

# Logistic Regression for Classifying Asteroid Impact Risk Using Orbital Features

*[The title should clearly and succinctly reflect the main content of the research and be appealing to academic readers. It must consist of 8–12 words (excluding the definite article the and the indefinite articles a, an). Use concise noun or verb phrases rather than full sentences.]*

**First Author<sup>1</sup>, Second Author<sup>2</sup>, ...**

Email: [email1@mail.smds.ac.id](mailto:email1@mail.smds.ac.id), [email2@mail.smds.ac.id](mailto:email2@mail.smds.ac.id),

*[In the submission process (Step 3. Entering the Submission's Metadata), Author(s) must fill out authors' information that consist of author First, Middle and Last Name (if author only uses one name, fill the first name and last name with the same name), Gender, Initial, Username, Password, Affiliation (University / or Institution), E-mail, Mailing Address (complete address of affiliation), Country and Bio Statement.]*

**Abstract** – Penelitian ini mengembangkan model klasifikasi yang transparan untuk mengidentifikasi asteroid berisiko tinggi bertumbukan dengan Bumi berdasarkan karakteristik fisik dan parameter orbit. Data yang digunakan berasal dari himpunan terbuka Possible Asteroid Impacts with Earth yang memuat informasi mengenai skenario kemungkinan tumbukan, diameter, kecepatan, energi tumbukan, serta parameter orbit utama termasuk jarak minimum perpotongan orbit dengan Bumi. Setelah proses penggabungan dan pembersihan, dilakukan feature engineering untuk membentuk variabel yang merepresentasikan intensitas skenario tumbukan, besaran energi, dan kedekatan lintasan. Ketidakseimbangan kelas antara asteroid berisiko tinggi dan rendah ditangani melalui pembobotan kelas dan pembagian data secara berstrata, sementara seluruh prediktor numerik dinormalisasi agar berada pada skala yang sebanding. Model utama yang digunakan adalah regresi logistik dengan regularisasi, dengan hiperparameter ditetapkan melalui grid search dan validasi silang. Hasil evaluasi pada data uji menunjukkan bahwa model mencapai akurasi 0,8676, balanced accuracy 0,8787, dan nilai AUC 0,947, yang mengindikasikan kemampuan diskriminasi yang sangat baik antara asteroid berisiko tinggi dan rendah. Analisis koefisien memperlihatkan bahwa banyaknya skenario tumbukan yang tercatat,

diameter asteroid, kecepatan, serta kombinasi kecepatan dengan jarak minimum perpotongan orbit merupakan determinan utama peningkatan risiko. Temuan ini menunjukkan bahwa regresi logistik dapat menjadi pendekatan dasar yang efektif dan mudah diinterpretasikan untuk screening awal asteroid potensial berbahaya, meskipun penyesuaian ambang keputusan dan pengayaan dengan informasi dinamis masih diperlukan untuk mengurangi kegagalan deteksi objek berisiko tinggi.

**Keywords:** Asteroid impact prediction; Near-Earth objects; Logistic regression; Risk classification; Machine learning

## I. Introduction

Asteroid dan objek dekat Bumi (*near-Earth objects*, NEO) merupakan salah satu sumber risiko alam yang mendapat perhatian besar dalam beberapa dekade terakhir. Berbagai peristiwa historis, mulai dari tumbukan skala purba hingga ledakan udara seperti insiden Chelyabinsk, menunjukkan bahwa kejadian yang relatif jarang ini dapat menimbulkan dampak signifikan terhadap lingkungan dan aktivitas manusia. Perkembangan program survei langit dan katalogisasi NEO membuat jumlah asteroid yang terdeteksi terus meningkat, sehingga diperlukan pendekatan

kuantitatif yang mampu mengelola informasi dalam jumlah besar dan membantu mengidentifikasi objek mana yang perlu diprioritaskan untuk pemantauan lebih lanjut.

Selama ini, penilaian risiko tumbukan terutama mengandalkan pemodelan dinamika orbit berbasis fisika dan simulasi numerik yang kompleks. Pendekatan tersebut krusial untuk menghasilkan perhitungan lintasan yang presisi, namun tidak selalu dirancang untuk memberikan gambaran ringkas mengenai faktor-faktor yang secara statistik paling berkaitan dengan peningkatan risiko. Di sisi lain, pemanfaatan teknik *machine learning* untuk deteksi dan klasifikasi objek astronomi mulai berkembang, tetapi banyak model yang bersifat *black box* sehingga sulit diinterpretasikan oleh peneliti maupun pengambil kebijakan. Dalam konteks mitigasi risiko, model yang tidak hanya akurat, tetapi juga transparan dan mudah dijelaskan menjadi sangat penting agar dasar pengambilan keputusan dapat dipertanggungjawabkan.

Penelitian ini memanfaatkan dataset terbuka *Possible Asteroid Impacts with Earth* untuk membangun model klasifikasi risiko tumbukan berbasis regresi logistik. Informasi yang digunakan mencakup karakteristik fisik asteroid, seperti diameter, kecepatan, dan energi tumbukan, serta parameter orbit utama termasuk jarak minimum perpotongan orbit dengan Bumi. Melalui serangkaian tahapan prapengolahan dan *feature engineering*, variabel-variabel tersebut diolah menjadi himpunan fitur yang merepresentasikan ukuran dan dinamika gerak asteroid sekaligus intensitas skenario tumbukan yang tercatat. Regresi logistik dipilih karena mampu menghasilkan estimasi probabilitas risiko dan menyediakan koefisien yang dapat langsung ditafsirkan sebagai pengaruh relatif masing-masing faktor terhadap peningkatan risiko.

Secara khusus, penelitian ini bertujuan untuk menjawab dua pertanyaan utama. Pertama, sejauh mana model regresi logistik yang dilengkapi penanganan ketidakseimbangan kelas dan penyetaraan skala mampu membedakan asteroid berisiko tinggi dan berisiko rendah pada dataset tersebut. Kedua, faktor-faktor fisik dan orbit mana yang

memberikan kontribusi terbesar terhadap peningkatan probabilitas risiko menurut model yang diestimasi. Untuk menjawab pertanyaan tersebut, artikel ini menyajikan uraian mengenai metode pengolahan data dan pemodelan, hasil evaluasi kinerja model pada data uji, serta interpretasi koefisien utama beserta implikasinya bagi pemantauan asteroid potensial berbahaya.

## II. Research Methods

### A. LOAD DATA

Data dalam penelitian ini bersumber dari dataset sekunder *Possible Asteroid Impacts with Earth* yang terdiri atas dua bagian utama: data potensi tumbukan dan data parameter orbit asteroid. Kedua bagian tersebut dianalisis terlebih dahulu, kemudian digabungkan setelah nama objek asteroid distandarkan (misalnya dengan menghapus spasi berlebih dan tanda kurung) sehingga setiap asteroid memiliki penanda yang konsisten.

Setelah penggabungan, dilakukan pemeriksaan nilai hilang dan duplikasi pada variabel-variabel kunci, seperti probabilitas kumulatif tumbukan, jumlah kemungkinan tumbukan, karakteristik fisik (kecepatan, magnitudo, diameter, energi tumbukan), dan parameter orbit (eksentrisitas, inklinasi, sumbu orbit, serta jarak minimum perpotongan orbit dengan Bumi). Observasi yang tidak memiliki informasi lengkap pada variabel-variabel tersebut dihapus sehingga analisis selanjutnya hanya menggunakan data yang utuh dan konsisten, guna meminimalkan distorsi pada hasil pemodelan.

### B. PEMAHAMAN DATASET

Setelah kedua berkas data digabungkan, tahap selanjutnya adalah menganalisis struktur dan isi dataset yang menjadi dasar pemodelan. Setiap baris merepresentasikan satu asteroid yang berpotensi mendekati atau bertumbukan dengan Bumi, dengan informasi yang mencakup karakteristik fisik dan parameter orbit sehingga memberikan gambaran cukup lengkap mengenai kondisi dinamis setiap objek.

Variabel utama yang digunakan meliputi probabilitas kumulatif tumbukan (*Cumulative Impact Probability*), jumlah skenario tumbukan yang tercatat (*Possible Impacts*), kecepatan relatif (*Asteroid Velocity*), magnitudo atau kecerahan semu (*Asteroid Magnitude*), diameter dalam kilometer (*Asteroid Diameter (km)*), serta energi tumbukan dalam megaton (*Impact Energy (Mt)*). Di samping itu, digunakan pula parameter orbit penting seperti eksentrisitas (*Orbit Eccentricity*), inklinasi (*Orbit Inclination (deg)*), sumbu orbit (*Orbit Axis (AU)*), dan jarak minimum perpotongan orbit dengan Bumi (*Minimum Orbit Intersection Distance (AU)*). Kombinasi variabel-variabel ini memungkinkan pemodelan risiko tumbukan yang mempertimbangkan sekaligus aspek fisik dan dinamika gerak asteroid.

Untuk menjaga kualitas data, keberadaan nilai hilang dan duplikasi diperiksa kembali pada variabel-variabel kunci tersebut. Hanya observasi dengan informasi lengkap yang dipertahankan dalam dataset akhir, sementara penanda objek asteroid ditinjau untuk menghindari kemunculan nama ganda yang dapat mengganggu analisis. Dengan cara ini, dataset yang digunakan diharapkan merefleksikan kondisi fisik dan orbit asteroid secara konsisten sehingga hasil pemodelan risiko tumbukan menjadi lebih optimal.

### C. PREPROCESSING DATA

Sebelum pemodelan, data gabungan terlebih dahulu melalui beberapa tahap prapengolahan agar kualitasnya memadai. Langkah pertama adalah menelusuri keberadaan nilai hilang pada setiap variabel dan menghitung jumlahnya per kolom. Observasi yang tidak memiliki informasi lengkap pada variabel-variabel kunci dikeluarkan dari dataset untuk mengurangi bias dan ketidakpastian akibat banyaknya data tidak lengkap, sehingga analisis selanjutnya hanya menggunakan baris data yang utuh.

Berikutnya, penanda nama objek diperiksa untuk memastikan setiap asteroid hanya tercatat satu kali. Jika ditemukan lebih dari satu baris dengan nama objek yang sama, rekaman tersebut ditinjau dan disesuaikan sehingga satu asteroid direpresentasikan oleh satu entri yang

konsisten. Langkah ini penting untuk mencegah satu objek memberikan pengaruh ganda terhadap hasil pemodelan.

Setelah data bersih dari nilai hilang dan duplikasi, dipilih variabel-variabel yang paling relevan dengan tujuan penelitian, yaitu variabel yang menggambarkan potensi tumbukan (probabilitas kumulatif dan jumlah kemungkinan tumbukan), karakteristik fisik asteroid (kecepatan, magnitudo, diameter, energi tumbukan), serta parameter orbit utama (eksentrisitas, inklinasi, sumbu orbit, dan jarak minimum perpotongan orbit dengan Bumi). Pemilihan ini membantu menyederhanakan struktur data tanpa menghilangkan informasi esensial, sehingga diperoleh dataset yang bersih dan terstruktur untuk tahap pemodelan statistik.

### D. FEATURE ENGINEERING

Setelah data dibersihkan dan variabel utama dipilih, dilakukan tahap *feature engineering* (lihat Tabel 1) untuk memperkaya informasi yang masuk ke model dan menangkap pola yang tidak tampak jelas pada variabel asli. Langkah pertama adalah menghapus variabel yang berpotensi menimbulkan *data leakage*, yaitu kolom yang secara langsung merangkum tingkat risiko, seperti skala Palermo kumulatif dan maksimum, skala Torino maksimum, serta probabilitas kumulatif tumbukan. Jika variabel-variabel ini disertakan, model berisiko “mencuri jawaban” sehingga kinerjanya tampak terlalu optimistis dan tidak mencerminkan kemampuan prediktif yang sebenarnya.

Selanjutnya, sejumlah variabel fisik ditransformasikan dan dikombinasikan sebagaimana dirangkum pada Tabel 1. Transformasi logaritmik diterapkan pada jumlah kemungkinan tumbukan, diameter, kecepatan, dan energi tumbukan untuk meredam pengaruh nilai ekstrem dan menstabilkan sebaran data. Selain itu, dibentuk variabel interaksi yang menggabungkan diameter, kecepatan, jumlah kemungkinan tumbukan, dan jarak minimum perpotongan orbit (MOID) guna merepresentasikan bahwa risiko tumbukan ditentukan oleh kombinasi ukuran, kecepatan, dan frekuensi skenario tumbukan sekaligus. Sejumlah variabel juga

dipangkatkan dua (misalnya kecepatan, diameter, dan eksentrisitas) untuk menonjolkan perbedaan antara kondisi yang moderat dan yang benar-benar ekstrem. Seluruh fitur turunan tersebut kemudian diseleksi kembali, dan hanya variabel yang relevan secara fisik serta stabil secara statistik yang digunakan sebagai masukan akhir bagi model klasifikasi.

Table 1. Summary of Feature Engineering

Feature	Formula
log_impacts	$\log(\text{Possible Impacts})$
log_diameter	$\log(\text{Asteroid Diameter (km)})$
log_velocity	$\log(\text{Asteroid Velocity})$
log_energy	$\log(\text{Impact Energy (Mt)})$
diameter_x_velocity	$\text{Diameter} \times \text{Velocity}$
velocity_x_MOID	$\text{Velocity} \times \text{MOID}$
diameter_x_impacts	$\text{Diameter} \times \text{Possible Impacts}$
velocity_x_impacts	$\text{Velocity} \times \text{Possible Impacts}$
velocity2	$\text{Velocity}^2$
diameter2	$\text{Diameter}^2$
ecc2	$\text{Orbit Eccentricity}^2$
inclination2	$\text{Orbit Inclination}^2$
axis2	$\text{Orbit Axis}^2$
MOID2	$\text{MOID}^2$
perihelion	$q = a(1 - e)$
peri_aphe_ratio	$Q = a(1 + e)$
q_norm_axis	$q/a$
Q_norm_axis	$Q/a$
sin_inclination	$\sin(\text{Inclination} \times \pi/180)$
cos_inclination	$\cos(\text{Inclination} \times \pi/180)$

#### E. PEMILIHAN FITUR UNTUK MODEL

Setelah berbagai variabel asli dan turunan dibentuk pada tahap *feature engineering*, langkah berikutnya adalah memilih himpunan fitur yang benar-benar digunakan sebagai masukan model. Pemilihan ini didasarkan pada dua pertimbangan utama, yaitu relevansi substantif terhadap risiko tumbukan dan kelayakan statistik. Dari sisi substansi, hanya variabel yang secara logis berkaitan dengan besarnya risiko yang dipertahankan, seperti ukuran dan kecepatan asteroid, frekuensi skenario tumbukan, serta kedekatan lintasan orbit dengan Bumi. Dari sisi statistik,

diupayakan agar fitur yang dipilih tidak duplikatif secara informasi, tidak memuat indikator risiko yang sudah “jadi”, dan memiliki data yang relatif lengkap setelah proses pembersihan.

Secara garis besar, kandidat fitur dikelompokkan menjadi beberapa jenis. Pertama, variabel hasil transformasi logaritmik (log diameter, log kecepatan, log jumlah kemungkinan tumbukan, dan log energi tumbukan) yang membantu menstabilkan sebaran data dan membuat perbedaan antarasteroid lebih proporsional. Kedua, parameter orbit inti yang menggambarkan geometri lintasan, meliputi sumbu orbit, eksentrisitas, inklinasi, jarak minimum perpotongan orbit dengan Bumi, jarak perihelion dan aphelion, serta periode orbit tahunan. Ketiga, variabel interaksi yang menggabungkan karakteristik fisik dan kedekatan orbit, seperti diameter  $\times$  kecepatan, diameter  $\times$  jumlah kemungkinan tumbukan, kecepatan  $\times$  jumlah kemungkinan tumbukan, serta kombinasi besaran fisik dengan jarak minimum perpotongan orbit. Selain itu, digunakan pula sejumlah variabel polinomial (kuadrat eksentrisitas, inklinasi, sumbu orbit, kecepatan, diameter, dan jarak minimum perpotongan orbit) untuk menonjolkan perbedaan antara kondisi orbit yang moderat dan ekstrem.

erakhir, dipertahankan beberapa variabel turunan yang menggambarkan dinamika orbit secara lebih halus, seperti rasio perihelion–aphelion, komponen sinus dan kosinus dari sudut inklinasi, serta perihelion dan aphelion yang dinormalisasi terhadap sumbu orbit utama. Seluruh kandidat ini kemudian dirangkum ke dalam satu daftar variabel penjelas, dan hanya fitur yang tersedia secara konsisten di dataset bersih serta relevan secara fisik yang akhirnya digunakan dalam pemodelan. Dengan demikian, model regresi logistik dibangun di atas himpunan fitur yang kaya informasi, relevan secara fisik, bebas dari kebocoran informasi risiko, dan cukup stabil secara statistik untuk klasifikasi risiko tumbukan asteroid.

#### F. TRAIN-TEST SPLIT

Setelah himpunan fitur dan variabel target dibentuk, data dibagi menjadi dua bagian, yaitu data latih (*training set*) dan data uji (*test set*). Sekitar 80% observasi dialokasikan sebagai data latih untuk membangun dan menyetel model, sedangkan sisanya digunakan sebagai data uji yang disimpan terpisah dan hanya dipakai pada tahap akhir sebagai tolok ukur kinerja model. Pemisahan ini penting untuk menilai kemampuan generalisasi model pada data yang belum pernah dilihat sebelumnya dan menghindari penilaian yang terlalu optimistis.

Pembagian dilakukan secara berstrata dengan mempertahankan proporsi asteroid berisiko tinggi dan rendah pada data latih dan data uji agar tetap sebanding dengan distribusi awal. Nilai bilangan acak awal diatur tetap sehingga proses *train-test split* dapat direplikasi. Setelah pembagian, distribusi kelas pada kedua subset diperiksa kembali untuk memastikan bahwa data latih tetap representatif dan bahwa evaluasi pada data uji benar-benar mencerminkan kemampuan model dalam membedakan kedua kelompok risiko.

#### G. SCALING

Setelah data dibagi menjadi kelompok latih dan uji, seluruh variabel prediktor numerik diseragamkan skalanya melalui proses *scaling*. Tujuannya adalah agar setiap variabel berada pada rentang nilai yang sebanding, sehingga tidak ada satu variabel pun yang mendominasi pemodelan hanya karena perbedaan satuan atau orde besaran. Tanpa penyetaraan skala, variabel dengan nilai sangat besar, seperti jarak orbit dalam satuan astronomi, dapat menutupi pengaruh variabel lain yang juga penting, seperti diameter atau kecepatan asteroid.

Dalam penelitian ini digunakan pendekatan *standardization*, yaitu mengubah nilai tiap variabel numerik pada data latih sehingga memiliki rata-rata nol dan simpangan baku satu. Parameter rata-rata dan simpangan baku dihitung hanya dari data latih, kemudian digunakan untuk menstandarkan data uji, sehingga tidak terjadi kebocoran informasi dari data uji ke proses pelatihan. Penyetaraan skala ini penting karena model yang digunakan adalah regresi logistik dengan regularisasi, yang sensitif

terhadap perbedaan skala antarfaktor penjelas. Melalui tahapan *scaling* ini, diperoleh himpunan prediktor yang lebih seimbang dan siap digunakan dalam pemodelan klasifikasi risiko tumbukan asteroid.

#### H. IMBALANCE HANDLING

Pada tahap berikutnya, perhatian diarahkan pada masalah ketidakseimbangan kelas pada variabel target. Secara alami, jumlah asteroid yang dikategorikan berisiko tinggi jauh lebih sedikit dibandingkan yang berisiko rendah. Jika kondisi ini dibiarkan, model cenderung “bermain aman” dengan hampir selalu memprediksi kelas berisiko rendah: nilai akurasi tampak tinggi, tetapi kemampuan mengenali asteroid berisiko tinggi justru rendah. Situasi ini tidak sejalan dengan tujuan penelitian, karena secara praktis lebih penting meminimalkan kegagalan mendeteksi objek yang berpotensi membahayakan Bumi.

Untuk mengurangi masalah tersebut, model dibuat lebih peka terhadap kesalahan pada kelas berisiko tinggi dengan memberikan bobot yang lebih besar pada misklasifikasi asteroid berisiko tinggi dibandingkan berisiko rendah. Dengan cara ini, algoritma “diingatkan” bahwa salah memprediksi asteroid berbahaya sebagai tidak berbahaya jauh lebih serius daripada kesalahan sebaliknya. Pendekatan pembobotan ini melengkapi *stratified train-test split* yang telah diterapkan sebelumnya, sehingga distribusi kelas yang tidak seimbang tetap realistis tetapi pengaruhnya terhadap proses pelatihan dan evaluasi dapat dikendalikan.

Selain itu, teknik *oversampling* pada data latih dipertimbangkan secara eksploratif untuk menambah representasi kelas minoritas, sehingga model memperoleh lebih banyak contoh asteroid berisiko tinggi selama pembelajaran. Setiap penyesuaian terhadap distribusi kelas dievaluasi secara hati-hati agar tidak menimbulkan pola yang terlalu artifisial. Melalui kombinasi pengaturan ini, diharapkan model tidak hanya mencapai kinerja global yang baik, tetapi juga memiliki kemampuan yang lebih seimbang dalam membedakan asteroid berisiko tinggi dan berisiko rendah.

### I. TRAINING LOGISTIC REGRESSION DAN HPO (GRIDSEARCH)

Pada tahap ini, penelitian memanfaatkan regresi logistik sebagai model utama untuk mengklasifikasikan asteroid ke dalam kelompok berisiko tinggi dan berisiko rendah. Pemilihan regresi logistik didasarkan pada dua pertimbangan utama. Pertama, model ini secara alami menghasilkan probabilitas bahwa suatu asteroid termasuk kelas berisiko tinggi, sehingga hasilnya mudah diinterpretasikan dalam konteks pengambilan keputusan. Kedua, koefisien regresi memberikan informasi langsung mengenai arah dan besarnya pengaruh setiap variabel penjelas terhadap peningkatan atau penurunan risiko tumbukan.

Pelatihan model dilakukan hanya pada data latih yang telah melalui proses *scaling* dan penanganan ketidakseimbangan kelas. Untuk mengurangi risiko *overfitting* dan meningkatkan kemampuan generalisasi, digunakan regresi logistik dengan regularisasi, yaitu penambahan penalti terhadap koefisien yang terlalu besar. Kekuatan penalti ini diatur oleh hiperparameter yang tidak ditetapkan secara sembarang, melainkan disetel melalui proses *hyperparameter optimization* dengan pendekatan *grid search*.

Dalam *grid search*, disusun sekumpulan kombinasi nilai hiperparameter yang dianggap masuk akal, kemudian masing-masing kombinasi diuji menggunakan validasi silang (*cross-validation*) pada data latih. Data latih dibagi menjadi beberapa lipatan, model dilatih pada sebagian lipatan dan diuji pada lipatan sisanya, lalu kinerja berdasarkan akurasi dan luas kurva ROC (AUC) dirata-ratakan. Kombinasi hiperparameter dengan kinerja rata-rata terbaik dipilih sebagai konfigurasi akhir. Model regresi logistik kemudian dilatih ulang pada seluruh data latih dengan pengaturan tersebut, sehingga dihasilkan satu model akhir yang siap dievaluasi pada data uji dan digunakan untuk menilai kemampuan klasifikasi serta peran masing-masing fitur dalam menjelaskan risiko tumbukan asteroid..

## III. Results and Discussion

### A. OVERALL MODEL PERFORMANCE

Table 1. Classification Report

Metric	Value
Accuracy	0.8676
95% CI for accuracy	0.7989 – 0.9196
No Information Rate	0.6618
p-value (Accuracy > NIR)	$4.01 \times 10^{-8}$
Kappa	0.7193
McNemar's test p-value	0.0339
Sensitivity	0.8444
Specificity	0.9130
Positive predictive value (PPV)	0.9500
Negative predictive value (NPV)	0.7500
Prevalence of positive class	0.6618
Detection rate	0.5588
Detection prevalence	0.5882
Balanced accuracy	0.8787
Area under ROC curve (AUC)	0.9473

Tabel 1 merangkum kinerja keseluruhan model regresi logistik pada data uji. Model menghasilkan akurasi 0,8676, jauh di atas *No Information Rate* sebesar 0,6618 dengan nilai  $p$   $4,01 \times 10^{-8}$ , sehingga peningkatan ini signifikan secara statistik. Interval kepercayaan 95% untuk akurasi (0,7989–0,9196) relatif sempit dan tetap berada di atas *No Information Rate*, yang mengindikasikan estimasi yang stabil. Selain itu, nilai Kappa 0,7193 dan *balanced accuracy* 0,8787 menunjukkan kesesuaian yang kuat antara prediksi model dan kelas aktual, tidak hanya mengikuti dominasi kelas mayoritas. Secara umum, hasil ini menegaskan bahwa model yang dibangun memberikan performa klasifikasi yang andal sebagai dasar untuk analisis lebih lanjut.

## B. CLASS-WISE PERFORMANCE & CONFUSION MATRIX

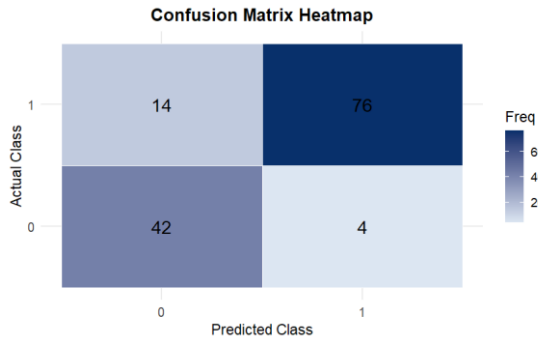


Figure 1. Confusion Matrix Heatmap

Kinerja model pada masing-masing kelas ditunjukkan oleh confusion matrix pada Gambar 1. Model berhasil mengklasifikasikan 76 asteroid berisiko tinggi secara benar (*true positive*) dan 42 asteroid berisiko rendah secara benar (*true negative*), sementara masih terdapat 14 asteroid berisiko tinggi yang diprediksi sebagai berisiko rendah (*false negative*) dan 4 asteroid berisiko rendah yang diprediksi berisiko tinggi (*false positive*). Pola ini konsisten dengan nilai sensitivitas 0,8444 dan spesifisitas 0,9130, di mana model cukup baik dalam mengenali kedua kelas, namun masih meloloskan sebagian kecil objek berisiko tinggi. Secara praktis, keberadaan *false negative* ini menunjukkan bahwa pemilihan ambang probabilitas perlu mempertimbangkan trade-off antara peningkatan sensitivitas dan jumlah peringatan yang dihasilkan.

## C. ROC CURVU AND DISCRIMINATIVE ABILITY

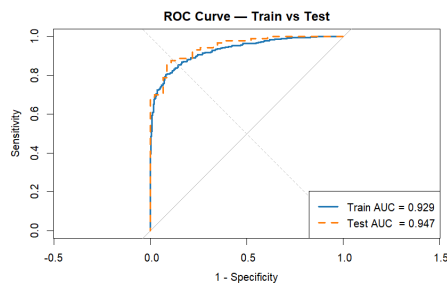


Figure 2. ROC Curve – Train vs Test

Kemampuan diskriminatif model diilustrasikan pada kurva ROC pada Gambar 1. Baik untuk data latih maupun data uji, kurva berada jauh di atas garis diagonal, yang menunjukkan bahwa model mampu membedakan asteroid berisiko tinggi dan rendah pada berbagai ambang probabilitas. Nilai AUC sebesar 0,929 pada data latih dan 0,947 pada data uji mengindikasikan kualitas klasifikasi yang sangat baik sekaligus tidak menunjukkan indikasi *overfitting*, karena performa pada data uji justru sedikit lebih tinggi. Dengan kata lain, model secara konsisten memberikan probabilitas lebih besar kepada asteroid yang benar-benar berisiko tinggi, sehingga layak dipertimbangkan sebagai alat bantu *screening* awal.

## D. IMPLICATION, COMPARATION, AND LIMITATION

Table 2. Top 5 Koefisien Fitur Logistic Regression

Feature	Coef
log_impacts	2.24677834
log_diameter	1.08292464
diameter_x_velocity	0.84225898
velocity2	0.75958655
velocity_x_MOID	0.56657751

Tabel 2 menampilkan lima koefisien positif terbesar dari model regresi logistik. Variabel *log\_impacts* memiliki koefisien paling besar, diikuti oleh *log\_diameter*, yang menunjukkan bahwa banyaknya skenario tumbukan yang tercatat dan ukuran asteroid merupakan penentu utama peningkatan risiko. Interaksi antara diameter dan kecepatan, kuadrat kecepatan (*velocity*<sup>2</sup>), serta kombinasi kecepatan dengan jarak minimum perpotongan orbit (*velocity* × *MOID*) juga berkontribusi positif, sehingga asteroid yang lebih besar, lebih cepat, dan melintas lebih dekat dengan orbit Bumi cenderung diklasifikasikan sebagai berisiko tinggi. Pola ini sejalan dengan penalaran fisik bahwa energi tumbukan dan kedekatan lintasan menjadi faktor kunci potensi bahaya.

Temuan tersebut mengindikasikan bahwa regresi logistik tidak hanya memberikan kinerja klasifikasi yang baik, tetapi juga mampu mengidentifikasi faktor-faktor risiko yang mudah diinterpretasikan oleh peneliti dan pengambil kebijakan. Meski demikian, masih ditemukannya *false negative* dan keterbatasan pada penggunaan parameter orbit yang bersifat statis menunjukkan bahwa model ini sebaiknya dipandang sebagai alat *screening* awal yang perlu dilengkapi dengan pemodelan dinamika orbit yang lebih rinci atau model lain yang lebih konservatif dalam mendeteksi asteroid berisiko tinggi.

#### IV. Conclusion

Penelitian ini mengembangkan model regresi logistik untuk mengklasifikasikan asteroid ke dalam kategori berisiko tinggi dan berisiko rendah dengan memanfaatkan dataset *Possible Asteroid Impacts with Earth*. Setelah melalui tahapan prapengolahan, *feature engineering*, penanganan ketidakseimbangan kelas, dan *scaling*, model yang dihasilkan menunjukkan kinerja yang baik dengan akurasi 0,8676, *balanced accuracy* 0,8787, dan nilai AUC 0,947 pada data uji. Hasil ini mengindikasikan kemampuan diskriminasi yang kuat antara kedua kelas dan menegaskan bahwa pendekatan statistik yang relatif sederhana dapat memberikan performa klasifikasi yang kompetitif. Analisis koefisien menunjukkan bahwa banyaknya skenario tumbukan yang tercatat, diameter asteroid, kecepatan, serta kombinasi kecepatan dengan jarak minimum perpotongan orbit merupakan faktor utama yang meningkatkan probabilitas risiko.

Meskipun demikian, keberadaan sejumlah *false negative* dan keterbatasan pada penggunaan parameter fisik dan orbit yang bersifat statis menunjukkan bahwa model ini belum memadai sebagai satu-satunya sistem peringatan dini. Model regresi logistik yang disusun lebih tepat diposisikan sebagai alat *screening* awal yang transparan dan mudah diinterpretasikan untuk membantu memprioritaskan asteroid potensial berbahaya. Penelitian selanjutnya disarankan untuk mengeksplorasi penyesuaian ambang keputusan, mengintegrasikan informasi dinamis lintasan, serta membandingkan pendekatan ini

dengan model *machine learning* lain yang lebih kompleks guna *further* meningkatkan sensitivitas terhadap asteroid berisiko tinggi.

#### Acknowledgement

The authors must acknowledge any sources of funding that supported the research and may acknowledge the outside reviewers of their drafts. Please provide after passing the review process.

#### References

The references must consist of 80% from relevant and recent primary sources (such as article of journal or conference from last 5 years). Citations and References are adapted from IEEE style and using reference manager software (Mendeley, Zotero, etc). Guidelines to IEEE style used for in-text citations and references is available <http://www.ieee.org/documents/ieeecitationref.pdf> However, some adjustments are made to suit the need of CommIT. A quick guide as example is stated as follows. The samples below are taken randomly from various sources and for example purposes only. They are intended neither as ads and promos nor as the Editorial Board's viewpoint or preferences

- [1] D. Broman et al., "The company approach to software engineering project courses," IEEE Trans. Educ., vol. 55, no. 4, pp. 445–452, Nov. 2012.
- [2] "Computer science curriculum 2008—An interim revision of CS 2001," Interim Review Task Force, 2008.
- [3] E. Carr et al., "Living in the knowledge society community building project," in Proc. 40th ACM Tech. Symp. Comput. Sci. Educ., Chattanooga, TN, USA, 2009, Rev. 47, pp. 777–780.
- [4] W. Chung, E. Fox, S. Sheetz, and S. Yang, "LIKES: Educating the next generation of knowledge society builders," in Proc. AMCIS, San Francisco, CA, USA, 2009, pp. 1–10.
- [5] Bologna Working Group on Qualifications Frameworks, A Framework for Qualifications of the European Higher Education Area. Last accessed Jun. 2015. [Online]. Available: <http://www.bologna->



bergen2005.no/Docs/00Main\_doc/050218  
\_QF\_EHEA.pdf