

PENERAPAN POHON KEPUTUSAN C4.5 UNTUK MEMPREDIKSI TREN POPULARITAS GAME DI PLATFORM STEAM

Mhd Ilyasyah Drilanang, Tri Sapta Warman Zai, Feby Juliana Silalahi, Fanny Ramadhani

Prodi Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Medan

Jl. William Iskandar Ps. V Medan, Sumatera Utara 20221

mdrilanang.4233550027@mhs.unimed.ac.id

ABSTRAK

Pengembang dan penerbit game menghadapi kesulitan dalam menentukan game mana yang dapat menjadi populer atau trending seiring dengan meningkatnya jumlah game yang tersedia di platform distribusi digital seperti Steam. Banyak faktor yang memengaruhi popularitas suatu game, serta kurangnya alat prediksi yang efektif untuk membantu membuat keputusan tentang pengembangan dan pemasaran produk, menyebabkan masalah ini. Dengan menggunakan algoritma pohon keputusan C4.5, tujuan penelitian ini adalah untuk membangun model yang dapat memprediksi tren popularitas game. Pengumpulan data dari platform Steam digunakan. Data ini termasuk jumlah rekomendasi, ulasan positif, pencapaian game, dan harga. Setelah itu, data diproses melalui tahapan preprocessing, pemilihan fitur, dan pembagian dataset. Untuk pelatihan dan pengujian, rasio 80:20 digunakan. Untuk mencegah overfitting, model dibangun menggunakan DecisionTreeClassifier dengan parameter optimasi seperti kriteria pemisahan entropy dan kedalaman maksimum lima level. Hasil evaluasi menunjukkan bahwa model memiliki akurasi sebesar 92 persen, dengan jumlah rekomendasi dan ulasan positif yang paling penting. Studi ini dapat membantu pengembang game dan analis pasar memahami faktor utama yang memengaruhi popularitas game. Ini juga dapat menjadi dasar untuk membuat strategi pemasaran dan pengembangan produk yang lebih baik. Disarankan untuk ditambahkan fitur baru, seperti kategori game, dan teknik ensemble untuk meningkatkan prediksi.

Kata kunci : Algoritma C4.5, Pohon Keputusan, Prediksi Tren, Popularitas Game, Steam.

1. PENDAHULUAN

Dalam beberapa dekade terakhir, industri game telah mengalami pertumbuhan yang luar biasa. Dengan ribuan title game dengan berbagai genre dan konsep, platform distribusi digital seperti Steam telah menjadi pusat utama bagi para pemain di seluruh dunia. Dengan meningkatnya jumlah game yang tersedia di Steam, pengembang dan penerbit game menghadapi tantangan untuk mengidentifikasi tren permainan yang akan diminati di masa depan. Memprediksi game mana yang akan menjadi populer—juga dikenal sebagai "trending"—menjadi sangat penting untuk keputusan bisnis, perencanaan strategi pemasaran, dan alokasi sumber daya pengembangan dalam lingkungan persaingan yang ketat. Penelitian ini bertujuan untuk menganalisis game – game apa saja yang pengguna mainkan di platform penyedia game Steam berdasarkan dataset yang disediakan oleh Steam, kemudian mengimplementasikan algoritma c4.5 untuk menentukan apakah sebuah game akan trending atau tidak dan mengevaluasi performa model klasifikasi yang digunakan dalam penelitian. Rumusan masalah yang didapat yaitu bagaimana cara menganalisis pola game pengguna di platform Steam berdasarkan dataset dari Steam, bagaimana algoritma c4.5 dapat digunakan untuk mengklasifikasikan game apa saja yang berpotensi menjadi trending kedepannya, dan seberapa akurat model klasifikasi yang dihasilkan dalam memprediksi tren popularitas game di platform Steam. Berdasarkan pada dataset yang disediakan Steam, dengan menggunakan teknik data mining dan

algoritma Decision Tree (C4.5) kami dapat menganalisis komponen utama yang mempengaruhi trending suatu game di platform tertentu. Penelitian ini diharapkan menemukan pola data yang jelas yang akan membantu pengembang game, penerbit, dan analis pasar memahami komponen apa yang berkontribusi terhadap popularitas sebuah game. Harapan yang ingin kami capai adalah bahwa penelitian ini akan memberikan pemahaman yang lebih mendalam tentang komponen utama yang mempengaruhi tren game dan untuk memberi pengembang game panduan untuk mengoptimalkan strategi pemasaran dan pengembangan produk mereka.

2. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Penelitian kali ini mengambil informasi dan penelitian terdahulu untuk digunakan sebagai referensi. Beberapa penelitian terdahulu yang menggunakan pendekatan algoritma c4.5 yaitu:

"Optimization of ID3 Algorithm in Educational Data Classification", penelitian dari Universitas Delhi, India. Studi ini menyelidiki seberapa efektif algoritma ID3 dalam memprediksi masa studi dengan mengklasifikasikan data akademik siswa. Hasil penelitian menunjukkan bahwa algoritma ID3 memiliki kemampuan untuk membuat sistem klasifikasi yang mudah dipahami dan akurasi tinggi dalam analisis data pendidikan [1].

Penelitian dari Universitas Potensi Utama di Indonesia berjudul "Decision Tree C4.5 for Predicting ATM Cash Availability" menggunakan algoritma

C4.5 untuk mengoptimalkan distribusi uang pada mesin ATM berdasarkan data transaksi. Hasil tes menunjukkan tingkat akurasi sebesar 85% untuk metode validasi cross-fold K-Fold [2].

Penelitian sebelumnya yang dilakukan di Universitas Teknokrat Indonesia membahas penerapan algoritma C4.5 untuk membantu mengklasifikasikan mahasiswa penerima program bantuan pendidikan. Dalam penelitian tersebut, data yang digunakan mencakup IPK, keikutsertaan dalam kompetisi, dan aktivitas organisasi mahasiswa. Dengan menerapkan metode Ten-Fold Cross Validation, model yang dibangun mampu mencapai akurasi, presisi, dan recall sebesar 87%. Hasil ini menunjukkan bahwa algoritma C4.5 cukup efektif dalam membantu pengambilan keputusan terkait pemberian beasiswa [13].

Sementara itu, penelitian lain dari Universitas Singaperbangsa Karawang mencoba menerapkan algoritma C4.5 untuk memprediksi penyakit paru-paru. Penelitian ini memanfaatkan data pasien dari platform Kaggle yang berjumlah sekitar 30.000 entri, dan diproses menggunakan RapidMiner. Setelah melalui serangkaian tahapan KDD, pengujian dilakukan dengan pembagian data 90% untuk pelatihan dan 10% untuk pengujian. Hasilnya cukup mengesankan, dengan akurasi sebesar 89,77%, recall 78,61%, dan precision mencapai 100%. Temuan ini menunjukkan potensi besar algoritma C4.5 dalam bidang medis, khususnya dalam prediksi penyakit berbasis data [15].

Adapun penelitian dari Universitas Amikom Yogyakarta mengangkat isu keterlambatan pembayaran uang sekolah. Dengan menggunakan algoritma C4.5 yang diimplementasikan melalui bahasa pemrograman Python, peneliti mencoba memprediksi potensi keterlambatan pembayaran berdasarkan data siswa seperti penghasilan orang tua, jumlah tanggungan keluarga, dan latar belakang pendidikan orang tua. Dataset diambil dari SMA Kolese De Britto Yogyakarta. Hasil pengujian menunjukkan bahwa model ini mampu mencapai akurasi sebesar 73%, dengan presisi dan recall yang sama-sama berada pada angka 71%. Meskipun angkanya belum seoptimal penelitian sebelumnya, pendekatan ini tetap menunjukkan bahwa algoritma C4.5 dapat diterapkan secara praktis dalam konteks pendidikan untuk mendukung pengambilan keputusan administrasi sekolah [16].

2.2. Data Mining

Data mining adalah proses penambangan informasi penting dari database. Ini adalah proses yang kompleks yang melibatkan artificial intelligence, teknik statistik, ilmu matematika, pembelajaran mesin, dan berbagai teknik lainnya. Teknik-teknik ini digunakan untuk menemukan dan mengekstraksi informasi berharga dari database yang sangat besar. Data mining adalah cabang computer science dan statistika yang digunakan untuk mengekstraksi

informasi dari database yang sangat besar. Ini adalah cara untuk menguraikan dan menemukan pengetahuan yang terkandung dalam suatu database. proses untuk memilih dan "menambang" informasi dari kumpulan data yang besar. Data mining adalah proses menemukan informasi penting dari database yang besar dengan mengeksplorasi dan menganalisis data yang besar untuk menemukan pola dan aturan. Dalam arti lain, data mining digunakan untuk mengekstrak informasi dari data yang besar (big data) untuk membantu pengambilan Keputusan [4].

2.3. Decision Tree

Pohon keputusan adalah salah satu metode klasifikasi yang paling umum karena mudah diinterpretasikan oleh manusia. Mengubah data menjadi pohon keputusan dan aturan-aturan keputusan adalah konsep dasar dari pohon keputusan. Salah satu manfaat utama dari penggunaan pohon keputusan adalah membuat proses pengambilan keputusan yang kompleks menjadi lebih sederhana. Ini memungkinkan pengambil keputusan untuk lebih baik memahami solusi masalah. Pohon keputusan juga dapat digunakan untuk menyelidiki data dan menemukan hubungan tersembunyi antara sebuah variabel target dan jumlah variabel potensial input [10].

2.4. Algoritma C4.5

Algoritma C4.5 adalah salah satu jenis yang terkenal dari Pohon Keputusan yang dirancang untuk menangani berbagai macam jenis atribut, baik yang berbentuk numerik maupun kategorikal. C4.5 merupakan algoritma yang digunakan dalam proses klasifikasi atau pengelompokan data. Fundamental algoritma C4.5 adalah pembuatan pohon keputusan.

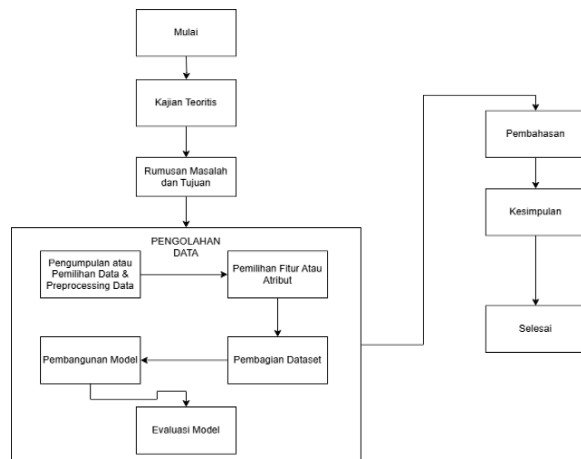
Algoritma C4.5 adalah evolusi dari ID3 dan mencakup pengembangan untuk mengatasi nilai yang tidak ada, terus data, dan pruning. Algoritma C4.5 memiliki input berupa pelatihan contoh dan contoh; pelatihan contoh adalah data contoh yang akan digunakan untuk membangun pohon yang telah diuji kebenarannya, dan pelatihan contoh adalah field-field data yang akan digunakan sebagai parameter dalam melakukan klasifikasi data [3].

3. METODE PENELITIAN

3.1. Tahapan Penelitian

Tahapan pada penelitian ini dibagi menjadi:

- Kajian Teoritis, dengan menyusun teori yang berhubungan dengan penelitian sebagai kerangka berpikir dan dasar teoritis
- Rumusan masalah dan tujuan, yaitu analisis, potensi dan akurasi
- Pengolahan data, bagaimana data dari platform Steam akan diolah
- Pembahasan
- Kesimpulan dari penelitian yang sudah dilakukan.



Gambar 1. Diagram alir tahapan penelitian

3.2. Metode Pengumpulan Data

Dalam penelitian ini, data yang dikumpulkan, dengan mengamati tren game di platform Steam, seperti pola pembelian, jumlah pemain aktif, dan ulasan pengguna. Berdiskusi dengan pengembang game, analis industri, atau pemain aktif untuk mengetahui faktor-faktor yang memengaruhi popularitas game, dan mengumpulkan informasi dari pengguna tentang preferensi mereka untuk game, pilihan pembelian, dan pengalaman bermain.

3.3. Metode Analisa Data

Metode Decision Tree (C4.5) digunakan untuk melakukan analisis data penelitian ini untuk mengklasifikasikan tren game berdasarkan fitur numerik yang tersedia. Tujuan analisis ini adalah untuk mendapatkan pemahaman tentang komponen utama yang mempengaruhi popularitas game di platform distribusi digital. Data yang dikumpulkan untuk analisis dalam penelitian ini mencakup berbagai game yang tersedia di platform digital seperti Steam, yang memiliki kategori yang berbeda, mulai dari indie hingga AAA, dengan variasi dalam jumlah pemain, ulasan, dan harga. Fitur-fitur yang mempengaruhi tren game, seperti rekomendasi (jumlah rekomendasi pengguna), ulasan positif (jumlah ulasan positif), harga (harga game), pencapaian (jumlah pencapaian dalam game), dan tingkat CCU tertinggi (jumlah Untuk memastikan relevansi tren saat ini, data yang digunakan diambil dari platform distribusi game digital seperti Steam API, yang memberikan informasi tentang statistik dan performa game. Analisis ini dilakukan dengan mengumpulkan data yang diformat dan dibersihkan sesuai dengan model setelah dikumpulkan dari API dan sumber lainnya. Untuk menemukan pola klasifikasi berdasarkan informasi gain, model C4.5 digunakan. Kemudian, untuk mengukur akurasi, ketepatan, recall, dan skor F1, model diuji menggunakan laporan klasifikasi dan matriks keakuratan. dan visualisasi pohon keputusan digunakan untuk mempelajari komponen utama yang memengaruhi tren game.

4. HASIL DAN PEMBAHASAN

4.1. Pengumpulan dan Processing Data

Dataset diambil dari platform HuggingFace.co dalam bentuk “.csv” dengan total data yaitu 2474 data dan 13 atribut awal, yaitu; AppID (Id Game), Name, Release date, Estimated owners (Perkiraan Banyaknya Pemilik Game), Peak CCU (Puncak Angka Tertinggi Pemain Game), Required age, Price, About the game (Deskripsi Game), Positive (Ulasan Positif), Negative (Ulasan Negatif), Achievements (Pencapaian Game), Recommendations (Banyaknya Orang Yang Merekomendasikan), dan Genres. Tahap preprocessing meliputi:

- Konversi Tipe Data: Kolom numerik seperti Price dan Positive awalnya berbentuk string. Menggunakan `pd.to_numeric()`, nilai-nilai tersebut diubah ke tipe float. Contoh: nilai "\$19.99" dikonversi menjadi 19.99.
 - Lalu, memilah atribut - atribut yang akan dipakai, totalnya hanya ada 7 Atribut berbentuk numerikal yang akan dipakai dengan menghilangkan atribut yang tidak penting seperti AppId, Required Age, Release Date, Genres, Deskripsi Game, dan Nama Game. Tetapi atribut nama game ini akan tetap dipakai tetapi tidak akan dipakai sebagai atribut data testing maupun data training.
 - Penanganan Missing Value: Data kosong (NaN) diisi dengan nilai 0 menggunakan `df.fillna(0)`, mengasumsikan bahwa ketiadaan nilai merepresentasikan tidak adanya kontribusi dari fitur tersebut (misal: Achievements = 0 berarti game tidak memiliki pencapaian).
 - Pembuatan Label Target: Atribut Peak CCU dan Estimated Owners akan menjadi target. Target yang dimaksud adalah untuk dijadikan dasar prediksi diakhir jika suatu game itu akan Trending atau Tidak. Maka akan menyisakan 5 atribut utama. Kita tidak mengikutkan atribut Estimated owners (Perkiraan Banyaknya Pemilik Game), Peak CCU (Puncak Angka Tertinggi Pemain Game) itu dikarenakan atribut ini sebenarnya sudah bisa membuat target Trending secara mudah dilihat dari besarnya angka atau nilai dari kedua atribut tersebut.
- Penelitian Kami berfokus pada apakah kelima atribut yang lain dapat memprediksi secara akurat seperti kedua atribut Peak CCU dan Estimated Owners.**
- Kolom "Trending" dibuat dengan kriteria logika $(\text{Peak CCU} > 5000) \mid (\text{Estimated owners} > 100000)$. Threshold ini dipilih berdasarkan analisis distribusi data untuk mengidentifikasi game dengan trafik pengguna yang signifikan.

4.2. Pemilihan Fitur/Atribut Target

Atribut yang digunakan adalah Price, Positive, Negative, Achievements, dan Recommendations. 5 Atribut yang akan menjadi atribut utama dalam memprediksi Target Trending.

4.3. Pembagian Dataset

Data yang ada didalam file “.csv” berjumlah 2474 total. Dataset dibagi dengan rasio 80:20 (training:testing). Artinya 20% atau sekitar 496 data akan dijadikan Data Testing, lalu sekitar 1979 data atau 80% dari data akan dijadikan Data Training.

Kita menggunakan `train_test_split` dengan parameter `random_state=42` (nilai konstan) untuk memastikan hasil yang konsisten setiap kali kode dijalankan. Pembagian ini menghasilkan variasi data yang seimbang antara kedua kelas (Trending dan Not Trending).

4.4. Pembangunan Model

Algoritma C4.5 diimplementasikan melalui `DecisionTreeClassifier` dengan parameter:

- `criterion="entropy"` untuk memaksimalkan gain informasi.
- `max_depth=5` untuk mencegah overfitting dan memudahkan interpretasi visual.
- Model dilatih dengan data training (`X_train`, `y_train`), dan struktur pohon keputusan di-generate untuk analisis lebih lanjut

4.5. Evaluasi Model

Evaluasi dilakukan dengan metrik:

- Akurasi: Proporsi prediksi benar dari total data.
- Classification Report: Menghitung precision, recall, dan F1-score untuk kedua kelas.
- Confusion Matrix: Memvisualisasikan jumlah prediksi benar dan salah.
- Visualisasi pohon keputusan diekspor menggunakan `plot_tree` untuk memahami pola keputusan model.

4.6. Menyimpan Hasil Output dan Dataset Kembali ke Bentuk .csv

- File CSV ini berisi data yang telah melalui preprocessing (konversi tipe data, penanganan missing value, dan kolom target Trending).
- Digunakan sebagai input untuk analisis lanjutan atau prediksi menggunakan model yang sudah disimpan.

4.7. Implementasi Prediksi dan Penyaringan File .csv Yang Sudah Disimpan Tadi

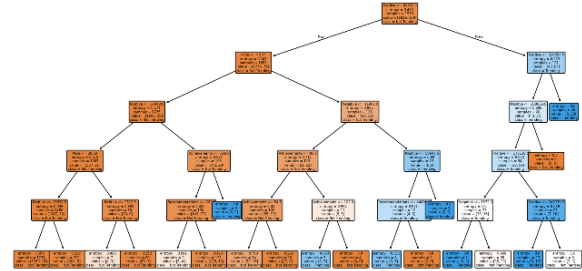
- Data difilter berdasarkan nilai Prediction untuk memisahkan game trending dan tidak trending.
- Manfaat Implementasi:
 - Efisiensi: Model tidak perlu dilatih ulang—cukup dimuat sekali untuk prediksi massal.
 - Transparansi: Hasil prediksi dapat diverifikasi langsung melalui file CSV, memudahkan analisis manual.
 - Skalabilitas: Kode dapat diadaptasi untuk dataset baru dengan struktur fitur yang sama.

Classification Report:				
	precision	recall	f1-score	support
0	0.94	0.99	0.97	456
1	0.77	0.26	0.38	39
accuracy			0.94	495
macro avg	0.85	0.62	0.68	495
weighted avg	0.93	0.94	0.92	495

Confusion Matrix:	
[[453 3]	
[29 10]]	

Gambar 2. Output program python

Output dari program python yang sudah dijelaskan



Gambar 3. Visualisasi pohon keputusan

Confusion Matrix dari hasil prediksi Data Testing yang berjumlah total 495 menunjukkan bahwa :

- Sample data yang diprediksi sebagai kelas 0 (Tidak Trending) dan Actual Kelasnya memang 0 juga, ada 453 sampel data yang menjadikannya TN (True Negative)
- Sample data yang diprediksi sebagai kelas 0 (Tidak Trending) tetapi Actual atau kelas Sebenarnya itu kelas 1 (Trending) ada 29 sampel data yang menjadikannya FN (False Negative)
- Sample data yang diprediksi sebagai kelas 1 (Trending) tetapi Actual kelas sebenarnya adalah kelas 0 (Tidak Trending) itu ada 3 sampel data (False Positive)
- Sample data yang diprediksi sebagai kelas 1 (Trending) dan Actual Kelasnya memang juga 1, ada 10 data yang menjadinya TP (True Positive)

Tabel 1. Tabel predicted dan actual

	PREDICTED: 0	PREDICTED: 1
ACTUAL: 0	453 (TN)	3 (FP)
ACTUAL: 1	29 (FN)	10 (TP)

Lalu masuk ke hasil Classification Report, ditunjukkan bahwa:

Precision (Persentase Prediksi Yang Benar) yang dihitung dari rumus :

$$\frac{TP}{TP + FP} \text{ atau } \frac{TN}{TN + FN}$$

Didapat bahwa kelas 0 (Tidak Trending) memiliki Persentase 0.94 atau 94 % dari 456 sampel data karena 29 sampel data (6%) masuk sebagai FN (False Negative). Lalu yang kelas 1 persentase 77% dari 10 sampel data dengan 3 sampel data (23%) termasuk FP (False Positive).

Recall (Persentase data kelas tertentu yang berhasil terdeteksi (terprediksi benar) dari semua data yang benar-benar berada di kelas tersebut) dihitung dari rumus:

$$\frac{TP}{TP + FN} \text{ atau } \frac{TN}{TN + FP}$$

Classification Report:				
	precision	recall	f1-score	support
0	0.94	0.99	0.97	456
1	0.77	0.26	0.38	39
accuracy			0.94	495
macro avg	0.85	0.62	0.68	495
weighted avg	0.93	0.94	0.92	495

Gambar 6. Hasil classification report untuk precision dan recall

Didapat bahwa kelas 0 mempunyai persentase 99% berhasil diprediksi dari total 456 sampel data, lain halnya pada kelas 1 hanya 26 % dari 39 sampel data berhasil diprediksi.

F1-Score (Rata-rata harmonik antara precision dan recall) F1-score tinggi menunjukkan keseimbangan baik antara precision dan recall. Dihitung dari rumus:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Didapat bahwa f1 score dari kelas 0 bernilai 97 % dan kelas 1 bernilai 38 %. Kelas 1 bernilai rendah karena perbedaan nilai yang signifikan dari target Trending dan Tidak Trending yaitu 456 dan 39.

Lalu accuracy (Persentase Keseluruhan Keakuratan) dihitung dari rumus

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Didapat keakuratan 94 % dari total sampel Data Testing 495.

Lalu untuk macro average(Rata-rata Precision, Recall, dan F1-score di antara semua kelas tanpa memperhatikan perbedaan jumlah data tiap kelas) dihitung dengan rumus (Rumus sama untuk recall dan f1-score)

$$\text{MacroAvg} = \frac{\text{Precision}_0 + \text{Precision}_1}{2}$$

Classification Report:				
	precision	recall	f1-score	support
0	0.94	0.99	0.97	456
1	0.77	0.26	0.38	39
accuracy			0.94	495
macro avg	0.85	0.62	0.68	495
weighted avg	0.93	0.94	0.92	495

Gambar 7. Hasil classification report untuk f1-score, accuracy, macro avg dan weighted avg

Didapat bahwa Macro average dari precision, recall, dan f1 score secara berurutan 85 %, 62 %, dan 68 %

Dan untuk weighted avg menggunakan rumus (Rumus sama untuk recall dan f1-score) dan S adalah Support (Jumlah sampel data)

$$\text{WeightedAvg} = \frac{S_0 \times \text{Precision}_0 + S_1 \times \text{Precision}_1}{S_0 + S_1}$$

Didapat Weighted Average dari precision, recall, dan f1 score secara berurutan 93 %, 94 %, dan 92%.

4.8. Analisis Pohon Keputusan

Visualisasi pohon mengungkap kriteria utama yang digunakan model:

a. Root Node (Node Akar)

Terletak di paling atas (misalnya, membagi data berdasarkan Positive <= 82752.0). Node ini dipilih oleh algoritma C4.5 karena memiliki information gain tertinggi di antara semua fitur.

Artinya, membagi data berdasarkan nilai atribut tersebut (misalnya Positive) memberikan pemisahan paling baik antara kelas Not Trending (cokelat) dan Trending (biru).

b. Cabang (Branch)

Setiap cabang keluar dari suatu node menunjukkan hasil dari kondisi pada atribut. Misalnya, cabang kiri jika Positive <= 82752.0, dan cabang kanan jika Positive > 82752.0. Proses pemilihan atribut diulang secara rekursif di setiap node, hingga data mencapai satu kelas atau memenuhi kondisi penghentian (seperti max_depth=5).

c. Leaf Node (Node Daun).

Node daun adalah node tanpa cabang lebih lanjut. Setiap leaf merepresentasikan kelas akhir (Not Trending atau Trending) atau distribusi kelas di data latih. Biasanya di dalam node daun ditunjukkan jumlah sampel yang masuk ke node tersebut dan probabilitas (atau mayoritas) kelasnya.

4.9. Hasil Output Data Dijadikan File “.csv”

Hasil Output data ini akan dimasukkan dan diubah lagi kedalam bentuk file “.csv” Dengan penambahan Target Atribut baru yaitu “Trending” dengan nilai 0 (Tidak Trending) dan nilai 1 (Trending). Juga penambahan kembali atribut yang tidak dipakai tadi seperti atribut AppId, Required Age, Release Date, Genres, dan Deskripsi Game (About the game).

AppId	Name	Release date	Estimated owners	Peak CCU	Required age	Price
578080	PUBG: BATTLEGROUNDS	Dec 21, 2017	0	275374	0	0
301520	Robocraft	Aug 24, 2017	0	302	0	0

About the game	Positive	Negative	Achievements	Recommendation	Genres	Trending
LAND, LOOT, SURVIVE	1154855	895978	37	1616422	Action, Adventure	1
Build insane, full	86140	32448	43	2542	Action, Free to play	0

Gambar 8. Hasil Output Data

4.10. Ketidakseimbangan Kelas (Imbalanced Dataset).

Dari 495 data uji, hanya 39 yang benar-benar Trending. Hal ini membuat recall kelas Trending

sangat rendah (0.26). Model cenderung lebih “aman” memprediksi Not Trending karena mayoritas data memang Not Trending.

4.11. Recall Rendah pada Kelas Trending

Dikarenakan rendahnya recall akibat dari ketidakseimbangan kelas Trending dan Tidak Trending, ada juga data game yang sebenarnya Trending tetapi diprediksi Not Trending (FN). Untuk aplikasi nyata, hal ini bisa berakibat melewatkan potensi game yang sebenarnya akan populer.

4.12. Definisi Trending Bergantung Atribut Tertentu

Label Trending ditentukan oleh Peak CCU > 5000 atau Estimated Owners > 100000. Tidak mengikutkan atribut Peak CCU dan Estimated Owner adalah fokus projek ini, Meskipun ini dilakukan untuk menghindari data leakage, model menjadi kurang optimal jika data lain tidak cukup kuat menjelaskan status Trending.

4.13. Overfitting Berpotensi Terjadi

Walau max_depth=5 membantu mencegah pohon menjadi terlalu dalam, masih ada kemungkinan model ini Overfitting (Model terlalu cocok dengan data pelatihan, sehingga tidak dapat memprediksi data baru dengan baik). Terdapat node-node yang mungkin sangat spesifik pada data training, sehingga kinerjanya pada data lain belum tentu sama

5. KESIMPULAN DAN SARAN

Kesimpulan dari penelitian kali ini, Algoritma C4.5 menghasilkan model dengan akurasi 92%, menjadikannya layak untuk prediksi tren game di Steam. Rekomendasi (Recommendations) dan ulasan positif (Positive) adalah fitur paling kritis, mencerminkan pentingnya interaksi komunitas dalam popularitas game. Implementasi model ini dapat dikembangkan dengan menambahkan fitur temporal (Release Date) atau kategori (Genres) untuk meningkatkan generalisasi. Dan Rekomendasi untuk Penelitian Lanjutan yaitu Eksperimen dengan algoritma ensemble (Random Forest) untuk mengurangi bias dan Analisis korelasi antara Price dan Trending untuk strategi penetapan harga optimal.

DAFTAR PUSTAKA

- [1] SAUDDIN, A., & IDA, N. (2022). Penerapan Pohon Keputusan Dalam Memprediksi Masa Studi Mahasiswa Uin Alauddin Makassar. *Jurnal INSTEK (Informatika Sains dan Teknologi)*, 7(2), 201-210.
- [2] Syahputra, F., Hartono, H., & Rosnelly, R. (2021). Penerapan Algoritma C4. 5 Dalam Memprediksi Ketersediaan Uang Pada Mesin ATM. *Jurnal Media Informatika Budidarma*, 5(2), 556-563.
- [3] Prasetyo, A., Hasibuan, M. H., & Sitompul, P. (2021). Simulasi Penerapan Metode Decision Tree (C4. 5) Pada Penentuan Status Gizi Balita. *Jurnal Nasional Komputasi dan Teknologi Informasi*, 4(3).
- [4] Telaumbanua, D., & Kurniawati, I. (2022). Penerapan Algoritma C4. 5 Untuk Klasifikasi Kepuasan Pelanggan Pada Jasa Layanan Pengiriman. *JoMMiT: Jurnal Multi Media dan IT*, 6(1).
- [5] Wanto, A., & Gunawan, I. (2022). Penerapan Algoritma Decision Tree C4. 5 untuk Klasifikasi Tingkat Kesejahteraan Keluarga pada Desa Tiga Dolok. *Krisnadana Journal*, 1(2), 21-32.
- [6] Putri, S. U., Irawan, E., & Rizky, F. (2021). Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4. 5. *Kesatria: Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen)*, 2(1), 39-46.
- [7] Andarista, R. R., & Jananto, A. (2022). Penerapan Data Mining Algoritma C4. 5 Untuk Klasifikasi Hasil Pengujian Kendaraan Bermotor. *Jurnal Tekno Kompak*, 16(2), 29-43.
- [8] Putro, A. W. G. P., & Setiadi, T. (2023). Penerapan Klasifikasi Decision Tree (C4. 5) untuk memprediksi kelulusan siswa sekolah dasar di Kecamatan Juai. *Jurnal Format*, 12(2), 151-157.
- [9] Subekti, I. A., Andriano, C. B., Nurdiansyah, D., & Hidayat, R. (2025). PENERAPAN ALGORITMA C4. 5 UNTUK MEMPREDIKSI HARGA RUMAH. *Just IT: Jurnal Sistem Informasi, Teknologi Informasi dan Komputer*, 15(2), 350-355.
- [10] Komarudin, R., Yudha, P. V. B., Maulana, Y. I., Afni, N., Salim, A., & Carolina, I. (2021). Penerapan Metode Algoritma C4. 5 Dalam Klasifikasi Diagnosa Penyakit Umum Menggunakan WEKA. *Bianglala Informatika*, 9(2), 95-102.
- [11] Suriani, U. (2023). Penerapan data mining untuk memprediksi tingkat kelulusan mahasiswa menggunakan algoritma decision tree C4. 5. *Journal of Computer and Information Systems Ampera*, 4(2), 55-65.
- [12] Ruliansyah, M. J., & Betty, M. (2024). Penerapan Metode C4. 5 dalam Prediksi Penjualan Tim Bev 1 pada PT. Surya Pangan Sejahtera Bekasi Jawa Barat. *Jurnal Indonesia: Manajemen Informatika dan Komunikasi*, 5(2), 1269-1278.
- [13] Adinda, P. R. (2022). Pohon keputusan c4. 5 algoritma untuk klasifikasi program bantuan belajar. *Jurnal Portal Data*, 2(9).
- [14] Harryanto, F. F., & Hansun, S. (2017). Penerapan Algoritma C4. 5 untuk Memprediksi Penerimaan Calon Pegawai Baru di PT WISE. *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, 3(2), 95-103.
- [15] Sofyan, F. M. A., Voutama, A., & Umaidah, Y. (2023). Penerapan Algoritma C4. 5 Untuk Prediksi Penyakit Paru-Paru Menggunakan

- Rapidminer. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 7(2), 1409-1415.
- [16] Ginting, V. S., Kusriani, K., & Luthfi, E. T. (2020). Penerapan Algoritma C4. 5 Dalam Memprediksi Keterlambatan Pembayaran Uang Sekolah Menggunakan Python. *Jurnal Teknologi Informasi*, 4(1).
- [17] Chandra, A. (2017). Penerapan data mining menggunakan pohon keputusan dengan algoritma c4. 5 dalam menentukan kecelakaan penerbangan. *SEMNAS TEKNOMEDIA ONLINE*, 5(1), 2-1.