Journal of Digital Ecosystem for Natural Sustainability (JoDENS)
Vol.4, No.2, Desember 2024, pp. 100~104
e-ISSN: 2798-6179
DOI https://doi.org/10.63643/jodens.v4i2.258

Implementasi Regresi Linear Berganda dalam Prediksi Waktu Tiba Bus Transjakarta Menuju Destinasi Akhir

Duandi [1*], Vivie Triyanti [2], Steven Tang [3]
[1,2,3] Teknik Perangkat Lunak, Universitas Universal
**Corresponding author E-mail: duandiduan22@email.com*

Article Info

Article history:

Received 29-11-2024
Revised 10-12-2024
Accepted 27-12-2024

Keyword:

Data Mining, Regresi Linear Berganda, Transjakarta

ABSTRACT

Transjakarta, the main public transportation system in Jakarta, plays a strategic role in meeting the growing mobility needs of urban residents. This study aims to apply multiple linear regression to predict bus arrival times at the final destination using simulation transaction data that includes tap-in and tap-out information, bus routes, travel time, as well as additional variables such as traffic conditions, route distance, and operating hours. Multiple Linear Regression is a statistical method used to understand the relationship between two or more independent variables and a dependent variable. In the context of this study, multiple linear regression is used to model bus arrival times based on a number of influencing factors, such as traffic conditions, route distance, and operating hours. This model assumes that changes in the independent variables will affect the bus arrival time linearly, thus allowing for accurate predictions in various scenarios. The analysis results show that the developed predictive model has high accuracy, with traffic conditions and route distance identified as significant factors. The application of this model has the potential to assist Transjakarta operators in improving route management efficiency, reducing passenger wait times, and providing more accurate arrival time estimates through supporting applications. This research also lays the foundation for developing data-driven predictive systems for public transportation management.

I. PENDAHULUAN

TransJakarta merupakan sistem transportasi *Bus Rapid Transit* (BRT) pertama di Asia Tenggara dan Selatan yang memiliki jalur lintasan terpanjang di dunia, yaitu 208 km. Sistem ini dirancang dengan mengadopsi konsep TransMilenio di Bogota, Kolombia. TransJakarta mulai beroperasi secara resmi pada 1 Februari 2004, bertujuan menyediakan moda transportasi massal yang efisien dan terjangkau bagi masyarakat Jakarta [1].

Kemacetan lalu lintas di Jakarta menjadi salah satu masalah utama yang memengaruhi mobilitas warga. Dalam konteks ini, ketepatan waktu keberangkatan dan kedatangan TransJakarta menjadi isu penting, mengingat sistem ini saat ini melayani 244 rute dengan 14 koridor utama serta 8 jenis layanan, yaitu: 51 rute BRT, 61 rute angkutan umum integrasi, 94 rute mikrotrans, 5 rute bus wisata, 1 layanan TransJakarta Cares, 13 rute Royaltrans, 10 rute Transjabodetabek, dan 19 rute menuju kawasan rumah susun. Ketidaksesuaian jadwal perjalanan dengan waktu yang direncanakan dapat memengaruhi kepercayaan pengguna terhadap layanan ini [2][3].

Penelitian terdahulu, seperti *Analisis Model Pemilihan Moda Transportasi Umum Travel dengan Rute Sumbawa-Mataram* dan *Model Bangkitan Transportasi pada Perumahan Korpri Kecamatan Sungai Kunjang Samarinda*, telah berhasil menggunakan metode statistik untuk memprediksi dan menganalisis pola transportasi. Penelitian pertama menggunakan regresi linier untuk menentukan preferensi moda transportasi berdasarkan faktor waktu dan biaya perjalanan, sedangkan penelitian kedua memanfaatkan regresi untuk menganalisis bangkitan perjalanan berdasarkan lokasi perumahan. Kedua penelitian ini menunjukkan potensi regresi linier dalam menganalisis dan memprediksi data transportasi, yang relevan dengan penelitian ini [4][5].

Regresi Linier Berganda adalah metode statistik yang digunakan untuk menganalisis hubungan antara satu variabel terikat dengan dua atau lebih variabel bebas. Metode ini sering diterapkan untuk memprediksi nilai tertentu berdasarkan data yang melibatkan beberapa objek dalam satu periode waktu. Misalnya, penelitian mengenai pengaruh *Capital Adequacy Ratio* (CAR), *Financing to Deposit Ratio* (FDR), Biaya Operasional terhadap Pendapatan Operasional (BOPO), dan *Non-Performing Financing* (NPF) terhadap Profitabilitas (ROA) Bank Syariah di Indonesia pada tahun 2015 menggunakan regresi linier berganda untuk menganalisis hubungan sebab akibat. Dengan kemampuan prediksi yang dimiliki, regresi linier berganda menjadi metode yang sesuai untuk diterapkan pada konteks transportasi seperti TransJakarta [6].

Penelitian ini bertujuan untuk memprediksi waktu perjalanan bus TransJakarta menuju destinasi akhir. Dengan menggunakan input seperti stasiun awal, waktu keberangkatan, dan stasiun tujuan, penelitian ini diharapkan dapat memberikan estimasi waktu kedatangan yang lebih akurat sehingga dapat meningkatkan kenyamanan dan kepercayaan pengguna layanan TransJakarta.

II. METODE

Penelitian ini menggunakan metode regresi linier berganda dengan tujuan memprediksi waktu perjalanan antar rute bus TransJakarta berdasarkan data historis, sehingga dapat memberikan estimasi waktu tiba ke destinasi secara lebih akurat.

A. *Pengumpulan Data*

Data yang digunakan dalam penelitian ini merupakan data yang didapatkan dari https://www.kaggle.com/datasets/dikisahkan/transjakarta-transportation-transaction/data. Data tersebut dipublish oleh dengan nama Diki Renanda pada bulan April 2023. Datasetnya tersusun atas 23 nama variabel dan 189500 data transaksi bus dengan sumber dari *website* https://ppid.transjakarta.co.id/pusat-data/data-terbuka/transjakarta-gtfs-feed [7].

B. *Data Preprocessing*

Pada tahap ini, untuk mengurangi waktu komputasi, dipilih secara acak 2.000 data sebagai data latih dan 2.000 data sebagai data uji dari total dataset yang tersedia, yang kemudian akan digunakan untuk pengujian model. Setelah itu, dilakukan pembersihan data dengan fokus pada penanganan nilai kosong (*missing values*) pada variabel yang teridentifikasi. Proses pembersihan ini bertujuan agar data yang digunakan dalam analisis lebih konsisten dan dapat dipercaya. Namun, pada penelitian ini, tidak dilakukan pembersihan data lebih lanjut, seperti penghapusan *outlier* atau imputasi data, karena fokus utama adalah pada penanganan nilai kosong saja.

Dari total 22 variabel yang tersedia dalam dataset, empat variabel utama yang dipilih untuk analisis regresi linear berganda adalah sebagai berikut:

Tabel 1. Keterangan variabel

Nama Variabel	Keterangan
TapInTime	Waktu awal keberangkatan
TapInStopsName	Nama halte keberangkatan
TapOutTime	Waktu sampai tujuan
TapOutStopsName	Nama halte tujuan

Dari keempat variabel tersebut, variabel yang akan diprediksi adalah *TapOutTime*, yaitu waktu sampai ke halte tujuan. Proses pemilihan variabel ini dilakukan dengan mempertimbangkan relevansi dan keterkaitan variabel tersebut dengan tujuan penelitian, yaitu untuk memprediksi waktu kedatangan bus di halte tujuan berdasarkan waktu keberangkatan dan lokasi awal keberangkatan bus [8].

C. *Pemrograman R*

R adalah bahasa pemrograman yang sering digunakan dalam penelitian ilmiah, analisis statistik, serta visualisasi data. Dalam konteks analisis regresi linear berganda, R menawarkan berbagai fitur unggulan yang memungkinkan peneliti untuk menangani data dalam skala besar secara efisien. Salah satu kekuatan utama R adalah kemampuannya untuk melakukan pembersihan dan transformasi data, serta membangun model statistik yang kompleks menggunakan paket-paket khusus.

Penggunaan R dalam analisis regresi linier memungkinkan peneliti untuk memodelkan hubungan antara variabel-variabel independen dan dependen, serta mengukur kekuatan hubungan ini dengan menghitung koefisien determinasi (R^2). Selain itu, R menyediakan alat untuk mengevaluasi validitas model, misalnya melalui analisis multikolinearitas untuk memeriksa apakah terdapat hubungan linier yang terlalu kuat antara variabel-variabel independen.

Berbagai pustaka yang tersedia di R memungkinkan pengguna untuk mempercepat analisis data dengan efisien, menghemat waktu, dan meningkatkan akurasi model prediktif [9][10].

D. *Metode Regresi Linear Berganda*

Regresi linier merupakan satu cara prediksi yang menggunakan garis lurus untuk menggambarkan hubungan diantara dua variabel atau lebih. Variabel adalah besaran yang berubah-ubah nilainya. Selanjutnya variabel tersebut terbagi atas dua jenis yaitu variabel pemberi pengaruh dan variabel terpengaruh regresi linier berganda

merupakan model persamaan yang menjelaskan hubungan satu variabel tak bebas/*response* (Y) dengan dua atau lebih variabel bebas/*predictor* (X1, X2, ..., Xn). memprediksi nilai variabel tak bebas/ response (Y) apabila nilai-nilai variabel bebasnya/*predictor* (X1, X2, ..., Xn) diketahui. Disamping itu juga untuk dapat mengetahui bagaimanakah arah hubungan variabel tak bebas dengan variabel-variabel bebasnya [10].

Persamaan dari regresi linear berganda adalah sebagai berikut

$$Y = a + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Y = variabel tak bebas (nilai variabel yang akan diprediksi)
a = konstanta
β = nilai koefisien regresi
x = variabel bebas

E. Evaluasi Prediksi

Evaluasi model regresi linear berganda dilakukan untuk menilai seberapa baik model mampu memprediksi nilai variabel dependen (Y) berdasarkan variabel-variabel independen (X). Dua metrik utama yang sering digunakan adalah *Root Mean Square Error* (RMSE) dan Koefisien Determinasi (R^2). RMSE digunakan untuk mengukur rata-rata kesalahan prediksi dalam satuan yang sama dengan data, sementara R^2 mengukur seberapa baik model dapat menjelaskan variasi data [12][13].

1. *Root Mean Square Error* (RMSE)

RMSE adalah ukuran kesalahan prediksi rata-rata dalam skala asli variabel dependen. Dalam regresi linear berganda, RMSE dihitung untuk mengetahui seberapa jauh prediksi model ($\hat{Y}$) berbeda dari nilai aktual (Y). Nilai RMSE yang lebih kecil menunjukkan model yang lebih akurat.

Rumus RMSE:

$$RMSE = \sqrt{\frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

y_i = Nilai aktual data ke-i
$\hat{y}_i$ = Nilai prediksi data ke-i
n = Jumlah total observasi

RMSE memberikan informasi yang intuitif karena hasilnya berada dalam satuan yang sama dengan variabel dependen. Namun, metrik ini sensitif terhadap outlier, sehingga nilai yang besar bisa mengindikasikan adanya data ekstrem.

2. Koefisien Determinasi (R^2)

Koefisien determinasi (R^2) mengukur proporsi variabilitas variabel dependen (Y) yang dapat dijelaskan oleh kombinasi variabel independen (X). Nilai R^2 berada dalam rentang 0 hingga 1:

- $R^2 = 1$: Model menjelaskan semua variabilitas dalam data.
- $R^2 = 0$: Model tidak menjelaskan variabilitas data sama sekali.

Rumus R^2:

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \underline{y})^2}$$

$\Sigma(y_i - \hat{y}_i)^2$: *Residual Sum of Square (RSS)*, kesalahan prediksi
$\Sigma(y_i - \underline{y})^2$: *Total Sum of Squares (TSS)*, total variabilitas data aktual.

III. HASIL DAN PEMBAHASAN

Import Dataset

Tahap awal penelitian dimulai dengan mengimpor dataset yang diperoleh dari *Kaggle* dalam format CSV. Dataset yang digunakan berisi informasi tentang tap-in dan tap-out pada sistem Transjakarta. Dataset diakses dan dibaca menggunakan fungsi *read.csv* di R.

```
data_aktual <-
read.csv('/content/dfTransjakarta180kRows_aktual.csv')

data_aktual <- data_aktual[, c("tapInStopsName",
"tapInTime", "tapOutStopsName", "tapOutTime")]
```

Dataset kemudian dibersihkan melalui beberapa langkah untuk memastikan kualitas data yang optimal. Langkah pembersihan meliputi:

Menghapus baris yang mengandung nilai kosong (NA) atau nol pada kolom-kolom yang relevan.
Menghapus spasi yang tidak diperlukan pada nama halte.
Memastikan data waktu dikonversi menjadi numerik, dan nama halte dikonversi menjadi faktor.

```
data_aktual <- na.omit(data_aktual)

data_aktual$tapInStopsName <-
trimws(data_aktual$tapInStopsName)
data_aktual$tapOutStopsName <-
trimws(data_aktual$tapOutStopsName)

data_aktual <- data_aktual[data_aktual$tapInTime != 0 &
```

```
data_aktual$tapOutTime != 0, ]

data_aktual <-
data_aktual[trimws(data_aktual$tapInStopsName) != "" &
trimws(data_aktual$tapOutStopsName) != "", ]

data_aktual$tapInTime <- as.numeric(data_aktual$tapInTime)
data_aktual$tapOutTime <-
as.numeric(data_aktual$tapOutTime)

data_aktual$tapInStopsName <-
factor(data_aktual$tapInStopsName)
data_aktual$tapOutStopsName <-
factor(data_aktual$tapOutStopsName)
```

Pemodelan Regresi Linear Berganda

Setelah data selesai dibersihkan,model regresi linear berganda dibangun untuk memprediksi waktu *tap-out* berdasarkan variabel-variabel independen, yaitu *tap-in time*, nama halte *tap-in*, dan nama halte *tap-out*. Model dilatih menggunakan fungsi *lm()*.

```
model <- lm(tapOutTime ~ tapInStopsName + tapInTime +
tapOutStopsName, data = data_aktual)
```

Model ini menghasilkan parameter-parameter yang menggambarkan hubungan antara variabel independen dan dependen, yang nantinya digunakan untuk prediksi.

Prediksi dengan Data Uji

Dataset uji diolah menggunakan prosedur yang sama seperti dataset pelatihan. Dataset ini mencakup langkah pembersihan dan konversi data ke format yang sesuai.

```
datauji <-
read.csv('/content/dfTransjakarta180kRows_datauji.csv')

datauji <- datauji[, c("tapInStopsName", "tapInTime",
"tapOutStopsName", "tapOutTime")]

datauji$tapInStopsName <- trimws(datauji$tapInStopsName)
datauji$tapOutStopsName <-
trimws(datauji$tapOutStopsName)

datauji <- na.omit(datauji)

datauji$tapInTime <- as.numeric(datauji$tapInTime)
datauji$tapOutTime <- as.numeric(datauji$tapOutTime)

datauji <- datauji[datauji$tapInTime != 0 & datauji$tapOutTime
!= 0, ]

datauji$tapInStopsName <- factor(datauji$tapInStopsName)
datauji$tapOutStopsName <- factor(datauji$tapOutStopsName)
```

Dengan model yang sudah terlatih, kita bisa melakukan prediksi untuk data uji dengan *predict()*.

```
predictions <- predict(model, newdata = datauji)
```

Evaluasi Model

Evaluasi model dilakukan menggunakan dua metrik utama:

***Root Mean Squared Error* (RMSE):** Mengukur kesalahan prediksi rata-rata dalam satuan yang sama dengan variabel dependen.

R-squared **(R^2):** Menilai seberapa baik model dapat menjelaskan variabilitas data.

Berikut adalah penghitungan kedua metrik:

```
RMSE <- sqrt(mean((predictions - datauji$tapOutTime)^2))
cat("Root Mean Squared Error (RMSE):", RMSE, "\n")

r_squared <- summary(model)$r.squared
cat("R-squared:", r_squared, "\n")
```

Prediksi dengan Input Pengguna

Bagian berikutnya adalah fitur yang memungkinkan pengguna untuk memasukkan data secara langsung dan memperoleh prediksi waktu *tap-out*. Pengguna diminta untuk memasukkan nama halte *tap-in*, waktu *tap-in* dalam format jam:menit, dan nama halte *tap-out*.

```
convert_seconds_to_time <- function(seconds) {
  hours <- floor(seconds / 3600)
  minutes <- floor((seconds %% 3600) / 60)
  return(sprintf("%02d:%02d", hours, minutes))
}

convert_time_string_to_seconds <- function(time_string) {
  time_parts <- unlist(strsplit(time_string, ":"))
  hours <- as.numeric(time_parts[1])
  minutes <- as.numeric(time_parts[2])
  total_seconds <- (hours * 3600) + (minutes * 60)
  return(total_seconds)
}

cat("Masukkan nama tapInStopsName: ")
tapInStopsName_input <- readline()
cat("Masukkan waktu tapInTime format (HH:MM): ")
tapInTime_input <- readline()
cat("Masukkan nama tapOutStopsName: ")
tapOutStopsName_input <- readline()

new_data <- data.frame(
  tapInStopsName = factor(tapInStopsName_input, levels =
levels(data_aktual$tapInStopsName)),
  tapInTime =
convert_time_string_to_seconds(tapInTime_input),
  tapOutStopsName = factor(tapOutStopsName_input, levels
= levels(data_aktual$tapOutStopsName))
)

predicted_tapOutTime <- predict(model, newdata =
new_data)
predicted_tapOutTime_time <-
convert_seconds_to_time(predicted_tapOutTime)
cat("Predicted tapOutTime (as hour):",
predicted_tapOutTime_time, "\n")
```

Fitur ini memberikan kemudahan bagi pengguna untuk melihat estimasi waktu perjalanan berdasarkan input waktu dan halte yang mereka pilih. Jika input valid, sistem akan memberikan prediksi waktu *tap-out* dalam format jam:menit.

```
Masukkan nama tapInStopsName: Polsek Makasar
Masukkan waktu tapInTime format (HH:MM): 17:14
Masukkan nama tapOutStopsName: Sbr. Jln. Man 2
Predicted tapOutTime (as hour): 18:57
```

Hasil Prediksi dan Pembahasan

Dari hasil model regresi linear, kami memperoleh nilai RMSE sebesar 464.1927 dan *R-squared* sebesar 0.9997828 Ini menunjukkan bahwa model memiliki akurasi tinggi dalam memprediksi waktu keluar berdasarkan variabel input. Meskipun model ini memberikan prediksi yang cukup baik, masih ada ruang untuk perbaikan, terutama dalam menangani variabilitas data yang lebih besar.

Untuk kasus prediksi individual, pengguna dapat memasukkan nama halte dan waktu *tap-in* dalam format yang telah disediakan, dan model akan memprediksi waktu *tap-out* sesuai dengan input tersebut.

```
Root Mean Squared Error (RMSE): 464.1927
R-squared: 0.9997828
```

IV. KESIMPULAN

Penelitian ini berhasil membangun model regresi linier berganda untuk memprediksi waktu kedatangan bus TransJakarta berdasarkan variabel nama halte keberangkatan, waktu keberangkatan, dan nama halte tujuan. Evaluasi model menunjukkan nilai *Root Mean Squared Error* (RMSE) sebesar 464,19 detik, yang berarti rata-rata selisih atau perbedaan antara prediksi model dan data aktual adalah sekitar 464 detik (7 menit 44 detik).

Selain itu, nilai koefisien determinasi (R^2) sebesar 0,9998 menunjukkan bahwa model memiliki tingkat akurasi yang sangat tinggi. Nilai ini mengindikasikan bahwa model mampu menjelaskan sekitar 99,98% dari variabilitas data aktual, sehingga hanya 0,02% variabilitas data yang tidak dapat dijelaskan oleh model.

Meskipun demikian, terdapat peluang untuk meningkatkan kinerja model lebih lanjut, terutama jika model diterapkan pada data dengan variabilitas yang lebih besar atau data di luar cakupan dataset pelatihan. Fitur prediksi interaktif yang dikembangkan juga memberikan kemudahan bagi pengguna untuk memperkirakan waktu kedatangan bus berdasarkan input waktu dan halte, yang diharapkan dapat meningkatkan kenyamanan dan kepercayaan pengguna terhadap layanan TransJakarta.

DAFTAR PUSTAKA

[1] Dudung KS, “Transjakarta || Tentang,” *Transjakarta.co.id*, 2024. https://transjakarta.co.id/tentang/sejarah

[2] H. B. Alexander, “20 Tahun Transjakarta, Keberhasilan Sistem BRT Terpanjang di Dunia Halaman all - Kompas.com,” *KOMPAS.com*, Jan. 14, 2024. https://lestari.kompas.com/read/2024/01/14/160000586/20-tahun-transjakarta-keberhasilan-sistem-brt-terpanjang-di-dunia?page=all (accessed Nov. 22, 2024).

[3] Antara and Iqbal Muhtarom, “TomTom Traffic Index Tunjukkan Tingkat Kemacetan Jakarta Kian Memburuk,” *Tempo*, Apr. 06, 2023. https://www.tempo.co/arsip/tomtom-traffic-index-tunjukkan-tingkat-kemacetan-jakarta-kian-memburuk--200673 (accessed Nov. 22, 2024).

[4] Muhammad Jazir Alkas, “MODEL BANGKITAN TRANSPORTASI PADA PERUMAHAN KORPRI KECAMATAN SUNGAI KUNJANG SAMARINDA,” *MODEL BANGKITAN TRANSPORTASI PADA PERUMAHAN KORPRI KECAMATAN SUNGAI KUNJANG SAMARINDA*, vol. 2, no. 1, Apr. 2019.

[5] Erlina Damayanti and Eti Kurniati, “ANALISIS MODEL PEMILIHAN MODA TRANSPORTASI UMUM TRAVEL DENGAN RUTE SUMBAWA-MATARAM,” *Hexagon*, vol. 4, no. 2, pp. 96–113, 2023, doi: https://doi.org/10.36761/hexagon.v4i2.3257.

[6] Felinda Arumningtyas, “Regresi Linear Berganda Dan Regresi Data Panel, Ini Dia Cara Membedakannya! - Exsight,” *Exsight*, Mar. 31, 2022. https://exsight.id/blog/2022/03/31/perbedaan-regresi-linear-berganda-panel/ (accessed Nov. 22, 2024).

[7] dikisahkan, “Transjakarta - Public Transportation Transaction,” *Kaggle.com*, 2023. https://www.kaggle.com/datasets/dikisahkan/transjakarta-transportation-transaction/data (accessed Nov. 22, 2024).

[8] Techietory, “Data Preprocessing Techniques: Essential Data Cleaning Methods and Tools,” *Techietory.com*, Sep. 11, 2024. https://techietory.com/data-science/data-preprocessing-techniques-data-cleaning/ (accessed Nov. 22, 2024).

[9] Anggie Irfansyah, “Executive Class Pengelolaan Keamanan Informasi,” *Eduparx Blog*, Oct. 07, 2022. https://eduparx.id/blog/insight/r-programming-untuk-analisis-dan-visualisasi-data/ (accessed Nov. 22, 2024).

[10] R Core Team, “R: A Language and Environment for Statistical Computing,” *R Foundation for Statistical Computing*, 2024. https://www.r-project.org/

[11] W. A. L. Sinaga, S. Sumarno, and I. P. Sari, “The Application of Multiple Linear Regression Method for Population Estimation Gunung Malela District,” *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, vol. 1, no. 1, pp. 55–64, Mar. 2022, doi: https://doi.org/10.55123/jomlai.v1i1.143.

[12] Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice*. Available: https://otexts.com/fpp3/

[13] D. R. Legates and G. J. McCabe, “Evaluating the use of ‘goodness-of-fit’ Measures in hydrologic and hydroclimatic model validation,” *Water Resources Research*, vol. 35, no. 1, pp. 233–241, Jan. 1999, doi: https://doi.org/10.1029/1998wr900018.