

A COMPARISON OF BBKA STOCK PRICE DIRECTION CLASSIFICATION USING LOGISTIC REGRESSION AND LINEAR DISCRIMINATION ANALYSIS

KOMPARASI KLASIFIKASI ARAH HARGA SAHAM BBKA DENGAN MENGGUNAKAN REGRESI LOGISTIK DAN *LINEAR DISCRIMINATION ANALYSIS*

Deanisa Syifa Azzahra^{1†}, Yasmin Alya Rahma², Daffa Al'bethlin³, Muhammad Farid Mufid Thorif⁴, and Ruhiyat⁵

^{1,2,3,4,5}Division of Economic, Financial, and Actuarial Mathematics, School of Data Science, Mathematics, and Informatics, IPB University, Indonesia

[†]corresponding author: deanisa.azzahra@gmail.com

Copyright © 2026 Deanisa Syifa Azzahra, Yasmin Alya Rahma, Daffa Al'bethlin, Muhammad Farid Mufid Thorif, and Ruhiyat. This is an open-access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

The volatile stock market requires investors to have the ability to accurately predict the direction of stock price movements, particularly for prime blue-chip stocks such as PT Bank Central Asia Tbk (BBKA). This study aims to compare the classification performance of the binomial logistic regression model (through the generalized linear model/GLM approach) and linear discriminant analysis (LDA) in predicting the direction of BBKA stock prices (up/down). The predictor variables evaluated include the JISDOR exchange rate, inflation, BI-Rate, transaction volume, and lag 1 to lag 5. To avoid overfitting and maintain the sequential structure of the data, the model's performance was tested using the time-series cross-validation (TSCV) technique with an expanding window scheme. The modeling results indicate that only the lag 1 variable (the previous day's return) has a statistically significant effect on the direction of stock prices. Based on the evaluation metrics, the logistic regression model demonstrated slightly superior performance with an accuracy rate of 57.143% and an area under the curve (AUC) of 58.2%. On the other hand, the LDA model yielded an accuracy of 57.143% and an AUC of 58.11%. Comprehensively, both algorithms present relatively equal and competent classification capabilities. The marginal advantage of logistic regression is suspected to occur due to the violation of the normal distribution assumption in the lag 1 data, which is an essential requirement for building the LDA model.

Keywords: Binomial Logistic Regression, Classification, Linear Discriminant Analysis, Stock Price Direction, Time-Series Cross Validation.

Abstrak

Pasar saham yang fluktuatif menuntut investor untuk memiliki kemampuan dalam memprediksi arah pergerakan harga saham secara akurat, khususnya pada saham *blue chip* primadona seperti PT Bank Central Asia Tbk (BBCA). Penelitian ini bertujuan untuk membandingkan performa model klasifikasi regresi logistik binomial (melalui pendekatan *generalized linear model* (GLM) dan *linear discriminant analysis* (LDA) dalam memprediksi arah pergerakan harga saham BBCA (naik/turun). Variabel prediktor yang dievaluasi meliputi kurs JISDOR, inflasi, BI-Rate, volume transaksi, serta *lag* 1 hingga *lag* 5. Evaluasi model dilakukan menggunakan *time-series cross validation* dengan skema *expanding window*. Hasil pemodelan menunjukkan bahwa hanya peubah *lag* 1 (*return* satu hari sebelumnya) yang berpengaruh signifikan secara statistik terhadap arah harga saham. Berdasarkan metrik evaluasi, model regresi logistik menunjukkan performa yang sedikit lebih unggul dengan tingkat akurasi 57.143% dan *area under the curve* (AUC) sebesar 58.2%. Di sisi lain, model LDA menghasilkan akurasi sebesar 57.143% dan AUC sebesar 58.11%. Kedua model menunjukkan performa klasifikasi yang relatif setara. Regresi logistik sedikit lebih unggul karena asumsi kenormalan pada LDA tidak terpenuhi.

Kata Kunci: Arah Harga Saham, Klasifikasi, *Linear Discriminant Analysis*, Regresi Logistik Binomial, *Time-Series Cross Validation*.

1. Pendahuluan

Saham merupakan instrumen pasar keuangan yang paling diminati di Indonesia, dengan PT Bank Central Asia Tbk (BBCA) dari sektor perbankan sebagai salah satu saham *blue chip* primadona berkapitalisasi pasar terbesar (Ramadhan & Pamuji, 2022). Mengingat pasar saham sangat fluktuatif, kemampuan memprediksi arah pergerakan harga saham secara akurat menjadi sangat berharga bagi investor. Arah harga saham dipengaruhi oleh faktor makroekonomi dan data historis seperti kurs *Jakarta Interbank Spot Dollar Rate* (JISDOR), inflasi, BI Rate, volume transaksi, dan *return* dari periode sebelumnya.

Untuk menyelesaikan masalah analisis multivariat ini, metode regresi logistik dengan pendekatan *generalized linear model* (GLM) dan *linear discriminant analysis* (LDA) dapat digunakan. Sementara itu, LDA digunakan untuk membedakan kelompok berdasarkan kombinasi linear variabel prediktor (Yang et al., 2026). Kedua metode ini menawarkan karakteristik pendekatan yang berbeda dalam mengklasifikasikan data

real.

Evaluasi model dilakukan menggunakan teknik *time-series cross validation* (TSCV) dengan skema *expanding window*. Dari pengujian sekuensial tersebut, tingkat kesalahan prediksi diekstraksi melalui metrik *mean squared error* (MSE), di mana nilai galat terkecil menunjukkan akurasi peramalan terbaik (Mustafidah & Rohman, 2023). Melalui pendekatan uji statistik, visualisasi *boxplot*, dan evaluasi TSCV ini, penelitian bertujuan membandingkan akurasi GLM binomial dan LDA untuk menentukan model klasifikasi terbaik dalam memprediksi arah pergerakan harga saham BBKA.

2. Metodologi

2.1 Model Regresi Logistik Binomial

Regresi logistik binomial merupakan metode *supervised learning* yang digunakan untuk mengklasifikasikan dua kemungkinan kejadian berdasarkan peubah prediktor. Regresi logistik menggunakan fungsi logistik untuk memodelkan peluang kejadian biner sehingga peluang yang dihasilkan berada pada interval $[0,1]$ (Starbuck, 2023). Berikut adalah persamaan untuk model regresi logistik berganda.

$$\log\left(\frac{p(X)}{1-p(x)}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p \quad (1)$$

Di mana $p(X)$ adalah peluang kejadian biner, $\beta_0, \beta_1, \dots, \beta_p$ adalah koefisien peubah dan $X_1, \dots, X_p$ adalah peubah prediktor.

2.2 Model *Linear Discrimination Analysis*

Linear discriminant analysis (LDA) merupakan metode klasifikasi parametrik yang mengasumsikan data menyebar normal multivariat dan memiliki matriks kovarian yang sama antarkelompok.

2.3 Ukuran Performa Model

2.3.1 Matriks Klasifikasi

Matriks klasifikasi, yang sering disebut sebagai *confusion matrix*, adalah tabel evaluasi yang memetakan perbandingan antara nilai prediksi model dengan nilai aktual pada data uji. Struktur matriks ini terdiri dari empat indikator fundamental, yaitu *true positive* (TP), *false negative* (FN), *false positive* (FP), dan *true negative* (TN) (Rashad et al., 2023). Performa model dievaluasi menggunakan *confusion matriks* melalui metrik akurasi. Berdasarkan elemen-elemen tersebut, kualitas

model dapat diukur secara kuantitatif melalui beberapa metrik utama. Akurasi digunakan untuk melihat proporsi seluruh prediksi benar terhadap total data dengan menggunakan formula:

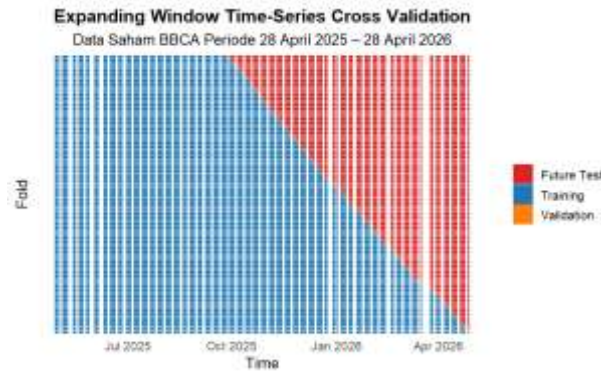
$$\frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

2.3.2 ROC dan AUC

Receiver operating characteristic (ROC) dan *area under the ROC curve* (AUC) digunakan untuk mengevaluasi kemampuan model dalam membedakan kedua kelas. Nilai AUC yang semakin mendekati 1 mengindikasikan performa diskriminasi model yang semakin baik (Carrington et al., 2023). Secara fungsional, penggunaan ROC dan AUC sangat efektif untuk membandingkan berbagai algoritma klasifikasi secara objektif karena hasil pengukurannya bersifat menyeluruh dan tidak bergantung pada pemilihan satu nilai *threshold* tertentu.

2.4 Time Series Cross-Validation

Time series cross-validation (TSCV) adalah metode validasi model pada data deret waktu yang mempertahankan sifat sekuensial data, yaitu menggunakan pengamatan masa lalu untuk memprediksi nilai pada periode berikutnya sehingga dapat menghindari *overfitting* dan meningkatkan kemampuan generalisasi model terhadap data pada waktu yang berbeda (Widodo & Santoso, 2026). Evaluasi performa model dilakukan menggunakan skema *rolling forecasting origin* dengan metode *expanding window*, yaitu data *training* diperluas secara bertahap dengan menambahkan data historis pada setiap iterasi sementara data *testing* diambil dari periode berikutnya (Widodo & Santoso, 2026). Pada iterasi pertama digunakan 100 data pertama sebagai data *training* dan 1 data berikutnya sebagai data *validation/testing*. Pada iterasi selanjutnya ukuran data *training* terus bertambah hingga observasi terakhir.



Gambar 1: Skema validasi *expanding window* pada *time-series cross-validation*

Gambar 1 menunjukkan proses *expanding window* pada data saham BBCA periode 28 April 2025 hingga 28 April 2026. Area biru menunjukkan data *training* yang terus bertambah pada setiap iterasi, area oranye menunjukkan data *validation/testing*, sedangkan area merah menunjukkan data yang belum digunakan pada iterasi tersebut.

2.5 Masalah Potensial

Masalah potensial biasanya cukup memberikan distraksi pada data dalam melakukan pemodelan. Beberapa masalah potensial tersebut adalah multikolinearitas, autokorelasi, dan pencilan. Multikolinearitas adalah kondisi ketika terdapat hubungan linear yang kuat antar dua atau lebih variabel prediktor dalam suatu model regresi. Multikolinearitas dideteksi menggunakan *variance inflation factor* (VIF) dengan nilai VIF melebihi 10 (atau 5 dalam kriteria yang lebih ketat) dianggap bermasalah (Gujarati & Porter, 2009). Kemudian, autokorelasi adalah kondisi ketika residual (kesalahan) dari suatu model regresi berkorelasi dengan residual pada periode waktu sebelumnya. Autokorelasi dideteksi menggunakan uji Durbin-Watson dengan nilai statistik uji d yang berkisar antara 0 hingga 4, dimana nilai d mendekati 0 mengindikasikan autokorelasi positif, nilai d mendekati 4 mengindikasikan autokorelasi negatif, dan nilai $d \approx 2$ mengindikasikan tidak adanya autokorelasi. Terakhir, pencilan (*outlier*) adalah observasi yang memiliki nilai ekstrem atau sangat berbeda dibandingkan dengan sebagian besar data lainnya. Deteksi *outlier* dapat dilakukan melalui metode *studentized residual*, yaitu residual yang dibagi dengan standar *error* nya, di mana nilai absolut yang melebihi 2 atau 3 mengindikasikan potensi adanya *outlier*.

2.6 Uji Shapiro-Wilk

Uji Shapiro-Wilk digunakan untuk menguji hipotesis apakah suatu data menyebar (H_0) normal atau tidak (H_1). Nilai statistik W yang mendekati 1 menunjukkan data cenderung menyebar normal. Dalam pengujian hipotesis, W dibandingkan dengan nilai kritis, jika nilai W kurang dari nilai kritis atau p -value kurang dari α , maka tolak H_0 sehingga data tidak menyebar normal (Ahadi & Zain, 2023).

2.7 Tahapan Analisis Data dan Pemodelan

Terdapat langkah-langkah yang dilakukan dalam membandingkan kedua model klasifikasi yaitu sebagai berikut.

- 1.) Mengumpulkan data,
- 2.) Melakukan eksplorasi data menggunakan *boxplot*,
- 3.) Membentuk model regresi logistik atau *generalized linear model* (GLM) binomial menggunakan variabel makroekonomi dan lag return sebagai prediktor. Dalam model regresi logistik, dapat dilakukan identifikasi peubah yang signifikan berdasarkan nilai p -value < 0.05 ,
- 4.) Melakukan pemodelan menggunakan regresi logistik dan *linear discriminant analysis* (LDA) dengan pendekatan *time-series cross validation*,
- 5.) Menghitung nilai akurasi serta membuat plot kurva ROC dan menghitung AUC sebagai evaluasi kinerja model dan interpretasikan hasil akhirnya.

3. Hasil dan Pembahasan

Dalam studi komparasi ini, data yang digunakan merupakan data sekunder meliputi data harga saham harian BBKA dengan rentang waktu 17 April 2025 hingga 28 April 2026 diperoleh dari (Fusion Media Limited, 2026). Pada periode yang sama, data nilai kurs JISDOR dan data BI-Rate diperoleh dari (Bank Indonesia, 2026). Kemudian, data tingkat inflasi bulanan pada bulan April 2025 sampai Maret 2026 dari (Bank Indonesia, 2026) dan pada bulan April 2026 dari (Trading Economics, 2026). Selanjutnya, *return* saham BBKA dihitung dengan menggunakan persamaan berikut.

$$r = \frac{S_t - S_0}{S_0} \quad (3)$$

Di mana, r adalah *return* saat ini, S_t adalah harga saham saat ini, dan

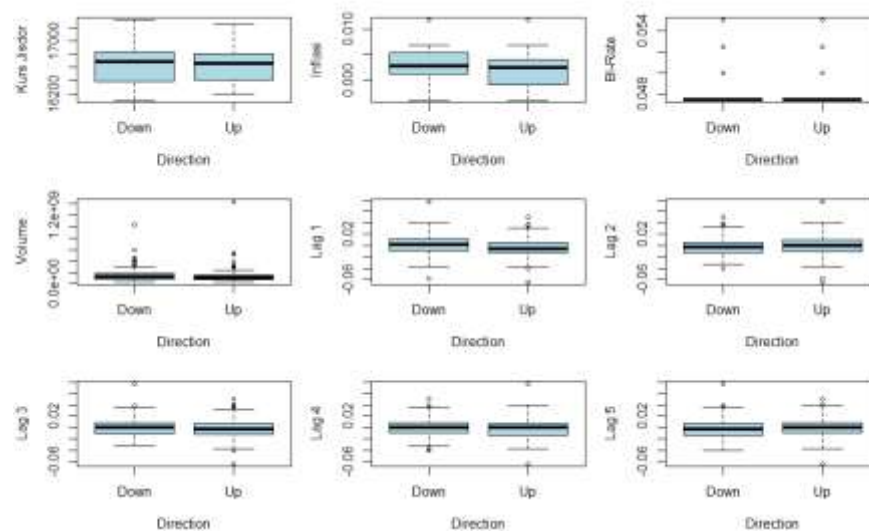
S_0 adalah harga saham sebelumnya. Dapat diketahui bahwa jika $r > 0$ maka harga saham naik dan jika $r < 0$ maka harga saham turun. Selanjutnya, untuk data *lag* ke-1 hingga ke-5 yang merupakan data *return* sebelumnya. Misalkan *lag* ke-1 adalah *return* 1 hari sebelumnya, *lag* ke-2 adalah *return* 2 hari sebelumnya, dan seterusnya. Sebagai tambahan, Tabel 1 berikut memuat keterangan peubah respons dan peubah prediktor yang akan digunakan.

Tabel 1: Keterangan peubah respons dan prediktor

Peubah	Jenis Peubah	Keterangan
Y	Kategorik	Arah harga saham BBKA (0 = <i>down</i> , 1 = <i>up</i>)
X_1	Numerik	Nilai kurs JISDOR
X_2	Numerik	Tingkat inflasi indeks harga konsumen
X_3	Numerik	Suku bunga BI-Rate
X_4	Numerik	Volume transaksi
X_5	Numerik	<i>Lag</i> 1 (<i>return</i> 1 hari sebelumnya)
X_6	Numerik	<i>Lag</i> 2 (<i>return</i> 2 hari sebelumnya)
X_7	Numerik	<i>Lag</i> 3 (<i>return</i> 3 hari sebelumnya)
X_8	Numerik	<i>Lag</i> 4 (<i>return</i> 4 hari sebelumnya)
X_9	Numerik	<i>Lag</i> 5 (<i>return</i> 5 hari sebelumnya)

3.1 Eksplorasi Data dengan Menggunakan *Boxplot*

Sebagai tahap eksplorasi awal, digunakan *boxplot* untuk menganalisis distribusi masing-masing peubah prediktor terhadap peubah *direction*. Analisis ini bertujuan untuk mengidentifikasi seberapa berpengaruh tiap peubah dalam membedakan kedua kelas.



Gambar 2: *Boxplot* variabel prediktor terhadap *direction*

Berdasarkan Gambar 2, terlihat bahwa sebagian besar variabel menunjukkan distribusi yang tumpang tindih antara kelas *up* dan *down*. Namun, variabel inflasi dan *lag 1* memperlihatkan pola pemisahan yang lebih jelas sehingga berpotensi menjadi prediktor yang akan digunakan dalam pemodelan.

3.2 Pemodelan Regresi Logistik dengan Pendekatan *Time-Series Cross Validation*

Model regresi logistik binomial digunakan untuk memodelkan pengaruh peubah lag *return* saham (*lag 1* hingga *lag 5*), kurs, inflasi, BI Rate, dan volume transaksi terhadap arah harga saham BBKA yang bersifat kategorik biner, yaitu naik (*up*) dan turun (*down*). Dengan menggunakan bantuan *software* RStudio, dilakukan pendugaan parameter regresi logistik dan diperoleh hasil seperti pada Tabel 2 berikut.

Tabel 2: Hasil dugaan koefisien dan *p-value* model regresi logistik binomial

Peubah	β	<i>p-value</i>
Kurs JISDOR	-0.000138	0.7982
Inflasi	-75.34	0.037
BI-Rate	75.17	0.2879
Volume Transaksi	-0.000000001206	0.2513

<i>Lag 1</i>	-20.18	0.0189
<i>Lag 2</i>	1.256	0.8768
<i>Lag 3</i>	-11.60	0.1652
<i>Lag 4</i>	-11.40	0.1694
<i>Lag 5</i>	1.420	0.8601

Berdasarkan Tabel 2, hanya peubah inflasi dan *lag 1* yang signifikan pada taraf signifikansi 5% karena memiliki *p-value* kurang dari 0.05, yaitu sebesar 0.037 untuk peubah inflasi dan 0.0189 untuk peubah *lag 1*. Kemudian, berdasarkan pengujian asumsi, hasilnya menunjukkan bahwa tidak terdapat multikolinearitas dan autokorelasi, serta tidak ditemukan outlier yang signifikan meskipun terdapat satu observasi dengan nilai *studentized residual* terbesar sebesar 2.03. Setelah itu, dilakukan pemodelan regresi logistik dengan TSCV yang menghasilkan *true positive* sebesar 41, *true negative* sebesar 39, *false positive* sebesar 32, dan *false negative* sebesar 28 sehingga model regresi logistik memiliki akurasi sebesar 57.143%.

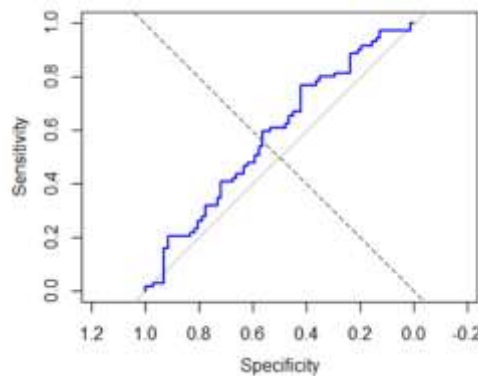
3.3 Pemodelan *Linear Discrimination Analysis* (LDA) dengan Pendekatan *Time-Series Cross Validation*

Berdasarkan pengujian asumsi kenormalan dengan uji Shapiro-Wilk pada data, didapatkan bahwa data inflasi dan *lag 1* tidak menyebar normal karena nilai *p-value* kurang dari 0.05. Namun, data tetap diasumsikan normal sebagai bahan perbandingan terhadap model regresi logistik binomial. Kemudian, didapatkan hasil ringkasan pada LDA di mana proporsi awal pada kelas *down* sebesar 49.58% dan kelas *up* sebesar 50.42%. Hal ini mengindikasikan bahwa pada data awal, proporsi kedua kelas relatif seimbang. Selain itu, rata-rata tingkat inflasi pada kelas *down* sebesar 0.323% dan kelas *up* sebesar 0.242%. Sedangkan, rata-rata *lag 1* pada kelas *down* sebesar 0.099% dan kelas *up* sebesar -0.365%. Hal ini menunjukkan bahwa rata-rata inflasi dan *lag 1* berbeda antar kelas, sehingga berpotensi membedakan arah harga saham. Selanjutnya, koefisien peubah inflasi dan *lag 1* pada fungsi diskriminan berturut-turut ialah sebesar -165.69865 dan -47.12699. Dengan menggunakan fungsi diskriminan tersebut, dapat diklasifikasi suatu kejadian termasuk kelas *down* atau *up* dengan *threshold*. Kemudian, untuk menentukan akurasi model LDA, didapatkan hasil dalam bentuk matriks klasifikasi sebagai berikut dan didapatkan akurasi sebesar 57.143% dengan *true positive*

sebesar 41, *true negative* sebesar 39, *false positive* sebesar 32, dan *false negative* sebesar 28.

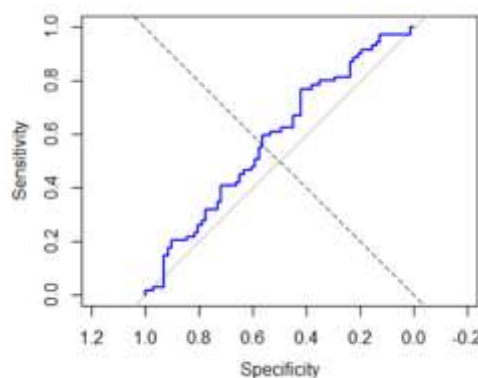
3.4 Perbandingan Kurva ROC dan AUC Model Regresi Logistik dan LDA

Kemampuan model dalam membedakan kedua kelas dievaluasi menggunakan kurva *receiver operating characteristic* (ROC). Berikut adalah kurva ROC yang diperoleh berdasarkan model regresi logistik.



Gambar 3: Plot *receiver operating characteristic* (ROC) pada model regresi logistik

Berdasarkan kurva ROC diperoleh nilai AUC (*area under the curve*) sebesar 58.2%. Gambar 4 adalah kurva ROC pada model LDA.



Gambar 4: Plot *receiver operating characteristic* (ROC) pada model LDA

Pada Gambar 4, didapatkan hasil AUC atau *area under the curve* sebesar 58.11%. Dibandingkan dengan model regresi logistik, nilai AUC pada model LDA sedikit lebih rendah, hal ini dapat disebabkan perbedaan

probabilitas prediksi pada model LDA dan regresi logistik serta asumsi yang tidak terpenuhi pada model LDA di mana data inflasi dan *lag* 1 tidak menyebar normal.

4. Simpulan dan Saran

Melalui estimasi parameter pada model regresi logistik, ditemukan bahwa hanya variabel *lag* 1 yang merupakan peubah yang signifikan. Evaluasi reliabilitas model dengan teknik *time-series cross validation* (TSCV) melalui skema *expanding window* menunjukkan tingkat akurasi regresi logistik mencapai 57.143% dan nilai AUC sebesar 58.2%. Di sisi lain, model LDA menghasilkan akurasi 57.143% dan AUC sebesar 58.11%. Kedua model menunjukkan performa klasifikasi yang relatif setara. Studi ini memiliki batasan pada durasi observasi yang singkat serta deviasi asumsi normalitas peubah inflasi dan *lag* 1 pada metode LDA. Rekomendasi untuk penelitian mendatang mencakup integrasi peubah teknikal seperti *moving average* maupun *relative strength index* (RSI) dan penggunaan data yang memenuhi asumsi kenormalan sebagai prasyarat model LDA.

Daftar Pustaka

- Ahadi, G. D., & Zain, N. N. L. E. (2023). The simulation study of normality test using Kolmogorov-Smirnov, Anderson-Darling, and Shapiro-Wilk. *Eigen Mathematics Journal*, 6(1): 11-19. doi: <https://doi.org/10.29303/emj.v6i1.131>
- Bank Indonesia (2026). Informasi Kurs (JISDOR). Available at: <https://www.bi.go.id/id/statistik/informasi-kurs/jisdor/Default.aspx> (Accessed: 29 April 2026)
- Bank Indonesia (2026). BI-Rate. Available at: <https://www.bi.go.id/id/statistik/indikator/BI-Rate.aspx> (Accessed: 29 April 2026)
- Carrington, A. M., Manuel, D. G., Fieguth, P. W., Ramsay, T., Osmani, V., Wernly, B., Bennett, C., Hawken, S., Magwood, O., Sheikh, Y., McInnes, M., & Holzinger, A. (2023). Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 329-341.
- Fusion Media Limited (2026). Bank Central Asia Tbk Stock Price Today. Available at: <https://www.investing.com/equities/bnk-central-as> (Accessed: 22 April 2026)
- Gujarati, D. N., & Porter, D. C. (2009). *Basic Econometrics* (5th ed.). Boston: McGraw-Hill Irwin.
- Mustafidah, H. & Rohman, S. N. (2023). *Mean square error* pada metode *random* dan *Nguyen Widrow* dalam jaringan syaraf tiruan, *Sainteks*, 20(2):

133-142.

Ramadhan, V. P., & Pamuji, F. Y. (2022). Analisis perbandingan algoritma *forecasting* dalam prediksi harga saham LQ45 PT Bank Mandiri Sekuritas (BMRI). *Jurnal Teknologi Dan Manajemen Informatika*, 8(1): 39–45. doi: <https://doi.org/10.26905/jtmi.v8i1.6092>

Rashad, I., Isnanto, R. R., & Widodo, C. E. (2023). Klasifikasi penyakit jantung menggunakan algoritma analisis diskriminan linier. *Jurnal Sistem Informasi Bisnis*, 13(1): 29–36.

Starbuck, C. (2023). *The Fundamentals of People Analytics: With Applications in R*. Springer Nature.

Trading Economics (2026). Indonesia Inflation Rate MoM. Available at: <https://id.tradingeconomics.com/indonesia/inflation-rate-mom> (Accessed: 4 May 2026)

Widodo, S., & Santoso, B. (2026). Optimizing time series forecasting for AI job market dynamics: a comparative study of SARIMA and Prophet with cross-validation. *International Journal of Informatics and Information Systems*, 9(2): 295–310.

Yang, M. H., Lin, H. R., Liu, C. Y., Chen, L. K., Lee, T. Y., & Chen, K. H. (2026). Using discriminant analysis to predict frailty in community-dwelling older adults in Taiwan. *The Journal of Nursing Research*, 34(1): e431.