

Analisis Performa Siswa Menggunakan K-Means dan Random Forest pada Ekstrakurikuler Robotik

Aldyansyah Tarigan^{1*}, Achmad Maezar Bayu Aji²,

^{1,2}Informatika, Universitas Bina Sarana Informatika, Indonesia

Email: ^{1*}aldytrgn@email.com, ²achmad.azb@bsi.ac.id

Abstrak

Penilaian performa siswa pada kegiatan ekstrakurikuler robotik umumnya masih dilakukan secara manual sehingga memerlukan waktu yang relatif lama dan berpotensi menimbulkan subjektivitas dalam proses evaluasi. Penelitian ini bertujuan menganalisis performa siswa menggunakan pendekatan hybrid yang mengombinasikan algoritma K-Means dan Random Forest. Dataset yang digunakan terdiri atas 1.369 data siswa dengan atribut *Practical Skill*, *Academic Skill*, *Life Skill*, dan persentase kehadiran. Tahapan penelitian meliputi transformasi data, *preprocessing*, proses *clustering* menggunakan algoritma K-Means dengan tiga cluster, kemudian hasil *clustering* digunakan sebagai label pada proses klasifikasi menggunakan algoritma Random Forest. Evaluasi model dilakukan menggunakan metode *Stratified 10-Fold Cross Validation* dengan metrik *Accuracy*, *Precision*, *Recall*, *F1-Score*, *Area Under the Curve* (AUC), dan *Matthews Correlation Coefficient* (MCC). Hasil penelitian menunjukkan bahwa algoritma K-Means berhasil mengelompokkan siswa ke dalam tiga kelompok performa dengan karakteristik yang berbeda. Model Random Forest memperoleh nilai *Accuracy* sebesar 89,8%, *Precision* sebesar 90,9%, *Recall* sebesar 89,8%, *F1-Score* sebesar 88,0%, AUC sebesar 98,1%, dan MCC sebesar 69,1%. Hasil tersebut menunjukkan bahwa kombinasi K-Means dan Random Forest mampu menghasilkan analisis performa siswa yang efektif serta dapat mendukung proses evaluasi dan pengambilan keputusan dalam pembinaan siswa pada kegiatan ekstrakurikuler robotik.

Kata Kunci: K-Means; Random Forest; Machine Learning; Analisis Performa Siswa; Ekstrakurikuler Robotik.

Abstract

Student performance assessment in robotics extracurricular activities is generally conducted manually, making the evaluation process time-consuming and potentially subjective. This study aims to analyze student performance using a hybrid approach that combines the K-Means and Random Forest algorithms. The dataset consisted of 1,369 student records with four main attributes: Practical Skill, Academic Skill, Life Skill, and attendance percentage. The research stages included data transformation, preprocessing, clustering using the K-Means algorithm with three clusters, and classification using the Random Forest algorithm, where the clustering results served as class labels. Model evaluation was performed using Stratified 10-Fold Cross Validation with Accuracy, Precision, Recall, F1-Score, Area Under the Curve (AUC), and Matthews Correlation Coefficient (MCC) as evaluation metrics. The results indicate that the K-Means algorithm successfully grouped students into three performance categories with distinct characteristics. The Random Forest model achieved an Accuracy of 89.8%, Precision of 90.9%, Recall of 89.8%, F1-Score of 88.0%, an AUC of 98.1%, and an MCC of 69.1%. These findings demonstrate that the integration of K-Means and Random Forest provides an effective approach for analyzing student performance and can support evaluation and decision-making processes in robotics extracurricular programs.

Keywords: K-Means; Random Forest; Machine Learning; Student Performance Analysis; Robotics Extracurricular.

*Corresponding Author:

Aldyansyah Tarigan, Universitas Bina Sarana Informatika, Indonesia

Email: aldytrgn@email.com

1. PENDAHULUAN

Perkembangan teknologi informasi telah mendorong pemanfaatan *machine learning* dalam berbagai bidang, termasuk sektor pendidikan[1]. Perkembangan *machine learning* dalam bidang pendidikan telah mendorong pemanfaatan teknik analisis data untuk mendukung evaluasi hasil belajar, identifikasi

karakteristik siswa, serta pengambilan keputusan yang lebih objektif dan berbasis data[2]. Analisis performa siswa menjadi aspek penting karena dapat memberikan informasi mengenai kemampuan akademik, keterampilan praktis, serta kemampuan sosial yang dimiliki siswa selama mengikuti proses pembelajaran.

Kegiatan ekstrakurikuler robotik merupakan salah satu media pembelajaran yang berperan dalam mengembangkan kemampuan siswa pada aspek *problem solving*, kreativitas, kerja sama tim, komunikasi, serta kemampuan berpikir logis. Proses penilaian dalam kegiatan tersebut umumnya dilakukan berdasarkan berbagai indikator, seperti *Design and Construct, Analysing and Evaluation, Problem Solving, Knowledge, Creativity and Imagination, Logical Thinking, Responsibility, Communication Skill, Teamwork and Leadership*, serta tingkat kehadiran. Banyaknya indikator penilaian menyebabkan proses evaluasi menjadi cukup kompleks sehingga membutuhkan waktu yang relatif lama dan berpotensi menimbulkan subjektivitas dalam menentukan kategori performa siswa. Oleh karena itu, diperlukan suatu pendekatan yang mampu mengolah data penilaian secara otomatis sehingga hasil evaluasi menjadi lebih efektif dan konsisten.

Teknik *data mining* merupakan salah satu pendekatan yang banyak digunakan untuk menemukan pola tersembunyi dari sekumpulan data[3]. Salah satu teknik yang banyak digunakan dalam analisis data adalah *clustering*, karena mampu mengelompokkan data berdasarkan karakteristik yang serupa sehingga memudahkan proses identifikasi pola performa siswa[4]. Dalam penelitian ini digunakan algoritma *K-Means* untuk mengelompokkan siswa berdasarkan tingkat kemiripan karakteristik performa[5], sedangkan algoritma *Random Forest* digunakan untuk membangun model klasifikasi berdasarkan hasil pengelompokan tersebut[6]. Kombinasi kedua algoritma diharapkan mampu menghasilkan proses analisis yang lebih komprehensif, dimana *K-Means* berperan dalam membentuk kelompok performa siswa, sedangkan *Random Forest* digunakan untuk memprediksi kategori performa dengan tingkat akurasi yang tinggi.

Beberapa penelitian terdahulu telah menerapkan algoritma *K-Means* maupun *Random Forest* pada bidang pendidikan. Penelitian oleh Romelyan dkk. menunjukkan bahwa algoritma *K-Means* efektif dalam mengelompokkan prestasi siswa berdasarkan karakteristik akademik[4]. Selanjutnya, Fatunnisa dan Marcos memanfaatkan algoritma *Random Forest* untuk memprediksi kelulusan siswa dengan hasil klasifikasi yang baik[7]. Selain itu, Ermatita dan Hafyz mengombinasikan algoritma *K-Means* dan *Random Forest* untuk memprediksi kelulusan mahasiswa[8]. Meskipun demikian, penerapan pendekatan *hybrid* untuk menganalisis performa siswa pada kegiatan ekstrakurikuler robotik masih terbatas. Oleh karena itu, penelitian ini mengusulkan kombinasi algoritma *K-Means* dan *Random Forest* untuk menghasilkan analisis performa siswa yang lebih komprehensif.

Meskipun berbagai penelitian telah menunjukkan bahwa algoritma *K-Means* maupun *Random Forest* memiliki performa yang baik pada bidang pendidikan, sebagian besar penelitian masih berfokus pada analisis data akademik formal, seperti prestasi belajar, prediksi kelulusan, maupun evaluasi hasil pembelajaran. Penelitian yang secara khusus menganalisis performa siswa pada kegiatan ekstrakurikuler robotik masih relatif terbatas. Selain itu, sebagian besar penelitian klasifikasi masih menggunakan label kelas yang ditentukan secara manual sehingga proses pelabelan bergantung pada penilaian ahli dan berpotensi menimbulkan subjektivitas.

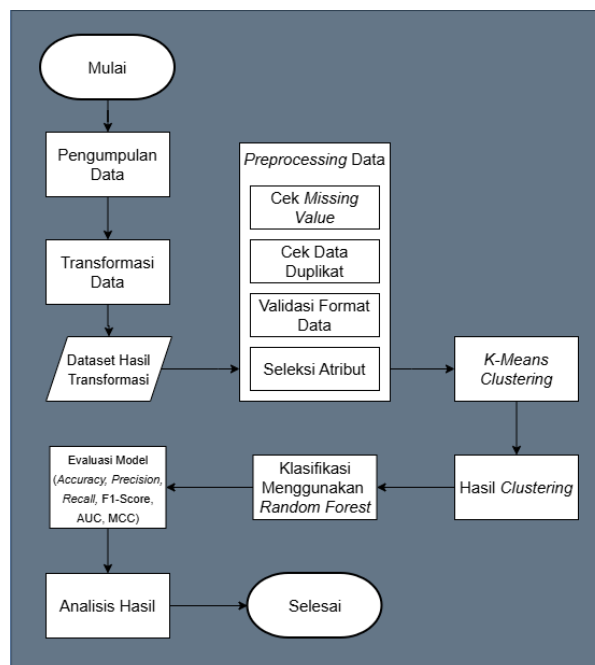
Berdasarkan permasalahan dan *research gap* tersebut, penelitian ini mengusulkan pendekatan *hybrid* yang mengombinasikan algoritma *K-Means* dan *Random Forest* untuk menganalisis performa siswa pada kegiatan ekstrakurikuler robotik. Berbeda dengan penelitian sebelumnya yang umumnya menggunakan label manual atau hanya menerapkan satu algoritma, penelitian ini memanfaatkan hasil *clustering K-Means* sebagai label pada proses klasifikasi menggunakan *Random Forest*. Model dievaluasi menggunakan metode *Stratified 10-Fold Cross Validation* dengan metrik *Accuracy, Precision, Recall, F1-Score, Area Under the Curve (AUC)*, dan *Matthews Correlation Coefficient (MCC)*. Pendekatan tersebut diharapkan mampu menghasilkan proses evaluasi yang lebih objektif, konsisten, serta memberikan model prediksi yang memiliki performa klasifikasi yang baik.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif untuk menganalisis performa siswa ekstrakurikuler robotik melalui kombinasi algoritma *K-Means* dan *Random Forest*. Tahapan penelitian meliputi pengumpulan data, *preprocessing*, proses *clustering* menggunakan algoritma *K-Means*, klasifikasi menggunakan algoritma *Random Forest*, serta evaluasi model menggunakan *Stratified 10-Fold Cross Validation*. Seluruh proses pengolahan data dilakukan menggunakan perangkat lunak *Orange Data Mining*.

2.1. Tahapan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan mengombinasikan algoritma *K-Means* dan *Random Forest* untuk menganalisis performa siswa ekstrakurikuler robotik. Tahapan penelitian meliputi proses pengumpulan data, transformasi data, *preprocessing*, *clustering*, klasifikasi, evaluasi model, dan analisis hasil.



Gambar 1. Tahapan Penelitian

Berdasarkan Gambar 1, penelitian diawali dengan proses pengumpulan data penilaian siswa ekstrakurikuler robotik dari TBOTICS Education. Data yang diperoleh kemudian ditransformasikan menjadi empat variabel utama, yaitu *Practical Skill*, *Academic Skill*, *Life Skill*, dan *Persentase Kehadiran*. Selanjutnya dilakukan tahap *preprocessing* yang meliputi pemeriksaan *missing value*, data duplikat, validasi format, dan seleksi atribut. Dataset hasil *preprocessing* kemudian dikelompokkan menggunakan algoritma *K-Means* untuk membentuk tiga kelompok performa siswa. Hasil *clustering* selanjutnya digunakan sebagai label kelas pada proses klasifikasi menggunakan algoritma *Random Forest*. Tahap terakhir adalah evaluasi model menggunakan *Stratified 10-Fold Cross Validation* dengan metrik *Accuracy*, *Precision*, *Recall*, *F1-Score*, *Area Under the Curve (AUC)*, dan *Matthews Correlation Coefficient (MCC)*, kemudian dilakukan analisis terhadap hasil penelitian.

2.2. Dataset

Dataset yang digunakan pada penelitian ini merupakan data penilaian siswa ekstrakurikuler robotik di TBOTICS Education sebanyak 1.369 data siswa. Data diperoleh dari hasil penilaian mentor terhadap kemampuan siswa berdasarkan beberapa indikator, yaitu *Design and Construct*, *Analysing and Evaluation*, *Problem Solving*, *Knowledge*, *Creativity and Imagination*, *Logical Thinking*, *Responsibility*,

Communication Skill, Teamwork and Leadership, serta persentase kehadiran. Selanjutnya, atribut-atribut tersebut ditransformasikan menjadi empat variabel utama, yaitu *Practical Skill, Academic Skill, Life Skill*, dan Persentase Kehadiran yang digunakan sebagai input pada proses analisis.

2.3. Preprocessing Data

Tahap *preprocessing* dilakukan untuk meningkatkan kualitas data sebelum proses analisis. Proses ini meliputi pemeriksaan nilai yang hilang (*missing value*), penghapusan data duplikat, serta pemilihan atribut yang relevan[9]. Setelah proses pembersihan data selesai, dilakukan transformasi seperti pada Tabel 1, beberapa indikator penilaian menjadi empat variabel utama agar data lebih representatif terhadap kemampuan siswa dan siap digunakan pada tahap *clustering* maupun klasifikasi. Tahap *preprocessing* merupakan proses yang sangat penting dalam *data mining* karena berperan dalam meningkatkan kualitas data sebelum dilakukan proses analisis. Tahapan ini bertujuan untuk mengurangi inkonsistensi data, memperbaiki nilai yang tidak sesuai, serta menghasilkan dataset yang lebih siap digunakan pada proses *clustering* maupun klasifikasi sehingga mampu meningkatkan performa model yang dibangun[10].

Tabel 1. Transformasi Atribut Penilaian Siswa

Nomor	Variabel Baru	Atribut Pembentuk
1	<i>Practical Skill</i>	<i>Design & Construct, Analysing & Evaluation, Problem Solving</i>
2	<i>Academic Skill</i>	<i>Knowledge, Creativity & Imagination, Logical Thinking</i>
3	<i>Life Skill</i>	<i>Responsibility, Communication Skill, Teamwork & Leadership</i>

2.4. Penentuan Jumlah Cluster

Penentuan jumlah *cluster* merupakan salah satu tahap penting dalam penerapan algoritma *K-Means* karena akan memengaruhi kualitas hasil pengelompokan. Pada penelitian ini, jumlah *cluster* ditetapkan sebanyak tiga ($k = 3$). Pemilihan nilai tersebut didasarkan pada pertimbangan kualitas hasil clustering serta kemudahan interpretasi terhadap karakteristik performa siswa. Tiga *cluster* merepresentasikan kategori performa rendah, sedang, dan tinggi sehingga hasil pengelompokan lebih mudah digunakan sebagai dasar evaluasi dalam kegiatan ekstrakurikuler robotik.

Untuk mengevaluasi kualitas hasil pengelompokan, digunakan *Silhouette Coefficient*. Nilai *silhouette* seluruh *cluster* bernilai positif, yaitu 0,406 pada *Cluster C1*, 0,223 pada *Cluster C2*, dan 0,601 pada *Cluster C3*. Nilai positif tersebut menunjukkan bahwa setiap *cluster* memiliki tingkat *cohesion* dan *separation* yang cukup baik, sehingga jumlah *cluster* sebanyak tiga dinilai mampu merepresentasikan struktur data dengan baik dan layak digunakan sebagai dasar pembentukan label pada proses klasifikasi menggunakan *Random Forest*. Meskipun nilai *silhouette* pada *Cluster C2* lebih rendah dibandingkan *cluster* lainnya, nilainya tetap berada pada rentang positif sehingga masih menunjukkan bahwa sebagian besar anggota *cluster* telah berada pada kelompok yang sesuai dengan karakteristiknya.

2.5. Clustering Menggunakan K-Means

Proses *clustering* dilakukan menggunakan algoritma *K-Means* untuk mengelompokkan siswa berdasarkan tingkat kemiripan karakteristik performa[5]. Setelah jumlah *cluster* ditentukan pada tahap sebelumnya, proses *clustering* dilakukan menggunakan algoritma *K-Means* dengan nilai $k = 3$ untuk mengelompokkan siswa berdasarkan tingkat kemiripan karakteristik performa. Tiga *cluster* tersebut merepresentasikan kategori performa rendah, sedang, dan tinggi sehingga hasil pengelompokan lebih mudah diinterpretasikan dan dimanfaatkan sebagai label pada proses klasifikasi menggunakan algoritma *Random Forest*. Algoritma *K-Means* dipilih karena mampu mengelompokkan data berdasarkan tingkat kemiripan karakteristik secara efisien serta banyak diterapkan pada berbagai penelitian di bidang pendidikan[11]. Algoritma bekerja dengan menghitung jarak setiap data terhadap centroid menggunakan metode *Euclidean Distance*, kemudian memperbarui posisi centroid secara

iteratif hingga diperoleh kelompok yang stabil. Hasil *clustering* selanjutnya digunakan sebagai label kelas pada proses klasifikasi menggunakan algoritma *Random Forest*.

2.6. Klasifikasi Menggunakan *Random Forest*

Tahap klasifikasi dilakukan menggunakan algoritma *Random Forest* dengan memanfaatkan hasil *clustering* sebagai label target. Algoritma membangun sejumlah *decision tree*[6] berdasarkan proses *bootstrap sampling* dan pemilihan atribut secara acak (*random feature selection*). Hasil prediksi diperoleh melalui mekanisme *majority voting* dari seluruh *decision tree* sehingga menghasilkan model klasifikasi yang lebih stabil dan memiliki kemampuan generalisasi yang baik. *Random Forest* telah banyak digunakan pada bidang pendidikan karena mampu menghasilkan performa klasifikasi yang tinggi pada data multikelas serta memiliki ketahanan terhadap *overfitting*[12].

2.7. Evaluasi Model

Evaluasi model dilakukan menggunakan metode *Stratified 10-Fold Cross Validation*. Kinerja model diukur menggunakan beberapa metrik evaluasi, yaitu *Accuracy*, *Precision*, *Recall*, *F1-Score*, *Area Under the Curve* (AUC), dan *Matthews Correlation Coefficient* (MCC). Penggunaan beberapa metrik tersebut bertujuan untuk memperoleh gambaran yang lebih komprehensif mengenai kemampuan model dalam melakukan klasifikasi performa siswa.

3. HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil penerapan algoritma *K-Means* dan *Random Forest* dalam menganalisis performa siswa ekstrakurikuler robotik di TBOTICS Education. Hasil penelitian meliputi proses *clustering* menggunakan algoritma *K-Means*, analisis karakteristik setiap cluster, evaluasi model klasifikasi menggunakan algoritma *Random Forest*, serta pembahasan terhadap hasil yang diperoleh. Evaluasi dilakukan untuk mengetahui kemampuan model dalam mengelompokkan dan mengklasifikasikan performa siswa berdasarkan variabel *Practical Skill*, *Academic Skill*, *Life Skill*, dan Persentase Kehadiran.

3.1. Hasil *Clustering* Menggunakan *K-Means*

Proses *clustering* dilakukan menggunakan algoritma *K-Means* dengan jumlah cluster sebanyak tiga ($k = 3$) berdasarkan empat variabel, yaitu *Practical Skill*, *Academic Skill*, *Life Skill*, dan *Attendance*. Hasil pengelompokan menghasilkan tiga cluster yang merepresentasikan kategori performa siswa dengan karakteristik yang berbeda. Distribusi data pada setiap cluster disajikan pada Tabel 2.

Tabel 2. Distribusi Data Hasil Clustering

Cluster	Jumlah Data	Persentase
C1	169	12,35%
C2	122	8,91%
C3	1078	78,74%
Total	1369	100%

Berdasarkan Tabel 2, *Cluster* C3 memiliki jumlah anggota terbanyak, yaitu 1.078 siswa (78,74%), sedangkan *Cluster* C2 memiliki jumlah anggota paling sedikit, yaitu 122 siswa (8,91%). Sementara itu, *Cluster* C1 terdiri atas 169 siswa (12,35%). Distribusi tersebut menunjukkan bahwa sebagian besar siswa berada pada kelompok dengan karakteristik performa sedang sehingga hasil *clustering* dapat digunakan sebagai dasar pada proses klasifikasi menggunakan algoritma *Random Forest*.

Distribusi data yang tidak seimbang antarcluster menunjukkan bahwa sebagian besar siswa berada pada kelompok dengan karakteristik performa sedang, sedangkan jumlah siswa pada kelompok performa tinggi dan rendah relatif lebih sedikit. Kondisi ini mengindikasikan bahwa karakteristik performa siswa di lingkungan ekstrakurikuler robotik cenderung terkonsentrasi pada tingkat performa menengah. Dari sisi pembelajaran mesin, distribusi yang tidak seimbang berpotensi menyebabkan model klasifikasi lebih banyak mempelajari pola dari cluster mayoritas. Oleh karena itu, penelitian ini

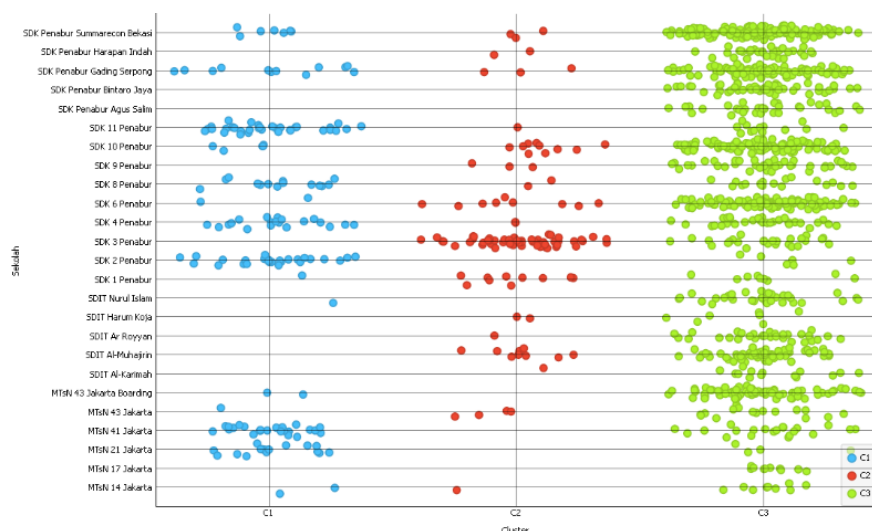
menggunakan metode *Stratified 10-Fold Cross Validation* sehingga proporsi setiap *cluster* tetap terjaga pada setiap *fold* pengujian. Pendekatan tersebut membantu mengurangi potensi bias akibat ketidakseimbangan distribusi data dan menghasilkan evaluasi model yang lebih representatif.

Tabel 3. Hasil Evaluasi Silhouette Coefficient

Cluster	Nilai Silhouette
C1	0.406
C2	0.223
C3	0.601

Kualitas hasil *clustering* dievaluasi menggunakan metrik *Silhouette Coefficient*. Berdasarkan Tabel 3, seluruh *cluster* memiliki nilai *silhouette* positif, yaitu 0,406 pada *Cluster* C1, 0,223 pada *Cluster* C2, dan 0,601 pada *Cluster* C3. Nilai tersebut menunjukkan bahwa proses pengelompokan telah menghasilkan pemisahan antarcluster yang cukup baik. *Cluster* C3 memperoleh nilai *silhouette* tertinggi, sehingga memiliki tingkat kekompakan (*cohesion*) dan pemisahan (*separation*) yang lebih baik dibandingkan dua *cluster* lainnya. Dengan demikian, hasil *clustering* dinilai layak untuk digunakan sebagai label pada proses klasifikasi menggunakan algoritma *Random Forest*. Nilai *Silhouette Coefficient* yang bernilai positif pada seluruh *cluster* menunjukkan bahwa hasil pengelompokan memiliki kualitas yang cukup baik. Semakin tinggi nilai *silhouette*, semakin baik tingkat kekompakan (*cohesion*) dalam suatu *cluster* serta pemisahan (*separation*) antarcluster, sehingga hasil *clustering* dapat digunakan sebagai dasar analisis pada tahap berikutnya [13].

Hasil evaluasi menggunakan *Silhouette Coefficient* menunjukkan bahwa penggunaan tiga *cluster* menghasilkan kualitas pengelompokan yang cukup baik. Seluruh *cluster* memiliki nilai *silhouette* positif sehingga menunjukkan bahwa sebagian besar data berada pada kelompok yang sesuai dengan karakteristiknya. *Cluster* C3 memperoleh nilai *silhouette* tertinggi (0,601), sedangkan *Cluster* C2 memiliki nilai *silhouette* terendah (0,223). Perbedaan tersebut menunjukkan bahwa tingkat kekompakan dan pemisahan antarcluster tidak sepenuhnya seragam, namun secara keseluruhan struktur pengelompokan telah memenuhi kriteria untuk digunakan sebagai dasar pembentukan label pada proses klasifikasi menggunakan *Random Forest*.



Gambar 2. Visualisasi Hasil Clustering Menggunakan Algoritma K-Means

Gambar 2 menunjukkan persebaran data hasil proses *clustering* menggunakan algoritma *K-Means*. Terlihat bahwa ketiga *cluster* memiliki distribusi yang berbeda berdasarkan kombinasi variabel yang digunakan. Meskipun masih terdapat beberapa data yang berada pada batas antarcluster, sebagian besar data telah membentuk kelompok yang relatif terpisah sehingga mendukung hasil evaluasi *Silhouette Coefficient* yang menunjukkan kualitas pengelompokan yang cukup baik.

3.2. Hasil Klasifikasi Menggunakan *Random Forest*

Hasil *clustering* yang diperoleh dari algoritma *K-Means* selanjutnya digunakan sebagai label kelas pada proses klasifikasi menggunakan algoritma *Random Forest*. Pengujian model dilakukan menggunakan metode *Stratified 10-Fold Cross Validation* untuk memperoleh evaluasi yang lebih objektif terhadap kemampuan model dalam mengklasifikasikan performa siswa. Hasil evaluasi model disajikan pada Gambar 3.

Test and Score

Thu Jun 18 26, 16:21:25

Settings

Sampling type: Stratified 10-fold Cross validation

Target class: None, show average over classes

Scores

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	0.981	0.898	0.880	0.909	0.898	0.691

Gambar 3. Hasil Evaluasi Model Random Forest

Pada Gambar 3, model *Random Forest* menghasilkan nilai *Accuracy* sebesar 89,8%, yang menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan benar. Selain itu, nilai *Precision* sebesar 90,9%, *Recall* sebesar 89,8%, dan *F1-Score* sebesar 88,0% menunjukkan bahwa model memiliki keseimbangan yang baik dalam mengidentifikasi setiap kategori performa siswa. Nilai *Area Under the Curve* (AUC) sebesar 98,1% mengindikasikan kemampuan model yang sangat baik dalam membedakan setiap kelas, sedangkan nilai *Matthews Correlation Coefficient* (MCC) sebesar 69,1% menunjukkan tingkat korelasi yang kuat antara hasil prediksi dan data aktual. Secara keseluruhan, hasil evaluasi tersebut menunjukkan bahwa algoritma *Random Forest* memiliki performa yang baik dalam mengklasifikasikan performa siswa ekstrakurikuler robotik.

		Predicted			Σ
		C1	C2	C3	
Actual	C1	122	0	47	169
	C2	0	33	89	122
	C3	3	0	1075	1078
	Σ	125	33	1211	1369

Gambar 4. Confusion Matrix Algoritma Random Forest

Gambar 4 menunjukkan hasil *Confusion Matrix* dari model *Random Forest*. Sebagian besar data pada setiap cluster berhasil diklasifikasikan dengan benar. Model mampu mengklasifikasikan 122 data pada Cluster C1, 33 data pada Cluster C2, dan 1.075 data pada Cluster C3 secara tepat. Meskipun masih terdapat sejumlah kecil kesalahan klasifikasi antarcluster, jumlahnya relatif rendah sehingga tidak memberikan pengaruh yang signifikan terhadap performa model secara keseluruhan. Hasil ini sejalan

dengan nilai *Accuracy* dan AUC yang tinggi, sehingga menunjukkan bahwa model memiliki kemampuan klasifikasi yang baik terhadap data performa siswa.

3.3. Pembahasan

Pada penelitian ini, algoritma *K-Means* berperan sebagai metode *unsupervised learning* yang digunakan untuk membentuk label kelas secara otomatis berdasarkan kemiripan karakteristik performa siswa. Label hasil *clustering* tersebut kemudian dimanfaatkan sebagai target pada proses klasifikasi menggunakan algoritma *Random Forest*. Dengan pendekatan ini, proses pelabelan tidak bergantung pada penilaian manual, tetapi dibentuk secara sistematis berdasarkan pola yang terdapat pada data. Pendekatan tersebut membantu mengurangi subjektivitas dalam proses evaluasi sekaligus menghasilkan kategori performa yang lebih konsisten.

Setelah memperoleh label dari proses *clustering*, algoritma *Random Forest* mempelajari hubungan antara atribut *Practical Skill*, *Academic Skill*, *Life Skill*, dan Persentase Kehadiran terhadap setiap kelompok performa. Model kemudian membangun sejumlah *decision tree* untuk mengenali pola yang membedakan masing-masing *cluster*. Dengan memanfaatkan mekanisme *ensemble learning*, *Random Forest* menghasilkan prediksi yang lebih stabil karena keputusan akhir diperoleh melalui mekanisme voting dari banyak *decision tree*. Pendekatan *ensemble* tersebut membantu mengurangi risiko *overfitting* yang umumnya terjadi pada penggunaan satu *decision tree* tunggal, sehingga model memiliki kemampuan generalisasi yang lebih baik.

Pendekatan *hybrid* tersebut tidak hanya menghasilkan pengelompokan siswa berdasarkan karakteristik performanya, tetapi juga membangun model klasifikasi yang dapat digunakan untuk memprediksi kategori performa siswa baru tanpa perlu melakukan proses *clustering* ulang. Dengan demikian, model yang dihasilkan memiliki potensi untuk diterapkan sebagai sistem pendukung evaluasi performa siswa pada kegiatan ekstrakurikuler robotik secara lebih efisien dan konsisten.

Hasil penelitian menunjukkan bahwa kombinasi algoritma *K-Means* dan *Random Forest* mampu memberikan analisis performa siswa yang efektif pada kegiatan ekstrakurikuler robotik. Algoritma *K-Means* berhasil mengelompokkan 1.369 data siswa ke dalam tiga *cluster* berdasarkan variabel *Practical Skill*, *Academic Skill*, *Life Skill*, dan *Attendance*. Selanjutnya, hasil *clustering* dimanfaatkan sebagai label kelas pada proses klasifikasi menggunakan algoritma *Random Forest*. Model yang dibangun menghasilkan nilai *Accuracy* sebesar 89,8%, *Precision* sebesar 90,9%, *Recall* sebesar 89,8%, *F1-Score* sebesar 88,0%, AUC sebesar 98,1%, dan MCC sebesar 69,1%.

Ketidakseimbangan jumlah anggota pada masing-masing *cluster* juga menjadi karakteristik penting dari dataset penelitian ini. Meskipun *Cluster C3* memiliki jumlah anggota yang jauh lebih besar dibandingkan *Cluster C1* dan *C2*, hasil evaluasi menunjukkan bahwa model *Random Forest* tetap mampu menghasilkan nilai *Accuracy* sebesar 89,8% dan AUC sebesar 98,1%. Hal ini menunjukkan bahwa penggunaan *Stratified 10-Fold Cross Validation* membantu menjaga proporsi data pada setiap proses pelatihan dan pengujian sehingga model tetap memiliki kemampuan generalisasi yang baik terhadap seluruh kategori performa siswa.

Hasil tersebut menunjukkan bahwa pendekatan *hybrid* mampu menggabungkan keunggulan algoritma *K-Means* dalam proses pengelompokan data dengan kemampuan algoritma *Random Forest* dalam melakukan klasifikasi. Melalui pendekatan ini, informasi mengenai karakteristik setiap kelompok siswa dapat dimanfaatkan untuk membangun model prediksi yang memiliki performa klasifikasi yang tinggi.

Hasil penelitian ini sejalan dengan penelitian Fatunnisa dan Marcos yang menunjukkan bahwa algoritma *Random Forest* memiliki performa yang baik dalam proses klasifikasi data pendidikan [7]. Selain itu, penelitian Rahmah Dila juga membuktikan bahwa algoritma *K-Means* efektif dalam mengelompokkan data berdasarkan karakteristik yang dimiliki siswa [14]. Kesamaan hasil tersebut menunjukkan bahwa kedua algoritma memiliki kemampuan yang baik ketika diterapkan secara terpisah pada proses pengelompokan maupun klasifikasi data pendidikan.

Berbeda dengan penelitian-penelitian sebelumnya, penelitian ini mengombinasikan algoritma *K-Means* dan *Random Forest* dalam satu alur analisis. Hasil *clustering* digunakan sebagai label pada proses klasifikasi sehingga model tidak hanya mampu mengidentifikasi karakteristik setiap kelompok siswa,

tetapi juga membangun model prediksi berdasarkan kelompok yang telah terbentuk. Pendekatan ini menghasilkan analisis yang lebih komprehensif dibandingkan penggunaan satu algoritma saja [15].

Pendekatan *hybrid* yang mengombinasikan K-Means dan Random Forest memberikan keunggulan karena mampu mengintegrasikan proses pengelompokan dan klasifikasi dalam satu alur analisis. K-Means berperan dalam membentuk label berdasarkan karakteristik data, sedangkan Random Forest memanfaatkan label tersebut untuk membangun model prediksi yang memiliki kemampuan generalisasi yang baik. Dengan demikian, penelitian ini tidak hanya menghasilkan informasi mengenai karakteristik setiap kelompok siswa, tetapi juga menyediakan model yang dapat dimanfaatkan untuk memprediksi performa siswa baru secara lebih objektif dan konsisten.

Dari sisi implementasi, hasil penelitian ini dapat dimanfaatkan oleh mentor maupun pengelola TBOTICS Education untuk mengidentifikasi kelompok siswa berdasarkan tingkat performanya. Informasi tersebut dapat digunakan sebagai dasar dalam menyusun strategi pembinaan, pemberian materi yang lebih sesuai dengan kemampuan siswa, serta melakukan evaluasi perkembangan belajar secara lebih objektif dan konsisten.

4. KESIMPULAN

Penelitian ini berhasil menerapkan pendekatan *hybrid* yang mengombinasikan algoritma K-Means dan Random Forest untuk menganalisis performa siswa pada kegiatan ekstrakurikuler robotik. Algoritma K-Means berhasil mengelompokkan 1.369 data siswa ke dalam tiga *cluster* berdasarkan atribut *Practical Skill*, *Academic Skill*, *Life Skill*, dan Persentase Kehadiran. Hasil *clustering* kemudian dimanfaatkan sebagai label pada proses klasifikasi menggunakan algoritma Random Forest. Model yang dihasilkan memperoleh nilai *Accuracy* sebesar 89,8%, *Precision* sebesar 90,9%, *Recall* sebesar 89,8%, *F1-Score* sebesar 88,0%, *Area Under Curve (AUC)* sebesar 98,1%, dan *Matthews Correlation Coefficient (MCC)* sebesar 69,1%, sehingga menunjukkan bahwa pendekatan yang digunakan mampu menghasilkan performa klasifikasi yang baik.

Kontribusi penelitian ini terletak pada penerapan pendekatan *hybrid* yang memanfaatkan hasil *clustering* K-Means sebagai label otomatis pada proses klasifikasi menggunakan Random Forest. Pendekatan ini memungkinkan proses evaluasi performa siswa dilakukan tanpa memerlukan pelabelan manual oleh evaluator, sehingga dapat mengurangi subjektivitas dalam penentuan kategori performa serta menghasilkan proses evaluasi yang lebih objektif, konsisten, dan berbasis data.

Dari sisi implementasi, model yang dihasilkan berpotensi dimanfaatkan sebagai sistem pendukung keputusan bagi mentor maupun pengelola kegiatan ekstrakurikuler robotik dalam mengidentifikasi tingkat performa siswa, menyusun strategi pembinaan yang sesuai dengan karakteristik setiap kelompok, serta melakukan evaluasi perkembangan belajar secara lebih efisien dan konsisten.

Penelitian selanjutnya disarankan untuk mengeksplorasi metode penentuan jumlah *cluster* sebagai pembanding terhadap *Silhouette Coefficient*, seperti *Elbow Method* atau *Davies-Bouldin Index*, serta mengevaluasi performa pendekatan *hybrid* ini menggunakan algoritma klasifikasi lain, seperti XGBoost, Support Vector Machine (SVM), atau LightGBM. Selain itu, pengujian pada dataset dari institusi pendidikan yang berbeda juga perlu dilakukan untuk mengetahui kemampuan generalisasi model pada konteks yang lebih luas.

REFERENCES

- [1] R. S. Nurhalizah, R. Ardianto, and P. Purwono, "Analisis Supervised dan Unsupervised Learning pada Machine Learning: Systematic Literature Review," *J. Ilmu Komput. dan Inform.*, vol. 4, no. 1, pp. 61–72, 2024, doi: 10.54082/jiki.168.
- [2] H. A. Maulana, N. Fitriyah, D. El Baradei, F. Ardiyanto, and M. Arifin, "Analisis Data Pendidikan Menggunakan Metode Data Mining dengan Algoritma Decision Tree dan K-Means Clustering," vol. 6, no. 1, 2026, doi: 10.58794/jekin.v6i2.2131.
- [3] E. Saputri, "Teknik dan aplikasi data mining di Indonesia: tinjauan literatur satu dekade (2015-2024)," *IT-Explore J. Penerapan Teknol. Inf. dan Komun.*, vol. 4, no. 2, pp. 138–149, 2025, doi: 10.24246/itexplore.v4i2.2025.pp138-149.
- [4] N. Romelyan, I. Kanedi, and A. Fikri Sallaby, "Penerapan Metode K-Means Untuk Menampilkan Tingkat Prestasi Siswa SDN 08 Kabupaten Lebong," *J. Technopreneursh. Inf. Syst.*, vol. 6, no. 3, pp. 59–66, 2023, doi:

- 10.36085/jtis.v6i3.5919.
- [5] N. A. Rahmadenti, R. A. Purnama, T. Alfian, and A. Sandi, "Analisis Pola Dan Tren Kualitas Udara DKI Jakarta 2022–2025 Menggunakan K-Means Clustering," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 3S1, pp. 2830–7062, 2025, [Online]. Available: <https://journal.eng.unila.ac.id/index.php/jitet/article/view/8141>
 - [6] M. Mahendra Alvanof and R. Kesuma Dinata, "Penerapan Algoritma Random Forest dalam Deteksi dan Klasifikasi Ransomware," *J. Elektron. dan Teknol. Inf.*, vol. 5, no. 2, pp. 2721–9380, 2024.
 - [7] A. Fatunnisa and H. Marcos, "Prediksi Kelulusan Tepat Waktu Siswa SMK Teknik Komputer Menggunakan Algoritma Random Forest Prediction of On-Time Graduation for Computer Engineering Vocational School Students Using the Random Forest Algorithm," *J. Manaj. Inform.*, vol. 14, no. April, pp. 101–111, 2024, [Online]. Available: <https://ojs.unikom.ac.id/index.php/jamika/article/view/12114>
 - [8] H. Ermita, "Prediksi Kelulusan Mahasiswa Menggunakan Metode K-Means dan Random Forest Fakultas Ilmu Komputer Universitas Sriwijaya , Indonesia Sales Prediction of Student Graduation using K-Means and Random Forest Methods," *J. Pendidik. dan Teknol. Indones. Vol.*, vol. 5, no. 1, pp. 209–214, 2025.
 - [9] A. Agung, A. Daniswara, I. Kadek, and D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," *J. Informatics Comput. Sci.*, vol. 05, no. Vol. 5 No. 01 (2023), pp. 97–100, 2023, [Online]. Available: <https://ejournal.unesa.ac.id/index.php/jinacs/article/view/55395>
 - [10] T. Gori, A. Sunyoto, and H. Al Fatta, "Preprocessing Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 1, pp. 215–224, 2024, doi: 10.25126/jtiik.20241118074.
 - [11] R. Awalia, "Penerapan Metode K-Means Clustering untuk Pengelompokan Prestasi Siswa menggunakan Orange Data Mining : Studi Kasus di MTs Muhammadiyah Tallo Makassar," *Mechatronics J. Prof. Entrep.*, vol. 6, no. 2, pp. 36–41, 2024.
 - [12] Y. Priantama and T. A. Yoga Siswa, "Optimasi Correlation-Based Feature Selection Untuk Perbaikan Akurasi Random Forest Classifier Dalam Prediksi Performa Akademik Mahasiswa," *JIKO (Jurnal Inform. dan Komputer)*, vol. 6, no. 2, p. 251, 2022, doi: 10.26798/jiko.v6i2.651.
 - [13] T. Rahmawati, Y. Wilandari, and P. Kartikasari, "Analisis Perbandingan Silhouette Coefficient Dan Metode Elbow Pada Pengelompokan Provinsi Di Indonesia Berdasarkan Indikator Ipm Dengan K-Medoids," *J. Gaussian*, vol. 13, no. 1, pp. 13–24, 2024, doi: 10.14710/j.gauss.13.1.13-24.
 - [14] R. Dila, S. Defit, and S. Arlis, "Analisis Algoritma K-Means Clustering Dalam Pengelompokan Prestasi Belajar Siswa Menengah Atas (SMA)," *Bull. Comput. Sci. Res.*, vol. 5, no. 5, pp. 1113–1119, 2025, doi: 10.47065/bulletincsr.v5i5.751.
 - [15] Ridwan, "IMPLEMENTASI DATA MINING MENGGUNAKAN ALGORITMA K-MEANS CLUSTERING DAN RANDOM FOREST UNTUK PREDIKSI STATUS KESEHATAN KARYAWAN," vol. 4, no. 1, pp. 1–10, 2025.