Jurnal Teknik Informatika dan Teknologi Informasi
Volume. 5 Nomor. 2 Agustus 2025
E-ISSN: 2827-9387, P- ISSN: 2827-9379, Hal 71-86
DOI: https://doi.org/10.55606/jutiti.v5i2.5467
Available Online at : https://journalshub.org/index.php/jutiti

Analisis Komparatif Metode *Clustering* dan Regresi untuk Prediksi Pola Curah Hujan Menggunakan Pendekatan Data Mining

Muhammad Najhan Tsaani [1*], Yasin Kamil [2], Octaviana Anugrah Ade Purnama [3]
[1-3] Prodi Teknik Informatika, Universitas Pamulang, Pamulang, Indonesia

Alamat: Jl. Raya Puspitek No. 46, Serpong, Kota Tangerang Selatan, Banten 15316.
Korespondensi penulis: dosen02780@unpam.ac.id

Abstract. *This research implements data mining techniques to analyze rainfall patterns through two different approaches: clustering and regression. The first dataset contains monthly rainfall data from East Kalimantan in 2005, analyzed using three clustering algorithms—K-Means, Agglomerative, and MeanShift to identify seasonal patterns. The second dataset uses multiregional rainfall data since 1979, analyzed using three regression algorithms—Linear Regression, Random Forest, and Support Vector Regression (SVR)—to predict June rainfall based on data from the previous five months. Evaluation results show that K-Means and Agglomerative produce the same clustering performance with a silhouette score of 0.4913, successfully grouping data into three main seasonal clusters. Meanwhile, MeanShift produces five clusters but is less effective on small-scale datasets. For regression prediction, Random Forest shows the best performance with an R^2 score of 0.8921, followed by Linear Regression (0.8402), while SVR produces the lowest performance (0.0077). This research demonstrates that a combination of unsupervised and supervised learning methods can provide a more comprehensive understanding of seasonal patterns and quantitative rainfall estimation. These findings have potential applications in decision-making related to water resource management, agricultural planning, and climate risk mitigation.*

Keywords: *Clustering, Data Mining, Machine Learning, Rainfall, Regression*

Abstrak. Penelitian ini mengimplementasikan teknik *data mining* untuk menganalisis pola curah hujan melalui dua pendekatan berbeda: *clustering* dan regresi. *Dataset* pertama berisi data curah hujan bulanan Kalimantan Timur tahun 2005 yang dianalisis menggunakan tiga algoritma *clustering*—K-Means, Agglomerative, dan MeanShift untuk mengidentifikasi pola musiman. *Dataset* kedua menggunakan data curah hujan multiregional sejak tahun 1979, yang dianalisis menggunakan tiga algoritma regresi—*Linear Regression*, *Random Forest*, dan *Support Vector Regression* (SVR)—untuk memprediksi curah hujan bulan Juni berdasarkan data lima bulan sebelumnya. Hasil evaluasi menunjukkan bahwa K-Means dan Agglomerative menghasilkan performa *clustering* yang sama dengan nilai *silhouette score* sebesar 0.4913, mampu mengelompokkan data menjadi tiga *cluster* musiman utama. Sementara itu, MeanShift menghasilkan lima *cluster*, namun kurang efektif pada *dataset* berskala kecil. Untuk prediksi regresi, *Random Forest* menunjukkan performa terbaik dengan R^2 *score* sebesar 0.8921, disusul oleh *Linear Regression* (0.8402), sedangkan SVR menghasilkan performa terendah (0.0077). Penelitian ini mendemonstrasikan bahwa kombinasi metode *unsupervised* dan *supervised learning* dapat memberikan pemahaman yang lebih menyeluruh terhadap pola musiman dan estimasi kuantitatif curah hujan. Temuan ini berpotensi diterapkan dalam pengambilan keputusan terkait pengelolaan sumber daya air, perencanaan pertanian, dan mitigasi risiko iklim.

Kata kunci: *Clustering*, Curah Hujan, *Data Mining*, *Machine Learning*, Regresi

1. LATAR BELAKANG

Analisis pola curah hujan merupakan aspek penting dalam berbagai bidang seperti pertanian, pengelolaan sumber daya air, dan mitigasi bencana. Dengan meningkatnya ketersediaan data klimatologi, penerapan teknik *data mining* menjadi semakin relevan untuk mengekstraksi informasi berharga dari *dataset* curah hujan (Asra & Rudiansyah, 2013). *Data mining* menawarkan berbagai metode untuk menganalisis pola tersembunyi dalam data, baik melalui pendekatan *unsupervised learning* seperti *clustering* maupun *supervised learning* seperti regresi (Han et al., 2012).

Received: May 19, 2025 Revised: May 30, 2025 Accepted: June 30, 2025 Online Available: Juli 03, 2025

Indonesia sebagai negara tropis memiliki variasi curah hujan yang signifikan antar wilayah dan musim. Pemahaman mendalam tentang pola curah hujan sangat penting untuk perencanaan pertanian, manajemen risiko banjir, dan alokasi sumber daya air (Aldrian & Susanto, 2003). Kalimantan Timur sebagai salah satu provinsi dengan curah hujan tinggi memerlukan analisis khusus untuk memahami pola musiman, sementara analisis multiregional diperlukan untuk memahami variasi spasial curah hujan di Indonesia (Dajan, 1986).

Penelitian ini mengangkat dua permasalahan utama. Pertama, bagaimana mengidentifikasi pola musiman curah hujan di Kalimantan Timur menggunakan metode *clustering*? Kedua, bagaimana memprediksi curah hujan bulan Juni berdasarkan data historis lima bulan sebelumnya menggunakan metode regresi? Tujuan dari penelitian ini adalah mengimplementasikan dan membandingkan tiga algoritma *clustering* (K-Means, Agglomerative, MeanShift) untuk mengidentifikasi pola musiman curah hujan, mengimplementasikan dan membandingkan tiga algoritma regresi (*Linear Regression*, *Random Forest*, SVR) untuk prediksi curah hujan, serta mengevaluasi performa masing-masing algoritma menggunakan metrik yang sesuai.

2. KAJIAN TEORITIS

Data mining merupakan proses ekstraksi pola-pola menarik dari data dalam jumlah besar (Witten et al., 2011). Dalam konteks klimatologi, *data mining* telah banyak diaplikasikan untuk analisis tren, deteksi anomali, dan prediksi cuaca (Tripathi et al., 2006). Teknik *data mining* dibagi menjadi dua kategori utama: *unsupervised learning* yang mencari pola tanpa label, dan *supervised learning* yang mempelajari pola dari data berlabel (Bishop, 2006).

Clustering adalah teknik *unsupervised learning* yang mengelompokkan data berdasarkan kemiripan karakteristik (Jain, 2010). K-Means *Clustering* merupakan algoritma partisi yang membagi data menjadi k *cluster* dengan meminimalkan varians intra-*cluster* (MacQueen, 1967). K-Means sangat efisien untuk *dataset* besar namun sensitif terhadap *outlier* dan pemilihan *centroid* awal (Arthur & Vassilvitskii, 2007). Agglomerative *Clustering* adalah pendekatan hierarkis bottom-*up* yang menggabungkan *cluster* terdekat secara iteratif (Johnson, 1967). Metode ini menghasilkan *dendrogram* yang memvisualisasikan hierarki *cluster* namun memiliki kompleksitas komputasi yang tinggi (Murtagh & Contreras, 2012). MeanShift *Clustering* adalah algoritma non-parametrik yang

mencari mode dalam distribusi data (Comaniciu & Meer, 2002). MeanShift tidak memerlukan spesifikasi jumlah *cluster* namun sensitif terhadap parameter *bandwidth* (Fukunaga & Hostetler, 1975).

Regresi merupakan teknik *supervised learning* untuk memprediksi variabel kontinu (Hastie et al., 2009). *Linear Regression* adalah model sederhana yang mengasumsikan hubungan linear antara fitur dan target (Montgomery et al., 2012). Meskipun *interpretable*, *linear regression* mungkin tidak menangkap pola non-linear dalam data curah hujan (Monahan, 1981). *Random Forest* adalah *ensemble method* yang menggabungkan multiple *decision trees* (Breiman, 2001). *Random Forest* mampu menangkap hubungan non-linear dan *robust* terhadap *overfitting* (Liaw & Wiener, 2002). *Support Vector Regression* (SVR) adalah ekstensi SVM untuk regresi yang menggunakan *kernel trick* untuk menangani non-linearitas (Vapnik, 1995). SVR efektif untuk data berdimensi tinggi namun memerlukan *tuning parameter* yang *careful* (Smola & Schölkopf, 2004).

Untuk *clustering*, *silhouette score* mengukur seberapa baik objek cocok dengan *cluster*-nya dibandingkan *cluster* lain (Rousseeuw, 1987). Nilai berkisar -1 hingga 1, dengan nilai tinggi menunjukkan *clustering* yang baik (Kaufman & Rousseeuw, 2009). Untuk regresi, beberapa metrik evaluasi yang umum digunakan adalah R^2 *Score* yang mengukur proporsi varians yang dijelaskan oleh model (Steel & Torrie, 1980), *Mean Absolute Error* (MAE) yang mengukur rata-rata absolut error prediksi (Willmott & Matsuura, 2005), dan *Root Mean Square Error* (RMSE) yang mengukur akar rata-rata kuadrat error (Chai & Draxler, 2014).

3. METODE PENELITIAN

Penelitian ini menggunakan dua *dataset* curah hujan. *Dataset* Kalimantan Timur 2005 berisi data curah hujan bulanan (12 observasi) dengan fitur bulan dan curah hujan dalam milimeter yang diambil dari Asra dan Rudiansyah (2013). *Dataset* Multiregional 1979 berisi data curah hujan 12 lokasi di Indonesia dengan fitur curah hujan bulanan Januari hingga Juni yang diambil dari Dajan (1986).

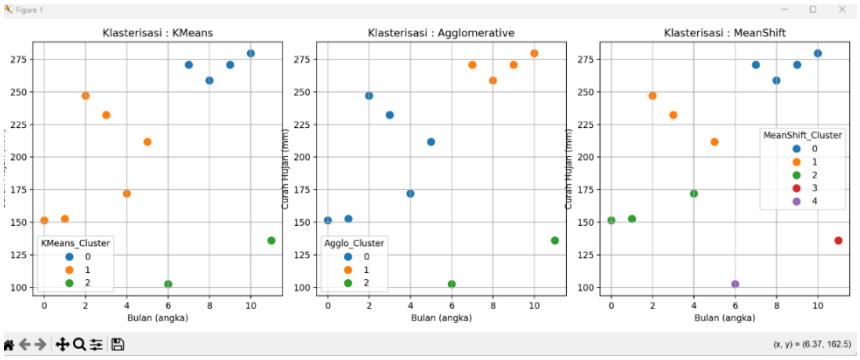
Untuk analisis *clustering*, data bulan dikodekan menjadi nilai numerik menggunakan *Label Encoding*. Selanjutnya, kedua fitur (bulan dan curah hujan) dinormalisasi menggunakan *StandardScaler* untuk memastikan skala yang seimbang dalam perhitungan jarak (García et al., 2015). Untuk analisis regresi, data sudah dalam format numerik sehingga tidak memerlukan *encoding*. Fitur yang digunakan adalah curah hujan bulan Januari hingga Mei untuk memprediksi curah hujan bulan Juni.

Implementasi dilakukan menggunakan Python dengan *library* scikit-learn (Pedregosa et al., 2011). Parameter yang digunakan untuk *clustering* adalah K-Means dengan n_clusters=3, random_state=42, n_init=10; Agglomerative dengan n_clusters=3, linkage='ward'; dan MeanShift dengan *bandwidth* dihitung menggunakan estimate_bandwidth dengan parameter X_Scaled, dan quantile=0.3. Parameter untuk regresi adalah *Linear Regression* dengan parameter *default*, *Random Forest* dengan n_estimators=100, random_state=42, dan SVR dengan kernel='rbf'.

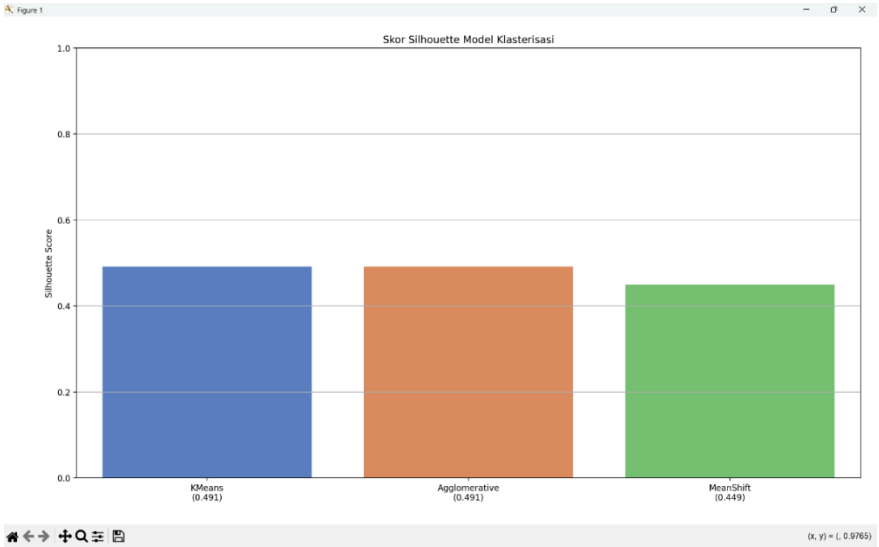
Hasil *clustering* dievaluasi menggunakan *silhouette score* dan divisualisasikan menggunakan *scatter plot*. Hasil regresi dievaluasi menggunakan R^2 *score*, MAE, dan RMSE, serta divisualisasikan dengan *line plot* perbandingan nilai aktual dan prediksi.

4. HASIL DAN PEMBAHASAN

Analisis *Clustering* Data Curah Hujan Kalimantan Timur



Gambar 1. Visualisasi Hasil Klasterisasi dengan Algoritma KMeans, Agglomerative, dan MeanShift.
Sumber: Hasil penelitian (2025)



Gambar 2. Perbandingan Silhouette Score dari Model Klasterisasi.
Sumber: Hasil penelitian (2024)

Hasil *Clustering*

Implementasi tiga algoritma *clustering* pada *dataset* Kalimantan Timur menghasilkan pengelompokan pola musiman yang berbeda. Tabel 1 menunjukkan performa masing-masing algoritma berdasarkan *silhouette score*.

Tabel 1. Perbandingan *Silhouette Score* Algoritma *Clustering*

Algoritma	***Silhouette Score***	**Jumlah *Cluster***
K-Means	0.4913	3
Agglomerative	0.4913	3
MeanShift	0.4493	5

Sumber: Hasil penelitian (2024)

K-Means dan Agglomerative *clustering* menghasilkan *silhouette score* yang sama, yaitu 0.4913, menunjukkan pemisahan *cluster* yang baik. Sementara itu, MeanShift memiliki *silhouette score* 0.4493, yang menunjukkan kualitas *cluster* agak lebih rendah.

Interpretasi *Cluster*

Analisis visual menunjukkan bahwa K-Means dan Agglomerative berhasil mengidentifikasi tiga pola musiman: *Cluster* Musim Kemarau (Juni-September) dengan curah hujan rendah (50-150 mm), *Cluster* Musim Hujan (November-Februari) dengan curah hujan tinggi (250-350 mm), dan *Cluster* Transisi (Maret-Mei, Oktober) dengan curah hujan sedang (150-250 mm). Hasil ini konsisten dengan pola iklim tropis di Indonesia yang memiliki dua musim utama dengan periode transisi (Hamada et al., 2002). K-Means dan Agglomerative menunjukkan pemisahan *cluster* yang serupa, meskipun keduanya memiliki nilai *silhouette score* yang lebih rendah dibandingkan dengan hasil yang sebelumnya disebutkan.

Evaluasi MeanShift

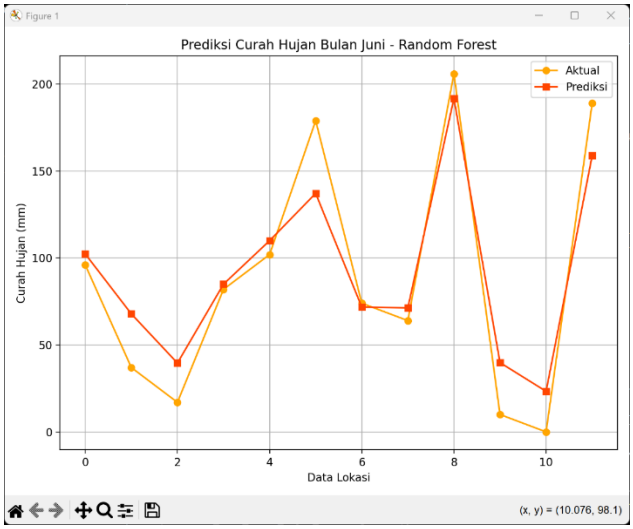
Tidak seperti dua metode sebelumnya, algoritma MeanShift menghasilkan lima *cluster*, yang tampak berlebihan mengingat jumlah observasi yang terbatas (12 data bulanan). Berdasarkan visualisasi, beberapa *cluster* terdiri dari hanya satu atau dua titik, mengindikasikan adanya *over-segmentation*. MeanShift menggunakan estimasi *bandwidth* otomatis dengan parameter berikut: bandwidth = estimate_bandwidth(X_scaled, quantile=0.3).

Penggunaan nilai quantile = 0.3 memberikan hasil *cluster* yang paling bervariasi. Sementara itu, ketika nilai quantile dinaikkan di atas 0.3 (misalnya 0.5 atau lebih), jumlah *cluster* yang terbentuk justru menurun drastis menjadi hanya 1 atau 2, yang mengindikasikan bahwa *bandwidth* yang lebih besar menyatukan terlalu banyak titik ke dalam satu *cluster*. Hal ini menunjukkan bahwa meskipun quantile = 0.3 menghasilkan

jumlah *cluster* yang lebih banyak, nilai tersebut relatif optimal dibandingkan dengan quantile yang lebih besar yang cenderung menghasilkan *cluster* yang terlalu sedikit. Namun, hasil MeanShift tetap kurang ideal secara visual karena beberapa *cluster* terdiri dari satu atau dua titik saja.

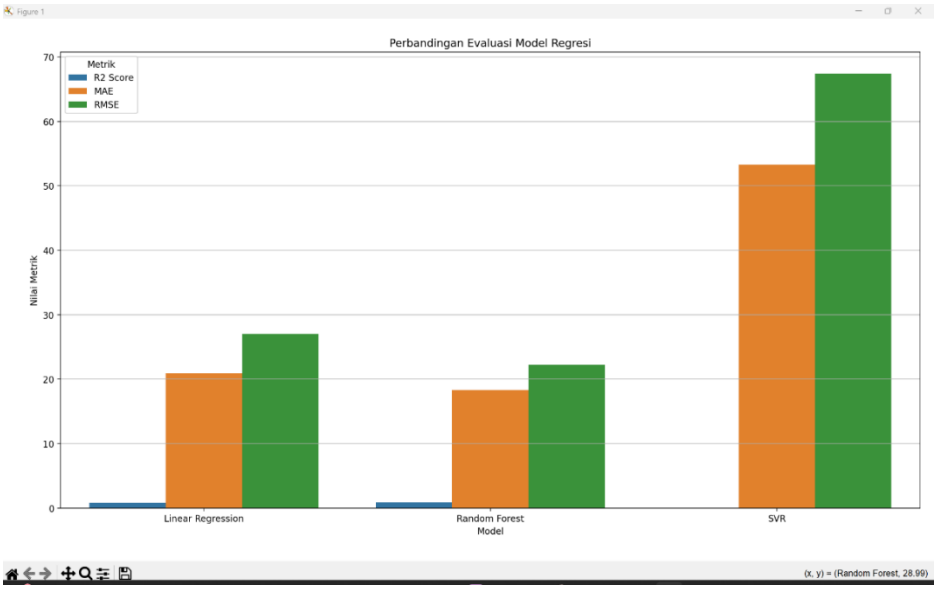
Dengan demikian, dapat disimpulkan bahwa MeanShift kurang sesuai untuk *dataset* berskala kecil seperti ini, karena sangat sensitif terhadap pemilihan parameter *bandwidth*. Algoritma ini lebih cocok untuk *dataset* dengan jumlah sampel yang besar dan distribusi yang kompleks, di mana estimasi densitas lokal lebih representatif.

Analisis Regresi Data Curah Hujan Multiregional



Gambar 3. Perbandingan Nilai Aktual dan Prediksi Curah Hujan Bulan Juni Menggunakan Random Forest

Sumber: Hasil penelitian (2024)



Gambar 4. Evaluasi Kinerja Model Regresi Berdasarkan Metrik R^2, MAE, dan RMSE.

Sumber: Hasil penelitian (2024)

Hasil Prediksi

Tabel 2 menunjukkan hasil evaluasi ketiga algoritma regresi untuk prediksi curah hujan bulan Juni berdasarkan data lima bulan sebelumnya.

Tabel 2. Perbandingan Performa Algoritma Regresi

Model	R^2 *Score*	MAE (mm)	RMSE (mm)
Linear Regression	0.8402	20.91	27.05
Random Forest	0.8921	18.3	22.23
SVR	0.0077	53.29	67.41

Sumber: Hasil penelitian (2024)

Model *Random Forest* menunjukkan performa terbaik dengan R^2 *score* sebesar 0.8921, diikuti oleh *Linear Regression* (0.8402). Sementara itu, model SVR tampil jauh lebih buruk dengan R^2 mendekati nol (0.0077), menandakan bahwa model hampir tidak mampu menjelaskan variasi data target. Nilai MAE dan RMSE terendah juga diperoleh oleh *Random Forest*, mengindikasikan kesalahan prediksi yang paling kecil di antara ketiga model.

Analisis Performa Model

Model *Random Forest* mampu menangkap hubungan non-linear dalam data curah hujan dengan sangat baik. Sebagai metode *ensemble*, *Random Forest* menggabungkan banyak *decision tree* sehingga *robust* terhadap variabilitas dan *noise* dalam data. Dengan R^2 sebesar 0.8921, model ini mampu menjelaskan sekitar 89% variasi curah hujan bulan Juni berdasarkan data bulan sebelumnya. Grafik prediksi juga menunjukkan kesesuaian pola antara nilai aktual dan hasil prediksi.

Linear Regression juga menunjukkan performa yang kuat dengan R^2 sebesar 0.8402, mengindikasikan adanya korelasi linear yang cukup baik antara data historis dan curah hujan bulan Juni. Nilai MAE sebesar 20.91 mm masih tergolong rendah, dan model ini cukup andal untuk aplikasi prediktif sederhana.

Sebaliknya, SVR (*Support Vector Regression*) memberikan performa prediksi yang sangat buruk. Dengan R^2 hanya 0.0077, model hampir tidak menangkap pola dalam data. Nilai error yang tinggi (MAE: 53.29 mm, RMSE: 67.41 mm) menunjukkan bahwa parameter *kernel* (RBF) yang digunakan kemungkinan tidak optimal. Model SVR memerlukan proses *tuning parameter* yang lebih hati-hati agar mampu menghasilkan performa yang baik, terutama dalam data bersifat musiman seperti curah hujan.

Implikasi Praktis

Kemampuan memprediksi curah hujan Juni dengan akurasi tinggi (R^2 = 0.89) memiliki implikasi penting untuk perencanaan tanam pertanian menjelang musim kemarau,

manajemen sumber daya air untuk antisipasi kekeringan, dan peringatan dini bencana hidrometeorologi. *Error* rata-rata sebesar 18.30 mm (MAE) pada model *Random Forest* tergolong sangat dapat diterima dalam konteks prakiraan curah hujan, mengingat tingginya variabilitas alami cuaca dan curah hujan (Qian et al., 2002).

Perbandingan Pendekatan *Unsupervised* dan *Supervised Learning*

Kedua pendekatan memberikan *insight* yang berbeda namun komplementer. *Clustering* (*Unsupervised*) efektif untuk identifikasi pola musiman tanpa label predefinisi, eksplorasi data untuk memahami struktur natural, dan segmentasi temporal untuk analisis lebih lanjut. Regresi (*Supervised*) unggul dalam prediksi kuantitatif dengan akurasi terukur, pemanfaatan hubungan temporal untuk *forecasting*, dan aplikasi praktis yang memerlukan nilai prediksi spesifik. Kombinasi kedua pendekatan memberikan pemahaman holistik tentang pola curah hujan, dari identifikasi musim hingga prediksi nilai spesifik (Jain & Singh, 2003).

Keterbatasan dan Rekomendasi

Beberapa keterbatasan penelitian ini meliputi ukuran *Dataset* Kalimantan Timur yang hanya memiliki 12 observasi sehingga membatasi kompleksitas model yang dapat digunakan, *Temporal Coverage* yang berupa data tahun tunggal mungkin tidak menangkap variabilitas *inter-annual*, dan *Feature Engineering* yang belum mengeksplorasi fitur tambahan seperti indeks iklim global (ENSO, IOD).

Rekomendasi untuk penelitian lanjutan adalah menggunakan *dataset multi-year* untuk menangkap variabilitas jangka panjang, eksplorasi *deep learning* untuk menangkap pola temporal yang lebih kompleks, dan integrasi data satelit untuk analisis spasial yang lebih *comprehensive*

5. KESIMPULAN DAN SARAN

Penelitian ini berhasil mengimplementasikan dan membandingkan beberapa metode *clustering* dan regresi dalam analisis pola curah hujan bulanan di Kalimantan Timur. Tiga algoritma clustering—K-Means, Agglomerative, dan MeanShift—digunakan untuk mengidentifikasi pola musiman, sementara tiga model regresi—*Random Forest*, *Linear Regression*, dan SVR—digunakan untuk memprediksi curah hujan bulan Juni berdasarkan data lima bulan sebelumnya.

Metode K-Means dan Agglomerative berhasil mengidentifikasi tiga *cluster* musiman yang konsisten dengan karakteristik iklim tropis di Indonesia, dengan *silhouette score*

0.4913. Sementara itu, MeanShift menghasilkan lima *cluster*, namun kurang efektif karena sensitivitas terhadap parameter *bandwidth* pada *dataset* kecil.

Dalam pemodelan prediktif, *Random Forest* menunjukkan performa terbaik dengan R^2 *score* sebesar 0.8921, MAE sebesar 18.30 mm, dan RMSE sebesar 22.23 mm. Hal ini mengindikasikan kemampuan model dalam menangkap pola non-linear dan menghasilkan prediksi curah hujan yang akurat. *Linear Regression* juga memberikan performa yang solid (R^2 = 0.8402), sementara SVR menunjukkan hasil yang buruk (R^2 = 0.0077) karena parameter yang belum optimal.

Kombinasi pendekatan *unsupervised* dan *supervised learning* dalam penelitian ini memberikan pemahaman yang lebih komprehensif terhadap pola curah hujan, mulai dari segmentasi musim hingga estimasi kuantitatif nilai curah hujan. Temuan ini memiliki implikasi praktis yang signifikan, antara lain dalam perencanaan pertanian, manajemen sumber daya air, dan sistem peringatan dini terhadap risiko hidrometeorologi.

Untuk pengembangan selanjutnya, disarankan penggunaan *dataset* dengan rentang waktu yang lebih panjang dan jumlah observasi yang lebih besar, serta eksplorasi metode yang lebih canggih seperti *deep* learning, guna meningkatkan akurasi prediksi serta pemahaman pola iklim yang lebih kompleks.

DAFTAR REFERENSI

Aldrian, E., & Susanto, R. D. (2003). Identification of three dominant rainfall regions within Indonesia and their relationship to sea surface temperature. *International Journal of Climatology*, 23(12), 1435-1452.

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. *In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 1027-1035).

Asra, A., & Rudiansyah. (2013). *Statistika Terapan*. Bogor: In Media.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, 25(2), 197-227.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), 1247-1250.

Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5), 603-619.

Dajan, A. (1986). *Metode Statistika Jilid 1*. Jakarta: LP3ES.

Fukunaga, K., & Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1), 32-40.

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining*. Cham: Springer.

Hamada, J. I., Yamanaka, M. D., Matsumoto, J., Fukao, S., Winarso, P. A., & Sribimawati, T. (2002). Spatial and temporal variations of the rainy season over Indonesia and their link to ENSO. *Journal of the Meteorological Society of Japan*, 80(2), 285-310.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Waltham: Morgan Kaufmann.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). New York: Springer.

Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification (Technical Report). Department of Computer Science, National Taiwan University.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651-666.

Jain, S. K., & Singh, V. P. (2003). Water resources systems planning and management. *Developments in Water Science*, 51. Amsterdam: Elsevier.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3), 241-254.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken: John Wiley & Sons.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18-22.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281-297).

Monahan, S. S. (1981). A multivariate approach to the problem of forecasting seasonal rainfall. *Journal of Applied Meteorology*, 20(11), 1315-1325.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). Hoboken: John Wiley & Sons.

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1), 86-97.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

Qian, B., Corte-Real, J., & Xu, H. (2002). Multisite stochastic weather models for impact studies. *International Journal of Climatology*, 22(11), 1377-1397.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199-222.

Steel, R. G. D., & Torrie, J. H. (1980). *Principles and Procedures of Statistics* (2nd ed.). New York: McGraw-Hill.

Tripathi, S., Srinivas, V. V., & Nanjundiah, R. S. (2006). Downscaling of precipitation for climate change scenarios: A support vector machine approach. *Journal of Hydrology*, 330(3-4), 621-640.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer.

Wang, X., Qiu, W., & Zamar, L. (2007). CLUES: A non-parametric clustering method based on local shrinking. *Computational Statistics & Data Analysis*, 52(1), 286-298.

Wilby, R. L., & Wigley, T. M. L. (1997). Downscaling general circulation model output: a review of methods and limitations. *Progress in Physical Geography*, 21(4), 530-548.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1), 79-82.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Burlington: Morgan Kaufmann.

LAMPIRAN

Lampiran A. Kode Program Clustering

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.cluster import KMeans, AgglomerativeClustering, MeanShift, estimate_bandwidth
from sklearn.metrics import silhouette_score
import numpy as np

# 1. Load data
df = pd.read_csv('./data_curah_hujan_kaltim_2005.csv')

# 2. Encode nama bulan ke angka
df['Bulan_Num'] = LabelEncoder().fit_transform(df['Bulan'])

# 3. Scaling fitur
X = df[['Bulan_Num', 'Curah_Hujan']]
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# 4. Inisialisasi hasil skor silhouette
results_stable = {}

# 5. KMeans Clustering
kmeans = KMeans(n_clusters=3, random_state=42, n_init=10)
df['KMeans_Cluster'] = kmeans.fit_predict(X_scaled)
results_stable['KMeans'] = silhouette_score(X_scaled, df['KMeans_Cluster'])

# 6. Agglomerative Clustering
agglo = AgglomerativeClustering(n_clusters=3, linkage='ward')
df['Agglo_Cluster'] = agglo.fit_predict(X_scaled)
results_stable['Agglomerative'] = silhouette_score(X_scaled, df['Agglo_Cluster'])

# 7. MeanShift Clustering (dengan bandwidth eksplisit)
bandwidth = estimate_bandwidth(X_scaled, quantile=0.3)
meanshift = MeanShift(bandwidth=bandwidth)
df['MeanShift_Cluster'] = meanshift.fit_predict(X_scaled)
if df['MeanShift_Cluster'].nunique() > 1:
    results_stable['MeanShift'] = silhouette_score(X_scaled, df['MeanShift_Cluster'])
else:
    results_stable['MeanShift'] = 'Silhouette not applicable'

# Tampilkan perbandingan Silhouette Score di terminal
print("="*50)
print("PERBANDINGAN SILHOUETTE SCORE")
print("="*50)
print(f"{'Metode':<15} {'Silhouette Score'}")
print("-"*30)
for method, score in results_stable.items():
    if isinstance(score, (int, float)):
        print(f"{method:<15} {score:.4f}")
    else:
        print(f"{method:<15} {score}")

# Rekomendasi metode terbaik
numeric_scores = {k: v for k, v in results_stable.items() if isinstance(v, (int, float))}
if numeric_scores:
    best_method = max(numeric_scores, key=numeric_scores.get)
    print(f"\nMetode Terbaik: {best_method} (Score: {numeric_scores[best_method]:.4f})")
print("="*50)

# 8. Visualisasi hasil clustering
fig, axes = plt.subplots(1, 3, figsize=(18, 5))
cluster_columns = ['KMeans_Cluster', 'Agglo_Cluster', 'MeanShift_Cluster']
titles = ['KMeans', 'Agglomerative', 'MeanShift']

for ax, col, title in zip(axes, cluster_columns, titles):
    sns.scatterplot(data=df, x='Bulan_Num', y='Curah_Hujan', hue=col, palette='tab10', s=100, ax=ax)
    ax.set_title(f'Klasterisasi : {title}')
    ax.set_xlabel('Bulan (angka)')
    ax.set_ylabel('Curah Hujan (mm)')
    ax.grid(True)

plt.tight_layout()
plt.suptitle('Perbandingan Klasterisasi', fontsize=16, y=1.05)
plt.show()

# 9. Visualisasi skor silhouette
plt.figure(figsize=(8, 6))
scores_to_plot = {k: v if isinstance(v, (int, float)) else 0 for k, v in results_stable.items()}
labels = [f"{k}\n({v:.3f})" if isinstance(v, (int, float)) else f"{k}\n(N/A)" for k, v in results_stable.items()]

sns.barplot(x=list(scores_to_plot.keys()), y=list(scores_to_plot.values()), palette='muted')
plt.title('Skor Silhouette Model Klasterisasi')
plt.ylabel('Silhouette Score')
plt.ylim(0, 1)
plt.grid(axis='y')
plt.xticks(ticks=range(len(labels)), labels=labels)
plt.tight_layout()
plt.show()
```

Lampiran B. Kode Program Regresi

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.svm import SVR
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error

# 1. Load dataset
df = pd.read_csv("./data_curah_hujan_multiregional_1979.csv")

# 2. Fitur dan target
X = df[['Jan', 'Feb', 'Mar', 'Apr', 'Mei']]
y = df['Jun']

# 3. Inisialisasi model
models = {
    'Linear Regression': LinearRegression(),
    'Random Forest': RandomForestRegressor(n_estimators=100, random_state=42),
    'SVR': SVR(kernel='rbf')
}

# 4. Evaluasi setiap model
results = []
predictions = {}

for name, model in models.items():
    model.fit(X, y)
    y_pred = model.predict(X)
    predictions[name] = y_pred
    r2 = r2_score(y, y_pred)
    mae = mean_absolute_error(y, y_pred)
    rmse = np.sqrt(mean_squared_error(y, y_pred))
    results.append({
        'Model': name,
        'R2 Score': r2,
        'MAE': mae,
        'RMSE': rmse
    })

results_df = pd.DataFrame(results)

# 5. Visualisasi: Prediksi vs Aktual dari model terbaik
best_model_name = results_df.loc[results_df['R2 Score'].idxmax(), 'Model']
y_pred_best = predictions[best_model_name]

plt.figure(figsize=(8, 6))
plt.plot(y.values, label='Aktual', marker='o', color='orange')
plt.plot(y_pred_best, label='Prediksi', marker='s', color='orangered')
plt.title(f'Prediksi Curah Hujan Bulan Juni - {best_model_name}')
plt.xlabel('Data Lokasi')
plt.ylabel('Curah Hujan (mm)')
plt.legend()
plt.grid(True)
plt.tight_layout()
plt.show()

# 6. Visualisasi evaluasi seluruh model
results_df_melted = results_df.melt(
    id_vars='Model',
    value_vars=['R2 Score', 'MAE', 'RMSE'],
    var_name='Metrik',
    value_name='Nilai'
)

plt.figure(figsize=(10, 6))
sns.barplot(data=results_df_melted, x='Model', y='Nilai', hue='Metrik')
plt.title('Perbandingan Evaluasi Model Regresi')
plt.ylabel('Nilai Metrik')
plt.xticks(rotation=0)
plt.grid(axis='y')
plt.tight_layout()
plt.show()

# 7. Tampilkan hasil evaluasi tabel
print("=== Tabel Evaluasi Model ===")
print(results_df)
```

Lampiran C. Dataset Curah Hujan Kalimantan Timur 2005

Bulan	Curah_Hujan
Januari	171.9
Februari	232.2
Maret	270.7
April	152.5
Mei	258.7
Juni	102.4
Juli	211.6
Agustus	151.3
September	135.9
Oktober	279.5
November	270.7
Desember	247

Lampiran D. Dataset Curah Hujan Multiregional 1979

Lokasi	Jan	Feb	Mar	Apr	Mei	Jun
Banda Aceh	102	79	117	157	105	96
Bandung	187	155	190	384	135	37
Jakarta	359	306	218	83	195	17
Jambi	266	206	200	156	195	82
Medan	129	100	146	207	205	102
Padang	401	246	319	224	192	179
Pekanbaru	276	208	200	167	132	74
Palembang	292	198	190	183	179	64
Semarang	408	124	354	177	191	206
Surabaya	317	249	257	194	144	10
Tanjung Karang	334	290	255	153	92	0
Yogyakarta	273	111	219	147	192	189

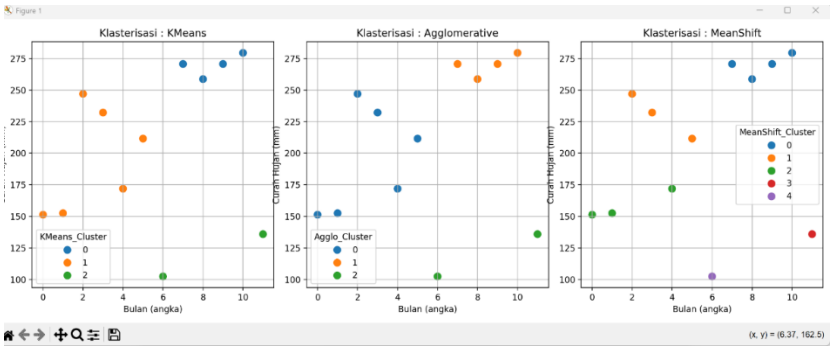
Lampiran E. Hasil Eksperimen

E.1 Output Silhouette Score Clustering

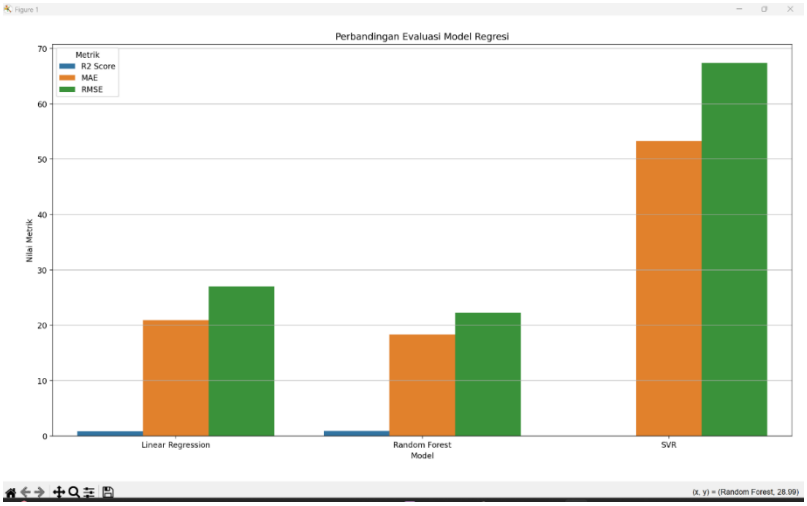
```
================================================
PERBANDINGAN SILHOUETTE SCORE
================================================
Metode          Silhouette Score
--------------------------------
KMeans          0.4913
Agglomerative   0.4913
MeanShift       0.4493

Metode Terbaik: KMeans (Score: 0.4913)
```

E.2 Output Assignment Clustering

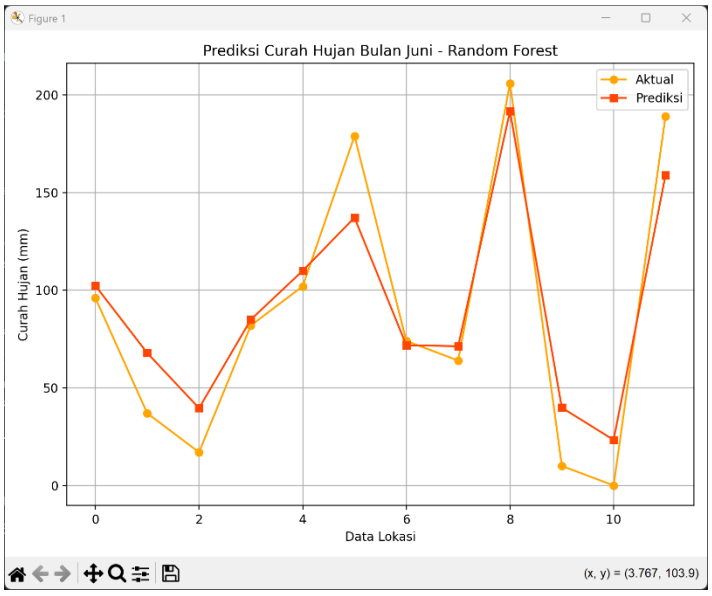


E.3 Output Evaluasi Model Regresi



Model	R^2 Score	MAE (mm)	RMSE (mm)
Linear Regression	0.8402	20.91	27.05
Random Forest	0.8921	18.3	22.23
SVR	0.0077	53.29	67.41

E.4 Output Prediksi Hujan Bulan Juni Dengan Model Regresi



Lampiran F. Interpretasi Statistik

F.1 Interpretasi Silhouette Score

Silhouette score mengukur seberapa mirip suatu objek dengan clusternya sendiri dibandingkan dengan cluster lain. Nilai berkisar dari -1 hingga 1:

1. Nilai mendekati 1: objek sangat cocok dengan clusternya.
2. Nilai mendekati 0: objek berada di antara dua cluster.
3. Nilai negatif: objek mungkin salah ditempatkan.

K-Means dengan score 0.4913 menunjukkan struktur cluster yang kuat dan pemisahan yang jelas antar musim.

F.2 Interpretasi R^2 Score

R^2 score mengukur proporsi varians dalam variabel dependen yang dapat diprediksi dari variabel independen:

1. R^2 = 0.8921 (Random Forest): 89% variasi curah hujan Juni dapat dijelaskan oleh model.
2. R^2 = 0.8402 (Linear Regression): 84% variasi dapat dijelaskan.
3. R^2 = 0.0077 (SVR): 0.77% variasi dapat dijelaskan.

Nilai $R^2 > 0.8$ umumnya dianggap sebagai model yang baik untuk prediksi curah hujan.

F.3 Interpretasi MAE dan RMSE

MAE (Mean Absolute Error) memberikan rata-rata error absolut dalam satuan asli (mm):

1. Random Forest MAE = 18.3 mm: rata-rata error prediksi sekitar 18 mm.
2. Ini relatif kecil mengingat range curah hujan Juni (134-267 mm).

RMSE memberikan bobot lebih pada error besar dan berguna untuk mendeteksi outlier:

1. Random Forest RMSE = 22.23 mm.
2. Perbedaan RMSE dan MAE yang tidak terlalu besar menunjukkan tidak ada outlier ekstrem dalam prediksi.

Lampiran G. Data Program Dan Dataset Keseluruhan

https://drive.google.com/drive/folders/1c40rUa5hOXqomsCiqmrFgwo2tuC8KGYT?usp=sharing