JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 5, Oktober 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i5.5021
Hal 1669−1678
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Pengembangan Model Machine Learning Regresi sebagai Web Service untuk Prediksi Harga Pembelian Mobil dengan Metode CRISP-DM

Ahmad Maulana Malik Fattah[*], Apriade Voutama, Nono Heryana, Nina Sulistiyowati

Fakultas Ilmu Komputer, Sistem Informasi, Universitas Singaperbangsa Karawang, Karawang, Indonesia
Email: [1,*]ahmad.maulana19003@student.unsika.ac.id, [2]apriade.voutama@staff.unsika.ac.id, [3]nono@unsika.ac.id, [4]nina.sulistio@unsika.ac.id
Email Penulis Korespondensi: ahmad.maulana19003@student.unsika.ac.id
Submitted **24-10-2022**; Accepted **31-10-2022**; Published **31-10-2022**

Abstrak

Seiring meningkatnya kebutuhan masyarakat terhadap moda transportasi mobil, aktivitas pelaku bisnis penjualan mobil pun turut meningkat. Upaya untuk tetap eksis dan kompetitif dilakukan seperti dengan penerapan model machine learning untuk menentukan harga jual mobil berdasarkan spesifikasinya. Pelaku bisnis juga dapat menstimulus peningkatan penjualan dengan memberikan iklan atau penawaran secara aktif pada pelanggan. Penawaran yang aktif dan masif dapat ditingkatkan efektifitasnya dengan melakukan personalisasi terhadap penawaran yang diberikan. Penelitian ini melakukan pendekatan berbasis machine learning untuk mempelajari data profil pelanggan untuk memprediksi harga mobil yang akan dibeli pelanggan tersebut. Penelitian dilakukan dengan mengadopsi metode CRISP-DM serta dikerjakan melalui platform Google Colaboratory dan Azure Machine Learning. Tahap pemodelan menghasilkan enam model regresi yaitu regresi linear, Lasso, Ridge, Random Forest Regressor, Elastic-net, dan Support Vector Regressor (SVR). Setelah melalui tahap evaluasi, model regresi Lasso dengan performa R-squared (R^2) sebesar 0,99958 dan Mean Absolute Error (MAE) sebesar 2.284.865,29 disebarkan sebagai web service endpoint sehingga dapat diakses secara waktu nyata (real-time). Web service tersebut memerlukan data "Gender, Age, Annual Salary, Credit Card Debt, dan Net Worth" pelanggan untuk mengembalikan response berupa prediksi rentang harga mobil yang direkomendasikan untuk dibeli pelanggan tersebut. Pada pengembangan lebih lanjut, prediksi yang diperoleh melalui web service dapat diimplementasikan pada aplikasi publik untuk menampilkan penawaran atau laman penjualan mobil yang terpersonalisasi berdasarkan profil pelanggan.

Kata Kunci: CRISP-DM; Machine Learning; Model Regresi; Pembelian Mobil; Web Service

Abstract

Along with the increasing public demand for car transportation modes, car sales businesses are also increasing. Efforts to exist and be competitive are carried out such as by applying machine learning models to determine the car's selling price based on its specification. Businesses can also stimulate sales by actively offering customers. The effectiveness of the active and massive offerings can be increased by personalizing the offers provided. This research uses a machine learning-based approach to learn customer profile data to predict the car's price they would buy. The research was conducted by adopting the CRISP-DM framework and developed using the Google Colaboratory and Azure Machine Learning platforms. The modeling stage developed six regression models, those are linear regression, Lasso, Ridge, Random Forest Regressor, Elastic-net, and Support Vector Regressor (SVR). After the evaluation stage, the Lasso regression model with the performance of R-squared (R^2) of 0,99958 and Mean Absolute Error (MAE) of 2.284.865,29 deployed as a web service endpoint so it could be accessed in real-time. The web service required the customer's "Gender, Age, Annual Salary, Credit Card Debt, and Net Worth" to return a response of the recommended car price range prediction for the customer to buy. In further development, predictions obtained through web services can be implemented in public applications to display personalized car sales offers or pages based on customer profiles.

Keywords: CRISP-DM; Machine Learning; Regression Model; Car Purchase; Web Service

1. PENDAHULUAN

Peningkatan aktivitas bisnis dan kebutuhan mobilitas mendorong masyarakat dari berbagai kalangan untuk memiliki moda transportasi yang dapat memenuhi kebutuhan mereka, salah satunya yaitu mobil. Pertumbuhan minat masyarakat ini mendorong meningkatnya bisnis penjualan mobil dengan pasar yang kompetitif. Setiap *showroom* penjualan mobil berusaha untuk meningkatkan penjualan mereka sehingga mampu mempertahankan bisnis dan bersaing dengan kompetitor lain [1]. Di antara cara yang dilakukan pelaku bisnis untuk dapat bersaing adalah dengan memberikan iklan-iklan dan menetapkan harga jual mobil yang optimal berdasarkan spesifikasi mobil. Pendekatan menggunakan model pembelajaran mesin (*machine learning*) dapat digunakan untuk mengimplementasikan kedua solusi tersebut.

Machine learning adalah studi yang mempelajari algoritma dan model statistik yang dapat digunakan oleh sistem komputer untuk melakukan tugas tertentu tanpa memerlukan instruksi eksplisit [2]. Salah satu jenis pemodelan dalam *machine learning* adalah model regresi yang dapat memprediksi nilai numerik dengan mempelajari data lampau [3]. Penelitian yang dilakukan [4] mengembangkan model dengan metode regresi linear berganda yang dapat memprediksi harga jual mobil bekas berdasarkan parameter seperti merek, selisih tahun berjalan dengan tahun perakitan, harga awal, dan kondisi mobil. Pengujian pada sampel penjualan mobil "Brio" tahun 2014 yang dijual pada tahun 2019 menghasilkan estimasi harga jual sebesar Rp114.000.000 dengan nilai aktual sebesar Rp112.000.000. Pengembangan model *machine learning* untuk kasus prediksi yang sama juga dilakukan oleh [1] dengan metode *Deep Neural Network (DNN)*. Data yang digunakan merupakan kumpulan data penjualan mobil bekas di India yang diperoleh melalui *platform* Kaggle. Model prediksi yang dibangun memanfaatkan fitur-fitur di antaranya merek, model, tahun produksi, dan spesifikasi mobil. Model tersebut menggunakan 3 *hidden layer* dan menghasilkan *Mean Absolute Error* (MAE) sebesar 501.232 dan R^2 sebesar

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 5, Oktober 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i5.5021
Hal 1669−1678
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

0,88. Penelitian yang dilakukan oleh [3] memprediksi harga jual mobil menggunakan model *Random Forest* dan *Extra Tree Regression* dengan hasil evaluasi yang tidak disebutkan. Model yang telah dibangun disebarkan sebagai sebuah aplikasi web melalui *platform* Heroku. Pengguna dapat memasukkan parameter yang dibutuhkan melalui antarmuka web, kemudian aplikasi akan menghasilkan prediksi harga jual mobil berdasarkan parameter tersebut.

Temuan dari penelitian-penelitian di atas dapat dijadikan dasar untuk implementasi sistem penetapan harga jual mobil oleh para pelaku bisnis. Selain penetapan harga jual yang optimal, cara lain yang dapat dilakukan pelaku bisnis untuk menarik minat pelanggan adalah dengan melakukan pengiklanan secara aktif. Namun, apabila dilihat dari sisi pelanggan, pemberian iklan-iklan yang dilakukan pelaku pasar juga berpotensi menghadirkan masalah tersendiri, yaitu terlalu banyak dan variatif informasi penawaran yang beredar. Hal ini dapat menyulitkan pelanggan dalam memutuskan mobil mana yang cocok untuk dibeli hingga bahkan mengabaikan penawaran yang diberikan [5]. Pelanggan dapat diberikan penawaran atau rekomendasi yang terpersonalisasi dibanding dengan sekadar menerima penawaran produk secara aktif dan masif namun acak.

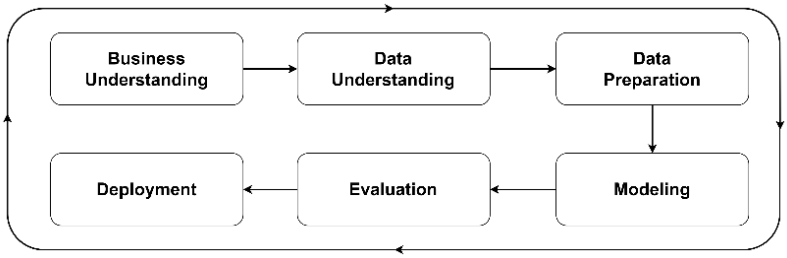
Sistem yang dikembangkan [6] dapat memberikan rekomendasi penawaran produk kepada pelanggan web lokapasar (*e-commerce*). Sistem mengimplementasikan algoritma *K-Means Clustering* untuk mengelompokkan pelanggan berdasarkan karakteristik riwayat transaksi yang dilakukan. Hasil penelitian menemukan terdapat 3 *cluster* pelanggan dengan karakteristik rentang umur dan penjualan yang berbeda. Temuan tersebut diimplementasikan pada halaman produk berbasis web sehingga tiap pengguna mendapatkan rekomendasi produk yang terpersonalisasi. Sementara itu, [7] mengimplementasikan metode *Content-based Filtering* dengan *cosine similarity* untuk menampilkan daftar rekomendasi musik berdasarkan riwayat pemutaran musik pengguna. Dengan memanfaatkan lebih dari 11.000 lagu dan 130.000 riwayat pemutaran, model yang dibangun memiliki performa *recall* antara 0,15 hingga 0,2 dan *precision* antara 0,015 hingga 0,125. Namun, sistem yang dibangun dapat mengalami penurunan performa apabila ukuran *dataset* semakin bertambah besar.

Kajian penerapan teknologi *big data* yang dilakukan [8] menghimpun implementasi model-model prediksi di industri *travel* dan pariwisata (*tourism*) yang menggunakan perjalanan udara. Salah satu penerapan tersebut adalah harga dinamis tiket pesawat yang dapat berubah untuk tiap penumpang dalam satu penerbangan untuk kursi yang sama. Apabila dilihat dari sisi penyedia maskapai, maka harga yang dihasilkan dapat memberikan profit maksimum. Di saat yang sama, harga tersebut juga merupakan harga tiket dengan biaya minimum yang harus dikeluarkan pelanggan. Model lain dalam kajian tersebut memprediksi waktu yang optimal bagi pelanggan untuk membeli tiket. Pemodelan dengan teknik *nu-SRV ridge regression*, *decision tree*, dan *PLS regression* dapat membuat pelanggan berhemat 69-75,3% untuk membeli tiket.

Pada penelitian ini, personalisasi rekomendasi harga pembelian mobil dilakukan dengan memanfaatkan beberapa data profil pelanggan, seperti usia, jenis kelamin, pendapatan per tahun, jumlah utang, dan total aset kekayaan. Melalui pemodelan *machine learning*, variabel-variabel tersebut dapat dimanfaatkan sebagai variabel prediktor untuk dapat memprediksi harga mobil yang relevan untuk ditawarkan. Dengan target prediksi berupa data kontinu, maka model *machine learning* yang dibangun akan menerapkan metode regresi. *Dataset* yang digunakan dalam penelitian berasal dari data publik pada *platform* Kaggle [9] dengan ukuran *dataset* sebesar 500 baris dan 9 kolom. Model prediksi yang telah dibangun disebarkan sebagai sebuah *web service endpoint* yang dapat dimanfaatkan oleh pengguna manusia maupun program komputer. *Web service* tersebut dirancang untuk dapat memberikan rekomendasi rentang harga mobil yang relevan untuk ditawarkan pada pelanggan.

2. METODOLOGI PENELITIAN

Penelitian dilakukan dengan mengadaptasi metodologi *Cross-Industry Standard Process for Data Mining* (CRISP-DM). CRISP-DM merupakan sebuah standar yang digunakan dalam melakukan penambangan data (*data mining*). Metode ini digagas oleh Daimler Chrysler (Daimler-Benz), SPSS (ISL), dan NCR yang selanjutnya dikembangkan oleh 300 organisasi hingga model ini dipublikasikan pada tahun 1999 dengan nama CRISP-DM 1.0 [10]. Meski awalnya ditujukan spesifik untuk *data mining*, standar ini juga relevan digunakan dalam lingkup lain di bidang data seperti pengembangan *machine learning*. CRISP-DM terdiri dari proses siklis sebagaimana tertera pada Gambar 1.



Gambar 1. Alur CRISP-DM

3.1 *Business Understanding*

Tahap *Business Understanding* meliputi aktivitas-aktivitas untuk memahami kasus dari *dataset* yang digunakan. Pemahaman yang diperoleh akan dijadikan dasar untuk menentukan model *machine learning* yang dapat dibangun dari

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 5, Oktober 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i5.5021
Hal 1669−1678
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

dataset tersebut. *Dataset* yang digunakan bersumber utama dari *platform* Kaggle dengan nama asli "*Car Purchase Price (beginner dataset)*" [9]. Pada bagian selanjutnya dari artikel ini, *dataset* ini disebut *dataset* "*Otomotic*".

3.2 *Data Understanding*

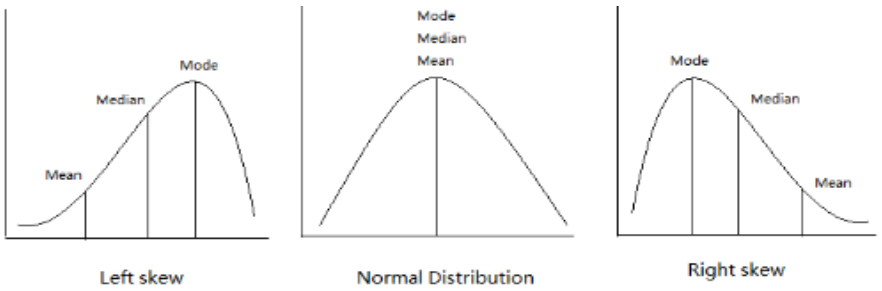
Tahap *Data Understanding* dilakukan dengan menggunakan teknik *Exploratory Data Analysis* (EDA). EDA merupakan proses menganalisis data untuk mendapatkan pemahaman yang lebih baik mengenai data tersebut [11]. EDA dilakukan pada *dataset* "Otomotic" untuk menggali *insight* yang dapat digunakan pada proses selanjutnya. Aktivitas yang dilakukan dengan teknik EDA adalah menganalisis statistik deskriptif dari *dataset*, mengidentifikasi *missing value*, menganalisis korelasi dan distribusi data, dan mengidentifikasi data pencilan (*outlier*).

2.2.1 Analisis Statistik Deskriptif dan Identifikasi *Missing Value*

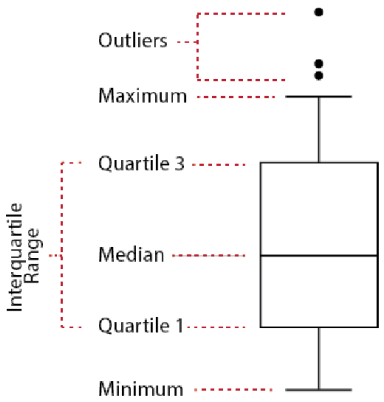
Analisis statistik deskriptif dilakukan untuk mengetahui karakteristik *dataset*. Pada analisis ini, ditampilkan nilai-nilai statistik seperti jumlah data, rerata, standar deviasi, kuartil, serta nilai minimum dan maksimum pada setiap kolom/fitur. Hasil analisis juga digunakan untuk menetapkan kolom-kolom prediktor dan kolom target prediksi. Sementara itu, identifikasi *missing value* dilakukan untuk memastikan data yang digunakan saat pembangunan model tidak mengandung data tanpa nilai.

2.2.2 Analisis Distribusi Data dan Identifikasi *Outlier*

Analisis distribusi data dilakukan guna mengetahui apakah data pada suatu fitur tersebar secara normal. Distribusi normal adalah distribusi simetris dengan posisi modus, median, dan rerata (*mean*) berada di garis tengah kurva atau mendekati garis tersebut [12]. Apabila suatu fitur memiliki data yang terdistribusi normal, maka sisi kanan dan kiri kurva akan memiliki luas mendekati 50% dari luas kurva [13].



Gambar 2. Tipe distribusi data [14]



Gambar 3. Bagian-bagian *boxplot* [15]

Fitur dengan distribusi tak normal dapat mengindikasikan adanya data pencilan (*outlier*). *Outlier* merupakan data yang terletak di luar distribusi normal dan jauh dari pusat data. Deteksi data *outlier* dapat dilakukan dengan memvisualisasikan fitur *dataset* dengan *boxplot* [16]. Pada visualisasi *boxplot*, data dengan nilai yang melewati 1,5 kali rentang interkuartil atas maupun bawah dianggap sebagai *outlier* dan divisualisasikan sebagai titik data sebagaimana pada Gambar 3. *Outlier* yang ditemukan akan ditangani dengan metode tertentu, seperti menghapus atau mengimputasi baris data terkait maupun mengeksklusi fitur terkait dari prediktor.

3.3 *Data Preparation*

Pada tahap ini, beberapa transformasi diterapkan terhadap *dataset* sehingga seluruh *input* yang diberikan pada model adalah data bertipe numerik sehingga dapat diproses sistem dengan lebih mudah [17][18]. Fitur dengan data yang terlalu

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 5, Oktober 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i5.5021
Hal 1669−1678
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

variatif atau tak dapat digeneralisasi akan dihilangkan selama proses ini. Pembagian data latih (*train*) dan data uji (*test*) juga dilakukan dengan proporsi data latih banding data uji sebesar 8:2 sebagaimana direkomendasikan dalam [19].

3.4 *Modeling* dan *Evaluation*

Hasil analisis yang dilakukan pada tahap *Business Understanding* menjadi dasar pembangunan model. Berdasarkan hasil analisis yang diperoleh, maka ditentukan bahwa model yang akan dibangun merupakan model regresi. Pemodelan regresi memiliki luaran berupa bilangan riil sebagai hasil prediksi. Persamaan dasar yang digunakan dalam pemodelan regresi ditulis sebagai [3]:

$$\hat{y} = w[0] * x[0] + w[1] * x[1] + \cdots + w[i] * x[i] + b \quad (1)$$

Di mana ŷ sebagai target prediksi, x[i] berupa fitur-fitur data, w[i] dan b sebagai parameter yang dicari selama pelatihan model.

Terdapat bermacam-macam algoritma yang dapat digunakan dalam pemodelan regresi, di antaranya adalah regresi linear, Lasso, Ridge, *Random Forest Regressor, Elastic-net, dan Support Vector Regressor (SVR)*. Pada penelitian ini, masing-masing algoritma tersebut dibangun menjadi model *machine learning* untuk mempelajari pola harga mobil yang akan dibeli oleh seorang pelanggan berdasarkan data profil yang dimilikinya. Dengan mempelajari pola tersebut, model dapat digunakan untuk memberikan rekomendasi rentang harga mobil yang relevan untuk ditawarkan pada pelanggan lain yang memiliki data profil berbeda dengan data yang dijadikan dasar pembangunan model.

Performa model yang telah dibangun diukur menggunakan matriks *R-squared* (R^2) dan *Mean Absolute Error* (MAE) dengan dasar kemudahan dalam interpretasi. Matriks evaluasi R^2 digunakan untuk mengetahui seberapa besar variabel dependen dapat direpresentasikan oleh variabel independen. Semakin tinggi nilai R^2 yang dicapai oleh model *machine learning*, maka semakin baik performa dari model tersebut. Persamaan matriks R^2 dituliskan sebagai berikut [20].

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\Sigma_i (y_i - \hat{y}_i)^2}{\Sigma_i (y_i - \bar{y}_i)^2} \quad (2)$$

Matriks evaluasi MAE digunakan untuk mengevaluasi seberapa besar *error* yang muncul antara hasil prediksi dengan hasil aktual. Berkebalikan dengan R^2, semakin rendah nilai MAE yang dicapai oleh suatu model *machine learning*, maka semakin baik performa dari model tersebut. Persamaan matriks MAE dituliskan sebagai berikut.

$$MAE = \frac{\Sigma_{i=1}^{n} |y_i - x_i|}{n} \quad (3)$$

Hasil evaluasi terhadap model digunakan untuk menentukan model terbaik untuk dilanjutkan pada tahap *Deployment* [21].

3.5 *Deployment*

Model dengan performa terbaik dipilih untuk dilanjutkan pada tahap *Deployment*. Pada tahap ini, model dirancang agar dapat dimanfaatkan dengan kebutuhan prediksi secara waktu nyata (*real-time*). Berdasarkan kebutuhan tersebut, maka model diunggah sebagai sebuah *web service endpoint* melalui layanan analisis prediktif yang tersedia pada *platform* Microsoft Azure, yaitu Azure Machine Learning (ML) [22].

Azure ML menyediakan fitur *endpoint deployment* untuk mengunggah model dan membuatnya *online* sebagai sebuah *RESTful web service* . Dengan demikian, model prediksi yang telah dibangun dapat dikonsumsi oleh pengguna atau layanan (*service*) lain, misalnya sebuah aplikasi berbasis web atau *mobile*. Melalui fitur tersebut, *endpoint* dirancang untuk menerima data dengan format *Javascript Object Notation* (JSON) yang berisi data yang dibutuhkan model untuk melakukan prediksi. Respon yang diberikan *endpoint* juga dirancang dalam format JSON yang berisi hasil prediksi dari model.

3. HASIL DAN PEMBAHASAN

Tahapan pada metode CRISP-DM dilakukan melalui *platform* Google Colaboratory dan Azure Machine Learning. Selain itu, pada tahap pemahaman data, metode-metode statistika juga digunakan di antaranya untuk menganalisis distribusi data dan deteksi *outlier*. Beberapa model *machine learning* regresi dibangun dan dievaluasi dengan matriks evaluasi untuk kasus regresi. Pada tahap akhir dilakukan *deployment* terhadap model dengan performa terbaik sebagai sebuah *web service endpoint*.

3.1 *Business Understanding*

Dataset "Otomotic" merepresentasikan profil pelanggan seperti nama, usia, jenis kelamin, utang, total kekayaan beserta harga mobil yang pernah dibelinya. Nilai uang pada *dataset* tersebut tidak dalam nilai mata uang Rupiah sehingga dilakukan konversi nilai mata uang ke Rupiah. *Dataset* memiliki total data sebanyak 500 baris dan terdiri dari fitur-fitur berikut.

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 5, Oktober 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i5.5021
Hal 1669−1678
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Tabel 1. Fitur-fitur *dataset*

Fitur	Representasi Nilai	Tipe Data		Nilai Unik	Jumlah *Missing*
		DataFrame	Statistika	Nominal	*Value*
Customer Name	Nama pelanggan	object	Nominal	498	0
Customer e-mail	Alamat *email* pelanggan	object	Nominal	500	0
Country	Negara	object	Nominal	1	0
Gender	Jenis kelamin {'0' = wanita, '1' = pria}	int64	Nominal	2	0
Age	Usia pelanggan	int64	Diskrit	-	0
Annual Salary	Pendapatan per tahun	float64	Kontinu	-	0
Credit Card Debt	Total utang pada kartu kredit	float64	Kontinu	-	0
Net Worth	Kekayaan bersih	float64	Kontinu	-	0
Car Purchase Amount	Harga mobil yang dibeli	float64	Kontinu	-	0

3.2 *Data Understanding*

Tahap *Data Understanding* meliputi aktivitas analisis statistik deskriptif data, identifikasi *missing value*, analisis distribusi data, dan identifikasi data pencilan (*outlier*).

3.2.1 Analisis Statistik Deskriptif dan Identifikasi *Missing Value*

Pada tahap ini digunakan teknik *Exploratory Data Analysis* (EDA). Berdasarkan analisis yang dilakukan, diperoleh nilai-nilai statistik deskriptif seperti jumlah data, rerata, standar deviasi, kuartil, serta nilai minimum dan maksimum pada setiap fitur numerikal. Sementara pada fitur kategorikal, diperoleh nilai jumlah data tiap label.

Tabel 2. Statistik deskriptif fitur numerikal

	Gender	*Age*	*Annual Salary*	*Credit Card Debt*	*Net Worth*	*Car Purchase Amount*
count	500.0	500.0	500.0	500.0	500.0	500.0
mean	0.506	46.224	575646158.617	89773835.334	100792726.703	516370454.87
std	0.500	7.99	136695457.699	32602972.419	40538186.281	125830727.73217237
min	0.0	20.0	83600000.0	934400.0	4672000.0	105120000.0
25%	0.0	41.0	485298293.650	69122387.56	70038932.157	439517185.775
50%	1.0	46.0	584853005.35	90216652.345	99688828.185	513894110.0
75%	1.0	52.0	668976628.25	110248617.775	130190998.225	598655007.175
max	1.0	70.0	1018000000.0	186880000.0	233600000.0	934400000.0

Tabel 3. Statistik deskriptif fitur kategorikal

	Customer Name	*Customer e-mail*	*Country*
count	500	500	500
unique	498	500	1

Pada tahap identifikasi *missing value*, tidak ditemukan *missing value* di fitur manapun dari *dataset* sebagaimana disajikan dalam Tabel 1. Dengan demikian, seluruh baris pada *dataset* telah memiliki nilai. Selanjutnya, berdasarkan hasil analisis statistik deskriptif yang dilakukan, *dataset* "*Otomotic*" teridentifikasi memiliki fitur yang dapat digunakan sebagai target prediksi, yaitu fitur "*Car Purchase Amount*". Oleh sebab fitur "*Car Purchase Amount*" bertipe data numerik, maka disimpulkan bahwa pemodelan yang akan dilakukan terhadap *dataset* adalah pemodelan regresi. Fitur-fitur yang digunakan sebagai prediktor terdiri dari "*Gender*, *Age*, *Annual Salary*, *Credit Card Debt*, dan *Net Worth*". Kelimanya dipilih karena memiliki nilai berbasis numerik dan dapat digeneralisasi, tidak terlalu unik seperti fitur "*Customer Name*" dan "*Customer e-mail*", dan cukup variatif tidak seperti fitur "*Country*" yang hanya memiliki satu nilai unik sebagaimana tertera pada Tabel 3.

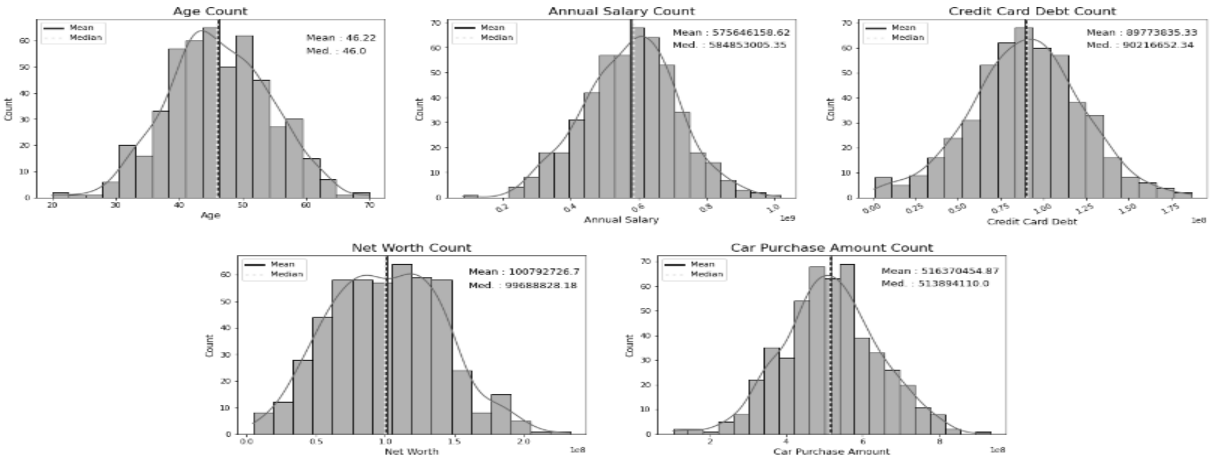
3.2.2 Analisis Distribusi Data

Analisis distribusi data dilakukan pada fitur yang secara statistika bertipe data kontinu. Fitur-fitur tersebut terdiri dari "*Age*, *Annual Salary*, *Credit Card Debt*, *Net Worth*, dan *Car Purchase Amount*". Analisis distribusi data dilakukan dengan memvisualisasikan histogram setiap fitur, dimana sumbu X merepresentasikan nilai fitur dan sumbu Y merepresentasikan banyak data pada setiap nilai di sumbu X.

Berdasarkan visualisasi yang dilakukan terhadap fitur "*Age*", diketahui bahwa fitur tersebut memiliki nilai rerata (*mean*) sebesar 46,22 dan median sebesar 46 dengan puncak grafik berada di tengah. Sementara itu, fitur "*Annual Salary*" memiliki nilai rerata sebesar 575.646.158,62 dan median sebesar 584.853.005,35 dengan puncak grafik yang juga berada di tengah. Selanjutnya, visualisasi dilakukan terhadap fitur "*Credit Card Debt*" dan "*Net Worth*". Berdasarkan hasil visualisasi, diketahui bahwa fitur "*Credit Card Debt*" memiliki nilai rerata sebesar 89.773.835,33 dan median sebesar 90.216.652,34 dengan puncak grafik berada di tengah. Sementara itu, fitur "*Net Worth*" memiliki nilai rerata sebesar

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 5, Oktober 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i5.5021
Hal 1669−1678
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

100.792.726,7 dan median sebesar 99.688.828,19 dengan puncak grafik berada di tengah. Visualisasi terakhir dilakukan terhadap fitur "*Car Purchase Amount*" dan diperoleh informasi bahwa fitur tersebut memiliki nilai rerata sebesar 516.370.454,87 dan median sebesar 513.894.110 dengan puncak grafik berada di tengah.



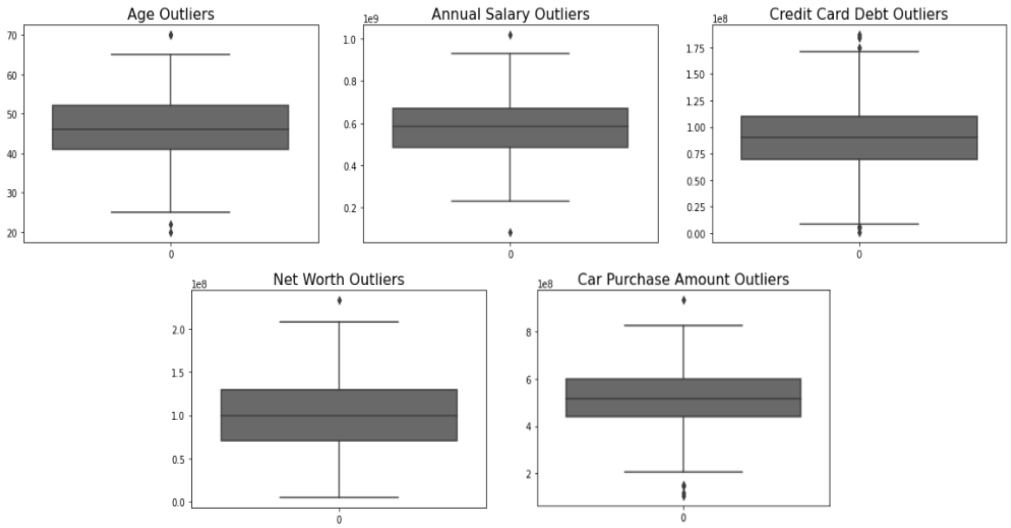
Gambar 4. Distribusi Data Kontinu

Berdasarkan visualisasi distribusi data pada kelima fitur di atas, diketahui bahwa nilai rerata (*mean*) dan median pada setiap fitur tidak memiliki selisih yang jauh. Selain itu, puncak grafik pada kelima fitur juga cenderung berada di tengah sumbu X. Dengan demikian, dapat disimpulkan bahwa persebaran data pada fitur "*Age*, *Annual Salary*, *Credit Card Debt*, *Net Worth*, dan *Car Purchase Amount*" cenderung normal.

3.2.3 Identifikasi Data Pencilan (*Outlier*)

Identifikasi *outlier* dilakukan menggunakan visualisasi *boxplot*. Pada proyek ini, deteksi *outlier* dilakukan dengan memvisualisasikan data dengan *boxplot*. Pada fitur "*Age*", visualisasi *boxplot* menampilkan 1 titik data di luar batas atas dan 2 titik data di luar batas bawah. Hal ini mengindikasikan terdapat setidaknya 3 baris pada fitur "*Age*" yang mengandung *outlier*. Sementara itu, visualisasi pada fitur "*Annual Salary*" menampilkan 1 titik data *outlier* di luar batas atas maupun bawah yang berarti terdapat setidaknya 2 baris pada fitur "*Annual Salary*" yang mengandung *outlier*.

Identifikasi *outlier* juga dilakukan pada fitur "*Credit Card Debt*" dan "*Net Worth*". Berdasarkan hasil visualisasi, pada fitur "*Credit Card Debt*", terdeteksi 3 titik data di luar batas atas dan 2 titik data di luar batas bawah *boxplot*. Sementara pada fitur "*Net Worth*", terdapat 1 titik data di luar batas atas *boxplot*. Hal ini berarti terdapat setidaknya 5 data *outlier* pada fitur "*Credit Card Debt*" dan 1 data *outlier* pada fitur "*Net Worth*". Deteksi *outlier* juga dilakukan terhadap fitur target "*Car Purchase Amount*" dan diperoleh 1 titik data di luar batas atas dan 3 titik data di luar batas bawah *boxplot*. Ini berarti terdapat setidaknya 4 baris pada fitur "*Car Purchase Amount*" yang mengandung *outlier*.



Gambar 5. Identifikasi *Outlier* dengan *Boxplot*

Hasil deteksi *outlier* mengindikasikan diketahui terdapat *outlier* pada fitur "*Age*, *Annual Salary*, *Credit Card Debt*, *Net Worth*, maupun *Car Purchase Amount*". Pada awal pengembangan proyek, transformasi diberlakukan pada data *outlier* tersebut dengan mengubahnya ke nilai *mean* atau median dari masing-masing fitur. Namun, pada tahap

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 5, Oktober 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i5.5021
Hal 1669−1678
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

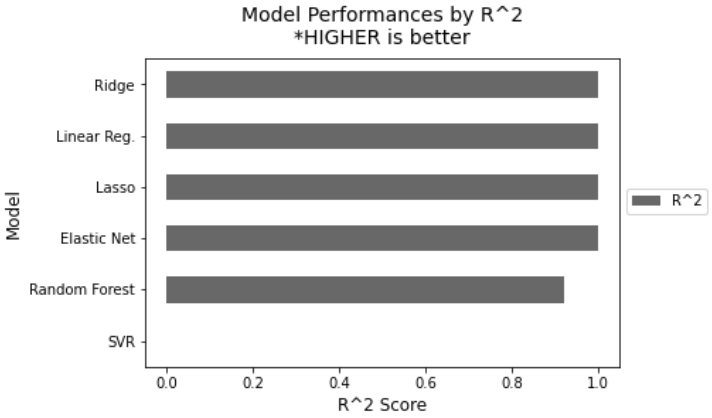
pengembangan lebih lanjut, diketahui bahwa transformasi tersebut justru menurunkan performa model. Oleh sebab itu, proses transformasi terhadap *outlier* pada kelima fitur di atas tidak lagi dilakukan dan data *outlier* dibiarkan sebagaimana adanya.

3.3 *Data Preparation*

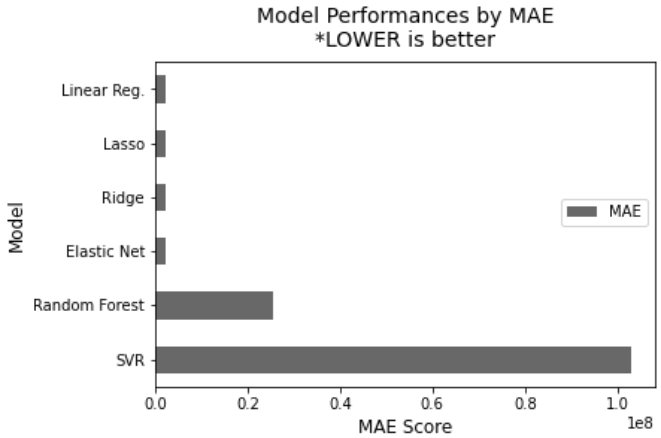
Aktivitas pertama pada tahap *data preparation* adalah membuang fitur-fitur yang tidak digunakan sebagai prediktor maupun target. Fitur-fitur tersebut adalah "*Customer Name*, *Customer e-mail*, dan *Country*". Pembagian data latih dan data uji kemudian dilakukan dengan komposisi 8:2. Dengan demikian, *dataset* yang semulanya memiliki 500 baris dibagi menjadi data latih sebanyak 400 baris dan data uji sebanyak 100 baris. Pada tahap ini juga fitur target dipisahkan dari fitur-fitur prediktornya. Fitur prediktor yang digunakan adalah adalah "*Gender*, *Age*, *Annual Salary*, *Credit Card Debt*, dan *Net Worth*". Sementara itu, fitur target yang digunakan adalah "*Car Purchase Amount*".

3.4 *Modeling* dan *Evaluation*

Dataset "Otomotic" memiliki fitur target bertipe data numerikal sehingga pemodelan *machine learning* yang relevan digunakan adalah pemodelan regresi. Algoritma yang dikembangkan terdiri dari regresi linear, Lasso, Ridge, *Elastic-net*, *Support Vector Regression* (SVR), dan *Random Forest Regressor*. Performa model dievaluasi pada data uji dengan dua matriks, yaitu R-*squared* (R^2) dan *Mean Absolute Error* (MAE).



Gambar 6. Perbandingan performa R^2 model



Gambar 7. Perbandingan performa MAE model

Hasil analisis performa model yang dilakukan dengan matriks evaluasi R^2 menunjukkan bahwa model Ridge, regresi linear, dan Lasso memiliki nilai R^2 yang sama baiknya pada data uji, yaitu berkisar 0,99958. Model *Elastic-net* dan *Random Forest Regressor* juga memiliki performa yang terbilang tinggi, masing-masing sebesar 0,999572 dan 0,921394. Sementara itu, model SVR memiliki performa terendah dengan nilai R^2 sebesar -0.000526.

Analisis performa dengan menggunakan matriks MAE menunjukkan bahwa model regresi linear memiliki nilai MAE terendah, yaitu sebesar 2.284.865,23. Model selanjutnya dengan nilai MAE terendah adalah model Lasso, Ridge, dan *Elastic-net*, masing-masing sebesar 2.284.865,29; 2.284.994,66; dan 2.332.398,66. Pada model *Random Forest Regressor* dan SVR, nilai MAE yang dihasilkan terbilang cukup tinggi, yaitu sebesar 25.483.869,81 dan 102.936.050,59.

Berdasarkan hasil perbandingan performa model regresi linear, Lasso, Ridge, *Elastic-net*, *Random Forest Regressor*, dan SVR dengan matriks R^2 dan MAE, diketahui bahwa model regresi linear memiliki performa terbaik dengan nilai R^2 sebesar 0,999580 dan MAE sebesar 2.284.865,23. Model Lasso, Ridge, dan *Elastic-net* juga memiliki

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 5, Oktober 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i5.5021
Hal 1669−1678
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

performa yang tinggi dengan masing-masing model memiliki nilai MAE sebesar 2.284.865,29; 2.284.994,66; dan 2.332.398,66. Hal ini menunjukkan bahwa model regresi linear, Lasso, Ridge, dan *Elastic-net* memiliki performa yang tinggi sehingga dapat digunakan untuk memprediksi harga mobil yang akan dibeli oleh pelanggan berdasarkan profil yang ia miliki.

3.5 *Deployment*

Tahap akhir pengembangan model *machine learning* adalah *model deployment*. Berdasarkan analisis performa model yang dilakukan, ditetapkan bahwa model yang akan diunggah adalah model Lasso. Ini diputuskan sebab pemodelan Lasso merupakan salah satu model dengan performa yang tinggi selama tahap *modeling*. Model ditargetkan dapat dikonsumsi pada aplikasi dengan berbagai basis (*multiplatform*) dengan kemampuan prediksi secara waktu nyata (*real-time*).

Berdasarkan target kebutuhan tersebut, maka model diunggah sebagai sebuah *web service endpoint* melalui layanan Azure Machine Learning. Model yang telah diunggah dapat diakses melalui *Representational State Transfer* (REST) *endpoint*. *Endpoint* tersebut akan menerima data dengan format *Javascript Object Notation* (JSON) yang berisi data profil pengguna, meliputi "*Customer Name*, *Customer e-mail*, *Country*, *Gender*, *Age*, *Annual Salary*, *Credit Card Debt*, dan *Net Worth*". Model akan menggunakan data tersebut sebagai dasar prediksi "*Car Purchase Amount*". *Endpoint* akan mengembalikan *response* berupa nilai hasil prediksi, hasil prediksi yang dikenakan MAE, dan pesan status prediksi sukses atau gagal. Tabel berikut menyajikan skema *response* yang dihasilkan *endpoint*.

Tabel 4. Skema *response* dari *endpoint*

Key	*Value*	*Subvalue*
Result	pred_value	Nilai aktual hasil prediksi model.
	min_value	0 (nol).
	max_value	Hasil penjumlahan "pred_value" dengan MAE model.
Status	*Successed, Failed*	-

Pada *subvalue* "min_value", nilai yang dikembalikan berupa 0 (nol), bukan selisih hasil pengurangan nilai "pred_value" dengan skor performa MAE model. Hal ini didasarkan agar *endpoint* dapat digunakan untuk menyaring data mobil dari *database* dengan rentang harga dari 0 (nol) sampai dengan "max_value".

Gambar berikut menunjukkan *request* ke *endpoint* dan *response* yang diberikan apabila prediksi berhasil dilakukan.

```
{
    "data": {
        "Customer Name": "Foo Bar",
        "Customer e-mail": "foobar@baz.com",
        "Country": "Indonesia",
        "Gender": 1,
        "Age": 24,
        "Annual Salary": 120000000,
        "Credit Card Debt": 12000000,
        "Net Worth": 300000000
    }
}
```

Gambar 8. Request ke endpoint dengan data lengkap

```
{
    "Result": {
        "pred_value": 329871260.17534614,
        "min_value": 0,
        "max_value": 332156125.17534614
    },
    "Status": "Successed"
}
```

Gambar 9. Response prediksi sukses dari endpoint

Pada Gambar 8, dilakukan request dengan data yang menyertakan seluruh fitur yang diperlukan oleh model prediksi, yaitu "*Customer Name*, *Customer e-mail*, *Country*, *Gender*, *Age*, *Annual Salary*, *Credit Card Debt*, dan *Net Worth*". Gambar 9 menunjukkan *endpoint* mengembalikan *response* dalam format JSON yang berisi hasil prediksi dan status bahwa prediksi berhasil dilakukan. Nilai aktual hasil prediksi disimpan pada "pred_value" dengan nilai 329.871.260,17. Nilai "min_value" berupa "0", bukan sebagai selisih nilai aktual prediksi (329.871.260,17) dengan nilai MAE (2.284.865,29). Sementara itu, nilai "max_value" merupakan hasil penjumlahan nilai aktual prediksi dengan nilai

JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 5, Oktober 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i5.5021
Hal 1669−1678
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

MAE, yaitu 332.156.125,17.

Pada gambar berikut, dilakukan *request* ke *endpoint* dan *response* yang diberikan apabila prediksi gagal dilakukan.

```
{
    "data": {
        "Customer Name": "Foo Bar",
        "Customer e-mail": "foobar@baz.com",
        "Country": "Indonesia",
        "Gender": 1,
        "Age": 24,
        "Net Worth": 300000000
    }
}
```

Gambar 10. *Request* ke *endpoint* dengan data tak lengkap

```
{
    "Result": "X has 3 features, but Lasso is expecting 5 features as input.",
    "Status": "Failed"
}
```

Gambar 11. *Response* prediksi gagal dari *endpoint*

Pada Gambar 10, data *request* tidak menyertakan fitur "*Annual Salary*" dan "*Credit Card Debt*". *Response* "*Status*" yang diberikan oleh *endpoint* pada Gambar 11 menunjukkan bahwa prediksi gagal dilakukan. Kegagalan prediksi ini disebabkan jumlah fitur yang diberikan pada model tidak lengkap sebagaimana tertera pada *response* "*Result*".

4. KESIMPULAN

Mayoritas model *machine learning* yang dibangun pada tahap *Modeling* memiliki performa tinggi, salah satunya model regresi Lasso. Model regresi Lasso memiliki performa R^2 sebesar 0,99958 dan MAE sebesar 2.284.865,29. Model tersebut disebarkan (*deploy*) sebagai sebuah *web service endpoint* melalui *platform* Azure Machine Learning. *Endpoint* dapat diakses secara waktu nyata (*real-time*) dan memerlukan data dari fitur-fitur prediktor untuk dapat mengembalikan *response* berupa hasil prediksi. Apabila *request* yang dikirim tidak menyertakan fitur prediktor secara lengkap, maka *endpoint* mengembalikan *response* yang menyatakan prediksi gagal dilakukan. Pada penelitian dan pengembangan lebih lanjut dapat menggunakan data dalam jumlah yang lebih besar dan telah memiliki nilai mata uang Rupiah tanpa perlu melakukan konversi nilai mata uang. Selain itu, model dengan performa lebih tinggi masih potensial untuk dikembangkan. Implementasi *endpoint* pada aplikasi yang digunakan publik juga dapat dilakukan sehingga meningkatkan nilai manfaat praktis dari penelitian ini.

REFERENCES

[1] B. Kriswantara, Kurniawati, and H. F. Pardede, "Prediksi Harga Mobil Bekas dengan Machine Learning," *Syntax Lit. J. Ilm. Indones.*, vol. 6, no. 5, pp. 2100–2110, 2021, doi: http://dx.doi.org/10.36418/syntax-literate.v6i5.2716.

[2] M. A. Aditya, R. D. Mulyana, I. P. Eka, and S. R. Widianto, "Penggabungan Teknologi Untuk Analisa Data Berbasis Data Science," in *Seminar Nasional Teknologi Komputer & Sains (SAINTEKS)*, 2020, pp. 51–56. [Online]. Available: https://prosiding.seminar-id.com/index.php/sainteks

[3] A. Pandey, V. Rastogi, and S. Singh, "Car's Selling Price Prediction using Random Forest Machine Learning Algorithm," *SSRN Electron. J.*, 2020, doi: 10.2139/ssrn.3702236.

[4] E. Dewi, S. Mulyani, F. Mulady, D. Ramadhan, A. Ariyantono, and D. Ramdani, "Estimasi Harga Jual Mobil Bekas Menggunakan Metode Regresi Linier Berganda," *e-Jurnal JUSITI (Jurnal Sist. Inf. dan Teknol. Informasi)*, vol. 9, no. 1, pp. 1–8, 2020, doi: 10.36774/jusiti.v9i1.649.

[5] I. Purnamasari, "Klasifikasi Pelanggan Produk IndiHome Menggunakan Naive Bayes Classifier Dengan Seleksi Fitur Algoritma Genetik," *J. Tek. Komput.*, vol. 4, no. 1, pp. 8–16, 2018.

[6] F. Naufal, Y. H. Chrisnanto, and A. K. Ningsih, "Sistem Rekomendasi Penawaran Produk Pada Online Shop Menggunakan K-Means Clustering," *Informatics Digit. Expert* -, vol. 1, pp. 10–17, 2022.

[7] A. I. Putra and R. R. Santika, "Implementasi Machine Learning dalam Penentuan Rekomendasi Musik dengan Metode Content-Based Filtering," *Edumatic J. Pendidik. Inform.*, vol. 4, no. 1, pp. 121–130, 2020, doi: 10.29408/edumatic.v4i1.2162.

[8] P. P. Yanti, "A Survey : Application of Big Data in the Travel and Tourism Industry," *ITEJ (Information Technol. Eng. Journals)*, vol. 5, no. 1, pp. 1–13, 2020, doi: 10.24235/itej.v5i1.38.

[9] Y. Khandelwal, "Car Purchase Price (beginner dataset)," 2020. https://www.kaggle.com/datasets/yashk07/car-purchase-price-beginner-dataset (accessed Jan. 04, 2022).

[10] A. P. Fadillah, "Penerapan Metode CRISP-DM untuk Prediksi Kelulusan Studi Mahasiswa Menempuh Mata Kuliah (Studi Kasus Universitas XYZ)," *J. Tek. Inform. dan Sist. Inf.*, vol. 1, no. 3, pp. 260–270, 2015, doi: 10.28932/jutisi.v1i3.406.

[11] M. Radhi, A. Amalia, D. R. H. Sitompul, S. H. Sinurat, and E. Indra, "Analisis Big Data Dengan Metode Exploratory Data Analysis (EDA) dan Metode Visualisasi Menggunakan Jupyter Notebook," *J. Sist. Inf. dan Ilmu Komput. Prima(JUSIKOM PRIMA)*, vol. 4, no. 2, pp. 23–27, 2022, doi: 10.34012/jurnalsisteminformasidanilmukomputer.v4i2.2475.

[12] Nuryadi, T. D. Astuti, E. S. Utami, and M. Budiantara, *Dasar-Dasar Statistik Penelitian*, 1st ed. Yogyakarta: SIBUKU MEDIA, 2017.

[13] M. Nurudin, M. N. Mara, and D. Kusnandar, "Ukuran Sampel dan Distribusi Sampling Dari Beberapa Variabel Random

Kontinu," *Bul. Ilm. Mat. Stat. dan Ter.*, vol. 03, no. 1, pp. 1–6, 2014, doi: 10.26418/bbimst.v3i01.4461.
[14] A. Sen, "Statistics — quick reference," 2021. https://medium.com/analytics-vidhya/statistics-quick-reference-4cad05eebd45 (accessed Jan. 09, 2022).
[15] A. Kreiley, "Spatial and Temporal Variability of the Saline Intrusion in the Lower Charles River," no. August, 2020, doi: 10.13140/RG.2.2.27771.34087.
[16] F. I. Kurniadi, D. Satyananda, E. Santika, and P. D. Larasati, "Multi-output Regression untuk memprediksi Luas Wilayah, Kualitas Padi dan Produksi Padi pada Pulau Jawa," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 5, no. 2, pp. 18–23, 2022, doi: 10.47970/siskom-kb.v5i2.269.
[17] J. Pendidikan and D. Konseling, "Clustering Menggunakan Algoritma K-Means Pada Penyakit ISPA di Puskesmas Kabupaten Karawang," *J. Pendidik. dan Konseling*, vol. 4, no. 5, pp. 462–473, 2022, doi: 10.31004/jpdk.v4i5.6632.
[18] F. Muhammad, N. M. Maghfur, and A. Voutama, "Sentiment Analysis Dataset on COVID-19 Variant News," *SYSTEMATICS*, vol. 4, no. 1, pp. 382–391, 2022.
[19] A. Gholamy, V. Kreinovich, and O. Kosheleva, "Why 70/30 or 80/20 Relation Between Training and Testing Sets : A Pedagogical Explanation," 2018. [Online]. Available: https://scholarworks.utep.edu/cs_techrep/1209/
[20] E. Zuccarelli, "Performance Metrics in Machine Learning — Part 2: Regression," 2021. https://towardsdatascience.com/performance-metrics-in-machine-learning-part-2-regression-c60608f3ef6a (accessed Jan. 04, 2022).
[21] A. Yoga Pratama *et al.*, "Analisis Sentimen Media Sosial Twitter Dengan Algoritma K-Nearest Neighbor Dan Seleksi Fitur Chi-Square (Kasus Omnibus Law Cipta Kerja)," *J. Sains Komput. Inform. (J-SAKTI*, vol. 5, no. 2, pp. 897–910, 2021.
[22] H. Ariesta and M. A. Kartawidjaja, "Feature Selection pada Azure Machine Learning untuk Prediksi Calon Mahasiswa Berprestasi," *TESLA J. Tek. Elektro*, vol. 20, no. 2, pp. 186–195, 2018.