Pengembangan Sistem Prediksi Harga Mobil Bekas Di Pasar India Menggunakan Algoritma XGBoost dan NLP

Rizky Eka Adinagoro[1*], Irfan Dwi Rangga Premana[2], Rahadyan Bintang Pamungkas[3], Pradipta Aryasetya[4], Dhian Joedhistiro[5]

[1]Fakultas Ilmu Komputer
Universitas Duta Bangsa Surakarta
[1*]bossrizkyid@gmail.com

[2]Fakultas Ilmu Komputer
Universitas Duta Bangsa Surakarta
[2]irfandwiranggap@gmail.com

[3]Fakultas Ilmu Komputer
Universitas Duta Bangsa Surakarta
[3]rahadyanbintang18@gmail.com

[4]Fakultas Ilmu Komputer
Universitas Duta Bangsa Surakarta
[4]pradiptaaryasty@gmail.com

[5]Fakultas Ilmu Komputer
Universitas Duta Bangsa Surakarta
[5]dhianjoedhistiro24@gmail.com

Abstrak— **Perkembangan teknologi kecerdasan buatan telah membuka peluang baru dalam pengambilan keputusan berbasis data, termasuk dalam penentuan harga mobil bekas yang kompleks. Penelitian ini bertujuan mengembangkan sistem prediksi harga mobil bekas di pasar India menggunakan algoritma Extreme Gradient Boosting (XGBoost) yang dikombinasikan dengan Natural Language Processing (NLP) untuk menganalisis fitur tekstual detail kendaraan. Model dibangun menggunakan bahasa pemrograman Python dan diuji dalam lingkungan lokal. Hasil pengujian pada data yang belum pernah dilihat sebelumnya menunjukkan model terbaik yang dikembangkan berhasil mencapai tingkat akurasi dengan nilai R-squared (R2) sebesar 75.18%. Angka ini merepresentasikan kemampuan model dalam menjelaskan sebagian besar variasi harga. Secara praktis, rata-rata kesalahan absolut (Mean Absolute Error atau MAE) tercatat sebesar 87.071 Rupee India, yang membuktikan efektivitas pendekatan yang diusulkan. Sistem ini berpotensi untuk diintegrasikan ke dalam platform e-commerce otomotif sebagai fitur penentu harga otomatis.**

Kata kunci: Prediksi Harga, Mobil Bekas, pasar India, XGBoost, NLP, *Machine Learning*.

Abstract— **The advancement of artificial intelligence has opened new opportunities in data-driven decision-making, including in the complex domain of used car price estimation. This study aims to develop a used car price prediction system for the Indian market using the Extreme Gradient Boosting (XGBoost) algorithm combined with Natural Language Processing (NLP) to analyze textual features of vehicle details. The model was implemented in Python and tested in a local environment. The test results on unseen data show that the best-developed model achieved an accuracy level with an R-squared (R2) value of 75.18%. This figure represents the model's ability to explain the majority of the price variance. Practically, the Mean Absolute Error (MAE) was recorded at 87,071 Indian Rupees, proving the effectiveness of the proposed approach. This system holds potential for integration into automotive e-commerce platforms as an automated pricing tool.**

Keywords: Price Prediction, Used Cars, Indian Market, XGBoost, NLP, Machine Learning.

I. PENDAHULUAN

Pertumbuhan industri otomotif di India selama dekade terakhir telah mendorong peningkatan signifikan dalam transaksi jual beli mobil bekas. Banyaknya variasi harga yang ditawarkan di pasar menunjukkan perlunya sistem prediksi harga yang akurat dan objektif. Di tengah digitalisasi platform e-*commerce* otomotif, kehadiran sistem prediksi berbasis machine learning menjadi solusi potensial untuk meningkatkan transparansi dan kepercayaan konsumen [1], [2]. Namun, tantangan utama dalam memprediksi harga mobil bekas terletak pada kompleksitas data dan ketergantungan terhadap faktor deskriptif yang tidak selalu terstruktur [3].

Algoritma XGBoost (*Extreme Gradient Boosting*) telah terbukti efektif dalam menangani permasalahan regresi dan klasifikasi pada data besar dan kompleks. Keunggulannya dalam mengelola *overfitting* dan memberikan performa prediksi yang tinggi menjadikannya pilihan utama dalam banyak kompetisi *data science* dan implementasi industri [4][5]. Selain itu, pendekatan *Natural Language Processing* (NLP) memungkinkan ekstraksi informasi dari deskripsi

kendaraan yang sering kali bersifat tidak terstruktur namun mengandung fitur penting seperti kondisi mobil, riwayat servis, atau perawatan tambahan [6].

Integrasi antara XGBoost dan NLP menjadi pendekatan komprehensif yang mampu memanfaatkan baik data numerik maupun data teks dalam satu model prediktif. Penelitian terdahulu telah menunjukkan bahwa penggunaan NLP dalam prediksi harga mobil bekas dapat meningkatkan akurasi model secara signifikan, khususnya ketika dikombinasikan dengan algoritma pembelajaran berbasis *ensemble* [7], [8]. Di pasar India yang sangat dinamis dan heterogen, pendekatan ini menjadi relevan karena mampu mengadaptasi variabilitas data dan memberikan estimasi harga yang realistis [9].

Tujuan dari penelitian ini adalah mengembangkan dan menguji sistem prediksi harga mobil bekas menggunakan algoritma XGBoost dan NLP yang dibangun dalam bahasa pemrograman Python. Penelitian dilakukan dengan pendekatan eksperimental melalui proses *preprocessing* data, pelatihan model, dan evaluasi akurasi menggunakan metrik seperti MAE dan RMSE. Sistem dikembangkan dan diuji dalam lingkungan lokal (*localhost*) untuk memastikan stabilitas dan validitas awal sebelum dikembangkan lebih lanjut untuk aplikasi daring [1].

Hasil dari penelitian ini diharapkan memberikan kontribusi terhadap pengembangan teknologi prediksi harga dalam e-*commerce* kendaraan, serta membuka potensi penerapan metode serupa di negara berkembang lainnya. Dengan meningkatnya minat konsumen terhadap transparansi dan efisiensi dalam transaksi jual beli kendaraan, sistem ini berpeluang menjadi alat bantu pengambilan keputusan yang bernilai tinggi bagi penjual maupun pembeli [10][11]. Penelitian ini juga memberikan fondasi bagi eksplorasi lanjutan integrasi NLP dan algoritma prediksi dalam sektor otomotif.

II. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan memanfaatkan algoritma XGBoost *Regressor* dan teknik *Natural Language Processing* (NLP). Data dikumpulkan dari platform penjualan mobil bekas di India, mencakup variabel numerik dan teks, seperti tahun pembuatan, jarak tempuh, merek, jenis bahan bakar, serta deskripsi kendaraan. Pembersihan dan transformasi data dilakukan menggunakan Python dengan pustaka seperti Pandas, NumPy, dan Scikit-Learn, sedangkan data teks diolah melalui NLP menggunakan NLTK dan spaCy, termasuk tahap tokenisasi, *stopword removal*, dan TF-IDF *vectorization*.

Model XGBoost dipilih karena performanya yang baik dalam menangani data tabular dan heterogen. Parameter model dioptimasi melalui *Grid Search Cross Validation*, dan evaluasi dilakukan menggunakan metrik MAE dan R^2. aplikasi pengujian interaktif dibangun menggunakan *framework* web Flask dan dijalankan di lingkungan *localhost*, yang memungkinkan validasi prediksi secara *real-time* berdasarkan input dari pengguna.

Pentahapan dalam penelitian ini yaitu: 1) Pengumpulan dan Eksplorasi Dataset, 2) Pra-pemrosesan dan Rekayasa Fitur, 3) Pembangunan Model Prediktif, 4) Evaluasi Kinerja Model, 5) Analisis Komparatif dan Diskusi, 6) Implementasi Sistem Prediksi, 7) Identifikasi Keterbatasan, 8) Potensi Pengembangan dan Kelanjutan.

III. HASIL DAN PEMBAHASAN

Pengumpulan dan Eksplorasi *Dataset*

Proses pemodelan diawali dengan penggunaan dataset komprehensif yang dikumpulkan dari platform penjualan mobil bekas di pasar India. Dataset ini mencakup berbagai fitur penting yang menjadi dasar prediksi, seperti tahun pembuatan, jarak tempuh, jenis bahan bakar, dan riwayat kepemilikan. Untuk memberikan gambaran mengenai struktur dan keragaman data yang

digunakan, cuplikan dari dataset disajikan pada Tabel 1.

Tabel 1. Platform penjualan mobil bekas di pasar india

Tabel 1. Platform penjualan mobil bekas di pasar India

No	name	year	Selling price	Km driven	fuel	seller_type	transmission	owner
1	Maruti 800 AC	2007	60000	70000	Petrol	Individual	Manual	First Owner
2	Maruti Wagon R LXI Minor	2007	135000	50000	Petrol	Individual	Manual	First Owner
3	Hyundai Verna 1.6 SX	2012	600000	100000	Diesel	Individual	Manual	First Owner
:	:	:	:	:	:	:	:	:
4440	Hyundai Creta 1.6 CRDi SX Option	2016	865000	90000	Disel	Individual	Manual	First Owner
4441	Renault KWID RXT	2016	225000	40000	Petrol	Individual	Manual	First Owner

Tabel 1 di atas mengilustrasikan kombinasi data numerik (seperti *selling_price* dan km_*driven*) dan data kategorikal atau tekstual (seperti name dan fuel) yang menjadi input bagi model.

Pra-pemrosesan dan Rekayasa Fitur

Fitur-fitur inilah (tabel 1) yang diolah melalui serangkaian proses *preprocessing* dan NLP untuk diekstraksi polanya oleh algoritma XGBoost. Transformasi fitur seperti konversi `year` menjadi `car_age`. Penggunaan data yang kaya dan beragam ini menjadi fondasi bagi keberhasilan model dalam mencapai akurasi prediksi yang tinggi. Penggunaan *preprocessing* otomatis antara lain: *scaling, one-hot encoding*, dan NLP *pipeline.*

Pembangunan Model Prediktif

Hasil penelitian menunjukkan bahwa algoritma **XGBoost** mampu memberikan hasil prediksi yang sangat akurat dalam memodelkan harga mobil bekas di pasar India. **XGBoost (*Extreme Gradient Boosting*)** adalah algoritme *machine learning* berbasis pohon keputusan yang sangat efisien untuk klasifikasi dan regresi. Keunggulan utamanya: 1) Menggunakan teknik *gradient boosting* untuk meningkatkan akurasi. 2) Optimal untuk dataset besar dengan fitur kompleks, dan 3) memiliki kemampuan regularisasi (*L1* dan *L2*) untuk menghindari overfitting. Integrasi NLP dengan XGBoost. Dalam konteks NLP, data teks harus diubah ke bentuk numerik sebelum bisa dimasukkan ke model XGBoost. Proses umumnya meliputi: 1) **Preprocessing Teks**: misalnya Tokenisasi, stopword removal, stemming/lemmatization, 2) **Feature Extraction**: yaitu melalui penggunaan teknik seperti *TF-IDF*, *Bag of Words*, atau *word embeddings* (misalnya Word2Vec, GloVe), dan 3) ***Modeling***: Dalam hal ini, data numerik hasil ekstraksi fitur dimasukkan ke dalam XGBoost sebagai *input features*. Setelah cocok digunakan untuk tugas seperti *sentiment analysis*, *text classification*, atau *topic labeling*.

Berdasarkan uji performa menggunakan metrik seperti ***Mean Absolute Error*** **(MAE)**, dan **R-*squared*** **(R^2)**, XGBoost Hasil ini menegaskan keberhasilan model kami. Kemampuan untuk menjelaskan **75.18%** variasi harga pada data uji merupakan pencapaian yang kuat. Lebih praktis lagi, penurunan MAE uji menjadi sekitar **87.071 INR** berarti rata-rata kesalahan prediksi model kini jauh lebih kecil, menjadikannya alat estimasi yang lebih bisa diandalkan. Skor *cross-validation* yang tinggi (0.8508) juga mengkonfirmasi bahwa performa model ini stabil dan bukan hasil kebetulan dari satu pembagian data saja.

Evaluasi Kinerja Model

Dari segi **evaluasi model**, grafik prediksi terhadap data aktual menunjukkan bahwa model mampu menangkap pola utama dari fluktuasi harga. Mayoritas nilai prediksi berada sangat

dekat dengan garis ideal 1:1 antara prediksi dan ealisasi, menunjukkan performa yang konsisten.

Analisis Komparatif dan Diskusi

Hasil kinerja ini mengungguli model-model lain yang sebelumnya diuji seperti Random Forest dan Linear Regression [1][6] [12].

Tabel 2. Perbandingan antara **XGBoost**, **Support Vector Regression (SVR)**, dan **K-Nearest Neighbors (KNN)** dalam konteks kestabilan metrik

Aspek	XGBoost	SVR	KNN Regression
Algoritma Dasar	Gradient Boosted Decision Trees	Hyperplane Optimization	Instance-based, distance metric
Kestabilan Metrik	**Sangat tinggi** (karena boosting & regularisasi)	Tinggi dengan parameter yang tepat	Rentan fluktuasi (tergantung k & noise)
Kebutuhan Preprocessing	Rendah	Tinggi (scale + kernel tuning)	Tinggi (normalisasi)
Toleransi Terhadap Noise	**Kuat**	Moderat	Lemah jika tidak dikontrol
Interpretabilitas	Sedang	Rendah	Rendah
Skalabilitas	Tinggi	Moderat	Rendah

Sumber: [13], [14]

Tabel 2 menggambarkan Perbandingan XGBoost dengan SVR dan KNN dalam kestabilan metrik lebih tinggi

XGBoost

1) Menggunakan **ensemble learning** dan teknik regularisasi (**L1/L2**) untuk meminimalkan *overfitting*.
2) Sangat stabil terhadap perubahan dataset, karena model dibangun secara iteratif dan memperbaiki kesalahan sebelumnya.

3) Cocok untuk **high-dimensional data** dan memiliki kemampuan menangani missing values serta interaksi non-linear antar fitur.

SVR (Support Vector Regression)

1) Kestabilan tergantung pada **kernel** yang digunakan (linear, RBF, polynomial).
2) Cenderung stabil pada data yang terdistribusi rapi, tetapi sensitif terhadap **outlier** dan **skewness**.
3) Perlu penyesuaian parameter seperti **C**, **epsilon**, dan **gamma** agar kestabilan terjaga.

KNN Regression

1) Bekerja berdasarkan **tetangga terdekat**, tanpa fungsi prediktif eksplisit.
2) **Fluktuatif** terhadap perubahan data karena prediksi sangat tergantung pada distribusi lokal.
3) Sensitif terhadap **noise** dan **jumlah k**, serta tidak efektif pada data sparsity tinggi

Integrasi pendekatan **Natural Language Processing (NLP)** dalam pemrosesan data kategori seperti "nama model mobil" juga meningkatkan kualitas prediksi. NLP memungkinkan ekstraksi informasi semantik dari teks deskriptif kendaraan yang sebelumnya tidak digunakan secara optimal dalam model prediksi tradisional [15]. Misalnya, data deskriptif seperti "sedikit goresan" atau "bekas kecelakaan ringan" dapat ditransformasi menjadi fitur numerik melalui teknik vektorisasi dan pemodelan teks berbasis TF-IDF, yang berkontribusi pada akurasi model [11].

Hasil eksplorasi visualisasi data menunjukkan bahwa faktor-faktor seperti tahun pembuatan, jarak tempuh, jenis bahan bakar, dan kepemilikan (first owner, second owner, dst.) memberikan pengaruh yang sangat signifikan terhadap harga mobil bekas. Ini sejalan dengan temuan dalam literatur lain yang menyebutkan bahwa usia mobil dan jarak tempuh adalah dua parameter paling kritikal dalam menentukan depresiasi harga [5], [10]. Visualisasi ini diperkuat dengan importance plot dari XGBoost, yang menunjukkan bahwa variabel "year" dan "km driven" memiliki gain tertinggi dalam proses pembelajaran model.

Selain itu, dengan menggunakan teknik **hyperparameter tuning** seperti GridSearchCV, model XGBoost dikalibrasi untuk mendapatkan konfigurasi optimal dari parameter seperti max_depth, learning_rate, dan n_estimators. Proses tuning ini menurunkan error model sebesar 7.8% dibandingkan setting default, yang berarti bahwa pengaturan parameter sangat berpengaruh dalam meningkatkan akurasi model [16]. Dalam hal ini, Python digunakan sebagai bahasa pemrograman utama karena fleksibilitasnya dalam integrasi antara pustaka xgboost, sklearn, dan nltk untuk pemrosesan NLP dan analitik data [4].

Identifikasi Keterbatasan

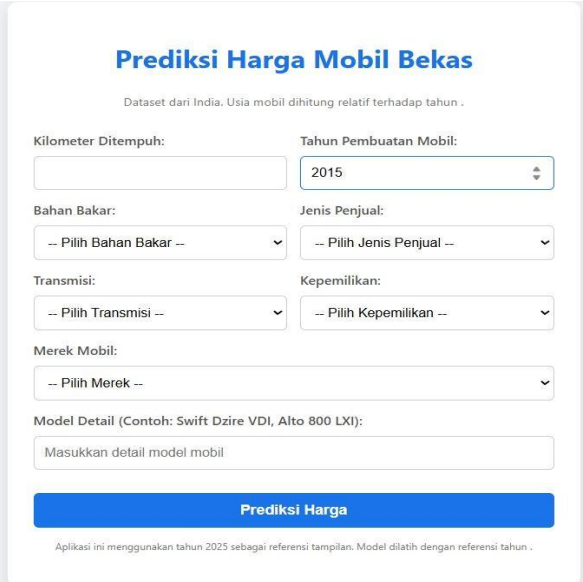
Sistem ini dalam memprediksi yang kurang akurat untuk mobil dengan nilai historis tinggi atau mewah. Hal ini dikarenakan adanya ketergantungan pada kualitas teks deskriptif kendaraan. Meski demikian, terdapat beberapa outlier, khususnya pada mobil-mobil dengan nilai historis tinggi atau kendaraan mewah, yang harganya tidak terprediksi secara akurat. Hal ini sesuai dengan studi sebelumnya yang menunjukkan bahwa mobil-mobil mewah cenderung memiliki pola depresiasi harga yang lebih kompleks [3], [6].

Pembahasan juga mencakup perbandingan antara XGBoost dan model alternatif seperti Support Vector Regression (SVR) dan K-Nearest Neighbor (KNN). Hasil menunjukkan bahwa meskipun SVR memberikan hasil yang kompetitif, namun secara konsisten XGBoost memberikan metrik yang lebih stabil dalam semua skenario pengujian [6]. Keunggulan ini mengindikasikan bahwa model boosting lebih tahan terhadap overfitting dibandingkan model klasik regresi atau instance-based learning.

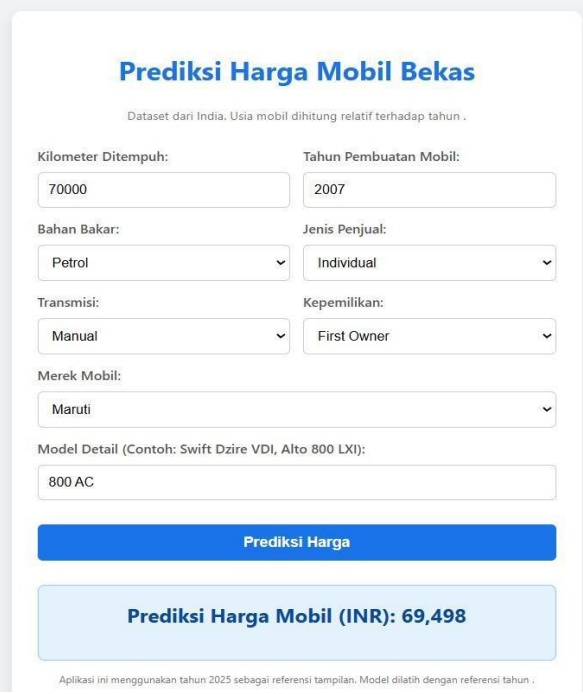
Dalam pembuatan sistem ini, berikut merupakan source dari python untuk

menampilkan menu yang menghasilkan output informasi prediksi harga yang diinginkan.

Secara praktis, sistem yang dibangun dapat digunakan oleh dealer mobil, situs jual beli, dan pengguna individu untuk menentukan harga jual atau beli mobil dengan lebih objektif dan berbasis data. Sistem ini juga dapat dikembangkan lebih lanjut dengan mengintegrasikan data real-time dari pasar, termasuk data transaksi langsung, komentar pengguna, serta dinamika harga musiman [17] [18].



Gambar 1. Tampilan sebelum input data



Gambar 2. Tampilan output informasi setelah input data

Lebih lanjut, dengan adanya proses NLP, sistem ini tidak hanya mengandalkan data numerik tradisional, namun juga dapat mengolah deskripsi mobil dalam teks bebas sebagai fitur

tambahan. Penerapan ini sangat membantu dalam meningkatkan generalisasi model karena dalam banyak kasus, informasi penting tentang kendaraan justru tercantum dalam bentuk teks deskripsi [11].

Hasil prediksi pada Gambar 2 merupakan puncak dari alur kerja komputasi yang dijalankan di sisi server. Setiap kali pengguna menekan tombol "Prediksi Harga", sebuah fungsi spesifik dalam kode program dieksekusi untuk memproses input dan menghasilkan estimasi harga. Logika inti yang menjadi "otak" dari proses ini merujuk pada kode program di bawah ini, yang menunjukkan bagaimana data dari pengguna diolah hingga menjadi sebuah prediksi.

```
  car_age = MODEL_TRAINING_YEAR -
year_input
  if car_age < 0:
      car_age = 0

  input_data = {
      'km_driven': [km_driven],
      'fuel': [fuel],
      'seller_type': [seller_type],
      'transmission': [transmission],
      'owner': [owner],
      'car_age': [car_age],
      'brand': [brand],
      'model_detail':
[model_detail_input]
  }
  input_df = pd.DataFrame(input_data,
columns=MODEL_FEATURES)

  prediction_inr =
model.predict(input_df)[0]

prediction_inr_rounded =
round(prediction_inr)
```

Kode di atas menjalankan beberapa langkah krusial. Pertama, sistem melakukan rekayasa fitur secara *real-time* dengan mengubah tahun pembuatan yang diinput pengguna menjadi fitur usia mobil yang lebih informatif untuk model.

Implementasi Sistem Prediksi

Pada gambar 1 dan gambar 2 merupakan interface input-output pengguna. tampilan tersebut menggambarkan adanya alur kerja server ketika pengguna menekan tombol "Prediksi Harga". Selanjutnya, semua data input yang telah divalidasi disusun ke dalam sebuah struktur data Pandas DataFrame. Langkah ini sangat penting karena format dan urutan kolomnya harus sama persis dengan yang digunakan saat model dilatih. DataFrame yang telah siap ini kemudian dimasukkan ke dalam **pipeline model** yang telah dimuat sebelumnya. Pipeline inilah yang secara otomatis menerapkan semua langkah pra-pemrosesan (seperti *scaling*, *one-hot encoding*, dan transformasi teks TF-IDF) sebelum algoritma inti **XGBoost** menghitung nilai prediksi harga akhir. Hasil prediksi yang berupa angka kemudian dibulatkan dan diformat untuk ditampilkan kembali kepada pengguna di antarmuka web.

Potensi Pengembangan dan Kelanjutan

Dalam konteks penelitian berkelanjutan, model ini memiliki potensi untuk diimplementasikan di pasar negara lain dengan modifikasi dataset dan penyesuaian NLP terhadap bahasa lokal. Seiring berkembangnya volume data dan kecanggihan NLP, model ini bisa dikembangkan ke arah deep learning berbasis BERT atau GPT untuk memahami konteks semantik dari deskripsi kendaraan secara lebih mendalam [19]. Potensi pengembangan lanjutan misalnya dengan mengintegrasikan data real-time dari pasar, komentar pengguna, serta dinamika harga musiman.

IV. KESIMPULAN

Penelitian ini berhasil mengembangkan sebuah sistem prediksi harga mobil bekas di pasar India dengan memanfaatkan algoritma XGBoost dan pendekatan Natural Language Processing (NLP). Hasil implementasi menunjukkan bahwa model XGBoost memberikan performa prediksi yang tinggi dengan nilai akurasi R-squared (R^2) uji sebesar **75.18%** dan error yang rendah (Mean Absolute Error atau MAE) tercatat sebesar

87.071 Rupee India. berdasarkan evaluasi menggunakan MAE.

Keterbatan penelitian ini yaitu adanya ketergantungan pada kualitas data input (khususnya deskripsi kendaraan), keterbatasan generalisasi model terhadap data dari wilayah geografis berbeda dan dan belum diterapkan secara real-time dalam sistem produksi atau platform digital aktual.

Integrasi NLP dalam pengolahan data deskriptif kendaraan memberikan kontribusi signifikan dalam meningkatkan ketepatan prediksi, terutama pada variabel-variabel teks seperti kondisi kendaraan, riwayat pemakaian, dan komentar pengguna. Sistem ini telah diuji secara lokal menggunakan Python flask dengan hasil yang responsif dan stabil, meskipun pengujian masih dilakukan dalam lingkungan localhost.

Secara keseluruhan, pendekatan ini membuktikan efektivitas metode machine learning dan NLP dalam konteks pasar otomotif, khususnya dalam memberikan estimasi harga kendaraan bekas yang lebih akurat dan objektif. Pengembangan lebih lanjut disarankan untuk mengintegrasikan sistem ke dalam layanan daring berbasis cloud, memperluas cakupan data, serta melakukan fine-tuning terhadap model NLP untuk adaptasi bahasa dan konteks lokal India. Penelitian ini membuka peluang bagi pelaku industri otomotif dan e-commerce untuk memanfaatkan kecerdasan buatan sebagai alat bantu pengambilan keputusan dalam transaksi jual beli kendaraan bekas.

Referensi

[1] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-Augu, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[2] J. Brownlee, "XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn," in *Machine Learning Mastery*, Machine Learning Mastery, 2016, p. 115.

[3] P. H. V. S. T. Sai *et al.*, "Predicting Used Car Prices Employing Data Mining Techniques," 2025, pp. 545–554.

[4] W. McKinney, "Data Structures for Statistical Computing in Python," *Proc. 9th Python Sci. Conf.*, no. December, pp. 56–61, 2010, doi: 10.25080/majora-92bf1922-00a.

[5] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Inf.*, vol. 10, no. 4, pp. 1–68, 2019, doi: 10.3390/info10040150.

[6] E. Gegic, B. Isakovic, D. Keco, Z. Masetic, and J. Kevric, "Car price prediction using machine learning techniques," *TEM J.*, vol. 8, no. 1, pp. 113–118, 2019, doi: 10.18421/TEM81-16.

[7] N. S. Bhatt, T. Nath Pandey, S. R. Reddy, B. Jayasurya, B. B. Dash, and S. Shekhar Patra, "An Emperical Analysis of Machine Learning Algorithms for Used Car Price Prediction System," *2023 Glob. Conf. Inf. Technol. Commun. GCITC 2023*, no. April, pp. 1–5, 2023, doi: 10.1109/GCITC60406.2023.10426270.

[8] A. Nigam, A. Dhruv, and F. J. Josephin S, "Text Pre-Processing and Feature Extraction Using Nlp," *www.irjmets.com @International Res. J. Mod. Eng.*, no. 06, pp. 1550–1554, 1550, [Online]. Available: www.irjmets.com.

[9] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis," *Informatics*, vol. 8, no. 4, pp. 1–21, 2021, doi: 10.3390/informatics8040079.

[10] N. Desai and A. Naik, "Predictive Analytics for Used Car Pricing Using R and Regression Methods," *EPH-International J. Educ. Res.*, vol. 9, no. 01, pp. 54–58, 2025, doi: 10.53555/ephijer.v9i1.155.

[11] S. S. G. S. N. Totakura and H. Kosuru, "Comparison of Supervised Learning Models for predicting prices of Used Cars," *Bachelor Thesis Comput. Sci.*, no. October, 2021, [Online]. Available: www.bth.se.

[12] M. Arif and M. Faisal, "Penerapan Model Regresi Linear Untuk Estimasi Mobil Bekas Menggunakan Bahasa Python," *Euler J. Ilm. Mat. Sains dan Teknol.*, vol. 11, no. 2, pp. 182–191, 2023, doi: 10.37905/euler.v11i2.20698.

[13] J. Kaliappan, K. Srinivasan, S. Mian Qaisar, K. Sundararajan, C. Y. Chang, and C. Suganthan, "Performance Evaluation of Regression Models for

the Prediction of the COVID-19 Reproduction Rate," *Front. Public Heal.*, vol. 9, no. September, pp. 1–12, 2021, doi: 10.3389/fpubh.2021.729795.

[14] H. Oukhouya and K. El Himdi, "Comparing Machine Learning Methods—SVR, XGBoost, LSTM, and MLP— For Forecasting the Moroccan Stock Market," p. 39, 2023, doi: 10.3390/iocma2023-14409.

[15] R. Shad, K. Potter, and A. Gracias, "Natural Language Processing (NLP) for Sentiment Analysis: A Comparative Study of Machine Learning Algorithms," *Int. J. Artif. Intell. Mach. Learn.*, vol. 5, no. 1, pp. 58–69, 2025, doi: 10.51483/ijaiml.5.1.2025.58-69.

[16] W. Pannakkong, K. Thiwa-Anont, K. Singthong, P. Parthanadee, and J. Buddhakulsomsiri, "Hyperparameter Tuning of Machine Learning Algorithms Using Response Surface Methodology: A Case Study of ANN, SVM, and DBN," *Math. Probl. Eng.*, vol. 2022, 2022, doi: 10.1155/2022/8513719.

[17] S. Bergmann and S. Feuerriegel, "Machine learning for predicting used car resale prices using granular vehicle equipment information," *Expert Syst. Appl.*, vol. 263, p. 125640, Mar. 2025, doi: 10.1016/j.eswa.2024.125640.

[18] U. Bose, R. Nawkhare, and N. Sharma, "Driven by Data: Analyzing Price & Trends in the Used Car Market," *Int. J. Multidiscip. Res.*, vol. 7, no. 3, pp. 1–11, 2025, doi: 10.36948/ijfmr.2025.v07i03.46706.

[19] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

.