PREDIKSI PENJUALAN GAME MENGGUNAKAN ALGORITMA LINEAR REGRESSION

Narendra Saputra[1)], Shifaul Akmal Dzauqi[2)], Dwika Mahendra[3)], Diana Laily Fithri[4)]
Program Studi Sistem Informasi, Fakultas Teknik, Universitas Muria Kudus[1),2),3),4)]

marendraru@gmail.com[1)], grafikakmal@gmail.com[2)], dwikamahendra6@gmail.com[3)], Diana.Laily@umk.ac.id[4)]

Abstrak

Penelitian ini bertujuan untuk meramalkan penjualan global video game menggunakan algoritma Linear Regression berdasarkan beberapa fitur seperti platform, genre, dan tahun rilis. Dataset yang digunakan memiliki 1.659 entri dan 11 atribut. Data mengalami proses preprocessing seperti pembersihan nilai kosong dan one-hot encoding. Hasil pengujian menunjukkan bahwa Linear Regression tidak cocok digunakan pada dataset ini, karena nilai R^2 yang negatif (-0.1111) dan MSE yang cukup tinggi (7.5634). Penelitian ini merekomendasikan penggunaan algoritma machine learning lain yang lebih kompleks [1].

Kata kunci: *Linear Regression, Video Game Sales, Prediksi, Machine Learning*

1. Pendahuluan

Industri game merupakan salah satu sektor hiburan yang mengalami pertumbuhan pesat dalam beberapa dekade terakhir [2]. Seiring dengan meningkatnya permintaan pasar dan pesatnya perkembangan teknologi, para pengembang dan publisher game dituntut untuk memahami tren penjualan agar dapat mengambil keputusan bisnis yang lebih tepat sasaran. Salah satu pendekatan yang digunakan untuk memahami pola dan memprediksi penjualan adalah dengan menerapkan teknik data mining, khususnya algoritma Linear Regression [1].

Linear Regression atau regresi linier merupakan metode statistik yang umum digunakan dalam peramalan, terutama ketika target yang diprediksi berbentuk data numerik kontinu. Dalam konteks penjualan game, variabel seperti tahun rilis (Year), platform, dan genre dapat digunakan sebagai fitur input untuk memprediksi besarnya penjualan global (Global_Sales). Metode ini dipilih karena sifatnya yang sederhana, efisien, dan mampu memberikan interpretasi yang jelas terhadap pengaruh tiap fitur terhadap target.

Dengan demikian, penggunaan regresi linier tetap memberikan gambaran awal tentang hubungan fitur terhadap penjualan, namun perlu ditingkatkan baik dari segi pemilihan fitur maupun pemodelan agar hasil prediksi menjadi lebih akurat. Penelitian lebih lanjut dengan penambahan fitur eksternal atau penggunaan model yang lebih kompleks seperti Random Forest atau XGBoost sangat disarankan [3].

2. Kajian Pustaka dan Pengembangan Hipotesis

2.1 Implementasi Data Mining untuk Game Android

Yulianti et al. menerapkan algoritma Apriori untuk menentukan game Android yang paling diminati berdasarkan pola asosiasi antar fitur. Meskipun berbeda pendekatan, studi ini memperlihatkan pentingnya data-driven decision making dalam industri game [5].

2.2 Analisis Big Data Penjualan Video Game

Husni et al. menggunakan metode EDA untuk menganalisis big data penjualan video game. Hasil analisis membantu perusahaan dalam menentukan strategi pemasaran berdasarkan genre dan platform dengan penjualan tertinggi dan terendah [6].

2.3 Prediksi Penjualan dengan Simple Linear Regression

Duran et al. menguji regresi linier sederhana untuk memprediksi penjualan produk. Studi ini menunjukkan relevansi metode regresi untuk peramalan penjualan, namun juga menekankan pentingnya fitur yang kuat [7].

2.4 Teori Linear Regression dalam Data Mining

Dijelaskan sebagai teknik statistik klasik dalam data mining yang efektif untuk target prediksi numerik [1].

2.5 Nilai R^2 dalam Evaluasi Model

Model Nilai R^2 negatif menandakan bahwa model berkinerja lebih buruk dari model baseline yang hanya memprediksi rata-rata target [3].

2.6 Evaluasi Komparatif Model Regresi

Royan dan Aathis membandingkan beberapa metode regresi untuk memprediksi penjualan video game, menunjukkan bahwa model linear memiliki keterbatasan dalam menangani pola non-linear [8].

2.7 Pengaruh Harga dan Ulasan dalam Penjualan Game

Jamil dan Sulianta menggunakan K-Means untuk menganalisis pengaruh harga dan ulasan terhadap penjualan video game, menunjukkan bahwa faktor eksternal seperti ulasan pengguna dapat mempengaruhi penjualan [9].

2.8 Pemanfaatan Regresi Berganda

Sari et al. mengembangkan model regresi berganda untuk menganalisis faktor-faktor yang mempengaruhi penjualan produk digital, termasuk video game [10].

2.9 Pengaruh Popularitas Platform terhadap Penjualan

Handayani et al. (2023) membahas bagaimana preferensi pengguna terhadap platform game (seperti PS, Xbox, atau PC) turut memengaruhi prediksi penjualan [11].

2.10 Implementasi XGBoost dalam Prediksi Penjualan

Ramadhani et al. menunjukkan keunggulan XGBoost dibandingkan Linear Regression untuk prediksi penjualan produk retail yang datanya kompleks [12].

3. Metode Penelitian

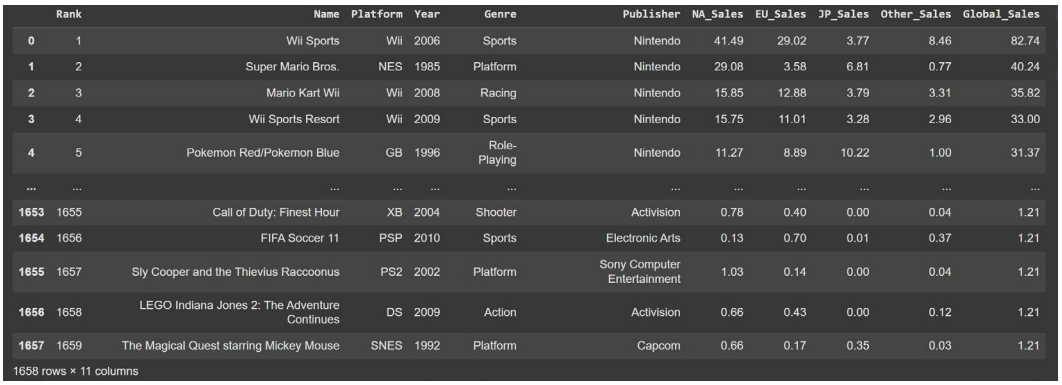
3.1 Algoritma Linear Regression

Linear regression adalah salah satu teknik *machine learning* yang digunakan untuk memodelkan hubungan antara variabel independen (fitur) dengan variabel dependen (target) yang bersifat kontinu.

Dengan menggunakan linear regression, kita bisa memperoleh wawasan tentang pengaruh masing-masing faktor terhadap penjualan global dan membuat prediksi yang berguna untuk keputusan bisnis di masa depan

3.2 Dataset

Dataset diperoleh dari situs Kaggle melalui tautan:
https://www.kaggle.com/datasets/anandshaw2001/video-game-sales
Dataset ini berisi informasi penjualan video game dari berbagai platform dan wilayah, serta data tambahan seperti tahun rilis dan genre



	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	1	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00
4	5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37
...	...	...	...	...	...	...	...	...	...	...	...
1653	1655	Call of Duty: Finest Hour	XB	2004	Shooter	Activision	0.78	0.40	0.00	0.04	1.21
1654	1656	FIFA Soccer 11	PSP	2010	Sports	Electronic Arts	0.13	0.70	0.01	0.37	1.21
1655	1657	Sly Cooper and the Thievius Raccoonus	PS2	2002	Platform	Sony Computer Entertainment	1.03	0.14	0.00	0.04	1.21
1656	1658	LEGO Indiana Jones 2: The Adventure Continues	DS	2009	Action	Activision	0.66	0.43	0.00	0.12	1.21
1657	1659	The Magical Quest starring Mickey Mouse	SNES	1992	Platform	Capcom	0.66	0.17	0.35	0.03	1.21

1658 rows × 11 columns

Gambar 1 Dataset mentah

- **Jumlah baris**: 1.659 baris
- **Jumlah kolom**: 11 kolom
- **Nama kolom dan tipe datanya**:
    - Rank — Integer
    - Name — String
    - Platform — String
    - Year — Float/Integer (tergantung isian)
    - Genre — String
    - Publisher — String
    - NA_Sales — Float (penjualan di North America dalam juta unit)
    - EU_Sales — Float (penjualan di Europe dalam juta unit)
    - JP_Sales — Float (penjualan di Japan dalam juta unit)
    - Other_Sales — Float (penjualan di wilayah lain)
    - Global_Sales — Float (penjualan global)

3.3 Preprocessing Dataset

Sebelum data digunakan dalam proses pemodelan atau analisis, diperlukan tahap preprocessing atau pra-pemrosesan data. Tahap ini sangat penting untuk memastikan bahwa data yang digunakan bersih, relevan, dan dalam format yang dapat dipahami oleh algoritma *machine learning*. Tanpa preprocessing, model yang dibangun bisa menghasilkan hasil yang tidak akurat atau bahkan gagal berfungsi. Adapun langkah-langkah preprocessing yang dilakukan pada penelitian ini adalah sebagai berikut:

1. **Menghapus baris dengan nilai kosong di kolom Year dan Publisher**
Nilai kosong (missing values) pada kolom tahun rilis (Year) dan penerbit (Publisher) dapat menyebabkan error atau bias dalam proses pelatihan model. Oleh karena itu, baris-baris yang memiliki nilai kosong pada kolom ini dihapus agar hanya data valid yang digunakan dalam analisis.
2. **Memilih fitur: Year, Platform, Genre dan target: Global_Sales**
Dari keseluruhan dataset, hanya kolom yang relevan dengan tujuan prediksi yang dipilih. Kolom Year, Platform, dan Genre dipilih sebagai fitur (variabel independen) yang diasumsikan berpengaruh terhadap penjualan, sedangkan Global_Sales dipilih sebagai target (variabel dependen) yang ingin diprediksi.
3. **Melakukan encoding terhadap kolom kategori (Platform dan Genre) menggunakan *one-hot encoding***
Algoritma seperti Linear Regression hanya dapat memproses input numerik. Karena Platform dan Genre merupakan data kategori (teks), maka dilakukan *one-hot encoding* untuk mengubahnya menjadi bentuk numerik. Dengan metode ini, setiap kategori akan direpresentasikan sebagai kolom biner (0 atau 1), tanpa memberi urutan atau bobot yang tidak semestinya.

4. Hasil dan Pembahasan

4.1 Pengujian Dataset

Untuk menguji seberapa baik model Linear Regression memprediksi nilai Global_Sales, dilakukan serangkaian langkah berikut:

4.1.1 Pemilahan Fitur dan Target

Langkah pertama adalah memisahkan antara fitur (variabel input) dan target (variabel output).
Kolom Global_Sales digunakan sebagai target (y) dan sisa kolom yang telah di encoding menjadi fitur (X) untuk pelatihan model.

4.1.2 *Split Data* untuk Training dan Testing

Dataset kemudian dibagi menjadi dua bagian, yaitu:

```
# 5. Splitting Data
# Memisahkan fitur (X) dan target (y)
X = df_selected.drop(['Global_Sales'], axis=1)
y = df_selected['Global_Sales']

# Membagi data menjadi training dan testing set (80:20) dengan random_state
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Gambar 2 *Splitting Data*

- **80% untuk data pelatihan (training)**
- **20% untuk data pengujian (testing)**

Pembagian ini dilakukan secara acak namun konsisten dengan menggunakan parameter random_state=42 agar hasil eksperimen dapat direproduksi.

4.1.3 Mencari Nilai Koefisien

Sebelum menghitung menggunakan rumus Linear Regression yaitu mencari koefisien tiap fitur dan disini kami menggunakan bantuan Google Collaboratory dan diapatkan hasilnya seperti gambar dibawah:

```
Model Koefisien:
Year: -0.1587
Platform_3DS: 6.0013
Platform_DC: 2.1837
Platform_DS: 6.2763
Platform_GB: 5.3856
Platform_GBA: 3.5129
Platform_GC: 3.1729
Platform_GEN: 1.9992
Platform_N64: 3.9987
Platform_NES: 2.8196
Platform_PC: 4.2929
Platform_PS: 3.0117
Platform_PS2: 3.9722
Platform_PS3: 5.2694
Platform_PS4: 6.0624
Platform_PSP: 4.1978
Platform_PSV: 3.9051
Platform_SAT: 2.2288
Platform_SCD: 0.4426
Platform_SNES: 2.9427
Platform_Wii: 7.3680
Platform_WiiU: 6.0348
Platform_X360: 5.1764
Platform_XB: 3.1652
Platform_XOne: 5.0381
Genre_Adventure: -0.7190
Genre_Fighting: -0.1613
Genre_Misc: -0.3789
Genre_Platform: 0.8774
Genre_Puzzle: -0.1526
Genre_Racing: 0.5300
Genre_Role-Playing: 0.6349
Genre_Shooter: 0.7331
Genre_Simulation: -0.4177
Genre_Sports: 0.1855
Genre_Strategy: -0.9569
```

Gambar 3 Hasil koefisien tiap fitur

- Setelah diketahui nilai koefisienya kita bisa memasukannya menggunakan rumus Linear Regression. Koefisien positif ➔ Fitur tersebut meningkatkan nilai prediksi Global_Sales.

- Koefisien negatif ➔ Fitur tersebut menurunkan nilai prediksi Global_Sales.

4.1.4 Pelatihan Model Linear Regression

Model Linear Regression dari pustaka *scikit-learn* digunakan dan dilatih menggunakan data training yang telah disiapkan.

```python
# 6. Training Model Linear Regression
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

model = LinearRegression()
model.fit(X_train, y_train)
```

Gambar 4 Uji model

Setelah proses ini, model siap digunakan untuk melakukan prediksi pada data uji (X_test), dan hasil prediksinya dapat dibandingkan dengan nilai sebenarnya (Y_test) untuk

mengevaluasi kinerja model menggunakan metrik seperti *Mean Squared Error* (MSE)*, Mean Absolute Error* (MAE)*,* dan *R² Score.*

4.1.5 Evaluasi Model Linear Regression

Sebelumnya model telah dievaluasi dengan bantuan Google Collaboratory untuk menghitung nilai-nilai akurasi seberapa baik modelnya dan didapatkan nilainya seperti gambar dibawah.

```
Evaluasi Model:
R2 Score: -0.11
Mean Squared Error: 7.56
Root Mean Squared Error: 2.75
Mean Absolute Error: 1.78
```

Gambar 5 Evaluasi model

- *R² Score* = -0.1111 ➔ Ini menunjukkan bahwa model tidak mampu menjelaskan variasi data sama sekali. Nilai negatif berarti model lebih buruk dibanding sekadar menebak rata-rata.

- MSE (*Mean Squared Error*) = 7.5634 ➔ Artinya, rata-rata kuadrat error prediksi cukup besar. Ini menunjukkan bahwa prediksi model sering meleset jauh dari nilai sebenarnya.

- RMSE 2.75 menunjukkan bahwa rata-rata, prediksi model Anda meleset sekitar 2.75 unit dari nilai aktual 'Global_Sales'.

- MAE 1.78 menunjukkan bahwa rata-rata, prediksi model Anda meleset sekitar 1.78 unit dari nilai aktual 'Global_Sales'. MAE lebih mudah diinterpretasikan daripada MSE karena menggunakan satuan yang sama dengan variabel target.

4.2 Rumus Metode

Setelah proses pelatihan model menggunakan algoritma Linear Regression, langkah selanjutnya adalah memahami bagaimana model menghasilkan prediksi secara matematis. Linear Regression memprediksi nilai target (dalam hal ini Global_Sales) berdasarkan kombinasi nilai fitur dan bobot (koefisien) yang diperoleh saat pelatihan. Berikut adalah bentuk umum rumus Linear Regression:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

di mana:

- $\hat{y}$ = nilai prediksi Global Sales
- β_0 = intercept
- β_n = koefisien fitur
- x_n = nilai fitur ke-n

4.3 Uji Coba Perhitungan Dataset

Misalkan kita ingin memprediksi Global_Sales untuk sebuah game dengan ciri sebagai berikut:

- Tahun rilis: 2008
- *Platform*: Wii
- *Genre*: Sports

Maka fitur *one-hot encoding* yang aktif:

- *Year* = 2008
- *Platform_Wii* = 1 (fitur lain bernilai 0)
- *Genre_Sports* = 1 (fitur lain bernilai 0)

Substitusi ke dalam rumus:

$$\hat{y} = 316.5448 + (-0.1587 \times 2008) + (7.3680 \times 1) + (0.1855 \times 1)$$

Perhitungan:

- $-0.1587 \times 2008 = -318.758$
- $7.3680 \times 1 = 7.368$
- $0.1855 \times 1 = 0.1855$

Gabungkan hasil:

$$\hat{y} = 316.5448 - 318.758 + 7.368 + 0.1855 = \mathbf{5.34\ juta\ unit\ (dibulatkan)}$$

4.4 Hasil Dataset

Setelah proses pelatihan model selesai dilakukan, tahap selanjutnya adalah menguji dataset untuk menghasilkan prediksi nilai Global_Sales. *Dataset* yang telah melalui proses preprocessing kemudian digunakan untuk menghasilkan prediksi menggunakan model *Linear Regression*. Hasil prediksi ini menunjukkan estimasi penjualan global dari masing-masing game berdasarkan fitur yang dimiliki.

Berikut adalah cuplikan hasil prediksi Global_Sales dari model terhadap 5 data pertama serta beberapa data di bagian akhir:

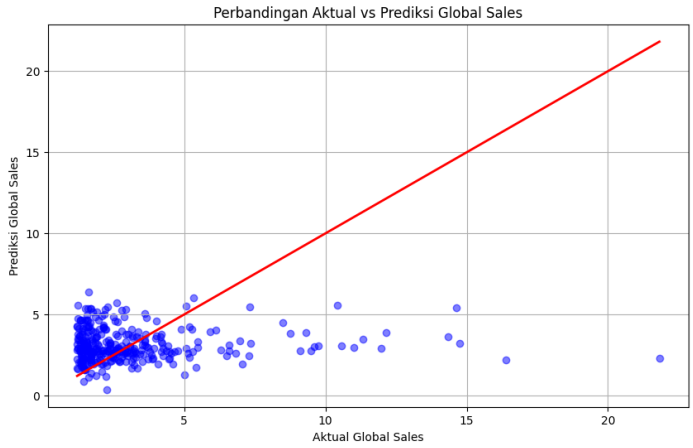
```
[51] # Memprediksi hasil pada data set
     y_pred = model.predict(X)

# Create a DataFrame from y_pred
y_pred_df = pd.DataFrame({'Predicted Global Sales': y_pred})

# Display the DataFrame as a table
y_pred_df
```

	Predicted Global Sales
0	5.236151
1	4.897585
2	5.744361
3	5.236151
4	5.747249
...	...
1632	2.439029
1633	2.352141
1634	3.349355
1635	4.077879
1636	4.032904

1637 rows × 1 columns

Gambar 6 Hasil prediksi menggunakan model

Setelah mengetahui nilai prediksi kita bandingkan dengan nilai aktualnya dan didapatkan hasilnya seperti gambar grafik dibawah



Gambar 7 Perbandingan nilai aktual vs nilai predikasi

5 Kesimpulan dan Saran

5.1 Kesimpulan

Linear Regression mampu memberikan gambaran kasar tentang hubungan fitur dengan penjualan game, akan tetapi tidak cocok untuk data ini karena

- Relasi antara fitur (Year, Platform, Genre) dan target (Global_Sales) kemungkinan tidak linear [3].
- Fitur-fitur yang digunakan belum cukup menjelaskan penjualan global suatu game.

Akurasi model masih rendah, menunjukkan bahwa prediksi Global_Sales memerlukan:

- Fitur tambahan seperti popularitas game, harga, marketing, dsb.
- Model lebih cocok menggunakan metode lain seperti Random Forest atau XGBoost [4][12].

5.2 Saran

Langkah selanjutnya dapat berupa:
- Menambahkan fitur baru.
- Menggunakan model yang lebih kompleks.
- Melakukan tuning hyperparameter.

Referensi

[1] Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Elsevier.

[2] Newzoo. (2023). *Global Games Market Report*. Retrieved from https://newzoo.com

[3] Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

[4] Abdullah, H. A., Putra, D. R. D., & Azhar, Y. (2022). Analisa Penjualan Video Game Menggunakan Metode Ensemble. JUST IT: Jurnal Sistem Informasi, Teknologi Informasi dan Komputer, 13(2), 97–102. https://jurnal.umj.ac.id/index.php/just-it/article/view/6746

[5] Yulianti, Y., Utami, D. Y., & Hikmah, N. (2019). Implementasi Data Mining Menentukan Game Android Paling Diminati Dengan Algoritma Apriori. Jurnal Paradigma, 21(2), 131–138. http://ejournal.bsi.ac.id/ejurnal/index.php/paradigma/article/view/4941

[6] Husni, D. T., Sitompul, D. R. H., Sinurat, S. H., et al. (2022). Analisis Big Data Penjualan Video Games Menggunakan EDA. Jurnal TEKINKOM, 5(1), 43–47. https://jurnal.murnisadar.ac.id/index.php/Tekinkom/article/view/517

[7] Duran, P. A., Vitianingsih, A. V., Riza, M. S., Maukar, A. L., & Wati, S. F. A. (2024). Data Mining Untuk Prediksi Penjualan Menggunakan Metode Simple Linear Regression. TEKNIKA: Jurnal Teknik Informatika, 13(1), 27–34. https://doi.org/10.34148/teknika.v13i1.712

[8] Royan, K., & Aathis, R. (2020). Perbandingan Pemodelan Linear dan Regresi pada Prediksi Penjualan Konsol Video Game. Universitas Multimedia Nusantara.

[9] Jamil, H. M., & Sulianta, F. (2023). Analisis Pengaruh Harga dan Ulasan Terhadap Penjualan Video Game Menggunakan Algoritma K-Means Clustering. ResearchGate.

[10] Sari, N. M., et al. (2021). Analisis Penjualan Produk Digital Menggunakan Regresi Linier Berganda. Jurnal Teknologi Informasi dan Ilmu Komputer, 8(2), 150–158.

[11] Handayani, R., Prasetyo, Y. L., & Fauzan, A. (2023). Analisis Preferensi Platform Game Terhadap Penjualan Game Digital. Jurnal Teknologi dan Informatika, 12(1), 45–50.

[12] Ramadhani, N., Fitri, L. R., & Zahra, A. (2022). Prediksi Penjualan Produk Retail Menggunakan XGBoost. Jurnal Informatika dan Komputer, 10(1), 22–29.