

Penerapan CRISP-DM untuk Deteksi Eksoplanet menggunakan Algoritma Regresi Logistik dan K-Nearest Neighbors

Livanty Efatania Dendy¹, Valencia Elcheiana Irawan², dan Rinabi Tanamal³

^{1,2,3} Program Studi Sistem Informasi Bisnis, Universitas Ciputra Surabaya
CitraLand CBD Boulevard, Surabaya, Indonesia, 60219

Korespondensi: Livanty Efatania Dendy (lefatania@student.ciputra.ac.id)

Received: 24 Juli 2024 – *Revised:* 31 Agustus 2024 - *Accepted:* 05 Sept 2024 - *Published:* 10 Sept 2024

Abstrak. Studi ini bertujuan untuk mengeksplorasi penggunaan model machine learning dalam klasifikasi eksoplanet dengan memprediksi apakah objek luar angkasa adalah eksoplanet berdasarkan intensitas cahayanya. Penemuan eksoplanet merupakan salah satu perkembangan paling menarik dalam astrofisika modern, hingga kini ribuan eksoplanet telah dikonfirmasi dan akan terus bertambah. Secara khusus, studi ini mengikuti metodologi CRISP-DM, yang mencakup pemahaman bisnis, pemahaman data, persiapan data, pemodelan, dan evaluasi serta menggunakan model Regresi Logistik dan K-Nearest Neighbors untuk menganalisis data dari Teleskop Luar Angkasa Kepler NASA. Tujuan utamanya adalah meningkatkan akurasi klasifikasi eksoplanet dan membantu dalam mengidentifikasi planet yang memiliki potensi untuk mendukung kehidupan. Model Regresi Logistik menunjukkan akurasi sebesar 0.53, *precision* sebesar 0.98, *recall* sebesar 0.53, dan *F1 Score* sebesar 0.68, menunjukkan kinerja sedang dengan fokus pada presisi. Sebaliknya, model K-Nearest Neighbors mencapai akurasi sebesar 0.99, *precision* sebesar 0.98, *recall* sebesar 0.99, dan *F1 Score* sebesar 0.99, menunjukkan kinerja superior di semua metrik. Studi ini menyimpulkan bahwa model K-Nearest Neighbors jauh lebih efektif untuk deteksi eksoplanet dibandingkan dengan Regresi Logistik. Studi selanjutnya disarankan untuk mempertimbangkan penambahan variabel tambahan dan memperluas ukuran sampel untuk meningkatkan validitas hasil serta mengeksplorasi penggunaan algoritma machine learning lainnya untuk meningkatkan akurasi prediksi.

Kata kunci: Eksoplanet, Machine Learning, CRISP-DM, Regresi Logistik, K-Nearest Neighbors

Citation Format: Dendy, L.E., Irawan, V.E., & Tanamal, R. (2024). Penerapan CRISP-DM untuk Deteksi Eksoplanet menggunakan Algoritma Regresi Logistik dan K-Nearest Neighbors. *Prosiding SENAM 2024: Seminar Nasional Sistem Informasi & Informatika Universitas Ma Chung*. 4, 160-169. Malang: Ma Chung Press.

PENDAHULUAN

Tata surya kita hanyalah salah satu dari milyaran tata surya di Bima Sakti, dan banyak di antaranya kemungkinan memiliki planet mirip Bumi (Loeb, 2021). Penemuan eksoplanet merupakan salah satu perkembangan paling menarik dalam astrofisika modern. Eksoplanet ialah planet yang berada di luar tata surya kita dan mengorbit bintang lain. Penemuan eksoplanet pertama kali terjadi pada tahun 1992 ketika dua eksoplanet ditemukan mengorbit pulsar PSR 1257+12 (Bruna *et al.*, 2023). Sejak deteksi eksoplanet pertama pada tahun 1992, ribuan eksoplanet telah dikonfirmasi dan akan terus bertambah.

Hal ini hanya sebagian kecil dari sampel galaksi secara keseluruhan. Eksoplanet ini menawarkan wawasan yang tak ternilai tentang keragaman sistem planet dan potensi keberadaan kehidupan di luar Bumi. Pencarian untuk eksoplanet yang dapat ditinggali adalah lebih dari sekadar sebuah usaha ilmiah; ini adalah perlombaan melawan waktu saat kita mencari alternatif atau pelengkap untuk sumber daya Bumi yang semakin berkurang dan tantangan lingkungan yang semakin meningkat (Pyne, 2022). Impey juga menekankan urgensi pencarian planet yang dapat dihuni karena ancaman eksistensial yang kita hadapi di Bumi, seperti pengurasan sumber daya dan degradasi lingkungan. Menurutnya, menemukan dan mungkin eksplorasi planet lain bisa menjadi penting untuk kelangsungan hidup jangka panjang umat manusia (Impey, 2023).

Menurut jurnal “Mencari Bumi yang Baru” pengklasifikasian planet terbagi menjadi dua, yaitu menurut ukuran planetnya dan menurut kelayakhuniannya. Menurut ukuran planetnya, planet dapat dikategorikan menjadi dua jenis, yaitu planet raksasa (*giant planets*) dan planet kerdil (*small size planets*) dimana planet kerdil tergolong menjadi planet terrestrial (*terrestrial planets*). Sedangkan menurut kelayakhuniannya, planet diklasifikasikan menjadi planet layak huni (*habitable planets*) dan planet mirip bumi (*earth-like planets*) (Nurcresia *et al.*, 2018).

Pada umumnya, sebagian besar planet yang telah ditemukan bukan tempat yang ramah untuk kehidupan, umumnya mereka memiliki atmosfer yang eksotik hingga temperatur yang ekstrim. Namun, di antara ribuan eksoplanet tersebut, beberapa diantaranya juga tidak menutup kemungkinan dalam menimbulkan harapan akan adanya kehidupan (Yustika *et al.*, 2021).

Seiring dengan kemajuan teknologi dan melalui beberapa metode, seperti metode deteksi dan metode kecepatan radial, akan meningkatkan kemampuan para ilmuwan untuk menemukan dan mempelajari eksoplanet dengan lebih detail. Metode transit yaitu metode untuk mengukur penurunan dalam intensitas cahaya bintang yang disebabkan oleh sebuah planet yang melintas di depannya. Penurunan intensitas cahaya ini memberikan informasi tentang ukuran dan orbit eksoplanet tersebut. Intensitas cahaya memainkan peran penting dalam berbagai studi, termasuk dalam mempelajari atmosfer eksoplanet. Studi menunjukkan bahwa intensitas cahaya memiliki dampak signifikan terhadap berbagai organisme (Dewi *et al.*, 2023). Melalui analisis ini, jejak penyerapan atau pantulan cahaya oleh molekul-molekul di atmosfer dapat memberikan informasi tentang komposisi atmosfer, distribusi vertikal gas-gas, serta sifat fisik atmosfer seperti tekanan, suhu, dan

kelembaban. Namun perlu diperhatikan juga intensitas cahaya yang terlalu rendah dapat membuat fotosintesis menjadi tidak mungkin, yang penting untuk kehidupan seperti yang kita kenal, sebaliknya, intensitas cahaya yang terlalu tinggi dapat menyebabkan penguapan air dari permukaan planet, yang membuatnya kering dan tidak cocok untuk kehidupan (Kaltenegger, 2017).

MASALAH

Berdasarkan pendahuluan yang telah diuraikan di atas, dapat dirumuskan beberapa permasalahan yaitu:

- Apakah model Regresi Logistik lebih akurat dibandingkan model *K-Nearest Neighbors* (KNN) dalam klasifikasi eksoplanet berdasarkan intensitas cahaya?
- Bagaimana kinerja model *K-Nearest Neighbors* dibandingkan dengan model Regresi Logistik dalam parameter evaluasi utama (*Accuracy, Precision, Recall, dan F1 Score*)?

METODE PELAKSANAAN

Studi ini dilaksanakan dengan mengadopsi sebuah standar proses data mining yang dikenal sebagai *Cross-Industry Standard Process for Data Mining* (CRISP-DM). CRISP-DM adalah standar untuk pemrosesan data mining yang telah dikembangkan, di mana data yang ada akan melewati setiap fase yang terstruktur dan terdefinisi dengan jelas dan efisien (Hasanah *et al.*, 2021). Metodologi ini terdiri dari lima fase utama: pemahaman masalah, pemahaman data, persiapan data, pemodelan, evaluasi. Setiap fase ini dirancang untuk memastikan bahwa proses data mining dilakukan dengan cara yang sistematis dan dapat diulang, memungkinkan untuk mengidentifikasi pola dan informasi yang berguna dari data yang ada dengan efektif.



Gambar 1. Standar proses model CRISP DM

Fase Pemahaman Masalah

Pada fase pemahaman masalah dalam deteksi eksoplanet menggunakan Regresi Logistik dan *K Nearest Neighbors*, tujuan utamanya adalah untuk mencari tingkat akurasi dan efisiensi dalam mengidentifikasi eksoplanet melalui metode analitik. Kebutuhan ini diidentifikasi dalam konteks peningkatan kemampuan deteksi yang dapat menghemat biaya dan waktu dalam observasi astronomi, serta mempercepat penemuan planet baru. Tujuan ini diterjemahkan menjadi formula *data mining* dengan menggunakan Regresi Logistik untuk analisis klasifikasi dan *K-Nearest Neighbors* untuk pengelompokan data yang besar dan kompleks, memungkinkan identifikasi pola yang lebih tepat dalam data astronomi. Strategi awal untuk mencapai tujuan ini meliputi pengumpulan dan pre-processing data, pengujian model Regresi Logistik dan *K-Nearest Neighbors*, serta evaluasi hasil untuk iterasi dan perbaikan model. Dengan pendekatan ini, studi ini bertujuan untuk memberikan kontribusi signifikan dalam ilmu astronomi dan eksplorasi luar angkasa, sekaligus membuka peluang baru dalam inovasi teknologi terkait.

Fase Pemahaman Data

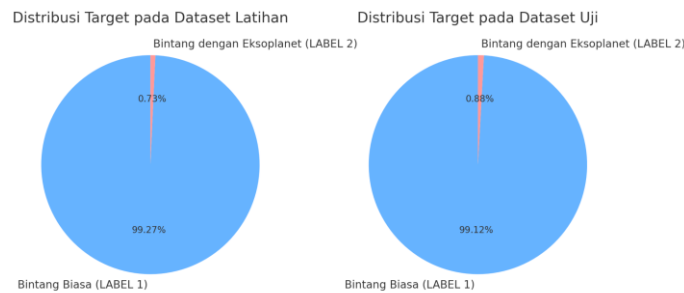
Dalam studi ini, peneliti menggunakan dataset yang bersumber dari Kaggle, yang dimana dataset ini telah dibersihkan. Peneliti menggunakan dataset studi dari Teleskop Luar Angkasa Kepler NASA yang berjudul "*Exoplanet Hunting in Deep Space*" (Kaggle.com, n.d.). Dataset ini memiliki nama kolom dan tipe data seperti yang ditunjukkan pada Tabel 1.

Tabel 1. Nama Kolom dan Tipe Data

Nama Kolom	Tipe Data
LABEL	Int64
FLUX1	Float
FLUX2	Float
⋮	⋮
FLUX3197	Float

Dataset ini dibagi menjadi dua bagian yaitu dataset latihan (*exoTrain*) dan dataset uji (*exoTest*) dengan proporsi pembagian data masing-masing 90% untuk dataset latihan dan 10% untuk dataset uji dengan total 5.657 data yang terdiri dari 3.197 variabel dependen dan 1 variabel independen. Distribusi target dalam dataset menunjukkan ketidakseimbangan kelas yang signifikan. Dalam dataset latihan, 99.27% entri adalah bintang biasa (LABEL 1) dan hanya 0.73% yang merupakan bintang dengan eksoplanet (LABEL 2). Pola yang sama terlihat dalam dataset uji dengan 99.12% bintang biasa dan

0.88% bintang dengan eksoplanet seperti yang tertera pada Gambar 2.



Gambar 2. Distribusi Dataset

Fase Persiapan Data

Dataset yang telah diperoleh akan diolah melalui pemrosesan pada fase persiapan data. Proses ini meliputi penggantian nilai dari kolom label, dimana data yang memiliki angka 1 akan digantikan dengan angka 0, sedangkan data yang memiliki angka 2 akan digantikan dengan angka 1.

Fase Pemodelan

a. Regresi Logistik

Regresi Logistik adalah metode klasifikasi dalam *statistical machine learning* yang populer digunakan dalam data mining karena kemampuannya yang baik dalam menangani data skala besar. Metode ini menghubungkan output biner dengan variabel-variabel independen berdasarkan probabilitas untuk memprediksi nilai dari variabel dependen, yang dimana pada hasil akhirnya akan mengklasifikasikan data ke dalam dua kategori yang berbeda.

$$l = \log p \left(\frac{p}{1-p} \right)^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n} \quad (1)$$

Keterangan:

l : log-odds, p adalah basis dari logaritma

β_n : adalah parameter model

p_y : adalah probabilitas dari kejadian tersebut

Istilah "variabel target" digunakan untuk merujuk pada variabel dependen dalam pembelajaran mesin. Variabel prediktor atau fitur adalah nama lain untuk variabel independen.

b. K-Nearest Neighbors

K-Nearest Neighbors merupakan metode untuk mengklasifikasikan data baru berdasarkan posisi jarak terdekat antara data baru tersebut dengan data lain

atau data tetangga terdekat. objek berdasarkan data pembelajaran yang paling dekat dengan objek tersebut. Ada banyak cara untuk mengukur jarak kedekatan antara data baru dengan data lama (*data training*), diantaranya *euclidean distance* dan *manhattan distance (city block distance)*, yang paling sering digunakan adalah *euclidean distance* (Bramer, 2007), yaitu:

$$d(x_1, x_2) = \sqrt{\sum_{i=1}^n (x_{1i} - x_{2i})^2} \quad (2)$$

Keterangan:

d = jarak

i = variabel data

n = dimensi data

x1 = sampel data

x2 = data uji

Namun, dalam menggunakan metode ini, harus lebih memperhatikan satuan data yang ingin dianalisis dikarenakan biasanya data yang digunakan memiliki variasi satuan sehingga menyebabkan hasil analisis tidak akurat. Oleh karena itu, standardisasi atau transformasi data perlu dilakukan sebelum klasifikasi. Dalam studi ini penulis menggunakan *min-max normalization* karena data terdiri dari skala pengukuran numerik dan kategorikal.

$$x_{baru} = \frac{x_1 - \min(x_1)}{\max(x_1) - \min(x_1)} \quad (3)$$

Setelah data dikonversi ke satuan yang sama, proses selanjutnya yaitu pemilihan nilai *k* (*dataset/nearest neighbour*) yang dapat mempengaruhi tingkat *accuracy* model prediksi.

Fase Evaluasi

Untuk menilai seberapa baik model yang diajukan dalam memprediksi dan memperoleh nilai proporsi yang tepat dibandingkan dengan nilai asli dari data yang tersedia, dapat dilihat dari nilai *Confusion Matrix* seperti pada Tabel 2.

Tabel 2. *Confusion Matrix*

Nilai Prediksi	Nilai Sebenarnya	
	Positif	Negatif
Positif	<i>True Positives</i>	<i>False Positives</i>
Negatif	<i>False Negatives</i>	<i>True Negatives</i>

Tabel 2 mengilustrasikan konsep *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) dalam evaluasi model. TP terjadi ketika prediksi

dan data keduanya positif, TN terjadi ketika prediksi dan data keduanya negatif, FP terjadi ketika model memprediksi positif namun data sebenarnya negatif, dan FN terjadi ketika model memprediksi negatif namun data sebenarnya positif. Berikut adalah rumus untuk menghitung nilai *accuracy* pada persamaan (4), *precision* pada persamaan (5), *recall* pada persamaan (6) dan *F1 score* pada persamaan (7).

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

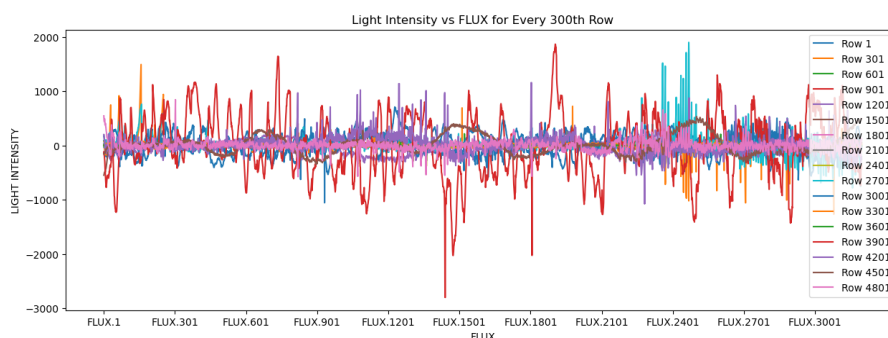
$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F1\ Score = \frac{2 * (recall * precision)}{recall + precision} \quad (7)$$

HASIL DAN PEMBAHASAN

Dataset ini terdiri menjadi dua bagian yaitu dataset latihan (*exoTrain*) dan dataset uji (*exoTest*). Statistik deskriptif menunjukkan bahwa nilai fluks di berbagai titik pengukuran sangat bervariasi dengan rentang nilai yang sangat luas. Median dari mayoritas nilai fluks mendekati nol, yang menunjukkan banyaknya nilai kecil dalam dataset. Pola statistik antara dataset latihan dan uji sangat mirip, menunjukkan konsistensi data antara kedua set tersebut.

Secara keseluruhan, analisis eksplorasi data ini menunjukkan bahwa dataset memiliki ketidakseimbangan kelas yang signifikan dan variasi besar dalam nilai fluks. Dalam konteks astronomi, fluks sering kali digunakan untuk mengukur jumlah energi yang diterima oleh alat pengamatan, seperti teleskop, dari objek astronomi tertentu. Pada Gambar 3 ini menunjukkan hubungan antara intensitas cahaya (*Light Intensity*) dan fluks (*FLUX*) untuk setiap baris ke-300 dalam dataset eksoplanet. Setiap kurva pada grafik ini mewakili variasi intensitas cahaya terhadap fluks pada titik waktu yang berbeda untuk sampel data yang berbeda.



Gambar 3. Hubungan Intensitas Cahaya dan Fluks

Dalam studi ini, peneliti telah menerapkan dua algoritma berbeda, yaitu Regresi Logistik dan *K-Nearest Neighbors*, untuk mendeteksi eksoplanet. Masing-masing algoritma dievaluasi berdasarkan beberapa metrik performa, termasuk *accuracy*, *precision*, *recall*, dan *F1 score* seperti pada Tabel 3.

Tabel 3. Hasil Model Regresi Logistik dan *K-Nearest Neighbors*

Model	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F1 Score</i>
Regresi Logistik	0.53	0.98	0.53	0.68
<i>K-Nearest Neighbors</i>	0.99	0.98	0.99	0.99

Model Regresi Logistik menunjukkan hasil *accuracy* sebesar 0.53 menunjukkan bahwa model ini hanya benar dalam 53% dari semua prediksinya, menandakan bahwa model ini kurang mampu mengklasifikasikan eksoplanet dengan benar secara keseluruhan. *Precision* sebesar 0.98 menunjukkan bahwa ketika model memprediksi suatu objek sebagai eksoplanet, prediksinya hampir selalu benar, sehingga model sangat baik dalam menghindari *false positives*. Namun, *recall* sebesar 0.53 menunjukkan bahwa model hanya mendeteksi 53% dari semua eksoplanet yang sebenarnya, berarti model cenderung menghasilkan banyak *false negatives*, gagal mengidentifikasi sebagian besar eksoplanet. *F1 Score* sebesar 0.68, yang merupakan rata-rata harmonis dari *precision* dan *recall*, menunjukkan bahwa secara keseluruhan, performa model dalam mendeteksi eksoplanet adalah sedang, dengan keseimbangan yang lebih berat ke *precision* daripada *recall*. Secara keseluruhan, model Regresi Logistik kurang efektif untuk mendeteksi eksoplanet karena meskipun *precision* tinggi, *recall* rendah hal ini menunjukkan bahwa banyak eksoplanet yang tidak terdeteksi.

Untuk model *K-Nearest Neighbors*, peneliti menggunakan parameter k sebesar 5, yang berarti bahwa klasifikasi dilakukan berdasarkan mayoritas dari 5 tetangga terdekatnya. Parameter ini dipilih untuk menyeimbangkan antara bias dan varians dalam model. Dengan menggunakan $k=5$, model diharapkan dapat menangkap pola yang cukup kompleks tanpa terlalu sensitif terhadap variasi atau anomali dalam data.

Model *K-Nearest Neighbors* menunjukkan hasil *accuracy* sebesar 0.99 menunjukkan bahwa model ini benar dalam 99% dari semua prediksinya, menandakan bahwa model ini sangat efektif dalam mengklasifikasikan eksoplanet dengan benar secara keseluruhan. *Precision* sebesar 0.98 menunjukkan bahwa ketika model memprediksi suatu objek sebagai eksoplanet, prediksinya hampir selalu benar, mirip dengan Regresi Logistik. *Recall* sebesar 0.99 menunjukkan bahwa model ini mendeteksi 99% dari semua eksoplanet

yang sebenarnya, berarti model sangat sedikit menghasilkan *false negatives*, dengan hampir semua eksoplanet berhasil diidentifikasi. *F1 Score* sebesar 0.99, yang merupakan rata-rata harmonis dari *precision* dan *recall*, menunjukkan bahwa model ini sangat seimbang dan efektif dalam mendeteksi eksoplanet, dengan performa yang sangat tinggi baik dalam *precision* maupun *recall*. Secara keseluruhan, model *K-Nearest Neighbors* sangat efektif untuk mendeteksi eksoplanet karena memiliki *accuracy*, *precision*, *recall*, dan *F1 score* yang sangat tinggi, menunjukkan bahwa model ini mampu mengidentifikasi hampir semua eksoplanet dengan sedikit kesalahan prediksi.

KESIMPULAN

Dengan studi ini, bisa dilihat bahwa penggunaan data *flux* dapat digunakan untuk memprediksi eksoplanet di masa yang akan datang. Kemudian berdasarkan hasil di atas, model *K-Nearest Neighbors* jauh lebih efektif dibandingkan dengan model Regresi Logistik dalam mendeteksi eksoplanet. Meskipun *precision* dari kedua model sangat tinggi namun *K-Nearest Neighbors* juga memiliki *recall* dan *accuracy* yang jauh lebih baik, menjadikannya pilihan yang lebih baik untuk menentukan eksoplanet. Hasil yang diperoleh dari *K-Nearest Neighbors* adalah *Accuracy* sebesar 0.99, *Precision* sebesar 0.98, *Recall* sebesar 0.99, dan *F1 Score* sebesar 0.99.

Studi ini menyimpulkan bahwa metode *K-Nearest Neighbors* memberikan hasil terbaik dalam memprediksi keberadaan eksoplanet dibandingkan dengan Regresi Logistik. Studi ini juga merekomendasikan penambahan algoritma atau variabel dan perluasan sampel dalam studi selanjutnya untuk meningkatkan validitas hasil. Selain itu, disarankan untuk menggunakan algoritma *machine learning* lainnya dan menggabungkan kedua metode untuk meningkatkan akurasi prediksi.

DAFTAR PUSTAKA

- Bruna, M., Cowan, N. B., Sheffler, J., Haggard, H. M., Bourdon, A., & Mâlin, M. (2023). Combining photometry and astrometry to improve orbit retrieval of directly imaged exoplanets. *Monthly Notices of the Royal Astronomical Society*, 519(1), 460–470. <https://doi.org/10.1093/mnras/stac3521>
- Dewi, R., Winanto, T., Haryono, F. E. D., Marhaeni, B., Hanifa, G., Nabila, D., Muis, D. R., & Khalisa, S. (2023). Potensi Klorofil dan Karotenoid Fitoplankton *Dunaliella salina* sebagai Sumber Antioksidan. *Buletin Oseanografi Marina*, 12(1), 125–132. <https://doi.org/10.14710/buloma.v12i1.49006>

- Hasanah, M. A., Soim, S., & Handayani, A. S. (2021). Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir. *Journal of Applied Informatics and Computing*, 5(2), 103–108. <https://doi.org/10.30871/jaic.v5i2.3200>
- Impey, C. (2023). *Worlds Without End Exoplanets, Habitability, and the Future of Humanity*. The MIT Press.
- Kaggle.com. (n.d.). *Exoplanet Hunting in Deep Space*. Kaggle.Com. Retrieved June 10, 2024, from <https://www.kaggle.com/datasets/keplersmachines/kepler-labelled-time-series-data>
- Kaltenegger, L. (2017). How to Characterize Habitable Worlds and Signs of Life. *Annual Review of Astronomy and Astrophysics*, 55, 433–485. <https://doi.org/10.1146/annurev-astro-082214-122238>
- Loeb, A. (2021). *Extraterrestrial: The First Sign of Intelligent Life Beyond Earth*. Houghton Mifflin Harcourt.
- Nurcresia, B., Simbolon, T. R., & Setiawan, J. (2018). Mencari Bumi yang Baru. *Academia*.
- Pyne, S. J. (2022). *The Pyrocene How We Created an Age of Fire, and What Happens Next*. University of California Press.
- Yustika, S. I., Utama, J. A., Arifin, M., & ... (2021). Karakteristik Eksoplanet Laik Huni Di Sistem Multiplanet. *Prosiding Seminar Nasional Fisika 7.0 (2021)* 311-317. <http://proceedings2.upi.edu/index.php/sinafi/article/view/1849>



© 2024 by authors. Content on this article is licensed under a Creative Commons Attribution 4.0 International license. (<http://creativecommons.org/licenses/by/4.0/>).