Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.1148503
Vol. 11, No. 4, Agustus 2024, hlm. 771-776 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

IMPLEMENTASI ALGORITMA CATBOOST DAN SHAPLEY ADDITIVE EXPLANATIONS (SHAP) DALAM MEMPREDIKSI POPULARITAS GAME INDIE PADA PLATFORM STEAM

Mohammad Teddy Syamkalla[*1], Siti Khomsah[2], Yohani Setiya Rafika Nur[3]

[1,2,3]Institut Teknologi Telkom Purwokerto, Purwokerto
Email: [1]20110013@ittelkom-pwt.ac.id, [2]siti@ittelkom-pwt.ac.id, [3]yohani@ittelkom-pwt.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 12 Februari 2024, diterima untuk diterbitkan: 09 Agustus 2024)

Abstrak

Meningkatnya popularitas game indie di pasar *game* mewajibkan para pengembang *game indie* bersaing untuk membuat *game* nya diminati oleh para pengguna dengan berbagai cara agar dapat meningkatkan potensi popularitasnya. Penelitian sebelumnya telah mencoba menggunakan algoritma *logistic regression* dan *random forest* untuk meramalkan popularitas game indie di *platform* Steam, namun hasil model menggunakan berbagai macam metode masih rendah. Selain itu masih belum memberikan pengetahuan yang cukup kepada pengembang tentang apa yang mempengaruhi popularitasnya.Karena data *game indie* yang diambil dari *platform* steam yang digunakan dalam studi ini memiliki tipe kategorikal dan *non-linear*, maka digunakan pendekatan lain dengan memanfaatkan Algoritma *CatBoost* yang dalam beberapa penelitian lain terbukti memiliki kinerja dan kemampuan yang lebih baik dalam menangani data kategorikal dan non-linear. Metode *Shapley Additive Explanations (SHAP*) juga digunakan untuk mengartikan kontribusi dan pengaruh dari setiap fitur terhadap hasil prediksi. Hasil evaluasi pada data *game indie* dari *platform* steam hasil scraping yang terdiri dari 52627 baris dan 11 fitur menunjukkan bahwa model *CatBoost* memiliki akurasi 81%, presisi 0.83, *recall* 0.77, *F1-score* 0.80 menunjukkan kemampuan model yang seimbang dalam membedakan kelas popularitas. Hal tersebut didukung dengan nilai *AUC* 0.88 dimana kurva cenderung mendekati 90 derajat. Metode *SHAP* mengungkapkan pengaruh fitur terhadap hasil prediksi. Keberadaan kategori *steam trading cards*, genre *RPG* dan kompartibel pada sistem operasi *mac* akan meningkatkan popularitas. Hal tersebut juga terjadi pada semakin tinggi harga dan *achievements* yang disediakan. Namun keberadaan genre *casual* akan mengurangi popularitas. Dengan hasil penelitian ini diharapkan dapat membantu pengembang *indie* dalam mengetahui faktor yang berkemungkinan mempengaruhi popularitas game mereka

Kata kunci: *prediksi, popularitas game indie. CatBoost, SHAP, steam*

IMPLEMENTATION OF CATBOOST AND SHAPLEY ADDITIVE EXPLANATIONS (SHAP) ALGORITHMS IN PREDICTING THE POPULARITY OF INDIE GAMES ON THE STEAM PLATFORM

Abstract

The increasing popularity of indie games in the gaming market requires indie game developers to compete to make their games attractive to users in various ways in order to increase their potential popularity. Previous research has tried to use logistic regression and random forest algorithms to forecast the popularity of indie games on the Steam platform, However, the model results using various methods are still low. Since the indie game data taken from the steam platform used in this study is categorical and non-linear, another approach is used by utilizing the CatBoost Algorithm which in several other studies has proven to have better performance and ability in handling categorical and non-linear data. The Shapley Additive Explanations (SHAP) method is also used to interpret the contribution and influence of each feature to the prediction results. Evaluation results on indie game data from the steam platform scraping results consisting of 52627 rows and 11 features show that the CatBoost model has 81% accuracy, precision 0.83, recall 0.77, F1-score 0.80 indicating a balanced model ability in distinguishing popularity classes. This is supported by the AUC value of 0.88 where the curve tends to approach 90 degrees. The SHAP method reveals the influence of features on prediction results. The existence of steam trading cards category, RPG genre and compatibility on mac operating system will increase the popularity. This also happens with the higher prices and achievements provided. However, the presence of the casual genre will reduce

popularity. With the results of this study, it is hoped that it can help indie developers in knowing the factors that are likely to affect the popularity of their games.

Keywords: *prediction, popularity of indiee game, CatBoost, SHAP, steam*

1. PENDAHULUAN

Istilah *Indie* atau indi dalam Kamus Besar Bahasa Indonesia (KBBI) merujuk kepada perusahaan yang bersifat kecil dan mandiri. Awalnya, istilah ini sering digunakan dalam industri perfilman dan musik untuk menggambarkan karya yang diproduksi di luar sistem studio besar. Seiring waktu, industri game juga mengadopsi istilah ini untuk membedakan pengembang game yang tidak tergantung pada studio besar. Meskipun konsep ini relatif baru dalam sejarah *game* digital, "indie" telah menjadi istilah umum yang digunakan untuk mengidentifikasi jenis *game* digital dan pengembang tertentu (Parker, 2013). Platform layanan digital, tempat para pengembang *game indie* mempublikasikan karya mereka, semakin populer di antara para penggemar. Namun, di antara berbagai platform tersebut, Steam muncul sebagai salah satu yang paling diminati.

Steam, yang merupakan platform distribusi game yang dikembangkan oleh *Valve Corporation*, telah menjadi tempat utama untuk jual-beli permainan video game di seluruh dunia. Hingga tahun 2022, lebih dari 10.000 permainan telah diterbitkan di platform ini (*steamspy*, 2022). Pertumbuhan signifikan terjadi setelah diperkenalkannya proyek *greenlight* yang memungkinkan pengembang merilis permainan tanpa memerlukan penerbit. Meskipun proyek tersebut kini telah dihapus dan digantikan oleh *Steam Direct.*

Peningkatan jumlah game indie yang dirilis mengakibatkan persaingan yang semakin ketat bagi para pengembang. Antusiasme dalam mengembangkan *game indie* tercermin dari jumlah pengembang yang terus meningkat. Sebagai contoh, pada tahun 2018, ratusan studio indie muncul di seluruh Spanyol, meskipun hanya sedikit yang berhasil meraih keuntungan dari karya mereka (Stefen T. Wright, 2018)..

Tantangan utama bagi pengembang game indie adalah bagaimana membuat produk mereka populer di pasar. Oleh karena itu, model prediksi popularitas *game indie* pada platform Steam menggunakan algoritma *CatBoost* dengan kinerja tinggi menjadi krusial sebelum merilis game ke Steam. Dengan mengetahui fitur yang mempengaruhi popularitas *game indie* melalui Metode *SHAP*, para pengembang dapat mengoptimalkan strategi mereka dan meningkatkan potensi kesuksesan serta popularitas game mereka di platform Steam secara lebih efisien.

2. METODE PENELITIAN

Langkah metode penelitian yang dilakukan pada studi ini terinspirasi dari (Permatasari, 2022) yang dimulai dari Pengumpulan data, lalu melakukan *Exploratory Data Analysis* (*EDA)* untuk mengetahui insight dan informasi terkait pada data. Lalu dilanjutkan dengan *Preprocessing Data* agar data yang digunakan terhindar dari *noise* yang tidak diperlukan. *Feature Engineering* juga dilakukan agar data terhindar dari *imbalanced,* dan dapat digunakan untuk tujuan awal. Selanjutnya dilakukan Evaluasi dan Analisis Hasil setelah dilakukan pemodelan menggunakan *CatBoost* dan *Shapley Additive Analysis (SHAP).*

3. TINJAUAN PUSTAKA

Tidak banyak penelitian yang membahas tentang prediksi popularitas *game indie* yang pernah dilakukan sebelumnya. Prediksi popularitas game indie menggunakan algoritma *logistic regression* dan *random forest* (Jiang & Wang, 2021). Namun, performa model masih rendah meskipun telah dilakukan penyetelan *hyperparameter*. Penelitian lain (Ibrahim et al., 2020) yang membandingkan *CatBoost* dengan beberapa algoritma lain menyimpulkan bahwa *CatBoost* memiliki kinerja yang lebih baik dalam pembuatan model prediksi.

Pemilihan algoritma machine learning harus disesuaikan dengan karakteristik data yang digunakan. Data di platform Steam mencakup ribuan game dengan berbagai fitur, seperti nama, tanggal rilis, perkiraan pemilik, usia minimal, harga, bahasa, situs web, kategori, genre, dan tag, dengan mayoritas data bersifat kategorikal dan numerik. CatBoost dipilih karena kemampuannya menangani kedua jenis data tersebut tanpa mengurangi kecepatan dan efisiensi model (Louis Owen, 2022).

Metode *Shapley Additive Explanations (SHAP)* digunakan untuk mengidentifikasi fitur yang berpengaruh pada hasil prediksi model. *SHAP* menyediakan pendekatan kuat untuk mengungkap kontribusi setiap fitur terhadap hasil prediksi (Rodríguez-Pérez & Bajorath, 2020). Penelitian sebelumnya (Permatasari et al., 2022; L. Wang et al., 2021) menunjukkan bahwa penerapan *SHAP* pada *CatBoost* mampu memberikan interpretasi kontribusi fitur yang baik pada prediksi diabetes melitus dan Analisis Stabilitas Seismik yang Efisien pada Lereng Tanggul yang Mengalami Perubahan Muka Air.

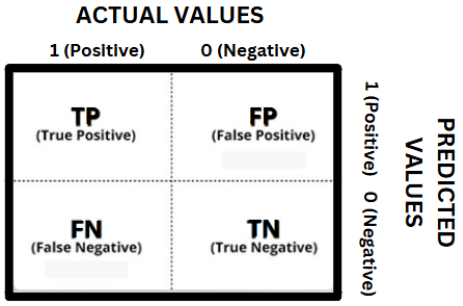
3.1 Scraping Data

Scraping adalah teknik otomatisasi untuk memperoleh informasi dari situs web tanpa perlu menyalinnya secara manual. Tujuan utama dari *scraping* adalah mengumpulkan data spesifik (Deviacita, 2019). Teknik ini menitikberatkan pada pengambilan dan ekstraksi data. Proses pengambilan Teknik *scraping* berasal dari internet, umumnya

berupa halaman web dalam bahasa *markup* seperti HTML atau XHTML, dan menganalisis dokumen tersebut untuk diambil data tertentu dari halaman tersebut untuk digunakan bagi kepentingan lain.

3.2 Matriks Konfusi

Menilai performa suatu sistem klasifikasi menjadi hal yang penting karena hal tersebut dapat mengindikasikan sejauh mana kemampuan sistem dalam mengklasifikasikan data (Baldi, 2000). Salah satu metode yang umum digunakan untuk mengukur kinerja klasifikasi adalah Matriks Konfusi. Matriks Konfusi adalah tabel yang mencatat hasil dari proses klasifikasi. Berikut adalah contoh tabel Matriks Konfusi untuk sistem klasifikasi dengan dua kelas yang ditunjukan pada Gambar 1:



Gambar 1. Matriks Konfusi (Caelen, 2017)

Akurasi, Presisi, *Recall*, dan *F1-score* merupakan beberapa matrik evaluasi yang digunakan untuk mengukur keakuratan model membedakan kelas dalam melakukan prediksi atau klasifikasi. Merupakan bagian dari matriks konfusi yang secara perhitungan ditunjukkan persamaan dibawah:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Presisi = \frac{TP}{TP + FP} \quad (2)$$

$$Recalll = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - score = 2 \times \frac{Presisi + Recall}{Presisi + Recall} \quad (4)$$

dimana nilai *TP*,*TN*,*FP*, dan *FN* pada persamaan diatas adalah :

a. *TP* (*True Positive*) merupakan jumlah observasi yang benar – benar termasuk ke dalam kelas positif dan diprediksi benar oleh model.
b. *TN* (*True Negative*) merupakan jumlah observasi yang benar – benar termasuk ke dalam kelas negatif dan diprediksi dengan benar oleh model.
c. *FP (False Positive*) merupakan jumlah observasi yang diprediksi sebagai bagian dari kelas positif oleh model, tetapi sebenarnya observasi tersebut termasuk dalam kelas negatif.
d. *FN* (*False Negative*) merupakan jumlah observasi yang diprediksi sebagai bagian dari kelas negatif oleh model, tetapi sebenarnya observasi tersebut termasuk dalam kelas positif.

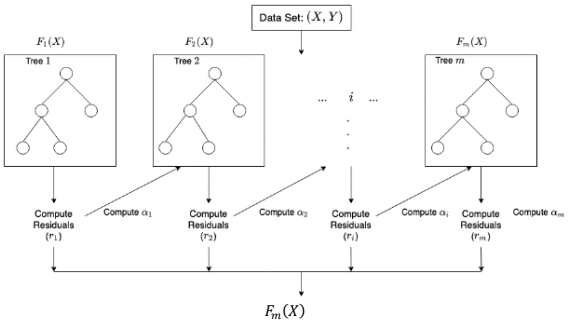
3.3 Exploratory Data Analysis

Tahap *Exploratory Data Analysis (EDA)* dilaksanakan untuk mendapatkan pemahaman mendalam mengenai data yang akan digunakan, termasuk struktur data, identifikasi nilai yang hilang, dan peninjauan distribusi data untuk mengenali adanya *outlier*. *EDA* dijalankan sebelum membangun model prediktif atau melakukan analisis lebih lanjut.

Proses *EDA* dapat memberikan wawasan tentang struktur data, termasuk jumlah baris dan fitur, jenis data dari setiap fitur, serta deteksi keberadaan data yang hilang atau duplikat. Selain menggambarkan struktur data, analisis statistik deskriptif juga dilakukan untuk memahami distribusi data. Informasi seperti jumlah data, rata-rata, deviasi standar, kuartil pertama, kedua, dan ketiga, nilai minimum, serta nilai maksimum dari fitur numerik akan diungkapkan melalui analisis statistik deskriptif (G. Wang, 2022).

3.4 Modelling CatBoost

Dalam tahap ini, peneliti menggunakan algoritma *CatBoost* untuk membuat model prediksi popularitas *game indie* dengan menggunakan data yang telah diolah pada langkah sebelumnya. Algoritma *CatBoost* dipilih karena memiliki kemampuan yang lebih unggul dalam memprediksi, terutama pada data yang memiliki fitur kategorikal, dibandingkan dengan algoritma lain yang telah digunakan penelitian sebelumnya pada (Ibrahim., 2020) dan (Mohamed Saber, 2021). Algoritma *CatBoost* mengameplementasikan *Gradient Boosting Decision Tree (GBDT)* untuk melakukan prediksi. *Gradien Boosting* merupakan suatu teknik ensemble yang melibatkan pembentukan model akhir dengan menggabungkan beberapa model indievidual, dengan tujuan untuk mengoptimalkan fungsi kerugian (*loss function*). Ilustrasi cara kerja Gradien Boosting sendiri ditunjukan pada Gambar 2 berikut :



Gambar 2. Cara Kerja Gradien Boosting (Amazon SageMaker, 2023)

Pada Gambar 2, $F_m(X)$ merupakan prediksi model final, prediksi final didapatkan menggunakan persamaan (5):

$$F_m(X) = F_{m-1}(X) + \alpha_m h_m(X, r_{m-1}) \quad (5)$$

α_i merupakan parameter learning_rate, dan r_{m-i} adalah residual dari model sebelumnya dan h_i model residual ke i.

Berbeda dengan jenis Gradient boosting lainnya seperti *XGBoost* dan *LightGBM* yang menggunakan pohon yang tidak simetris, *CatBoost* menggunakan pohon simetris untuk mencapai kecepatan prediksi yang optimal. Keputusan ini memberikan beberapa keunggulan yang sangat berguna, seperti kemampuan mengurangi risiko *overfitting*, waktu pelatihan yang lebih singkat, dan peningkatan efisiensi penggunaan *GPU* (Liudmila Prokhorenkova, 2018.). Keuntungan-keuntungan ini muncul karena *CatBoost* menggunakan kondisi yang seragam di setiap nodenya dalam pembentukan pohon keputusan, sehingga setiap bagian dari pohon keputusan yang dihasilkan menggunakan algoritma yang lebih efisien dalam mengurangi kesalahan dibandingkan dengan bagian sebelumnya (Louis Owen, 2022).

Selain keunggulan yang telah dijelaskan sebelumnya, keunggulan utama dari algoritma ini terletak pada kemampuannya untuk secara otomatis mengatasi berbagai jenis data, termasuk data numerik, teks, dan khususnya data kategorikal (Anna Veronika Doroogush, 2018). Dengan hanya mendefinisikan fitur-fitur yang bersifat kategorikal, *CatBoost* secara bawaan akan mengubah data pada fitur-fitur kategorikal tersebut menjadi data numerik, (Liudmila Prokhorenkova, 2018):

$$ctr_i = \frac{countInClass + prior}{totalCount + 1} \quad (6)$$

Pada persamaan 6, ctr_i merupakan data ke-i pada fitur kategorikal. $countInClass$ menunjukkan berapa kali nilai label melebihi i untuk objek dengan nilai fitur kategorikal pada saat ini (Veronika Dorogush, 2018). Fungsi ini hanya menghitung objek yang sudah memiliki nilai ini, dan perhitungan dilakukan berdasarkan urutan objek setelah dilakukan pengocokan. $totalCount$ adalah jumlah total objek yang memiliki nilai fitur yang sesuai dengan nilai fitur pada saat ini. Sedangkan $prior$ adalah suatu angka konstan yang ditetapkan oleh parameter awal.

Beberapa pengaturan seperti jumlah maksimum iterasi yang digunakan, kedalaman maksimum dari Pohon Keputusan, dan jumlah maksimum kombinasi fitur kategorikal untuk meningkatkan kinerja model juga dapat diterapkan pada algoritma ini (Liudmila Prokhorenkova, 2018.).

3.5 Shapley Additive Explanations (SHAP)

Pendekatan S*hapley Additive Explanations (SHAP)* adalah metode yang memungkinkan untuk menginterpretasikan model prediksi pembelajaran mesin yang bersifat *blackbox* atau sulit dipahami (Bahador Parsa, 2019). Tujuan dari *SHAP* adalah untuk menjelaskan prediksi dari fitur x dengan menghitung kontribusi dari setiap fitur terhadap prediksi(Kannangara, 2022). Metode *SHAP* menghitung nilai Shapley mirip dengan teori permainan koalisi (Permatasari, 2022). Nilai-nilai fitur dari sebuah contoh data bertindak sebagai pemain dalam sebuah koalisi. Nilai *Shapley* mendistribusikan kontribusi prediksi secara adil di antara fitur-fitur. Seorang pemain dapat berupa nilai fitur individual, memungkinkan pemahaman yang lebih mendalam tentang faktor-faktor yang memengaruhi hasil prediksi.

pendekatan *Shapley Additive Explanations* (*SHAP*) menggunakan rumus yang ditunjukan pada Persamaan (6):

$$\phi_i = \sum_{S \subseteq \{1,\dots,p\}\{i\}} \frac{|S|!(p-|S|-1)!}{p!} \times [Val(S \cup \{i\}) - Val(S)] \quad (7)$$

ϕ_i adalah nilai shapley dari anggota suatu fitur terhadap hasil prediksi. $Val(S)$adalah output dari model ML yang akan dijelaskan menggunakan satu set fitur S, dan p adalah jumlah keseluruhan dari semua fitur.

Kontribusi akhir atau nilai Shapley dari fitur $i(\phi i)$ didefinisikan juga sebagai rata-rata dari marginal kontribusinya di seluruh permutasi yang mungkin dari set fitur . Proses ini menggunakan perhitungan nilai Shapley dengan memasukan semua kemungkinan kombinasi fitur dan mengukur bagaimana kontribusi setiap fitur berubah saat fitur-fitur lain berubah.

4. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan data primer yang berbentuk ekstensi json, yang diperoleh melalui proses pengambilan data dari halaman web Steam menggunakan bahasa pemrograman Python. Teknik pengumpulan data yang diterapkan dalam penelitian ini adalah teknik pengambilan data dengan teknik *scraping*, yaitu metode pengumpulan menarik semua data pada situs web dengan cara mengambil informasi dari tampilan halaman web menggunakan perintah wget! dan kemudian menyimpannya dalam format yang dapat diakses oleh komputer dengan nama games.json (Deviacita, 2019). Pemilihan teknik *scraping dalam* mengambil data dilakukan daripada teknik *crawling* karena terdapat beberapa fitur yang tidak bisa didapatkan apabila menggunakan teknink *crawling.*

Selanjutnya data yang telah terhimpun disortir hingga hanya mencakup *game indie* dari tahun 2012 dan fitur yang relevan. Setelah itu, data yang telah disaring dijadikan sebagai dataset baru yang akan menjadi fokus utama dalam penelitian ini.

Sementara itu, untuk langkah berikutnya, hanya akan menggunakan beberapa fitur yang memiliki tipe data numerik dan kategorikal, sehingga terdapat 52627 baris data dan 11 fitur yang akan digunakan. Fitur-fitur tersebut adalah *name*, *release_date*, *required_age*, *price*, *windows*, *mac*, *linux*, *achievements*, *category*, *genre*,*supported_languages*, *full_audio_languages*, dan E*stimated_owners.* Fitur *Estimated_owners* akan berfungsi sebagai variabel

target dalam penelitian ini karena mencerminkan tingkat popularitas suatu game pada data yang tersedia ditunjukan pada Tabel 1.

Tabel 1. Fitur dataset *game indie*

Fitur	Tipe data	Deskripsi
required_age	Numerikal	Usia minimal untuk memainkan game
Price	Numerikal	Harga game
windows	Boolean	Dapat berjalan di sistem windows
Mac	Boolean	Dapat berjalan di sistem mac
Linux	Boolean	Dapat berjalan di sistem linux
achievements	Numerikal	Jumlah penghargaan yang bisa di dapatkan pemain
category	Kategorikal	Kategori game
Genre	Kategorikal	Genre game
supported_ languages	Kategorikal	Bahasa teks yang tersedia
full_audio_ languages	Kategorikal	Bahasa audio yang tersedia
Estimated_ owners	Kategorikal	Range dari total pembeli game

Tahap *Exploratory Data Analysis (EDA)* dilakukan untuk mendapatkan pemahaman mendalam mengenai data yang akan digunakan, termasuk struktur data, identifikasi nilai yang hilang, dan peninjauan distribusi data untuk mengenali adanya outlier. *EDA* dijalankan sebelum membangun model prediktif atau melakukan analisis lebih lanjut.

Proses *EDA* dapat memberikan wawasan tentang struktur data, termasuk jumlah baris dan fitur, jenis data dari setiap fitur, serta deteksi keberadaan data yang hilang atau duplikat yang ditunjukkan pada tabel 2.

Tabel 2. Struktur data

column	*null_count*	*Dtype*
required_age	0	Integer
Price	0	Float
windows	0	Boolean
Mac	0	Boolean
Linux	0	Boolean
achievements	0	Integer
category	47	Object
Genre	0	Object
supported_languages	10	Object
full_audio_languages	30622	Object
Estimated_owners	0	Object

Pada table 2 diketahui bahwa data masih memiliki nilai *null* pada beberapa kolom, dan juga terdiri dari 2 kolom bertipe *integer*, 1 *float*, 3 *boolean*, dan 5 *object*

Selain menggambarkan struktur data, analisis statistik deskriptif juga dilakukan untuk memahami distribusi data. Informasi seperti jumlah data, rata-rata, deviasi standar, kuartil pertama, kedua, dan ketiga, nilai minimum, serta nilai maksimum dari fitur numerik akan diungkapkan melalui analisis statistik deskriptif yang ditunjukan pada tabel 3.

Tabel 3. Statistik deskriptif

	required_age	price	achievements
count	52657.00	52657.00	52657.00
mean	0.16	6.68	23.32
std	1.63	11.08	183.31
min	0.00	0.00	0.00
25%	0.00	0.99	0.00
50%	0.00	4.99	7.00
75%	0.00	9.99	21.00
max	21.00	999.98	9821.00

Berdasarkan statistik deskriptif di tabel 3, diketahui bahwa jumlah data adalah 52.657. Tabel ini juga memberikan gambaran tentang tiga fitur numerik dalam data. Pertama, fitur required_age yang menunjukkan usia minimum yang diperlukan untuk memainkan sebuah gim, memiliki rata-rata sekitar 0,16 dengan deviasi standar sekitar 1,63. Rentang usia ini berkisar dari 0 hingga 21, menunjukkan variasi yang signifikan. Namun, kuartil pertama dan ketiga sama-sama bernilai 0, yang mengindikasikan bahwa sebagian besar data berada pada nilai minimum, menunjukkan bahwa mayoritas produk tidak memiliki batasan usia.

Kedua, fitur price yang mencerminkan harga produk, menunjukkan rata-rata sekitar 6,68 dengan deviasi standar 11,08. Harga minimum adalah 0, sementara harga maksimum mencapai 999,98, menunjukkan variasi yang luas. Kuartil pertama berada pada 0,99, kuartil kedua sebesar 4,99, dan kuartil ketiga mencapai 9,99. Ini menunjukkan bahwa sebagian besar harga cenderung rendah, namun adanya harga maksimum yang tinggi (999,98) mengindikasikan adanya produk dengan harga yang sangat tinggi.

Terakhir, fitur achievements yang merepresentasikan jumlah prestasi yang bisa didapatkan dalam gim memiliki rata-rata sekitar 23,32 dengan deviasi standar yang tinggi yaitu 183,31. Distribusi data pada fitur ini sangat bervariasi, dengan nilai minimum 0 dan maksimum 9821. Kuartil pertama dan median adalah 0, sementara kuartil ketiga mencapai 21. Ini menunjukkan bahwa sebagian besar gim memiliki sedikit atau tidak ada prestasi sama sekali, namun ada beberapa gim dengan jumlah prestasi yang sangat tinggi.

Langkah selanjutnya adalah *Preprocessing Data*, karena data yang kurang berkualitas dapat menghasilkan model yang tidak optimal, dan keputusan yang baik selalu didasarkan pada data yang baik pula. Tahap awal pembersihan data melibatkan penanganan nilai yang hilang dan eliminasi redundansi data.

Pertama-tama, nilai yang hilang diatasi dengan mengubahnya menjadi "None". Hal ini dilakukan karena kekosongan nilai pada variabel data Steam bukanlah suatu kesalahan data, melainkan mencerminkan bahwa game tersebut memang tidak memiliki kriteria tertentu pada suatu fitur, seperti audio atau subtitle.

Selanjutnya, untuk mengurangi redundansi pada data kategorikal yang merupakan kelompok minoritas, dilakukan penghapusan pada entitas yang tidak termasuk dalam 10 teratas dari masing-masing fitur. Setelah itu, dilakukan eliminasi data duplikat

untuk memastikan setiap entitas game memiliki informasi yang unik pada setiap fiturnya.

Langkah selanjutnya dalam proses ini adalah pelabelan, yang melibatkan penentuan kelas berdasarkan hasil *EDA* yang ditunjukan pada tabel 3.

Tabel 3. Persebaran angka *Estimated_owners*

No	*Estimated_owners*	Frekuensi
1	0 - 999	3.406
2	1000 – 20.000	35.153
3	20.001 – 50.000	6.117
4	50.001 – 100.000	2.868
5	100.001 – 200.000	1.786
6	200.001 – 500.000	1.410
7	500.001 – 1.000.000	518
8	1.000.001 – 2.000.000	211
9	2.000.001 – 5.000.000	131
10	5.000.001 – 10.000.000	34
11	10.000.0001 – 20.000.000	12
12	20.000.0001 – 50.000.000	11

Fitur yang berfungsi sebagai label dalam penelitian ini adalah *estimated_owner* dan diberikan perlakuan transformasi nilai, yaitu nilai 0 untuk tingkat di bawah 20000, dan 1 untuk tingkat di atas 20000, terinspirasi dari (Jiang & Wang, 2021), dan juga berdasarkan nilai dari fitur estimated_owners agar tidak terlalu *imbalance* untuk labelnya.

Feature Engineering diterapkan untuk mengubah atau memodifikasi fitur-fitur yang ada dalam dataset. Proses ini mencakup pemisahan tipe kolom object menjadi kategorikal, bertujuan untuk memastikan bahwa data siap untuk digunakan dalam pemodelan dan tidak berbentuk *list*.

Untuk menangani ketidakseimbangan distribusi kelas dalam dataset, oversampling diterapkan menggunakan fungsi random oversampling. Hal ini dilakukan untuk mencapai keseimbangan antar kelas, karena ketidakseimbangan dapat mempengaruhi kinerja model yang akan dibangun, Hasil penanganan imbalanced data ditunjukkan pada Tabel 4.

Tabel 4. Jumlah label sesudah random oversampling

No	Label Popularitas	Frekuensi
1	0 (Popular)	39.312
2	1 (Tidak popular)	39.312

Feature Selection menjadi tahap penting dalam mengidentifikasi fitur-fitur yang akan digunakan oleh model untuk memprediksi target. Dengan menghapus fitur yang tidak relevan atau sudah diberi perlakuan sebelumnya pada dataset, peneliti dapat memastikan bahwa dataset telah diolah dengan baik dan menjadi dasar yang kokoh untuk memulai proses pemodelan ditunjukkan pada tabel 5.

Tabel 5. Fitur sebelum dan sesudah feature selection

No	Fitur sebelum *feature selection*	Fitur sesudah *feature selection*
1	*required_age, price, windows, mac, linux, achievements, supported_languages, full_audio_languages, categories, genres, popularity, genre_other, genre_Action, genre_Strategy, genre_Early Access, genre_Free to Play, genre_Racing, genre_Adventure, genre_Simulation, genre_Sports, genre_RPG, genre_Casual, category_Steam Cloud, category_Steam Trading Cards, category_Full controller support, category_other, category_Single-player, category_PvP, category_Steam Achievements, category_Partial Controller Support, category_Multi-player, category_Co-op, category_Steam Leaderboards, language_other, language_Japanese, language_Korean, language_Portuguese - Brazil, language_German, language_English, language_Russian, language_Italian, language_French, language_Spanish - Spain, language_Simplified Chinese, audio_other, audio_Japanese, audio_Portuguese - Brazil, audio_German, audio_English, audio_Traditional Chinese, audio_Spanish - Spain, audio_French, audio_Russian, audio_Simplified Chinese*	*genre_RPG, genre_Casual, category_Steam Cloud, category_Steam Trading Cards, category_Full controller support, category_other, category_Single-player, category_PvP, category_Steam Achievements, category_Partial Controller Support, category_Multi-player, category_Co-op, category_Steam Leaderboards, language_other, language_Japanese,language_Korean, language_Portuguese - Brazil, language_German,language_English, language_Russian, language_Italian, language_French, language_Spanish - Spain,language_Simplified Chinese, audio_other, audio_Japanese, audio_Portuguese - Brazil, audio_German, audio_English, audio_Traditional Chinese, audio_Spanish - Spain, audio_French, audio_Russian, audio_Simplified Chinese*

Proses pelatihan menggunakan 78624 baris data dan 50 kolom setelah melewati proses *preprocessing data* dan *feature engineering*, akan dilakukan pemodelan menggunakan algoritma *CatBoost* digunakan dala penelitian ini karena dapat menangani dua hal tersebut. *CatBoost* adalah algoritma penguatan gradien yang dirancang khusus untuk menangani berbagai jenis data, termasuk data dengan hubungan non-linear. *CatBoost* memanfaatkan teknik seperti *boosting* untuk membangun serangkaian model prediktif yang kuat.

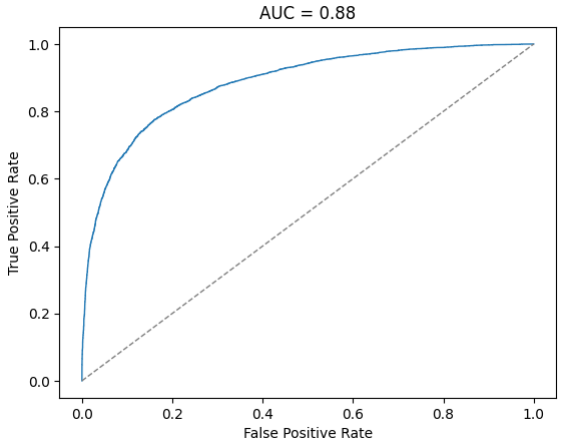
Model *CatBoost* yang digunakan pada data *testing* setelah dibagi sebesar 20% dari data *training* untuk memprediksi popularitas game indie menunjukkan kinerja yang yang baik pada data testing. Dengan tingkat akurasi sebesar 81% yang ditunjukan pada Tabel 6.

Tabel 6. Hasil penngujian model CatBoost

Label	Presisi	*Recall*	*F1-score*	*Support*
0	0.79	0.85	0.81	7911
1	0.83	0.77	0.80	7814
Accuracy			0.81	15725

Selain menggunakan metrik konfusi, evaluasi kinerja model juga dilakukan dengan mempertimbangkan Kurva *AUC* pada *ROC*. Kurva ini membantu mengukur sejauh mana model dapat membedakan antara kelas positif dan negatif. Secara umum, Kurva *AUC* pada *ROC* memberikan gambaran komprehensif tentang kinerja model pada berbagai

nilai *threshold*. Gambar 3. menampilkan hasil dari kurva *AUC* pada *ROC* yang dihasilkan dari pengujian model pada data testing.



Gambar 3. Kurva *AUC* pada *ROC* pemodelan

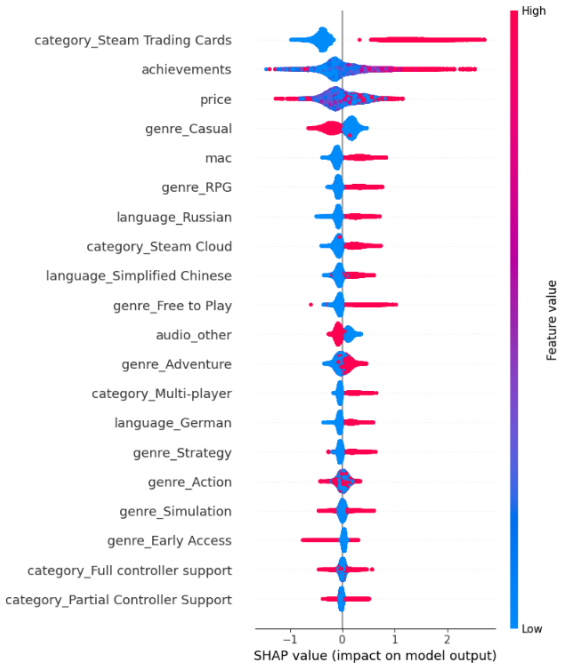
Gambar 3 menunjukkan kurva Area Under Curve (AUC) pada Receiver Operating Characteristic (ROC) dengan nilai sebesar 88%. Hal ini berarti bahwa model memiliki kemampuan yang baik dalam membedakan kelas-kelas yang ada dalam data. Kurva AUC yang mendekati nilai 1, atau dalam hal ini 88%, mengindikasikan bahwa model cenderung mendekati sudut 90 derajat pada grafik ROC, menunjukkan tingkat akurasi yang tinggi dalam klasifikasi. Dengan kata lain, model efektif dalam memisahkan antara kelas positif dan negatif, memberikan kepercayaan yang tinggi terhadap kinerja prediktifnya.

Setelah melakukan evaluasi model menggunakan metrik konfusi akurasi dan *AUC-ROC*, peneliti juga menerapkan metode *Shapley Additive Explanations (SHAP*) pada model *CatBoost* yang digunakan untuk memprediksi popularitas game *indie*. Tujuannya adalah untuk memperoleh pemahaman yang lebih mendalam mengenai kontribusi setiap fitur terhadap hasil prediksi model tersebut. Melalui analisis *SHAP*, hasil interpretasi nilai *Shapley* pada setiap fitur dalam menentukan prediksi popularitas *game indie* pada model *CatBoost* ditampilkan dalam Gambar 4.

Urutan fitur dari atas kebawah pada visualisasi *SHAP* merupakan urutan fitur dari yang paling berpengaruh hingga paling tidak berpengaruh yang diukur berdasarkan jumlah besaran nilai *shapley* pada semua instansi, dan konsistensi nilai *shapley* dari setiap instansi fitur terhadap hasil prediksi model dari Gambar 4.

Dapat dilihat bahwa fitur yang paling berpengaruh terhadap prediksi model adalah *category_Steam Trading Cards, achievements, price, genre_Casual, mac, genre_RPG*, dan seterusnya.

Nilai *Shapley* pada *category_Steam Trading Cards* menunjukkan bahwa keberadaan kartu perdagangan Steam dalam *game indie* memiliki kontribusi positif yang signifikan terhadap popularitasnya. Hal ini mungkin disebabkan oleh tingginya minat pemain terhadap fitur ini, yang memungkinkan mereka mengumpulkan dan memperdagangkan kartu, sehingga menambah nilai replay dan interaktivitas dalam permainan.



Gambar 4. *SHAP* pemodelan

Demikian juga, fitur *achievements* memiliki nilai *Shapley* yang tinggi, yang menunjukkan bahwa sistem pencapaian dalam *game indie* memberikan kontribusi positif terhadap popularitasnya. Pencapaian ini dapat meningkatkan keterlibatan pemain dengan memberikan tujuan tambahan dan tantangan dalam permainan. Selain itu, *price* atau harga game juga berkontribusi positif, menandakan bahwa *game indie* dengan harga yang kompetitif atau diskon cenderung lebih populer.

Ketersediaan *game* di *platform mac* juga memiliki nilai *Shapley* yang tinggi, menunjukkan bahwa ketersediaan *game* untuk pengguna *Mac* dapat meningkatkan popularitasnya. Hal ini mungkin karena basis pengguna Mac yang setia yang menghargai game yang kompatibel dengan perangkat mereka. *genre_RPG* atau genre permainan peran juga menunjukkan kontribusi positif, mengindikasikan bahwa *game indie* dengan genre *RPG* cenderung lebih diminati oleh pemain karena alur cerita yang mendalam dan pengalaman bermain yang kompleks.

Namun, tidak semua fitur memiliki dampak positif. Misalnya, *genre_Casual* menunjukkan nilai *Shapley* negatif, yang berarti bahwa *game indie* dengan genre kasual cenderung kurang populer dibandingkan dengan genre lain. Hal ini mungkin disebabkan oleh persepsi bahwa game kasual kurang menantang atau mendalam dibandingkan genre lain, sehingga menarik audiens yang lebih kecil.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian, peneliti meraih beberapa kesimpulan yang signifikan. Pertama, model algoritma CatBoost untuk prediksi popularitas game indie berhasil mencapai tingkat akurasi sebesar 81%. Namun, perlu diperhatikan adanya indiekasi overfitting selama proses training, yang memerlukan perhatian lebih lanjut dalam pengembangan model.

Kedua, evaluasi menggunakan metrik presisi, recall, dan F1-score menunjukkan keseimbangan yang baik antara kemampuan model dalam memprediksi antara kelas 0 (dibawah 20000) dan kelas 1 (diatas 20000), memberikan keyakinan terhadap kehandalan model dalam membedakan kategori popularitas game indie.

Ketiga, kemampuan model dalam membedakan antara kedua kelas tersebut didukung oleh hasil kurva *AUC* pada *ROC*, yang memiliki nilai sebesar 88%. Angka ini mencerminkan kemampuan model untuk secara efektif memisahkan kedua kelas, dengan kurva mendekati sudut 90 derajat, menandakan performa yang baik dalam memprediksi popularitas game indie.

Terakhir, analisis fitur Shapley menyoroti beberapa faktor yang memiliki pengaruh signifikan terhadap popularitas game indie. Fitur seperti kategori Steam Trading Cards dan penghargaan (achievements) memberikan kontribusi positif, sementara harga yang lebih tinggi juga memiliki dampak positif. Di sisi lain, keberadaan genre Casual cenderung memiliki dampak negatif, sementara game indie yang dapat dimainkan di sistem operasi Mac dan memiliki genre RPG cenderung memberikan dampak positif.

Temuan ini dapat menjadi dasar untuk pengambilan keputusan lebih lanjut dalam pengembangan dan pemasaran game indie, memungkinkan fokus pada elemen-elemen yang paling berpengaruh dalam meningkatkan daya tarik di pasar Steam.

Peneliti memberikan beberapa saran untuk penelitian selanjutnya berdasarkan temuan dalam penelitian ini. Pertama, diharapkan penelitian mendatang mempertimbangkan penggunaan jenis fitur teks, seperti detailed_description, short_description, dan about_the_game, yang tersedia pada steam untuk memberikan informasi tambahan yang dapat meningkatkan kualitas pembelajaran model.

Kedua, disarankan untuk melakukan percobaan dengan kombinasi perlakuan yang berbeda dari variabel penelitian ini, guna mengamati potensi perbedaan hasil evaluasi model yang dapat ditemukan.

Ketiga, penelitian selanjutnya diharapkan mengeksplorasi variasi dalam hyperparameter tuning untuk melihat pengaruhnya terhadap evaluasi model, sehingga dapat ditemukan konfigurasi optimal yang meningkatkan kinerja model.

Terakhir, peneliti merekomendasikan agar penelitian ini dapat diarahkan ke tahap deployment, sehingga model yang dikembangkan dapat digunakan secara langsung oleh pengembang game indie. Hal ini akan membantu menerapkan temuan penelitian dalam praktik, mendukung pengembangan dan pemasaran game indie secara efektif.

DAFTAR PUSTAKA

AMAZON SAGEMAKER. 2023. *Amazon SageMaker* Panduan Developer. *Retrieved from https://docs.aws.amazon.com/sagemaker/*

ANNA VERONIKA DOROOGUSH. 2018. *CatBoost gradient boosting with categorical features Support*. arXiv. https://doi.org/10.48550/arXiv.1810.11363

BAHADOR PARSA, A., MOVAHEDI, A., TAGHIPOUR, H., DERRIBLE, S., & MOHAMMADADIAN, A. 2019. *Toward Safer Highways, Application of XGBoost and SHAP for Real-Time Accident Detection and Feature Analysis*. Accident Analysis & Prevention, 125, 105405. https://doi.org/10.1016/j.aap.2019.105405

BALDI, P., BRUNAK, S., CHAUVIN, Y., ANDERSEN, C. A. F., & NIELSEN, H. 2000. *Assessing the Accuracy of Prediction Algorithms for Classification: An Overview*. Bioinformatics, 16(5), 412-424. https://doi.org/10.1093/bioinformatics/16.5.412

CAELEN, O. 2017. *A Bayesian Interpretation of the Confusion Matrix. Annals of Mathematics and Artificial Intelligence*, 81(3-4), 289-301. https://doi.org/10.1007/s10472-017-9564-8

CATBOOST.AI. n.d. *Catboost Doc*. Retrieved July 5, 2023, from https://catboost.ai/en/docs/

DEVIACITA, D., SASTY, H., MUHARDI, H., NAWAWI, D. H. H., & LAUT, B. 2019. Implementasi *Web Scraping* untuk Pengambilan Data pada Situs *Marketplace*. Jurnal Ilmu Komputer dan Informasi, 7(4), 45-54. http://dx.doi.org/10.26418/justin.v7i4.30930

DOROGUSH, A. V. 2018. *CatBoost: Gradient Boosting with Categorical Features Support*. Yandex. https://doi.org/10.48550/arXiv.1810.11363

IBRAHIM, A. A., RIDWAN, R. L., MUHAMMED, M. M., ABDULAZIZ, R. O., & SAHEED, G. A. 2020. *Comparison of the CatBoost Classifier with Other Machine Learning Methods. International Journal of Advanced Computer Science and Applications* (IJACSA), 11(11), 45-53. Retrieved from www.ijacsa.thesai.org

JIANG, Z., & WANG, Y. 2021. *PREDICTING THE POPULARITY OF INDEPENDENT VIDEO GAMES ON THE STEAM PLATFORM*. arXiv. https://doi.org/10.17615/aaj0-x494

KANNANGARA, K. K. P. M., ZHOU, W., DING, Z., & HONG, Z. 2022. *Investigation of Feature Contribution to Shield Tunneling-Induced*

Settlement Using Shapley Additive Explanations Method. Journal of Rock Mechanics and Geotechnical Engineering, 14(4), 1052-1063. https://doi.org/10.1016/j.jrmge.2022.01.002

LOUIS OWEN. 2022. *Boost your machine learning model's performance via hyperparameter tuning*. Packt publishing Ltd.

PARKER, F. 2013. *Indie Game Studies Year Eleven. Proceedings of DiGRA 2013 Conference: DeFragging Game Studies*. https://doi.org/10.48550/arXiv.1810.11363

PERMATASARI, N., ASY SYAHIDAH, S., LEOFIRO IRFIANSYAH, A., & AL-HAQQONI, M. G. 2022b. *PREDICTING DIABETES MELLITUS USING CATBOOST CLASSIFIER AND SHAPLEY ADDITIVE EXPLANATION (SHAP) APPROACH. BAREKENG*: Jurnal Ilmu Matematika Dan Terapan, 16(2), 615–624. https://doi.org/10.30598/barekengvol16iss2pp615-624

RODRÍGUEZ-PÉREZ, R., & BAJORATH, J. 2020. *Interpretation of Machine Learning Models Using Shapley Values: Application to Compound Potency and Multi-Target Activity Predictions. Journal of Computer-Aided Molecular Design*, 34(10), 1013-1026. https://doi.org/10.1007/s10822-020-00314-0

SABER, M., BOULMAIZ, T., GUERMOUI, M., ABDRABO, K. I., SUMI, T., BOUTAGHANE, H., NOHARA, D., & MABROUK, E. 2021. *Examining LightGBM and CatBoost Models for Wadi Flash Flood Susceptibility Prediction. Arabian Journal of Geosciences*, 14(21), 1-14. https://doi.org/10.1080/10106049.2021.1974959

STEAMSPY. *Retrieved July* 1, 2023, from steamspy.com

STEFEN T. WRIGHT. 2018. *There are too many video games. What now? - Polygon . Retrieved from* https://www.polygon.com/2018/9/28/17911372/there-are-too-many-video-games-what-now-indiepocalypse

WANG, G., ZHAO, B., WU, B., ZHANG, C., & LIU, W. 2022. *Intelligent Prediction of Slope Stability Based on Visual Exploratory Data Analysis of 77 In Situ Cases. International Journal of Mining Science and Technology*, 32(4), 845-854. https://doi.org/10.1016/j.ijmst.2022.07.002

WANG, L., WU, J., ZHANG, W., WANG, L., & CUI, W. (2021. *Efficient Seismic Stability Analysis of Embankment Slopes Subjected to Water Level Changes Using Gradient Boosting Algorithms. Frontiers in Earth Science*, 9, 807317. https://doi.org/10.3389/feart.2021.807317

Halaman ini sengaja dikosongkan.