PERBANDINGAN RANDOM FOREST DAN SUPPORT VECTOR MACHINE UNTUK KLASIFIKASI CAPAIAN BELAJAR SISWA PADA DATA TIDAK SEIMBANG

Beby Vera Gabriel Silaban[1*], Harmi Sugiarti[2]
[1,2)]Program Studi Statistika, Universitas Terbuka, Indonesia

**Penulis korespondensi: 030271753@ecampus.ut.ac.id*

ABSTRAK

Ketidakseimbangan data (*imbalanced data*) pada capaian belajar siswa dapat memengaruhi performa model klasifikasi terutama dalam mengenali kelas minoritas. Tujuan riset ini mengevaluasi kinerja Random Forest pada klasifikasi capaian belajar dengan analisis komparatif terhadap model Support Vector Machine (SVM). Data yang digunakan merupakan data sekunder dari *Student Performance Analytics Dataset* yang terdiri dari 10.000 observasi dan 14 variabel sebelum tahap *preprocessing*. Rangkaian alur riset ini dimulai dengan menyiapkan dataset yang terdiri dari variabel input (X) dan variabel output (Y); melakukan tahap *preprocessing data* yang mencakup pembersihan data, penghapusan variabel yang tidak relevan serta sinkronisasi tipe data; melakukan analisis deskriptif untuk mengetahui karakteristik dan distribusi data; memisahkan *dataset* dilakukan dengan menetapkan proporsi 80:20 masing-masing untuk kebutuhan *training data* dan *testing data* menggunakan teknik *stratified sampling;* menyiapkan dua data skenario penelitian, yaitu tanpa *balancing* dan dengan *balancing* menggunakan metode *upsampling* pada data latih; membangun model klasifikasi menggunakan Random Forest; melakukan proses *centering* dan *scaling* pada Support Vector Machine (SVM); menjalankan teknis *kernel* Radial Basis Function (RBF) pada SVM; menguji model menggunakan data uji (*testing data)* untuk memprediksi capaian belajar; dan tahap akhir melakukan evaluasi pada model Random Forest tanpa *balancing,* Random Forest dengan *balancing,* SVM tanpa *balancing,* SVM dengan *balancing.* Evaluasi model disajikan dalam sebuah tabel *confusion matrix* yang menampilkan nilai *accuracy* dan *kappa* untuk mengukur kinerja klasifikasi secara keseluruhan sedangkan *precision, recall* dan *F-1 score* untuk mengevaluasi kemampuan model lebih rinci pada setiap kelas. Ditemukan nilai terbaik pada tabel *confusion matrix* dengan perolehan *accuracy* sebesar 95,89%, *kappa* 0,9341 dan *F1-score* 0,8235. Maka dapat disimpulkan bahwa analisis Support Vector Machine (SVM) lebih efektif dalam klasifikasi kondisi *imbalanced data.*

Kata kunci: Random Forest, Support Vector Machine, *Imbalance Data, Upsampli*ng, Capaian Belajar Siswa.

1 PENDAHULUAN

Penilaian capaian belajar siswa dalam sistem pendidikan merupakan metode efektif untuk mengukur perkembangan kemajuan siswa dalam setiap proses pembelajaran yang telah dilalui. Sarana tersebut memberikan informasi kepada pihak sekolah, guru, orangtua maupun wali sehingga mendapatkan informasi yang dapat dijadikan evaluasi untuk memperbaiki strategi pembelajaran dan pendampingan yang lebih tepat. Capaian belajar siswa dipengaruhi oleh berbagai faktor yang jika diteliti saling berkaitan baik dari aspek akademik maupun nonakademik. Penelitian pada bidang *Educational Data Mining* mengungkapkan bahwa nilai tugas, tingkat kehadiran, serta faktor lainnya memberikan pengaruh terhadap hasil capaian belajar siswa (Albreiki et al., 2021; Gusnina et al., 2022).

Menurut Zhao et al. (2024) serta Budiasto et al. (2025), perkembangan teknologi dan analisis data telah menjadi pendorong utama pemanfaatan metode *machine learning* untuk menganalisis capaian belajar siswa. James et al. (2021) dan Tan et al. (2019) mencatat bahwa regresi, *decision tree*, serta *naïve Bayes* merupakan metode yang cukup dominan dalam berbagai kajian terdahulu. Pemanfaatan Random Forest dalam dataset pendidikan sering kali disandingkan dengan performa Support Vector Machine (SVM), karena keduanya merupakan pendekatan yang paling populer di bidang ini. Adriansyah et al. (2022) serta Mustakim et al. (2025) mengungkapkan bahwa algoritma *machine learning* mampu memberikan performa yang optimal dalam mengklasifikasikan capaian belajar siswa. Random Forest mampu meningkatkan akurasi melalui penggabungan banyak pohon keputusan, sedangkan SVM efektif dalam membentuk batas keputusan yang optimal, terutama pada data berdimensi tinggi (James et al., 2021; Sen, 2022; Syarip et al., 2025).

Namun demikian, performa kedua metode tersebut sangat dipengaruhi oleh karakteristik data yang digunakan, khususnya pada kondisi ketidakseimbangan data (*imbalanced data*). Pada kondisi tersebut, model klasifikasi dapat lebih banyak mengenali kelas mayoritas sehingga kemampuan menjadi kurang optimal dalam mendeteksi kelas minoritas (He & Gracia, 2009). Admassu et al. (2024) serta Khairy et al. (2024) mencatat bahwa dalam situasi tersebut, model klasifikasi sering kali tidak objektif dengan condong ke arah kelas mayoritas, yang mengakibatkan penurunan tingkat akurasi dalam mendeteksi kelas minoritas. Meskipun Alamri et al. (2020) dan Syarip et al. (2025) telah melakukan studi Random Forest dengan membandingkan model Support Vector Machine (SVM) dalam ranah pendidikan, mayoritas riset tersebut cenderung menggunakan data yang proporsional atau kurang memerhatikan penanganan ketidakseimbangan data secara terstruktur.

Tujuan utama riset ini adalah mengklasifikasikan capaian belajar siswa dengan mengevaluasi kinerja Random Forest sebagai algoritma *ensemble* kemudian dilakukan pula analisis komparatif terhadap model Support Vector Machine (SVM). Evaluasi model dijalankan menggunakan metrik yang merepresentasikan ketepatan seperti *accuracy* dan *kappa*, serta metrik yang berfokus pada keseimbangan kelas seperti *precision, recall* dan *F-1 score.* Hasil penelitian dapat menjadi bahan pertimbangan dalam pemilihan metode klasifikasi yang tepat pada data pendidikan terutama pada kondisi *imbalanced data.*

2 METODE

2.1 Data dan Variabel

Untuk kebutuhan analisis, penelitian ini mengandalkan data sekunder diambil dari *Student Performance Analytics Dataset* di platform Kaggle, yang dapat diakses secara publik melalui tautan: [*https://www.kaggle.com/datasets/borovai0/student-performance-analytics-dataset*]. Secara keseluruhan, terdapat 12 variabel yang diseleksi dari 10.000 observasi dalam *dataset* ini guna melakukan klasifikasi terhadap tingkat capaian belajar siswa. Seluruh rangkaian proses pengolahan hingga analisis data diselesaikan dengan bantuan *software* RStudio. Variabel penelitian terdiri atas variabel independen (X) berupa faktor akademik, kebiasaan belajar, karakteristik siswa, serta variabel dependen (Y) berupa tingkat capaian belajar siswa (*grade)* dirangkum pada Tabel 1.

Tabel 1. Variabel Penelitian

No	Jenis Variabel	Nama Variabel	Keterangan	Skala Data
1	Akademik	assignment_score	Nilai tugas siswa	Rasio
		midterm_score	Nilai UTS	Rasio
		final_exam_score	Nilai UAS	Rasio
		participation_score	Partisipasi siswa	Rasio
2	Kebiasaan Belajar	study_hours_per_day	Jam belajar per hari	Rasio
		attendance_percentage	Persentase kehadiran	Rasio
		sleep_hours	Jam tidur	Rasio
		extra_classes	Kelas tambahan	Nominal
		internet_access	Akses internet	Nominal
3	Karakteristik	gender	Jenis kelamin	Nominal
		parent_education	Pendidikan orang tua	Ordinal
4	Capaian Belajar	grade	Tingkat capaian belajar	Ordinal

2.2 *Preprocessing* Data

Langkah persiapan data atau *data preprocessing* merupakan tahapan sebelum pemodelan dimulai, bertujuan untuk memperbaiki standar kualitas data sekaligus mereduksi kemungkinan kesalahan selama analisis berlangsung (Admassu et al., 2024). Pada tahap ini, variabel *student_id* dihapus karena hanya berfungsi sebagai identitas, sedangkan *overall_score* dihapus untuk menghindari *data leakage* akibat keterkaitannya dengan variabel target yaitu tingkat capaian belajar siswa (*grade*). Selanjutnya, variabel kategorik seperti *gender*, *internet_access*, *extra_classes*, *parent_education*, serta variabel target *grade* diubah menjadi tipe *factor* agar dapat digunakan dalam proses klasifikasi. Setelah itu dilakukan pemeriksaan struktur data untuk memastikan data siap digunakan dalam proses pemodelan. Normalisasi menggunakan metode *centering* dan *scaling* diterapkan pada tahap pemodelan Support Vector Machine (SVM) karena metode tersebut sensitif terhadap perbedaan skala data.

2.3 Pembagian Data

Pemisahan *dataset* merupakan langkah awal yang perlu dilakukan dengan membagi menjadi dua *subset*, yakni data latih (*training data)* sebagai pengembangan model dan untuk evaluasi menyeluruh menggunakan data uji (*testing data)*. Penggunaan metode *stratified sampling* dimaksudkan untuk menjaga keseimbangan distribusi kelas pada variabel target (*grade*) dalam kedua kelompok data, sehingga representasi capaian belajar siswa tetap konsisten selama proses pemisahan dataset. Langkah ini diambil demi menjamin distribusi kelas pada masing-masing *dataset* tetap mencerminkan karakteristik data penelitian secara utuh. Dengan demikian, hasil penilaian model yang diperoleh nantinya dapat memberikan gambaran yang jauh lebih akurat terkait kinerja klasifikasi. Proses pembagian data dilakukan dengan menetapkan proporsi 80:20, masing-masing untuk kebutuhan *training data* dn *testing data*. Data latih diolah menggunakan pendekatan Random Forest, yang kemudian disandingkan dengan teknik Support Vector Machine (SVM). Hasil dari kedua model ini nantinya akan dievaluasi melalui data uji guna mengukur ketepatan klasifikasi capaian belajar siswa.

2.4 Penanganan *Imbalanced Data*

Data pada penelitian ini memiliki kondisi ketidakseimbangan data (*imbalanced data*), sehingga penelitian dilakukan menggunakan dua skenario, yaitu tanpa *balancing* dan dengan *balancing*. Pada skenario tanpa *balancing*, proses pemodelan dilakukan menggunakan data asli. Sementara itu, pada skenario dengan *balancing*, dilakukan penyeimbangan kelas pada data latih (*training data*) menggunakan metode *upsampling*. Metode *upsampling* dilakukan dengan menambahkan sampel pada kelas minoritas hingga setiap kelas memiliki jumlah observasi yang sama. Melalui proses ini, sebaran kelas dapat dibuat lebih merata, sehingga diharapkan model menunjukkan tingkat sensitivitas yang lebih tinggi dalam mengidentifikasi kelas minoritas.

2.5 Pemodelan Random Forest

Sebagai pengembangan dari *ensemble learning*, Random Forest bekerja dengan mengumpulkan hasil prediksi dari berbagai pohon keputusan (*decision trees*) yang disusun secara paralel (Breiman, 2001). Struktur ini tidak hanya bertujuan untuk mencapai akurasi yang lebih tinggi tetapi juga untuk menjaga model agar tetap stabil dari ancaman *overfitting*. Keputusan akhir dari model ini secara sederhana ditentukan melalui mekanisme pemungutan suara terbanyak (*majority voting*) dari seluruh pohon.

Prediksi final $\hat{Y}$ ditentukan melalui pengambilan modus dari seluruh prediksi pohon keputusan, yang dirumuskan sebagai:

$$\hat{Y} = mode\,\{h_1(x), h_2(x), \dots, h_k(x)\}$$

Dimana $h_k(x)$ merepresentasikan keluaran dari pohon keputusan ke-k, dan k merupakan total keseluruhan pohon keputusan yang digunakan dalam model.

Implementasi model Random Forest dalam studi ini memanfaatkan *library* randomForest, dengan menetapkan parameter *ntree* atau jumlah pohon sebanyak 500 untuk menghasilkan klasifikasi yang optimal.

2.6 Pemodelan Support Vector Machine

Cortes dan Vapnik (1995) memperkenalkan Support Vector Machine (SVM) sebagai metode klasifikasi yang mengoptimalkan penentuan *hyperplane* sebagai batas pemisah antar kelas. Dalam SVM, formulasi untuk fungsi *hyperplane* tersebut dapat dirumuskan melalui persamaan berikut:

$$f(x) = w^T x + b$$

Dimana w merepresentasikan *Weight vector*, b nilai penambah atau bias dan vektor data yang menjadi masukan model adalah x.

Pemilihan *kernel* Radial Basis Function (RBF) dalam penelitian ini didasarkan pada kemampuannya untuk mengolah data nonlinier secara efektif. Rumusan fungsi *kernel* RBF tersebut adalah sebagai berikut:

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

$K(x_i, x_j)$ berperan sebagai fungsi *kernel* yang mengukur tingkat kemiripan di antara dua observasi, yaitu x_i dan x_j. Selain itu, γ berfungsi sebagai parameter yang menentukan sensitivitas dari *kernel* tersebut, sedangkan $\left\| x_i - x_j \right\|^2$ merepresentasikan kuadrat jarak Euclidean yang mencerminkan jarak antar pasangan data yang sedang dianalisis.

Normalisasi data menggunakan teknik *centering* dan *scaling* wajib dilakukan sebelum memasuki fase pemodelan SVM, mengingat sifat algoritma ini yang sensitif terhadap perbedaan rentang skala variabel (Alamri et al., 2020; Syarip et al., 2025). Selanjutnya, mengimplementasikan SVM dengan *kernel* RBF sebagai konfigurasi utama.

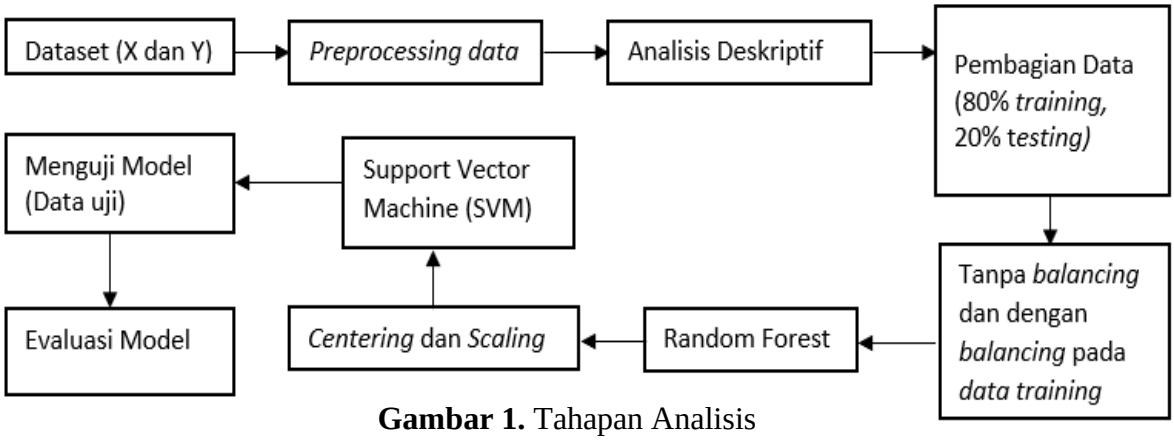
2.7 Evaluasi Model

Penggunaan *testing data* ditujukan untuk melihat kinerja klasifikasi capaian belajar siswa pada algoritma Random Forest dan prosedur yang sama juga dilakukan pada Support Vector Machine (SVM). Untuk menjalankan evaluasi model, digunakan metrik yang merepresentasikan ketepatan seperti *accuracy* dan *kappa,* serta metrik yang berfokus pada keseimbangan kelas seperti *precision, recall* dan *F-1 score* pada empat skenario penelitian yaitu Random Forest tanpa *balancing*, Random Forest dengan *balancing*, SVM tanpa *balancing*, SVM dengan *balancing*. Hasil prediksi dari model *balancing* maupun tanpa *balancing* dibandingkan pada satu tabel *confusion matrix*. Dengan pendekatan ini, performa model dapat dinilai secara objektif baik dari sisi ketepatan klasifikasi maupun stabilitas prediksinya pada kondisi data yang berbeda.

2.8 Tahapan Analisis

Secara ringkas, tahapan penelitian ditunjukkan pada Gambar 1 meliputi:

1. menyiapkan dataset terdiri dari variabel input (X) dan variabel output (Y),
2. melakukan tahap *preprocessing data* yang mencakup pembersihan data, penghapusan variabel yang tidak relevan, serta sinkronisasi tipe data,
3. melakukan analisis deskriptif untuk mengetahui karakteristik dan distribusi data,
4. memisahkan *dataset* dilakukan dengan menetapkan proporsi 80:20, masing-masing untuk kebutuhan *training data* dan *testing data* menggunakan teknik *stratified sampling*,
5. menyiapkan dua data skenario penelitian, yaitu tanpa *balancing* dan dengan *balancing* menggunakan metode *upsampling* pada data latih,
6. membangun model klasifikasi menggunakan Random Forest,
7. melakukan proses *centering* dan *scaling* pada data untuk pemodelan Support Vector Machine (SVM),
8. mengonstruksi model klasifikasi berbasis Support Vector Machine (SVM), dengan pemanfaatan teknis *kernel* Radial Basis Function (RBF),
9. menguji model menggunakan data uji (*testing data)* untuk memprediksi capaian belajar,
10. mengevaluasi model dilakukan melalui metrik *accuracy* serta *kappa*, yang kemudian diperkuat dengan analisis *precision*, *recall*, serta *F1-score*.



Gambar 1. Tahapan Analisis

3 HASIL DAN PEMBAHASAN

3.1 Analisis Deskriptif

Melalui tahapan *preprocessing*, dataset yang terkumpul berjumlah 10.000 observasi dengan 12 variabel. Sebaran data untuk variabel target kemudian dirangkum pada Tabel 2.

Tabel 2. Distribusi Variabel Target (*Grade*)

Grade	Jumlah	Persentase
A	154	1,54%
B	2704	27,04%
C	5073	50,73%
D	2008	20,08%
F	61	0,61%
Total	**10000**	**100%**

Distribusi variabel target (*grade*) menunjukkan kondisi ketidakseimbangan data (*imbalanced data*). Kelas C memiliki jumlah data terbanyak sedangkan kelas F memiliki jumlah data paling sedikit. Ketidakseimbangan data antar kelas berpotensi menurunkan efektivitas model klasifikasi, khususnya dalam kemampuan model untuk mengidentifikasi kategori minoritas dengan tepat.

Tabel 3. Analisis Deskriptif Variabel Numerik

Variabel	Mean	Minimum	Maximum	Standar Deviasi
study_hours_per_day	5,47	1	10	2,59
attendance_percentage	70,49	40	100	17,30
assignment_score	64,72	30	99,99	20,05
midterm_score	62,32	25	100	21,71
final_exam_score	64,98	30,01	100	20,22
participation_score	55,10	10	99,99	25,96
sleep_hours	6,51	4	9	1,46

Berdasarkan nilai standar deviasi pada Tabel 3, variabel *participation_score* memiliki nilai standar deviasi terbesar sehingga menunjukkan bahwa tingkat partisipasi siswa memiliki sebaran data yang paling beragam dibandingkan variabel lainnya. Adapun untuk variabel *sleep_hours*, nilai standar deviasi yang paling kecil menandakan bahwa variasi durasi tidur di antara siswa cenderung minim atau lebih homogen. Selain itu, rentang nilai minimum dan maksimum pada variabel *assignment_score, midterm_score,* dan *final_exam_score* menunjukkan adanya perbedaan capaian belajar siswa dirangkum pada Tabel 4.

Tabel 4. Analisis Deskriptif Variabel Kategorik

Variabel	Kategori	Jumlah	Persentase
Gender	Male	5013	50,13%
	Female	4987	49,87%
Internet Access	Yes	5061	50,61%
	No	4939	49,39%
Extra Classes	Yes	5087	50,87%
	No	4913	49,13%

Parent Education	Bachelor	2527	25,27%
	High School	2528	25,28%
	Master	2487	24,87%
	PhD	2458	24,58%

Distribusi variabel kategorik menunjukkan bahwa setiap kategori memiliki proporsi yang relatif seimbang. Variabel *gender* memperlihatkan jumlah siswa laki-laki hampir menyamai jumlah siswa perempuan. Pada variabel *internet_access*, jumlah siswa yang memiliki akses internet juga relatif seimbang dengan siswa yang tidak memiliki akses internet. Variabel *extra_classes* menunjukkan bahwa jumlah siswa yang mengikuti kelas tambahan hampir sama dengan siswa yang tidak mengikuti kelas tambahan. Selain itu, distribusi *parent_education* juga terlihat merata pada setiap kategori pendidikan orang tua.

3.2 Pembagian Data

Pemisahan dataset dilakukan dengan membagi menjadi dua subset, yakni data latih (*training data*) sebagai pengembangan model dan untuk evaluasi menyeluruh menggunakan data uji (*testing data*). Dataset dipisahkan melalui teknik *stratified sampling* menggunakan rasio 80:20. Metode ini mempertahankan proporsi masing-masing kelas pada variabel target, yaitu tingkat capaian belajar siswa (*grade*), dirangkum pada Tabel 5.

Tabel 5. Hasil Pembagian Data

Jenis Data	Jumlah Data	Persentase
Data Latih	8003	80%
Data Uji	1997	20%
Total	10000	100%

Berdasarkan Tabel 5, diperoleh sebanyak 8.003 *training data* dan 1.997 *testing data*. Proses pemisahan ini dilaksanakan sebelum tahap pemodelan untuk memastikan bahwa pengembangan model klasifikasi serta evaluasi kinerjanya dilakukan menggunakan data yang belum pernah diolah sebelumnya, dengan perincian distribusi variabel target (*grade*) pada kedua kelompok data tersebut dirangkum pada Tabel 6.

Tabel 6. Distribusi Grade pada Data Training dan Testing

Grade	Training Data	Persentase	Testing Data	Persentase
A	124	1,55%	30	1,50%
B	2164	27,04%	540	27,04%
C	4059	50,72%	1014	50,78%
D	1607	20,08%	401	20,08%
F	49	0,61%	12	0,60%

Berdasarkan Tabel 6, distribusi kelas pada data latih dan data uji relatif sama dengan distribusi pada dataset awal. Hasil tersebut menunjukkan bahwa metode *stratified sampling*

berhasil mempertahankan proporsi masing-masing kelas sehingga data yang digunakan dalam proses pemodelan tetap representatif.

3.3 Penyeimbangan Data (*Data Balancing)*

Berdasarkan hasil analisis deskriptif sebelumnya, distribusi variabel target, yaitu tingkat capaian belajar siswa (*grade*) menunjukkan kondisi ketidakseimbangan data (*imbalanced data*), terutama pada kelas A dan F yang memiliki kuantitas data yang jauh lebih sedikit dibandingkan kategori lainnya. Penyeimbangan data dilakukan dengan menggunakan metode *random oversampling* (*upsampling*) pada data latih (*training data*). Metode ini dilakukan dengan cara menambahkan sampel secara acak pada kelas jumlah terendah. Hasil distribusi data sebelum dan sesudah penerapan *upsampling* dirangkum pada Tabel 7.

Tabel 7. Distribusi Data Sebelum dan Sesudah *Upsampling*

Grade	Sebelum Upsampling	Sesudah Upsampling
A	124	4059
B	2164	4059
C	4059	4059
D	1607	4059
F	49	4059
Total	8003	20295

Berdasarkan Tabel 7, jumlah data latih meningkat dari 8.003 menjadi 20.295 observasi setelah dilakukan *upsampling*. Sebelum proses *upsampling*, distribusi kelas masih tidak seimbang dengan jumlah data kelas minoritas yang jauh lebih sedikit dibandingkan kelas mayoritas. Setelah proses *upsampling*, seluruh kelas memiliki jumlah data yang sama yaitu sebanyak 4.059 observasi. Hasil tersebut menunjukkan bahwa metode *upsampling* berhasil menghasilkan distribusi kelas yang lebih seimbang (*balanced data*). Dengan demikian, metode ini diharapkan menjadi lebih handal dalam mengenali kelas minoritas saat memasuki fase pemodelan tahap selanjutnya.

3.4 Pemodelan Random Forest

Pemodelan Random Forest dijalankan melalui dua skenario, yaitu tanpa menerapkan *balancing* dan dengan menerapkan *balancing* menggunakan metode *upsampling*. Model memanfaatkan algoritma Random Forest dengan menetapkan parameter pohon keputusan (*ntree*) sebanyak 500. Evaluasi model kemudian diukur menggunakan *accuracy*, *kappa*, *precision*, *recall*, dan *F1-score* dirangkum pada Tabel 8 dan Tabel 9.

Tabel 8. Hasil Evaluasi Random Forest

Metrik	Tanpa Balancing	Balancing
Accuracy	0,9089	0,9144
Kappa	0,8511	0,8622
F1-Score	0,6364	0,6957

Tabel 9. Hasil Evaluasi Random Forest Per Kelas

Metrik	Kelas A	Kelas B	Kelas C	Kelas D	Kelas F
Precision Tanpa Balancing	1,0000	0,9364	0,8851	0,9396	1,0000
Precision Balancing	1,0000	0,9256	0,9133	0,9010	0,7500

Recall Tanpa Balancing	0,4667	0,8722	0,9724	0,8529	0,1667
Recall Balancing	0,5333	0,8981	0,9448	0,9077	0,2500
F1-Score Tanpa Balancing	0,6364	0,9032	0,9267	0,8941	0,2857
F1-Score Balancing	0,6957	0,9117	0,9287	0,9043	0,3750

Berdasarkan Tabel 8, penerapan *upsampling* pada ketiga metrik yang diamati menunjukkan adanya peningkatan masing-masing nilai. Sebelum diterapkan *balancing*, model memberikan hasil yang baik pada kelas mayoritas pada kelas C dengan nilai *recall* 0,972 dan *F1-score* yaitu 0,9267. Sedangkan kemampuan model dalam mengenali kelas minoritas A dan F masih rendah. Namun setelah dilakukan *upsampling*, nilai mengalami peningkatan dari sebelumnya. Hasil ini menunjukkan *upsampling* membantu Random Forest dalam menangani kelas minoritas pada *imbalanced data.*

3.5 Pemodelan Support Vector Machine (SVM)

Implementasi algoritma ini dijalankan melalui dua skenario, yaitu tanpa menerapkan *balancing* dan dengan menerapkan *balancing* menggunakan metode *upsampling*. Sebelum proses pemodelan, data numerik dinormalisasi menggunakan metode *centering* dan *scaling*. Evaluasi model kemudian diukur menggunakan *accuracy*, *kappa*, *precision*, *recall*, dan *F1-score* dirangkum pada Tabel 10 dan Tabel 11.

Tabel 10. Hasil Evaluasi SVM

Metrik	Tanpa Balancing	Balancing
Accuracy	0,9589	0,9474
Kappa	0,9341	0,9177
F1-Score	0,8235	0,8169

Tabel 11. Hasil Evaluasi SVM Per Kelas

Metrik	Kelas A	Kelas B	Kelas C	Kelas D	Kelas F
Precision Tanpa Balancing	1,0000	0,9679	0,9576	0,9481	1,0000
Precision Balancing	0,7073	0,9453	0,9854	0,8991	0,6875
Recall Tanpa Balancing	0,7000	0,9500	0,9803	0,9576	0,2500
Recall Balancing	0,9667	0,9593	0,9290	0,9776	0,9167
F1-Score Tanpa Balancing	0,8235	0,9589	0,9688	0,9529	0,4000
F1-Score Balancing	0,8169	0,9522	0,9563	0,9367	0,7857

Berdasarkan Tabel 10, model SVM tanpa *balancing* memperoleh nilai *accuracy* sebesar 0,9589 dan *kappa* sebesar 0,9341. Setelah dilakukan *upsampling*, pada nilai *accuracy* dan nilai *kappa* sedikit menurun menjadi 0,9474 dan 0,9177, sedangkan nilai *F1-score* yaitu 0,8169. Hasil ini menunjukkan bahwa model memperlakukan data menjadi lebih seimbang baik kelas minoritas maupun pada kelas mayoritas. Hasil ini menunjukkan *upsampling* membantu SVM dalam menangani kelas minoritas pada *imbalanced data.*

3.6 Perbandingan Model

Perbandingan performa model dilakukan untuk mengidentifikasi algoritma yang menunjukkan efektivitas tertinggi dalam mengklasifikasikan tingkat capaian belajar siswa. Untuk mengukur performa, digunakan metrik yang merepresentasikan ketepatan seperti *accuracy* dan *kappa*, serta metrik yang berfokus pada keseimbangan kelas seperti *precision, recall*, dan *F1-score* pada empat skenario penelitian, yaitu Random Forest tanpa *balancing*, Random Forest dengan *balancing*, SVM tanpa *balancing*, dan SVM dengan *balancing* dirangkum pada Tabel 12 dan Tabel 13.

Tabel 12. Perbandingan Performa Model

Model	Accuracy	Kappa	F1-Score
Random Forest Tanpa *Balancing*	0,9089	0,8511	0,6364
Random Forest Dengan *Balancing*	0,9144	0,8622	0,6957
SVM Tanpa *Balancing*	0,9589	0,9341	0,8235
SVM Dengan *Balancing*	0,9474	0,9177	0,8169

Berdasarkan Tabel 12, penggunaan metrik *accuracy*, *kappa*, dan *F1-score* menunjukkan bahwa prosedur *balancing* mampu mengoptimalkan performa model dalam mengidentifikasi kelas minoritas secara lebih efektif. Walaupun penerapan *upsampling* pada metode SVM mengakibatkan sedikit penurunan pada nilai *accuracy*, *kappa*, dan *F1-score*, namun tetap menunjukkan kinerja klasifikasi yang tangguh, terutama dalam menangani kendala ketidakseimbangan data (*imbalanced data*). Secara umum, algoritma Support Vector Machine (SVM) memberikan keunggulan dalam hal hasil performa klasifikasi dari Random Forest.

4 KESIMPULAN

Penelitian ini menyimpulkan bahwa algoritma Support Vector Machine (SVM) serta pengembangan model berbasis Random Forest dapat digunakan untuk mengklasifikasikan tingkat capaian belajar siswa pada *Student Performance Analytics Dataset*. Berdasarkan penelitian dengan penerapan *upsampling* pada data, hasil memberikan dampak yang berbeda pada kedua metode yang digunakan. Pada Random forest, *upsampling* mampu meningkatkan nilai evaluasi model jika dibandingkan dengan data sebelum dilakukan *balancing.* Sementara itu, *upsampling* pada SVM membantu model dalam mengenali kelas minoritas dengan adanya nilai penururan pada nilai metrik *accuracy*, *kappa*, dan *F1-score.* Pada tabel yang disajikan oleh *confusion matrix,* hasil menunjukkan model SVM tanpa *balancing* memberikan nilai *accuracy* 95,89%, kappa 0,9341 dan *F1-score* 0,8235. Nilai tersebut merupakan hasil terbaik jika dibandingkan dengan model lainnya. Maka dapat disimpulkan, model SVM lebih efektif dalam mengklasifikasi capaian belajar siswa. Untuk itu, ada beberapa saran penulis pada penelitian berikutnya yaitu pada penelitian dengan kondisi ketidakseimbangan data dapat diterapkan metode lainnya seperti SMOTE dan ADASYN; pada pemodelan Random Forest dan Support Vector Machine (SVM) dapat diterapkan nilai parameter yang berbeda pula seperti *hyperparameter tuning;* selain itu dapat pula menerapkan metode seleksi fitur untuk mengidentifikasi variabel yang paling berpengaruh terhadap klasifikasi capaian belajar siswa.

UCAPAN TERIMAKASIH

Syukur dan terima kasih atas kemurahan hati Tuhan Yesus Kristus melalui Kongregasi Suster-Suster Cinta Kasih St Carolus Borromeus dan Yayasan Tarakanita, para suster, komunitas, keluarga, sahabat yang telah memberikan atensi dan doa-doa yang menyertai penulis selama proses penelitian.

DAFTAR PUSTAKA

Admassu, T., Salau, A., Chhabra, G., & Kaushik, K. (2024). Evaluation of random forest and support vector machine models in educational data mining. *Proceedings of the 2nd International Conference on Advancement in Computation & Computer Technologies*. https://doi.org/10.1109/InCACCT61598.2024.10551110

Adriansyah, I., Mahendra, M. D., Rasywir, E., & Pratama, Y. (2022). Perbandingan metode random forest classifier dan SVM pada klasifikasi kemampuan level beradaptasi pembelajaran jarak jauh siswa. *Bulletin of Informatics and Data Science*, 1(2), 98–103. https://ejurnal.pdsi.or.id/index.php/bids/article/view/49

Agung, G. M., Zuama, R. A., & Budi, E. S. (2026). Analysis of student academic performance using random forest and support vector machines. *Computer Science (CO-SCIENCE)*, 6(1), 57–65. https://doi.org/10.31294/co-science.v6i1.10123

Alamri, A., Alharbi, A., Almutiry, M., & Alotaibi, Y. (2020). Predicting student academic performance using support vector machine and random forest. In *Proceedings of the International Conference on Computing and Information Technology*. https://doi.org/10.1145/3446590.3446607

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of students' performance prediction using machine learning techniques. *Education Sciences*, 11(9), 552. https://doi.org/10.3390/educsci11090552

Arifuddin, N. A., Andriyani, W., Dahlan, A., & Insani, C. N. (2025). *Machine learning*. Lingkar Edukasi Indonesia.

Balcıoğlu, Y. S., & Artar, M. (2025). Predicting academic performance of students with machine learning. *Information Development*, 41(3), 896–915. https://doi.org/10.1177/02666669231213023

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Budiasto, J., Tallulembang, T. M., Pare, S., & Hasbi, M. (2025). *Machine learning untuk pemula: Konsep dan implementasi*. Penerbit Buku Indonesia.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Gusnina, M., Wiharto, W., & Salamah, U. (2022). Student performance prediction in Sebelas Maret University based on the random forest algorithm. *Ingénierie des Systèmes d'Information*, 27(3). https://doi.org/10.18280/isi.270317

Handayani, S., Devianto, Y., Niqotaini, Z., Rahmawati, R., & Putri, A. (2025). *Machine learning di dunia pendidikan*. PT Penamuda Media.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer.

Khairy, D., Alharbi, N., Amasha, M. A., Areed, M. F., Alkhalaf, S., & Abougalala, R. A. (2024). Prediction of student exam performance using data mining classification algorithms. *Education and Information Technologies*, 29, 21621–21645. https://doi.org/10.1007/s10639-024-12619-w

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence, 5*, 221–232. https://doi.org/10.1007/s13748-016-0094-0

Mustakim, M., Sari, W. J., & Ulfa, F. (2025). Enhancing student performance classification through dimensionality reduction and feature selection in machine learning. *Indonesian Journal of Artificial Intelligence and Data Mining*, 8(3), 717–727.

Novianto, E., Suhirman, & Prasetyo, D. (2024). Perbandingan metode klasifikasi random forest dan support vector machine dalam memprediksi capaian studi mahasiswa. *Jurnal Ilmiah Penelitian dan Pembelajaran Informatika*, 9(4), 1821–1833.

Ruliana, R., Rais, Z., & Sudirman, L. M. R. (2024). The support vector machine (SVM) and random forest methods for classification graduation rate. *ARRUS Journal of Engineering and Technology*, 4(2), 203–210. https://doi.org/10.35877/jetech3436

Sen, J. (2022). *Machine learning: Algorithms, models, and applications*. arXiv.

Syarip, A., Sugito, S., & Rahmawati, R. (2025). Comparison of random forest and support vector machine classification methods for predicting the accuracy level of madrasah data. *Media Statistika*, 18(1), 37–48. https://doi.org/10.14710/medstat.18.1.37-48

Tan, P.-N., Steinbach, M., & Kumar, V. (2019). *Introduction to data mining* (2nd ed.). Pearson.

Zhao, M., Zhang, H., & Zhou, W. (Eds.). (2024). *Application of machine learning and data mining*. MDPI.