Vol. 1 No. (1) (Juli 2023) 19-26

Journal of Data Insights

http://journalnew.unimus.ac.id/index.php/jodi

Perbandingan Hasil Klasifikasi Data Iris menggunakan Algoritma *K-Nearest Neighbor* dan *Random Forest*

Budiono Rahman [1], **Fatkhurokhman Fauzi**[2], **Saeful Amri**[3]
[1]Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Muhammadiyah Semarang, Indonesia
[2]Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Muhammadiyah Semarang, Indonesia
[3]Program Studi Sains Data, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Muhammadiyah Semarang, Indonesia

DOI:

Info Artikel

Sejarah Artikel:
Disubmit 6 Mei 2023
Direvisi 16 Mei 2023
Disetujui 03 Juni 2023

Keywords:
Akurasi; F1-Score; K-Nearest Neighbor; Random Forest.

Abstrak

Data mining merupakan suatau metode yang baik untuk menangani data skala besar. Performasi menjadi penting dalam metode data mining. Dua metode yang memiliki performasi terbaik diantaranya *K-Nearest Neighbor* (KNN) dan *Random Forest* (RF). Artikel ini membahas terkait perbandingan performasi K-NN dan RF. Data yang digunakan pada penelitian ini adalah Iris. Data dibagi menjadi 80% data *training* dan 20% data *testing*. Validasi performasi menggunakan nilai akurasi dan *F1-Score*. Berdasarkan nilai. Berdasarkan hasil yang didapat metode RF lebih baik dibandingkan dengan metode K-NN. Nilai akurasi yang didapat oleh metode RF adalah 1.00 atau 100% dan nilai *F1-Score* sebesar 1.00.

Abstract

Data mining is a good method for dealing with large scale data. Performance becomes important in data mining methods. The two methods that have the best performance are K-Nearest Neighbor (KNN) and Random Foreast (RF). This article discusses the performance comparison of K-NN and RF. The data used in this study is Iris. The data is divided into 80% training data and 20% testing data. Performance validation uses accuracy value and F1-Score. Based on value. Based on the results obtained, the RF method is better than the K-NN method. The accuracy value obtained by the RF method is 1.00 or 100% and the F1-Score value is 1.00.

✉ Alamat Korespondensi:
E-mail: fatkhurokhmanf@unimus.ac.id

e-ISSN:

PENDAHULUAN

Data iris merupakan data yang teridiri dari dari 150 bunga yang diidentifikasi berdasarkan panjang mahkota, lebar mahkota, panjang kelopak dan lebar kelopak (Hussain et al., 2020; Thirunavukkarasu et al., 2018). Dari 150 data tersebut pada umumnya peneliti-peneliti sebelumnya mengelompokkan menjadi tiga kelompok bunga, yaitu iris setosa, iris virginica dan iris versi color(Azmi, 2014). Pengujian metode pengklasteran oleh peneliti-peneliti sebelumnya yang menggunakan data iris, karena data iris merupakan data sederhana yang mudah didapat. Salah satu cabang ilmu yang dapat digunakan untuk mengelompokkan data menjadi beberapa kelompok data, diantaranya adalah dengan menggunakan konsep data mining.

Data mining adalah adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan didalam daftar data. Data mining merupakan proses yang menggunakan teknik statistik, matematika, kecerdasan buatan dan machine learning untuk mengekstrasi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai daftar data besar(Agustina et al., 2020; Layton, 2017; Zhao, 2012). Machine learning merujuk pada sebuah metode yang membuat komputer memiliki kemampuan dalam mempelajari dan melakukan sebuah pekerjaan secara otomatis. Proses machine learning dilakukan melalui algoritma tertentu, sehingga pekerjaan yang diperintahkan kepada komputer dapat dilakukan secara otomatis (Primajaya & Sari, 2018).

Klasifikasi digunakan untuk kelompok data yang bersifat superviced, dimana data-data terawasi atau sudah diketahui kelasnya. Tujuan klasifikasi adalah untuk mendekrispikan suatu data atau objek baru kedalam kelas tertentu berdasarkan kemiripan karakteristik datanya. Klasifikasi dapat menggunakan clustering untuk mengelompokkan data yang didasari pada kemiripan antar data, sehingga data dengan kemiripan paling dekat berada dalam satu cluster sedangkan data yang berbeda dalam kelompok lainnya (Widyadhana et al., 2021). Adapun metode klasifikasi yang sering digunakan dalam beberapa penelitian sebelumnya, yaitu menggunakan metode K-Neirs Neighbors dan Random Forest.

K-Nearest Neighbor (KNN) diperkenalkan oleh Fix dan Hodges pada tahun 1951. Metode KNN merupakan suatu metode non parametrik untuk klasifikasi atas dasar kebutuhan untuk melakukan analisis diskriminan ketika nilai estimasi parametrik dari fungsi peluangnya tidak diketahui atau sulit untuk ditentukan. Selain itu, metode tersebut menjadi terkenal karena kesederhanaannya dan kekonvergenannya relatif tinggi (Mumtaz et al., 2023). Metode KNN adalah satu yang paling banyak digunakan di dunia untuk persoalan pengklasifikasian (Fauzi, 2017). Sedangkan *Random Forest* (RF) merupakan metode *bagging*, yaitu metode yang membangkitkan sejumlah *tree* dari data sample dimana pembuatan satu *tree* pada saat *training* tidak bergantung pada *tree* sebelumnya, kemudian keputusan diambil berdasarkan voting terbanyak.

Penelitian sebelumnya mengenai klasifikasi pernah dilakukan oleh Sanrang (2017), dalam penelitian tersebut menggunakan metode RF untuk mengklasifikasi curah hujan bulanan di Kabupaten Indramayu. Hasil penelitian tersebut menunjukkan bahwa klasifikasi dengan metode RF maupun *rotation forest* mampu menghasilkan akurasi klasifikasi yang cukup konsisten walaupun nilai yang dihasilkan setiap menjalankan algoritmenya berbeda-beda.

Selain itu, penelitian yang dilakukan oleh Mulyana (2017), dimana penelitian dilakukan dengan pendeteksian warna kulit untuk segmentasi tangan dan pengenalan gestur menggunakan klasifikasi *K-Nearest Neighbor* (KNN). Hasil pengujian dalam penelitian yang telah dilakukan diperoleh nilai rata-rata akurasi sebesar 83,05% untuk pengujian menggunakan beberapa user. Selain itu, diperoleh nilai akurasi 93,94% untuk pengujian gestur huruf dan 92,86% untuk pengujian gestur angka.

Berdasarkan kedua penelitian diatas terlihat bahwa metode RF dan KNN memiliki performasi baik. Oleh karena itu pada penelitian ini akan membandingkan performasi metode RF dan metode KNN. Performasi dari kedua metode akan dilihat dari nilai *F1-Score* dan nilai akurasinya.

METODE

Penelitian ini berfokus membandingkan antara metode *K-Nearest Neighbor* (KNN) dan *Random Foreast* (RF). Data yang digunkan untuk membandingkan kedua metode tersebut adalah data Iris yang bersumber pada website: https://www.kaggle.com/datasets/saurabh00007/iriscsv. Perbandingan performasi KNN dan RF dilihat dari nilai akurasi dan *F1-Score*. Nilai akurasi dan *F1-Score* tertinggi merupakan metode terbaik. Berikut langkah analisis yang dilakukan:

1. Statistik Deskriptif.
2. Memisahkan *features* dan label.
3. Membagi data *training* dan *testing* sebesar 80%:20%.
4. Kategorisasi data (label).
5. Standarisasi data.
6. Klasifikasi menggunakan metode KNN.
7. Klasifikasi menggunakan metode RF.
8. Perbandingan metode KNN dan RF
9. Memperoleh metode terbaik.

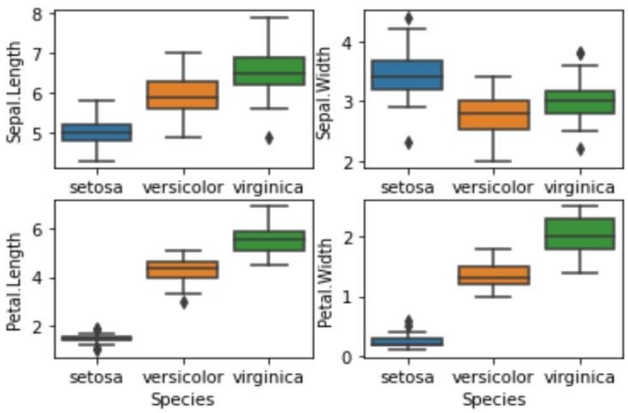
HASIL DAN PEMBAHASAN

Statistik deskriptif bertujuan untuk menjelaskan karakteristik data yang digunakan, sehingga diperoleh hasil sebagai berikut:

Tabel 1. Statistik Deskriptif

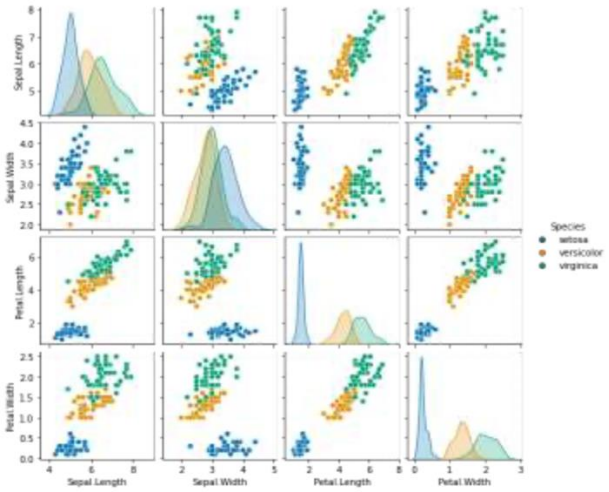
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
N	150	150	150	150
Rata-rata	5.84	3.06	3.76	1.2
Standar deviasi	0.83	0.46	1.77	0.76
Minimum	4.3	2.00	1.00	0.10
Maksimum	7.90	4.40	6.90	2.50
Kuartil 1	5.10	2.80	1.60	0.30
Median	5.80	3.00	4.35	1.30
Kuartil 3	6.40	3.30	5.10	1.80

Berdasarkan Tabel 1 menunjukkan bahwa pada masing-masing atribut data berjumlah 150 data. Selain itu, semua atribut yang digunakan juga memiliki nilai mean dan standar deviasi yang beragam untuk masing-masing atribut. Adapun perbedaan tersebut dapat divisualisasikan sebagai berikut:



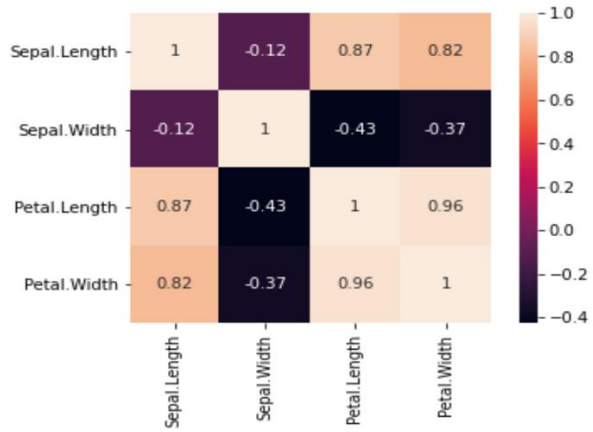
Gambar 1. Deskripsi karakteristik pada masng-masing spesies

Gambar 1 menunjukkan bahwa perbedaan kelompok titik data yang dijelaskan pada Tabel 1 adalah spesies bunga setosa. Ukuran nilai atribut pada bunga setosa lebih kecil dan kurang menyebar dibandingkan dengan dua spesies lainnya. Selain itu, pada Tabel 1 juga menjelaskan bahwa spesies bunga versicolor memiliki nilai rata-rata yang lebih rendah daripada virginica. Selanjutnya, untuk mengetahui korelasi antara variabel (*species*) diperlihatkan pada Gambar 2 di bawah ini.



Gambar 2. Hubungan antar variabel (*species*)

Gambar 2 menunjukkan korelasi antar variabel, diketahui bahwa spesies *iris setosa* selalu terpisah dari kelas yang lain, artinya saat melakukan klasifikasi terdapat kemungkinan besar bahwa model akan selalu dapat membedakan spesies *setosa* dengan baik. Selain itu, distribusi data untuk *petal-length* pada spesies *setosa* terpisah dari kelas yang lain. Selanjutnya, jika dilihat persebaran datanya pada diagram pencar, sebagian besar kombinasi atribut miliki korelasi *Pearson* yang positif, artinya *features* yang terdapat pada dataset ini baik untuk digunakan untuk membuat sebuah model. Selain itu, dengan korelasi yang tinggi, maka model klasifikasi dapat diubah menjadi model regresi untuk melakukan peramalan. Kemudian, secara kuantitatif matriks korelasi dapat dijelaskan pada Gambar 3 di bawah ini.



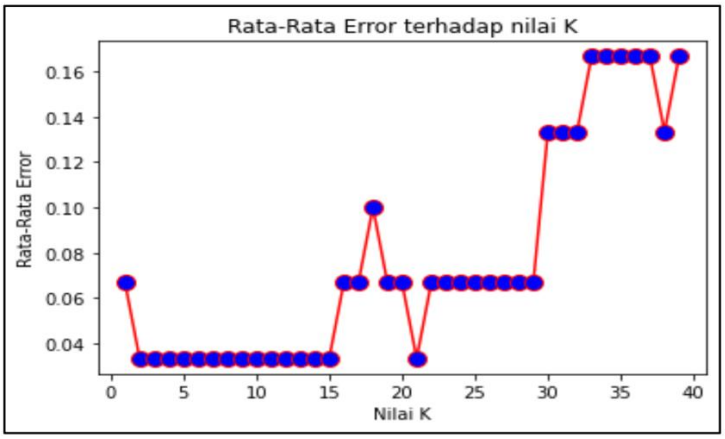
Gambar 3. Matriks korelasi

Berdasarkan gambar 3, diketahui bahwa fitur *sepal-length* memiliki korelasi yang sangat positif. Sedangkan pada *sepal-width* tidak berkorelasi. Selain itu, pada fitur *petal-length* juga memiliki korelasi yang relatif tinggi dengan *sepal- length*, tetapi tidak berkorelasi dengan dengan *sepal-width*.

Tahap selanjutnya adalah *pre-processing* data sebelum dilakukan klasifikasi menggunakan metode *K-Nearest Neighbor* (KNN) dan *Random Foreast* (RF). Tahap pertama *pre-processing* data adalah *slicing*. Tahap *Slicing* yaitu memtransformasi bentuk *dataframe* menjadi *array*. Tahap kedua adalah membagi data menjadi data training (80%) dan data testing (20%). Tahap ketiga adalah *label Encoding*. Tahap ini mengkategorisasikan label kedalam skala nominal. Tahap keempat adalah standarisasi data. Standarisasi data bertujuan untuk menyamakan skala antar variabel.

K-Nearest Neighbor

Model *K-Nearest Neighbors* membutuhkan suatu nilai konstanta *k* untuk menentukan berapa banyak tetangga yang akan digunakan oleh model. Kode di bawah ini akan melakukan training sebanyak 40 kali dengan nilai *k* dari 1 hingga 40, dimana angka 40 ini bersifat bebas dan dapat diganti dengan angka lain.



Gambar 4. Akurasi KNN terhadap nilai *k*

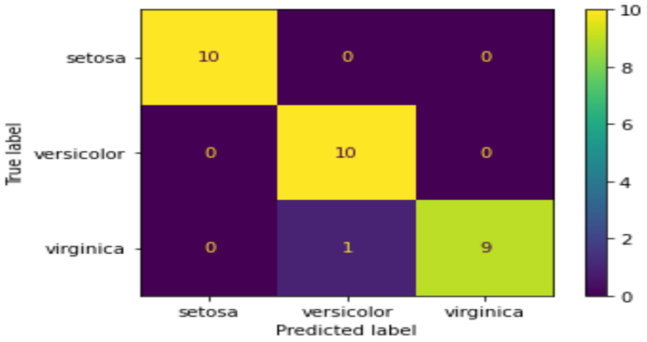
Berdasarkan Gambar 4, menunjukkan bahwa terdapat beberapa nilai *k* yang dapat meminimalisasi nilai *error*. Nilai *k* awal yang digunakan sebesar 40, dapat dilihat pada grafik bahwa nilai *k* sebesar 40 memiliki tingkat *error* sebesar 0,17. Sedangkan nilai *k* yang kecil seperti 1, 2, 4 dan 5 memiliki tingkat *error* yang kecil. Selanjutnya akan dilakukan proses *training,* dimana akan menggunakan nilai konstanta *k* = 5 (setelah dilakukan *trial* dan *error*).

Setelah melakukan proses *training*, kemudian model akan dievaluasi untuk mengetahui performasi dari metode KNN. Adapun hasil yang diperoleh sebagai berikut:

Tabel 2. Performasi KNN

Performasi	Nilai
F1-Score	0.98
Akurasi	0.97

Berdasarkan Tabe 2, secara keseluruhan model ini berhasil mendapatkan akurasi sebesar 97%, sehingga dapat dinyatakan bahwa model klasifikasi yang dihasilkan sangat baik. Selain itu, mengacu pada nilai *F1-score* yang tinggi, menandakan bahwa sebaran hasil klasifikasi seimbang pada setiap kelas. Hasil klasifikasi juga dapat digambarkan melalui plot *confusion matrix* (Gambar 5).



Gambar 5. *Confusion Matrix*

***Random Forest* (RF)**

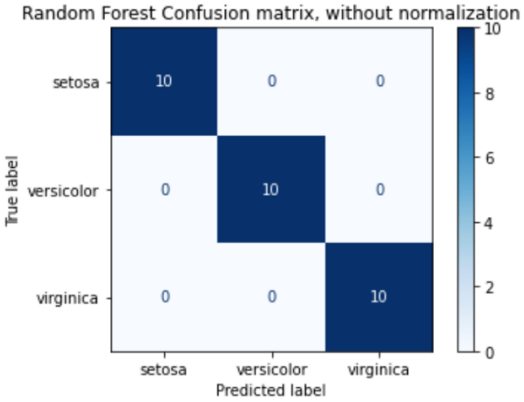
Random forest memprediksi respons suatu amatan dengan cara menggunakan semua hasil prediksi *L* pohon keputusan. Penggunaan algoritma *random forest* dalam kasus klasifikasi digunakan Teknik suara terbanyak untuk menentukan hasil prediksi, yaitu kategori yang paling sering muncul sebagai hasil prediksi dari *L* pohon klasifikasi. Pemodelan klasifikasi dalam makalah ini menggunakan program *Python* dan *library scikit-learn.* Tahapan yang dilakukan pada *preprocessing* data dilakukan proses yang sama seperti pada analisis pada KKN, selanjutnya tahapan klasifikasi dilakukan sebagai berikut:

1. Membuat Model Random Klasifikasi Terhadap *Training set*
   Proses ini dilakukan untuk melatih algoritma menggunakan data training yang dibentuk pada tahapan *prerocessing data,* dimana parameter jumlah pohon (*n_estimator*) yang akan dibentuk sebanyak 100. Adapun kriteria yang digunakan yaitu *entropy*. Setelah melakukan proses *training*, kemudian model akan dievaluasi untuk mengetahui apakah model yang dibuat memiliki akurasi yang baik atau belum.
2. Evaluasi Model
   Berdasarkan Tabel 3, diketahui bahwa model dapat membedakan semua data pada spesies *iris setosa, iris versicolor* dan *iris virginica* dengan sempurna. Adapun secara keseluruhan model ini berhasil mendapatkan akurasi sebesar 100%, sehingga dapat dinyatakan bahwa model klasifikasi yang dihasilkan sangat baik. Selain itu, mengacu pada nilai *F1-score* yang tinggi, menandakan bahwa sebaran hasil klasifikasi seimbang pada setiap kelas.

Tabel 3. Performasi RF

Performasi	Nilai
F1-Score	1.00
Akurasi	1.00

Hasil klasifikasi juga dapat digambarkan melalui plot *confusion matrix*.. Gambar 6 menunjukkan ketepatan akurasi hasil prediksi untuk setiap kelas (*species*). Hal ini juga dapat dinyatakan bahwa model yang dihasilkan mampu memprediksi setiap kelas dengan tepat. Secara teori algoritma *random forest* menggunakan nilai *majority* dari keseluruhan *tree* atau pohon yang terbentuk.



Gambar 6. *Confusion Matrix*

Perbandingan Klasifikasi KNN dan RF

Perbandingan dilakukan untuk mengetahui performasi terbaik dari metode KNN dan RF. Secara garis besar kedua metode tersbut menghasilkan performasi yang sangat baik. Hal tersebut dapat dilihat berdasarkan nilai *F1-Score* dan nilai akurasi yang mendekati nilai 1. Jika dilihat lebih detail lagi metode RF lebih baik dibandingkan dengan metode KNN, terbukti dengan nilai akurasi adalah 1 atau 100% berdasarkan Tabel 4.

Tabel 4. Perbandingan KNN dan RF

Metode	F1-Score	Akurasi
KNN	0.98	0.97
RF	1.00	1.00

KESIMPULAN

Model klasifikasi menggunakan algoritma *K-Nearest Neighbors* (KNN) dapat memprediksi setiap kelas dengan nilai akurasi sebesar 97%. Selain itu, terjadi kesalahan prediksi pada spesies iris versicolor dan iris virginica. Hal ini dapat dilihat pada nilai precision dan nilai recall yang tidak mencapai angka 1.Model klasifikasi menggunakan algoritma random forest dapat memprediksi setiap kelas dengan nilai akurasi sebesar 100%, artinya tidak ada kesalahan prediksi pada spesies iris setosa, iris versicolor dan iris virginica. Hal ini dapat dilihat pada nilai precision dan nilai recall sebesar 1,00. Berdasarkan hasil klasifikasi menggunakan algoritma *K-Nearest Neighbors* (KNN) dan Random Forest, maka dapat disimpulkan bahwa metode random forest lebih baik dalam memprediksi setiap kelas yang diuji..

DAFTAR PUSTAKA

Agustina, D., Putri, E., Fauzi, F., Alawiyah, S. N., & Wasono, R. (2020). METODE SUPPORT VECTOR MACHINE (SVM) UNTUK KLASIFIKASI DATA EKSPRESI GEN MICROARRAY. *EDUSAINTEK 4*, 1–10.

Azmi, M. (2014). KOMPARASI METODE JARINGAN SYARAF TIRUAN SOM DAN LVQ UNTUK MENGIDENTIFIKASI DATA BUNGA IRIS. *Jurnal TEKNOIF*, *2*(1), 65–70.

Fauzi, F. (2017). K-Nearset Neighbor (K-NN) dan Support Vector Machine (SVM) untuk Klasifikasi Indeks Pembangunan Manusia Provinsi Jawa Tengah Info Artikel. *Jurnal MIPA*, *40*(2). http://journal.unnes.ac.id/nju/index.php/JM

Hussain, Z. F., Ibraheem, H. R., Alsajri, M., Ali, A. H., Ismail, M. A., Kasim, S., & Sutikno, T. (2020). A new model for iris data set classification based on linear support vector machine parameter's optimization. *International Journal of Electrical and Computer Engineering*, *10*(1), 1079–1084. https://doi.org/10.11591/ijece.v10i1.pp1079-1084

Layton, R. (2017). *Learning Data Mining with Python*. Packt Publishing. https://books.google.co.id/books?id=8UEwDwAAQBAJ

Mumtaz, G., Akram, S., Iqbal, W., Ashraf, M. U., Almarhabi, K. A., Alghamdi, A. M., & Bahaddad, A. A. (2023). Classification and Prediction of Significant Cyber Incidents (SCI) using Data Mining and Machine Learning (DM-ML). *IEEE Access*, 1. https://doi.org/10.1109/ACCESS.2023.3249663

Primajaya, A., & Sari, B. N. (2018). Random Forest Algorithm for Prediction of Precipitation. *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)*, *1*(1), 27–31.

Thirunavukkarasu, K., Singh, A. S., Rai, P., & Gupta, S. (2018). Classification of IRIS Dataset using Classification Based KNN Algorithm in Supervised Learning. *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, 1–4. https://doi.org/10.1109/CCAA.2018.8777643

Widyadhana, D., Hastuti, R. B., Kharisudin, I., & Fauzi, F. (2021). Perbandingan Analisis Klaster K-Means dan Average Linkage untuk Pengklasteran Kemiskinan di Provinsi Jawa Tengah. *PRISMA, Prosiding Seminar Nasional Matematika*, *4*, 584–594. https://journal.unnes.ac.id/sju/index.php/prisma/

Zhao, Y. (2012). *R and Data Mining: Examples and Case Studies*.

Sanrang, M. H. 2017. *Pemodelan Klasifikasi Curah Hujan Bulanan di Kabupaten Indramayu dengan Metode Random Forest* dan *Rotation Forest.* Departemen Statistika. Fakultas Matematika dan Ilmu Pengetahuan Alam. Bogor: Institut Pertanian Bogor.