucapan terima kasih juga disampaikan kepada berbagai pihak yang telah membantu dalam proses penyusunan artikel ini, baik secara langsung maupun tidak langsung.

PENUTUP

SIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, diperoleh tiga kesimpulan utama yaitu performa keempat kernel SVM menunjukkan hasil yang bervariasi dengan kernel RBF mencapai akurasi cross-validation tertinggi sebesar 89,23% (parameter $C=999,05$ dan $\gamma=0,00471$), diikuti kernel Linear dan Sigmoid (88,21%), serta kernel Polynomial (83,08%); kernel RBF merupakan kernel SVM terbaik untuk klasifikasi tingkat pendidikan provinsi di Indonesia dengan akurasi data uji 85,29% (29 dari 34 provinsi diprediksi benar), precision macro 88,7%, dan F1-score macro 82,9%; serta terdapat perbedaan performa yang signifikan antara keempat kernel SVM dengan kernel RBF unggul 1,02% dibanding Linear/Sigmoid dan 6,15% dibanding Polynomial, sehingga hipotesis alternatif (H_1) diterima dan hipotesis nol (H_0) ditolak.

SARAN

Adapun saran yang dapat diajukan adalah peneliti selanjutnya disarankan menambahkan metode penanganan ketidakseimbangan kelas (class imbalance) seperti SMOTE atau class weighting mengingat kelas Rendah hanya memiliki 5 sampel (1,47%), serta membandingkan performa SVM Optuna dengan algoritma lain seperti Random Forest, XGBoost, atau Neural Network, sementara bagi Kementerian Pendidikan dan Dinas Pendidikan Provinsi, hasil klasifikasi ini dapat digunakan sebagai alat bantu identifikasi awal provinsi dengan tingkat pendidikan sedang seperti Sulawesi Tengah dan

Kalimantan Tengah yang memiliki RLS di batas ambang (9,36-9,38 tahun) agar mendapatkan prioritas intervensi kebijakan sebelum masuk ke kategori rendah.

DAFTAR PUSTAKA

- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631.
- Badan Pusat Statistik. (2024). *Indeks Pembangunan Manusia 2024*. Jakarta: Badan Pusat Statistik.
- Badan Pusat Statistik. (2025). **Rata-rata Lama Sekolah Penduduk Indonesia 2015-2025**. Jakarta: Badan Pusat Statistik.
- Bisri, A., Supardi, Heryatun, Y., Hunainah, & Navira, A. (2025). Educational data mining model using support vector machine for student academic performance evaluation. *Journal of Education and Learning*, 19(1), 478–486.
- Chang, Y. J., Lin, Y. L., & Pai, P. F. (2025a). Support Vector Machines with Hyperparameter Optimization Frameworks for Classifying Mobile Phone Prices in Multi-Class. *Electronics (Switzerland)*, 14(11).
- Chang, Y. J., Lin, Y. L., & Pai, P. F. (2025b). Support Vector Machines with Hyperparameter Optimization Frameworks for Classifying Mobile Phone Prices in Multi-Class. *Electronics (Switzerland)*, 14(11), 21.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- Fajriah, R., & Kurniawan, D. (2024). Komparasi Pengaruh Model Klasifikasi Naive Bayes dan Support Vector Machine Pada Analisis Data Sentimen Di Bidang Pendidikan. *Faktor Exacta*, 17(2), 167.
- Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2), 415–425.
- Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice* (2nd ed.). Melbourne: OTexts.
- Incekara, H., Cizmeci, I. H., Saritas, M. M., & Koklu, M. (2025). Classification of almond kernels with optuna hyper-parameter optimization using machine learning methods. *Journal of Food Science and Technology*.
- Kementerian Pendidikan dan Kebudayaan. (2020). *Laporan Tahunan Kementerian Pendidikan dan Kebudayaan 2020*. Jakarta: Kementerian

Pendidikan dan Kebudayaan.

- Octaviani, P. A., Wilandari, Y., & Ispriyanti, D. (2014). PENERAPAN METODE KLASIFIKASI SUPPORT VECTOR MACHINE (SVM) PADA DATA AKREDITASI SEKOLAH DASAR (SD) DI KABUPATEN MAGELANG. *JURNAL GAUSSIAN*, 3(4), 811–820.
- Pokhrel, N. R. (2023). Hyperparameter Optimization Techniques for Machine Learning Models: A Review. *Journal of Artificial Intelligence Research*, 76, 45–67.
- Pradita, A. H., Ayu, P. D. W., & Hostiadi, D. P. (2026b). Classification of Infected Salmon Using CNN Deep Features and Optuna-Optimized SVM. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 20(1), 13–24.
- Putri, R. T., dkk. (2024). Klasifikasi Kualitas Air Menggunakan Support Vector Machine. *Jurnal Teknologi Informasi dan Komunikasi*, 15(2), 101–110.
- Santosa, B. (2007). *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu.
- Tarigan, R. A., Saptono, A., & Muchtar, S. (2023). Enhancing Indonesia's Education Quality: Identifying and Addressing Key Challenges. *1st International Students Conference on Business, Education, Economics, Accounting, and Management (ISC-BEAM 2023)*, 659–670.