# ANALISIS KOMPARATIF METODE *FEATURE SELECTION FILTER, WRAPPER, EMBEDDED* PADA MODEL *MACHINE LEARNING* UNTUK KLASIFIKASI STRES PELAJAR

**SKRIPSI**

Program Studi Sarjana Informatika
Jurusan Informatika


Oleh:
**MUHAMMAD HAIKAL ASHARI**
NIM D1041221067






**FAKULTAS TEKNIK**
**UNIVERSITAS TANJUNGPURA**
**PONTIANAK**
**2026**

## HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini:

Nama : Muhammad Haikal Ashari

NIM : D1041221067

menyatakan bahwa dalam SKRIPSI yang berjudul "Analisis Komparatif Metode *Feature Selection Filter, Wrapper, Embedded* pada Model *Machine Learning* untuk Klasifikasi Stres Pelajar" tidak terdapat karya yang pernah diajukan untuk memperoleh gelar sarjana di suatu perguruan tinggi manapun. Sepanjang pengetahuan Saya, tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam Daftar Rujukan.

Demikian pernyataan ini dibuat dengan sebenar-benarnya. Saya sanggup menerima konsekuensi akademis dan hukum di kemudian hari apabila pernyataan yang dibuat ini tidak benar.

Pontianak,15 Juni 2026



Muhammad Haikal Ashari
NIM D1041221067

KEMENTERIAN PENDIDIKAN TINGGI, SAINS, DAN TEKNOLOGI
**UNIVERSITAS TANJUNGPURA**
**FAKULTAS TEKNIK**
Jalan Prof. Dr. H. Hadari Nawawi Pontianak 78124
Telepon (0561) 740186, WA: +6282152280907
Email : ft@untan.ac.id Website : http://teknik.untan.ac.id

## HALAMAN PENGESAHAN

# ANALISIS KOMPARATIF METODE *FEATURE SELECTION FILTER, WRAPPER, EMBEDDED* PADA MODEL *MACHINE LEARNING* UNTUK KLASIFIKASI STRES PELAJAR

Jurusan Informatika
Program Studi Sarjana Informatika

Oleh:

Muhammad Haikal Ashari
NIM. D1041221067

Telah dipertahankan di depan Penguji Skripsi pada tanggal 15 Juni 2026
dan diterima sebagai salah satu persyaratan untuk memperoleh gelar sarjana

Susunan Penguji Skripsi

| | |
|---|---|
| Dosen Pembimbing Utama | : Rina Septiriana, S.T., M.Cs.<br>NIP. 198709232020122001 |
| Dosen Pembimbing Kedua | : Niken Candraningrum, S.T., M.Cs.<br>NIP. 199309272022032010 |
| Dosen Penguji Utama | : Prof. Dr. Herry Sujaini, S.T., M.T.<br>NIP. 196806291997021001 |
| Dosen Penguji Kedua | : Tursina, S.T., M.Cs.<br>NIP. 197801152002122003 |

Pontianak, 15 Juni 2026

Dekan

Dr.-Ing. Ir. Slamet Widodo, M.T., IPM
NIP. 196712231992031002

Pembimbing Utama

Rina Septiriana, S.T., M.Cs.
NIP. 198709232020122001

## HALAMAN PERSEMBAHAN

Skripsi ini didedikasikan kepada kedua orang tuaku Bapak Khairul Barry dan Ibu Herlina dan saudara-saudariku. Apa yang mereka telah berikan melebihi daripada apa yang pernah aku inginkan.

Dari titisan ilmu-Mu kupersembahkan goresan tanganku bagi ilmu pengetahuan Indonesia

# PRAKATA

Puji syukur kehadirat Allah SWT atas kehendak-Nya maka penelitian dan penulisan skripsi dengan judul "Analisis Komparatif Metode *Feature Selection Filter*, *Wrapper*, *Embedded* pada Model *Machine Learning* untuk Klasifikasi Stress Pelajar" ini dapat diselesaikan dengan baik.

Penelitian ini menjelaskan tentang perbandingan empiris dari tiga pendekatan seleksi fitur yaitu *Filter, Wrapper,* dan *Embedded*, tiga pendekatan tersebut diwakili oleh empat metode seleksi fitur, yaitu ANOVA, Mutual Information, RFECV, dan LASSO, dalam meningkatkan performa model klasifikasi tingkat stres pelajar. Dengan memanfaatkan dataset *Student Stress Factors* yang terdiri dari 1.100 sampel dan 20 fitur prediktor, penelitian ini mengevaluasi pengaruh masing-masing metode seleksi fitur terhadap akurasi, generalization gap, serta kestabilan model *Random Forest* dan *Support Vector Machine* (SVM) pada dua skenario pengujian, yakni parameter rujukan dan *Bayesian Optimization*.

Secara khusus penulis menyampaikan terima kasih yang sebesar-besarnya kepada Rina Septiriana, S.T., M.CS. sebagai pembimbing utama dan Niken Candraningrum, S.T., M.CS. pembimbing pendamping yang telah memberikan bimbingan, masukan, dan diskusi yang intensif khususnya tentang metode seleksi fitur dan model machine learning yang digunakan. Terima kasih disampaikan pula kepada seluruh dosen dan staf Program Studi Informatika Fakultas Teknik Universitas Tanjungpura yang telah memberikan dukungan selama proses penelitian, serta kepada keluarga dan rekan-rekan mahasiswa yang senantiasa memberikan semangat dan bantuan hingga skripsi ini dapat terselesaikan.

Penulis berharap penelitian ini dapat bermanfaat sebagai referensi dalam literatur pengembangan sistem deteksi dini stres pelajar berbasis *machine learning*, khususnya dalam pemilihan metode seleksi fitur yang tepat guna meningkatkan kualitas dan generalisasi model. Penulis mengharapkan masukan yang konstruktif agar dapat menyempurnakan penelitian ini.

Pontianak, 12 Maret 2026

Muhammad Haikal Ashari

# ABSTRAK

Tingginya prevalensi stres pada pelajar di Indonesia yang mencapai 36,7–71,6% menjadi permasalahan kesehatan mental yang serius dan memerlukan penanganan terstruktur, salah satunya melalui deteksi dini berbasis machine learning. Namun, penerapan *machine learning* pada data survei stres pelajar yang berdimensi tinggi seringkali menghadapi tantangan berupa redundansi fitur, multikolinearitas, dan noise akibat jawaban subjektif, sehingga dibutuhkan teknik seleksi fitur untuk menyaring fitur yang paling relevan dan informatif guna memperoleh model yang efektif. Penelitian ini bertujuan untuk membandingkan pengaruh empat metode seleksi fitur dari tiga pendekatan berbeda, yaitu ANOVA dan *Mutual Information* (Filter), RFECV (Wrapper), serta LASSO (Embedded), terhadap performa model machine learning *Random Forest* dan *Support Vector Machine* (SVM) dalam mengklasifikasikan tingkat stres pelajar. Dataset yang digunakan adalah *Student Stress Factors* yang terdiri dari 1.100 sampel dengan 20 fitur prediktor, dan pengujian dilakukan pada dua skenario konfigurasi parameter, yaitu parameter rujukan dan *Bayesian Optimization*. Evaluasi dilakukan berdasarkan akurasi, presisi, recall, F1-score, serta *generalization gap* untuk mengukur kemampuan generalisasi model. Hasil penelitian menunjukkan bahwa setiap metode seleksi fitur menghasilkan jumlah fitur yang berbeda, di mana RFECV menyisakan paling sedikit fitur (3–4 fitur), *Mutual Information* mempertahankan 11 fitur, ANOVA 12 fitur, dan LASSO 18 fitur; seluruh metode secara konsisten memilih fitur *blood_pressure* sebagai fitur yang relevan. Hasil pengujian menunjukkan bahwa metode seleksi fitur yang mampu mempertahankan informasi antar fitur seperti *Mutual Information*, yang diiringi dengan pemilihan algoritma model dan parameter yang sesuai, cenderung memberikan performa yang lebih stabil pada dataset yang digunakan. Hal ini dipengaruhi oleh karakteristik data survei yang didominasi fitur ordinal dan memiliki hubungan linear maupun non-linear antar variabel, sehingga reduksi fitur yang terlalu agresif dan subset fitur yang kurang relevan pada beberapa skenario menyebabkan penurunan stabilitas performa model. Penelitian ini menyimpulkan bahwa kombinasi terbaik adalah model SVM dengan parameter rujukan dan seleksi fitur *Mutual Information* (11 fitur), karena menghasilkan keseimbangan yang baik antara akurasi tertinggi (88,00%) dan *generalization gap* yang rendah (3,88%).

Kata kunci: stres pelajar, seleksi fitur, *machine learning*, klasifikasi, *filter, wrapper, embedded*

# ABSTRACT

The high prevalence of stress among students in Indonesia, ranging from 36.7% to 71.6%, has become a serious mental health issue that requires structured handling, one of which is through early detection using machine learning. However, the implementation of machine learning on high-dimensional student stress survey data often faces challenges such as feature redundancy, multicollinearity, and noise caused by subjective responses, thereby requiring feature selection techniques to identify the most relevant and informative features in order to obtain effective classification models. This study aims to compare the influence of four feature selection methods from three different approaches, namely ANOVA and Mutual Information (Filter), RFECV (Wrapper), and LASSO (Embedded), on the performance of Random Forest and Support Vector Machine (SVM) machine learning models in classifying student stress levels. The dataset used was the Student Stress Factors dataset consisting of 1,100 samples with 20 predictor features, and the experiments were conducted using two parameter configuration scenarios, namely reference parameters and Bayesian Optimization. Model evaluation was performed using accuracy, precision, recall, F1-score, and generalization gap to measure the generalization capability of the models. The results showed that each feature selection method produced a different number of selected features, where RFECV retained the fewest features (3–4 features), Mutual Information retained 11 features, ANOVA retained 12 features, and LASSO retained 18 features. All methods consistently selected the blood_pressure feature as a relevant feature. The experimental results also showed that feature selection methods capable of preserving information among features, such as Mutual Information, when accompanied by the selection of appropriate algorithms and parameter configurations, tended to provide more stable performance on the dataset used. This was influenced by the characteristics of the survey data, which were dominated by ordinal features and contained both linear and non-linear relationships among variables, causing overly aggressive feature reduction and less relevant feature subsets in several scenarios to reduce the stability of model performance. This study concludes that the best combination was achieved by the SVM model with reference parameters and Mutual Information feature selection (11 features), as it produced a good balance between the highest accuracy of 88.00% and a low generalization gap of 3.88%.



Keyword: student stress, feature selection, machine learning, classification, filter, wrapper, embedded

# DAFTAR ISI

HALAMAN PERNYATAAN ..... i
HALAMAN PENGESAHAN ..... ii
HALAMAN PERSEMBAHAN ..... iii
PRAKATA ..... iv
ABSTRAK ..... v
ABSTRACT ..... vi
DAFTAR ISI ..... vii
DAFTAR GAMBAR ..... x
DAFTAR TABEL ..... x
DAFTAR KODE PROGRAM ..... xiii
GLOSARIUM ..... xv
BAB I PENDAHULUAN ..... 1
1.1 Latar Belakang ..... 1
1.2 Rumusan Masalah ..... 5
1.3 Tujuan Penelitian ..... 5
1.4 Pembatasan Masalah ..... 5
1.5 Sistematika Penulisan ..... 6
BAB II TINJAUAN PUSTAKA ..... 8
2.1 Penelitian Terdahulu ..... 8
2.2 Landasan Teori ..... 13
2.2.1 Stres Pelajar ..... 13
2.2.2 Machine Learning ..... 13
2.2.3 Exploratory Data Analysis (EDA) ..... 14
2.2.4 Standard Scaler ..... 14
2.2.5 Seleksi Fitur ..... 15
2.2.6 Metode Filter ..... 16
2.2.7 Metode Wrapper ..... 16
2.2.8 Metode Embedded ..... 16
2.2.9 ANOVA ..... 17
2.2.10 Mutual Information ..... 18

2.2.11 RFECV ..... 18
2.2.12 LASSO ..... 19
2.2.13 Random Forest ..... 20
2.2.14 SVM ..... 20
2.2.15 Hyperparameter Tuning ..... 20
2.2.16 Bayesian Optimization ..... 21
2.2.17 Matriks Evaluasi ..... 21
2.2.18 Generalization Gap ..... 22
BAB III METODOLOGI PENELITIAN ..... 23
3.1 Data Penelitian ..... 23
3.2 Perangkat Penelitian ..... 23
3.3 Metodologi Penelitian ..... 23
3.3.1 Pengumpulan Data ..... 25
3.3.2 Exploratory Data Analysis (EDA) ..... 27
3.3.3 Pra-pemrosesan Data ..... 27
3.3.4 Seleksi Fitur ..... 28
3.3.5 Implementasi Model ..... 33
3.3.6 Evaluasi ..... 35
3.3.7 Analisis Hasil ..... 39
BAB IV HASIL DAN PEMBAHASAN ..... 41
4.1 Exploratory Data Analysis ..... 41
4.2 Pra-pemrosesan ..... 49
4.3 Seleksi Fitur ..... 51
4.3.1 ANOVA ..... 52
4.3.2 Mutual Information ..... 54
4.3.3 RFECV ..... 56
4.3.4 LASSO ..... 59
4.4 Implementasi Model ..... 60
4.4.1 Random Forest ..... 61
4.4.2 SVM ..... 62
4.5 Evaluasi ..... 63
4.6 Analisis Hasil ..... 68
BAB V PENUTUP ..... 111

5.1 Kesimpulan ..... 111
5.2 Saran ..... 111
DAFTAR RUJUKAN ..... 112

## DAFTAR GAMBAR

Gambar 2. 1 Tiga Kategori Seleksi Fitur ..... 16
Gambar 3. 1 Metodologi Penelitian ..... 24
Gambar 3. 2 Pra-pemrosesan ..... 28
Gambar 3. 3 Proses ANOVA ..... 29
Gambar 3. 4 Proses *Mutual Information* ..... 30
Gambar 3. 5 Proses RFECV ..... 31
Gambar 3. 6 Proses LASSO ..... 32
Gambar 4. 1 Karakteristik Data ..... 41
Gambar 4. 2 Hasil Pengecekan Duplikasi ..... 42
Gambar 4. 3 Hasil Pengecekan Distribusi Kelas Target ..... 42
Gambar 4. 4 Hasil Boxplot 1 ..... 43
Gambar 4. 5 Hasil Boxplot 2 ..... 43
Gambar 4. 6 Hasil Boxplot 3 ..... 44
Gambar 4. 7 Hasil Boxplot 4 ..... 44
Gambar 4. 8 Hasil Distribusi Fitur terhadap Fitur Target ..... 46
Gambar 4. 9 Hasil Heatmap ..... 48
Gambar 4. 10 Hasil *Splitting* Data *Train* dan Data *Test* ..... 49
Gambar 4. 11 Hasil Normalisasi *Standard Scaler* 1 ..... 50
Gambar 4. 12 Hasil Normalisasi *Standard Scaler* 2 ..... 51
Gambar 4. 13 Hasil Normalisasi *Standard Scaler* 3 ..... 51
Gambar 4. 14 Hasil Normalisasi *Standard Scaler* 4 ..... 51
Gambar 4. 15 Hasil rata-rata *F-score* ..... 52
Gambar 4. 16 Hasil Total Fitur Terpilih ANOVA ..... 53
Gambar 4. 17 Hasil Seleksi Fitur ANOVA ..... 53
Gambar 4. 18 Hasil rata-rata *Mutual information* ..... 54
Gambar 4. 19 Hasil Total Fitur Tepilih *Mutual Information* ..... 55
Gambar 4. 20 Hasil Seleksi Fitur *Mutual Information* ..... 55
Gambar 4. 21 Hasil Total Fitur Terpilih RFECV-RF ..... 56
Gambar 4. 22 Hasil Seleksi Fitur RFECV-RF ..... 57
Gambar 4. 23 Hasil Total Fitur Terpilih RFECV-SVM ..... 58
Gambar 4. 24 Hasil Seleksi Fitur RFECV-SVM ..... 58
Gambar 4. 25 Hasil Total Fitur Terpilih LASSO ..... 60
Gambar 4. 26 Hasil Seleksi Fitur LASSO ..... 60
Gambar 4.27 Perbandingan Jumlah Fitur Pada Seleksi Fitur ..... 68
Gambar 4.28 Rangkuman Fitur Terpilih ..... 73
Gambar 4.29 Perbandingan Evaluasi Model *Random Forest* ..... 76
Gambar 4. 30 Hasil *Confusion Matrix* Model *Random Forest Bayesian Optimization* ..... 81
Gambar 4. 31 Hasil *Confusion Matrix* Model *Random Forest* Parameter Statis ..... 85
Gambar 4. 32 Perbandingan Evaluasi Model SVM ..... 88
Gambar 4. 33 Hasil *Confusion Matrix* Model SVM *Bayesian Optimization* ..... 93
Gambar 4. 34 Hasil *Confusion Matrix* Model SVM Parameter Statis ..... 96

Gambar 4.35 Perbandingan Akurasi Data *Train-Test* pada *Random Forest Bayesian Optimization* ..... 99
Gambar 4.36 Perbandingan Akurasi Data *Train-Test* pada *Random Forest* Parameter Statis ..... 101
Gambar 4.37 Perbandingan Akurasi Data *Train-Test* pada SVM *Bayesian Optimization* ..... 104
Gambar 4. 38 Perbandingan Akurasi Data *Train-Test* pada SVM Parameter Statis 106
Gambar 4. 39 Perbandingan Akurasi dengan Penelitian Amalia ..... 110

# DAFTAR TABEL

Tabel 2. 1 Penelitian Terdahulu ..... 9
Tabel 3. 1 Penjelasan Data ..... 25
Tabel 3. 2 *Hyperparameter Random Forest* yang disetel ..... 33
Tabel 3. 3 *Hyperparameter* SVM yang disetel ..... 35
Tabel 3. 4 Hasil yang diharapkan untuk evaluasi seluruh seleksi fitur. ..... 36
Tabel 3. 5 Penjelasan Skenario ..... 37
Tabel 3. 6 Hasil yang diharapkan untuk evaluasi model ..... 38
Tabel 3. 7 Hasil yang diharapkan untuk *generalization gap* model ..... 38
Tabel 4. 1 Hasil *Hyperparameter Random Forest* ..... 62
Tabel 4. 2 Hasil *Hyperparameter* SVM ..... 63
Tabel 4. 3 Hasil Seleksi Fitur Semua Metode. ..... 64
Tabel 4. 4 Hasil Evaluasi Model ..... 66
Tabel 4. 5 Hasil Evaluasi *Generalization Gap* ..... 67
Tabel 4. 6 Selisih *Bayesian Optimazitaion* dan Parameter statis *Random Forest* ..... 79
Tabel 4. 7 Selisih *Bayesian Optimization* dan Parameter statis SVM ..... 91
Tabel 4. 8 Hasil *Generalization Gap Random Forest Bayesian Optimization* ..... 100
Tabel 4. 9 Hasil *Generalization Gap Random Forest* Parameter Statis ..... 102
Tabel 4. 10 Hasil *Generalization Gap* SVM *Bayesian Optimization* ..... 105
Tabel 4. 11 Hasil *Generalization Gap* SVM Parameter Statis ..... 107

## DAFTAR KODE PROGRAM

Kode Program 4. 1 Pengecekan Karakteristik pada Data ..... 41
Kode Program 4. 2 Pengecekan Duplikasi pada Data..... 42
Kode Program 4. 3 Split Data *Train* dan *Test*..... 49
Kode Program 4. 4 Normalisasi pada data dengan *Standard Scaler* ..... 50
Kode Program 4. 5 Implementasi ANOVA ..... 52
Kode Program 4. 6 Implementasi *Mutual Information*..... 54
Kode Program 4. 7 Implementasi RFECV-RF ..... 56
Kode Program 4. 8 Implementasi RFECV-SVM..... 57
Kode Program 4. 9 Implementasi LASSO..... 59
Kode Program 4. 10 Parameter *Bayesian Optimization Random Forest*..... 61
Kode Program 4. 11 Parameter *Bayesian Optimization* SVM..... 62
Kode Program 4. 12 Evaluasi Model ..... 65

# DAFTAR LAMPIRAN

**LAMPIRAN A**.....115

# GLOSARIUM

| | |
|---|---|
| ANOVA | Metode statistik berbasis filter yang mengukur perbedaan rata-rata antar kelompok kelas untuk menilai relevansi setiap fitur terhadap label target dalam proses seleksi fitur. |
| Akurasi | Proporsi prediksi yang benar terhadap keseluruhan data uji, digunakan sebagai salah satu metrik utama dalam evaluasi performa model klasifikasi. |
| Bayesian Optimization | Teknik optimasi hyperparameter yang memanfaatkan model probabilistik untuk memandu pencarian konfigurasi parameter terbaik secara efisien. |
| Dataset | Kumpulan data terstruktur yang digunakan sebagai bahan penelitian. Dalam penelitian ini digunakan dataset publik Student Stress Factors yang terdiri dari 1.100 sampel dengan 20 fitur prediktor dan 1 label target. |
| Dimensionalitas | Jumlah fitur atau atribut yang terdapat dalam sebuah dataset. Dimensionalitas yang tinggi dapat menyebabkan redundansi dan menurunkan performa model machine learning. |
| Embedded | Kategori metode seleksi fitur yang mengintegrasikan proses pemilihan fitur ke dalam proses pelatihan model. Dalam penelitian ini diwakili oleh metode LASSO. |
| Evaluasi Model | Proses pengukuran performa model klasifikasi menggunakan metrik seperti akurasi, presisi, recall, dan F1-score pada data uji. |
| F1-Score | Metrik evaluasi yang merupakan rata-rata harmonik dari presisi dan recall, digunakan untuk mengukur keseimbangan antara keduanya terutama pada kondisi distribusi kelas yang tidak seimbang. |
| Filter | Kategori metode seleksi fitur yang menilai relevansi fitur berdasarkan |

| | |
|---|---|
| | ukuran statistik secara independen dari algoritma machine learning. Dalam penelitian ini diwakili oleh ANOVA dan Mutual Information. |
| Fitur | Variabel atau atribut input yang digunakan sebagai prediktor dalam model machine learning. Dataset dalam penelitian ini memiliki 20 fitur prediktor. |
| Generalization Gap | Selisih antara akurasi pada data latih dan akurasi pada data uji. Generalization gap yang besar mengindikasikan risiko overfitting pada model. |
| Hyperparameter | Parameter konfigurasi model machine learning yang nilainya ditentukan sebelum proses pelatihan dan tidak dipelajari dari data, seperti jumlah pohon pada Random Forest atau nilai C dan gamma pada SVM. |
| Hyperparameter Tuning | Proses pencarian kombinasi hyperparameter terbaik untuk memaksimalkan performa model machine learning. |
| Klasifikasi | Tugas machine learning yang bertujuan memprediksi kategori atau kelas dari suatu data berdasarkan fitur-fitur inputnya. Dalam penelitian ini, klasifikasi dilakukan terhadap tingkat stres pelajar. |
| Label Target | Variabel output yang ingin diprediksi oleh model. Dalam penelitian ini berupa tingkat stres pelajar yang direpresentasikan sebagai kelas kategorikal. |
| LASSO *(Least Absolute Shrinkage and Selection Operator)* | Metode seleksi fitur berbasis embedded yang menerapkan regularisasi L1 sehingga koefisien fitur yang tidak relevan diciutkan hingga bernilai nol selama proses pelatihan model. |
| Machine Learning | Cabang kecerdasan buatan yang memungkinkan sistem komputer belajar dari data dan membuat keputusan atau prediksi tanpa diprogram secara eksplisit untuk setiap tugas. |
| Mutual Information | Metode seleksi fitur berbasis filter yang mengukur ketergantungan statistik non- |

| | |
|---|---|
| | linear antara setiap fitur dengan label target, mampu menangkap hubungan yang tidak bersifat linear. |
| Normalisasi | Proses transformasi nilai fitur ke dalam skala tertentu agar tidak ada fitur yang mendominasi proses pelatihan akibat perbedaan rentang nilai. Dalam penelitian ini digunakan Standard Scaler. |
| Overfitting | Kondisi di mana model terlalu menyesuaikan diri dengan data latih sehingga performanya menurun secara signifikan pada data baru yang belum pernah dilihat sebelumnya. |
| Parameter Statis | Skenario pengujian di mana model dijalankan menggunakan nilai hyperparameter yang tetap tanpa melalui proses optimasi otomatis. |
| Pra-pemrosesan Data | Serangkaian tahapan persiapan data sebelum digunakan untuk melatih model, mencakup pembersihan data, pembagian dataset, dan normalisasi. |
| Presisi *(Precision)* | Proporsi prediksi positif yang benar-benar positif, mengukur ketepatan model dalam memprediksi suatu kelas. |
| Random Forest | Algoritma machine learning berbasis ensemble yang membangun sejumlah pohon keputusan secara acak dan menggabungkan hasilnya melalui mekanisme voting untuk menghasilkan prediksi akhir. |
| Recall | Proporsi data positif yang berhasil diidentifikasi dengan benar oleh model, mengukur kemampuan model dalam menemukan seluruh data dari suatu kelas. |
| Reduksi Dimensi | Proses pengurangan jumlah fitur dalam dataset dengan tujuan menghilangkan fitur yang tidak relevan atau redundan sehingga meningkatkan efisiensi dan performa model. |
| Redundansi | Kondisi di mana terdapat dua atau lebih fitur yang mengandung informasi yang sama atau sangat berkorelasi satu sama lain dalam sebuah dataset. |
| RFECV *(Recursive Feature Elimination* | Metode seleksi fitur berbasis wrapper |

| | |
|---|---|
| *with Cross-Validation)* | yang secara iteratif mengeliminasi fitur paling tidak relevan menggunakan estimator tertentu (Random Forest atau SVM) dan menentukan jumlah fitur optimal melalui validasi silang. |
| Seleksi Fitur *(Feature Selection)* | Proses memilih subset fitur yang paling relevan dari keseluruhan fitur yang tersedia untuk meningkatkan performa model, mengurangi kompleksitas, dan meminimalkan risiko overfitting. |
| Standard Scaler | Teknik normalisasi yang mentransformasi nilai fitur sehingga memiliki rata-rata nol dan standar deviasi satu berdasarkan distribusi data latih, dikenal juga sebagai Z-score normalization. |
| Stres Pelajar | Kondisi tekanan psikologis yang dialami pelajar akibat tuntutan akademis dan faktor lingkungan yang dapat berdampak negatif terhadap kesehatan mental dan performa belajar. |
| Student Stress Factors | Dataset publik yang digunakan dalam penelitian ini, tersedia secara terbuka dan terdiri dari 1.100 sampel dengan 20 fitur prediktor yang merepresentasikan berbagai faktor penyebab stres pada pelajar. |
| Support Vector Machine (SVM) | Algoritma machine learning yang bekerja dengan mencari hyperplane optimal yang memisahkan kelas-kelas data dengan margin terbesar dalam ruang fitur berdimensi tinggi. |
| Train-Test Split | Proses pembagian dataset menjadi data latih dan data uji. Dalam penelitian ini digunakan rasio 75:25, di mana 75% data digunakan untuk melatih model dan 25% untuk pengujian. |
| Validasi Silang *(Cross-Validation)* | Teknik evaluasi model yang membagi data latih menjadi beberapa lipatan untuk memvalidasi performa model secara lebih robust dan mengurangi bias estimasi. |
| Wrapper | Kategori metode seleksi fitur yang mengevaluasi subset fitur menggunakan algoritma machine learning tertentu |

sebagai evaluator. Dalam penelitian ini diwakili oleh metode RFECV.

# BAB I
# PENDAHULUAN

## 1.1 Latar Belakang

Menurut laporan Indonesia National Adolescent Mental Health Survey (I-NAMHS) pada tahun 2022, satu dari tiga remaja di Indonesia teridentifikasi memiliki masalah kesehatan mental [1]. Temuan ini mengindikasikan bahwa kelompok remaja, yang mayoritas berstatus sebagai pelajar, sangat rentan terhadap permasalahan psikologis. Salah satu bentuk gangguan psikologis yang paling umum dan meluas dialami oleh pelajar adalah stres. Menurut Adawiyah dalam Yuda, prevalensi pelajar yang mengalami stres di Indonesia tercatat sebesar 36,7–71,6%, lebih tinggi dibandingkan rata-rata Asia yang berkisar 39,6–61,3% dan rata-rata global sebesar 38–71% [2]. Tingginya angka ini menunjukkan bahwa stres pada pelajar bukan sekadar persoalan individual, melainkan telah menjadi fenomena global yang serius dan memerlukan penanganan yang terstruktur. *World Health Organization* (WHO) mendefinisikan stres secara umum sebagai keadaan kekhawatiran atau ketegangan emosional yang muncul sebagai respons alami tubuh saat menghadapi situasi yang menantang atau sulit [3]. Bagi seorang pelajar, stres tidak hanya dialami melalui jalur akademik, tetapi juga rentan mengalami stres secara general akibat berbagai dinamika kehidupan sehari-hari yang kompleks [4]. Berbagai tekanan hidup yang tidak terkelola dengan baik pada pelajar ini dapat memberikan dampak buruk yang berujung pada penurunan performa akademis sekaligus menghambat proses pertumbuhan dan kualitas hidup mereka secara keseluruhan [5]. Lebih lanjut, jika kondisi stres berlebih ini dibiarkan berkepanjangan tanpa penanganan, dampaknya dapat memburuk menjadi gangguan kecemasan hingga depresi berat [6]. Mengingat rentetan dampak tersebut, deteksi dini terhadap tingkat stres pada pelajar sangat diperlukan. Langkah ini dibutuhkan agar pihak berwenang dapat segera memberikan intervensi dan penanganan yang cepat sebelum kondisi psikologis pelajar semakin memburuk [7].

Perkembangan teknologi saat ini menawarkan solusi baru dalam permasalahan deteksi stres pelajar secara dini, yaitu melalui pendekatan *Machine Learning* (ML). Secara sederhana, *Machine Learning* merupakan pendekatan dari bidang teknologi

yang memungkinkan komputer belajar mengenali pola dari data secara mandiri [8]. Metode ini telah terbukti dalam berbagai penelitian untuk melakukan pendeteksi dini kondisi stres pelajar berdasarkan data akademik maupun data pribadi mereka oleh Amalia [9]. Dibandingkan dengan metode manual yang cenderung memakan waktu, *Machine Learning* memiliki keunggulan dalam mengolah data secara otomatis. Kemampuan ini memungkinkan pihak yang berwenang untuk mendeteksi dini, sehingga penanganan yang tepat dapat diberikan sebelum kondisi pelajar memburuk.

Namun, penerapan *Machine Learning* pada data survei yang cukup banyak dan berdimensi tinggi seringkali menyebabkan tantangan teknis. Dataset survei stres pelajar umumnya memuat data dalam jumlah besar, dimana tidak semua pertanyaan dalam survei dapat berkontribusi terhadap perancangan ML. Penelitian oleh Lindner menyatakan bahwa data berasal dari kuesioner seringkali memiliki masalah redundansi dan variabel bebas yang memiliki hubungan sangat kuat sehingga menyebabkan informasi yang berulang [10]. Selain itu, penelitian Saifuddin menemukan bahwa data kesehatan mental rentan terhadap gangguan atau *noise* akibat jawaban yang subjektif oleh responden [11]. Temuan oleh Jamilah bahwa dataset survei dapat memiliki noise dan nilai yang hilang sehingga dapat berdampak kepada model jika tidak di proses dengan baik [12].

Untuk mengatasi permasalahan tersebut, teknik seleksi fitur atau *feature selection* dapat menjadi solusi. Seleksi fitur adalah teknik untuk mereduksi dimensi pada data melalui pembuangan fitur yang kurang informatif, sehingga model dari *machine learning* dapat fokus pada fitur yang paling informatif [13]. Studi oleh Amalia menunjukkan bahwa penerapan seleksi fitur pada data stres pelajar secara signifikan meningkatkan akurasi model dibandingkan dengan menggunakan seluruh fitur tanpa seleksi [9]. Seleksi fitur sendiri memiliki tiga pendekatan utama: metode *filter* yang menggunakan metode statistika dalam pemilihan fitur dan efisien dalam kecepatan; metode *wrapper* yang menggunakan sebuah model dalam penilaian fitur dan mengutamakan interaksi antar fitur untuk akurasi tinggi; metode *embedded* yang menyeimbangkan efisiensi dan performa model [14].

Dalam perkembangan riset terkini, dataset stres pelajar yang terbentuk dari survei mulai banyak dipakai sebagai objek evaluasi model deteksi stres karena cakupan fitur-fiturnya yang luas, salah satu datasetnya yaitu *Student Stress Factors* dataset yang diperoleh melalui survei oleh Acharya [15].Dataset ini mencakup banyak faktor yang dapat menyebabkan stres dari pelajar meliputi aspek akademik, sosial, dan lingkungan. Studi terbaru yang memanfaatkan dataset ini antara lain dilakukan oleh Amalia yang berfokus pada optimasi algoritma *Random Forest* menggunakan pendekatan seleksi fitur *wrapper* (*Backward Elimination*) [9].

Namun, tinjauan terhadap literatur tersebut menunjukkan bahwa eksplorasi metode seleksi fitur yang diterapkan masih bersifat tunggal pada satu metode. penelitian tersebut belum melibatkan pendekatan lainnya yaitu *filter dan embedded* sebagai pembanding utama dalam skenario pengujian yang sama. Kurangnya literatur dalam analisis komparatif yang menyandingkan ketiga pendekatan *filter*, *wrapper*, dan *embedded* secara bersama menciptakan kesenjangan mengenai metode mana yang paling baik untuk data stres tersebut.

Di sisi lain, penelitian yang dilakukan oleh Ariyoga berhasil dalam menyandingkan ketiga pendekatan utama dari seleksi fitur tersebut secara bersamaan dengan menggunakan model algoritma *Random Forest* dan SVM untuk klasifikasi data spektrum [16]. Penelitian tersebut membuktikan bahwa untuk menemukan model terbaik, penting dalam membandingkan langsung ketiga pendekatan tersebut secara berdampingan karena setiap metode memiliki karakteristik yang berbeda terhadap performa model. Namun, studi tersebut diterapkan pada data NIRS mangga, yang memiliki struktur data yang berbeda dengan data stres hasil survei.

Untuk menjawab tantangan tersebut, penelitian ini menerapkan ketiga pendekatan seleksi fitur yaitu *filter, wrapper, embedded* dengan empat metode seleksi fitur yang masing-masing difokuskan untuk menangani karakteristik spesifik data kuisioner atau data survei stres. Pertama, metode dengan kategori *filter* yaitu ANOVA digunakan sebagai standar untuk mengukur hubungan linear antara skor numerik dan kategori stres [17]. Kedua, metode dengan kategori *filter* yaitu *Mutual Information* (MI) dipilih untuk menangkap pola non-linear dan distribusi data yang tidak normal pada skala

kuesioner [18] Ketiga, metode dengan kategori *wrapper* RFECV diterapkan untuk mengevaluasi interaksi antar fitur yang saling mempengaruhi, pemilihan metode ini didasarkan pada kemampuannya untuk mengeliminasi fitur secara rekursif melalui validasi silang guna mencari subset paling optimal, pendekatan yang telah terbukti dalam studi deteksi stress oleh Ovi [19]. Terakhir, metode dengan kategori *embedded* yaitu LASSO digunakan untuk mengatasi masalah pertanyaan yang redundan dengan secara otomatis membuang fitur yang tidak efisien untuk menghasilkan model yang ringkas [20].

Model klasifikasi yang digunakan yaitu menggunakan algoritma SVM dan *Random Forest* yang telah terbukti efektif pada studi sejenis oleh Feibi dan Amalia [20][9]. Penggunaan model teruji ini bertujuan menetapkan standar kinerja yang stabil. Lebih lanjut, guna memastikan bahwa evaluasi metode seleksi fitur dilakukan secara maksimal, penelitian ini menerapkan teknik *hyperparameter tuning* sebagai alat validasi performa yang dihasilkan. Berbeda dengan penelitian Feibi yang menggunakan pendekatan *Grid Search* dalam memaksimalkan potensi model dalam klasifikasi data stres pelajar, penelitian ini mengimplementasikan *Bayesian Optimization* melalui Optuna, yang merupakan metode yang memfasilitasi proses pencarian yang lebih efisien dengan mempertimbangkan hasil dari uji coba sebelumnya. Hal ini memungkinkan identifikasi kombinasi hyperparameter yang terbaik [21]. Pendekatan ini berfungsi untuk mendorong model ke kondisi paling optimal, sehingga analisis dapat difokuskan sepenuhnya pada dampak metode seleksi fitur terhadap peningkatan performa model.

Oleh karena itu, penelitian ini bertujuan untuk melakukan analisis komparatif terhadap metode seleksi fitur berbasis *filter* (ANOVA dan *Mutual Information*), *wrapper* (RFECV), dan *embedded* (LASSO) pada dataset *Student Stress Factors* dengan model klasifikasi Random Forest dan SVM. Penelitian ini difokuskan untuk mengukur dampak penerapan ketiga pendekatan tersebut terhadap performa model klasifikasi, baik pada parameter rujukan maupun saat model dioptimasi menggunakan *Bayesian Optimization,* guna menentukan metode *feature selection* yang paling optimal untuk dataset stres pelajar. Hasil perbandingan komprehensif ini diharapkan

dapat menentukan metode *feature selection* terbaik dengan dataset stres pelajar yang digunakan.

## 1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, rumusan masalah penelitian ini adalah sebagai berikut.

1. Apa pengaruh penerapan metode seleksi fitur berbasis Filter (ANOVA, Mutual Information), Wrapper (RFECV), dan Embedded (LASSO) terhadap performa model klasifikasi stres pelajar pada algoritma *Random Forest* dan SVM dibandingkan tanpa seleksi fitur?
2. Metode seleksi fitur manakah yang menghasilkan performa model paling optimal terhadap model klasifikasi *Random Forest* dan SVM pada data survei stres pelajar?

## 1.3 Tujuan Penelitian

Berdasarkan rumusan maslaah yang telah disebutkan, tujuan penelitian ini adalah sebagai berikut.

1. Membandingkan pengaruh metode seleksi fitur berbasis Filter (ANOVA, Mutual Information), Wrapper (RFECV), dan Embedded (LASSO) terhadap performa model *Random Forest* dan SVM dalam klasifikasi tingkat stres pelajar.
2. Menentukan metode seleksi fitur berbasis Filter (ANOVA, Mutual Information), Wrapper (RFECV), dan Embedded (LASSO) yang paling optimal terhadap performa model *Random Forest* dan SVM dalam klasifikasi tingkat stres pelajar.

## 1.4 Pembatasan Masalah

Agar pengerjaan penelitian menjadi lebih terarah, maka dibuat batasan masalah pada penelitian ini, diantaranya yaitu:

1. Fokus utama penelitian ini adalah menganalisis komparasi performa metode seleksi fitur (*Feature Selection*) dari tiga pendekatan berbeda, yaitu pendekatan *filter* (menggunakan ANOVA dan *Mutual Information*), pendekatan *wrapper* (menggunakan *Recursive Feature Elimination with Cross-Validation* / RFECV), dan pendekatan *embedded* (*menggunakan Least Absolute Shrinkage*

*and Selection Operator* / LASSO).

2. Pengujian kinerja metode seleksi fitur tersebut diterapkan pada dua algoritma *machine learning*, yaitu *Support Vector Machine* (SVM) dan *Random Forest* serta penggunaan metode *hyperparameter tuning Bayesian Optimization*. Algoritma *machine learning* dan penggunaan *hyperparameter tuning* yang digunakan tidak menjadi objek utama penelitian, melainkan berperan sebagai instrumen evaluasi untuk mengukur dampak penerapan metode seleksi fitur terhadap kinerja model. Dengan demikian, penelitian ini tidak membahas optimasi sangat mendalam terhadap parameter maupun pengembangan arsitektur algoritma klasifikasi yang digunakan.
3. Data yang digunakan merupakan data sekunder publik berjudul *Student Stress Factors* oleh Acharya [15]. Dataset ini terdiri dari 1.100 baris dan 21 kolom, yang mencakup 20 fitur faktor penyebab stres dan 1 kolom label target (stress_level) yang memiliki 3 kelas kategori.

### 1.5 Sistematika Penulisan

Dalam penelitian ini, pembahasan disajikan menjadi lima bab dengan sistematika pembahasan sebagai berikut.

BAB I : PENDAHULUAN menjelaskan latar belakang dan motivasi dari penelitian ini serta memberikan ulasan singkat terkait permasalahan yang dihadapi. Selain itu, tujuan dan manfaat yang diharapkan dari penelitian ini juga dijelaskan pada bab ini.

BAB II : TINJAUAN PUSTAKA menguraikan literatur terkait mengenai metode seleksi fitur yang digunakan dalam perbandingan, algoritma klasifikasi yang digunakan serta teori-teori yang relevan dengan penelitian ini.

BAB III : METODOLOGI PENELITIAN merincikan tahap-tahap penelitian, mulai dari data yang digunakan, mempersiapkan dataset, apa saja yang dilakukan saat menjalankan seleksi fitur dan yang digunakan sebelum menjalankan pengklasifikasian menggunakan algoritma yang ditetapkan, sampai dengan evaluasi model yang telah dibuat.

BAB IV : HASIL DAN ANALISIS menyajikan hasil pengujian metode klasifikasi, kelebihan, dan kekurangan dari metode yang digunakan akan dibahas secara mendalam.

BAB V : PENUTUP menyimpulkan keseluruhan dari hasil penelitian yang terdapat jawaban terhadap rumusan masalah yang diajukan dalam penelitian. Selain itu, saran dan rekomendasi untuk penelitian lebih lanjut serta potensi peningkatan metode yang digunakan akan diuraikan.