

OPTIMISASI METODE PREDIKSI PADA DATA CUSTOMER ATTACK LIST (STUDI KASUS: PT AKASTRA TOYOTA)

Delly Asep Gustira¹, Ardiansyah Dores²
Fakultas Ilmu Komputer, Universitas Mercu Buana

ardian@mercubuana.ac.id

Abstrak

Kunjungan pelanggan di bengkel resmi TOYOTA untuk servis terdata oleh sistem aplikasi. Pengelolaan data pelanggan /data *attacklist* diperlukan oleh tim *general repair*, untuk melakukan tindak lanjut berupa konfirmasi pengingat-penjadwalan servis selanjutnya. Proses ini sudah berjalan selama kurun waktu 2 tahun, namun hasil penjadwalan kedatangan kunjungan servis pelanggan belum sesuai dari total data pelanggan (*attacklist*) yang dikonfirmasi. Penelitian ini dilakukan untuk memprediksi variabel rentang waktu servis, menggunakan metode eksplorasi dan analisa data (*EDA*), membuat model algoritma *Linear Regression*, *Random Forest Regressor* dan *Gradient Boosting Regressor* yang digunakan untuk prediksi. Dengan menggunakan metodologi *Cross-Industry Standard Process for Data Mining (CRISP-DM)*. Nilai prediksi yang dihasilkan oleh setiap model algoritma, untuk variabel *rentang_hari* (waktu servis selanjutnya) memiliki akurasi yang mendekati nilai aktual/sebenarnya. Sehingga penelitian ini merekomendasikan penggunaan model algoritma, yang dapat diterapkan pada sistem aplikasi yang ada. Sehingga optimasi data pelanggan/*attacklist* sinergis dengan proses bisnis yang berjalan.

Kata Kunci : *data mining, forecast, regression, random forest, gradient boosting, CRISP-DM.*

Abstract

Customer visits to the official TOYOTA workshop for servicing are recorded in a system. The service team uses customer data to schedule the next service. Even though this has been done for 2 years, the actual/real value number of customers coming in for service is not as expected. This study aims to predict how long customers wait between services using data analysis and machine learning models like linear regression, random forest, and gradient boosting. We found that these models can accurately predict when customers should need their next service. Therefore, we recommend using these models in the system to improve customer data management and align it better with the overall business process.

Keywords : *data mining, forecast, regression, random forest, gradient boosting, CRISP-DM.*

1. Pendahuluan

Pengelolaan data pelanggan yang dilakukan oleh PT.AKASTRA TOYOTA, dilakukan dengan tujuan agar Tim General Repair dapat melakukan tindak lanjut berupa konfirmasi pengingat, dan penjadwalan untuk servis selanjutnya. Berdasarkan informasi dari riwayat servis terakhir yang telah dilakukan oleh pelanggan tersebut, dimana selama kurun waktu 2 tahun proses ini tidak berjalan efektif dalam mendukung proses bisnis di bengkel

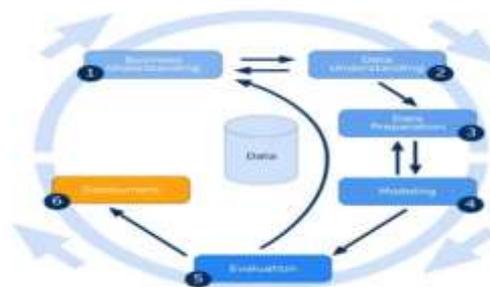
resmi TOYOTA. Dari total list jumlah pelanggan yang di hubungi (baca: Maintain) oleh Tim *General Repair*, dimana hal tersebut merupakan upaya pengingat /reminder dan penjadwalan kedatangan servis / attendance, tidak sesuai dengan prediksi yang diharapkan. Kondisi tersebut mengindikasikan perlunya analisis data yang efektif dan relevan dalam mendukung sistem pengambilan keputusan yang strategis dalam hal pemasaran, pada jurnal penelitian yang dibuat oleh Kumar organisasi perlu merencanakan dan mengeksekusi strategi pemasaran yang efisien

untuk meningkatkan kinerja dan daya saing mereka, pendekatan yang inovatif seperti penggunaan algoritma untuk meningkatkan akurasi dan efektivitas dalam pengambilan keputusan (P., 2020). Maka dari itu, perlu dikembangkannya analisis strategi prediksi yang baru yang ditampilkan dalam bentuk visualisasi data, dengan perkembangan teknologi informasi yang pesat, visualisasi data menjadi salah satu metode yang efektif dalam menganalisis dan memahami data secara menyeluruh. Visualisasi data adalah representasi grafis dari informasi dan data, dengan menggunakan elemen visual seperti bagan, grafik, dan peta, alat visualisasi data menyediakan cara yang mudah diakses untuk melihat dan memahami tren, pencila, dan pola dalam data (Tableau, 2022).

Beberapa penelitian terkait mengenai strategi *marketing* juga menjadi perhatian, akan tetapi sulit untuk diterapkan karena memiliki karakteristik data yang berbeda dan menggunakan metode konvensional, seperti *AHP* (Azza & Dores, 2018), *Naive Bayes* (Herdi & Dores, 2021), *Fuzzy Tsukamoto* (Meilina et al., 2017), dan Algoritma Apriori (Melaka, 2021). Pada penelitian kali ini, pembahasan mengenai metode-metode prediksi, dengan memperhatikan karakteristik data, menjadi perhatian khusus untuk mendapatkan hasil yang optimal, beberapa metode yang telah digunakan, yaitu *Linear Regression* (Praveen et al., 2022) (Moumen et al., 2023) (Raizada & Saini, 2021) (Muhammad saad bin ilyas et al., 2023), *Random Forest Regressor* (Praveen et al., 2022) (Moumen et al., 2023) (Raizada & Saini, 2021), *Gradient Boosting Regressor* (Moumen et al., 2023). Hingga pada akhirnya, metode evaluasi yang digunakan pun menyesuaikan dengan metode prediksi yang akan digunakan, seperti yang telah dilakukan pada penelitian-penelitian terkait menggunakan *MSE* (Raizada & Saini, 2021) (Muhammad saad bin ilyas et al., 2023), *RMSE* (Moumen et al., 2023), *MAE* (Moumen et al., 2023) (Raizada & Saini, 2021) (Muhammad saad bin ilyas et al., 2023), *R2* (Moumen et al., 2023).

2. Metodologi

Pendekatan penelitian ini menggunakan metodologi *Cross-Industry Standard Process for Data Mining (CRISP-DM)* sebagai kerangka kerja penelitian. Metode *CRISP-DM* ini mencakup enam fase: pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi dan penerapan (Rotty, 2023).



Gambar. 1. Siklus *CRISP-DM*

Pemahaman Bisnis (*Business Understanding*) Tahapan ini secara garis besar telah disampaikan pada bagian sebelumnya, ada beberapa hal yang perlu dipastikan pada tahapan ini, yaitu pemangku kepentingan/pihak yang terlibat dalam pengambilan keputusan (stakeholder) yaitu tim yang melakukan konfirmasi pengingat (reminder) kepada pelanggan dan pimpinan departemen pelayanan bengkel (*head of general repair*), daftar pelanggan yang dikonfirmasi tidak sesuai dengan kunjungan servis yang diperkirakan. Kondisi tersebut bisa dianalisa dari bentuk pelaporan dari data yang digunakan dalam proses operasional, pelaporan ini menghasilkan informasi dan rekomendasi kepada pemangku kepentingan/pengambil Keputusan untuk dilakukan tindakan lebih lanjut.

Pemahaman Data (*Data Understanding*)

Pada fase ini pengumpulan data pelanggan potensial (*Attack List*) didapat dari sistem aplikasi (MRSGENIUS) yang periodic update ke staging server source data Toyota Astra Motor DMS dan memanfaatkan fitur *broadcast whatsapp message* dari vendor.



Gambar. 2. Arsitektur aplikasi

Persiapan Data (Data Preparation)

Data dalam format (.CSV) dieksplorasi dan dianalisis menggunakan metode *EDA* (*Explanatory Data Analysis*), yang dilakukan pada *google colaboratory* dan menjadi dataset yang digunakan, dataset ini terdiri dari kolom *Vehicle Id Number (VIN)*, *JobDate*, *JobType* dan *JobCategory*.

```
1 # import dataset "dataset-servis.csv" menjadi data frame
2
3 df = pd.read_csv(path_data+"dataset-servis.csv", sep=',')
```

Gambar. 3. Proses *Import* datasets menggunakan pustaka *pandas*

```
1 # menampilkan 5 baris teratas data frame
2 df.head()
```

	VIN	JobDate	JobType	CategoryName
0	MHFG8GS2K0896470	03/10/2023 13:03	GR	Service GR
1	MHKA4DB3JGJ054415	03/10/2023 10:46	GR	Service GR
2	MHKM5EA2JLK080462	03/10/2023 10:58	SBE	Service GR
3	MHFJW8EM1G2303370	04/10/2023 09:24	SBE	Service GR
4	MR2BF3HK1L4009617	03/10/2023 07:48	SBE	Service GR

Gambar. 4. *Dataframe* (*EDA*)

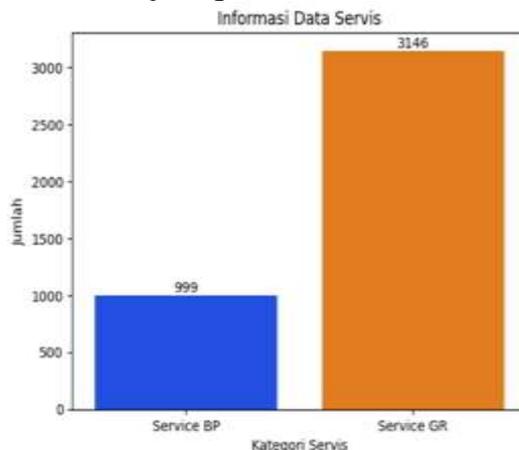
```
1 # cek informasi dari data set "df"
2 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4145 entries, 0 to 4144
Data columns (total 4 columns):
#   Column      Non-Null Count  Dtype
---  ---
0   VIN         4145 non-null   object
1   JobDate     4145 non-null   object
2   JobType     4145 non-null   object
3   CategoryName 4145 non-null   object
dtypes: object(4)
memory usage: 129.7+ KB
```

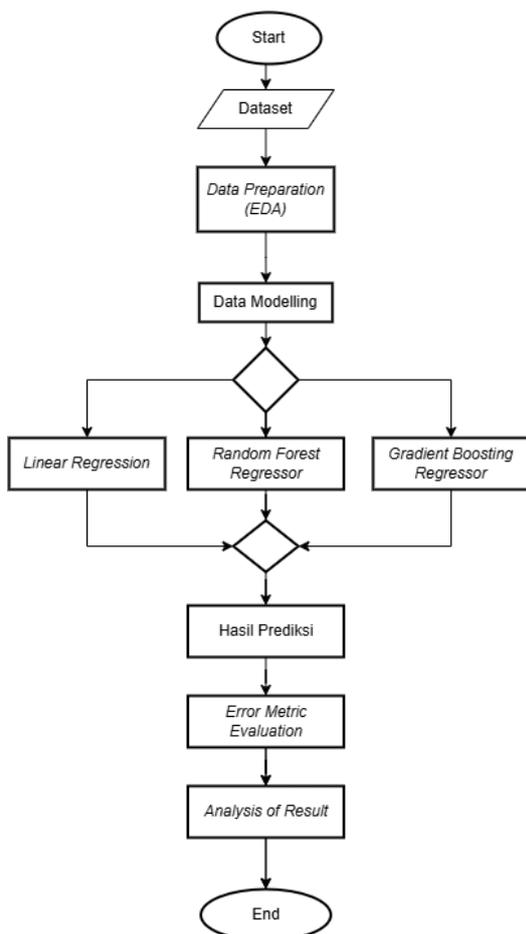
Gambar. 5. Bentuk *Dataframe* (*EDA*)

Proses *EDA* yang dilakukan mulai dari import dataset, terlihat pada gambar 3, selanjutnya terjadi proses pengecekan terhadap nilai yang

kosong pada tiap kolom, pengecekan duplikasi data, informasi umum *dataframe* (lihat gambar 4). Kemudian akan terlihat bentuk umum *dataframe* (lihat gambar 5), mengkonversi tipe data tanggal/waktu, memodifikasi dan transformasi kolom, visualisasi data dan konversi tipe data kategorikal/ordinal menjadi tipe data numerik (Priambodo et al., 2022), hasil dari proses *EDA* terlihat pada gambar 6.



Gambar. 6. Visualisasi Data (*EDA*)



Gambar. 7. Alur Proses Pemodelan Algoritma

3. Hasil dan Pembahasan

Setelah dataset/dataframe di eksplorasi melalui tahap data preparation maka tahap selanjutnya yaitu melakukan pemodelan prediktif dengan menggunakan dataset tersebut.

Pemodelan (Modelling)

Dalam hal ini ingin memprediksi berapa lama waktu yang dibutuhkan mobil untuk melakukan kebutuhan servis berikutnya (rentang_hari). Informasi lainnya yang dapat digunakan seperti jenis servis yang sebelumnya dilakukan (Service GR, Service BP), berapa lama sejak servis terakhir (rentang_bulan), dan kebiasaan servis dari pemilik (kebiasaan). Informasi-informasi ini adalah variabel prediktor (fitur) yang dimasukkan kedalam model untuk memprediksi variabel target (waktu hingga servis berikutnya).

Pada pemodelan data dibagi menjadi dua bagian yaitu untuk melatih model dan menguji

kinerja model tersebut, variabel X_{train} (variabel prediktor) yang digunakan untuk melatih model dan variabel X_{test} yang digunakan untuk mengevaluasi kinerja model yang sudah dilatih. Sedangkan variabel y_{train} berisi target (yang ingin diprediksi) yang sesuai dengan data latih dan y_{test} berisi target yang sesuai dengan data uji, yang digunakan untuk membandingkan dengan prediksi model. Menggunakan fungsi $train_test_split(x,y)$ untuk melakukan pemisahan variabel X dan Y.

Pengaturan parameter $test_size=0.2$ berarti 20% data akan digunakan untuk pengujian dan 80% sisanya untuk pelatihan. Sedangkan parameter $random_state=42$ memastikan bahwa data dibagi dengan cara yang sama. Setiap kali kode dijalankan untuk direproduksi maka akan mendapatkan hasil yang sama meskipun menjalankan kode beberapa kali, namun jika tidak ditentukan parameter $random_state$, maka pemisahan akan berbeda untuk setiap kali kode dijalankan yang berpotensi menyebabkan kinerja model bervariasi atau tidak konsisten (Dzulhijjah et al., 2023) (Maulana & Napitupulu, 2022), seperti pada gambar 8.

```

1 # 1. Memanggil Data
2 # 2. Definisikan variabel target dan prediktor
3 y = df['rentang_hari']
4 x = df[['Service BP', 'Service GR', 'rentang_bulan', 'kebiasaan']] # sesuai dengan nama prediktor
5 # 3. Bagi data menjadi data latih dan data uji
6 X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)
    
```

Gambar. 8. Definisi variabel x,y dan pembagian data latih dan uji.

Linear Regression

Regresi Linear adalah metode statistik yang digunakan untuk memodelkan hubungan antara satu variabel dependen (target) dan satu atau lebih variabel independen (prediktor). Tujuannya adalah untuk menemukan persamaan garis lurus yang paling sesuai dengan data, yang kemudian dapat digunakan untuk membuat prediksi.

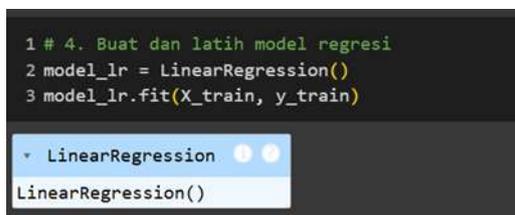
Membuat objek model Regresi Linear dan menetakannya ke variabel model_lr dengan menggunakan fungsi $LinearRegression()$ dari pustaka $sklearn.linear_model$ yang menyediakan alat untuk membangun model linear. Pelatihan dan pembelajaran untuk

model dari data menggunakan metode fit dari objek `model_lr` dengan memberikan dua informasi yaitu `x_train` (variabel prediktor/independen) dari dataset yang digunakan untuk pelatihan, dan `y_train` (variabel target) yang sesuai atau ingin diprediksi oleh model seperti pada gambar 9.

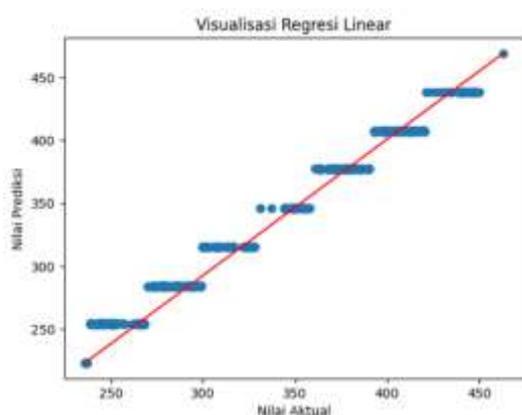
```

1 # 4. Buat dan latih model regresi
2 model_lr = LinearRegression()
3 model_lr.fit(X_train, y_train)

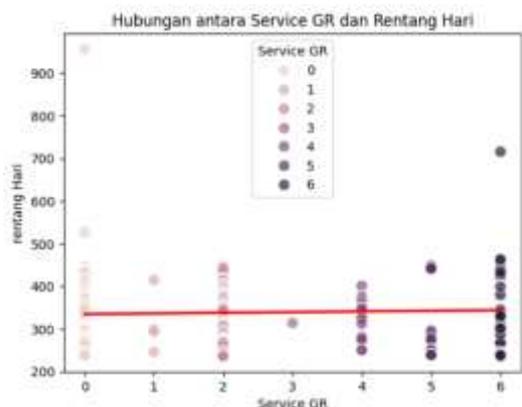
```



Gambar 9. Membuat model *Linear Regression*



Gambar 10. Visualisasi Regresi Linear.



Gambar 11. Visualisasi regresi service GR dan rentang hari



Gambar 12. Visualisasi regresi rentang bulan dan rentang hari.

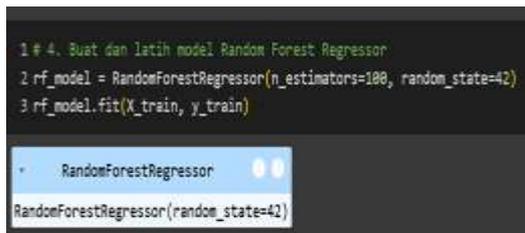
Random Forest Regressor

Model pembelajaran mesin yang digunakan untuk memprediksi nilai numerik yang kontinu, seperti harga rumah, suhu udara, atau jumlah penjualan. Digunakan untuk data yang memiliki hubungan *non-linear* atau interaksi kompleks antar variabel, model *random forest regressor* menggunakan banyak pohon keputusan untuk membuat prediksi. Setiap pohon keputusan dilatih menggunakan subset data dan fitur yang berbeda, prediksi akhir diambil dari rata-rata (atau nilai terbanyak) prediksi dari semua pohon keputusan.

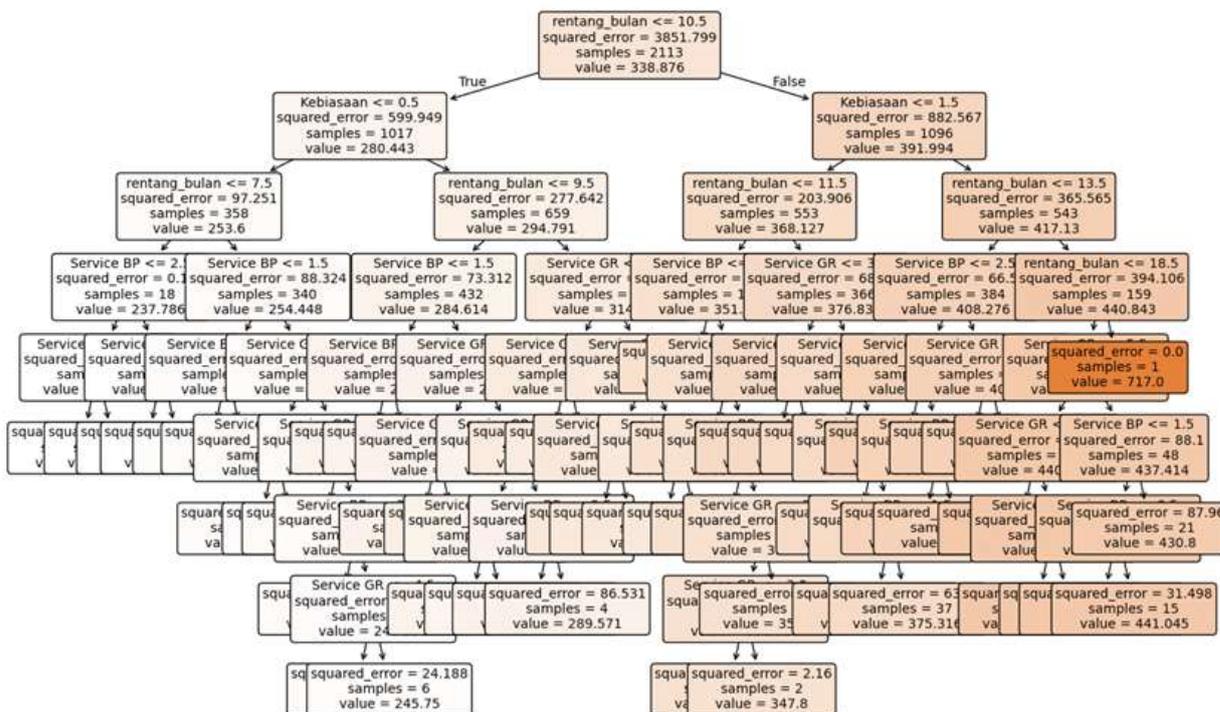
Membuat variabel `rf_model` agar mudah untuk dipanggil nanti dan `RandomForestRegressor(...)` adalah fungsi yang digunakan untuk membuat model 'Random Forest Regressor'. Parameter `n_estimator=100` menentukan bahwa model ini menggunakan 100 pohon keputusan, semakin banyak pohon semakin kompleks dan akurat model, tetapi juga membutuhkan waktu lebih lama untuk diproses. Parameter `random_state=42` memberi nomor acak pada model agar hasilnya tetap sama setiap kali dijalankan. Setelah model dibuat maka perlu untuk melatih dengan belajar dari data, melalui perintah `rf_model.fit(...)`. Data masukan berisi informasi yang akan digunakan model untuk belajar adalah variabel `X_train`, dan variabel `y_train` adalah data keluaran yang berisi hasil yang ingin di prediksi seperti pada gambar 13.

```

1 # 4. Buat dan latih model Random Forest Regressor
2 rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
3 rf_model.fit(X_train, y_train)
    
```



Gambar. 13. Membuat model *RandomForestRegressor*



Gambar. 14. Visualisasi individual *decision tree* dalam *Random Forest Regressor*

Gradient Boosting Regressor

Gradient boosting membangun model secara bertahap dengan menambahkan *weak learners* (biasanya *decision trees*) secara berurutan. Setiap *weak learner* baru difokuskan untuk memperbaiki kesalahan yang dibuat oleh *weak learner* sebelumnya. Proses ini terus berlanjut hingga mencapai jumlah *weak learners* yang ditentukan atau kinerja model tidak lagi meningkat secara signifikan.

Menggunakan fungsi dari modul *sklearn.ensemble* yang membuat objek model *GradientBoostingRegressor(...)*, model prediksi disimpan pada variabel *gb_model* dengan metode *Gradient Boosting*. Parameter pengaturan pada model ini menggunakan 100 tahapan *boosting* (*weak learners*) dan random state yang tetap untuk konsistensi.

Kemudian model ini dilatih menggunakan variabel *X_train* (fitur) dan variabel *y_train*

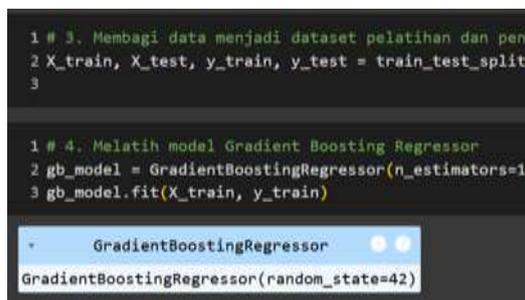
(nilai aktual yang akan diprediksi) sehingga model dapat belajar bagaimana membuat prediksi pada data baru yang belum pernah dilihat sebelumnya. Seperti pada gambar 15.

```

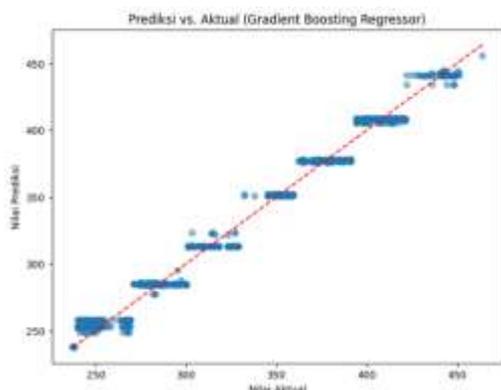
1 # 3. Membagi data menjadi dataset pelatihan dan peng
2 X_train, X_test, y_train, y_test = train_test_split(
3
    
```

```

1 # 4. Melatih model Gradient Boosting Regressor
2 gb_model = GradientBoostingRegressor(n_estimators=100,
3 gb_model.fit(X_train, y_train)
    
```



Gambar. 15. Membuat model *GradientBoostingRegressor*



Gambar. 16. Visualisasi dalam *GradientBoostingRegressor*.

meningkatkan efisiensi dan efektivitas dalam hal optimasi pengelolaan (data *attacklist*). Memberikan wawasan bagi departemen *General Repair (MRA)* perihal penanganan bisnis yang dijalankan.

Hasil Prediksi

Hasil prediksi dari setiap model algoritma yang digunakan terhadap variabel y_{train} (*rentang_hari*), dimana nilai yang dihasilkan oleh algoritma *Random Forest Regressor* dan *Gradient Boosting Regressor* memiliki akurasi yang mendekati nilai aktual variabel y_{train} . Lihat pada tabel 1.

Evaluasi (Evaluation)

Tahap evaluasi merupakan proses pengukuran tingkat presisi dan keakuratan model yang digunakan dengan analisis matrik kesalahan (*error metric*), seperti *Mean Absolute Error (MAE)*, *Mean Squared Error (MSE)*, *Root Mean Square Error (RMSE)* dan *R-squared (R2)* untuk menilai akurasi prediksi (Tarek et al., 2023) lihat pada tabel 2 .

Penerapan (Deployment)

Ini adalah fase terakhir dalam metode *CRISP-DM* dimana model yang telah dibangun dan dievaluasi selanjutnya dapat diterapkan di lingkungan operasional, yaitu dengan mengimplementasikan model kedalam sistem aplikasi internal data pelanggan potensial (*attack list*). Misalnya dengan mengembangkan *Application Programming Interface (API)* untuk akses model (Dzulhijjah et al., 2023).

4. Kesimpulan

Melalui penelitian ini merekomendasikan dari beberapa model algoritma machine learning yang digunakan, khususnya *Random Forest Regressor* dan *Gradient Boosting Regressor* mampu melakukan prediksi rentang waktu servis, secara signifikan sehingga mampu

Tabel. 1. Hasil prediksi model algoritma.

VIN	rentang_hari	LR	RF	GB
MHRRV3870NJ209191	240	254.78293761296322	248.52702532136686	249.5712061515334
MHFXW40G0E4505669	303	315.86187002754986	323.13192726391264	317.3357802042266
MHFZR69G3E3083885	240	254.78293761296322	248.52702532136686	249.5712061515334
MHFAB1BY9N0023893	240	255.14669773607812	257.70182556961555	256.66202482148714
JTNGF3DH4J801551	238	223.68914683882412	237.6589603174603	236.41145481084948
MHKAB1BA3PJ07383	238	223.68914683882412	237.6589603174603	236.41145481084948
MHFZ28H38H0028157	238	223.68914683882412	237.6589603174603	236.41145481084948
MHHHS1604HK959388	237	223.68914683882412	237.6589603174603	236.41145481084948
MHKAA1BA0MJ0O4035	237	223.85798779123544	237.1874801587301	236.45195416213005
MHFGB8GS8G0822784	240	408.4113924485366	408.70645622310326	408.429663405245

Tabel. 2. Matrik Kesalahan (Error Metric)

Model	MAE	MSE	RMSE	R ²
Linear Regression	7.259634667708646	71.25202360600109	8.441091375290348	0.9823012803676093
Random Forest Regressor	6.808287875437167	65.65752710178688	8.102933240610271	0.983690931076479
Gradient Boosting Regressor	6.788229986465548	65.30639851931119	8.081237437379945	0.983778150021593

5. Daftar Pustaka

- Azza, G. M., & Dores, A. (2018). Sistem Informasi Manajemen Marketing Tools Serta Penerapan Metode Ahp (Analytical Hierarchy Process) Pada Proses Uji Kualitas Barang (Studi Kasus : PT Edi Indonesia). *Jurnal Cendikia, XVI*, 107–114.
- Dzulhijjah, D. A., Herlambang, M. B., Haifan, M., Studi, P., Insinyur, P., & Selatan, T. (2023). Implementasi Framework CRISP-DM untuk Proses Data Mining Aplikasi Credit Scoring PT.XYZ. *Prosiding Seminar Nasional Teknologi Informasi Dan Bisnis, I(1)*, 238–251.
- Herdi, T., & Dores, A. (2021). Bayes interpretation for smoke-free area cities index. *Bulletin of Social Informatics Theory and Application, 5(1)*, 38–46. <https://doi.org/10.31763/businta.v5i1.286>
- Maulana, A., & Napitupulu, T. A. (2022). Lost Won Opportunity Prediction in Sales Pipeline B2B Crm Using Machine Learning. *Journal of Theoretical and Applied Information Technology, 100(10)*, 3486–3496.
- Meilina, P., Rosanti, N., & Astryani, N. (2017). Sistem Pendukung Keputusan Penentuan Jumlah Produksi Barang Dengan Metode Fuzzy Tsukamoto Berbasis Android. *Jurnal Seminar Nasional Sains Dan Teknologi, 4(7)*, 1–11. <https://jurnal.umj.ac.id/index.php/semnastek/article/view/2073>
- Melaka, C. (2021). *IC E C o n ~ 2 0 2 1 CONFERENCE ON EMERGING COMPUTATIONAL. i.*
- Moumen, I., Abouchabaka, J., & Rafalia, N. (2023). Adaptive traffic lights based on traffic flow prediction using machine learning models. *International Journal of Electrical and Computer Engineering, 13(5)*, 5813–5823. <https://doi.org/10.11591/ijece.v13i5.pp5813-5823>
- Muhammad saad bin ilyas, Atif Ikram, Muhammad Aadil Butt, & Iqra Tariq. (2023). Comparative Analysis of Regression Algorithms used to Predict the Sales of Big Marts. *Journal of Innovative Computing and Emerging Technologies, 3(1)*. <https://doi.org/10.56536/jicet.v3i1.53>
- P., D. K. (2020). Analysis of Neural Network Based Language Modeling. *Journal of Artificial Intelligence and Capsule Networks, 2(1)*, 53–63. <https://doi.org/10.36548/jaicn.2020.3.006>
- Praveen, R., Kumar, D. P., Sam, A. P., & Sundari, G. S. (2022). Sales prediction using machine learning. *AIP Conference Proceedings, 2444(June)*, 3619–3625. <https://doi.org/10.1063/5.0078390>
- Priambodo, B., Jumaryadi, Y., Rahayu, S., Ani, N., Ratnasari, A., Salamah, U., Putra, Z. P., & Otong, M. (2022). Predicting Employee Turnover in IT Industries using Correlation and Chi-Square Visualization. *International Journal of Advanced Computer Science and Applications, 13(12)*, 71–75. <https://doi.org/10.14569/IJACSA.2022.0131210>
- Raizada, S., & Saini, J. R. (2021). Comparative Analysis of Supervised Machine Learning Techniques for Sales Forecasting. *International Journal of Advanced Computer Science and Applications, 12(11)*, 102–110. <https://doi.org/10.14569/IJACSA.2021.0121112>
- Rotty, A. K. (2023). *Cross-Industry Standard Process for Data Mining (CRISP-DM) Approach in Determining the Most Significant Employee Engagement Drivers to Sales at X Car Dealership.* 3368–3379. <https://doi.org/10.46254/ap03.20220552>
- Tableau. (2022). *What Is Data Visualization? Definition, Examples, And Learning Resources.* Tableau Software, LLC, A Salesforce Company. <https://www.tableau.com/learn/articles/data-visualization>
- Tarek, Z., Shams, M. Y., Elshewey, A. M., El-Kenawy, E. S. M., Ibrahim, A., Abdelhamid, A. A., & El-Dosuky, M. A. (2023). Wind Power Prediction Based on Machine Learning and Deep Learning Models. *Computers, Materials and Continua, 74(1)*, 715–732. <https://doi.org/10.32604/cmc.2023.032533>

