PEMODELAN DAN ANALISA FAKTOR-FAKTOR YANG BERKONTRIBUSI TERHADAP PENJUALAN PROPERTI DI NYC MENGGUNAKAN REGRESI LINEAR

M. Agus Munandar*[1], Terttiaavini[2] , Khoirusy Syafaat[3] , Parhan Oktaria Putra[4]
Universitas Indo Global Mandiri[1,2,3,4]
Jl. Jend. Sudirman Km.4 No.629, 20 Ilir D.IV, Kec. Ilir Tim I, Kota Palembang
Email: agus@uigm.ac.id[1], avini.saputra@uigm.ac.id[2], khoirusysyafaat@students.uigm.ac.id[3], farhanop@students.uigm.ac.id[4]

ABSTRAK

Pasar penjualan properti di New York City (NYC) merupakan salah satu yang paling dinamis di dunia, dipengaruhi dengan berbagai macam faktor. Penelitian ini bertujuan untuk mengidentifikasi serta menganalisa faktor-faktor apa saja yang memengaruhi harga jual properti di NYC menggunakan pendekatan kuantitatif dengan menggunakan *Exploratory Data Anlysis* (EDA) dan juga menggunakan model regresi. Data yang digunakan bersumber dari NYC *property Sales Dataset* yang didapatkan dari situs kaggle, dimana mencakup transaksi penjualan selama satu tahun. Penelitian ini melalui pra-pemrosesan data dimana dilakukan untuk mengatasi data yang hilang ataupun outlier, selanjutnya yaitu EDA dilakukan untuk mengungkapkan atau mencari pola hubungan antar variabel, sedangkan pada proses model regresi yaitu digunakan untuk menunjukkan varaibel-variabel seperti borought, taxclass, buildingclass serta status bangunan prewar ataupun postwar memiliki pengaruh yang signifikan terhadap harga jual. Hasil dari penelitian ini menunjukkan properti yang ada di Manhattan, kelas pajak 2, dan yang termasuk pre-war cenderung memiliki harga yang cukup tinggi, dan juga variabel numerik seperti luas bangunan memiliki korelasi hubungan terhadap harga jual. Hasil model regresi yang dikembangkan hanya mampu mendapatkan hasil 21,9 % variasi harga jual, disebabkan ada faktor-faktor lain yang memengaruhi nya.

Kata kunci: EDA, NYC, penjualan properti, regresi linear.

ABSTRACT

The propery sales market in New York City (NYC) is among the most dynamic in the world, influenced by a wide range of factors. This study aims to identify and analyze the factors that influence property sales price in NYC using quantitative approach, employing Exploratory Data Analysis (EDA) and regression model. The data utilized in this reasearch is sourced from the NYC Property Sales Dataser available on Kaggle, which covers sales transactions over one year. The research process involves data preprocessing to address missing values and outliers, followed by EDA to uncover patterns and relationships between variables. Subsequently, regression modeling is applied to access the significant impact of variables such as borought, tacclass, buildingclass, and pre-war or post-war building status on sales prices. The findings reveal that properties located in Manhattan, classified under taxclass 2, and categorized as pre-war tend to have realtively high prices. Additionally, numerical variables such as building size show a significant correlation with sales prices. However, the developed regression model explains only 21,9 % of the variation in sales prices, indicating that other factors also influence these prices.

Keywords: EDA, NYC, sales property, linear regression.

1. PENDAHULUAN

Industri properti di New York City (NYC) merupakan salah satu sektor yang paling dinamis dan menjanjikan, dengan potensi keuntungan yang signifikan. Namun, penjualan properti sering kali mengalami perubahan yang kadang naik bahkan kadang turun yang dipengaruhi oleh berbagai faktor. Salah satu tantangan utama dalam transaksi properti adalah kesulitan dalam memperoleh data yang akurat mengenai harga jual. Kesenjangan informasi ini dapat disebabkan oleh berbagai macam faktor, termasuk infrastruktur yang tidak merata di berbagai wilayah di NYC dan kurangnya promosi yang efektif. Oleh karrna itu, sangat penting untuk menganalisis data penjualan properti yang mencakup lokasi, jenis properti, harga dan waktu penjualan untuk memahami dinamika pasar [1].

Melalui analisis ini, diharapkan dapat teridentifikasi area-area yang populer dan juga faktor-faktor yang mempengaruhi nilai properti, pada penelitian ini juga bertujuan untuk mencari tren-tren yang dapat menjadi masukan bagi para investor dan juga pengembang properti dalam pengambilan keputusan investasi di masa yang akan datang. Dengan demikian, penelitian ini tidak hanya akan memberikan analisis kinerja penjualan tetapi juga dapat merekomendasikan strategi untuk meningkatkan penjualan dimasa-masa yang akan datang.

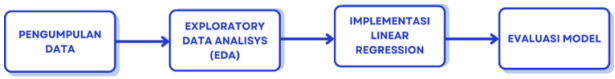
Beberapa penelitian terdahulu yang membahas penggunaan *machine learning* yaitu [2] [3] bagaimana menentukan sebuah model yang paling optimal dalam penggunaan algoritma *machine learning* yaitu Random Forest, Logistic Regression, Naive Bayes, SVM dan Neural Networks, serta bagaimana cara mendapatkan hasil evaluasi berdasarkan tingkat Accuracy, Precission, dan juga Recall dalam penghitungan pada *machine learning* tersebut.

Penelitian ini, menggunakan Eksplorasi Data Analisis (EDA) dan juga menggunakan model regresi, dimana EDA akan digunakan untuk mengeksplorasi data secara visual dan statistik yang digunakan untuk menemukan pola-pola yang ada. Selanjutnya, model regresi akan diterapkan untuk mengukur pengaruh berbagai variabel terhadap penjualan properti [4], dimana data penelitian ini diambil dari dataset yang ada di kaggle, penelitian ini juga akan menganalisis hubungan antara berbagai faktor dan tingkat penjualan properti. Dimana hasil dari penelitian ini diharapkan dapat tidak hanya untuk memberikan gambaran menyeluruh tentang kondisi pasar properti di NYC tetapi juga menghasilkan rekomendasi strategi untuk meningkatkan penjjualan properti [5].

Tujuan penelitian ini adalah untuk mengidentifikasi faktor-faktor apa saja yang mempengaruhi penjualan properti di NYC dengam menggunakan pendekatan *Exploratory Data Analysis* (EDA) dan juga menggunakan model algoritma regresi, dimana EDA akan digunakan untuk mengekplorasi data penjualan di NYC dengan berfokus pada variabel-variabel seperti lokasi, jenis properti, harga dan juga fasilitas yang tersedia. Melalui analisis ini nantinya diharapkan dapat ditemukan pola-pola yang relevan yang dapat memberikan wawasan bagi pengambil keputusan dalam industri properti tersebut.

2. METODE PENELITIAN

Metodologi penelitian ini dirancang untuk menganalisis faktor-faktor apa saja yang memengaruhi penjualan properti di New York City (NYC) dengan menggunakan pendekatan *Exploratory Data Analysis* (EDA) serta menggunakan model regresi. Adapun tahapan-tahapan yang dilakukan dalam penelitian ini dapat dilihat pada Gambar 2.1 dibawah ini.



Gambar 2.1. Langkah-langkah Penelitian

2.1 Pengumpulan Data

Penelitian ini menggunakan dataset yang diambil dari www.kaggle.com dengan judul NYC Property Sales, dimana dataset ini adalah sekumpulan data catatan setiap gedung ataupun unit bangunan yang terjual di kota New York dalam rentang waktu 1 tahun.

2.2 *Exploratory Data Analysis* (EDA)

EDA merupakan suatu metode yang digunakan untuk menganalisis suatu data dengan bentuk visual dimana untuk tujuan memperoleh suatu pemahaman yang lebih baik akan suatu informasi, EDA mempunyai peranan penting yaitu, pertama, EDA dapat membantu dalam pengumpulan dan pemahaman terkait suatu informasi, kedua, EDA dapat menghilangkan suatu kesalahan, mengetahui outliers, dan ketiga EDA dapat mengumpulkan data untuk dianalisis dengan hasil yang lebih spesifik dalam bentuk tampilan visual [6]. Adapun langkah-langkah yang dilakukan sebelum melakukan EDA adalah sebagai berikut [7] :

a. Data *Preparation*
Pada tahapan ini menggunakan data sekunder yaitu data yang didapatkan dari sumber kedua, yang sudah berbentuk dataset

b. Data *Cleaning*
Pada tahapan Data *cleaning* adalah proses perbaikan dari data yang ada ataupun pembersihan data berdasarkan kebutuhan, dimana proses ini yaitu membersihkan, memperbaiki data ataupun menghapus data sehingga data yang sudah dibersihkan dapat digunakan untuk keperluan analisis serat pengambilan keputusan yang lebih signifikan.

c. Visualisasi
Pada tahapan ini EDA digunakan untuk menjawab dari pertanyaan-pertanyaan serta membuat hipotesis untuk analisa lanjutan dimana akan ditampilan dalam bentuk visual [8]

2.3 Implementasi *Linear Regression*

Linear regression adalah salah satu metode yang ada pada sebuah *machine learning* yang merupakan suatu metode statistika yang dapat berfungsi untuk menguji hubungan atau korelasi antar sebab akibat dan faktor dependen terhadap faktor independen. Tujuan penggunaan analisis model regresi adalah untuk dapat mengetahui ataupun mengestimasi nalai dari variabel dependen terhadap variabel independen [9][10]. Adapun untuk rumus persamaan dari linear regression adalah sebagai berikut :

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

Persamaan diatas, untuk Y melambangkan variabel dependen yang bergantung pada variabel X yaitu variabel independen. Sedangkan untuk nilai a adalah suatu konstanta dan b merupakaan sebuah koofesien regresi yang berkaitan dengan varibel X [11][12]

2.4 Evaluasi Hasil

Tahap evaluasi model yaitu dengan menggunakan *Mean Squared Error* (MSE), *Mean Absolute Error* (MAE) dan *Root Mean Square Error* (RMSE) [13], [14]. Adapun

P-ISSN: 2337-8344
E-ISSN: 2623-1247

formula persamaan dalam evaluasi model tersebut adalah sebagai berikut :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^{real} - y_i^{pred}\right)^2$$

$$MAE = \frac{1}{n}\sum_{i=0}^{n}|y_i^{real} - y_i^{pred}|$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i^{real} - y_i^{pred}\right)^2}$$

3. HASIL DAN PEMBAHASAN

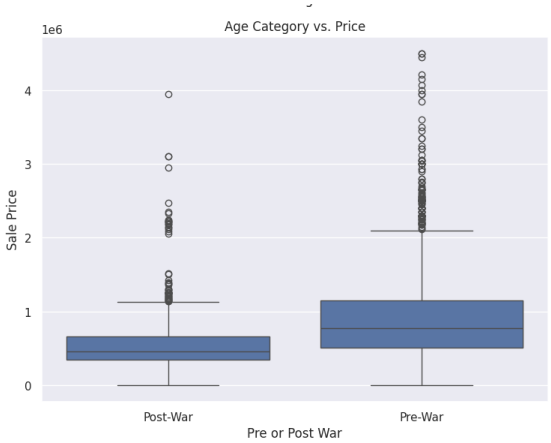
Hasil analisis penelitian ini menggunakan pemrograman python melalui platform google *collaboratory* untuk menganalisis faktor apa saja yang memengaruhi penjualan properti.

3.1. Analisis Explorasi Data (EDA)

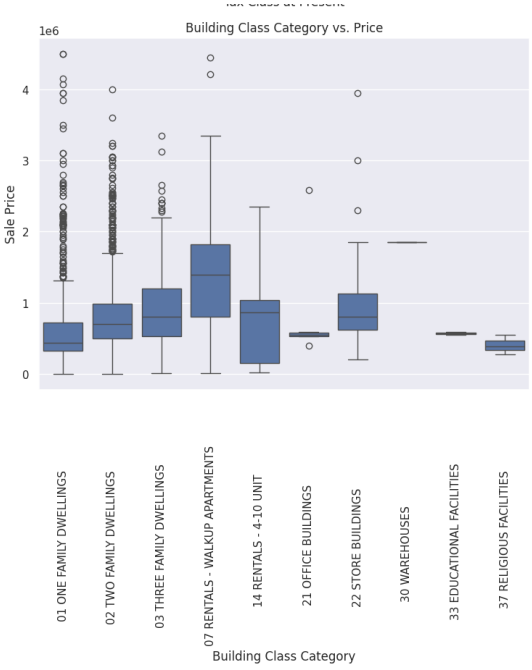
EDA adalah tahapan yang sangat dibutuhkan dalam menganalisis dari dataset yang ada, dimana langkah awal yang dilakukan adalah memastikan dataset yang ada apakah terdapat *missing value*, dan bagaimana menangani apabila terdapat *missing value*, dan juga untuk menangani apabila terdapat outliers dari data tersebut. Pada Gambar 3.1 dibawah ini menunjukkan hubungan korelasi antara *categorical features* terhadap *sales price.*

(a)

(b)



(c)



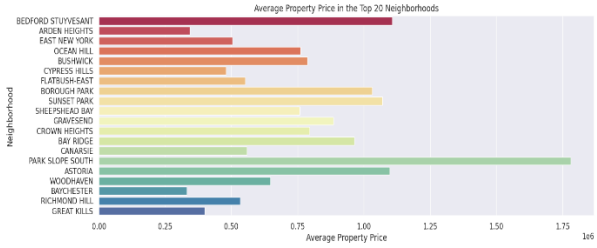
(d)

Gambar 3.1. Hubungan antara *Categorical Features* vs *Sales Price*

Hubungan borough (wilayah) terhadap *sales price* menunjukkan bahwa pada wilayah Manhattan memiliki harga jual yang paling tinggi, sedangkan brooklyn berada pada posisi ke dua, lalu diikuti oleh queens, sedangkan untuk the bronx dan staten menunjukkan harga yang paling rendah, sehingga dari data tersebut menunjukkan pada Manhattan adalah wilayah dengan harga properti yang tertinggi dan terdapat harga yang cukup signifikan di antar wilayah. Selanjutnya pada gambar tax class vs sales price kelas pajak 2 memiliki harga yang tertinggi, variasi data yang ditunjukkan pada gambar pada setiap kelas pajak cukup tinggi terlihat dari rentang boxplot yang cukup luas. Pada gambar Age Category vs Sales price menunjukkan bahwa harga

properti pre-war atau bangunan yang dibangun sebelum tahun 1945 mempunyai harga yang leibh tinggi apabila dibandingkan dengan properti post-war. Selanjutnya pada gambar Building class Category vs sales price memperlihatkan bahwa ada perbedaan berbagai kategori kelas bangunannya, dan pola yang sulit untuk diidentifikasi.

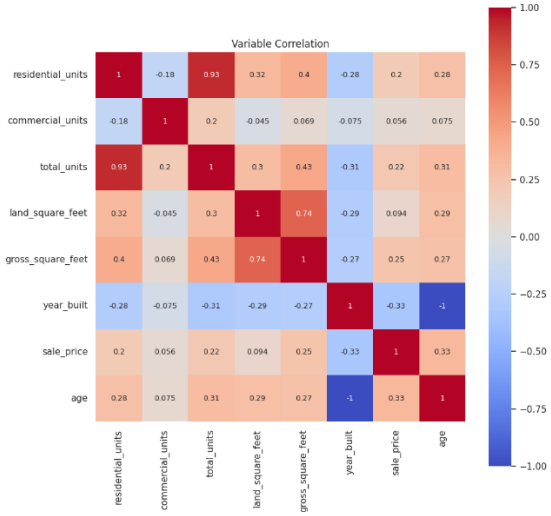
Gambar 3.2 dibawah ini menujukkan hasil analisis yang dilakukan di 20 lingkungan dengan harga rata-rata properti yang tertinggi, dimana hasil grafik yang ditunjukkan pada park slope south menduduki tempat tertinggi sebagai lingkungan dengan harga rata-rata tertinggi lalu dilanjutkan oleh Bedford Stuyvesant dan Astgoria



Gambar 3.2. Average Property Price

3.2. Korelasi Antar Variabel

Hubungan antar variabel numerik dalam dataset penjualan properti di new york city akan ditunjukkan pada Gambar 3.3 dibawah ini, dimana pada visualisasi menunjukkan koofesien nilai antara -1 hingga +1 dimana, +1 menunjukkan bahwa korelasi positif sempurna, variabel nya bergerak ke arah yang sama, sedangkan 0 menunjukkan bahwa tidak ada korelasi linear, dimana perubahan pada satu variabel tidak terkait dengan variabel yang lainnya, dan untuk nilai -1 menunjukkan korelasi negatif sempurna, dimana kedua varaibel nya bergerak kearah yang berlawanan dan juga untuk nilai absolut yang mendekati angak +1 ataupun -1 menunjukkan nilai-nilai absolut yang mengindikasikan kekuatan korelasi yang lebih kuat.



Gambar 3.3. Variabel Correlation

3.3. *Linear Regression*

Model yang dibangun untuk menentukan harga jual properti di new york city yaitu menggunakan model regresi linear, dimana sebelumnya setelah melakukan analisa hubungan antara variabel numerik dengan variabel kategorikal dengan harga jual. Adapun model regresi ini dibangun dengan menggunakan variabel-variabel numerik sebagai prediktor, yaitu : unit hunian, unit komersal, total unit,luas tanah, luas kotor, tahun pembangunan, dan usia dari bangunan. Pada model linear regression yang digunakan menggunakan data training (latih) dan juga data Testing (uji) untuk membangun model tersebut.

Proses evaluasi kinerja model menggunakan regresi linear yaitu beberapa metrik evaluasi, yaitu :

1. *Mean Absolute Error* (MAE) : yaitu mengukur rata-rata selisih nilai absolute anatara nilai prediksi dan nilai aktual.
2. *Mean Squared Error* (MSE) : yaitu mengukur rata-rata kuadrat selisih antara nilai prediksi dan nilai aktual, MSE akan memberikan bobot yang lebih besar pada kesalahan yang lebih besar.
3. *Root Mean Squared Error* (RMSE) : yaitu merupakan akar kuadrat dari MSE, memberikan ukuran suatu kesalahan dalam satuan yang sama dengan variabel target (harga jual).
4. *R-Squared* (R^2) : yaitu mengukur proporsi variasi daloam variabel target (harga jual), nilai R^2 berkisar antara 0 dan 1, dimana nilai yang lebih tinggi menunjukkan model yang lebih baik.

Berdasarkan dari model yang sudah dibangun dapat dibuatkan rumus persamaan sebagai berikut:

$$sale_{price} = \beta^0 + \beta^1 * residential_{units} + \beta^2 * commercial_{units} + \beta^3 * total_{units} + \beta^4 * land_{square_{feet}} + \beta^5 * gross_{square_{feet}} + \beta^6 * year_{built} + \beta^7 * age$$

Ket : $\beta^0 = Sale_{price} = 0$;
$\beta^1, \beta^2, \beta^3, \beta^4, \beta^5, \beta^6, \beta^7 =$ *Koofesien variabel independen* ;

Tabel 1 dibawah ini akan menunjukkan hasil dari MAE. MSE. RMSE, dan juga R^2 :

Tabel 1. Hasil Model Evaluasi

Model	Nilai
MAE	360313.8977874546
MSE	263022765190.83694
RMSE	512857.4511409939
R-Squared	0.21858531389210922

4.. KESIMPULAN DAN SARAN

4.1 Kesimpulan

Penelitian ini adalah untuk mengidentifikasi faktor-faktor apa saja yang akan memengaruhi penjualan properti di NYC menggunakan EDA dan juga model regresi dimana proses EDA dilakukan untuk mencari hubungan pola antar variabel, sedangkan model regresi digunakan untuk menghitung pengaruh dari faktor-faktor apa saja terhadap harga jual properti. Berdasarkan hasil yang didapatkan bahwa faktor yang memengaruhi penjualan properti di NYC adalah pertama, variabel kategorikal seperti borough, taxclass, dan juga buildingclass memiliki pengaruh yang sangat signifikan terhadap harga jual, properti yang ada di Manhattan memiliki harga yang lebih tinggi dibandingkan dengan yang lainnya, kedua varaibel numerik juga memperlihatkan pengaruh terhadap harga jual, semakin besar luas bangunannya maka akan berbanding lurus dengan harga jual, sehingga dapat disimpulkan properti yang lebih tua dengan luas yang tidak besar akan cenderung lebih mahal harga nya, dan juga berdasarkan hasil yang didapatkan menggunakan model regresi hanya mendapatkan hasil sekitar 21,9 % saja dari variasi harga jual, ini menandakan terdapat faktor-faktor lain yang belum atau tidak terukur dari model yang ada dalam hal yang mempengaruhi harga jual.

4.2 Saran

Berdasarkan dari hasil penelitian yang sudah dilakukan ada beberapa saran yang dapat dilakukan untuk pengembangan yang lebih baik yaitu dengan menambahkan variabel-variabel lain yang lebih dapat memperkuat hasil yang memengaruhi apa saja dan dapat menggunakan model yang lain untuk membandingkan tingkat keakuratannya.

DAFTAR PUSTAKA

[1] F. Husna I, "Implementasi Data Analytic Dalam Upaya Peningkatan Penjualan Properti Sebesar 10% di NYC Amerika Serikat," *Venus: Jurnal Publikasi Rumpun Ilmu Teknik*, vol. 2, no. 1, pp. 134–144, 2024.

[2] N. F. Sahamony, T. Terttiaavini, and H. Rianto, "Analisis Perbandingan Kinerja Model Machine Learning untuk Memprediksi Risiko Stunting pada Pertumbuhan Anak," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 2, pp. 413–422, Feb. 2024, doi: 10.57152/malcom.v4i2.1210.

[3] I. P. Putri, T. Terttiaavini, and N. Arminarahmah, "Analisis Perbandingan Algoritma Machine Learning untuk Prediksi Stunting pada Anak," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 1, pp. 257–265, Jan. 2024, doi: 10.57152/malcom.v4i1.1078.

[4] G. N. Ayuni and D. Fitrianah, "Penerapan Metode Regresi Linear Untuk Prediksi Penjualan Properti pada PT XYZ," *Jurnal Telematika*, vol. 14, no. 2, pp. 79–86, Apr. 2020, doi: 10.61769/telematika.v14i2.321.

[5] M. E. N. Lisna and A. Voutama, "Analisis Dan Visualisasi Data Penjualan Pada NYC Property Menggunakan EDA," *JELIKU (Jurnal Elektronik Ilmu Komputer Udayana)*, vol. 12, no. 4, p. 879, May 2024, doi: 10.24843/JLK.2024.v12.i04.p13.

[6] P. Sherly, M. Christopher Parulian, W. Lucky Raffael, P. Manisha Arie, A. Willyanto Dharmansyah, and N. Rahmi Yulia, "Implementasi Exploratory Data Analysis (EDA) Untuk Menganalisis Berbagai Faktor Risiko Penyakit Jantung Di Amerika Serikat," *https://jurnal.ulb.ac.id/index.php/JoSDIS/issue/view/363*, vol. 3, no. 2, pp. 108–124, Jul. 2023, doi: https://doi.org/10.36987/josdis.v3i2.4563.

[7] I. N. Rizki, D. Prayoga, M. L. Puspita, and M. Q. Huda, "IMPLEMENTASI EXPLORATORY DATA ANALYSIS UNTUK ANALISIS DAN VISUALISASI DATA PENDERITA STROKE KALIMANTAN SELATAN MENGGUNAKAN PLATFORM TABLEAU," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 1, Jan. 2024, doi: 10.23960/jitet.v12i1.3856.

[8] I. Hidayat, A. I. Tolago, R. D. R. Dako, and J. Ilham, "Analisis Data Eksploratif Capaian Indikator Kinerja Utama 3 Fakultas Teknik," *Jambura Journal of Electrical and Electronics Engineering*, vol. 5, no. 2, pp. 185–191, Jul. 2023, doi: 10.37905/jjeee.v5i2.18397.

[9] F. H. Hamdanah and D. Fitrianah, "Analisis Performansi Algoritma Linear Regression dengan Generalized Linear Model untuk Prediksi Penjualan pada Usaha Mikra, Kecil, dan Menengah," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 10, no. 1, p. 23, Apr. 2021, doi: 10.23887/janapati.v10i1.31035.

[10] S. Lestari, "Analisis Algoritma Regresi Linear Sederhana dalam Memprediksi Tingkat Penjualan Album KPOP," *INSOLOGI: Jurnal Sains dan Teknologi*, vol. 2, no. 1, pp. 199–209, Feb. 2023, doi: 10.55123/insologi.v2i1.1692.

[11] T. Indarwati, T. Irawati, and E. Rimawati, "PENGGUNAAN METODE LINEAR REGRESSION UNTUK PREDIKSI PENJUALAN SMARTPHONE," *Jurnal Teknologi Informasi dan Komunikasi (TIKomSiN)*, vol. 6, no. 2, Jan. 2019, doi: 10.30646/tikomsin.v6i2.369.

[12] U. Lathifah and R. Danar Dana, "IMPLEMENTASI METODE LINEAR REGRESSION UNTUK PREDIKSI HARGA PROPERTI REAL ESTATE

MENGGUNAKAN RAPIDMINER," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 1, pp. 1129–1137, Mar. 2024, doi: 10.36040/jati.v8i1.8919.

[13] M. R. Alifi, H. Hayati, and C. Fauzi, "Penerapan Algoritma Regresi Linier pada Prediksi Tarif Influencer Media Sosial," *Journal of Information System Research (JOSH)*, vol. 4, no. 1, pp. 210–218, Oct. 2022, doi: 10.47065/josh.v4i1.2361.

[14] M. D. H. Kusuma and S. Hidayat, "Penerapan Model Regresi Linier dalam Prediksi Harga Mobil Bekas di India dan Visualisasi dengan Menggunakan Power BI," *Jurnal Indonesia : Manajemen Informatika dan Komunikasi*, vol. 5, no. 2, pp. 1097–1110, May 2024, doi: 10.35870/jimik.v5i2.629.