Jurnal Komputer dan Teknologi Sains (KOMTEKS)
Vol. 3, No. 2, Oktober 2024, hlm. 13-17 e-ISSN: 2964-5247

ANALISIS PREDIKSI HARGA SMARTPHONE TAHUN 2023 MENGGUNAKAN MODEL RANDOM FOREST REGRESSION BERDASARKAN FITUR-FITUR SPESIFIKASI TEKNIS

Putu Satya Saputra*[1], I Putu Gede Abdi Sudiatmika[2]

[1,2]Jurusan Akuntansi, Politeknik Negeri Bali
[1]satya@pnb.ac.id, [2]sudiatmika.abdi@pnb.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 17 September 2024, diterima untuk diterbitkan: 13 Oktober 2024)

Abstrak

Penelitian ini bertujuan untuk menganalisis dan memprediksi harga smartphone tahun 2023 menggunakan model Random Forest Regression berdasarkan fitur-fitur spesifikasi teknis. Dataset yang digunakan berasal dari https://www.kaggle.com/datasets/howisusmanali/mobile-prices-2023, dengan jumlah data sebanyak 1837 dan 10 variabel, antara lain Phone Name, Rating ?/5, Number of Ratings, RAM, ROM/Storage, Back/Rare Camera, Front Camera, Battery, Processor, dan Price in INR. Tahap pertama adalah proses data collecting, diikuti dengan pembersihan data (data cleaning) dengan menghapus variabel yang tidak relevan, seperti Phone Name. Selanjutnya, dilakukan tahap data processing, di mana dataset dibagi menjadi data latih (90%) dan data uji (10%), kemudian dilakukan normalisasi dan standarisasi data. Hasil dari model Random Forest Regression yang dibangun menunjukkan Mean Squared Error (MSE) sebesar 9322848,510354057 dan R-squared sebesar 0.9546596207605785. Hasil evaluasi tersebut menunjukkan bahwa model memiliki tingkat akurasi yang tinggi dalam memprediksi harga smartphone berdasarkan fitur-fitur spesifikasi teknisnya.

Kata kunci: *random forest regression, model machine learning, evaluasi akurasi model*

ANALYSIS OF SMARTPHONE PRICE PREDICTION IN 2023 USING RANDOM FOREST REGRESSION MODEL BASED ON TECHNICAL SPECIFICATION FEATURES

Abstract

This research aims to analyze and predict the price of smartphones in 2023 using the Random Forest Regression model based on technical specification features. The dataset used is sourced from https://www.kaggle.com/datasets/howisusmanali/mobile-prices-2023, with a total of 1837 data points and 10 variables, including Phone Name, Rating ?/5, Number of Ratings, RAM, ROM/Storage, Back/Rare Camera, Front Camera, Battery, Processor, and Price in INR. The first stage involves data collecting, followed by data cleaning by removing irrelevant variables, such as Phone Name. Subsequently, the data processing stage divides the dataset into training data (90%) and testing data (10%), followed by data normalization and standardization. The results of the Random Forest Regression model show a Mean Squared Error (MSE) of 9322848,510354057 and an R-squared of 0.9546596207605785. These evaluation results indicate that the model has a high level of accuracy in predicting smartphone prices based on their technical specification features.

Keywords: *random forest regression, machine learning model, model accuracy evaluation*

1. PENDAHULUAN

Pada era digital ini, industri smartphone terus berkembang dengan pesat, menyajikan berbagai inovasi dan teknologi terbaru kepada konsumen. Dalam pasar yang begitu dinamis ini, prediksi harga smartphone memiliki peran yang krusial bagi produsen, pedagang, dan konsumen. Analisis prediksi harga dapat memberikan wawasan berharga tentang tren pasar, membantu dalam perencanaan strategis, dan meningkatkan daya saing perusahaan (Haque-Fawzi, 2022).

Random Forest Regression dapat menangani berbagai macam masalah regresi dengan baik, termasuk yang memiliki fitur yang kompleks dan korelasi yang tinggi (Ramadhansyah, 2022). Metode ini juga cenderung lebih tahan terhadap overfitting dibandingkan dengan pohon keputusan tunggal karena penggunaan banyak pohon yang berbeda.

Proses model *Random Forest Regression* dapat dijelaskan secara sederhana sebagai berikut (Religia et al., 2021). Pemilihan sampel: sejumlah sampel dari data latih diambil secara acak dengan penggantian (*bootstrap*), artinya beberapa data dapat diambil beberapa kali, sementara yang lain mungkin tidak diambil sama sekali. Pembangunan pohon keputusan: setelah sampel diambil, pohon keputusan dibangun menggunakan *subset* fitur yang dipilih secara acak dari seluruh fitur dataset. Proses ini dilakukan secara rekursif dengan membagi dataset menjadi *sub-dataset* menggunakan aturan pemisahan yang optimal untuk setiap node pohon. Pembuatan banyak pohon: langkah kedua diulangi beberapa kali untuk membuat sejumlah besar pohon keputusan, masing-masing dengan sampel dan fitur yang berbeda. Prediksi agregat: setelah semua pohon keputusan dibangun, prediksi dilakukan untuk setiap pohon. Hasil prediksi dari semua pohon tersebut kemudian diambil rata-ratanya untuk menghasilkan prediksi akhir.

Random Forest Regression adalah salah satu teknik dalam *machine learning* yang sering digunakan dalam masalah prediksi, termasuk dalam prediksi harga. Keunggulan utama dari model ini adalah kemampuannya untuk menangani *dataset* dengan banyak fitur dan observasi, serta mampu menangani non-linearitas dan interaksi antar fitur (Mustafa et al., 2024).

Namun, seperti halnya dengan setiap analisis prediksi, keterbatasan dataset yang digunakan perlu dipertimbangkan. Terdapat variabel yang tidak tersedia atau informasi yang hilang, yang dapat mempengaruhi kualitas prediksi. Oleh karena itu, pemahaman yang baik tentang dataset yang digunakan dan keterbatasannya sangat penting untuk memperoleh hasil yang akurat dan bermanfaat.

Evaluasi model dilakukan untuk mengukur seberapa baik model yang telah dibangun dalam memprediksi harga smartphone berdasarkan fitur-fitur spesifikasi teknisnya. Evaluasi model memberikan pemahaman tentang seberapa akurat dan dapat diandalkan model tersebut dalam menghasilkan prediksi harga (Raharja et al., n.d.). Penelitian ini menggunakan Mean Squared Error (MSE) dan R-squared untuk evaluasi model. Kedua metrik ini memiliki tujuan yang berbeda namun saling melengkapi dalam mengevaluasi kinerja model regresi. Mean Squared Error adalah metrik evaluasi yang mengukur rata-rata dari kuadrat selisih antara nilai prediksi dan nilai sebenarnya (Anam & Jakaria, 2023). Semakin kecil nilai MSE, semakin baik kualitas prediksi model. MSE dapat dihitung dengan menjumlahkan kuadrat selisih antara prediksi dan nilai sebenarnya, kemudian dibagi dengan jumlah total data (Noor, 2018). R-squared adalah metrik evaluasi yang menyatakan seberapa baik model regresi cocok dengan data yang diamati. R-squared memiliki rentang nilai antara 0 dan 1, di mana nilai 1 menunjukkan model yang sempurna cocok dengan data, sementara nilai 0 menunjukkan bahwa model tidak memberikan penjelasan apa pun terhadap variabilitas data (Azmi et al., 2020). Secara sederhana, *R-squared* adalah proporsi variabilitas dalam variabel respons yang dapat dijelaskan oleh model.

Dalam artikel ini, digunakan dataset spesifikasi teknis smartphone untuk tahun 2023. Data ini mencakup berbagai fitur teknis yang dapat memengaruhi harga smartphone. Penulis akan mengeksplorasi keunggulan model *Random Forest Regression* dalam memprediksi harga smartphone berdasarkan fitur-fitur spesifikasi tersebut. Tujuan akhir dari analisis ini adalah memberikan wawasan yang bermanfaat bagi para pemangku kepentingan dalam industri smartphone.

2. METODE PENELITIAN

Metode penelitian yang digunakan dalam penelitian ini adalah metode penelitian studi kasus dan analisis regresi. Alur penelitian dari tahap awal sampai tahap akhir penelitian ditujukkan pada Gambar 1.

Gambar 1. Alur Penelitian

Alur penelitian yang dibahas dalam artikel ini dijabarkan sebagai berikut:

1. Pengumpulan Data
   Data harga smartphone tahun 2023 dikumpulkan dari https://www.kaggle.com/datasets/howisusmanali/mobile-prices-2023. Dataset ini terdiri dari 1837 data dengan 10 variabel, yaitu Phone Name, Rating ?/5, Number of Ratings, RAM, ROM/Storage, Back/Rare Camera, Front Camera, Battery, Processor, dan Price in INR.

2. Pembersihan Data (Data Cleaning)
   Langkah pertama dalam pembersihan data adalah menghapus variabel 'Phone Name' karena variabel tersebut memiliki dampak minimal atau tidak relevan terhadap target prediksi. Selanjutnya, dilakukan pembersihan pada variabel lain sesuai kebutuhan, seperti mengubah tipe data, menghapus karakter tambahan, dan mengatasi data yang hilang atau tidak valid.

3. Data Processing
   Setelah data dibersihkan, dataset dibagi menjadi data latih (90%) dan data uji (10%). Pembagian ini penting untuk evaluasi model. Selanjutnya, dilakukan normalisasi dan standarisasi data numerik agar semua fitur memiliki skala yang seragam.

4. Pembangunan Model Random Forest Regression
   Model Random Forest Regression dibangun dengan menggunakan dataset latih. Proses ini melibatkan pembuatan banyak pohon keputusan dengan sampel dan fitur yang berbeda. Proses pembuatan pohon keputusan dilakukan secara rekursif dengan membagi dataset menjadi sub-dataset menggunakan aturan pemisahan yang optimal untuk setiap node pohon.

5. Evaluasi Model
   Model dievaluasi menggunakan metrik Mean Squared Error (MSE) dan R-squared. MSE mengukur rata-rata dari kuadrat selisih antara nilai prediksi dan nilai sebenarnya, sedangkan R-squared menyatakan seberapa baik model cocok dengan data yang diamati.

6. Analisis Hasil
   Hasil evaluasi model menunjukkan bahwa model Random Forest Regression memiliki tingkat akurasi yang tinggi dalam memprediksi harga smartphone berdasarkan fitur-fitur spesifikasi teknisnya.

3. HASIL DAN PEMBAHASAN

Dataset yang digunakan yaitu data harga *smartphone* tahun 2023 bersumber dari https://www.kaggle.com/datasets/howisusmanali/mobile-prices-2023. Jumlah data yaitu 1837 dengan 10 variabel yaitu Phone Name, Rating ?/5, Number of Ratings, RAM, ROM/Storage, Back/Rare Camera, Front Camera, Battery, Processor, Price in INR. Tahap ini termasuk dalam tahap *data collecting*.

Tabel 1. Dataset Harga Smartphone

```
Data columns (total 10 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Phone Name         1836 non-null   object
 1   Rating ?/5         1836 non-null   float64
 2   Number of Ratings  1836 non-null   object
 3   RAM                1836 non-null   object
 4   ROM/Storage        1662 non-null   object
 5   Back/Rare Camera   1827 non-null   object
 6   Front Camera       1435 non-null   object
 7   Battery            1826 non-null   object
 8   Processor          1781 non-null   object
 9   Price in INR       1836 non-null   object
dtypes: float64(1), object(9)
memory usage: 143.6+ KB
```

Tabel 2. Harga Smartphone

	Phone Name	Rating ?/5	Number of Ratings	RAM	ROM/Storage	Back/Rare Camera	Front Camera	Battery	Processor	Price in INR
0	POCO C50 (Royal Blue, 32 GB)	4.2	33,561	2 GB RAM	32 GB ROM	8MP Dual Camera	5MP Front Camera	5000 mAh	Mediatek Helio A22 Processor, Upto 2.0 GHz Pro...	₹5,649
1	POCO M4 5G (Cool Blue, 64 GB)	4.2	77,128	4 GB RAM	64 GB ROM	50MP + 2MP	8MP Front Camera	5000 mAh	Mediatek Dimensity 700 Processor	₹11,999
2	POCO C51 (Royal Blue, 64 GB)	4.3	15,175	4 GB RAM	64 GB ROM	8MP Dual Rear Camera	5MP Front Camera	5000 mAh	Helio G36 Processor	₹6,999
3	POCO C55 (Cool Blue, 64 GB)	4.2	22,621	4 GB RAM	64 GB ROM	50MP Dual Rear Camera	5MP Front Camera	5000 mAh	Mediatek Helio G85 Processor	₹7,749
4	POCO C51 (Power Black, 64 GB)	4.3	15,175	4 GB RAM	64 GB ROM	8MP Dual Rear Camera	5MP Front Camera	5000 mAh	Helio G36 Processor	₹6,999

Tabel 3. Data yang kosong pada setiap kolom

```
Data Kosong pada Setiap Kolom:
Phone Name             0
Rating ?/5             0
Number of Ratings      0
RAM                    0
ROM/Storage          174
Back/Rare Camera       9
Front Camera         401
Battery               10
Processor             55
Price in INR           0
Date of Scraping       0
dtype: int64
```

Langkah berikutnya yang dilakukan yaitu mengubah dataset menjadi dataframe. Selanjutnya proses pembersihan data atau *data cleaning* dengan menghapus variable ‘Phone Name’ karena variable tersebut memiliki dampak minimal atau tidak relevan terhadap target prediksi.

Tabel 4. Harga Smartphone setelah menghapus variable yang tidak relevan yaitu *Phone Name*

	Rating ?/5	Number of Ratings	RAM	ROM/Storage	Back/Rare Camera	Front Camera	Battery	Processor	Price in INR
0	4.2	33,561	2 GB RAM	32 GB ROM	8MP Dual Camera	5MP Front Camera	5000 mAh	Mediatek Helio A22 Processor, Upto 2.0 GHz Pro...	₹5,649
1	4.2	77,128	4 GB RAM	64 GB ROM	50MP + 2MP	8MP Front Camera	5000 mAh	Mediatek Dimensity 700 Processor	₹11,999
2	4.3	15,175	4 GB RAM	64 GB ROM	8MP Dual Rear Camera	5MP Front Camera	5000 mAh	Helio G36 Processor	₹6,999
3	4.2	22,621	4 GB RAM	64 GB ROM	50MP Dual Rear Camera	5MP Front Camera	5000 mAh	Mediatek Helio G85 Processor	₹7,749
4	4.3	15,175	4 GB RAM	64 GB ROM	8MP Dual Rear Camera	5MP Front Camera	5000 mAh	Helio G36 Processor	₹6,999

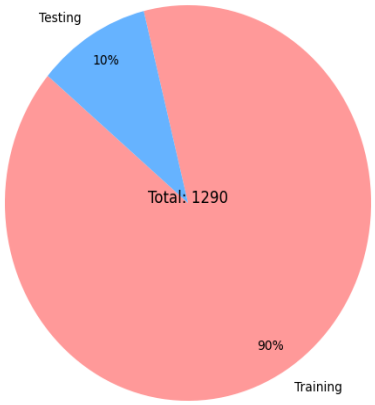
Pembersihan data pada setiap variabel dilakukan sesuai dengan kebutuhan. Pada variabel ‘RAM’, tahapan *data cleaning* dimulai dengan mengambil angka dari *string* di dalam kolom ‘RAM’ menggunakan *regular expression* dan mengubahnya menjadi tipe data *float*. Proses tersebut juga dilakukan untuk variabel ‘Rating ?/5’, ‘Number of Ratings’, ‘Front Camera’, ‘Battery’ , ‘Back/Rare Camera’ dengan menghapus tanda tambah (+) dan spasi serta variabel ‘Price in INR’ dengan menghapus tanda '₹' sebelum dilakukan konversi menjadi tipe data *float*.

Tahapan *data cleaning* pada variabel ‘ROM/Storage’ dengan tambahan konversi data ‘Expandable’ yang ada dalam kolom ‘ROM/Storage’ menjadi nilai ‘Nan’ sehingga ketika data tersebut diubah menjadi tipe *data float* maka hasilnya akan menjadi NaN (*Not a Number*), yang merupakan representasi dari nilai yang hilang atau tidak valid.

Variabel ‘Processor’ merupakan sebuah fitur kategorikal yang berisi nama prosesor smartphone sehingga dilakukan proses *label encoding* untuk diberi label numerik unik. Penggunaan *label encoding* dapat dengan mudah mengubah fitur 'Processor' yang berisi nama-nama prosesor menjadi nilai numerik. Hal ini memungkinkan model untuk memahami hubungan antara prosesor dengan variabel target (harga *smartphone*) saat melakukan proses *training*.

Tahap selanjutnya yaitu *data processing*. Dataset yang sudah diubah menjadi *dataframe* kemudian dilakukan tahap *processing* sehingga bisa diolah oleh model *machine learning*.

Sebelum melakukan normalisasi dan standarisasi data pada data processing, hasil *data cleaning* dibagi menjadi data latih (*training*) dan data uji (*testing*). Setelah melalui proses *data cleaning*, data yang tersisa yaitu 1290 yang dibagi menjadi 90% atau 1161 data latih dan 10% atau 129 data uji. Pembagian data menjadi data latih dan data uji penting dalam pembangunan model *machine learning* karena membantu dalam mengevaluasi seberapa baik model bekerja pada data yang belum pernah dilihat sebelumnya. Pemilihan pembagian data, seperti 90% untuk data latih dan 10% untuk data uji, bukan aturan yang baku dan dapat bervariasi tergantung pada ukuran dataset, kompleksitas model, dan tujuan analisis.



Gambar 2. Proporsi Data Training dan Data Testing

Setelah membagi data menjadi data latih dan data uji, selanjutnya tahap normalisasi dan standarisasi data numerik. Proses normalisasi dan standarisasi data dilakukan untuk memastikan bahwa semua fitur memiliki skala yang seragam, sehingga tidak ada satu fitur pun yang mendominasi dalam perhitungan model.

Tabel 5. Data variable setelah melalui proses *cleaning*, normalisasi dan standarisasi

Data Variabel Setelah Cleaning :

	Rating ?/5	Number of Ratings	RAM	ROM/Storage	Back/Rare Camera	Front Camera	Battery	Processor_LabelEncoded
0	0.875000	0.024998	0.001957	0.0625	0.04	0.083333	0.677419	0.574074
1	0.875000	0.057450	0.005871	0.1250	0.25	0.133333	0.677419	0.530864
2	0.895833	0.011303	0.005871	0.1250	0.04	0.083333	0.677419	0.324074
3	0.875000	0.016850	0.005871	0.1250	0.25	0.083333	0.677419	0.592593
4	0.895833	0.011303	0.005871	0.1250	0.04	0.083333	0.677419	0.324074

Mean Squared Error: 9322848.510354057
R-squared: 0.9546596207605785

Gambar 3. Hasil evaluasi model Random Forest Regression

Selanjutnya model yang telah dibuat akan dievaluasi dengan metrik evaluasi *Mean Squared Error* (MSE) dan *R-squared*. Menggunakan kedua metrik ini secara bersamaan memberikan gambaran yang lebih lengkap tentang kinerja model regresi. MSE memberikan informasi tentang akurasi prediksi secara keseluruhan, sedangkan R-squared memberikan informasi tentang seberapa baik model sesuai dengan data yang ada. Dengan demikian, menggunakan MSE dan R-squared membantu dalam mengevaluasi dan memahami kinerja model regresi secara holistik.

4. KESIMPULAN

Berdasarkan pengujian dan implementasi, Studi ini menunjukkan bahwa model Random Forest Regression efektif dalam memprediksi harga smartphone tahun 2023 berdasarkan fitur-fitur spesifikasi teknisnya. Hasil evaluasi menunjukkan tingkat akurasi yang tinggi, ditandai dengan MSE sebesar 9322848,510354057 dan R-squared sebesar 0.9546596207605785. Hal ini menunjukkan bahwa model ini dapat menjadi alat yang berguna bagi industri smartphone dalam merencanakan strategi harga dan mengantisipasi tren pasar.

Secara praktis, penelitian ini memberikan wawasan bagi produsen, pedagang, dan konsumen tentang faktor-faktor yang memengaruhi harga smartphone. Implikasi lainnya termasuk potensi untuk pengembangan model prediksi yang lebih kompleks dengan mempertimbangkan faktor-faktor eksternal seperti kondisi pasar global dan perkembangan teknologi.

Relevansi penelitian ini juga terletak pada kontribusinya terhadap literatur terkait. Meskipun banyak penelitian telah dilakukan dalam analisis harga smartphone, penelitian ini memberikan pendekatan yang lebih terfokus pada fitur-fitur spesifikasi teknis, yang dapat menjadi tambahan berharga bagi penelitian sebelumnya. Diharapkan penelitian ini dapat memberikan kontribusi yang berarti dalam pemahaman dan analisis harga smartphone untuk tahun 2023 dan menjadi landasan untuk penelitian lanjutan di masa depan.

Untuk pengembangan lebih lanjut, penelitian ini dapat diekspansi dengan mempertimbangkan faktor-faktor lain yang dapat mempengaruhi harga smartphone, seperti kondisi pasar global, tren teknologi, dan faktor ekonomi lainnya. Selain itu, penggunaan teknik pengolahan data yang lebih canggih dan model machine learning yang lebih kompleks juga dapat meningkatkan akurasi prediksi.

5. DAFTAR PUSTAKA

HAQUE-FAWZI, M.G., 2022. Strategi Pemasaran. Tangerang Selatan: Pascal Books.

CHANG HARTONO, P. AND DWIYOGA WIDIANTORO, A., 2024. Analisis Prediksi Harga Saham Unilever Menggunakan Regresi Linier Dengan RapidMiner. [online] Available at: https://journal-computing.org/index.php/journal-cisa/index [Accessed 18 Jan. 2025].

RAMADHANSYAH, D.S., 2022. Perbandingan Metode Seleksi Fitur Filter.

RELIGIA, Y., NUGROHO, A. AND HADIKRISTANTO, W., 2021. Klasifikasi Analisis Perbandingan Algoritma Optimasi Pada Random Forest Untuk Klasifikasi Data Bank Marketing. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 5(1), pp.187–192. doi:10.29207/resti.v5i1.2813.

MUSTAFA, W.F., HIDAYAT, S. AND FUDHOLI, D.H., 2024. Prediksi Retensi Pengguna Baru Shopee Menggunakan Machine Learning. Jurnal Media Informatika Budidarma, 8(1), p.612. doi:10.30865/mib.v8i1.7074.

RAHARJA, A.R., PRAMUDIANTO, A. and MUCHSAM, Y., n.d. Penerapan Algoritma Decision Tree Dalam Klasifikasi Data 'Framingham' Untuk Menunjukkan Risiko Seseorang Terkena Penyakit Jantung Dalam 10 Tahun Mendatang.

Anam, M.K. and Jakaria, D.A., 2023. Sistem Prediksi Harga Kripto Dengan Metode Regresi. Jurnal MDP, 10(2), pp.467–479. [online] Available at: http://jurnal.mdp.ac.id [Accessed 18 Jan. 2025].

NOOR, A., 2018. Perbandingan Algoritma Support Vector Machine Biasa Dan Support Vector Machine Berbasis Particle Swarm Optimization Untuk Prediksi Gempa Bumi.

AZMI, U., HADI, Z.N. and SORAYA, S., 2020. ARDL Method: Forecasting Data Curah Hujan Harian NTB. Jurnal Varian, 3(2), pp.73–82. doi:10.30812/varian.v3i2.627.

ROSITA, Y.D. and MOONLIGHT, L.S., 2024. Perbandingan Metode Prediksi Untuk Nilai Jual USD: Holt-Winters, Holt's, Dan Single Exponential Smoothing. Jurnal Teknologi Informasi Dan Multimedia, 5(4), pp.322–333. doi:10.35746/jtim.v5i4.473.