AICOMS
Applied Information Technology and Computer Science
e-ISSN: 2964-7703
Vol. 5 No. 1, Juni-2026, pp. 99-107

Komparasi Algoritma Ensemble Dan Linear Pada Klasifikasi Objek Astronomi SDSS17 Berbasis Visual Workflow

Muhammad Abdan Rafi[1], Hidayat[2*], Siti Ar-rachmi Ningrum[3], Febhi Aditiya[4]

[1] Fakultas TIK, Universitas Komputer Indonesia; abdan.10222024@mahasiswa.unikom.ac.id
[2] Fakultas Pascasarjana, Universitas Komputer Indonesia; hidayat@email.unikom.ac.id
[3] Fakultas TIK, Universitas Komputer Indonesia; siti.10222025@mahasiswa.unikom.ac.id
[4] Fakultas TIK, Universitas Komputer Indonesia; febhi.10222023@mahasiswa.unikom.ac.id
* Korespondensi: hidayat@email.unikom.ac.id

Info Artikel:
Dikirim: 09 April 2026
Direvisi: 20 Mei 2026
Diterima: 05 Juni 2026

Abstract: Modern astronomical surveys such as the Sloan Digital Sky Survey (SDSS) generate massive datasets that require efficient automated classification techniques. This study evaluates the performance of Random Forest, XGBoost, and Logistic Regression in classifying stars, galaxies, and quasars using the SDSS17 dataset. Unlike previous studies that primarily focused on classification accuracy, this research emphasizes model interpretability and data efficiency through a visual workflow approach implemented in Orange Data Mining. The experimental process includes data preprocessing, feature selection, and cross-validation on 100,000 samples. The results indicate that Random Forest achieved the best and most stable performance, reaching an accuracy of 97.7%. Learning curve analysis revealed model convergence when using approximately 60% of the training data, indicating optimal computational efficiency. Furthermore, feature importance analysis identified redshift as the most influential feature, which is scientifically consistent with Hubble's Law in distinguishing local and extragalactic objects. These findings demonstrate that ensemble-based models not only provide superior statistical performance but also offer meaningful astronomical explanations. The integration of performance evaluation and Explainable Artificial Intelligence (XAI) contributes to the development of transparent, interpretable, and reproducible astronomical classification systems.

Keywords: Machine Learning; Random Forest; SDSS17; Astronomical Classification; Redshift; Visual Workflow.

Intisari: Survei astronomi modern seperti *Sloan Digital Sky Survey* (SDSS) menghasilkan data masif yang memerlukan klasifikasi otomatis secara efisien. Penelitian ini mengevaluasi performa *Random Forest*, *XGBoost*, dan *Logistic Regression* dalam mengklasifikasikan bintang, galaksi, dan quasar menggunakan dataset SDSS17. Berbeda dengan penelitian sebelumnya yang fokus pada akurasi, studi ini menekankan pada interpretabilitas model dan efisiensi data menggunakan pendekatan *visual workflow* di Orange Data Mining. Tahapan eksperimen meliputi prapemrosesan data, seleksi fitur, dan validasi silang pada 100.000 sampel. Hasil penelitian menunjukkan bahwa *Random Forest* memberikan performa paling unggul dan stabil dengan akurasi mencapai 97,7%. Analisis *learning curve* mengungkapkan terjadinya konvergensi model sejak penggunaan 60% data latih, menandakan efisiensi komputasi yang optimal. Temuan krusial pada analisis *feature importance* mengidentifikasi *redshift* sebagai fitur paling dominan, yang secara saintifik selaras dengan Hukum Hubble dalam membedakan objek lokal dan ekstragalaksi. Penelitian ini membuktikan bahwa model berbasis *ensemble* tidak hanya unggul secara statistik, tetapi juga mampu memberikan penjelasan logis secara astronomis. Integrasi evaluasi performa dan aspek *Explainable AI* (XAI) ini

memberikan kontribusi penting bagi pengembangan sistem klasifikasi astronomi yang transparan dan reprodusibel.

Kata Kunci: machine learning; Random Forest; SDSS17; klasifikasi astronomi; redshift; visual workflow.

1. Pendahuluan

Perkembangan survei astronomi modern telah mengubah paradigma pengolahan data ilmiah dari skala kecil menjadi skala masif. Proyek *Sloan Digital Sky Survey* (SDSS) secara konsisten menghasilkan jutaan pengamatan fotometrik dan spektral setiap tahunnya, mencakup berbagai objek langit seperti bintang, galaksi, dan *Quasar*. Rilis terbaru SDSS17 menyediakan katalog observasi dengan resolusi tinggi dan cakupan luas yang menjadi rujukan utama dalam penelitian astronomi komputasional [1]. *Volume* data yang sangat besar tersebut menjadikan proses klasifikasi manual tidak lagi praktis, sehingga diperlukan sistem otomatis berbasis pembelajaran mesin untuk mengelompokkan objek secara cepat dan konsisten [2], [3].

Berbagai pendekatan *Machine Learning* (ML) telah diterapkan untuk klasifikasi objek astronomi. Model berbasis pohon keputusan seperti *Random Forest* dan *Gradient Boosting* banyak dilaporkan mampu menangkap pola *non-linear* pada data fotometrik dan spektral secara efektif [4], [5], [6]. Di sisi lain, model *linear* seperti *Logistic Regression* masih sering digunakan sebagai *baseline* karena kesederhanaan formulasi matematis dan interpretasi koefisiennya yang jelas [7], [8]. Namun demikian, sifat *linear* dari model tersebut sering kali tidak cukup untuk memisahkan kelas yang memiliki distribusi fitur kompleks. Sejumlah studi menunjukkan bahwa metode *ensemble* secara konsisten menghasilkan akurasi lebih tinggi dibandingkan pendekatan *linear* pada data SDSS [6], [5], [9]. Meskipun demikian, sebagian besar penelitian terdahulu cenderung berfokus pada peningkatan metrik performa seperti akurasi atau AUC semata. Aspek interpretabilitas model serta keterkaitannya dengan prinsip fisika astronomi jarang dianalisis secara mendalam. Padahal, dalam konteks ilmiah, keandalan model tidak hanya ditentukan oleh ketepatan prediksi, tetapi juga oleh kemampuan menjelaskan alasan di balik keputusan tersebut.

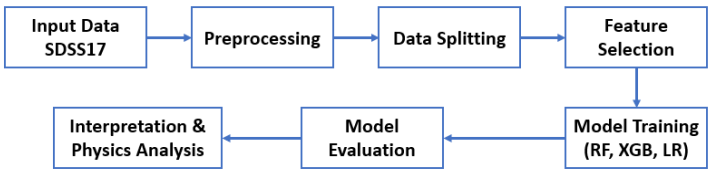
Konsep *Explainable Artificial Intelligence* (XAI) hadir untuk menjawab kebutuhan tersebut dengan menyediakan teknik analisis kontribusi fitur, seperti *feature importance* atau metode berbasis SHAP [10], [11], [12]. Pendekatan ini memungkinkan peneliti memahami apakah keputusan model selaras dengan teori fisika yang telah mapan. Dalam domain astronomi, keselarasan tersebut menjadi penting agar sistem prediktif tidak bersifat "*black box*", melainkan dapat diverifikasi secara konseptual. Selain itu, banyak implementasi model dilakukan melalui pemrograman khusus yang kompleks dan sulit direplikasi. Pendekatan berbasis *visual workflow* seperti *Orange Data Mining* menawarkan alternatif yang lebih transparan dan *reproducible*. Melalui antarmuka berbasis *widget*, seluruh tahapan eksperimen dapat divisualisasikan secara eksplisit tanpa pemrograman manual.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan mengevaluasi tiga algoritma klasifikasi, yaitu *Logistic Regression*, *Random Forest*, dan *XGBoost*, serta menganalisis interpretabilitasnya menggunakan dataset SDSS17. Penelitian ini berkontribusi melalui perbandingan komprehensif model *linear* dan *non-linear*, analisis pengaruh rasio data latih terhadap konvergensi model, serta interpretasi *feature importance* yang dikaitkan langsung dengan konsep fisika *redshift*. Meskipun berbagai penelitian telah menerapkan *Random Forest* atau *XGBoost* untuk klasifikasi objek SDSS, sebagian besar studi hanya melaporkan metrik akurasi tanpa mengevaluasi stabilitas model terhadap variasi jumlah data latih maupun mengaitkan hasil pembelajaran dengan interpretasi fisika astronomi. Selain itu, implementasi eksperimen umumnya berbasis pemrograman khusus yang menyulitkan reproduktibilitas. Hingga saat ini, masih terbatas penelitian yang secara sistematis menggabungkan evaluasi performa, analisis *learning curve*, serta interpretabilitas model dalam lingkungan *visual workflow* yang mudah direplikasi. Celah inilah yang menjadi fokus utama penelitian ini.

2. Metode

Bagian ini menjelaskan secara rinci rancangan eksperimen, karakteristik data, serta prosedur pengolahan dan evaluasi model yang diterapkan dalam penelitian. Seluruh proses dilakukan menggunakan pendekatan *visual workflow* pada perangkat lunak *Orange Data Mining*. Pendekatan ini dipilih untuk memastikan transparansi metodologi, kemudahan reproduksibilitas, serta menghindari ketergantungan pada pemrograman manual. Dengan memanfaatkan alur kerja berbasis *widget*, setiap tahapan eksperimen dapat divisualisasikan secara eksplisit sehingga mempermudah verifikasi dan validasi oleh peneliti lain.

Secara umum, penelitian ini mengikuti alur proses terstruktur yang dimulai dari memasukkan data SDSS17, *pre-processing*, *data splitting*, *feature selection*, *model training* (RF, XGB, LR), *model evaluation* dan *interpretation* serta *physics analysis*. Alur penelitian tersebut dapat ditampilkan pada Gambar 1.



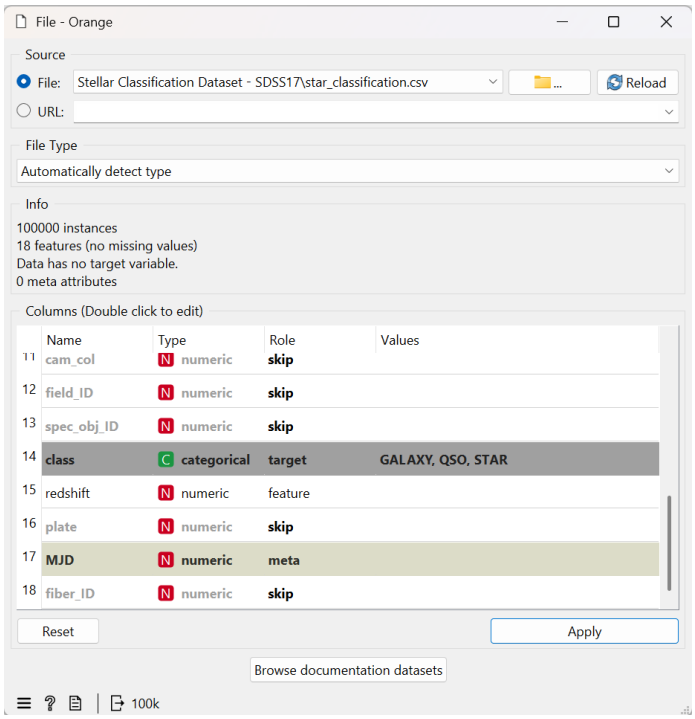
Gambar 1. Alur penelitian

Penelitian ini diawali dengan akuisisi data spektral SDSS17 yang kemudian diproses melalui tahapan *preprocessing* untuk pembersihan dan normalisasi, diikuti dengan pembagian data (*data partitioning*) menggunakan teknik *stratification* guna menjaga keseimbangan proporsi kelas. Selanjutnya, dilakukan seleksi fitur untuk mengidentifikasi variabel paling relevan sebelum diumpankan ke dalam tiga algoritma klasifikasi, yaitu *Random Forest*, *XGBoost*, dan *Logistic Regression*. Seluruh alur ini dijalankan menggunakan pendekatan *visual workflow* untuk memastikan proses yang transparan dan mudah direproduksi.

Pada tahap akhir, performa ketiga model diuji melalui metrik evaluasi komprehensif seperti akurasi dan *F1-score*, serta analisis *learning curve* untuk mengukur efisiensi data. Untuk memberikan validasi saintifik, dilakukan analisis interpretasi model melalui *Confusion Matrix* dan *Feature Importance*. Analisis ini bertujuan untuk menghubungkan fitur teknis seperti *redshift* dengan prinsip fisika astronomi, memastikan bahwa hasil klasifikasi tidak hanya unggul secara statistik tetapi juga memiliki dasar ilmiah yang kuat.

2.1 Dataset

Dataset yang digunakan berasal dari *Sloan Digital Sky Survey* Data Release 17 (SDSS17) [1]. Dataset ini terdiri dari 100.000 sampel dengan tiga kelas utama, yaitu *Star*, *Galaxy*, dan *Quasar*. Fitur-fitur yang tersedia mencakup pengukuran spektral dan fotometrik yang merepresentasikan karakteristik fisik objek. Data astronomi semacam ini umumnya memiliki distribusi *non-linear* serta korelasi kompleks antar fitur seperti pada penelitian [13], [14] sehingga memerlukan algoritma klasifikasi yang fleksibel. Untuk keperluan reproduksibilitas eksperimen, dataset yang digunakan dalam penelitian ini juga tersedia secara publik melalui *link website* berikut (https://www.kaggle.com/datasets/fedesoriano/stellar-classification-dataset-sdss17). Ringkasan karakteristik umum dataset SDSS17 yang digunakan dalam penelitian ini disajikan pada Gambar 2.



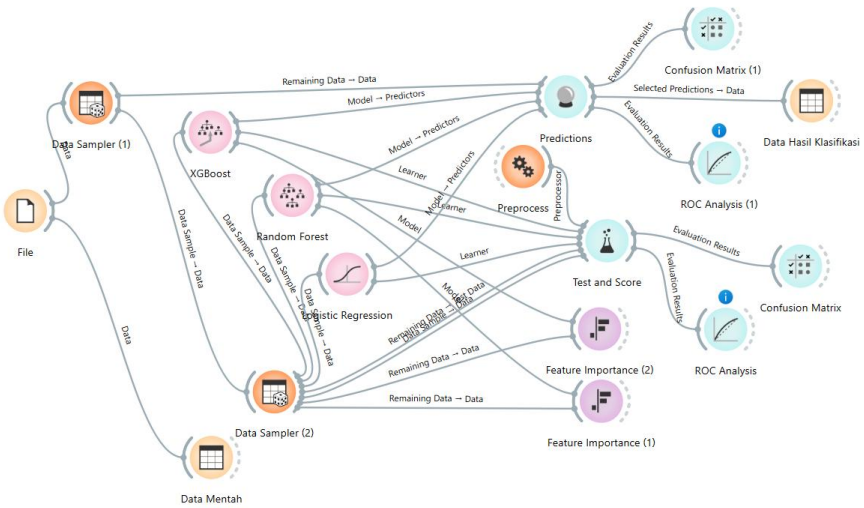
Gambar 2. Karakteristik umum dataset SDSS17

2.2 Lingkungan Eksperimen

Seluruh eksperimen dilakukan menggunakan perangkat lunak *Orange Data Mining*. *Orange* menyediakan antarmuka berbasis *widget* yang memungkinkan pengguna membangun *pipeline* analisis secara visual tanpa pemrograman manual. Pendekatan ini dipilih karena dua alasan utama. Pertama, *visual workflow* meningkatkan transparansi metodologi, sehingga alur eksperimen dapat dengan mudah direplikasi. Kedua, pendekatan ini menurunkan hambatan teknis bagi peneliti yang tidak memiliki latar belakang pemrograman intensif. Dalam penelitian ini, seluruh proses mulai dari impor data, prapemrosesan, pelatihan model, hingga evaluasi dilakukan melalui rangkaian *widget* yang saling terhubung.

2.3 Workflow Eksperimen

Pipeline eksperimen dimulai dari *widget File* untuk memuat dataset SDSS17. Data yang dimuat kemudian divisualisasikan menggunakan *Data Table* untuk memastikan struktur atribut telah sesuai. Selanjutnya, *widget Preprocess* digunakan untuk melakukan imputasi nilai hilang dengan rata-rata numerik serta normalisasi standar (mean = 0, standar deviasi = 1). Tahap berikutnya adalah seleksi atribut menggunakan *Select Columns* untuk memisahkan fitur relevan dan label kelas. Dataset yang telah dibersihkan kemudian dibagi menggunakan *Data Sampler* dengan teknik *stratified sampling* untuk menjaga proporsi kelas. Model klasifikasi dilatih dan dievaluasi menggunakan *Test & Score* yang menyediakan berbagai metrik kinerja secara otomatis. Hasil prediksi divisualisasikan melalui *Confusion matrix*, sedangkan kontribusi fitur dianalisis menggunakan *Feature Importance*. Seluruh tahapan tersebut membentuk *pipeline* terintegrasi yang konsisten dengan praktik eksperimen berbasis GUI. Alur lengkap proses eksperimen divisualisasikan melalui *workflow Orange Data Mining* sebagaimana ditunjukkan pada Gambar 3.



Gambar 3. *Workflow Orange*

2.4 Prapemrosesan Data

Tahap prapemrosesan mencakup tiga langkah utama. Pertama, imputasi nilai hilang dilakukan untuk menghindari bias akibat data yang tidak lengkap. Kedua, normalisasi standar diterapkan agar seluruh fitur berada pada skala yang sebanding, sehingga model *linear* maupun berbasis jarak tidak terdistorsi oleh perbedaan magnitudo. Ketiga, seleksi fitur bertujuan menghilangkan atribut teknis yang tidak relevan dengan sifat fisik objek. Praktik prapemrosesan seperti normalisasi dan seleksi fitur merupakan prosedur standar dalam pembelajaran mesin statistik modern [15].

2.5 Dasar Teoretis dan Karakteristik Model

Penelitian ini membandingkan tiga algoritma pembelajaran terawasi, yaitu *Logistic Regression*, *Random Forest*, dan *XGBoost*. Ketiga metode tersebut dipilih untuk merepresentasikan pendekatan *linear*, *ensemble* berbasis *bagging*, serta *ensemble* berbasis *boosting*, sehingga memungkinkan evaluasi komprehensif terhadap karakteristik model dengan tingkat kompleksitas yang berbeda. Pendekatan ini penting mengingat data astronomi bersifat *non-linear*, berdimensi tinggi, dan memiliki interaksi fitur yang kompleks [13], [14] .

Logistic Regression merupakan model klasifikasi *linear* yang memodelkan probabilitas keanggotaan suatu sampel terhadap kelas tertentu melalui kombinasi linier fitur yang ditransformasikan menggunakan fungsi logistik. Secara matematis, probabilitas prediksi dirumuskan pada rumus (1).

$$P(y|x) = \frac{1}{1+e^{-(w^T x+b)}} \quad (1)$$

Variabel w menyatakan bobot parameter dan b adalah bias. Transformasi sigmoid membatasi keluaran pada rentang nol hingga satu sehingga dapat diinterpretasikan sebagai probabilitas. Untuk kasus multikelas, fungsi ini diperluas menggunakan *softmax* sehingga menghasilkan distribusi probabilitas pada setiap kelas. Model ini unggul dari sisi kesederhanaan, efisiensi komputasi, serta interpretabilitas koefisien, namun batas keputusan yang dihasilkan bersifat *linear* sehingga kurang mampu menangkap hubungan *non-linear* kompleks pada data spektral [7], [8].

Random Forest merupakan metode *ensemble* berbasis teknik *bootstrap aggregation* atau *bagging* yang mengombinasikan banyak pohon keputusan independen [16]. Misalkan terdapat T pohon keputusan $h_t(x)$, maka prediksi akhir diperoleh melalui agregasi mayoritas atau rata-rata pada persamaan (2).

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \dots, h_T(x)\} \quad (2)$$

Proses pelatihan setiap pohon menggunakan subset data acak dan subset fitur acak sehingga menghasilkan model yang beragam. Strategi ini menurunkan varians model dan meningkatkan generalisasi. Secara intuitif, pohon keputusan membagi ruang fitur menjadi partisi-partisi *non-linear*, sehingga *Random Forest* mampu membentuk batas keputusan yang kompleks dan adaptif. Karakteristik tersebut menjadikannya sangat efektif untuk memodelkan distribusi objek astronomi yang heterogen dan tidak dapat dipisahkan secara *linear*.

XGBoost atau *Extreme Gradient Boosting* mengimplementasikan pendekatan *boosting* berbasis gradien dengan membangun model secara bertahap untuk meminimalkan fungsi objektif [17]. Prediksi model dinyatakan sebagai penjumlahan aditif sejumlah pohon menggunakan persamaan (3). Fungsi $f_k(x)$ merepresentasikan pohon keputusan ke-k. Fungsi objektif yang diminimalkan terdiri atas komponen kerugian prediksi dan regularisasi kompleksitas model pada persamaan (4). Pendekatan ini memungkinkan setiap pohon baru berfokus pada kesalahan residu model sebelumnya, sehingga pembelajaran menjadi lebih presisi. Penambahan regularisasi membantu mengendalikan *overfitting* sekaligus meningkatkan stabilitas. Kemampuan optimasi bertahap ini membuat *XGBoost* sangat efektif dalam menangkap interaksi fitur yang rumit dan pola *non-linear* pada data astronomi.

$$\hat{y} = \sum_{k=1}^{K} f_k(x) \quad (3)$$

$$\mathcal{L} = \sum l(y_i, \hat{y}_i) + \sum \Omega(f_k) \quad (4)$$

Berdasarkan karakteristik matematis tersebut, model berbasis *ensemble* seperti *Random Forest* dan *XGBoost* secara teoritis lebih sesuai untuk dataset SDSS dibandingkan model *linear* murni. Oleh karena itu, kedua metode tersebut diharapkan memberikan performa klasifikasi yang lebih tinggi serta batas keputusan yang lebih representatif terhadap struktur fisik objek langit.

2.6 Metode Evaluasi

Kinerja model dievaluasi menggunakan pendekatan *hold-out validation* melalui pembagian data latih dan data uji secara terpisah. Dataset terlebih dahulu dipisahkan menjadi data pelatihan dan pengujian menggunakan *Data Sampler* dengan teknik *stratified sampling* untuk menjaga proporsi kelas. Pada tahap akhir, model yang telah dilatih diuji pada data uji yang tidak pernah dilihat selama proses pelatihan (*unseen test data*), sehingga hasil evaluasi mencerminkan kemampuan generalisasi model secara realistis.

Pendekatan ini dipilih karena ukuran dataset yang besar (100.000 sampel) telah cukup representatif untuk membentuk distribusi pelatihan yang stabil, sehingga validasi silang tidak diperlukan dan justru meningkatkan beban komputasi tanpa manfaat signifikan. Evaluasi performa dihitung menggunakan metrik *Accuracy*, *Precision*, *Recall*, *F1-score*, dan AUC melalui *widget Test & Score*. Selain itu, *confusion matrix* digunakan untuk menganalisis pola kesalahan klasifikasi, sedangkan *feature importance* dimanfaatkan untuk memahami kontribusi masing-masing fitur terhadap keputusan model.

3. Hasil dan Pembahasan

Bagian ini menyajikan hasil eksperimen serta analisis mendalam terhadap performa model klasifikasi yang telah dibangun. Evaluasi tidak hanya difokuskan pada perbandingan metrik kuantitatif seperti akurasi dan AUC, tetapi juga mencakup analisis stabilitas model terhadap variasi jumlah data latih serta interpretasi kontribusi fitur terhadap keputusan prediksi. Pendekatan ini bertujuan untuk memperoleh pemahaman yang komprehensif, sehingga model tidak hanya dinilai berdasarkan ketepatan statistik, tetapi juga berdasarkan konsistensi ilmiah dan keterkaitannya dengan prinsip fisika astronomi.

Secara bertahap, pembahasan dimulai dari komparasi kinerja antar algoritma, dilanjutkan dengan analisis *learning curve* dan fenomena *diminishing returns*, evaluasi pola kesalahan melalui *confusion matrix*, hingga interpretasi *feature importance* sebagai bentuk implementasi *Explainable* AI. Dengan struktur ini, hasil penelitian diharapkan memberikan gambaran menyeluruh mengenai keunggulan, keterbatasan, serta makna ilmiah dari model yang dikembangkan.

3.1 Hasil Kinerja Model

Untuk mengevaluasi performa model secara komprehensif, pengujian dilakukan menggunakan tiga skenario pembagian data latih dan uji, yaitu 60:40, 70:30, dan 80:20. Pendekatan ini bertujuan menganalisis stabilitas model terhadap variasi jumlah data pelatihan sekaligus mengidentifikasi apakah penambahan data latih memberikan peningkatan kinerja yang signifikan. Seluruh metrik evaluasi dihitung menggunakan modul *Test & Score* pada *Orange Data Mining* dan dirangkum pada Tabel 1.

Tabel 1. Perbandingan Hasil Kinerja Ketiga Model pada Split Data yang Berbeda

Rasio Split	Model	AUC	CA	F1	Prec	Recall	MCC
60:40	*XGBoost*	**0.995**	0.975	0.974	0.974	0.975	0.955
	Random Forest	**0.995**	**0.977**	**0.977**	**0.977**	**0.977**	**0.960**
	Logistic Regression	0.986	0.950	0.950	0.951	0.950	0.913
70:30	*XGBoost*	**0.995**	0.975	0.975	0.975	0.975	0.956
	Random Forest	**0.995**	**0.977**	**0.977**	**0.977**	**0.977**	**0.960**
	Logistic Regression	0.986	0.951	0.951	0.951	0.951	0.914
80:20	*XGBoost*	**0.996**	0.976	0.975	0.975	0.976	0.957
	Random Forest	0.995	**0.977**	**0.977**	**0.977**	**0.977**	**0.960**
	Logistic Regression	0.986	0.952	0.951	0.952	0.952	0.915

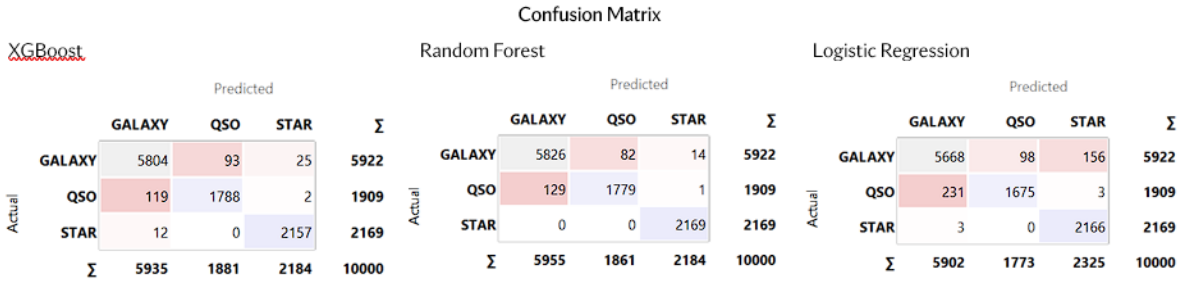
Berdasarkan hasil pada Tabel 1, terlihat bahwa model berbasis pohon keputusan secara konsisten mengungguli pendekatan *linear* pada seluruh skenario. *Random Forest* mencapai akurasi tertinggi dan paling stabil, yaitu sekitar 97,7% pada ketiga rasio pembagian data, diikuti oleh *XGBoost* dengan performa yang sangat mendekati. Sebaliknya, *Logistic Regression* menunjukkan performa yang lebih rendah dengan selisih sekitar dua hingga tiga persen. Dominasi metode *ensemble* ini sejalan dengan berbagai penelitian terdahulu yang melaporkan bahwa pendekatan berbasis pohon lebih efektif dalam menangani distribusi *non-linear* dan interaksi fitur kompleks pada data astronomi [2], [6], [18], [19].

Keunggulan tersebut dapat dijelaskan secara teoretis karena *Random Forest* dan *XGBoost* membentuk batas keputusan *non-linear* melalui kombinasi banyak pohon keputusan, sedangkan *Logistic Regression* hanya menghasilkan batas *linear* sederhana. Pada dataset spektral seperti SDSS yang memiliki korelasi fitur tinggi dan struktur distribusi heterogen [13], [14], fleksibilitas *non-linear* menjadi faktor kunci dalam meningkatkan akurasi prediksi. Selain itu, sifat agregasi pada *Random Forest* diketahui mampu menurunkan varians model dan meningkatkan stabilitas generalisasi [8].

Analisis terhadap variasi rasio pelatihan menunjukkan bahwa peningkatan proporsi data latih dari 60% menjadi 80% tidak menghasilkan kenaikan performa yang berarti. Perubahan nilai akurasi, *F1-score*, maupun AUC hanya berada pada kisaran sepersepuluh persen. Pola ini mengindikasikan adanya fenomena *diminishing returns*, yaitu kondisi ketika penambahan data pelatihan tidak lagi memberikan kontribusi signifikan terhadap peningkatan kemampuan prediksi. Fenomena serupa juga dilaporkan pada studi klasifikasi [6] dan [5], [9] astronomi berskala besar, di mana jumlah data yang sangat besar menyebabkan model telah mencapai konvergensi lebih awal [20], [21]. Dengan demikian, penggunaan data latih yang berlebihan hanya meningkatkan beban komputasi tanpa manfaat performa yang sepadan.

3.2 Evaluasi Kesalahan Klasifikasi

Distribusi kesalahan klasifikasi masing-masing model divisualisasikan melalui *confusion matrix* pada Gambar 4. *Confusion matrix* menunjukkan bahwa kesalahan prediksi terutama terjadi antara kelas *Galaxy* dan *Quasar*. Overlap karakteristik spektral kedua kelas tersebut telah dilaporkan dalam beberapa penelitian sebelumnya, seperti [22] dan [19] sehingga wajar jika model mengalami ambiguitas pada wilayah batas keputusan.

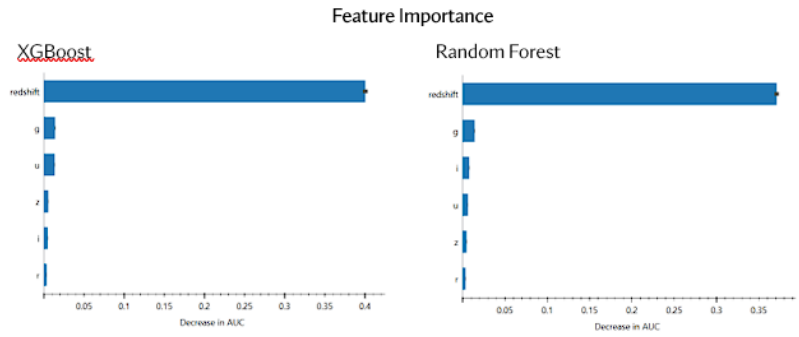


Gambar 4. Perbandingan hasil Confusion matrix pada ketiga model pada data uji

3.3 Interpretability dan Scientific Insight melalui Feature Importance

Kontribusi relatif setiap fitur terhadap keputusan model berbasis pohon (*tree-based*) dianalisis melalui mekanisme *feature importance*, sebagaimana ditunjukkan pada Gambar 5. Hasil analisis mengungkapkan bahwa **redshift** merupakan fitur paling dominan pada seluruh model *ensemble*. Dominasi fitur ini tidak hanya relevan secara statistik, tetapi juga memiliki landasan ilmiah yang kuat dalam bidang astronomi. Dalam kosmologi, *redshift* (pergeseran merah) menggambarkan perubahan panjang gelombang cahaya akibat ekspansi alam semesta. Berdasarkan Hukum Hubble, kecepatan resesi suatu objek langit berbanding lurus dengan jaraknya dari pengamat ($v = H_0 \cdot d$). Bintang yang terletak di dalam galaksi Bima Sakti (objek lokal) memiliki nilai *redshift* yang mendekati nol. Sebaliknya, galaksi dan quasar yang berada pada jarak ekstragalaksi menunjukkan nilai *redshift* yang signifikan [12]. Oleh karena itu, secara alami *redshift* menjadi indikator fundamental dalam membedakan objek lokal dan luar galaksi. Fakta bahwa model *machine learning* secara otomatis menempatkan *redshift* sebagai fitur terpenting membuktikan bahwa algoritma tersebut mampu menangkap prinsip fisika yang nyata, bukan sekadar korelasi data secara acak.

Interpretabilitas ini sejalan dengan prinsip *Explainable AI* (XAI), yang menekankan pentingnya transparansi keputusan model dalam aplikasi ilmiah [10], [11], [12], [21]. Model yang bersifat *explainable* memberikan tingkat kepercayaan (*trust*) yang lebih tinggi dan mempermudah validasi konseptual oleh para peneliti. Capaian akurasi sebesar 97,7% dalam penelitian ini sebanding dengan studi mutakhir berbasis *ensemble learning* pada data SDSS lainnya [4], [6], [9]. Hasil tersebut menegaskan bahwa pendekatan *visual workflow* mampu menghasilkan performa yang kompetitif dibandingkan implementasi berbasis kode manual (*hand-coded*), sekaligus menawarkan keunggulan dalam aspek transparansi dan reproduktibilitas eksperimen.



Gambar 5. Perbandingan Feature importance pada dua model tree-based

4. Kesimpulan

Penelitian ini berhasil mengevaluasi performa *Logistic Regression*, *Random Forest*, dan *XGBoost* dalam klasifikasi objek astronomi menggunakan data spektral SDSS17. Hasil eksperimen menunjukkan bahwa model berbasis *ensemble* secara konsisten mengungguli model linear, dengan *Random Forest* sebagai algoritma paling robust yang mencapai

akurasi stabil sebesar 97,7%. Analisis *learning curve* mengindikasikan bahwa konvergensi model telah tercapai pada penggunaan 60% data latih, sehingga penambahan volume data setelah titik tersebut tidak memberikan peningkatan performa yang signifikan. Hal ini menunjukkan efisiensi komputasi yang optimal dalam penggunaan dataset masif untuk klasifikasi bintang, galaksi, dan quasar.

Lebih lanjut, integrasi aspek *Explainable AI* (XAI) mengungkap bahwa *redshift* merupakan fitur paling dominan, sebuah temuan yang secara saintifik selaras dengan prinsip Hukum Hubble mengenai jarak kosmologis. Kesesuaian antara hasil model dan teori fisika ini menegaskan bahwa sistem tidak hanya akurat secara statistik, tetapi juga memiliki interpretasi ilmiah yang valid. Secara metodologis, penggunaan *visual workflow* terbukti efektif dalam menghasilkan eksperimen yang transparan, kompetitif, dan mudah direplikasi (*reproducible*). Dengan demikian, penelitian ini memberikan kontribusi praktis sebagai referensi pengembangan sistem klasifikasi astronomi yang tidak hanya berperforma tinggi, namun juga dapat dijelaskan secara teoritis.

Daftar Pustaka

[1] Abdurro'uf *et al.*, "The Seventeenth Data Release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar, and APOGEE-2 Data," *Astrophys. J. Suppl. Ser.*, vol. 259, no. 2, p. 35, 2022, doi: 10.3847/1538-4365/ac4414.

[2] O. J. Pérez Cruz, C. A. Martínez Pinto, S. G. Navarro Jiménez, L. J. Corral Escobedo, and M. M. Outeiro, "Analyzing Supervised Machine Learning Models for Classifying Astronomical Objects Using Gaia DR3 Spectral Features," *Appl. Sci.*, vol. 14, no. 19, 2024, doi: 10.3390/app14199058.

[3] J. V. Rodríguez, I. Rodríguez-Rodríguez, and W. L. Woo, "On the application of machine learning in astronomy and astrophysics: A text-mining-based scientometric analysis," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 12, no. 5, pp. 1–32, 2022, doi: 10.1002/widm.1476.

[4] J. L. Solorio-Ramírez, R. Jiménez-Cruz, Y. Villuendas-Rey, and C. Yáñez-Márquez, "Random forest Algorithm for the Classification of Spectral Data of Astronomical Objects," *Algorithms*, vol. 16, no. 6, 2023, doi: 10.3390/a16060293.

[5] B. Arroquia-Cuadros, N. Sánchez, V. Gómez, P. Blay, V. Martinez-Badenes, and L. Nieves-Seoane, "Photometric classification of quasars from Alhamra survey using random forest," *Astron. Astrophys.*, vol. 673, no. 45531, pp. 1–7, 2023, doi: 10.1051/0004-6361/202245531.

[6] F. Z. Zeraatgari *et al.*, "Machine learning-based photometric classification of galaxies , quasars , emission-line galaxies , and stars," *Mon. Not. R. Astron. Soc.*, vol. 4689, no. v, pp. 4677–4689, 2024.

[7] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. John Wiley & Sons, Inc., Hoboken, New Jersey, 2013.

[8] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. 2013.

[9] M. Wierzbiński, P. Pławiak, M. Hammad, and U. R. Acharya, "Development of accurate classification of heavenly bodies using novel machine learning techniques," *Soft Comput.*, vol. 25, no. 10, pp. 7213–7228, 2021, doi: 10.1007/s00500-021-05687-4.

[10] C. Molnar, "Interpretable machine learning: A guide for making black box models explainable . christophm. github. io/interpretable-ml-book," *Samek, W., Montavon, G., Lapuschkin, S., Anders, CJ, Müller, KR (2021). Explain. Deep neural networks beyond A Rev. methods Appl. Proc. IEEE*, vol. 109, no. 3, pp. 247–278, 2022.

[11] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Section 2, pp. 4766–4775, 2017.

[12] R. Dwivedi *et al.*, "Explainable AI (XAI): Core Ideas, Techniques, and Solutions," *ACM Comput. Surv.*, vol. 55, no. 9, 2023, doi: 10.1145/3561048.

[13] L. Rimoldini *et al.*, "Gaia Data Release 3: All-sky classification of 12.4 million variable sources into 25 classes," *Astron. Astrophys.*, vol. 674, no. 2018, 2023, doi: 10.1051/0004-6361/202245591.

[14] G. Li, Z. Lu, J. Wang, and Z. Wang, "Machine Learning in Stellar Astronomy: Progress up to 2024," no. Ml, pp. 1–14, 2025.

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. Springer, 2009.

[16] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010950718922.

[17] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-Augu, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[18] Y. W. Wu, "Machine Learning Classification of Stars, Galaxies, and Quasars," *MATTER Int. J. Sci. Technol.*, vol. 6, no. 3, pp. 102–122, 2021, doi: 10.20319/mijst.2021.63.102122.

[19] M. A. H. Bin Muhamad Suwaid, M. 'Ilyas A. Ab Karim, R. Hassan, and A. Abdul Aziz, "Automated Classification of Celestial Objects Using Machine Learning," *Int. J. Perceptive Cogn. Comput.*, vol. 11, no. 2, pp. 22–41, 2025, doi: 10.31436/ijpcc.v11i2.537.

[20] E. Scornet, "Trees, forests, and impurity-based variable importance in regression," *Ann. l'institut Henri Poincare Probab. Stat.*, vol. 59, no. 1, pp. 21–52, 2023, doi: 10.1214/21-AIHP1240.

[21] H. I. Aysel, X. Cai, and A. Prugel-Bennett, "Explainable Artificial Intelligence: Advancements and Limitations," *Appl. Sci.*, vol. 15, no. 13, pp. 1–26, 2025, doi: 10.3390/app15137261.

[22] P. A. C. Cunha and A. Humphrey, "Photometric redshift-aided classification using ensemble learning," *Astron. Astrophys.*, vol. 666, pp. 1–10, 2022, doi: 10.1051/0004-6361/202243135.