Analisa dan Prediksi Transportasi Transjakarta Menggunakan Algoritma Regresi Linear dan Pustaka Pyspark

Transjakarta Transportation Analysis and Prediction Using Linear Regression Algorithm and Pyspark Library

Muhamad Ridwan
Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa
Muhamad.ridwan@mhs.pelitabangsa.ac.id

Abstract

Transportation facilities are very important facilities for the community, because almost everyone uses transportation facilities to travel somewhere. Regression is a statistical method used to estimate the relationship between a related variable and one or more independent variables. This method can also be used to assess the strength of the relationship between variables and future forecasts. This study used a linear regression algorithm with RMSE values of 66473, MAE 36063 and showed high values and R-Squared of 0.627 and pyspark library for data analysis. Although it has limitations, including the datasets used and the use of Linear Regression algorithms, it is expected to contribute and be useful to future research.

Keywords*: Data Science, Data Analysis, Linear Regression, Transportation*

Abstrak

Sarana transportasi merupakan sarana yang sangat penting bagi masyarakat, pasalnya hampir semua orang menggunakan sarana transportasi untuk berpergian ke suatu tempat. Regresi adalah metode statistik yang dipakai untuk memperkirakan hubungan antara sebuah variabel terkait dengan satu variabel independen atau lebih. Metode ini juga bisa digunakan untuk menilai kekuatan hubungan antara variabel dengan perkiraan masa depan. Pada penelitian ini menggunakan algoritma regresi linier dengan nilai RMSE sebesar 66473, MAE 36063 dan menunjukan nilai yang tinggi dan dan R-Squared sebesar 0.627 dan pustaka pyspark untuk anlisa data. Meski memiliki keterbatasan, termasuk dataset yang digunakan dan penggunaan algoritma Regresi Linier, diharapkan dapat berkontribusi dan bermanfaat pada penelitian yang akan datang.

Kata kunci: Sains Data, Analisis Data, Algoritma Linear Regresi, Transportasi

Pendahuluan

Sarana transportasi merupakan sarana yang sangat penting bagi masyarakat, pasalnya hampir semua orang menggunakan sarana transportasi untuk berpergian ke suatu tempat, baik itu jauh atau dekat, bekerja atau bermain, dan banyak alasan lain untuk menggunakan transportasi. Terdapat 2 jenis transportasi, diantaranya transportasi pribadi merupakan sarana transportasi milik pribadi seperti sepeda motor pribadi, mobil pribadi, dan sepeda. Sedangkan transportasi umum merupakan transportasi yang digunakan bersama orang lain dan biasa nya berjumlah banyak seperti bus, kereta, angkot dan lain-lain.

Transportasi darat terdiri dari dua kategori: transportasi jalan raya (road transport) dan transportasi jalan rel (rail transport). Transportasi jalan raya adalah jenis transportasi yang digunakan oleh manusia, termasuk hewan, sepeda, sepeda motor, becak, mobil, bus, truk, dan kendaraan bermotor lainnya. Jalan umum terdiri dari jalan setapak, jalan tanah, jalan kerikil, dan jalan aspal. Transportasi jalan rel adalah jenis transportasi yang difungsikan seperti kereta api dengan rel baja dan digerakkan oleh uap, diesel, dan listrik. Tenaga penggerak yang digunakan termasuk manusia, hewan, uap, BBM, dan diesel[1], [2].

Jakarta merupakan kota dengan padat penduduk, dimana moda transportasi darat sangat diperlukan, namun dengan banyaknya jumlah penduduk jakarta, baik itu pribumi maupun perantauan. Transportasi pribadi menjadi kendala kemacetan yang terjadi. Maka dari itu perlu penggunaan transportasi umum untuk mengurangi kemacetan yang terjadi.

PT. Transjakarta mengembangkan jasa layanan angkutan umum yang meliputi jasa layanan angkutan umum pengumpan, layanan integrasi, layanan angkutan umum Transjabodetabek dan layanan angkutan umum lainnya yang memerlukan standar pelayanan minimal di dalam pengoperasian[3]. Konsep kenyamanan, keamanan, kecepatan, dan keramahan harus diperhatikan saat membuat pelayanan transportasi yang baik. Layanan transportasi memiliki dampak yang signifikan pada pelanggannya, terutama bagi masyarakat menengah ke bawah.

Bus Transjakarta menjadi moda transportasi yang banyak digunakan, karena harganya yang cukup terjangkau dan mempunyai banyak rute perjalanan. Terdapat beberapa jenis bus transjakarta, dan mempunyai banyak rute perjalanan menjadi hal yang sulit dalam menganalisa performa dari bus transjakarta itu sendiri. Banyak hal yang bisa didapatkan dari histori penumpang, seperti melihat jenis bus mana yang lebih banyak penumpangnya, rute mana yang sering mendapatkan penumpang terbanyak, dan masih banyak kuntungan yang bisa didapatkan.

Dalam menganalisa data perlu pengetahuan lebih lanjut tentang bagaimana mengelola data yang ada. Sains data melibatkan pengumpulan, pengolahan, dan analisis data untuk menghasilkan data bermanfaat[4]. Data science adalah bidang ilmu yang khusus mempelajari data, terutama data kuantitatif (angka), baik yang terstruktur maupun tidak terstruktur. Bidang ini mencakup semua proses yang berkaitan dengan data, seperti pengumpulan, analisis, pengolahan, manajemen, kearsipan, pengelompokan, penyajian, distribusi, dan cara mengubah data menjadi kumpulan data yang dapat dipahami dan digunakan. Ilmu data juga bagian dari ekonomi, khususnya ilmu bisnis[5].

Algoritma Regresi linier adalah teknik analisis statistik yang digunakan untuk menemukan hubungan fungsional antara dua variabel, di mana satu variabel (variabel independen) mempengaruhi atau memprediksi nilai variabel lainnya (variabel dependen). Tujuan regresi linier adalah untuk menemukan garis terbaik (best fit line) yang dapat digunakan untuk memprediksi nilai variabel dependen berdasarkan nilai variabel independen yang diberikan[6][7].

Regresi adalah metode statistik yang dipakai untuk memperkirakan hubungan antara sebuah variabel terkait dengan satu variabel independen atau lebih. Metode ini juga bisa digunakan untuk menilai kekuatan hubungan antara variabel dengan perkiraan masa depan.

Regresi linear sederhana merupakan suatu model persamaan yang menggamba rkan hubungan satu variabel bebas (X) dengan satu variabel tak bebas (Y)[8], [9], [10].

Persamaan regresi linear sederhana secar matematik diekspresikan oleh

$\hat{Y} = a + bX$

Dimana

$\hat{Y}$ = garis regresi

a = konstanta (intersep)

b = konstanta regresi (slope)

X = varabel bebas (predictor) Besarnya

Beberapa penelitian sebelumnya telah mengeksplorasi penerapan algoritma regresi linier dalam konteks prediksi. Studi yang dilakukan Ajeng Afifah Muhartini menunjukkan bahwa algoritma regresi linear berhasil memprediksi jumlah penerimaan mahasiswa baru. Studi ini menerapkan algoritma regresi linear pada data penerimaan mahasiswa dari tahun ke tahun dan mengidentifikasi pola penerimaan yang signifikan. Hasilnya akurasi dari model dengan algoritma regresi linear sebesar 96,556% .

Berdasarkan tinjauan literatur yang dilakukan, penerapan algoritma regresi linear memiliki akurasi yang cukup tinggi, sehingga menjadi pilihan pada penelitian ini. Diharapkan dengan dilakukan nya penelitian ini dapat berkontribusi pada penelitian yang mendatang.

Metode Penelitian

Penelitian yang dilakukan menggunakan algoritma regresi linier dan pustaka pyspark pada proses analisa datanya. Langkah-langkah yang dilakukan dalam penelitian ini adalah

1. Persiapan Data:
   a) Data digabungkan menjadi 1 file, karena data tersebut terpisah setiap bulan nya.
   b) Melakukan preprocessing data, seperti cleansing data, memperbaiki data yang salah
2. Analisa Data:
   a) Menampilkan visualisasi data yang sudah di preprocessing
   b) Menampilkan data berdasarkan kategori yang ingin dilihat
3. Pembuatan model machine learning dengan algoritma regresi linear untuk memprediksi data
4. Pengujian hasil prediksi

Sumber Data

Dataset yang digunakan adalah dataset publik dari open data jakarta dengan URL dataset transjakarta, yang merupakan sumber data publik. Dataset ini berisi informasi tentang jenis transjakarta, trayek, dan jumlah penumpang tiap bulan nya. Sehingga, dataset ini dapat memberikan pemahaman yang lebih baik tentang pola penumpang transjakarta dan mengetahui rute atau jenis transjakarta mana yang paling banyak penumpang. Dengan dataset ini, peneliti dapat melakukan analisis mendalam terhadap pola penumpang transjakarta. ALgoritma regresi linier diharapkan dapat ditrapkan pada dataset ini, sehingga dapat mengetahui kinerja algoritma dengan dataset mempunya kecocokan atau tidak.

Hasil dan Pembahasan

Percobaan dilakukan dalam beberapa tahap dan menggunakan python sebagai bahasa pemrograman nya dengan bantuan library pyspark sebagai alat bantu dalam menganalisa dan implementasi algoritma regresi linier. Berikut tahapan pelaksanaan percobaannya:

1. Menggunakan Python:

a) Python digunakan sebagai bahasa pemrograman untuk mengimplementasikan algoritma regresi linier.
b) Pustaka pyspark digunakan untuk manipulasi data, dan juga di pustaka pyspark terdapat algoritma regresi linier yang bisa digunakan.

Preprocessing data

Preprocessing data menggunakan pyspark dengan beberapa tahap yang dilakukan, berikut meruoakan sample data yang akan digunakan pada penelitian ini.

```
# Load the dataset into a PySpark DataFrame
file_path = "transjakarta.csv"
transjakarta_df = spark.read.csv(file_path, header=True, inferSchema=True)

# Display the schema and first few rows of the DataFrame
transjakarta_df.printSchema()
transjakarta_df.show(5, truncate=False)
```

```
root
 |-- tahun: integer (nullable = true)
 |-- bulan: integer (nullable = true)
 |-- jenis: string (nullable = true)
 |-- kode_trayek: string (nullable = true)
 |-- trayek: string (nullable = true)
 |-- jumlah_penumpang: integer (nullable = true)

+-----+-----+----------+-----------+-----------------------------------+----------------+
|tahun|bulan|jenis     |kode_trayek|trayek                             |jumlah_penumpang|
+-----+-----+----------+-----------+-----------------------------------+----------------+
|2021 |8    |Mikrotrans|JAK.88     |Terminal Tanjung Priok - Ancol Barat|20245           |
|2021 |8    |Mikrotrans|JAK.85     |Bintara - Cipinang Indah           |19989           |
|2021 |8    |Mikrotrans|JAK.84     |Terminal Kampung Melayu - Kapin Raya|33638           |
|2021 |8    |Mikrotrans|JAK.80     |Rawa Buaya - Rawa Kompeni          |46653           |
|2021 |8    |Mikrotrans|JA.77      |Tanjung Priok - Jembatan Item      |47157           |
+-----+-----+----------+-----------+-----------------------------------+----------------+
only showing top 5 rows
```

Gambar 1sampel data

Melakukan cleansing terhadap data dan memperbaiki data jika terdapat data yang hilang

```
# Check for missing values
missing_values = transjakarta_df.select([col(c).alias(c) for c in transjakarta_df.columns]).\
    select([col(c).isNull().cast("int").alias(c) for c in transjakarta_df.columns]).\
    describe()
```

```
# Convert missing values summary to a Pandas DataFrame for better display
missing_values.toPandas().transpose()
```

	0	1	2	3	4
summary	count	mean	stddev	min	max
tahun	1473	0.0	0.0	0	0
bulan	1473	0.0	0.0	0	0
jenis	1473	0.0	0.0	0	0
kode_trayek	1473	0.0	0.0	0	0
trayek	1473	0.0013577732518669382	0.03683545644565792	0	1
jumlah_penumpang	1473	0.0	0.0	0	0

Gambar 2 Pengecekan Data Yang Hilang

Pada gambar diatas, terdapat data yang tidak sesua pada kolom trayek, sehingga perlu penyesuaian data seperti menghapus data yang tidak sesuai

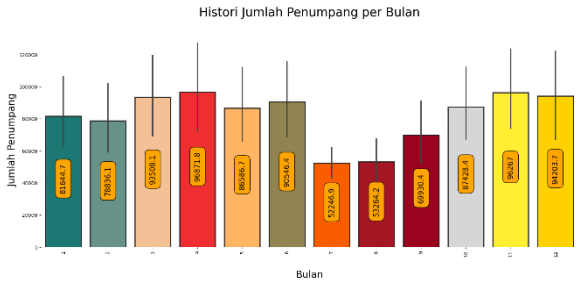
```
# Handle missing values

transjakarta_df = transjakarta_df.na.drop()  # Drop rows with any missing values
#count missing value
for column in transjakarta_df.columns:
    missing_count = transjakarta_df.filter(col(column).isNull()).count()
    print(f"{column}: {missing_count}")
```

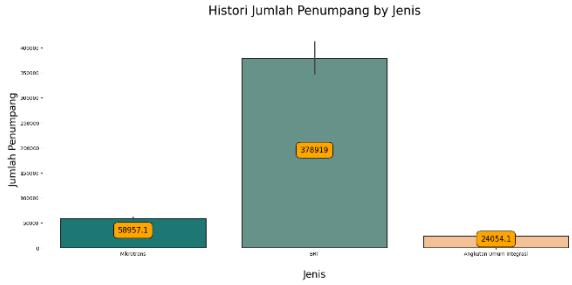
```
tahun: 0
bulan: 0
jenis: 0
kode_trayek: 0
trayek: 0
jumlah_penumpang: 0
```

Gambar 3 Data Yang Sudah Diperbaiki

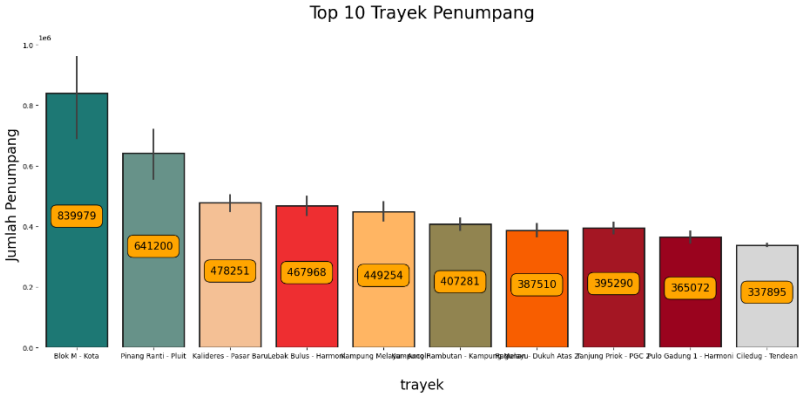
c) Pustaka matplotlib digunakan untuk merepresntasikan data secara visual, supaya memudahkan dalam menganalisa data.
Berikut merupakan grafik visualisai untuk memudahkan analisa



Gambar 4Grafik Histori Jumlah Penumpang



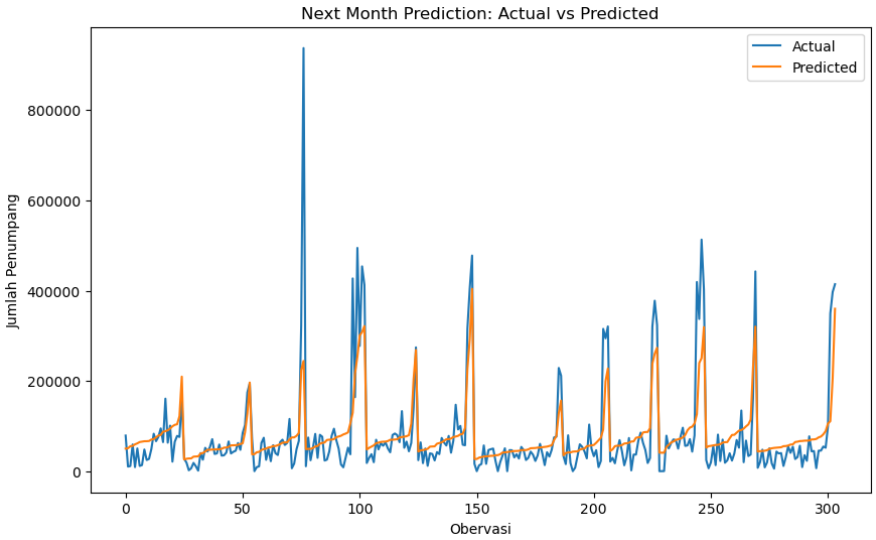
Gambar 5 Grafik Visualisasi Histori Jumlah Penumpang Berdasarkan Jenis



Gambar 6 Grafik Visualisasi Top 10 Trayek Penumpang

2. Analisis Algorima regresi linier.

Implementasi algoritma regresi linier pada dataset mendapatkan visualisai hasil sebagai berikut



Gambar 7 Grafik Hasil Prediksi dan Aktual

3. Interpretasi dan Analisis hasil kinerja algoritma regresi linier berdasarkan nilai MAE, RMSE, dan R-Squared

Berdasarkan algoritma yang telah digunakan, perlu menghitung kinerja dari model itu sendiri, maka perlu acuan untuk melihat apakah model terssebut merupakan model yang sesuai atau tida. MAE, RMSE dan R-squared digunakan sebagai acuan. Nilai dari analisa yang sudah dilakukan adalah sebagai berikut

```
from pyspark.ml.evaluation import RegressionEvaluator
# Evaluasi kinerja model
evaluator = RegressionEvaluator(
    labelCol="label", predictionCol="prediction", metricName="rmse"
)
rmse = evaluator.evaluate(predictions)
print(f"Root Mean Squared Error (RMSE) using Linear Regression: {rmse}")
```

```
Root Mean Squared Error (RMSE) using Linear Regression: 66473.04819569246
```

Gambar 8 Nilai RMSE

```
# Evaluasi kinerja model dengan Mean Absolute Error (MAE)
evaluator_mae = RegressionEvaluator(
    labelCol="label", predictionCol="prediction", metricName="mae"
)
mae = evaluator_mae.evaluate(predictions)
print(f"Mean Absolute Error (MAE) using Linear Regression: {mae}")

# Evaluasi kinerja model dengan R-squared
evaluator_r2 = RegressionEvaluator(
    labelCol="label", predictionCol="prediction", metricName="r2"
)
r2 = evaluator_r2.evaluate(predictions)
print(f"R-squared using Linear Regression: {r2}")
```

```
Mean Absolute Error (MAE) using Linear Regression: 36063.45821652545
R-squared using Linear Regression: 0.6277593819920417
```

Gambar 9 Nilai MAE dan R-Squared

Berdasarkan perhitungan yang sudah dilakukan, didapatkan bahwa nilai RMSE sebesar 66473, MAE 36063, dan R-Squared sebesar 0.627. beberapa faktor penyebab nilai yang tinggi tersebut dikarenakan variasi data yang sulit untuk dipelajari dengan algoritma regresi linear. Nilai dari jumlah penumpang yang memiliki variasi tinggi.

Kesimpulan

Dalam penelitian ini, algoritma Regresi Linier diterapkan untuk Analisa dan Prediksi Transportasi Transjakarta Menggunakan Algoritma Regresi Linear dan Pustaka Pyspark. Berdasarkan penlitian yang sudah dilakukan hasil dari penggunaan algoriotma regresi linier dapat melihat pola dan melakukan prediksi sesuai dengan aktualnya, namun pada perhitungan nilai RMSE sebesar 66473, MAE 36063 dan menunjukan nilai yang tinggi dan dan R-Squared sebesar 0.627, ini berarti kinerja model masih kurang baik, bisa disebabkan karena variasi data yang sulit untuk dipelajari dengan algoritma regresi linear dan nilai dari jumlah penumpang yang memiliki variasi tinggi. perlu diperhatikan bahwa analisis ini tidak menunjukkan hubungan sebab-akibat yang pasti. Namun, hasil ini memberikan wawasan berharga dalam analisa dan penggunaan algoritma regresi linear menggunakan pyspark. Meski memiliki keterbatasan, termasuk dataset yang digunakan dan penggunaan algoritma Regresi Linier, diharapkan dapat berkontribusi dan bermanfaat pada penelitian yang akan datang.

Daftar Rujukan

[1] I. Wulansari, “Penyuluhan Keselamatan Transportasi Darat Usia Transisi (Remaja ke Dewasa),” *Alfatina, J. Community Serv.*, vol. 1, no. 1, pp. 17–21, 2021, [Online]. Available: https://journal.inspire-kepri.org/index.php/JoCS.

[2] A. Faradibah and E. Suryani, “Pengembangan Model Simulasi Sistem Dinamik Untuk Transportasi,” vol. 11, no. 28, pp. 67–76, 2019.

[3] N. L. W. Rita Kurniati, “Dampak Ekonomi Pengoperasian Transjakarta Ditinjau dari Persepsi Pengguna,” *J. Penelit. Transp. Darat*, vol. 22, no. 2, pp. 194–205, 2021, doi: 10.25104/jptd.v22i2.1669.

[4] A. T. Sasongko, "Studi Literatur Konsep dan Implementasi Sains Data untuk Memaksimalkan Kinerja Industri Manufaktur," *J. Teknol. Dan Sist. Inf. Bisnis*, vol. 5, no. 2, pp. 90–94, 2023, doi: 10.47233/jteksis.v5i2.778.

[5] M. A. Aditya, R. D. Mulyana, I. P. Eka, and S. R. Widianto, "Penggabungan Teknologi Untuk Analisa Data Berbasis Data Science," *Semin. Nas. Teknol. Komput. Sains*, vol. 7, no. 3, pp. 51–56, 2020.

[6] M. Abdul, R. Wahid, A. Nugroho, and A. H. Anshor, "Prediksi Penyakit Kanker Paru-Paru Dengan Algoritma Regresi Linier," *Bull. Inf. Technol.*, vol. 4, no. 1, pp. 63–74, 2023.

[7] D. Purba and M. Purba, "Aplikasi Analisis Korelasi dan Regresi menggunakan Pearson Product Moment dan Simple Linear Regression," *Citra Sains Teknol.*, vol. 1, no. 2, pp. 97–103, 2022.

[8] A. D. Sidik and A. Ansawarman, "Prediksi Jumlah Kendaraan Bermotor Menggunakan Machine Learning," *Formosa J. Multidiscip. Res.*, vol. 1, no. 3, pp. 559–568, 2022, doi: 10.55927/fjmr.v1i3.745.

[9] R. Novita, I. Yani, and G. Ali, "Sistem Prediksi untuk Penentuan Jumlah Pemesanan Obat Menggunakan Regresi Linier," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 2, no. 1, pp. 62–70, 2022, doi: 10.57152/malcom.v2i1.198.

[10] F. O. Lusiana, I. Fatma, and A. P. Windarto, "Estimasi Laju Pertumbuhan Penduduk Menggunakan Metode Regresi Linier Berganda Pada BPS Simalungun," *J. Informatics Manag. Inf. Technol.*, vol. 1, no. 2, pp. 79–84, 2021, doi: 10.47065/jimat.v1i2.104.

Project : https://github.com/Riel77/data-science