Syntax Literate: Jurnal Ilmiah Indonesia p–ISSN: 2541-0849
e-ISSN: 2548-1398
Vol. 6, No. 5, Mei 2021

PREDIKSI HARGA MOBIL BEKAS DENGAN *MACHINE LEARNING*

Bambang Kriswantara, Kurniawati dan Hilman F. Pardede
Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Nusa Mandiri Jakarta dan Pusat Penelitian Informatika LIPI Bandung, Jawa Barat, Indonesia
Email: 14002426@nusamandiri.ac.id, 4002430@nusamandiri.ac.id dan hilman@nusamandiri.ac.id

Abstract

The price of a used car is built by several factors related to the car itself, such as the type of car, model, edition, year of production, transmission, fuel, engine capacity, and mileage. The price is also fluctuating and there is high competition between used car sellers, a tool is needed to predict car prices accurately and quickly. The purpose of this research is to help used car show rooms in predicting prices quickly with historical data using the Deep Neural Network (DNN) method with three hidden layers. In this research, two main approaches of research were conducted, namely qualitative approach and quantitative approach. The results of our study resulted in MAE = 501232, R2 = 0.88 which was better than previous researchers with the Random Forest method resulting in MAE = 521947, R2 = 0.82. So that the DNN method will improve the accuracy of the better predictions, although it does not increase significantly according to the research results.

Keywords: *machine learning; price prediction used cars; decision tree; random forest; DNN*

Abstrak

Harga mobil bekas dipengaruhi oleh beberapa faktor yang berkaitan dengan mobil itu sendiri, seperti jenis mobil, model, edisi, tahun produksi, transmisi, bahan bakar, kapasitas mesin, dan jarak tempuh. Harganya juga fluktuatif dan persaingan yang tinggi antar penjual mobil bekas, dibutuhkan alat untuk memprediksi harga mobil bekas secara akurat dan cepat. Tujuan dari penelitian ini untuk membantu *show room* mobil bekas dalam memprediski harga secara cepat dengan data *history* menggunakan metode *Deep Neural Network* (DNN) dengan tiga lapisan tersembunyi. Dalam Penelitian ini dilakukan dua pendekatan utama penelitian, yaitu pendekatan kualitatif dan pendekatan kuantitatif. Hasil penelitian kami menghasilkan MAE=501232, R2=0.88 yang lebih baik dari penelitian sebelumnya dengan metode *Random Forest* menghasilkan MAE=521947, R2=0.82. Sehingga Metode DNN akan meningkatkan akurasi prediski yang lebih baik, meskipun tidak naik secara signifikan mengacu pada hasil penelitian.

Kata Kunci: machine learning; prediksi harga mobil bekas; decision tree; random forest; DNN

How to cite: Kriswantara, Bambang., Kurniawati dan Hilman F. Pardede (2021) Prediksi Harga Mobil Bekas dengan *Machine Learning*. *Syntax Literate.* *6*(5). http://dx.doi.org/10.36418/syntax-literate.v6i5.2716
E-ISSN: 2548-1398
Published by: Ridwan Institute

Pendahuluan

Seiring dengan tingginya aktivitas dan bisnis, mobil sudah menjadi kebutuhan pokok. Disisi lain, harga mobil baru semakin tinggi dengan berbagai *feature* yang disematkan pada produk baru. Untuk memenuhi kesenjangan tersebut, masyarakat mencari alternatif untuk membeli mobil bekas yang kondisi masih baik dan layak digunakan. Tingginya minat masyarakat terhadap mobil bekas membuat bisnis ini semakin meningkat, hal ini ditandai dengan banyaknya *showroom* mobil bekas. Tak luput diantara *showroom* sangat kompetitif bersaing agar tetap eksis dalam bisnis mobil bekas. Salah satu masalah yang dihadapi semua s*howroom* adalah menentukan harga secara cepat dan akurat sehingga *showroom* bisa menjual mobil dagangannya dan segera mendapatkan *revenew*. Kondisi saat ini prediksi harga mobil pun masih popular (Pandey, Rastogi, & Singh, 2020). Berbagai *showroom* saling bersaing harga untuk mendapatkan pelanggan.

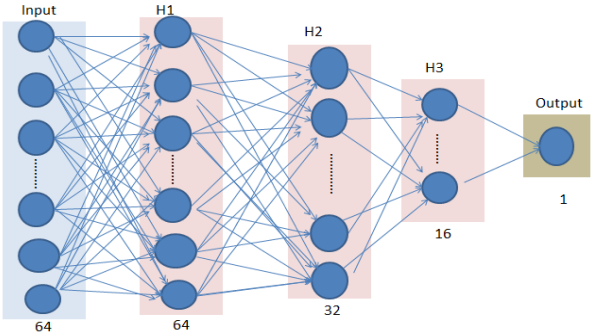
Machine Learning (ML) merupakan salah satu cabang dari *Artificial Intellegence* (AI) yang bisa memberikan keputusan berdasarkan data (Kaur & Kumari, 2020). Dengan sejumlah data latih, model ML dioptimasi pada data tersebut untuk menghasilkan model prediksi yang baik (Siji George C G, 2020), serta dapat digunakan pada data di masa yang akan datang.

Tujuan penelitian untuk mendapatkan akurasi terbaik dari model *machine learning* untuk prediksi harga mobil. Merujuk pada penelitian terdahulu (Vanshika, Singh, Sanika, 2020) dengan *Random forest* menghasilkan MAE=521947, R2=0.82 dengan waktu yang tidak disebutkan, dengan dataset yang sama pada penelitiaan ini menggunakan model DNN menghasilkan MAE=501232, R2=0.88 dengan epoch =111 dan waktu=0.51 detik. Sehingga ada novelty dari sisi tingkat error MAE turun sebesar 20715 dan R2 naik sebesar 0.6.

Penelitian ini sangat penting diera kompetisi persaingan harga pasar mobil bekas. Dengan kecepatan hasil prediski mobil secara marketing bisa segera dilihat jenis mobil yang paling laku dijual, tahun, transmisi dan jenis bahan bakar yang sangat diminati pembeli. Oleh sebab itu untuk kelangsungan eksistensi bisnis usaha penjualan mobil bekas harus segera dengan cepat bisa memprediski harga jual guna memperbaiki strategi marketing dan juga kondisi finansial perusahaan.

Deep Neural Network (DNN) adalah Jaringan syaraf tiruan adalah paradigma pemrosesan suatu informasi yang terispirasi oleh sistem sel syaraf biologi, sama seperti otak yang memproses suatu informasi (Murtadho, 2020). Neural network terdiri dari dua atau lebih lapisan, meskipun sebagian besar jaringanterdiri dari tiga lapisan: lapisan input, lapisan tersembunyi, dan lapisan output. Neural network juga terdiri dari neuron yang terdapat pada setiap layer dengan jumlah berbeda. Pada neural network terdiri dari banyak neuron di dalamnya. Neuron-neuron ini akan dikelompokkan ke dalam beberapa layer. Neuron yang terdapat pada tiap layer dihubungkan dengan neuron pada layer lainnya. Hal ini tentunya tidak berlaku pada layer input dan output, tapi hanya layer yang berada di antaranya. Informasi yang diterima di layer input dilanjutkan ke layer-layer dalam ANN secara satu persatu hingga mencapai layer terakhir/layer output. Layer

yang terletak di antara input dan output disebut sebagai hidden layer. Namun, tidak semua ANN memiliki hidden layer, ada juga yang hanya terdapat layer input dan output saja (Yusran, 2016). Arsitektur DNN dapat dilihat di gambar 1.



Gambar 1
DNN

Metode Penelitian

Dalam Penelitian ini dilakukan dua pendekatan utama penelitian, yaitu pendekatan kualitatif dan pendekatan kuantitatif. Pendekatan kualitatif digunakan untuk menganalisa kajian literatur yang berkenaan dengan variable-variabel yang memiliki pengaruh terhadap harga mobil bekas. Sedangkan pendekatatan kuantitif merupakan metode penelitian yang digunakan untuk meneliti populasi atau sampel tertentu terhadap dataset yang digunakan untuk prediksi harga mobil.

Metode pengumpulan data untuk mendapatkan sumber data yang digunakan adalah metode pengumpulan data sekunder yang diperoleh dari sumber kaggle.com dengan jumlah data 6500 baris dan 9 kolom. Tabel 1 adalah tampilan beberapa sample dari dataset:

Tabel 1
Sample dataset yang digunakan

Make	Model	Year	Transmission	Body type	Fuell type	Engine capacity	Mileage	Price
Volkswagen	Beetle	2015	Automatic	Concertible	Petrol	1400	16000	7500000
Suzuki	Alto	2016	Manual	Hatchback	Petrol	800	34500	1795000
Toyota	Corolla	1986	Manual	Station wagon	Petrol	1300	55000	425000
Nissan	March	2001	Automatic	Hatchback	Petrol	1000	126000	1625000
Suzuki	Liana	2003	Automatic	Saloon	Petrol	1500	140000	17800000

Sumber data set dari kaggle.com, data set tersebut merupakan sekumpulan jenis mobil bekas yang diproduksi di India dengan jumlah data 6500 *row* dan 9 *feature. Range Price* maksimal USD 62500000 dan minimal USD 27000, sedangkan *range* tahun produksi maksimal tahun 2020 dan minimal tahun 1936.

Pada data yang digunakan terdapat data *missing.* Pada penelitian ini, data tersebut dihilangkan karena jumlahnya tidak terlalu besar. Dari data awal sebanyak 6500 baris dan 9 kolom setelah dihapus data yang bernilai NAN menjadi 6493 baris dan 9 kolom. Tabel 2 adalah daftar jumlah data missing pada data ini:

Tabel 2
Feature missing

Make	1
Model	1
Year	1
Transmission	1
Body typeFuell type	7
Engine capacity	1
Mileage	1
Price	1
Dtype : int64	

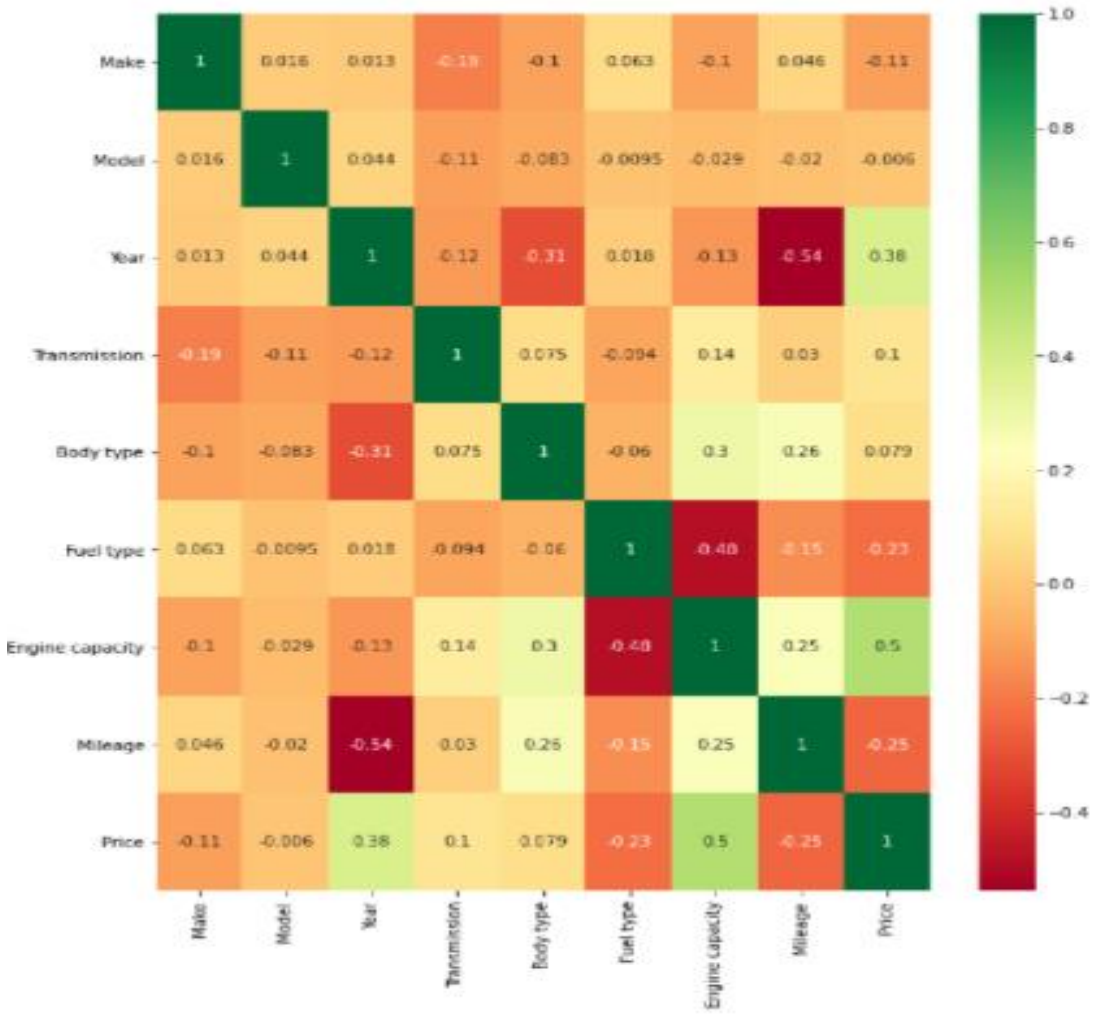
Proses *encoding*, yaitu transformasi fitur dari nominal ke numerik dilakukan pada penelitian ini. Fitur *make*, *model*, *transmission*, *body type* dan *fuel type* adalah fitur yang ditransformasikan. Hasil transformasi fitur dapat dilihat pada Tabel 3:

Tabel 3
Data Transformation

	Make	Model	Transmission	Body type	Fuel type
0	49	83	0	0	5
1	45	66	1	1	5
2	47	129	1	6	5
3	33	241	0	1	5
4	45	234	0	4	5

Feature korelasi menjelaskan hubungan antar *feature*. Jadi singkatnya sebuah fitur harus sangat berkorelasi dengan kelas dan tidak banyak berkorelasi ke fitur lain di kelas yang akan kita tuju (Gegic, Isakovic, Keco, Masetic, & Kevric, 2019). Pemilihan fitur merupakan langkah preprocessing data penting yang dilakukan sebelum algoritma pembelajaran diterapkan. Masalah yang harus dipertimbangkan saat mengusulkan metode pemilihan fitur adalah kompleksitas komputasinya (S Brunda, Nimish L, 2018).

Dalam penelitian ini beberapa korelasi terhadap price yang bernilai positif antara lain: *engine capacity*, *body type*, *transmission* dan *year*. Korelasi *price* dengan *year* adalah 0.38 artinya semakin muda tahun pembuatannya maka harga akan naik, berbanding lurus. Sedangkan *feature* yang berkorelasi negatif terhadap *price* antar lain: *maleage*, *fuel type*, *model* dan *make*. Korelasi *maleage* dengan *price* adalah -0.25 artinya semakin tinggi *maleage* atau daya tempuh mobil maka harga akan semakin turun, berbanding terbalik. *Heatmap* korelasi data digambarkan pada Gambar 2:



Gambar 2
Feature Corellation

Arsitektur yang kami usulkan pada penelitian ini adalah DNN dengan 3 hidden layer seperti dijelaskan pada tabel 4:

Tabel 4
Arsitektur DNN

Layer Input	Neuron	Leaarning rate	Activation	Batch Size
Layer Input	64	0.1	relu	40
Layer Hiden1	32	0.1	relu	40
Layer Hiden2	32	0.1	relu	40
Layer Hiden3	16	0.1	relu	40
Layer Output	1	0.1	relu	40

Untuk meningkatkan akurasi dalam neural network dalam penelitian ini adalah melakukan pengaturan jumlah neuron pada layer input, layer hidden dan layer output. Menggunakan activation relu, learning rate 0.01, epoch serta batch_size. Pengaturan hyperparameter untuk DNN pada tabel 5.

Tabel 5
Hyperparameter DNN

Hyperparameter	Nilai
epoch	15
Learning_rate	0.1
Batch_size	40
activation	relu

Sebagai pembanding, kami menerapkan random forest dan decision tree pada penelitian ini. Pohon keputusan atau dikenal dengan Decision Tree adalah salah satu metode klasifikasi yang menggunakan representasi struktur pohon (tree) dimana setiap node mempresentasikan nilai dari atribut, dan daun merepresentasikan kelas. Node yang paling atas dari deission tree disebut sebagai root (Pudaruth, 2014). Pohon keputusan adalah model pembelajaran mesin yang diawasi yang digunakan untuk memprediksi target dengan mempelajari aturan keputusan dari fitur (Doshi, 2014).

Dalam meningkatkan akurasi dalam decision tree bisa dilakukan dengan beberapa perubahan pada parameter seperti max_dept yang merupakan kedalaman layer setiap decision tre akan dibuat dan min_samples_split yang merupakan minimal sample dalam setiap node (Chen, 1959; Kaur & Kumari, 2020). Pengaturan *hyperparameter* untuk *decision tree* pada tabel 6:

Tabel 6
Hyperparameter decission tree

Hyperparameter	Nilai
max_dept	15
min_samples_split	3

Metode Random Forest (RF) adalah pengembangan dari decission tree, yaitu sekumpulan pohon keputusan (*decission tree*) yang dapat meningkatkan hasil akurasi, karena dalam membangkitkan simpul anak untuk setiap node dilakukan secara acak. Metod ini digunakan untuk membangun pohon keputusan yang terdiri dari root node, internal node, dan leaf node dengan mengambil atribut dan data secara acak sesuai ketentuan yang diberlakukan. Root node merupakan simpul yang terletak paling atas, atau bisa disebut sebagai akar dari pohon keputusan. Internal node adalah simpul percabangan, dimana node ini mempunyai output minimal dua dan hanya ada satu input. Sedangkan leaf node atau terminal node merupakan simpul terakhir yang hanya memiliki satu input dan tidak memiliki output (Botchkarev, 2018; Siji George C G, 2020).

Untuk meningkatkan akurasi pada random forest dengan melakukan hyperparameter diantaranya *n_estimators* dan *max_depth*. Parameter *n_estimators* ini yang akan menetukan berapa decission tree yang akan dibuat, sedangkan *max_depth* menyatakan seberapa dalam layer setiap decission tree akan dibuat (Syukron, Santoso, & Widiharih, 2020; Zhang, Yang, & Zhou, 2018). Tabel 7 adalah ringkasan pengaturan hyperparameter RF.

Tabel 7
Hyperparameter random forest

Hyperparameter	Nilai
max_depth	15
n_estimators	5

Dalam melakukan prediksi machine learning berdasarkan pembelajaran data (*supervised learning*) maka dalam penelitian ini berdasarkan data yang akan diteliti maka data akan dibagi menjadi dua bagian yaitu data training dan data testing.

Sesuai dengan pokok bahasan untuk prediksi harga mobil maka dalam penelitian ini yang menjadi feature X antara lain: *make, model, year, transmission, body type, fuel type* dan *eangine capacity*. Sedangkan untul label Y adalah price. Guna keperluan data training pada supervised machine learning, komposisi data latih 80% dari data bersih sebesar 5194 row sedangkan data test 20% sebesar 1299 row.

Evaluasi hasil eksperimen diukur menggunakan beberapa metrik. Pada penelitian ini, digunakan ukuran mean absolute error (MAE) dan mean square error (MSE) untuk mengukur tingkat kesalahan (Badrul, 2016; Blessie & Karthikeyan, 2012) Kedua metrik ini akan semakin baik jika nilainya semakin mendekati nol. Root square (R2) juga digunakan. Untuk R2, semakin mendekati 1 maka nilai prediksi akan semakin baik.

Hasil dan Pembahasan

A. Hasil Penelitian

Pada metode decision tree dengan hyperparameter pada max_dept dan min_samples_split. Dengan max_depth=15 diperoleh tingkat MAE=671632, R2=0.69 dengan waktu=0.2 detik. Sedangkan dengan memberikan nilai pada min_samples_split=3 diperolah tingkat MAE=669839, R2=0.72 dan waktu komputasi =0.2 detik (lihat Tabel 8).

Tabel 8
Nilai akurasi decission tree

Parameter		MSE	MAE	R2	Time (second)
max_depth					
	5	6233038238890.20	1066653	0.59	0.2
	10	5404309150881.32	766806	0.65	0.2
	15	4730371646659.62	671632	0.69	0.2
	18	4730371646659.62	697813	0.68	0.2
	20	5301763819167.29	708552	0.65	0.2
min_samples_split					
	2	4761858515075.70	686353	0.69	0.2
	3	4257164376378.25	669839	0.72	0.2
	5	4485489941663.47	697590	0.71	0.2

Dari hasil simulasi pada model decision tree dengan beberapa nilai max_depth yang tinggi tidak selalu menunjukkan akurasi yang baik sehingga memang harus dilakukan beberapa percobaan dengan nilai max_depth yang bervariasi sampai ditemukan nilai max_depth yang mempunyai akurasi yang baik. Begitupula pada nilai min_samples_split dilakukan percobaan dengan nilai bervariasi sampai mendapatkan nilai akurasi yang baik.

Sedangkan dengan menggunakan metode random forest dan melakukan hyperparameter pada *n_esimators*=5 diperolah tingkat MAE=565285, R2=0.83 dan waktu=0.2 detik. Sedangkan dengan memberikan *max_depth*=15 diperoleh tingkat MAE=521947, R2=0.84 dan waktu=0.2 detik (lihat Tabel 9).

Tabel 9
Nilai akurasi random forest

Parameter		MSE	MAE	R2	Time (second)
n_estimators	2	4653747155569.72	682865	0.69	0.2
	3	3270554589383.10	597116	0.79	0.2
	5	2655736319179.13	565285	0.83	0.2
	6	3410788263577.75	606339	0.78	0.2
max_depth					
	5	4051279078748.20	906654	0.73	0.2
	10	2676516745266.81	592193	0.82	0.2
	15	2466081929087.05	521947	0.84	0.2
	20	2518294731350.67	521685	0.83	0.2

Dari hasil simulasi ke beberapa nilai n_estimator yang tinggi tidak selalu menunjukkan akurasi yang baik sehingga memang harus dilakukan beberapa percobaan dengan nilai n_estimator yang bervariasi sampai ditemukan nilai n_estimator yang mempunyai akurasi yang baik. Begitupula pada nilai mx_depth dilakukan percobaan dengan nilai bervariasi sampai mendapatkan nilai akurasi yang baik

Metode yang diusulkan, DNN, memperoleh tingkat akurasi tertinggi pada metode ini diperoleh pada epoch =111 dengan tingkat MAE=501232, R2=0.88 dan waktu=0.51 detik, merupakan komputasi terlama dibandingkan metode yang lain. Hasil akurasi dapat dilihat pada Tabel 10:

Tabel 10
Nilai akurasi DNN

Parameter		MSE	MAE	R2	Time (second)
epoch	111	2329747618903.44	501232	0.88	0.51
	105	5802740777460.38	790568	0.62	0.51
	113	4006665539681.85	747402	0.74	0.51

Dari hasil penelitian dengan model DNN dengan banyak layer dan juga epoch untuk meningkatkan akurasi, disisi lain akan membutuhkan waktu proses yang

lama. DNN lebih cocok untuk proses data yang besar karena ada beberapa parameter yang mengatur pembagian data saat proses dengan parameter batch, sehingga data akan dilakukan proses secara bertahap sesuai pengaturan batch atas data yang akan diproses sampai data selesai.

Kesimpulan

Hasil penelitian yang dilakukan diperoleh tingkat akurasi prediksi dengan menggunakan model Decision Tree dan min_samples_split=3 dihasilkan akurasi 72% (0.72) dengan waktu proses 0.2 detik. Model Random Forest dan max_depth=15 dihasilkan akurasi 84% (0.84) dengan waktu proses 0.2 detik. Model DNN yang kami usulkan menghasilkan R2 0.88 dengan waktu proses 0.51 detik. Uji coba dalam penelitian ini menggunakan laptop dengan spesifikasi Processor: Intel Core™ i3-7100 CPU @ 2.40GHz, RAM: 8GHz, System Ops: windows 64 bit

Dapat disimpulkan bahwa dari hasil penelitian dengan menggunakan Random Forest Model lebih cepat waktu prosesnya dibandingkan Deep Neural Network dan secara akurasi keduanya cukup baik. Namun perlu diingat, walaupun perbedaan komputasi tidak terlalu besar karena tipe data yang digunakan sederhana, komputasi yang lebih besar dapat sangat mempengaruhi komputasi DNN.

Dimasa depan, menarik untuk diinvestigasi prediksi harga mobil berdasarkan faktor eksternal misalnya isu-isu lain yang tidak terkait dengan spesifikasi mobil itu sendiri. Pada saat pandemi saat ini terlihat bahwa isu-isu terkini, kondisi ekonomi dan lain-lain dapat memengaruhi fluktuasi harga mobil.Hal ini menarik untuk menjadi salah satu fitur dimasa depan.

BIBLIOGRAFI

Badrul, Mohammad. (2016). Optimasi Neural Network Dengan Algoritma Genetika Untuk Prediksi Hasil Pemilukada. *Bina Insani ICT Journal*, *3*(1), 229–242. Google Scholar

Blessie, E. Chandra, & Karthikeyan, E. (2012). Sigmis: A feature selection algorithm using correlation based method. *Journal of Algorithms & Computational Technology*, *6*(3), 385–394. Google Scholar

Botchkarev, Alexei. (2018). Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *ArXiv Preprint ArXiv:1809.03006*. Google Scholar

Chen, W. K. (1959). Linear Networks and Systems, Belmont, CA: Wadsworth, 1993. *JU Duncombe,"Infrared Navigation—Part I: An Assessment of Feasibility," IEEE Trans. Electron Devices*, *11*, 34–39. Google Scholar

Doshi, Mital. (2014). Correlation based feature selection (CFS) technique to predict student Perfromance. *International Journal of Computer Networks & Communications*, *6*(3), 197. Google Scholar

Gegic, Enis, Isakovic, Becir, Keco, Dino, Masetic, Zerina, & Kevric, Jasmin. (2019). Car price prediction using machine learning techniques. *TEM Journal*, *8*(1), 113. Google Scholar

Kaur, Harleen, & Kumari, Vinita. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*. Google Scholar

Murtadho, Ali. (2020). *Machine Learning Untuk Perbandingan Tingkat Akurasi Prediksi Penyakit Diabetes Dengan Supervised Learning*. Skripsi. Universitas 17 Agustus 1945 Surabaya. Google Scholar

Pandey, Abhishek, Rastogi, Vanshika, & Singh, Sanika. (2020). Car's Selling Price Prediction using Random Forest Machine Learning Algorithm. *5th International Conference on Next Generation Computing Technologies (NGCT-2019)*. Google Scholar

Pudaruth, Sameerchand. (2014). Predicting the price of used cars using machine learning techniques. *Int. J. Inf. Comput. Technol*, *4*(7), 753–764. Google Scholar

S Brunda, Nimish L, Chiranthan S. dan Arbaaz Khan. (2021). *Crop Price prediction using Random Forest and Decision Tree Regression*. Google Scholar

Siji George C G, B. Sumath. (2020). Grid Search Tuning of Hyperparameters in Random Forest Classifier for Customer Feedback Sentiment Prediction. *Turkish Journal*. Google Scholar

Syukron, Muhamad, Santoso, Rukun, & Widiharih, Tatik. (2020). Perbandingan Metode Smote Random Forest Dan Smote Xgboost Untuk Klasifikasi Tingkat Penyakit Hepatitis C Pada Imbalance Class Data. *Jurnal Gaussian*, *9*(3), 227–236. Google Scholar

Yusran, Yusran. (2016). Implementasi Jaringan Syaraf Tiruan (Jst) Untuk Memprediksi Hasil Nilai Un Menggunakan Metode Backpropagation. *Jurnal Ipteks Terapan*, *9*(4). Google Scholar

Zhang, Xingzhi, Yang, Yan, & Zhou, Zhurong. (2018). A novel credit scoring model based on optimized random forest. *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, 60–65. IEEE. Google Scholar