Perbandingan Kinerja Random Forest Dan Smote Random Forest Dalam Mendeteksi Dan Mengukur Tingkat Stres Pada Mahasiswa Tingkat Akhir

Vionita Oktaviani[1], Neny Rosmawarni[2], M. Panji Muslim[3]
[1,2,3]Fakultas Ilmu Komputer / S1 Informatika
[1,2,3]Universitas Pembangunan Nasional Veteran Jakarta
[1,2,3]Jl. RS. Fatmawati, Pondok Labu, Jakarta Selatan, DKI Jakarta, 12450, Indonesia
vnoktaviani@gmail.com[1], nenyrosmawarni@upnvj.ac.id[2], muhammadpanji@gmail.com[3]

Abstrak. Dalam kehidupan sehari-hari manusia, stres merupakan masalah nyata sehingga menjadi bagian yang tidak terpisahkan. Ketidaksiapan individu dalam menghadapi tuntutan akademis dapat mengakibatkan stres sebagai salah satu gangguan psikologis. Dalam hal ini, stres akademik merupakan stres yang dialami oleh mahasiswa, terutama mahasiswa tingkat akhir. Adanya banyak tekanan baik dari masalah ekonomi, akademik maupun kondisi sosial dapat menjadi pemicu stres bagi mahasiswa tingkat akhir. Penelitian ini bertujuan untuk mengklasifikasikan diagnosa tingkat stress mahasiswa tingkat akhir dengan membandingkan kinerja yang terbaik antara Random Forest dengan SMOTE Random Forest. Data yang diolah dalam penelitian ini merupakan data yang dihasilkan oleh kuesioner yang berisi 14 pertanyaan yang ditujukan pada mahasiswa tingkat akhir yang sedang melaksanakan skripsi. Adapun hasil dari penelitian ini, disimpulkan bahwasannya metode Random Forest dengan menggunakan SMOTE mampu mempengaruhi dan meningkatkan evaluasi klasifikasi studi kasus klasifikasi diagnosa mahasiswa tingkat akhir dengan akurasi sebesar 71%, precision sebesar 72% dan recall sebesar 71% pada pembagian 80% data latih, 20% data uji dengan nilai K=5.

Kata Kunci: Klasifikasi, Random Forest, Oversampling SMOTE, Stress.

1 Pendahuluan

Pendidikan merupakan bagian aspek penting yang akan menjadi bagian dari pertumbuhan seorang manusia, baik aspek kognitif, aspek sosial-emosional, maupun aspek moral dan etika. Mengingat pendidikan merupakan aspek terpenting dalam kehidupan manusia, maka setiap orang Indonesia berhak mendapatkannya dan diharapkan dapat melanjutkan pendidikan sepanjang hidupnya.

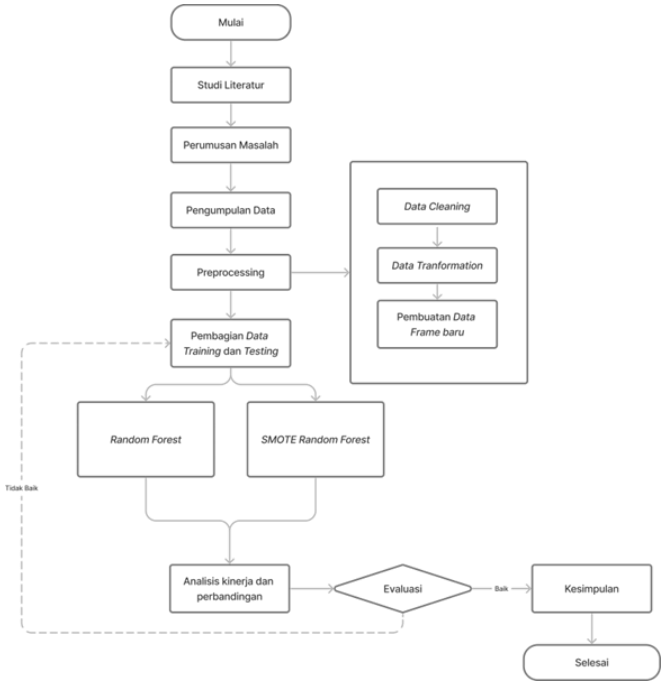
Jenjang perkuliahan merupakan salah satu jenjang pendidikan. Mahasiswa yang sedang berproses dalam mencapai kelayakan penyandang gelar yang akan diberikan pada akhir namanya dengan memenuhi beberapa persyaratan yang telah diajukan oleh pihak akademik maupun program studi sebelumnya seperti biaya administrasi dan penyelesaian SKS (Satuan Kredit Semester) merupakan mahasiswa tingkat akhir atau mahasiswa akhir [1][2].

Adapun tantangan yang umum dihadapi oleh mahasiswa akhir sebagian besar merasa kesulitan bahkan tidak jarang cukup banyak juga mahasiswa yang merasa tidak mampu untuk menyelesaikan persyaratan kelulusan untuk dapat mengajukan pembuatan tugas akhir, skripsi, tesis maupun disertasi sehingga berakhir dengan mengundurkan diri dari perkuliahannya [3][4].

Maka berdasarkan penjelasan sebelumnya akan dilakukan perbandingan antara Random Forest dan SMOTE Random Forest untuk melihat dan mengevaluasi kinerja diantara keduanya dalam studi kasus mendeteksi dan mengukur tingkat stress pada mahasiswa tingkat akhir.

2 Metodologi Penelitian

2.1 Tahapan Penelitian



Gambar. 1. Tahapan Penelitian

Studi Literatur

Tahapan awal serta tahapan yang paling penting dalam sebuah penelitian yaitu studi literatur. Dengan melakukan studi literatur dapat membuka wawasan serta informasi dan kajian - kajian peneliti terkait hasil yang sekiranya akan menjadi landasan teori dalam penelitian. Pada tahap ini peneliti mencari landasan penelitian dan informasi terkait penelitian yang relevan dengan penelitian mendeteksi dan mengukur tingkat stres mahasiswa tingkat akhir dengan beberapa algoritma machine learning[5][6][7].

Perumusan Masalah

Dalam studi literatur awal, setelah mengumpulkan informasi umum tentang subjek penelitian, langkah berikutnya adalah merumuskan masalah. Rumusan masalah tidak hanya memperkuat alasan penelitian, tetapi juga membimbing penelitian dengan merinci pertanyaan kunci, mengidentifikasi kekosongan pengetahuan, dan menyoroti relevansi serta potensi kontribusinya. Selain itu, rumusan masalah berperan dalam merumuskan tujuan penelitian, memberikan panduan untuk perencanaan metodologi, serta mengidentifikasi manfaat potensial baik untuk kemajuan ilmu pengetahuan maupun aplikasinya. Dalam proses ini, peneliti menetapkan batasan penelitian untuk mengelola cakupan secara efektif. Dengan demikian, rumusan masalah menjadi landasan kokoh yang membimbing penelitian menuju hasil yang lebih bermakna.

Pengumpulan Data

Penelitian ini menggunakan metode pengumpulan data melalui penyebaran kuesioner, yang diisi oleh 528 mahasiswa tingkat akhir. Tujuan dari pengumpulan data ini adalah untuk mengukur dan menganalisis tingkat stres yang dialami oleh mahasiswa pada tahap akhir studi mereka. Kuesioner mencakup sejumlah variabel respon yang dirancang untuk menggambarkan dan memahami faktor-faktor yang berkontribusi terhadap tingkat stres tersebut. Analisis terhadap hasil kuesioner ini akan memberikan gambaran yang lebih mendalam tentang dinamika stres pada populasi mahasiswa tingkat akhir yang menjadi fokus penelitian ini[8][9][10].

Preprocessing Data

Meninjau dan membersihkan data pada langkah ini untuk melihat apakah data tersebut mengandung nilai yang hilang, data yang tidak konsisten, atau variabel yang tidak relevan untuk pemodelan merupakan bagian dari tahapan preprocessing data. Pada tahap ini, pembagian interval juga digunakan untuk membagi proses transformasi data ke dalam setiap kelas atau jenis tingkat stress pada mahasiswa tingkat akhir. Tahapan preprocessing yang dilakukan pada penelitian ini, yaitu sebagai berikut[11].

Data Cleaning

Pada tahap ini data dibersihkan yakni dengan pengecekan missing value ataupun data yang duplikat serta dilakukan pengecekan korelasi antar data sehingga dapat mengapus menghapus kolom yang tidak digunakan dalam proses penelitian seperti timestamp, nama dan email.

Data Transformation
Pada tahap ini data diubah atau bertransformasi dengan cara mengubah data yang memiliki tipe kategorikal menjadi numerik dengan label encoding dan menormalisasi data dalam skala 0 - 4 pada beberapa kolom data yang belum memiliki skala dengan rentang tersebut agar tidak mendistorsi perbedaan dalam rentang nilai atau kehilangan informasi serta membuat model algoritma dengan benar.

Pembuatan Data Frame baru
Pembuatan data frame baru diharapkan agar mempermudah analisis, pemilihan fitur, pemodelan, meningkatkan kualitas data dan menyederhanakan struktur data.

Pembagian Data
Pada tahap ini, dilakukan pembagian data training dan testing dengan rasio 80:20 pada suatu data frame setelah melalui proses preprocessing. Preprocessing data melibatkan langkah-langkah seperti penanganan nilai yang hilang dan normalisasi untuk memastikan kualitas data. Metode train-test split digunakan sebagai pendekatan dalam pembagian data, di mana 80% dari dataset digunakan untuk melatih model, dan 20% untuk menguji kinerja model. Dengan mengoptimalkan rasio ini dan menerapkan train-test split, penelitian ini memastikan bahwa model dapat belajar dari data yang cukup dan diuji pada dataset independen untuk mengukur kinerjanya secara objektif, sambil memberikan fleksibilitas untuk validasi model jika diperlukan.

Pembuatan Model
Dalam langkah ini, pembuatan model dilakukan dengan memanfaatkan data yang telah terbagi menjadi data pelatihan (training) dan data pengujian (testing). Pustaka (library) Python, yaitu Random Forest Classifier, digunakan untuk melakukan pemodelan menggunakan algoritma Random Forest serta SMOTE Random Forest. Setelah model terbentuk, tahap berikutnya melibatkan pemeriksaan dan evaluasi kinerja model tersebut ketika data yang telah dibersihkan dimasukkan ke dalamnya. Proses ini memungkinkan untuk mengukur sejauh mana model dapat memberikan hasil yang akurat dan dapat diandalkan setelah melibatkan data yang telah disiapkan secara terstruktur.

Evaluasi
Dalam evaluasi hasil klasifikasi data, model yang telah terbentuk pada tahap sebelumnya akan dinilai melalui berbagai metrik performa, termasuk accuracy, recall/sensitivitas, precision, dan F-Score. Accuracy bertujuan untuk mengukur sejauh mana model mampu mengklasifikasikan data secara benar secara keseluruhan, sementara recall/sensitivitas berfokus pada kemampuan model untuk mengidentifikasi seluruh instance positif yang ada. Precision memberikan gambaran tentang seberapa tepat model dalam mengklasifikasikan instance positif, sedangkan F-Score menggabungkan kedua aspek ini menjadi satu metrik yang menyeluruh. Dengan menggunakan nilai-nilai metrik ini, evaluasi model dapat memberikan wawasan mendalam tentang seberapa baik model dapat menangani dan mengklasifikasikan data sesuai dengan kebutuhan yang diinginkan.

3. Hasil dan Penelitian

3.1 Data

Tahap pengumpulan data tingkat stres pada mahasiswa tingkat akhir pada penelitian ini didapatkan dengan menyebarkan kuesioner kepada responden dengan kriteria sebagai mahasiswa tingkat akhir yang sedang melaksanakan skripsi dengan 14 variabel pertanyaan yang merujuk pada Perceived Stress Scale (PSS) atau skala stres yang dirasakan. PSS adalah alat pengukuran yang digunakan untuk mengukur tingkat stres yang dirasakan oleh seseorang dalam kehidupan sehari-hari dan didapatkan 258 responden pada penelitian ini.

3.2 Preprocessing Data

1. Data Cleaning
   Pembersihan data atau yang dikenal dengan data cleaning dilakukan pada data ini dengan melakukan pengecekan terkait missing value dan duplikasi data. Berdasarkan data yang didapatkan dari hasil sebaran kuesioner, tidak didapati missing value dan duplikasi data. Pada tahap ini juga dilakukan

menghapus atau menghilangkan kolom yang tidak diperlukan dalam penelitian, seperti kolom Timestamp, Email Address, Nama, Nomor HP/E-Wallet.

2. Data Transformation
   Setelah dilakukan pembersihan data, dilakukan transformasi data dengan cara mengubah data yang bersifat kategorikal sebelumnya menjadi numerik dengan menggunakan label encoding dan melakukan normalisasi nilai antara 0 – 4 ke beberapa kolom, seperti Status tingal, Adanya perubahan pola tidur, dan Adanya peningkatan/penurunan nafsu makan secara signifikan.

3. Pembuatan DataFrame baru
   "Pembuatan DataFrame baru" mengacu pada proses membuat struktur data yang disebut DataFrame di Python, biasanya menggunakan pustaka pandas. DataFrame adalah struktur data dua dimensi (mirip dengan tabel dalam database, spreadsheet, atau matriks) yang digunakan untuk menyimpan dan mengelola data dalam baris dan kolom.

3.3 Pembagian Data

Pembagian data dilaksanakan guna memisahkan data menjadi data pelatihan, pengujian dan validasi. Data pelatihan akan dimanfaatkan oleh algoritma klasifikasi untuk konstruksi model klasifikasi, data pengujian digunakan untuk mengukur kinerja dan keakuratan model saat melakukan klasifikasi, sementara data validasi digunakan untuk mengevaluasi kinerja model dan melakukan penyetelan parameter selama proses pengembangan model machine learning. Dalam penelitian ini, pembagian antara data pelatihan dan data pengujian dipraktikkan dengan metode train_test_split. Sebelum melakukan pembagian data, fitur dan label akan disisihkan, dengan X mewakili fitur dan Y mewakili label dengan perbandingan 80:20 dalam konteks data diagnosa tingkat stres mahasiswa tingkat akhir yang termasuk dalam kelas "Normal", "Ringan", "Sedang", "Berat", dan "Sangat Berat" Adapun pembagian data pada algoritma Random Forest dijelaskan pada Tabel 1 Pembagian data Random Forest dan SMOTE Random Forest pada Tabel 2.

Tabel. 1 Pembagian data Random Forest.

Rasio Perbandingan	Total Data	Normal	Ringan	Sedang	Berat	Sangat Berat
	258	42	65	92	46	13
Data latih 80%	206	35	59	68	32	12
Data uji 20%	52	7	6	24	14	1

Tabel. 2 Pembagian data SMOTE Random Forest.

Rasio Perbandingan	Total Data	Normal	Ringan	Sedang	Berat	Sangat Berat
	460	92	92	92	92	92
Data latih 80%	368	79	79	71	72	67
Data uji 20%	92	13	13	21	20	25

3.4 Pembentukan Model

Pembentukan model Random Forest dilakukan dengan library scikit learn yaitu Random Forest Classifier. Pada pemodelan akan dilakukan pemanggilan fungsi klasifikasi Random Forest. Lalu data training akan dimasukkan pada fungsi RandomForestClassifier dengan n_estimator 100.

3.5 Evaluasi

Tahapan terakhir dari perbandingan kinerja Random Forest dan SMOTE Random Forest dalam mendeteksi dan mengukur tingkat stres pada mahasiswa tingkat akhir ialah melihat hasil evaluasi setiap model machine learning yang telah dibuat dengan tujuan untuk mengevaluasi performa model. Untuk mengevaluasi model, Confusion Matrix digunakan untuk menghitung nilai akurasi, precision, recall dan F1 Score menggunakan nilai

True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Berikut adalah lampiran evaluasi model dari percobaan train test split data rasio 80% data latih dan 20% dengan nilai K=5.

Adapun tabel berdasarkan gambar confussion matrix Random Forest dapat dijelaskan pada Tabel 3 berikut.

Gambar. 1. Confussion matrix Random Forest.

Tabel. 3 Pembagian CM

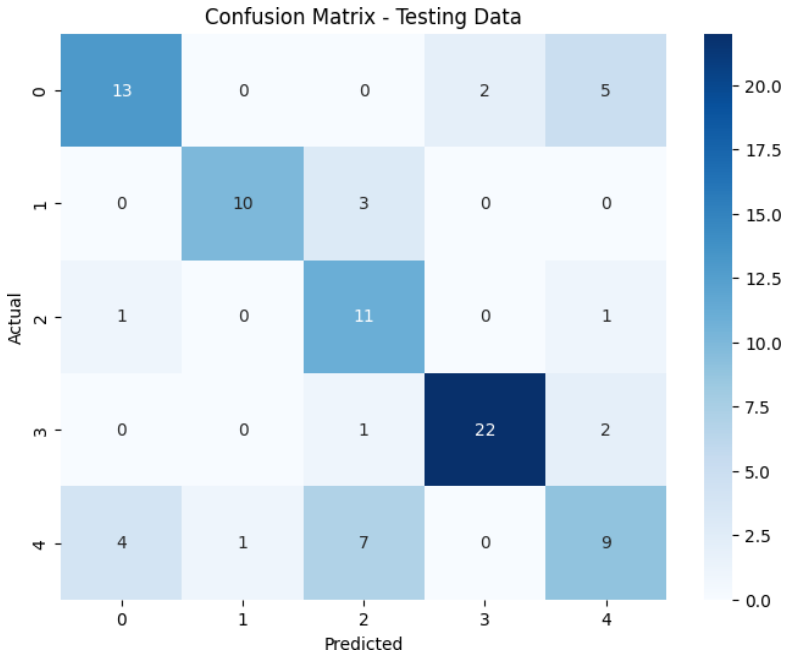
Kelas	Nilai Confussion Matrix				Jumlah
	True Positive	True Negative	False Positive	False Negative	
Berat	7	32	6	7	52
Normal	4	41	4	3	
Ringan	2	39	7	4	
Sangat Berat	1	51	0	0	
Sedang	14	21	7	10	

Berdasarkan nilai TP, TN, FP, dan FN diatas dapat dilakukan perhitungan untuk mendapatkan nilai accuracy, precision, recall, F1 Score (2.1).

Accuracy = (Total TP)/(Jumlah Data)
= (7+ 4 + 2 + 1 + 14)/52
= 28/52
= 0.538 ≈ 0.54

Pada tabel 4 juga akan diberikan penjelasan terkait precission, recall dan F1 Score.

Kelas	Precision	Recall	F1-Score
Berat	0.54	0.50	0.52
Normal	0.50	0.57	0.52
Ringan	0.22	0.33	0.27
Sangat Berat	1.00	1.00	1.00



Gambar. 2. Confussion matrix SMOTE Random Forest

Adapun tabel berdasarkan gambar confussion matrix SMOTE Random Forest dapat dijelaskan pada Tabel 4 berikut.

Tabel. 5. Pembagian data Uji SMOTE Random Forest

Kelas	Nilai Confussion Matrix				Jumlah
	True Positive	True Negative	False Positive	False Negative	
Berat	13	67	5	7	
Normal	10	78	1	3	
Ringan	11	68	11	2	92
Sangat Berat	22	65	1	3	
Sedang	9	63	8	12	

Tabel 6. Pembagian data uji SMOTE Random Forest.

Kelas	Precision	Recall	F1-Score
Berat	0.72	0.65	0.68
Normal	0.91	0.77	0.88
Ringan	0.50	0.85	0.63
Sangat Berat	0.92	0.88	0.90

Berdasarkan nilai TP, TN, FP, dan FN diatas dapat dilakukan perhitungan untuk mendapatkan nilai accuracy, precision, recall, F1 Score (2.1).

Accuracy = (Total TP)/(Jumlah Data)
= (13 + 10 + 11 + 22 + 9)/92
= 65/92
= 0.706 ≈ 0.71

Pada tabel 5 juga akan diberikan penjelasan terkait precission, recall dan F1 Score.

4 Kesimpulan

Berdasarkan analisis model Machine Learning pada klasifikasi diagnosa tingkat stress pada mahasiswa tingkat akhir menggunakan algoritma Random Forest, diperoleh bahwa:

1) Model dengan oversampling SMOTE menghasilkan performa terbaik pada rasio 80% data latih, 20% data uji dengan nilai K=5 mendapatkan akurasi sebesar 54%, precision sebesar 59% dan recall sebesar 60%.
2) Perbandingan akurasi yang didapatkan antara metode Random Forest tanpa oversampling SMOTE dan Random Forest dengan oversampling SMOTE memiliki perbedaan yang cukup signifikan. Hasil akurasi sebesar 71% menggunakan metode Random Forest dengan oversampling SMOTE mampu ditingkatkan pada data yang imbalaced yang sebelumnya tanpa menggunakan SMOTE memiliki akurasi sebesar 54%. Sehingga dapat disimpulkan bahwasannya metode tersebut berhasil untuk mempengaruhi dan meningkatkan evaluasi klasifikasi studi kasus klasifikasi diagnosa mahasiswa tingkat akhir.
3) Oversampling dengan metode SMOTE berhasil menyelesaikan tantangan ketidakseimbangan data pada kelas minoritas. Berdasarkan nilai recall, pendekatan oversampling SMOTE menunjukkan performa yang lebih unggul dalam mengevaluasi seberapa baik model memprediksi data secara akurat. Selain itu, pendekatan oversampling SMOTE tidak memerlukan volume besar data latih, karena metode ini menghasilkan data latih tambahan secara sintetis, secara efektif meningkatkan jumlah sampel dalam data latih.

Referensi

[1] Sudarsono, B. G., dan Lestari, S. P. Diagnosa Tingkat Depresi Mahasiswa Akhir Terhadap Penelitian Ilmiah Menggunakan Algoritma K-Nearest Neighbor. Vol. 4, pp. 1094–1099, 2020. https://doi.org/10.30865/mib.v4i4.2448

[2] Istammaroh, S. T. F. *Klasifikasi rekurensi pasien kanker payudara menggunakan metode Random Forest (RF)*. 2020, UIN Sunan Ampel Surabaya.

[3] Sultan Farel Syah Reza. *Implementasi Algoritma Random Forest Terhadap Prediksi Good Loan/Bad Loan Kredit Nasabah Bank Di Jakarta*. 2023. Universitas Pembangunan Nasional Veteran Jakarta.

[4] Pramana, S., Yuniarto, B., Mariyah, S., Santoso, I., dan Nooraeni, R. *Data mining dengan R konsep setara implementasi (1st ed.)*. Bogor : IN MEDIA, 2018.

[5] Google for Developers. *Machine Learning: Data Tidak Seimbang*. Retrieved December 5, 2023, from https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data?hl=id.

[6] Yarah, H. R. Perbandingan Random Forest Dan Smote Random Forest Pada Klasifikasi Berat Badan Lahir Rendah (BBLR), 2023.

[7] Giovanniello, J., Bravo-Rivera, C., Rosenkranz, A., dan Matthew Lattal, K. Stress, associative learning, and decision-making. *Neurobiology of Learning and Memory*, vol.204, pp.107812, 2023. https://doi.org/10.1016/J.NLM.2023.107812.

[8] Ambarwati, P. D., Pinilih, S. S., dan Astuti, R. T. Gambaran Tingkat Stres Mahasiswa. *Jurnal Keperawatan Jiwa* vol. 5, no.1.

[9] MA Putri, N Rosmawarni. Klasifikasi Kematangan Melinjo (Gnetum Gnemon Linn.) Berdasarkan Citra HSV dengan K-Nearest Neighbors. *Krea-TIF: Jurnal Teknik Informatika*, 2023

[10] N Rosmawarni, RD Amalia, Z Niqotaini. Pengaplikasian Penggunaan Microsoft Office Sebagai Media Pengajaran Dan Pembelajaran Bagi Guru Di Smks Mandiri Bojonggede Bogor. Jurnal Abdimas Bina Bangsa, 2023.

[11] TDR Octavia, N Rosmawarni, A Zaidiah. Implementasi Algoritma Multiple Linear Regression untuk Memprediksi Temperatur Udara Berdasarkan Kadar Zat Polutan di Kota Tangerang Selatan. JRSF Raya, 2024.