Perbandingan Prediksi Jumlah Transaksi Ojek *Online* Menggunakan Regresi Linier dan *Random Forest*

Dian Pramesti[1], Wiga Maulana Baihaqi[2]
[1,2]Teknologi Informasi, Fakultas Ilmu Komputer, Universitas Amikom Purwokerto
E-mail: [1]dianprmsti@gmail.com, [2]wiga@amikompurwokerto.ac.id
Coresponden Author: *dianprmsti@gmail.com*

Diterima Redaksi: 08 Juli 2023 Revisi Akhir: 16 Oktober 2023 Diterbitkan Online: 24 Oktober 2023

Abstrak *- Dalam era kemajuan teknologi saat ini, peralihan masyarakat dari ojek tradisional ke ojek online telah menciptakan perubahan signifikan dalam industri transportasi. Penelitian ini bertujuan untuk memprediksi tarif ojek online dengan mempertimbangkan berbagai feature yang relevan menggunakan dua metode, yaitu regresi linear dan Random Forest. Dataset yang digunakan adalah data ojek online Pontianak, Indonesia yang di peroleh dari Kaggle. Hasil analisis menunjukkan bahwa regresi linear memiliki kinerja yang lebih baik dalam hal prediksi tarif ojek online, dengan nilai Root Mean Square Error (RMSE) dan Mean Squared Error (MSE) yang lebih rendah, serta nilai Mean Absolute Percentage Error (MAPE) yang lebih rendah dibandingkan dengan model Random Forest. Meskipun hasil ini menunjukkan potensi model regresi linear untuk memberikan prediksi yang lebih akurat, penelitian selanjutnya diharapkan melakukan analisis yang mendalam guna mencapai perbaikan yang lebih signifikan dalam nilai RMSE, MSE, dan MAPE serta untuk memahami feature yang memengaruhi tarif ojek online dengan lebih baik.*

Kata Kunci *- Ojek Online, Pontianak, Prediksi, Random Forest, Regresi Linier*

Abstract - *In the current era of technological advances, the transition of society from traditional ojek to online ojek has created significant changes in the transportation industry. This research aims to predict online ojek rates by considering various relevant features using two methods, namely linear regression and Random Forest. The dataset used is online ojek data Pontianak, Indonesia obtained from Kaggle. The analysis results show that linear regression has better performance in terms of online ojek tariff prediction, with lower Root Mean Square Error (RMSE) and Mean Squared Error (MSE) values, as well as lower Mean Absolute Percentage Error (MAPE) values compared to the Random Forest model. Although these results show the potential of linear regression models to provide more accurate predictions, future research is expected to conduct in-depth analysis to achieve more significant improvements in RMSE, MSE, and MAPE values and to better understand the features that affect online ojek rates.*

Keywords *- Online Motorcycle Taxis, Pontianak, Predictions, Random Forest, Linear Regression*

1. PENDAHULUAN

Perkembangan teknologi telah memberikan dampak signifikan pada berbagai aspek kehidupan manusia dalam beberapa tahun terakhir, salah satunya adalah industri transportasi. Transformasi ini telah mengubah moda transportasi tradisional menjadi yang lebih canggih dan efisien. Transportasi didefinisikan sebagai tindakan memindahkan produk atau orang dari satu tempat ke tempat lain [1]. Kemunculan aplikasi pemesanan berbasis *mobile* telah mengubah cara masyarakat beraktivitas, terutama di daerah perkotaan yang padat penduduk. Sebelumnya, untuk menggunakan layanan ojek, seseorang harus pergi langsung ke pangkalan ojek, tetapi sekarang, layanan ojek *online* telah menjadikan penggunaan layanan ini lebih praktis dengan aplikasi *mobile*. Transportasi *online* bergantung pada teknologi untuk menemukan penumpang dan menyediakan sistem pembayaran yang mudah bagi penggunanya [2]. Ojek *online* telah menjadi salah satu jenis transportasi *online* yang populer dan dipercayai masyarakat. Keunggulannya adalah kemudahan penggunaan dan tarif yang lebih terjangkau dibandingkan dengan ojek konvensional. Namun, peralihan ini juga telah menimbulkan ketegangan dengan sebagian pengemudi transportasi konvensional yang merasa terganggu adanya transportasi *online*. Para pengemudi transportasi konvensional melakukan tindakan protes, penolakan, sabotase, dan demonstrasi massal sebagai bentuk penolakan terhadap keberadaan ojek *online*. Dengan pesatnya pertumbuhan industri ojek *online*, jumlah transaksi harian di platform ojek *online* pun meningkat secara signifikan. Untuk mengelola kualitas layanan dan menyusun strategi bisnis yang lebih baik, penyedia layanan ojek *online* perlu memahami *feature-feature* yang

mempengaruhi volume transaksi harian. Oleh karena itu, memahami *feature-feature* yang memengaruhi volume transaksi harian dalam industri ojek *online* menjadi kunci dalam manajemen kualitas layanan dan penyusunan strategi bisnis yang lebih baik. Penelitian ini berfokus pada algoritma regresi linier dan *random forest* dalam memprediksi jumlah transaksi ojek *online*. Garis lurus digunakan sebagai representasi hubungan regresi linier antara dua variabel [3]. Terdapat 2 jenis variabel dalam regresi linier yaitu *dependen* dan *independent*. Jika *dependen* keberadaannya dipengaruhi variabel lain, sedangkan *independent* tidak dan dinotasikan dengan simbol X [4]. Sedangkan *random forest* adalah algoritma yang menggunakan sejumlah pohon keputusan *(decision tree) independent* untuk menghasilkan prediksi. Data yang digunakan dalam penelitian ini diperoleh dari dataset *online taxis* (ojek *online*) Pontianak, Indonesia yang diperoleh dari *Kaggle*.

Penelitian sebelumnya oleh Andi Saiful, dkk [5] telah mengimplementasikan metode regresi linier untuk memprediksi harga rumah. Hasil penelitian tersebut menunjukkan tingkat akurasi mencapai 80%, namun nilai *RMSE* yang dihasilkan cukup tinggi yaitu 259171,91. Metode regresi linier berganda digunakan dalam penelitian oleh Kandari Puteri, dkk [6] yang memprediksi harga sembako. Hasil penelitian menunjukkan nilai *R-squared* sebesar 84,2% dengan 3 atribut yang paling berpengaruh yaitu tanggal, komoditas, dan pasar. Menggunakan metode yang berbeda yaitu *random forest*, Siti Saadah, dkk [7] mampu memprediksi harga bitcoin. Dengan data acak, nilai *MAPE* yang dihasilkan adalah 1,50% atau akurasi 98%. Penelitian lain oleh Yusuf Supriyanto, dkk [8] berhasil membandingkan metode regresi linier dan *random forest* dalam memprediksi harga minyak kelapa sawit. Hasil penelitian menunjukkan bahwa metode regresi linier lebih baik dengan nilai *RMSE* sebesar 31.174, sedangkan nilai *RMSE random forest* mencapai 32.850. Dalam penelitian tersebut, digunakan 80% dari data pelatihan dan 20% dari data pengujian. Selain itu penelitian oleh Evita Fitri, dkk [9] mengevaluasi prediksi harga saham yang dibuat dengan 3 metode yang berbeda yaitu regresi linier, *random forest regression*, dan *multilayer perceptron*. Dari hasil penelitian yang dilakukan, nilai *RMSE* paling rendah yaitu menggunakan metode regresi linier sebesar 0,010. Penelitian ini bertujuan untuk memprediksi jumlah transaksi pada layanan ojek *online* berdasarkan *feature-feature* yang memengaruhi. Harapannya, penelitian ini akan memberikan wawasan penting dalam mendukung penyedia layanan ojek *online* dalam pengambilan keputusan bisnis yang lebih efektif, meningkatkan efisiensi operasional, dan memastikan kepuasan pelanggan yang lebih tinggi.

2. METODE PENELITIAN

2.1. Prediksi

Prediksi atau peramalan adalah usaha untuk memperkirakan peristiwa di masa depan [10]. *Machine learning* merupakan salah satu metode yang sangat berguna dalam analisis prediksi, di mana data digunakan untuk mengidentifikasi pola dan informasi [11]. *Machine learning* digunakan untuk mengenali pola, membuat prediksi, mengklasifikasikan data, atau tugas lainnya berdasarkan pola dan fitur yang ditemukan dalam data tersebut. Semakin baik algoritma yang digunakan, semakin akurat keputusan dan prediksi yang dihasilkan oleh sistem tersebut. *Machine learning* bekerja dengan menganalisis data yang dimasukkan ke dalamnya. Melalui proses pelatihan, sistem dapat mengenali pola yang tersembunyi dalam data input dan output [12].

2.2 Regresi linier

Regresi linier adalah salah satu metode prediksi. regresi linier merupakan teknik kuantitatif yang digunakan dalam analisis *time series* yang menggunakan waktu sebagai dasar prediksi. *Regresi linier* dibedakan menjadi dua jenis yaitu regresi linier sederhana dan regresi linier berganda. Jika regresi linier sederhana, digunakan ketika ingin menggambarkan hubungan antara satu variabel *independent* dan satu variabel *dependen*. Berikut adalah persamaannya [13] :

$$Y = a + bX \tag{1}$$

$$a = \frac{\left(\sum X^2\right)(\sum Y) - (\sum XY)(\sum X)}{n(\sum X^2) - (\sum X)^2} \tag{2}$$

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2} \tag{3}$$

Keterangan :

Y = Variabel terikat

X = Variabel bebas

a = *Intercept*

b = Koefisien variabel X

n = Jumlah data

Sedangkan, regresi linier berganda digunakan ketika ingin mempelajari hubungan antara satu variabel *dependen* dan dua atau lebih variabel *independent*. Berikut adalah persamaannya [14] :

$$Y = a + b1X1 + b2X2 + \ldots + bnXn \quad (4)$$

Keterangan :

Y = Variabel terikat

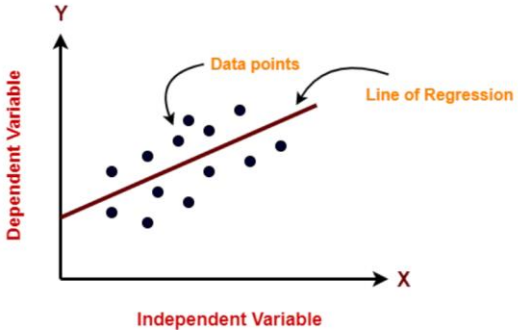
X = Variabel bebas

a = *Intercept*

b = Koefisien variabel X

Pembentukan model regresi linier umumnya melibatkan beberapa langkah, yaitu:

1. Membuat dataset (data *training* dan *testing*).
2. Membuat model regresi linier:
   a. Menentukan nilai X^2, Y^2, XY, serta jumlah dari setiap variabel.
   b. Menggunakan persamaan regresi linier untuk menghitung koefisien a.
   c. Berdasarkan estimasi koefisien yang telah dihitung, langkah selanjutnya membuat model persamaan untuk regresi linier.
   d. Melakukan prediksi atau menghitung nilai variabel *dependen* dan variabel *independent*.
3. Menguji kinerja model. Setelah model regresi linier dibentuk, perlu dilakukan pengujian menggunakan data uji yang terpisah. Beberapa metrik evaluasi yang umum digunakan dalam pengujian regresi linier adalah:
   a. Mean Square Error (MSE), jumlah perbedaan kuadrat antara nilai yang diharapkan dengan nilai sebenarnya. Nilai MSE yang lebih kecil mengindikasikan tingkat kesalahan yang lebih rendah [15].
   b. Root Mean Square Error (RMSE), adalah akar kuadrat dari MSE dan memberikan ukuran kesalahan rata-rata dalam skala yang sama dengan variabel dependen. Semakin kecil nilai RMSE, semakin baik hasil evaluasinya [16].
   c. Mean Absolute Persentase Error (MAPE), memberikan informasi seberapa akurat prediksi model dalam proporsi persentase [17]. Jika nilai MAPE <10% dianggap sangat baik, 10-20% dianggap baik, 20-50% dianggap cukup, dan >50% dianggap buruk.



Gambar 1. Linear Regression Graph

Gambar 1.*Linear regression graph* diatas merujuk pada grafik yang menggambarkan hasil dari analisis regresi linear. Grafik ini digunakan untuk memvisualisasikan hubungan linier antara satu atau lebih variabel *independen* (prediktor) dan variabel *dependen* (respons) dalam suatu studi atau analisis [18].

2.3 Random Forest Regression

Random Forest Regression atau RFR, adalah nama lain untuk teknik *random forest* yang digunakan dalam pemodelan regresi. Breiman memperkenalkan *random forest* pada tahun 2001 [9]. Dengan membuat simpul anak secara acak untuk setiap simpul (simpul di atasnya), algoritma ini dapat meningkatkan hasil akurasi. Algoritma ini terdiri dari beberapa bagian yaitu *root node* (simpul paling atas), *internal node* (simpul percabangan),

dan *leaf node* (simpul terakhir). Untuk menghitung *random forest*, yang pertama dilakukan adalah menghitung nilai *entropy* dan nilai *information gain*. Perhitungannya adalah sebagai berikut :

$$Entropy\ (Y)\ =\ -\sum p(c|Y)log^2p(c|Y) \quad (5)$$

Keterangan :

Y = Himpunan kasus

P(c|Y) = Proporsi nilai Y terhadap kelas c

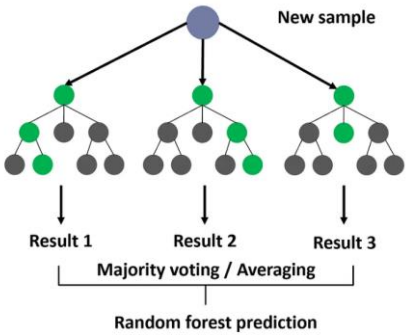
$$Information\ Gain\ (Y,a)\ =\ Entropy\ (Y)\ -\ \sum(v\varepsilon Values\ (a))\ [(|Y_v|\ /\ |Y_a|\ *\ Entropy\ (Y_{v)} \quad (6)$$

Keterangan :

Values(a) = Nilai yang mungkin dalam himpunan kasus a

Y_v = Subkelas dari Y dengan kelas v yang berhubungan dengan kelas a.

Y_a = Semua nilai yang sesuai dengan a

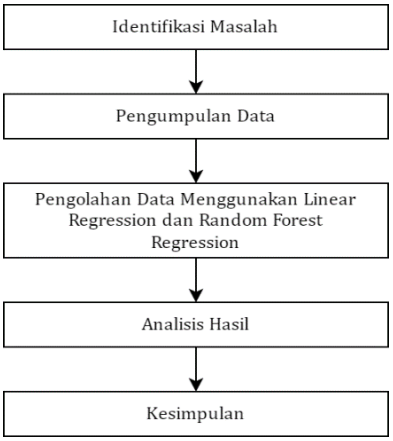


Gambar 2. Random Forest Prediction

Gambar 2. *Random forest prediction* diatas merupakan representasi cara algoritma *random forest* membuat prediksi baru pada dataset yang telah diinput. Gambar tersebut menggambarkan struktur pohon keputusan dalam *random forest*, yang terdiri dari berbagai node dan cabang, yang digunakan untuk memprediksi hasil akhir [19].

2.4 Metodologi

Beberapa langkah yang dilakukan dalam penelitian ini, antara lain:



Gambar 3. Alur Penelitian

Penjelasan dari Gambar 3.Alur penelitian diatas adalah sebagai berikut:

1. Identifikasi Masalah
   Menentukan atau identifikasi permasalahan yang akan dipecahkan, yaitu memprediksi jumlah transaksi ojek *online* menggunakan regresi linier dan *random forest*.
2. Pengumpulan Data
   Langkah selanjutnya adalah melakukan pencarian dan pemilihan dataset yang relevan dari sumber data Kaggle. Dataset tersebut akan menjadi dasar untuk melakukan analisis dan prediksi dalam penelitian ini.
3. Pengolahan Data Menggunakan regresi linier
   Setelah dataset terkumpul, tahap berikutnya adalah melakukan pengolahan data menggunakan regresi linier dan random forest. Proses ini memiliki beberapa tahap, seperti data *cleaning* (penanganan *outlier, transformasi log*, normalisasi data), *eksplorasi* data (menganalisis dan memahami data yang ada), data *preprocessing* (pemisahan data *training* dan data *testing*, model *devining* (mendefinisikan model regresi linier dan *random forest* yang akan digunakan), dan model *evaluation* (evaluasi kinerja model yang telah dilatih).
4. Analisis Hasil
   Setelah model dibentuk dan dievaluasi, tahap selanjutnya adalah menganalisis perbandingan hasil prediksi jumlah transaksi ojek *online* menggunakan regresi linier dan *random forest*.
5. Kesimpulan
   Kesimpulan diambil dari analisis kinerja metode regresi linier dan *random forest* dalam memprediksi jumlah transaksi ojek *online*. Evaluasi dilakukan berdasarkan keakuratan prediksi, kegunaan metode, dan saran untuk penelitian selanjutnya.

3. HASIL DAN PEMBAHASAN

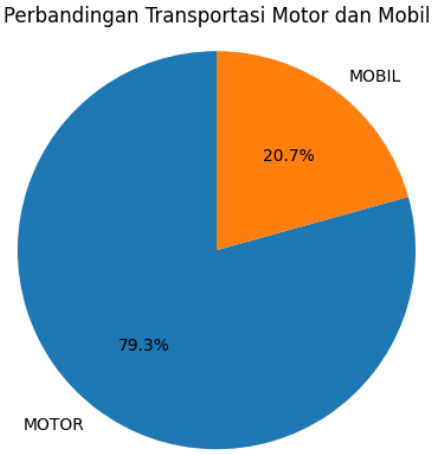
Dalam beberapa tahun terakhir, ojek *online* menjadi mode transportasi yang sangat popular di kalangan masyarakat. Kemudahan pemesanan melalui aplikasi smartphone telah membuat penggunaan ojek *online* menjadi praktis dan efisien. Karena keadaan ini, ojek *online* menjadi alternatif yang diinginkan masyarakat untuk layanan di Indonesia, khususnya di wilayah metropolitan yang padat [20]. Tujuan penelitian ini adalah untuk mengkaji unsur-unsur yang mempengaruhi keseluruhan biaya transaksi ojek *online* dan mengembangkan model prediksi yang memiliki tingkat akurasi tinggi berdasarkan data yang ada. Dataset yang digunakan adalah *online* taxis (ojek *online*) Pontianak Indonesia yang diambil dari Kaggle. Dataset terdiri dari 1017 baris dan 26 kolom yang memiliki tipe data *categorical, integer, dan numerical*. Untuk memprediksi total harga transaksi di Pontianak, perlu dipertimbangkan faktor-faktor spesifik yang berlaku di wilayah tersebut. Proses pertama yang dilakukan adalah data *loading*.

	id	date	mode	from_alamat	from_kelurahan	from_kecamatan	from_latlng	to_alamat	to_kelurahan	to_kecamatan	...	customer_gender	customer_birthdate	driver_id	driver_gender
0	1617	2019/03/09 20:45 s/d 2019/03/09 19:55	BIKE	Gang Ikhwan No.16, Sungai Jawi	DARAT SEKIP	PONTIANAK KOTA	-0,0303277,109,297753	Jl. Prof. M.Yamin No.3a, Sungai Bangkong	BENUA MELAYU LAUT	PONTIANAK SELATAN	...	P	1994-02-05T00:00:00	90	L
1	1297	2019/03/09 19:55 s/d 2019/03/09 19:54	FOOD	Neo Shabu-Shabu Steak & Shake, Johar, Jl. Joha...	SUNGAI BANGKONG	PONTIANAK KOTA	-0,02861,109,329253	Jl. Dare Nandong Villa Ria Indah, Tj. Hulu	BANJAR SERASAN	PONTIANAK TIMUR	...	L	2004-04-22T00:00:00	77	L
2	1394	2019/03/09 19:54 s/d 2019/03/09 18:56	SHOP	Alfamart Pontianak Mall, Jl. Teuku Umar	DARAT SEKIP	PONTIANAK KOTA	-0,0301863,109,3356331	Gg. Gn. Malabar No.21, Sungai Jawi	SUNGAI BELIUNG	PONTIANAK BARAT	...	L	2000-01-07T00:00:00	75	L
3	1120	2019/03/09 18:56 s/d 2019/03/09 12:28	FOOD	Parklife, Jl. Karimata No.64, Sungai Bangkong,...	MARIANA	PONTIANAK KOTA	-0,0305815,109,3264009	Unnamed Road, Pal IX	BANGKA BELITUNG DARAT	PONTIANAK TENGGARA	...	L	1987-08-02T00:00:00	82	L
4	2053	2019/03/09 12:28 s/d 2019/03/08 18:25	CAR	Jl. Tabrani Ahmad No.12, Sungai Jawi Dalam	PAL LIMA	PONTIANAK BARAT	-0,018461872,109,3075679	Pal IX, Kakap River	BANSIR LAUT	PONTIANAK TENGGARA	...	L	2004-01-23T00:00:00	109	P
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1012	1639	2018/09/10 18:28 s/d 2018/09/10 15:29	BIKE	Gg. Metal, Siantan Hulu	SIANTAN HILIR	PONTIANAK UTARA	-0,006927295,109,3522482	Gang Jaya Makmur, Kota Baru	KOTA BARU	PONTIANAK SELATAN	...	L	2004-04-22T00:00:00	77	L
1013	1123	2018/09/10 15:29 s/d 2018/09/10 14:27	FOOD	Mekuru Ramen House, Jl. Ketapang No.31, Benua ...	BENUA MELAYU LAUT	PONTIANAK SELATAN	-0,0331539,109,344327	Jl. Kebangkitan Nasional, Siantan Hulu	BATU LAYANG	PONTIANAK UTARA	...	P	1985-12-24T00:00:00	81	L
1014	1242	2018/09/10 14:27 s/d 2018/09/10 10:30	FOOD	Pondok Ale-ale, Gg. Suka Damai No.21, Sungai B...	SUNGAI BANGKONG	PONTIANAK KOTA	-0,03718,109,325818	Jl. Swadaya, Pal IX	BANGKA BELITUNG LAUT	PONTIANAK TENGGARA	...	L	2002-03-26T00:00:00	93	L
		2018/09/10													

Gambar 4. Data Loading

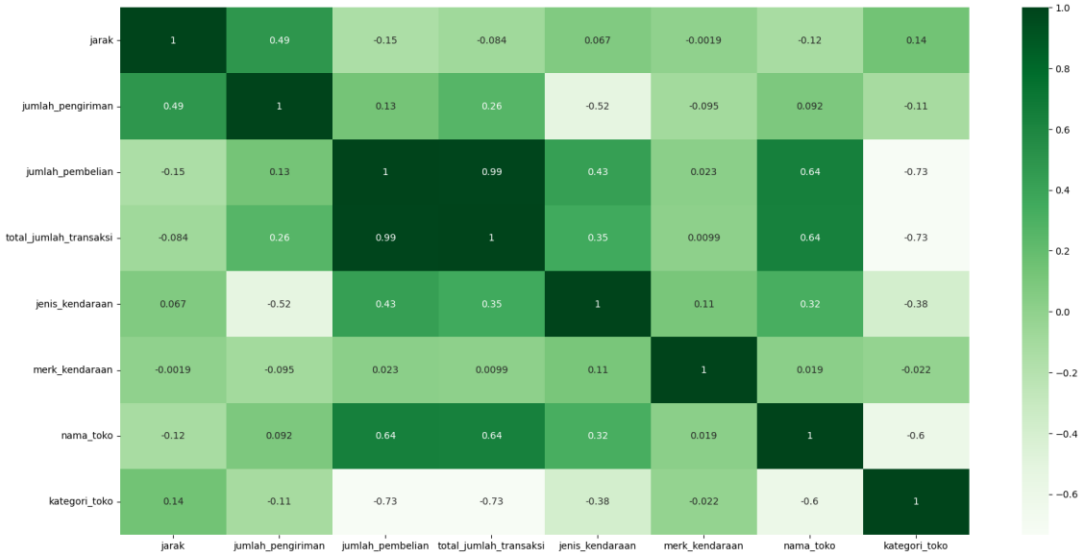
Gambar 4.data loading diatas merupakan proses pengambilan data dengan pustaka *pandas* pada *python*. Setelah dilakukan data *loading*, langkah selanjutnya adalah data *cleaning*. Proses ini meliputi penghapusan kolom yang tidak relevan, penanganan data yang duplikat, penanganan *missing values*, dan penanganan *outlier*. Pada dataset ini, tidak ditemukan adanya data yang duplikat. Namun, terdapat beberapa atribut yang memiliki *missing*

values, yaitu atribut merchant_name dengan 567 nilai yang hilang, to_alamat dengan 3 nilai yang hilang, merchant_id dengan 567 nilai yang hilang, dan merchant_category dengan 567 nilai yang hilang. Selanjutnya, kolom-kolom yang dianggap tidak penting seperti id, customer_id, driver_id, driver_birthdate, customer_birthdate, date, merchant_id, from_latlng, to_latlng, from_alamat, to_alamat, customer_gender, driver_gender, mode, from_kelurahan, to_kecamatan, from_kecamatan, dan to_kelurahan akan dihapus. Data yang mengandung *outlier* adalah data yang berada di luar pola umum dari kumpulan data, sering kali outlier muncul dalam model regresi linier. Data outlier ini melanggar asumsi keberadaan distribusi normal dalam regresi linier [21]. Ketika dilakukan pengecekan outlier, ditemukan bahwa dataset ini memiliki banyak outlier. Oleh karena itu, perlu penanganan outlier dengan metode IQR *(Interquartile Range).* Metode IQR *(Interquartile Range)* membagi data menjadi tiga bagian, yaitu Q1 (kuartil pertama), Q2 (median), hingga Q3 (kuartil ketiga) [22]. Setelah penanganan *outlier* dilakukan, jumlah baris dalam data frame mengalami penurunan dari 1014 baris menjadi 962, sedangkan jumlah kolom tetap 8. Proses tersebut menghasilkan penghapusan sekitar 356 baris data, atau sekitar 35,11% dari jumlah data awal. Setelah proses data *cleaning*, langkah selanjutnya adalah *eksplorasi* data. Tujuan dari tahap ini adalah memahami karakteristik data yang ada, mengidentifikasi pola atau tren, serta mengeksplorasi hubungan antara variabel-variabel yang relevan. Salah satu hasil eksplorasi data adalah diagram pie. Berikut diagram pie yang menggambarkan perbandingan penggunaan transportasi motor dan mobil dalam dataset.



Gambar 5. Pie Chart Perbandingan Transportasi Motor dan Mobil

Dari Gambar 5.pie chart diatas diketahui sekitar 79,3% pengguna menggunakan motor dan 20,7% sisanya menggunakan mobil. Dapat dilihat bahwa presentasi motor lebih tinggi dari pada mobil.



Gambar 6. Heatmap Plot

Berdasarkan visualisasi Gambar 6. heatmap plot di atas, dapat dilihat bahwa warna kotak yang lebih gelap menunjukkan adanya korelasinya lebih kuat, sementara warna yang lebih cerah menunjukkan korelasi yang lemah. Visualisasi ini memberikan gambaran mengenai tingkat hubungan antara fitur-fitur yang ada dalam dataset. Dapat dilihat bawha jarak memiliki korelasi positif dengan jumlah_pengiriman. Setelah melakukan *eksplorasi* data, langkah selanjutnya adalah melakukan *preprocessing* data. Tahap ini meliputi pembagian data menjadi data latih *(train)*, dan data uji *(test)*, transformasi data seperti normalisasi atau *encoding*, serta proses lain yang diperlukan. *Label Encoder* adalah salah satu teknik *encoding* yang dapat digunakan, yang mengubah nilai kategori dalam kolom menjadi nilai *numerik* berdasarkan label yang unik. Setelah proses *encoding* selesai, langkah berikutnya adalah melakukan *feature selection* atau memilih fitur-fitur yang paling relevan untuk memprediksi variabel target. Biasanya, data pelatihan dan pengujian didistribusikan dalam rasio tertentu, seperti 70:30 atau 80:20. Pada penelitian ini menggunakan rasio 70:30. Tahap selanjutnya adalah menentukan model yang akan digunakan yaitu regresi linier dan *random forest.* Setelah menentukan model, model kemudian dilatih menggunakan data *training*. Setelah proses pelatihan selesai, tahap terakhir adalah evaluasi model. Evaluasi model digunakan untuk mengukur kinerja dan akurasi model yang telah dilatih. Hasil evaluasi model akan memberikan gambaran tentang sejauh mana model mampu memprediksi variabel target dengan akurat. Berikut adalah hasil evaluasi model yang dihasilkan.

```
y_pred = lm_1.predict(X_test)
# R-squared data train
lr_train_r2 = lr_i.score(X_train, y_train)
# R-squared data test
lr_test_r2 = lr_i.score(X_test, y_test)
# Root Mean Squared Error (RMSE)
lr_rmse = np.sqrt(mean_squared_error(y_pred, y_test))
# Mean Squared Error (MSE)
mse = mean_squared_error(y_pred, y_test)
# Mean Absolute Percentage Error (MAPE)
mape = np.mean(np.abs((y_test - y_pred) / y_test)) * 100
print('Linear Regression train R Squared     = ', lr_train_r2)
print('Linear Regression test R Squared      = ', lr_test_r2)
print('Root Mean Square Error (RMSE)         = ', lr_rmse)
print('Mean Squared Error (MSE)              = ', mse)
print('Mean Absolute Percentage Error (MAPE) = ', mape)
```

```
Linear Regression train R Squared     =  1.0
Linear Regression test R Squared      =  1.0
Root Mean Square Error (RMSE)         =  1.6306667556190067e-11
Mean Squared Error (MSE)              =  2.659074067881018e-22
Mean Absolute Percentage Error (MAPE) =  1.1813074909714779e-13
```

Gambar 7. Evaluasi Model Regresi Linier

```
y_pred2 = rfr.predict(X_test)
# R-squared data train
rf_train_r2 = rfr_model.score(X_train,y_train)
# R-squared data test
rf_test_r2 = rfr_model.score(X_test, y_test)
# Root Mean Squared Error (RMSE)
rf_rmse = np.sqrt(mean_squared_error(y_pred2,y_test))
# Mean Squared Error (MSE)
mse = mean_squared_error(y_pred2, y_test)
# Mean Absolute Percentage Error (MAPE)
mape = np.mean(np.abs((y_test - y_pred2) / y_test)) * 100
print('Random Forest train R Squared         = %.3f' % rf_train_r2)
print('Random Forest test R Squared          = %.3f' % rf_test_r2)
print('Root Mean Squared Error (RMSE)        = %.3f' % rf_rmse)
print('Mean Squared Error (MSE)              =', mse)
print('Mean Absolute Percentage Error (MAPE) =', mape)
```

```
Random Forest train R Squared         = 1.000
Random Forest test R Squared          = 0.999
Root Mean Squared Error (RMSE)        = 921.195
Mean Squared Error (MSE)              = 848600.0553633218
Mean Absolute Percentage Error (MAPE) = 0.9108293813041904
```

Gambar 8. Evaluasi Model Random Forest

Dari Gambar 8. diatas, model regresi linier yang dilatih memiliki nilai R-squared sebesar 1.0 baik pada data latih maupun data uji. Selain itu, Root Mean Square Error (RMSE) sebesar 1.6 yang menunjukkan rata-rata kesalahan prediksi tarif ojek *online*. Semakin dekat hasil prediksi model dengan nilai sebenarnya menunjukkan semakin rendahnya nilai RMSE. Oleh karena itu, nilai RMSE sebesar 1.6 menandakan bahwa model memiliki tingkat kesalahan yang rendah. Nilai MSE sebesar 2.6 yang menunjukkan rata-rata kesalahan prediksi kuadrat tarif ojek *online*. MSE mengukur kesalahan prediksi rata-rata dalam bentuk kuadrat, dan semakin rendah nilai MSE, semakin baik model dalam mengestimasi nilai target. Serta Nilai Mean Absolute Percentage Error (MAPE) 1.1 yang menunjukkan bahwa kesalahan prediksi tarif ojek *online* rata-rata sekitar 1.1% dari nilai sebenarnya. Semakin

rendah nilai MAPE, semakin baik model dalam memprediksi secara akurat dalam persentase. Nilai MAPE sebesar 1.1% dapat dianggap baik. Sedangkan pada Gambar 6. model ke dua yaitu *random forest* memiliki nilai R-squared sebesar 1.0 pada data latih. Pada data uji memiliki nilai R-squared 0.999. Nilai Root Mean Square Error (RMSE) sebesar 921.195 yang menunjukkan rata-rata kesalahan prediksi tarif ojek *online*. Nilai MSE sebesar 848600.05 yang menunjukkan rata-rata kesalahan prediksi kuadrat tarif ojek *online*. Serta Nilai Mean Absolute Percentage Error (MAPE) 0.91 menunjukkan bahwa rata-rata kesalahan prediksi model adalah sekitar 0.91% dari nilai sebenarnya. Dalam hal nilai R-squared, kedua model (*Random Forest* dan regresi linier) memberikan nilai yang sangat tinggi, mendekati 1.0 pada data pelatihan dan data uji. Ini menunjukkan bahwa keduanya mampu menjelaskan variasi target variabel dengan baik. Model l regresi linier memiliki nilai RMSE yang sangat rendah, bahkan mendekati nol, sedangkan *random forest* memiliki nilai RMSE yang lebih tinggi (921.195). Dalam hal ini, model regresi linier memberikan kesalahan prediksi yang lebih kecil dibandingkan model *random forest*. Model regresi linier juga memberikan nilai MSE yang sangat rendah, bahkan mendekati nol, sedangkan *random forest* memiliki nilai MSE yang lebih tinggi. Selain itu kedua model memiliki nilai MAPE yang rendah, namun model regresi linier memiliki nilai MAPE yang lebih rendah, mendekati nol, yang menunjukkan kesalahan prediksi yang sangat kecil. Berdasarkan perbandingan metrik ini, model regresi linier memberikan kinerja yang lebih baik dalam hal kesalahan prediksi (RMSE dan MSE) serta MAPE yang lebih rendah dibandingkan dengan model *random forest*. Namun, keduanya memiliki nilai R-squared yang sangat tinggi, menunjukkan kemampuan yang baik dalam menjelaskan variasi target variabel. Nilai R-squared sebesar 1.0 dan RMSE sangat kecil pada data pengujian yang sama dengan data pelatihan bisa menjadi indikasi adanya overfitting. Overfitting terjadi ketika model terlalu "memorize" data pelatihan dan tidak mampu menggeneralisasi dengan baik pada data baru. Dengan menggunakan model yang telah dilatih sebelumnya, kita dapat melakukan prediksi tarif ojek *online* baru dengan memasukkan nilai-nilai atribut seperti jumlah_pembelian, kategori_toko, nama_toko, jenis_kendaraan, jumlah_pengiriman, jarak, dan merk_kendaraan. Model akan memproses nilai-nilai tersebut dan menghasilkan prediksi tarif yang sesuai. Berikut adalah contoh prediksi yang dihasilkan menggunakan ke dua model.

```
jumlah_pembelian = float(input('jumlah_pembelian:'))
kategori_toko = float(input('kategori_toko:'))
nama_toko = float(input('nama_toko:'))
jenis_kendaraan = float(input('jenis_kendaraan:'))
jumlah_pengiriman = float(input('jumlah_pengiriman:'))
jarak = float(input('jarak:'))
merk_kendaraan = float(input('merk_kendaraan:'))

x_input = [[jumlah_pembelian,kategori_toko, nama_toko, jenis_kendaraan, jumlah_pengiriman,jarak, merk_kendaraan]]
ypred_lm = lm_1.predict(x_input)
print('Hasil prediksi total harga transaksi baru menggunakan linear regression :')
for prediksi1 in ypred_lm:
    print(prediksi1)
```

```
jumlah_pembelian:50000
kategori_toko:2
nama_toko:1
jenis_kendaraan:4
jumlah_pengiriman:15000
jarak:8
merk_kendaraan:2
Hasil prediksi total harga transaksi baru menggunakan linear regression :
65000.00000000005
```

Gambar 9. Prediksi Tarif Baru regresi linier

```
jumlah_pembelian = float(input('jumlah_pembelian:'))
kategori_toko = float(input('kategori_toko:'))
nama_toko = float(input('nama_toko:'))
jenis_kendaraan = float(input('jenis_kendaraan:'))
jumlah_pengiriman = float(input('jumlah_pengiriman:'))
jarak = float(input('jarak:'))
merk_kendaraan = float(input('merk_kendaraan:'))

x_input = [[jumlah_pembelian,kategori_toko, nama_toko, jenis_kendaraan, jumlah_pengiriman,jarak, merk_kendaraan]]
ypred_rfr = rfr_model.predict(x_input)
print('Hasil prediksi total harga transaksi baru menggunakan random forest :')
for prediksi2 in ypred_rfr:
    print(prediksi1)
```

```
jumlah_pembelian:50000
kategori_toko:2
nama_toko:1
jenis_kendaraan:4
jumlah_pengiriman:15000
jarak:8
merk_kendaraan:2
Hasil prediksi total harga transaksi baru menggunakan random forest :
65000.00000000005
```

Gambar 10. Prediksi Tarif Baru Random Forest

Prediksi tarif baru yang dihasilkan adalah 65.000 dengan jumlah pembelian sebesar 50.000, dengan kategori toko 2, nama toko 1, jenis kendaraan 4, jumlah pengiriman 15.000, jarak 8 km, dan merk kendaraan 2. Hasil prediksi akan disesuaikan dengan nilai atribut yang dimasukkan, jika mengubah nilai atribut maka hasil

prediksi tarif baru akan mengikuti atribut yang dimasukkan. Dari hasil prediksi diatas menunjukkan kedua model memiliki prediksi tarif yang sama.

4. SIMPULAN

Perkembangan teknologi dalam beberapa tahun terakhir telah memiliki dampak besar dalam berbagai aspek kehidupan manusia, termasuk industri transportasi. Kemajuan teknologi telah mengubah moda transportasi tradisional menjadi lebih canggih dan efisien. Aplikasi pemesanan berbasis *mobile* ini telah memberikan kemudahan bagi masyarakat dalam mengakses layanan transportasi. Penelitian ini bertujuan melakukan prediksi jumlah transaksi ojek *online* menggunakan algoritma regresi linier dan *random forest*. Dataset yang digunakan adalah data ojek *online* Pontianak, Indonesia, yang diambil dari Kaggle yang terdiri dari 1017 baris dan 26 kolom. Penelitian mencakup berbagai langkah analisis data, yaitu data *loading*, data *cleaning*, *eksplorasi* data, data *preprocessing*, pendefinisian model, pelatihan model, evaluasi model, dan model *inference*. Berdasarkan perbandingan metrik, model regresi linier memberikan kinerja yang lebih baik dalam hal kesalahan prediksi (RMSE dan MSE) serta MAPE yang lebih rendah dibandingkan dengan model *random forest* yaitu nilai RMSE 1.6, MSE 2.6, dan MAPE 1.1. Sedangkan model *random forest* memperoleh nilai RMSE 921.19, MSE 948600.05, dan MAPE 0.91. Dalam melakukan prediksi tarif baru, tarif baru yang dihasilkan akan mengikuti data pada atribut yang kita masukkan pada model. *Feature* yang digunakan adalah jumlah_pembelian, kategori_toko, nama_toko, jenis_kendaraan, jumlah_pengiriman, jarak, dan merk_kendaraan. Dari prediksi di atas kedua model akan memperoleh prediksi tarif ojek *online* yang sama.

5. SARAN

Diharapkan penelitian selanjutnya dapat melakukan analisis yang lebih mendalam pada dataset seperti mempertimbangkan penambahan atribut yang dapat mempengaruhi tarif ojek *online*, penanganan outlier dengan metode yang berbeda selain IQR, dan eksplorasi algoritma lain. Sehingga diharapkan menghasilkan nilai MSE, RMSE, dan MAPE yang lebih baik.

DAFTAR PUSTAKA

[1] A. K. M. Sugianto, "Tingkat Ketertarikan Masyarakat Terhadap Transportasi Online, Angkutan Pribadi Dan Angkutan Umum Berdasarkan Persepsi," *J. Teknol. Transp. dan Logistik*, vol. 1, no. 2, pp. 51–58, 2020.

[2] C. Andini and D. Akbar, "Tantangan Pariwisata pada Wilayah Perbatasan dalam Era Disrupsi Teknologi: Studi Kasus Regulasi Transportasi Online di Kota Batam, Kepulauan Riau," *Indones. J. Tour. Leis.*, vol. 1, no. 2, pp. 73–81, 2020, doi: 10.36256/ijtl.v1i2.102.

[3] N. R. Lase and F. Riandari, "Perancangan Aplikasi Prediksi Jumlah Pendaftar Siswa Baru Dengan Metode Regresi Linier (Studi Kasus: SMA RK Deli Murni Bandar Baru)," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 3, no. 3, pp. 330–334, 2020, doi: 10.32672/jnkti.v3i3.2520.

[4] N. Suhandi, E. A. K. Putri, and S. Agnisa, "Analisis Pengaruh Jumlah Penduduk terhadap Jumlah Kemiskinan Menggunakan Metode Regresi Linear di Kota Palembang," *J. Ilm. Inform. Glob.*, vol. 9, no. 2, pp. 77–82, 2018, doi: 10.36982/jig.v9i2.543.

[5] A. Saiful, "Prediksi Harga Rumah Menggunakan Web Scrapping dan Machine Learning Dengan Algoritma Linear Regression," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 1, pp. 41–50, 2021, doi: 10.35957/jatisi.v8i1.701.

[6] K. Puteri and A. Silvanie, "Machine Learning Untuk Model Prediksi Harga Sembako Dengan Metode Regresi Linier Berganda," *J. Nas. Inform.*, vol. 1, no. 2, pp. 82–94, 2020.

[7] S. Saadah and H. Salsabila, "Prediksi Harga Bitcoin Menggunakan Metode Random Forest (Studi Kasus: Data Acak Pada Awal Masa Pandemic Covid-19)," *J. Komput. Terap.*, vol. 7, no. Vol. 7 No. 1 (2021), pp. 24–32, 2021, doi: 10.35143/jkt.v7i1.4618.

[8] S. Supriyanto, M. Ilhamsyah, and U. Enri, "Prediksi Harga Minyak Kelapa Sawit MenggunakanLinear Regression Dan Random Forest," *J. Ilm. Wahana Pendidik.*, vol. 8, no. 7, pp. 1–8, 2022, doi: 10.5281/zenodo.6559603.

[9] E. Fitri and D. Riana, "Analisa Perbandingan Model Prediction Dalam Prediksi Harga Saham Menggunakan Metode Linear Regression, Random Forest Regression Dan Multilayer Perceptron," *METHOMIKA J. Manaj. Inform. dan Komputerisasi Akunt.*, vol. 6, no. 1, pp. 69–78, 2022, doi: 10.46880/jmika.vol6no1.pp69-78.

[10] G. N. Ayuni and D. Fitrianah, "Penerapan metode Regresi Linear untuk prediksi penjualan properti pada PT XYZ," *J. Telemat.*, vol. 14, no. 2, pp. 79–86, 2019, [Online]. Available: https://journal.ithb.ac.id/telematika/article/view/321

[11] H. K. Pambudi, P. G. A. Kusuma, F. Yulianti, and K. A. Julian, "Prediksi Status Pengiriman Barang Menggunakan Metode Machine Learning," *J. Ilm. Teknol. Infomasi Terap.*, vol. 6, no. 2, pp. 100–109, 2020, doi: 10.33197/jitter.vol6.iss2.2020.396.

[12] M. Mahendra, R. Chandra Telaumbanua, A. Wanto, and A. Perdana Windarto, "Akurasi Prediksi Ekspor Tanaman Obat, Aromatik dan Rempah-Rempah Menggunakan Machine Learning," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol.

2, no. 6, pp. 207–215, 2022, doi: 10.30865/klik.v2i6.402.

[13] N. Almumtazah, N. Azizah, Y. L. Putri, and D. C. R. Novitasari, "Prediksi Jumlah Mahasiswa Baru Menggunakan Metode Regresi Linier Sederhana," *J. Ilm. Mat. Dan Terap.*, vol. 18, no. 1, pp. 31–40, 2021, doi: 10.22487/2540766x.2021.v18.i1.15465.

[14] M. Masruroh and K. F. Mauladi, "Penerapan Metode Regresi Linear Berganda Dalam Sistem Prediksi Nilai Ujian Nasional Siswa Smp," *J. Tek.*, vol. 12, no. 1, p. 1, 2020, doi: 10.30736/jt.v12i1.393.

[15] F. H. Hamdanah and D. Fitrianah, "Analisis Performansi Algoritma Linear Regression dengan Generalized Linear Model untuk Prediksi Penjualan pada Usaha Mikra, Kecil, dan Menengah," *J. Nas. Pendidik. Tek. Inform.*, vol. 10, no. 1, p. 23, 2021, doi: 10.23887/janapati.v10i1.31035.

[16] D. Saputro and D. Swanjaya, "Analisa Prediksi Harga Saham Menggunakan Neural Network Dan Net Foreign Flow," *Gener. J.*, vol. 7, no. 2, pp. 96–104, 2023, doi: 10.29407/gj.v7i2.20001.

[17] D. Swanjaya and D. Putra Pamungkas, "Analisa Hasil Prediksi Metode Least Square menggunakan Korelasi dan MAPE pada Toko PS," *Gener. J.*, vol. 5, no. 1, pp. 11–18, 2021, doi: 10.29407/gj.v5i1.15440.

[18] N. Verma, "Linear Regression Algorithm Explained in Less Than 5 Minutes." [Online]. Available: https://medium.com/@techynilesh/linear-regression-explained-in-less-than-5-minutes-5f90a26a33a8

[19] R. Yehoshua, "Random Forest." [Online]. Available: https://medium.com/@roiyeho/random-forests-98892261dc49

[20] D. G. Nugroho, Y. H. Chrisnanto, and A. Wahana, "Analisis Sentimen Pada Jasa Ojek Online ... (Nugroho dkk.)," pp. 156–161, 2015.

[21] N. A. Atamia, Y. Susanti, and S. S. Handajani, "Perbandingan Analisis Regresi Robust Estimasi-S dan Estimasi-M dengan Pembobot Huber dalam Mengatasi Outlier," *Pros. Semin. Nas. Mat.*, vol. 4, pp. 673–679, 2021, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/

[22] N. A. Iskandar, I. Ernawati, and Y. Widiastiwi, "Klasifikasi Diagnosis Penyakit Stroke Dengan Menggunakan Metode Random Forest," *Semin. Nas. Mhs. Ilmu Komput. dan Apl.*, pp. 432–441, 2022, [Online]. Available: https://conference.upnvj.ac.id/index.php/senamika/article/view/2190