

Analisis Faktor Dominan dalam Klasifikasi Objek Astronomi Berbasis Random Forest

Febriani Yolanda Tesselonika^{1*}, Siti Nur Izzati², Najwa Sofia³, Alfian Rizaldy⁴

^{1,2,3}Sains Data, Universitas Pembangunan Nasional “Veteran” Jawa Timur , Indonesia

Dikirimkan: 18Juni 2026
Diterbitkan: 10 Juli 2026

Keywords:

data mining;
random forest;
klasifikasi objek astronomi;
keyword3.

E-mail Penulis

korespondensi:

Febrianiyolanda63@gmail.co
m

Abstrak. Perkembangan survei astronomi modern menghasilkan data observasi dalam jumlah sangat besar sehingga diperlukan teknik data mining yang mampu mengolah dan mengekstraksi informasi penting secara efektif. Salah satu permasalahan yang dihadapi adalah identifikasi faktor-faktor yang paling berpengaruh dalam proses klasifikasi objek astronomi yang terdiri atas Galaxy, QSO, dan Star. Penelitian ini bertujuan untuk menganalisis faktor dominan dalam klasifikasi objek astronomi menggunakan algoritma Random Forest melalui pendekatan Feature Importance dan Permutation Importance. Dataset yang digunakan merupakan Star Classification Dataset yang terdiri atas 100.000 data observasi dengan 18 atribut, kemudian melalui tahapan preprocessing, seleksi fitur menggunakan Mutual Information, pembangunan model Random Forest, evaluasi menggunakan accuracy, precision, recall, F1-score, confusion matrix, serta pengujian robustness model. Hasil penelitian menunjukkan bahwa model Random Forest mampu menghasilkan accuracy sebesar 97,995%, precision sebesar 97,985%, dan recall sebesar 97,995%, yang menunjukkan kemampuan klasifikasi yang sangat baik terhadap objek Galaxy, QSO, dan Star. Analisis Feature Importance dan Permutation Importance secara konsisten mengidentifikasi atribut redshift sebagai faktor yang paling dominan, diikuti oleh atribut z, g, i, dan u, sehingga menunjukkan bahwa karakteristik spektral memiliki kontribusi terbesar dalam proses klasifikasi. Selain memiliki performa yang tinggi, model juga menunjukkan kestabilan pada berbagai ukuran data latih melalui analisis robustness. Temuan penelitian ini memberikan informasi yang lebih interpretatif mengenai faktor-faktor yang memengaruhi klasifikasi objek astronomi dan dapat menjadi dasar pengembangan sistem analisis data astronomi berbasis data mining yang lebih efektif dan akurat.

Abstract. The rapid growth of modern astronomical surveys has generated massive observational datasets, creating a need for data mining techniques capable of extracting meaningful information efficiently. One of the major challenges is identifying the most influential factors in the classification of astronomical objects, including galaxies, quasars (QSOs), and stars. This study aims to analyze the dominant factors affecting astronomical object classification using the Random Forest algorithm through Feature Importance and Permutation Importance approaches. The study utilizes the Star Classification Dataset consisting of 100,000 observations and 18 attributes. The research methodology includes data preprocessing, feature selection using Mutual Information, Random Forest model construction, performance evaluation using accuracy, precision, recall, F1-score, and confusion matrix, followed by robustness analysis to assess model stability. The experimental results demonstrate that the Random Forest model achieves an accuracy of 97.995%, a precision of 97.985%, and a recall of 97.995%, indicating excellent classification performance for Galaxy, QSO, and Star objects. Furthermore, Feature Importance and Permutation Importance

analyses consistently identify redshift as the most influential attribute, followed by the z, g, i, and u spectral attributes, highlighting the significant contribution of spectral characteristics to the classification process. In addition, the robustness analysis confirms that the proposed model maintains stable performance across different training data sizes. These findings provide a more interpretable understanding of the factors influencing astronomical object classification and demonstrate the effectiveness of Random Forest as a data mining technique for supporting large-scale astronomical data analysis.

1. Pendahuluan

Otomatisasi klasifikasi objek langit (bintang, galaksi, dan kuasar) menjadi instrumen krusial dalam astronomi modern untuk memproses data fotometrik dan spektroskopi bervolume masif seperti Sloan Digital Sky Survey (SDSS). Pengolahan data secara manual oleh para astronom sudah tidak efisien dan tidak mungkin lagi dilakukan karena pesatnya pertumbuhan data observasi [1]. Oleh karena itu, penerapan algoritma pembelajaran mesin menjadi solusi utama untuk mempercepat klasifikasi objek fisis di alam semesta tersebut.

Sejumlah penelitian terdahulu telah menerapkan berbagai algoritma untuk klasifikasi ini. Model *decision tree* mampu mengungguli *multinomial logistic regression* dengan akurasi mencapai 99% [2]. Teknik ensemble berupa voting classifier yang menggabungkan 15 model pada data SDSS Data Release 17 (DR17) mampu mencapai akurasi 98,36% [1]. Pendekatan multi-modal deep learning yang menggabungkan data citra dan tabular digunakan untuk mengatasi ketidakseimbangan kelas (imbalanced dataset) [3]. Selain itu, evaluasi komparatif algoritma seperti Random Forest, Gradient Boosting, dan Support Vector Machine (SVM) terbukti efektif memetakan signifikansi fitur pada dataset SDSS17 [4].

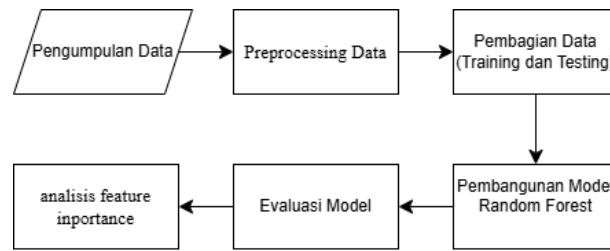
Eksplorasi pengklasifikasian objek langit berskala besar ini juga diperluas menggunakan algoritma Random Forest berbasis data fotometrik filter pita lebar dan sempit (*narrow- and broad-band filters*) [5], variasi algoritma *ensemble* pohon keputusan untuk memproses data spektral transien secara efisien [6], hingga perbandingan performa 10 algoritma pembelajaran mesin pada skenario ketiadaan data parameter tertentu [7], serta klasifikasi fotometrik tingkat lanjut untuk memisahkan galaksi garis emisi (*emission-line galaxies*) [8].

Meskipun model-model di atas menghasilkan akurasi yang tinggi, sebagian besar studi berfokus pada optimasi performa prediksi global tanpa memberikan interpretabilitas yang mendalam terkait kontribusi fisis dari setiap fitur parameter spektroskopi yang digunakan (*gap analysis*). Studi terdahulu cenderung menitikberatkan pada aspek pra-pemrosesan seperti penanganan pencilaan data [1], ataupun pada aspek arsitektur berupa integrasi multi-format data [3]. Ketidadaan transparansi dalam menjelaskan interaksi parameter fisis astronomi ini menjadi celah komputasi yang signifikan.

Penelitian ini diadakan untuk mengatasi keterbatasan tersebut dengan mengintegrasikan metode Explainable AI. Kebaruan (*novelty*) penelitian ini terletak pada penerapan analisis kontribusi fitur menggunakan algoritma Shapley Additive Explanations (SHAP) pada model *Random Forest*, *Gradient Boosting*, dan SVM menggunakan dataset SDSS17. Melalui pendekatan ini, performa model tidak hanya dinilai dari skor akurasi, melainkan dibongkar secara teoritis untuk membuktikan mengapa parameter fisis tertentu seperti redshift secara konsisten memiliki pengaruh paling signifikan [4], [7] dibandingkan dengan fitur koordinat posisi dalam keputusan klasifikasi objek langit.

2. Metode Penelitian

Penelitian ini disusun melalui serangkaian tahap yang dilakukan secara sistematis untuk mengklasifikasikan objek astronomi menggunakan algoritma Random Forest. Tahapan penelitian meliputi pengumpulan dataset, preprocessing data, pembagian data, pembangunan model Random Forest, evaluasi performa model, dan analisis feature importance. Seluruh tahapan tersebut dilakukan secara berurutan untuk memperoleh model klasifikasi yang mampu membedakan objek astronomi pada dataset dapat dilihat pada Gambar 1.



Gambar 1. Alur Penelitian

2.1 Pengumpulan Data

Dataset yang digunakan merupakan Star Classification dataset yang berisis 100.000 data observasi objek astronomi dengan 18 atribut. Dataset terdiri atas tiga kelas utama, yaitu bintang (Star), galaksi (Galaxy), dan quasar (Quasar). Setiap data observasi memiliki sejumlah atribut numerik yang merepresentasikan karakteristik spektral dan fotometrik objek astronomi, seperti nilai alpha, delta, redshift, magnitudo pada beberapa pita spektrum, serta informasi hasil pengamatan lain. Sebelum proses analisis dilakukan, data diperiksa untuk memastikan tidak terdapat duplikasi data maupun nilai yang hilang. Selanjutnya, atribut yang tidak memiliki kontribusi terhadap proses klasifikasi, seperti identitas unik objek, dieliminasi dari dataset.

Tabel 1. Variabel Penelitian

Variabel	Tipe Data	Keterangan
alpha	Numerik	Right Ascension, posisi objek pada koordinat langit
delta	Numerik	Declination, posisi objek pada koordinat langit
u	Numerik	Magnitudo pada pita ultraviolet (u-band)
g	Numerik	Magnitudo pada pita hijau (g-band)
r	Numerik	Magnitudo pada pita merah (r-band)
i	Numerik	Magnitudo pada pita inframerah dekat (i-band)
z	Numerik	Magnitudo pada pita inframerah (z-band)
redshift	Numerik	Pergeseran spektrum yang menunjukkan jarak relatif objek astronomi
class	Kategorikal	Kelas objek astronomi yang terdiri dari Star, Galaxy, dan Quasar

2.3 Preprocessing Data

Proses yang dilakukan meliputi pemeriksaan nilai kosong (missing value), penghapusan atribut yang tidak relevan, serta transformasi atribut kategorikal menjadi bentuk numerik menggunakan teknik label Encoding. Data kemudian dibagi menjadi data pelatihan (training set) dan data pengujian (testing set) menggunakan metode train-test split. Data pelatihan digunakan untuk membangun model, sedangkan data pengujian digunakan untuk mengukur kemampuan model dalam melakukan prediksi terhadap data.

2.4 Pembagian Data (Train dan Test)

Setelah proses pra-pemrosesan selesai, dataset dibagi menjadi data pelatihan (*training data*) dan data pengujian (*testing data*). Pembagian data dilakukan untuk melatih model menggunakan sebagian data serta menguji kemampuan generalisasi model menggunakan data yang belum pernah diproses. Jumlah data pelatihan dan data pengujian dihitung menggunakan Persamaan 1.

$$\begin{aligned}
 N_{train} &= p \times N \\
 &= (1 - p) \times N
 \end{aligned} \tag{1}$$

Dengan N_{train} adalah jumlah data pelatihan, N_{test} adalah jumlah data pengujian, N adalah jumlah keseluruhan data pada dataset, dan p adalah proporsi data pelatihan yang digunakan dalam proses pembagian data.

2.5 Pembangunan Metode Random Forest Classifier.

Metode klasifikasi yang digunakan dalam penelitian ini adalah Random Forest Classifier. Algoritma ini membangun sejumlah pohon keputusan (*decision tree*) menggunakan sampel data hasil *bootstrap sampling* dan menggabungkan seluruh hasil prediksi melalui mekanisme *majority voting*. Pendekatan tersebut memungkinkan model menghasilkan performa klasifikasi yang lebih stabil dibandingkan dengan penggunaan satu pohon keputusan tunggal. Prediksi akhir Random Forest diperoleh menggunakan Persamaan 2.

$$\hat{y} = \text{mode}(T_1(x), T_2(x), \dots, T_n(x)) \quad (2)$$

$\hat{y}$ adalah hasil prediksi akhir model Random Forest, $T_n(x)$ adalah prediksi yang dihasilkan oleh masing-masing pohon keputusan terhadap data masukan x , sedangkan n Adalah jumlah pohon keputusan yang digunakan dalam model. Pada proses pemisahan data pada setiap simpul pohon keputusan dilakukan menggunakan indeks Gini sebagaimana ditunjukkan pada Persamaan 3.

$$Gini = 1 - \sum_{i=1}^c p_i^2 \quad (3)$$

Gini adalah nilai ketidakmurnian (*impurity*) suatu simpul, p_i^2 adalah proporsi data pada kelas ke- i , dan c menunjukkan jumlah seluruh kelas yang terdapat pada dataset dan i merupakan indeks yang merepresentasikan setiap kelas dalam proses perhitungan. Semakin kecil nilai *Gini* yang diperoleh, semakin homogen data pada simpul tersebut sehingga atribut yang dipilih dianggap mampu menghasilkan pemisahan data.

2.6 Evaluasi Model

Untuk menguji kestabilan model, penelitian ini menerapkan metode Stratified K-Fold Cross Validation dengan jumlah lipatan (*fold*) sebanyak 5. Teknik ini membagi dataset menjadi lima bagian dengan proporsi kelas yang tetap terjaga pada setiap lipatan. Pada setiap iterasi, empat bagian digunakan sebagai data pelatihan dan satu bagian digunakan sebagai data validasi hingga seluruh bagian memperoleh giliran sebagai data validasi. Jumlah data pada setiap lipatan dihitung menggunakan Persamaan 4.

$$N_{fold} = \frac{N}{k} \quad (4)$$

N_{fold} adalah jumlah data pada setiap lipatan validasi, N adalah jumlah keseluruhan data yang digunakan dalam proses validasi silang, dan k adalah jumlah lipatan (*fold*) yang digunakan.

2.7 Analisis Feature Importance

Analisis *feature importance* dilakukan untuk mengetahui tingkat kontribusi masing-masing fitur dalam proses klasifikasi objek astronomi menggunakan model Random Forest. Nilai *feature importance* menunjukkan seberapa besar pengaruh suatu fitur terhadap kemampuan model dalam membedakan kelas Star, Galaxy, dan Quasar. Semakin tinggi nilai *feature importance*, semakin besar kontribusi fitur tersebut terhadap proses pengambilan keputusan model.

$$FI_j = \frac{1}{T} \sum_{t=1}^T \sum_{n \in t} \Delta Gini(n, j) \quad (5)$$

FI_j adalah nilai *feature importance* dari fitur ke- j , T adalah jumlah pohon keputusan (*decision tree*), $Gini(n, j)$ adalah penurunan nilai *Gini* yang dihasilkan oleh fitur ke- j pada simpul ke- n . Nilai *feature importance* diperoleh dari rata-rata akumulasi penurunan nilai *Gini* pada seluruh pohon Keputusan.

3. Hasil dan Pembahasan

3.1. Seleksi Fitur Menggunakan Mutual Information

Proses identifikasi faktor dominan diawali dengan seleksi fitur menggunakan metode Mutual Information seperti terlihat pada Tabel 2. Metode ini digunakan untuk mengukur tingkat keterkaitan antara setiap atribut dan kelas objek astronomi yang menjadi target klasifikasi. Nilai Mutual Information yang tinggi menunjukkan bahwa suatu atribut memiliki kemampuan yang lebih baik dalam membedakan objek Galaxy, Star, dan QSO. Hasil pengukuran ini digunakan sebagai dasar dalam menentukan fitur yang akan digunakan pada tahap pemodelan Random Forest. Informasi yang diperoleh diharapkan mampu meningkatkan efektivitas proses klasifikasi.

Tabel 2. Seleksi Fitur Mutual Information

No	Fitur	MI Score
1	redshift	0,801813
2	plate	0,274720
3	MJD	0,194624
4	z	0,145491
5	g	0,120400

Tabel 2 menunjukkan lima atribut dengan skor Mutual Information tertinggi yang diperoleh dari proses seleksi fitur. Variabel *redshift* menempati peringkat tertinggi sehingga menjadi atribut yang paling relevan terhadap kelas objek astronomi. Kondisi tersebut menunjukkan bahwa *redshift* memiliki kemampuan yang lebih baik dalam membedakan objek galaksi, bintang, dan QSO dibandingkan dengan atribut lainnya. Variabel *plate* dan *MJD* juga menunjukkan keterkaitan yang cukup kuat terhadap variabel target sehingga turut mendukung proses klasifikasi.

Temuan ini mengindikasikan bahwa karakteristik spektral memiliki peranan penting dalam identifikasi objek astronomi.

Perbedaan nilai Mutual Information antarfitur menunjukkan bahwa tidak semua atribut memiliki tingkat pengaruh yang sama terhadap variabel target. Fitur dengan nilai yang lebih tinggi cenderung memberikan informasi yang lebih besar dalam proses pembentukan model klasifikasi. Temuan tersebut menjadi dasar dalam pemanfaatan fitur-fitur terpilih pada algoritma Random Forest. Proses seleksi fitur juga membantu mengurangi atribut yang kurang informatif sehingga model dapat bekerja lebih efisien. Informasi yang diperoleh dari tahap ini selanjutnya digunakan untuk menganalisis faktor dominan berdasarkan nilai Feature Importance dan Permutation Importance.

3.2. Hasil Klasifikasi Menggunakan Random Forest

Klasifikasi objek astronomi dilakukan menggunakan algoritma Random Forest dengan fitur hasil seleksi sebelumnya. Evaluasi model dilakukan menggunakan metrik accuracy, precision, recall, dan F1-score untuk mengukur kemampuan model dalam mengklasifikasikan objek Galaxy, QSO, dan Star.

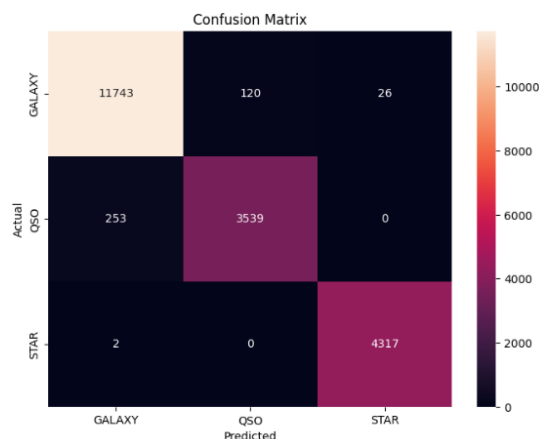
Tabel 3. Hasil Klasifikasi Random Forest

Metrik	Nilai
Accuracy	0,97995
Precision	0,97985
Recall	0,97995

Tabel 3 menunjukkan bahwa model Random Forest menghasilkan performa klasifikasi yang tinggi pada seluruh metrik evaluasi. Hasil tersebut menunjukkan bahwa sebagian besar objek astronomi berhasil diklasifikasikan sesuai dengan kelas sebenarnya. Nilai precision, recall, dan F1-score yang relatif seimbang mengindikasikan bahwa model mampu mempertahankan kualitas prediksi pada setiap kelas. Temuan ini menunjukkan bahwa atribut yang digunakan mampu memberikan informasi yang relevan terhadap proses klasifikasi. Analisis yang lebih rinci mengenai distribusi hasil prediksi selanjutnya dilakukan menggunakan confusion matrix.

3.3. Analisis Hasil Klasifikasi

Evaluasi lebih lanjut dilakukan menggunakan confusion matrix untuk mengetahui distribusi hasil prediksi pada masing-masing kelas objek astronomi. Analisis ini digunakan untuk mengidentifikasi jumlah data yang berhasil diklasifikasikan dengan benar maupun data yang masih mengalami kesalahan klasifikasi.



Gambar 2 .Hasil Confusion Matrix

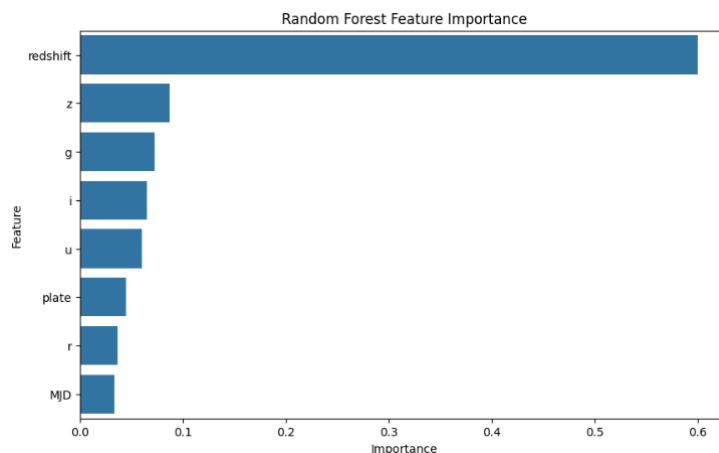
Gambar 2 menunjukkan bahwa sebagian besar observasi berada pada diagonal utama matriks, yang menandakan bahwa model berhasil mengklasifikasikan data sesuai dengan kelas sebenarnya. Jumlah prediksi yang benar pada setiap kelas jauh lebih besar dibandingkan jumlah kesalahan klasifikasi yang terjadi. Kesalahan prediksi masih ditemukan pada sebagian kecil data, terutama antara kelas Galaxy dan QSO yang memiliki karakteristik yang relatif mirip. Kelas Star menunjukkan tingkat ketepatan yang sangat tinggi karena hampir seluruh observasi berhasil diklasifikasikan dengan benar. Temuan ini menunjukkan bahwa model Random Forest memiliki kemampuan yang baik dalam membedakan objek Galaxy, QSO, dan Star berdasarkan atribut yang digunakan dalam penelitian.

Hasil confusion matrix mendukung temuan pada tahap evaluasi sebelumnya yang menunjukkan performa klasifikasi yang tinggi. Dominasi prediksi yang benar pada setiap kelas mengindikasikan bahwa model mampu

mengenali karakteristik utama masing-masing objek astronomi. Tingkat kesalahan yang relatif rendah menunjukkan bahwa atribut yang digunakan memiliki kemampuan diskriminasi yang baik terhadap variabel target. Kondisi tersebut memperlihatkan bahwa model mampu menghasilkan prediksi yang konsisten pada data pengujian. Analisis selanjutnya difokuskan pada identifikasi faktor dominan yang memengaruhi keputusan model menggunakan Feature Importance.

3.4. Analisis Faktor Dominan Berdasarkan Feature Importance

Analisis faktor dominan dilakukan menggunakan metode Feature Importance yang dihasilkan oleh model Random Forest. Metode ini digunakan untuk mengukur tingkat kontribusi setiap atribut terhadap proses pengambilan keputusan model dalam mengklasifikasikan objek astronomi. Nilai importance yang lebih tinggi menunjukkan bahwa suatu atribut memiliki pengaruh yang lebih besar terhadap hasil prediksi. Informasi tersebut digunakan untuk mengidentifikasi faktor-faktor yang paling dominan dalam membedakan objek Galaxy, QSO, dan Star.



Gambar 3. Hasil Random Forest Feature Importance

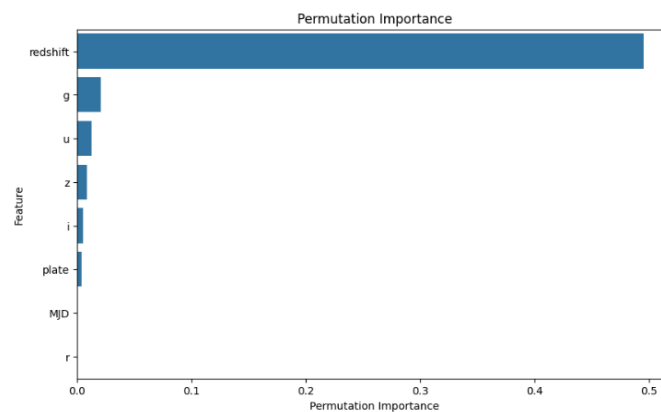
Gambar 3 menunjukkan bahwa setiap atribut memiliki tingkat kontribusi yang berbeda terhadap proses klasifikasi objek astronomi. Variabel *redshift* tampak memiliki kontribusi paling dominan dibandingkan atribut lainnya. Dominasi tersebut menunjukkan bahwa informasi mengenai pergeseran spektrum memiliki peranan penting dalam membedakan karakteristik objek Galaxy, QSO, dan Star. Atribut *z*, *g*, *i*, dan *u* juga memberikan kontribusi terhadap proses klasifikasi meskipun pengaruhnya relatif lebih kecil. Temuan ini menunjukkan bahwa karakteristik spektral dan fotometrik menjadi sumber informasi utama yang digunakan model dalam proses identifikasi objek astronomi.

Hasil Feature Importance menunjukkan pola yang sejalan dengan tahap seleksi fitur menggunakan Mutual Information. Atribut yang sebelumnya memiliki tingkat keterkaitan tinggi terhadap variabel target kembali muncul sebagai atribut dengan kontribusi terbesar dalam model Random Forest. Konsistensi tersebut menunjukkan bahwa faktor dominan yang diperoleh memiliki tingkat keandalan yang baik. Temuan ini mengindikasikan bahwa proses klasifikasi lebih banyak dipengaruhi oleh atribut tertentu dibandingkan dengan penggunaan seluruh atribut secara merata. Analisis lebih lanjut diperlukan untuk memvalidasi hasil tersebut melalui metode *Permutation Importance*.

3.5 Validasi Faktor Dominan Menggunakan *Permutation Importance*

Validasi faktor dominan dilakukan menggunakan *Permutation Importance* untuk mengukur pengaruh setiap atribut terhadap performa model. Metode ini bekerja dengan mengacak nilai suatu atribut dan mengamati perubahan kinerja model setelah proses pengacakan dilakukan. Penurunan performa yang lebih besar menunjukkan bahwa atribut tersebut memiliki kontribusi yang lebih penting terhadap hasil klasifikasi. Pendekatan ini digunakan untuk memastikan bahwa faktor dominan yang diperoleh tidak hanya bergantung pada mekanisme internal Random Forest.

Gambar 4 menunjukkan bahwa variabel *redshift* menghasilkan pengaruh terbesar terhadap performa model dibandingkan atribut lainnya. Kondisi tersebut menunjukkan bahwa kemampuan model dalam mengklasifikasikan objek astronomi sangat bergantung pada informasi yang terkandung dalam variabel tersebut. Atribut *g*, *u*, *z*, dan *i* juga memberikan kontribusi terhadap proses klasifikasi, meskipun pengaruhnya jauh lebih rendah dibandingkan *redshift*. Perbedaan kontribusi yang cukup besar menunjukkan bahwa tidak semua atribut memiliki peranan yang sama dalam pembentukan keputusan model. Temuan ini memperlihatkan bahwa sebagian besar informasi yang digunakan model berasal dari sejumlah atribut utama.

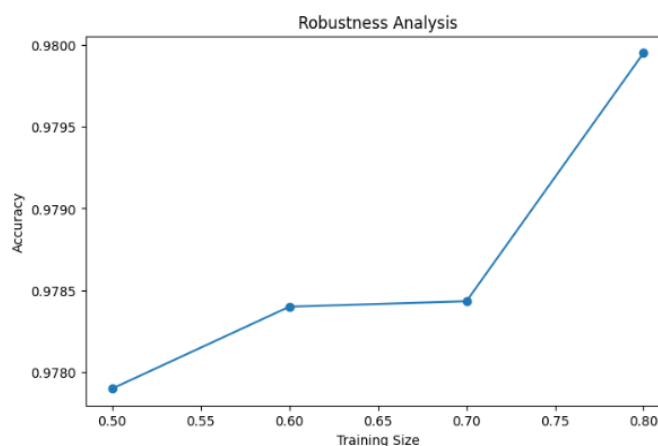


Gambar 4 . Hasil Permutation Importance

Hasil Permutation Importance menunjukkan pola yang konsisten dengan analisis Feature Importance pada subbab sebelumnya. Variabel *redshift* tetap muncul sebagai atribut yang paling berpengaruh terhadap proses klasifikasi, sehingga memperkuat posisinya sebagai faktor dominan dalam penelitian ini. Konsistensi hasil yang diperoleh dari kedua metode menunjukkan bahwa atribut tersebut memiliki kontribusi yang stabil terhadap performa model. Temuan ini mengindikasikan bahwa klasifikasi objek astronomi lebih banyak dipengaruhi oleh karakteristik yang direpresentasikan oleh *redshift* dibandingkan atribut lainnya. Hasil validasi tersebut memperkuat kesimpulan bahwa *redshift* merupakan faktor utama yang digunakan model Random Forest dalam membedakan objek Galaxy, QSO, dan Star.

3.6 Analisis Robustness Model

Analisis robustness dilakukan untuk mengevaluasi kestabilan performa model Random Forest pada berbagai ukuran data latih. Pengujian ini bertujuan untuk mengetahui pengaruh jumlah data yang digunakan selama proses pelatihan terhadap tingkat akurasi model. Performa model diamati melalui perubahan nilai akurasi pada beberapa proporsi data latih yang berbeda.



Gambar 5. Analisis Robustness

Gambar 5 menunjukkan bahwa nilai akurasi cenderung meningkat seiring bertambahnya ukuran data latih yang digunakan dalam proses pelatihan model. Pola tersebut mengindikasikan bahwa penambahan jumlah data memberikan informasi yang lebih banyak bagi model untuk mempelajari karakteristik objek astronomi. Perubahan nilai akurasi yang relatif kecil pada setiap ukuran data menunjukkan bahwa performa model tetap stabil selama pengujian dilakukan. Kondisi tersebut mengindikasikan bahwa model tidak mengalami penurunan performa yang signifikan ketika jumlah data latih berubah. Hasil ini menunjukkan bahwa Random Forest memiliki tingkat kestabilan yang baik dalam melakukan klasifikasi objek astronomi.

Konsistensi performa yang ditunjukkan pada berbagai ukuran data latih memperlihatkan bahwa model mampu mempertahankan kualitas prediksi secara relatif seragam. Tingkat akurasi yang tetap tinggi pada seluruh pengujian menunjukkan bahwa atribut yang digunakan telah mampu merepresentasikan karakteristik utama data dengan baik. Temuan ini mendukung hasil evaluasi sebelumnya yang menunjukkan performa klasifikasi yang tinggi dan

stabil. Hasil analisis robustness juga menunjukkan bahwa model memiliki kemampuan generalisasi yang baik terhadap data pengujian. Dengan demikian, model Random Forest yang dibangun dapat dianggap cukup andal untuk digunakan dalam klasifikasi objek astronomi.

4. Kesimpulan

Penelitian ini berhasil mengidentifikasi faktor dominan dalam klasifikasi objek astronomi menggunakan algoritma Random Forest. Hasil evaluasi menunjukkan bahwa model mampu melakukan klasifikasi objek Galaxy, QSO, dan Star dengan tingkat akurasi sebesar 97,995%, yang didukung oleh nilai precision, recall, dan F1-score yang relatif seimbang. Analisis Feature Importance menunjukkan bahwa atribut *redshift* merupakan faktor yang paling dominan dalam proses klasifikasi, sedangkan hasil Permutation Importance mengonfirmasi bahwa atribut tersebut memberikan pengaruh terbesar terhadap performa model dibandingkan atribut lainnya. Hasil confusion matrix menunjukkan bahwa sebagian besar objek berhasil diklasifikasikan sesuai dengan kelas sebenarnya, sementara analisis robustness menunjukkan bahwa model memiliki performa yang stabil pada berbagai ukuran data latih. Temuan ini mengindikasikan bahwa informasi yang terkandung dalam atribut *redshift* memiliki peranan penting dalam membedakan objek Galaxy, QSO, dan Star. Hasil penelitian dapat dimanfaatkan sebagai dasar dalam pengembangan sistem klasifikasi objek astronomi berbasis machine learning yang lebih efektif dan interpretatif. Penelitian selanjutnya dapat dilakukan dengan menggunakan dataset yang lebih besar atau membandingkan performa Random Forest dengan algoritma klasifikasi lainnya untuk memperoleh pemahaman yang lebih komprehensif mengenai faktor-faktor yang memengaruhi klasifikasi objek astronomi.

Daftar Rujukan

- [1] A. Estiak, "Classification of Galaxies, Stars, and Quasars from Sloan Digital Sky Survey Data Release 17 Using Different Machine Learning Techniques," Department of Physics, Shahjalal University of Science and Technology, 2026.
- [2] 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.
- [3] S. Ethiraj and B. K. Bolla, "Classification of Quasars, Galaxies, and Stars Using Multi-Modal Deep Learning," arXiv preprint, 2022, [Online]. Available: <https://arxiv.org/abs/2205.10745v1>
- [4] T. Zhou, "Comparison of machine learning algorithms and feature importance analysis for star classification," in Proc. Int. Conf. Mach. Learn. Autom., 2023, doi: 10.54254/2755-2721/30/20230111.
- [5] B. Arroquia-Cuadros, N. Sánchez, V. Gómez, P. Blay, V. Martínez-Badenes, and L. Nieves-Seoane, "Photometric classification of QSOs from ALHAMBRA survey using random forest," Astron. Astrophys., 2023, [Online]. Available: <https://arxiv.org/abs/2303.14567v1>
- [6] 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, p. 293, 2023, doi: 10.3390/a16060293.
- [7] D. Chatterjee and P. Ghosh, "Redshift-Agnostic Machine Learning Classification: Unveiling Peak Performance in Galaxy, Star, and Quasar Classification (Using SDSS DR17)," Astron. Nachr., vol. 346, no. 2, 2025, doi: 10.1002/asna.20240034.
- [8] F. Z. Zeraatgari, F. Hafezianzadeh, Y. Zhang, L. Mei, A. Ayubinia, A. Mosallanezhad, and J. Zhang, "Machine learning-based photometric classification of galaxies, quasars, emission-line galaxies, and stars," Mon. Not. R. Astron. Soc., vol. 527, no. 3, pp. 4677–4689, 2024, doi: 10.1093/mnras/stad3436.