OPEN ACCESS

Volume 9 (3), July-September 2025, 790-801

E-ISSN:2580-1643

Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)

DOI: https://doi.org/10.35870/jtik.v9i3.3490

Perbandingan Prediksi terhadap Peningkatan Jumlah Pelanggan Iconnet dengan Algoritma Regresi Linear dan Random Forest pada Wilayah Jabodetabek dan Banten

Fitrah Amelia Ramelan [1*], Lukman Hakim [2]

[1,2] Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana Jakarta, Kota Jakarta Pusat, Daerah Khusus Ibukota Jakarta, Indonesia.*

a r t i c l e i n f o

Article history:
Received 10 December 2024
Received in revised form
20 December 2024
Accepted 15 January 2025
Available online Juli 2025.

Keywords:
Linear Regression; Random Forest; Customer Prediction.

Kata Kunci:
Regresi Linear: Random Forest: Prediksi Pelanggan.

a b s t r a c t

This study compares the linear regression and random forest algorithms in predicting the number of Iconnet service customers in the Jabodetabek and Banten regions. The dataset comprises two years of sales data processed through filtering, cleaning, and labeling. Evaluation metrics include MAE, MAPE, and RMSE. The results show that linear regression performs better in predicting customer numbers, achieving MAE 369.85, MAPE 8.80%, and RMSE 388.89, compared to random forest with MAE 679.37, MAPE 16.95%, and RMSE 794.26. Conversely, random forest outperforms linear regression in bandwidth prediction (MAE 733.80, MAPE 26.61%, RMSE 860.20) and regional prediction (MAE 25607.49, MAPE 23.42%, RMSE 38177.12), as linear regression produces higher errors. The findings highlight the importance of selecting algorithms based on data characteristics and application needs. This research provides strategic insights for developing data-driven customer service solutions.

a b s t r a k

Penelitian ini membandingkan algoritma regresi linear dan random forest dalam memprediksi jumlah pelanggan layanan Iconnet di Jabodetabek dan Banten. Dataset terdiri dari data penjualan dua tahun terakhir yang melalui proses filterisasi, pembersihan, dan pelabelan. Evaluasi dilakukan menggunakan MAE, MAPE, dan RMSE. Hasil menunjukkan regresi linear lebih baik untuk memprediksi jumlah pelanggan dengan MAE 369.85, MAPE 8.80%, dan RMSE 388.89, dibandingkan random forest (MAE 679.37, MAPE 16.95%, RMSE 794.26). Sebaliknya, random forest unggul dalam prediksi bandwidth (MAE 733.80, MAPE 26.61%, RMSE 860.20) dan wilayah (MAE 25607.49, MAPE 23.42%, RMSE 38177.12), dibandingkan regresi linear yang menghasilkan kesalahan lebih besar. Kesimpulan menekankan pentingnya memilih algoritma berdasarkan jenis data dan kebutuhan aplikasi. Penelitian ini memberikan wawasan strategis untuk pengembangan layanan pelanggan berbasis data.

ACM Computing Classification System (CCS)
EBSCOhost
Communication and Mass Media Complete (CMMC)

Corresponding Author. Email: amelfitrah07@gmail.com [1].

Copyright 2025 by the authors of this article. Published by Lembaga Otonom Lembaga Informasi dan Riset Indonesia (KITA INFO dan RISET). This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

1. Pendahuluan

Perkembangan teknologi informasi dan tingginya permintaan masyarakat terhadap akses internet yang terus meningkat, mendorong perusahaan penyedia layanan internet untuk terus berinovasi dalam memperluas jangkauan dan meningkatkan kualitas layanan mereka. Salah satu perusahaan yang beroperasi di sektor ini adalah PT PLN ICON PLUS (ICON+). Peningkatan jumlah pelanggan menjadi salah satu faktor kunci yang menentukan keberhasilan perusahaan. Oleh karena itu, perusahaan-perusahaan dalam industri ini memerlukan pemahaman mendalam mengenai faktor-faktor yang mempengaruhi pertumbuhan pelanggan mereka. Dalam hal ini, analisis prediktif berfungsi sebagai alat yang efektif untuk meramalkan peningkatan jumlah pelanggan di masa depan. Prediksi adalah proses estimasi yang dilakukan secara sistematis dengan tujuan untuk menentukan kemungkinan kejadian di masa depan berdasarkan data yang tersedia, baik dari masa lalu maupun masa kini. Tujuan utama dari prediksi adalah untuk meminimalkan kesalahan atau selisih antara hasil prediksi dengan kenyataan yang terjadi (Kafil, 2019).

Dalam penelitian ini, metode yang digunakan untuk memprediksi data pelanggan adalah *Regresi Linear* dan *Random Forest*. *Regresi linear* merupakan metode statistik yang digunakan untuk menganalisis hubungan antara variabel dependen (misalnya jumlah pelanggan) dan variabel independen lainnya seperti harga, promosi, dan faktor eksternal lainnya (Widiastuti *et al*., 2022). Di sisi lain, *Random Forest* adalah algoritma dalam pembelajaran mesin yang digunakan untuk tugas klasifikasi dan regresi. Algoritma ini dikenal karena kemampuannya untuk menangani data yang kompleks dan fleksibilitasnya dalam memberikan solusi *prediksi* yang akurat. *Random Forest* dibangun dari kumpulan beberapa *decision tree*, yang berfungsi untuk memetakan data ke dalam ruang fitur guna meminimalkan kesalahan *prediksi* secara maksimal (Erlin *et al*., 2022). Beberapa penelitian terdahulu terkait *prediksi* menggunakan algoritma *regresi linear* dan *Random Forest* menunjukkan berbagai aplikasi dan hasil yang bervariasi. Salah satunya adalah penelitian mengenai *prediksi* tarif ojek online menggunakan kedua algoritma tersebut, yang menunjukkan bahwa *regresi linear* menghasilkan akurasi dengan *RMSE* 1.6, *MSE* 2.6, dan *MAPE* 1.1, sedangkan *Random Forest* menghasilkan *RMSE* 921.19, *MSE* 948600.05, dan *MAPE* 0.91 (Pramesti & Wiga Maulana Baihaqi, 2023). Penelitian lainnya membandingkan kedua algoritma untuk memprediksi kasus positif COVID-19, di mana *regresi linear* menghasilkan *RMSE* 3031.127 dan *MAPE* 47.66, sementara *Random Forest* menunjukkan *RMSE* 1886.555 dan *MAPE* 14.85 (Fachid & Triayudi, 2022). Penelitian serupa juga dilakukan dalam memprediksi harga bawang di kota Samarinda, yang menghasilkan akurasi lebih rendah untuk *regresi linear* (53.75) dibandingkan dengan *Random Forest* (753.36) (Aditya Pratama *et al*., 2024). Berdasarkan latar belakang tersebut, permasalahan utama yang perlu diteliti adalah bagaimana memprediksi pertumbuhan jumlah pelanggan Iconnet di masa mendatang. Hal ini bertujuan untuk mendukung proses pengambilan keputusan yang lebih akurat dan relevan. Oleh karena itu, penelitian ini difokuskan pada perbandingan *prediksi* peningkatan jumlah pelanggan Iconnet menggunakan algoritma *regresi linear* dan *Random Forest* pada wilayah Jabodetabek dan Banten.

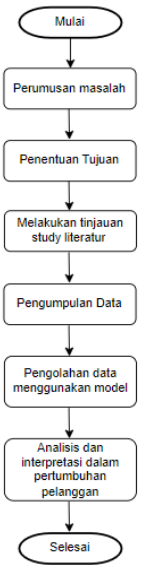
2. Metodologi Penelitian

Dalam era *big data* dan teknologi informasi yang berkembang pesat, analisis data menjadi komponen krusial dalam pengambilan keputusan di berbagai bidang, termasuk keuangan, kesehatan, dan pendidikan. Algoritma *machine learning* seperti *regresi linear* dan *random forest* telah menjadi metode andalan dalam prediksi dan analisis data karena kemampuan mereka dalam menangani beragam jenis data. *Regresi linear*, sebagai pendekatan statistik yang sederhana dan interpretatif, sangat efektif untuk memahami hubungan linier antar variabel. Di sisi lain, *random forest*, sebagai algoritma *ensemble* yang lebih kompleks, unggul dalam menangani data yang non-linier dan bersifat multivariat. Sebelum memasuki metode penelitian, berikut adalah tahapan penelitian untuk studi kasus ini.

Tahapan Penelitian

Tahapan pertama dimulai dengan perumusan masalah, yaitu mengidentifikasi masalah utama: bagaimana melakukan *prediksi* jumlah pelanggan Iconnet menggunakan algoritma *regresi linear* dan

random forest; bagaimana perbandingan hasil pemanfaatan algoritma *regresi linear* dengan *random forest* dalam melakukan *prediksi*; serta bagaimana perbandingan hasil akurasi yang diperoleh dari perhitungan MAE, MAPE, dan RMSE untuk algoritma *regresi linear* dibandingkan dengan *random forest* dalam *prediksi* tersebut.



Gambar 1. Tahapan Penelitian

Setelah masalah dirumuskan, tahap kedua adalah menentukan tujuan penelitian. Tujuan utama dari penelitian ini adalah untuk *memprediksi* hasil pertumbuhan pelanggan Iconnet dan untuk mengetahui tingkat akurasi terbaik di antara kedua algoritma dengan menggunakan perhitungan MAE, MAPE, dan RMSE. Hal ini bertujuan untuk menentukan seberapa optimal model tersebut dalam melakukan *prediksi*. Selanjutnya, dilakukan *Tinjauan Studi Literatur*, yaitu pengumpulan dan penelaahan literatur atau penelitian terdahulu yang relevan dengan perbandingan *prediksi* jumlah pelanggan serta metode *regresi linear* dan *random forest* yang diterapkan dalam penelitian ini. Tahap berikutnya adalah pengumpulan data, di mana dataset yang digunakan dalam penelitian ini diperoleh dari hasil penjualan Iconnet selama dua tahun terakhir yang mencakup berbagai parameter. Dataset tersebut akan menjadi studi kasus dalam penelitian ini. Setelah data berhasil dikumpulkan, langkah selanjutnya adalah *Pengolahan Data Menggunakan Model*, di mana data yang telah dikumpulkan akan diolah melalui *preprocessing* dan dianalisis menggunakan model dengan metode algoritma *regresi linear* serta *random forest*. Setelah data diolah, dilakukanlah *Analisis dan Interpretasi Hasil*, yang bertujuan untuk memahami implikasi dari data yang telah dianalisis dalam hal pertumbuhan pelanggan. Tahapan ini akan diakhiri dengan tahap *Selesai*.

$$Y = a + bX$$

Dimana :
Y = Variabel dependen
X = Variabel independen
a = Konstanta / *Intercept*
b = Koefisien regresi / *Slope*

Algoritma Random Forest

Algoritma random forest merupakan teknik dalam *machine learning* yang diterapkan untuk tugas klasifikasi, regresi, dan berbagai tugas lainnya yang berkaitan dengan *prediksi*. *Random forest* memiliki berbagai keunggulan, seperti kemampuan untuk menghasilkan tingkat kesalahan yang rendah, kinerja unggul dalam klasifikasi, efisiensi dalam mengolah dataset besar, serta efektivitas dalam mengestimasi data yang hilang (Pamuji & Ramadhan, 2021). Selain itu, *random forest* juga dikenal karena fleksibilitasnya yang tinggi dan kemampuan untuk menangani berbagai jenis data, baik yang bersifat linier maupun non-linier (Afdhal *et al.*, 2022).

***Mean Absolute Error* (MAE)**

MAE (Mean Absolute Error) adalah metrik yang digunakan untuk mengukur kesalahan dalam model *prediksi* atau regresi, dengan skala yang sama seperti data asli, sehingga mudah dipahami. Semakin kecil nilai *MAE*, semakin baik kinerja model, karena prediksi yang dihasilkan lebih mendekati nilai aktual (Mulyana & Marjuki, 2022). Berikut adalah rumus untuk *MAE*:

$$MAE = \sum \frac{|Y\prime - Y|}{n}$$

Dimana:
Y' = Nilai Prediksi
Y = Nilai Sebenarnya
n = Jumlah Data

***Mean Absolute Percentage Error* (MAPE)**

MAPE (*Mean Absolute Percentage Error*) adalah metrik yang digunakan untuk menilai akurasi model *prediksi* atau regresi dengan mengungkapkan kesalahan sebagai persentase dari nilai aktual. *MAPE* juga merupakan metode pengukuran akurasi dalam *forecasting* (Andrean *et al.*, 2021). *MAPE* dinyatakan dalam bentuk persentase, yang membuatnya lebih mudah dipahami dan dibandingkan antar berbagai model atau dataset. Secara umum, nilai *MAPE* yang lebih rendah menunjukkan kinerja model *prediksi* yang lebih baik, karena persentase kesalahan *prediksi* terhadap nilai aktual menjadi lebih kecil (Setiawan, 2021). Berikut adalah rumus untuk *MAPE*:

$$RMSE = \sqrt{\sum \frac{(Y1 \quad - \quad iY)2}{n}}$$

Dimana:

Y^1 = Nilai Prediksi
Y = Nilai Sebenarnya
n = Jumlah titik data

3. Hasil dan Pembahasan

Hasil

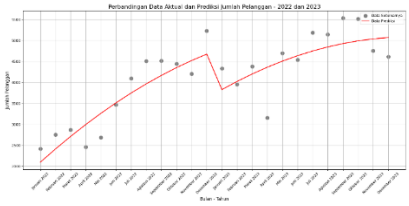
Algoritma Regresi Linear

Proses dimulai dengan menetapkan variabel independen (sumbu X), yaitu bulan, dan variabel dependen (sumbu Y), yaitu jumlah pelanggan. Selanjutnya, *algoritma regresi linear* diimpor untuk membangun model. Setelah model dibangun, nilai *intercept* dan *koefisien* dari fitur dataset ditampilkan untuk memberikan pemahaman lebih lanjut mengenai hubungan antara bulan dan jumlah pelanggan.

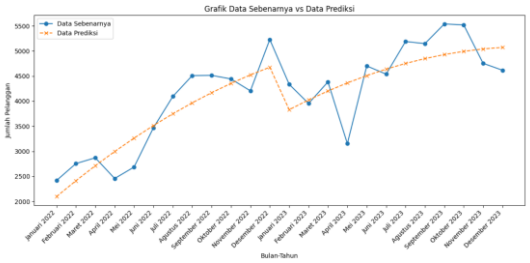
Tabel 1. Menggabungkan Data Pelanggan Sebenarnya Dan Prediksi Regresi Linear

	Tahun	Bulan	Data Sebenarnya	Data Prediksi
0	2022	1	2416	2095.687793
2	2022	2	2754	2411.702432
4	2022	3	2870	2711.440589
...	...	...	...	…
23	2023	12	4616	5072.087678

Tabel 1 memperlihatkan kombinasi antara data pelanggan aktual dan data hasil prediksi. Data yang telah diproses kemudian disajikan dalam bentuk diagram untuk mempermudah pemahaman dan interpretasi. Dapat dilihat dari gambar 2 dimana tahun 2022 pada bulan desember menunjukan jumlah pelanggan terbanyak baik nilai sebenarnya maupun prediksi, sedangkan tahun 2023 pada bulan September menunjukan jumlah pelanggan terbanyak pada data sebenarnya dan bulan desember menunjukan jumlah pelanggan terbanyak pada data prediksi.



Gambar 2. Visualisasi Data Pelanggan Sebenarnya dan Prediksi Regresi Linear Tahun 2022 dan 2023



Gambar 3. Visualisasi Data Pelanggan Perbulan Dalam 24 Bulan Regresi Linear

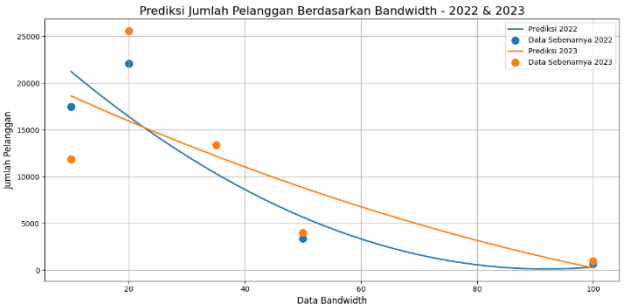
Bandwidth

Bandwidth merujuk pada kapasitas maksimum yang dapat dicapai oleh saluran komunikasi untuk mentransfer data dalam periode waktu tertentu. Pengukuran *bandwidth* yang tepat sangat penting dalam menganalisis kinerja jaringan, karena semakin besar *bandwidth*, semakin banyak data yang dapat ditransmisikan, yang pada gilirannya meningkatkan kualitas layanan yang diberikan kepada pelanggan.

Tabel 2. Menggabungkan Data *Bandwith* Sebenarnya Dan Prediksi Regresi Linear

	Tahun	*Bandwith*	Jumlah Pelanggan	Prediksi Jumlah Pelanggan
0	2022	10	17468	21238.138122
1	2022	20	22059	16403.792818
2	2022	50	3406	5668.082873
...	...	...	...	…
8	2023	100	987	246.847625

Tabel 2 menunjukkan kombinasi antara data pelanggan aktual dan hasil prediksi. Tahapan selanjutnya merepresentasikan data yang telah diproses menggunakan *algoritma regresi linear*. Berdasarkan Gambar 4, pada tahun 2022 dan 2023, hasilnya menunjukkan bahwa *bandwidth* 20 memiliki jumlah pelanggan terbanyak. Namun, pada garis prediksi, *bandwidth* 10 menunjukkan jumlah pelanggan terbanyak.



Gambar 4. Visualisasi Data *Bandwith* Sebenarnya dan Prediksi Regresi Linear Tahun 2022 dan 2023

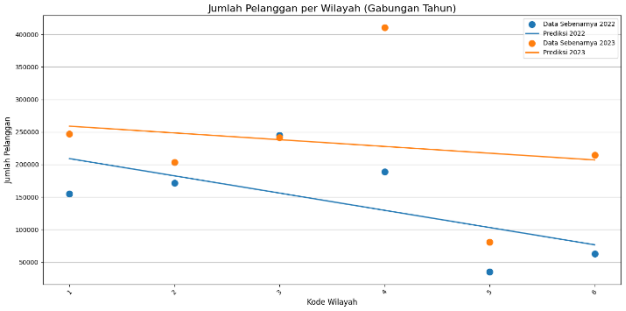
Wilayah

Pada tahap ini, data disusun untuk menganalisis jumlah pelanggan per bulan berdasarkan wilayah.

Tabel 3. Menggabungkan Data Wilayah Sebenarnya dan Prediksi

	Tahun	Kode Wilayah	Jumlah Pelanggan	Prediksi Jumlah Pelanggan
0	2022	1	155010	209068.571429
1	2022	2	171280	182585.142857
2	2022	3	244690	156101.714286
...	...	...	...	…
8	2023	6	214350	206981.190476

Pada tabel 3 memperlihatkan kombinasi antara data wilayah aktual dan data hasil prediksi. Ditunjukan pada gambar 5 pada tahun 2022 jumlah pelanggan terbanyak berada di wilayah 3 (area depok), namun pada garis prediksinya di wilayah 1 (area jakarta) memiliki jumlah pelanggan terbanyak. Sedangkan pada tahun 2023 jumlah pelanggan terbanyak berada di wilayah 4 (area Tangerang), namun pada garis prediksinya di wilayah 1 (area jakarta) memiliki jumlah pelanggan terbanyak.



Gambar 5. Visualisasi Data Wilayah Sebenarnya dan Prediksi Regresi Linear Tahun 2022 dan 2023

Algoritma Random Forest

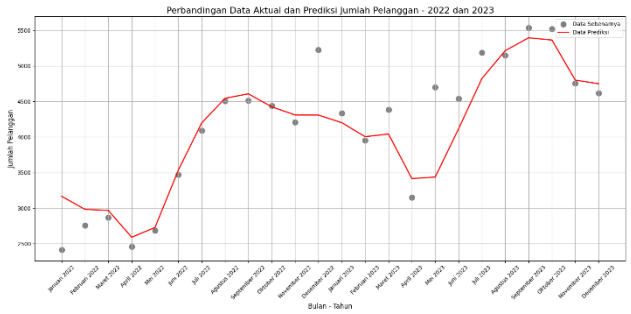
Dalam analisis prediksi jumlah pelanggan, *algoritma random forest* diterapkan untuk mengidentifikasi pola yang lebih kompleks dalam data yang mungkin tidak

dapat ditangkap oleh model linier sederhana. Algoritma ini menggunakan teknik *ensemble learning*, yang menggabungkan banyak *decision tree* untuk menghasilkan prediksi yang lebih akurat dan mengurangi variabilitas hasil prediksi. Penerapan *random forest* pada data jumlah pelanggan bertujuan untuk memperbaiki akurasi prediksi dengan mempertimbangkan berbagai faktor yang saling terkait.

Tabel 4. Menggabungkan Data Pelanggan Sebenarnya Dan Prediksi *Random Forest*

	Tahun	Bulan	Data Sebenarnya	Data Prediksi
0	2022	1	2416	3166.58
2	2022	2	2754	2982.90
4	2022	3	2870	2966.72
...	...	...	...	…
23	2023	12	4616	4749.95

Pada tabel 4 memperlihatkan kombinasi antara data pelanggan aktual dan data hasil prediksi.



Gambar 6. Visualisasi Data Pelanggan Sebenarnya dan Prediksi *Random Forest* Tahun 2022 dan 2023

Pada Gambar 6, grafik data menggambarkan perbedaan antara nilai data aktual dan data prediksi. Pada tahun 2022, grafik menunjukkan nilai yang sejalan dan relevan, namun terdapat perbedaan yang cukup signifikan pada bulan Desember, yang disebabkan oleh peningkatan ketertarikan pelanggan terhadap produk. Sementara itu, pada tahun 2023, grafik memperlihatkan adanya perbedaan yang lebih mencolok antara data aktual dan data prediksi, dengan ketidaksesuaian nilai yang lebih jelas. Perbedaan signifikan terjadi pada bulan April, yang diakibatkan oleh penurunan ketertarikan pelanggan terhadap produk, diikuti dengan peningkatan penjualan yang tercatat pada bulan Mei.

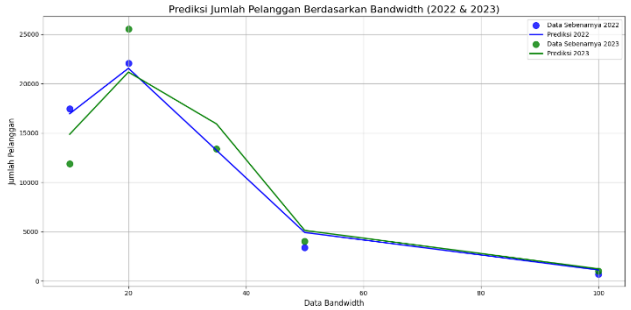
Bandwidth

Bandwidth merujuk pada jumlah data yang dapat ditransmisikan melalui saluran komunikasi dalam waktu tertentu. Sebagai salah satu parameter utama dalam kualitas layanan jaringan, *bandwidth* memainkan peran penting dalam menentukan seberapa cepat dan efisien informasi dapat diproses dan dikirimkan, yang berpengaruh langsung pada pengalaman pengguna dalam menggunakan layanan internet.

Tabel 5. Menggabungkan Data *Bandwith* Sebenarnya dan Prediksi *Random Forest*

	Tahun	*Bandwith*	Data Sebenarnya	Pre diksi Jumlah Pelanggan
0	2022	10	17468	16948.03
1	2022	20	22059	21555.28
2	2022	50	3406	4913.47
...	...	...	...	…
8	2023	100	987	1224.44

Pada tabel 5 memperlihatkan kombinasi antara data pelanggan aktual dan data hasil prediksi.



Gambar 7. Visualisasi Data *Bandwith* Sebenarnya dan Prediksi *Random Forest* Tahun 2022 dan 2023

Pada Gambar 7, grafik data menunjukkan perbedaan antara nilai data aktual dan data prediksi. Pada tahun 2022, grafik memperlihatkan nilai yang sejalan dan berbanding lurus, namun terdapat perbedaan pada *bandwidth* 50 yang disebabkan oleh penurunan ketertarikan pelanggan terhadap *bandwidth* yang lebih besar. Sementara itu, pada tahun 2023, grafik menunjukkan nilai yang relevan, meskipun ada perbedaan pada *bandwidth* 20, yang dipengaruhi oleh peningkatan ketertarikan pelanggan terhadap *bandwidth* tersebut.

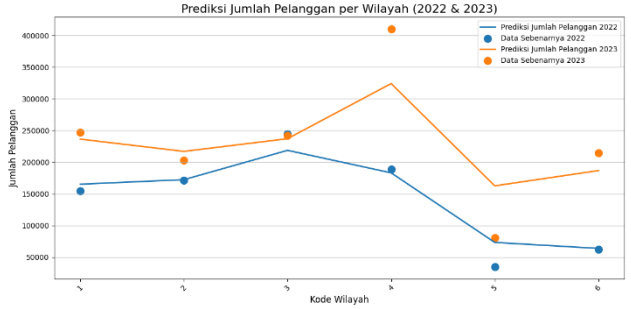
Wilayah

Dalam analisis data pelanggan, faktor wilayah memainkan peran penting dalam memahami distribusi layanan dan preferensi pelanggan. Setiap wilayah memiliki karakteristik yang berbeda, yang dapat mempengaruhi pola penggunaan layanan. Oleh karena itu, pemetaan dan analisis berdasarkan wilayah dapat memberikan wawasan yang lebih mendalam mengenai variabilitas permintaan dan perilaku pelanggan di berbagai lokasi.

Tabel 6. Menggabungkan Data Wilayah Sebenarnya dan Prediksi

	Tahun	Kode Wilayah	Jumlah Pelanggan	Prediksi Jumlah Pelanggan
0	2022	1	155010	165192.40
1	2022	2	171280	172351.20
2	2022	3	244690	218631.40
...	...	...	...	...
8	2023	6	214350	186686.55

Pada tabel 6 memperlihatkan kombinasi antara data pelanggan aktual dan data hasil prediksi.



Gambar 8. Visualisasi Data Wilayah Sebenarnya Dan Prediksi *Random Forest* Tahun 2022 dan 2023

Pada Gambar 8, grafik untuk tahun 2022 menunjukkan nilai yang relevan, meskipun terdapat perbedaan pada kode wilayah 3 (Area Depok), yang disebabkan oleh peningkatan ketertarikan pelanggan, serta pada kode wilayah 5 (Area Bekasi), yang dipengaruhi oleh penurunan ketertarikan pelanggan terhadap produk. Sementara itu, pada tahun 2023, grafik kembali menunjukkan nilai yang relevan, namun ada perbedaan pada kode wilayah 4 (Area Tangerang) akibat peningkatan ketertarikan pelanggan, dan pada kode wilayah 5 (Area Bekasi) yang menunjukkan penurunan ketertarikan pelanggan terhadap produk.

Analisis Hasil

Analisis hasil perbandingan antara model regresi linear dan *random forest* sebagai berikut:

Model	Kategori	RMSE	MAE	MAPE
Linear Regression	Prediksi Jumlah Pelanggan	388.89	369.85	8.80%
	Prediksi Bandwidth	4884.21	3906.78	51.63%
	Prediksi Wilayah	78211.56	56829.52	47.70%
Random Forest	Prediksi Jumlah Pelanggan	794.26	679.37	16.95%
	Prediksi Bandwidth	860.20	733.80	26.61%
	Prediksi Wilayah	38177.12	25607.49	23.42%

Gambar 9. Perhitungan Evaluasi Model Regresi Linear Dan *Random Forest*

Berdasarkan Gambar 9, hasil penelitian menunjukkan perbedaan signifikan antara performa *regresi linear* dan *random forest* dalam memprediksi berbagai variabel. Untuk *prediksi* jumlah pelanggan, *regresi linear* lebih unggul dengan *RMSE* 388.89, *MAE* 369.85, dan *MAPE* 8.80%, dibandingkan dengan *Random Forest* yang memiliki *RMSE* 794.26, *MAE* 679.37, dan *MAPE* 16.95%. Sebaliknya, pada *prediksi* bandwidth, *Random Forest* menunjukkan akurasi yang jauh lebih baik dengan *RMSE* 860.20, *MAE* 733.80, dan *MAPE* 26.61%, dibandingkan dengan *regresi linear* yang memiliki *RMSE* 4884.21, *MAE* 3906.78, dan *MAPE* 51.63%. Dalam *prediksi* wilayah, *Random Forest* kembali unggul dengan *RMSE* 38177.12, *MAE* 25607.49, dan *MAPE* 23.42%, dibandingkan dengan *regresi linear* yang memiliki *RMSE* 78211.56, *MAE* 56829.52, dan *MAPE* 47.70%. Secara keseluruhan, *regresi linear* lebih efektif untuk *prediksi* jumlah pelanggan, sementara *Random Forest* lebih baik dalam memprediksi bandwidth dan wilayah. Penelitian ini menggunakan dataset primer dari perusahaan PT PLN ICON PLUS (ICON+) yang mencakup periode 2022–2023. Fitur yang diambil dalam penelitian ini meliputi *bandwidth*, *month*, *year*, *wilayah*, *latitude*, dan *longitude*.

Tabel 7. *Dataset* Penjualan Iconnet

BW	*Month*	*Year*	Wilayah	*Lat*	...
10 Mbps	April	2022	Kab. Bekasi	-6.1546	...
20 Mbps	April	2022	Kab. Bekasi	-6.1564	...
10 Mbps	April	2022	Kab. Bekasi	-6.1567	...
10 Mbps	April	2022	Kab. Bekasi	-6.1569	...
20 Mbps	April	2022	Kab. Bekasi	-6.1583	...

Dataset ini terdiri dari 6 fitur dengan total 101.326 data, yang mencakup 4 data objek dan 2 data numerik.

Pre-Processing

Pre-processing adalah tahap untuk mentransformasi data agar sesuai dengan format yang tepat dan siap untuk diproses (Shevira *et al.*, 2022). Pada tahap *pre-processing*, data akan diolah melalui tiga tahapan utama, yaitu *filterisasi data*, *cleaning data*, dan *labeling data*.

Filterisasi Data

Tahapan pertama dalam *pre-processing* adalah *filterisasi data*, yang melibatkan pemilihan subset data yang memenuhi kriteria tertentu atau penghapusan data yang dianggap tidak relevan. Langkah ini bertujuan untuk menyaring data sehingga analisis dapat difokuskan pada informasi yang benar-benar penting. Pada tahap ini, data yang tidak diperlukan, seperti *Latitude* dan *Longitude*, akan dihapus untuk memastikan fokus analisis pada subset data yang relevan.

	BW	MONTH	YEAR	WILAYAH
0	10 Mbps	April	2022	Kab. Bekasi
1	20 Mbps	April	2022	Kab. Bekasi
2	10 Mbps	April	2022	Kab. Bekasi
3	10 Mbps	April	2022	Kab. Bekasi
4	20 Mbps	April	2022	Kab. Bekasi

Gambar 10. Filterisasi Data

Cleaning Data

Tahapan kedua adalah *cleaning data*, proses ini bertujuan untuk mengidentifikasi, memperbaiki, dan menghapus ketidakakuratan, ketidakkonsistenan, atau data yang tidak relevan dalam sebuah dataset. Langkah-langkahnya mencakup penanganan data yang terduplikasi, data yang hilang, atau data yang tidak memenuhi format yang ditentukan. Tahap ini mencakup:

1) Menghilangkan karakter *non* numerik pada kolom *bandwith*

```
   BW  MONTH  YEAR      WILAYAH
0  10  April  2022  Kab. Bekasi
1  20  April  2022  Kab. Bekasi
2  10  April  2022  Kab. Bekasi
3  10  April  2022  Kab. Bekasi
4  20  April  2022  Kab. Bekasi
```

Gambar 11. Menghilangkan Karakter *Non* Numerik Pada Kolom *Bandwith*

2) Mengubah nilai pada kolom bulan menjadi angka

	BW	MONTH	YEAR
0	10	4	2022
1	20	4	2022
2	10	4	2022
3	10	4	2022
4	20	4	2022

Gambar 12. Mengubah Nilai Pada Kolom Bulan Menjadi Angka

Labeling Data

Tahapan ketiga adalah *labeling data*, *Labeling data* adalah proses memberikan label pada data berdasarkan karakteristik tertentu. Pada tahap ini, dibuat kolom baru dengan nama "Kode Wilayah" untuk memberikan label wilayah sesuai dengan kode wilayah sebagai berikut:

3) Kode Wilayah 1 : Area Jakarta
   a. Kota Adm. Jakarta Pusat
   b. Kota Adm. Jakarta Barat
   c. Kota Adm. Jakarta Timur
   d. Kota Adm. Jakarta Selatan
   e. Kota Adm. Jakarta Utara

4) Kode Wilayah 2 : Area Bogor
   a. Kota Bogor
   b. Kab. Bogor

5) Kode Wilayah 3 : Area Depok
   a. Kota Depok

6) Kode Wilayah 4 : Area Tangerang
   a. Kota Tangerang
   b. Kota Tangerang Selatan
   c. Kab. Tangerang

7) Kode Wilayah 5 : Area Bekasi
   a. Kota Bekasi
   b. Kab. Bekasi

8) Kode Wilayah 6 : Area Banten
   a. Kota Serang
   b. Kab. Serang
   c. Kota Cilegon

	BW	MONTH	YEAR	Kode Wilayah
0	10	4	2022	5
1	20	4	2022	5
2	10	4	2022	5
3	10	4	2022	5
4	20	4	2022	5

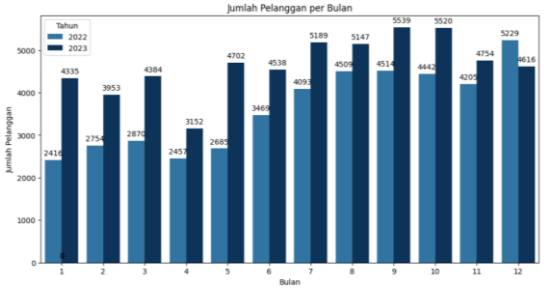
Gambar 13. Mengubah Wilayah Menjadi Kode Wilayah

Pembuatan Model

Pada tahap ini, model dibangun dengan membagi data menjadi dua subset, yaitu data *training* dan data *testing*. Data *training* digunakan untuk melatih model *prediksi*, sementara data *testing* digunakan untuk mengevaluasi kinerja model yang telah dilatih. Umumnya, proporsi data *training* lebih besar, dengan pembagian 80% untuk *training* dan 20% untuk *testing*. Proses ini juga bertujuan untuk menghindari masalah seperti *overfitting* atau *underfitting*, serta memastikan bahwa model dapat diterapkan secara efektif dalam situasi nyata.

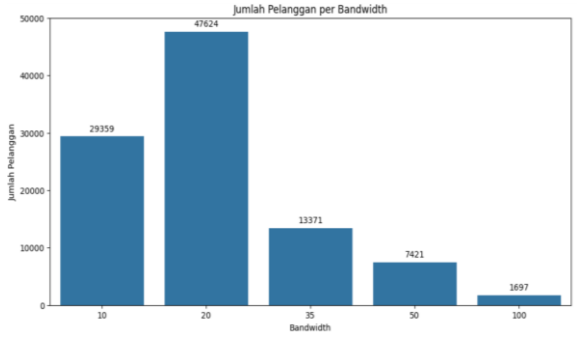
Visualisasi Data

Pada tahapan *visualisasi data*, grafik dibuat sebagai salah satu metode terbaik untuk mengeksplorasi dan menyampaikan hasil temuan. Grafik memungkinkan analisis yang lebih mudah dan intuitif, serta memudahkan pemahaman terhadap pola yang ada dalam data.



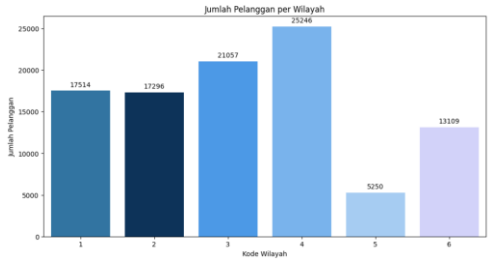
Gambar 14. Visualisasi Data Pelanggan Dengan Bulan

Pada gambar 14 menunjukkan visualisasi data pelanggan dengan nilai x adalah bulan, dan nilai y adalah jumlah pelanggan. Dimana menampilkan data jumlah terbanyak pada tahun 2022 di bulan 12 yaitu Desember dan pada tahun 2023 di bulan 9 yaitu September, dan jumlah data terkecil pada tahun 2022 di bulan 1 yaitu Januari dan pada tahun 2023 di bulan 4 yaitu April.



Gambar 15. Visualisasi Data Pelanggan Dengan *Bandwith*

Pada gambar 15 menunjukan visualisasi data *bandwidth* dengan nilai x adalah *bandwidth*, dan nilai y adalah jumlah pelanggan. Dimana data jumlah terbanyak di *bandwidth* 20 mbps, dan terkecil di 100 mbps.



Gambar 16. Visualisasi Data Pelanggan Dengan Wilayah

Pada gambar 16 menunjukan visualisasi data wilayah dengan nilai x adalah kode wilayah, dan nilai y adalah jumlah pelanggan. Dimana data jumlah terbanyak di wilayah 4 atau Area Tangerang, dan terkecil di wilayah 5 atau Area Bekasi. Data dirapihkan, dievaluasi pada visualisasi data. Dari data tersebut menghasilkan data untuk prediksi.

Pembahasan

Berdasarkan hasil yang diperoleh dari penerapan algoritma *regresi linear* dan *random forest*, terdapat perbedaan yang signifikan dalam kinerja kedua model dalam memprediksi berbagai variabel. Pada *prediksi* jumlah pelanggan, *regresi linear* menunjukkan kinerja yang lebih baik dengan *RMSE* 388.89, *MAE* 369.85, dan *MAPE* 8.80%. Hal ini sesuai dengan temuan yang dijelaskan oleh Galih *et al.* (2023), yang menunjukkan bahwa *regresi linear* sangat efektif untuk memprediksi variabel dengan hubungan yang relatif linier dan sederhana. Sebaliknya, *random forest* menghasilkan *RMSE* 794.26, *MAE* 679.37, dan *MAPE* 16.95%, yang lebih tinggi dibandingkan dengan *regresi linear*. Hal ini menunjukkan bahwa *random forest* mungkin lebih cocok untuk menangani data yang lebih kompleks, namun dalam hal ini tidak lebih akurat dibandingkan dengan *regresi linear* untuk *prediksi* jumlah pelanggan. Dalam *prediksi* bandwidth, *random forest* menunjukkan performa yang lebih baik dengan *RMSE* 860.20, *MAE* 733.80, dan *MAPE* 26.61%, sementara *regresi linear* menghasilkan *RMSE* 4884.21, *MAE* 3906.78, dan *MAPE* 51.63%. Temuan ini sejalan dengan penelitian Afdhal *et al.* (2022), yang menunjukkan bahwa *random forest* lebih unggul dalam menangani data yang lebih tidak terstruktur atau memiliki hubungan non-linier. Penggunaan *random forest* dalam kasus ini memungkinkan model untuk menangkap pola yang lebih kompleks dalam data *bandwidth*, yang dipengaruhi oleh berbagai faktor eksternal. Pada *prediksi* wilayah, *random forest* kembali menunjukkan keunggulan dengan *RMSE* 38177.12, *MAE* 25607.49, dan *MAPE* 23.42%, dibandingkan dengan *regresi linear* yang memiliki *RMSE* 78211.56, *MAE* 56829.52, dan *MAPE* 47.70%. Hasil ini menunjukkan bahwa *random forest* lebih efektif dalam memodelkan variabilitas yang ada dalam data wilayah, yang sering kali dipengaruhi oleh faktor eksternal seperti lokasi geografis dan demografis pelanggan. Hal ini sejalan dengan penelitian Fachid & Triayudi (2022), yang menunjukkan keunggulan *random forest* dalam menangani data dengan banyak variabel yang saling berinteraksi.

Secara keseluruhan, *regresi linear* terbukti lebih efektif dalam *prediksi* jumlah pelanggan, sementara *random forest* lebih unggul dalam memprediksi bandwidth dan wilayah, yang melibatkan variabel dengan hubungan yang lebih kompleks. Penelitian ini juga menegaskan pentingnya memilih algoritma yang sesuai dengan karakteristik data yang dianalisis. Seperti yang dijelaskan oleh Pramesti & Wiga Maulana Baihaqi (2023), pemilihan algoritma yang tepat sangat mempengaruhi hasil *prediksi*, terutama ketika data memiliki variabilitas yang tinggi. Dataset yang digunakan dalam penelitian ini adalah dataset primer dari PT PLN ICON PLUS (ICON+) untuk periode 2022–2023. Fitur yang digunakan dalam penelitian ini mencakup *bandwidth*, *month*, *year*, *wilayah*, *latitude*, dan *longitude*, yang memberikan gambaran komprehensif mengenai faktor-faktor yang mempengaruhi permintaan layanan di berbagai wilayah.

4. Kesimpulan

Berdasarkan temuan dari penelitian ini, dapat disimpulkan bahwa *regresi linear* menunjukkan performa terbaik dalam memprediksi jumlah pelanggan dengan hasil evaluasi yang mencakup *RMSE* sebesar 388.89, *MAE* sebesar 369.85, dan *MAPE* sebesar 8.80%. Sementara itu, dalam memprediksi *bandwidth* dan *wilayah*, *random forest* menunjukkan kinerja yang lebih unggul dengan *RMSE* 860.20, *MAE* 733.80, dan *MAPE* 26.61% untuk *bandwidth*, serta *RMSE* 38177.12, *MAE* 25607.49, dan *MAPE* 23.42% untuk *wilayah*.

Random forest secara keseluruhan memiliki tingkat kesalahan yang lebih rendah dibandingkan dengan *regresi linear* dalam kedua variabel tersebut. Kedua algoritma ini memiliki potensi besar untuk dioptimalkan lebih lanjut melalui penyesuaian parameter model dan penambahan fitur yang lebih relevan, guna meningkatkan akurasi *prediksi*. Pemilihan model yang tepat berpengaruh besar dalam mendukung pengambilan keputusan strategis, yang dapat meningkatkan kualitas layanan pelanggan, efisiensi operasional, serta efektivitas dalam pengelolaan data.

5. Daftar Pustaka

Afdhal, I., Kurniawan, R., Iskandar, I., Salambue, R., Budianita, E., & Syafria, F. (2022). Penerapan Algoritma Random Forest Untuk Analisis Sentimen Komentar Di YouTube Tentang Islamofobia. *Penerapan Algoritma Random Forest Untuk Analisis Sentimen Komentar Di YouTube Tentang Islamofobia*, *5*(1), 122-130.

Andrean, D., Susanto, R., & Hakim, L. (2021). Aplikasi Forecasting Supplier Sembako dengan Metode Least Square dan Double Exponential Smoothing. *Jurnal Teknik Informatika UNIKA Santo Thomas*, 386-394. https://doi.org/10.54367/jtiust.v6i2.1568.

Erlin, E., Desnelita, Y., Nasution, N., Suryati, L., & Zoromi, F. (2022). Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang. *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, *21*(3), 677-690. https://doi.org/10.30812/matrik.v21i3.1726.

Fachid, S., & Triayudi, A. (2022). Perbandingan Algoritma Regresi Linier dan Regresi Random Forest Dalam Memprediksi Kasus Positif Covid-19. *Jurnal Media Informatika Budidarma*, *6*(1), 68.

Galih, M., & Atika, P. D. (2022). Prediksi Penjualan Menggunakan Algoritma Regresi Linear pada Koperasi Karyawan Usaha Bersama. *Journal of Informatic and Information Security*, *3*(2), 193-202. https://doi.org/10.31599/jiforty.v3i2.1354.

Hidayat, R., Saputra, H. T., Husnah, M., Nabila, N., Hidayatullah, M. B., Nazhmi, M. N., ... & Rana, A. (2025). Implementasi Algoritma Random Forest Regression Untuk Memprediksi Penjualan Produksi di Supermarket. *Jurnal Sistem Informasi dan Sistem Komputer*, *10*(1), 101-109. https://doi.org/10.51717/simkom.v10i1.703.

Ibrahim, M. R., Susanto, R., & Nurchim, N. (2024). IMPLEMENTASI SISTEM APLIKASI PEMESANAN MAKANAN MENGGUNAKAN METODE COLLABORATIVE FILTERING PADA FOOD COURT BERBASIS WEB. *Jurnal Informatika Teknologi dan Sains (Jinteks)*, *6*(3), 612-619. https://doi.org/10.51401/jinteks.v6i3.4313.

Kafil, M. (2019). Penerapan Metode K-Nearest Neighbors Untuk Prediksi Penjualan Berbasis Web Pada Boutiq Dealove Bondowoso. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *3*(2), 59-66. https://doi.org/10.36040/jati.v3i2.860.

Mulyana, D. I. (2022). Optimasi Prediksi Harga Udang Vaname dengan Metode RMSE dan MAE Dalam Algoritma Regresi Linier. *JURNAL ILMIAH BETRIK: Besemah Teknologi Informasi dan Komputer*, *13*(1), 50-58.

Pamuji, F. Y., & Ramadhan, V. P. (2021). Komparasi Algoritma Random Forest dan Decision Tree

untuk Memprediksi Keberhasilan Immunotheraphy. *Jurnal Teknologi Dan Manajemen Informatika*, *7*(1), 46-50. https://doi.org/10.26905/jtmi.v7i1.5982.

Pramesti, D., & Baihaqi, W. M. (2023). Perbandingan Prediksi Jumlah Transaksi Ojek Online Menggunakan Regresi Linier dan Random Forest. *Generation Journal*, *7*(3), 21-30. https://doi.org/10.29407/gj.v7i3.20676.

Setiawan, I. (2021). Rancang Bangun Aplikasi Peramalan Persediaan Stok Barang Menggunakan Metode Weighted Moving Average (WMA) Pada Toko Barang XYZ. *Jurnal Teknik Informatika*, *13*(3), 1-9. https://doi.org/10.26905/jtmi.v7i1.5982.

Shevira, S., Made, I., Suarjaya, A. D., & Buana, P. W. (2022). Pengaruh Kombinasi dan Urutan Pre-Processing pada Tweets Bahasa Indonesia. *JITTER-Jurnal Ilm. Teknol. dan Komput*, *3*(2).

Sholeh, M., Nurnawati, E. K., & Lestari, U. (2023). Penerapan Data Mining dengan Metode Regresi Linear untuk Memprediksi Data Nilai Hasil Ujian Menggunakan RapidMiner. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *8*(1), 10-21. https://doi.org/10.14421/jiska.2023.8.1.10-21.

Widia, F., & Murniati, W. (2022). Penerapan Data Mining Untuk Memprediksi Penjualan Kain Tenun Mnggunakan Regresi Linear. *Jurnal Ilmiah Teknik Mesin, Elektro dan Komputer*, *2*(1), 27-39.