ISSN: 1979-2328

Rain Prediction Clustering in Australia Using the K-Means Algorithm in the WEKA and RStudio Application

Klasterisasi Prediksi Hujan di Australia dengan Menggunakan Algoritme K-Means pada Aplikasi WEKA dan RStudio

Dinar Ajeng Kristiyanti[1], Irwansyah Saputra[2], Rina[3]

[1,2] Ilmu Komputer, IPB University, Indonesia
[1] Teknologi Informasi, Universitas Bina Sarana Informatika, Indonesia
[2,3] Sistem Informasi, Universitas Nusa Mandiri, Indonesia

[1]dinarajengkristiyanti@apps.ipb.ac.id, [2]92irwansyah@apps.ipb.ac.id,
[3]11200244@nusamandiri.ac.id

Keywords:Clustering; K-Means; WEKA; Rstudio; Rain Australia

Abstract

Purpose: *The purpose of this study is how to create an ideal cluster in predicting rainfall in Australia based on the percentage of the sum of squares error (SSE) using the K-Means algorithm with WEKA and RStudio applications.*

Design/methodology/approach: *The method or stages applied in predicting rain in Australia are through several stages including Data Collection, Data Pre-processing (including Missing Value handling in it), Data Mining Modeling by applying the K-Means Clustering algorithm using WEKA and RStudio, Validation results with SSE as well as Data Visualization using plots.*

Findings/result: *Based on the results obtained, clusters of 2 with an SSE of 28.0% are ideal clusters for predicting rain in Australia. In the WEKA software, rain clusters are represented by blue nodes, and non-rainy clusters are represented by red nodes. While in the RStudio software, rain clusters are represented by black nodes and non-rainy clusters are represented by red nodes.*

Originality/value/state of the art: *Get the ideal cluster in predicting rainfall in Australia by comparing the results obtained using the WEKA and RStudio applications.*

ISSN: 1979-2328

Kata kunci:Klasterisasi; K-means; WEKA; RStudio; Hujan Australia

Abstrak

Tujuan: Tujuan dari penelitian ini adalah bagaimana cara menciptakan *cluster* yang ideal dalam memprediksi curah hujan di Australia berdasarkan presentase sum of squares error (SSE) menggunakan algoritme K-Means dengan aplikasi WEKA dan RStudio.

Perancangan/metode/pendekatan: Metode atau tahapan yang diterapkan dalam melakukan prediksi hujan di Australia yaitu melalui beberapa tahapan diantaranya Pengumpulan Data, Data Pre-processing (termasuk dilakukan penanganan *Missing Value* didalamnya), Pemodelan *Data Mining* dengan menerapkan algoritme *K-Means Clustering* menggunakan WEKA dan RStudio, Validasi hasil dengan SSE serta Visualiasi Data menggunakan plot.

Hasil: Berdasarkan hasil yang diperoleh, *cluster* yang berjumlah 2 dengan SSE 28.0% merupakan *cluster* ideal untuk memprediksi hujan di Australia. Pada *software* WEKA *cluster* hujan diwakili oleh *node* berwarna biru dan *cluster* tidak hujan diwakili oleh *node* berwarna merah. Sedangkan pada *software* RStudio *cluster* hujan diwakili oleh *node* berwarna hitam dan *cluster* tidak hujan diwakili oleh *node* berwarna merah.

Keaslian/ *state of the art*: Mendapatkan *cluster* yang ideal dalam memprediksi curah hujan di Australia dengan membandingkan hasil yang didapatkan menggunakan aplikasi WEKA dan RStudio.

1. Pendahuluan

Ketidakpastian data curah hujan merupakan salah satu masalah yang paling kompleks [1]. Untuk itu diperlukan suatu metode yang dapat digunakan untuk memprediksi curah hujan dengan hasil yang akurat. Saat ini, sebagian besar metode peramalan curah hujan tidak mampu mendeteksi pola tersembunyi atau tren non-linier dalam data curah hujan secara akurat dengan waktu tertentu [2]. Masalah yang kompleks terkait data curah hujan serta metode yang telah ada tersebut menyebabkan prediksi perkiraan menjadi salah, tidak akurat, tidak tepat waktu bahkan lebih jauh mengakibatkan kerugian yang besar bagi negara [1]. Prediksi curah hujan yang tepat dan akurat dapat mendukung keberhasilan suatu negara dalam perencanaan dan pengelolaan sumber daya air, kegiatan konstruksi terhadap tata wilayah kota maupun daerah, faktor penentu operasi penerbangan [3], atau bahkan penanggulangan dan persiapan bencana seperti peringatan banjir [4].

Prediksi curah hujan yang akurat telah menjadi perhatian besar di banyak negara, khususnya di Australia dimana memiliki iklim yang sangat bervariasi [5]. Seorang ilmuan data bernama Joe

ISSN: 1979-2328

Young [6] telah merilis data curah hujan di Australia yang telah diunggah ke situs Kaggle dengan judul *dataset rain in Australia*. Kaggle merupakan situs yang berisikan komunitas *online* dengan beragam ilmuan data dan praktisi *machine learning* yang diakuisisi oleh Google. Selain menjadi komunitas, situs ini juga berisi kumpulan data (*dataset*) berbagai kasus yang ditemukan dari seluruh dunia [7].

Salah satu metode untuk menemukan pola data yang tersembunyi agar menghasilkan pengetahuan baru atau disebut juga dengan *data mining.* Dalam *data mining* terdapat sebuah metode yang bernama *clustering* yaitu metode yang digunakan untuk membagi data yang memiliki karakteristik yang sama ke dalam satu kelompok dan data yang memiliki karakteristik berbeda ke dalam kelompok lain [8]. Algoritme *K-means* merupakan salah satu algoritme *clustering* [9].

K-means merupakan metode iteratif sederhana untuk membagi kumpulan data ke dalam sejumlah *cluster* k, dimana jumlah k telah ditentukan oleh pengguna [10]. Sebagai pembelajaran tanpa pengawasan atau *unsupervised learning*, iterasi dalam *K-means* akan mengelompokkan data yang memiliki karakteristik yang mirip ke dalam satu *cluster* dan data yang memiliki karakteristik berbeda dikelompokkan ke dalam *cluster* lain. Di setiap *cluster*, terdapat titik pusat (*centroid)* yang menunjukkan identifikasi unik dari *cluster* tersebut [11]. Beberapa penelitian terkait prediksi curah hujan telah banyak dilakukan menggunakan algoritme *K-means clustering* [5][12][13][14][15]. Sebuah algoritme *clustering* menemukan kelompok-kelompok contoh serupa di seluruh *dataset*. Aplikasi WEKA mendukung beberapa algoritme *clustering* seperti EM (*Expectation Maximization*), *Filtered Clusterer*, *Hierarchical Clusterer*, *Simple K-Means* dan sebagainya [16]. RStudio merupakan aplikasi yang dapat juga dilakukan untuk melakukan proses *clustering* [9]. Aplikasi WEKA maupun RStudio keduanya memilki performa tersendiri dalam mengelompokkan data (*clustering*) yang akurat.

Penelitian ini berfokus bagaimana cara menciptakan *cluster* yang ideal dalam memprediksi curah hujan di Australia. Idealitas *cluster* dapat dilihat menurut presentase *sum of squares error* (SSE). SSE adalah teknik statistik yang digunakan dalam analisis regresi untuk menentukan distribusi titik data. Tujuan dari analisis regresi adalah untuk menentukan tingkat kesesuaian antara kumpulan data dan fungsi, yang membantu menjelaskan bagaimana kumpulan data dihasilkan [17]. Kontribusi penelitian yang dihasilkan dari penelitian ini adalah mendapatkan *cluster* yang ideal dalam memprediksi curah hujan di Australia dengan membandingkan hasil yang didapatkan menggunakan aplikasi WEKA dan RStudio.

2. Metode/Perancangan

Penelitian tentang klasterisasi prediksi hujan di Australia dilakukan menggunakan algoritme *K-Means* dengan hasil visualisasi menggunakan aplikasi WEKA dan RStudio. Di mana metode tersebut akan mempartisi data ke dalam sejumlah *cluster* yang telah ditentukan. Untuk data yang memiliki karakteristik mirip akan dikelompokan menjadi satu *cluster* dan data yang memiliki karakteristik berbeda akan dikelompokan ke *cluster* lainnya.

2.1. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini adalah *Rain in Australia dari kaggle.com Repository*, dengan alamat url yaitu https://www.kaggle.com/jsphyg/weather-dataset-rattle-package. *Dataset* ini merupakan data yang berisikan pengamatan cuaca harian dari berbagai

stasiun cuaca Australia. Dimana *dataset* ini konteksnya adalah memprediksi apakah esok hari akan terjadi hujan atau tidak dengan melatih model klasifikasi biner pada target variabel RainTomorrow. Variabel target *RainTomorrow* berarti mengisyaratkan "Apakah hari berikutnya akan turun hujan? Ya atau Tidak".

Dataset ini berukuran 14 MB dengan format .csv, terdiri atas 1 file .csv bernama weatherAUS.csv dengan jumlah 142.193 record dan 24 variabel atau atribut atau field, diantaranya dapat dilihat dalam tabel 1 berikut ini. Setiap atribut dalam *dataset* ini dijelakan pada **Tabel 1.**

Tabel 1. Penjelasan Atribut dalam Dataset Rain in Australia

No.	*Feature*	Tipe Data	Deskripsi
1.	*Date*	*date*	Tanggal observasi prediksi pengamatan cuaca harian
2.	*Location*	*chr*	Nama umum lokasi stasiun cuaca
3.	*MinTemp*	*numeric*	Suhu minimum dalam derajat celsius
4.	*MaxTemp*	*numeric*	Suhu maksimum dalam derajat celsius
5.	*Rainfall*	*numeric*	Jumlah curah hujan yang tercatat untuk hari itu dalam mm
6.	*Evaporation*	*numeric*	Disebut juga penguapan kelas A (mm) dalam 24 jam hingga 9 pagi
7.	*Sunshine*	*numeric*	Jumlah jam sinar matahari cerah dalam satu hari
8.	*WindGustDir*	*chr*	Arah hembusan angin terkuat dalam 24 jam hingga tengah malam
9.	*WindGustSpeed*	*numeric*	Kecepatan (km/jam) hembusan angin terkuat dalam 24 jam hingga tengah malam
10.	*WindDir9am*	*chr*	Arah angin pada jam 9 pagi
11.	*WindDir3pm*	*chr*	Arah angin pada jam 3 sore
12.	*WindSpeed9am*	*numeric*	Kecepatan angin (km / jam) rata-rata lebih dari 10 menit sebelum jam 9 pagi
13.	*WindSpeed3pm*	*numeric*	Kecepatan angin (km / jam) rata-rata lebih dari 10 menit sebelum jam 3 sore
14.	*Humidity9am*	*numeric*	Kelembapan (persen) pada jam 9 pagi
15.	*Humidity3pm*	*numeric*	Kelembapan (persen) pada jam 3 sore
16.	*Pressure9am*	*numeric*	Tekanan atmosfer (hpa) berkurang hingga rata-rata permukaan laut pada pukul 9 pagi
17.	*Pressure3pm*	*numeric*	Tekanan atmosfer (hpa) berkurang hingga rata-rata permukaan laut pada pukul 3 sore
18.	*Cloud9am*	*numeric*	Bagian langit yang tertutup awan pada jam 9 pagi. Ini diukur dalam "oktas", yang merupakan satuan delapan. Ini mencatat berapa banyak awan
19.	*Cloud3pm*	*numeric*	Bagian langit yang tertutup awan pada jam 3 sore. Ini diukur dalam "oktas", yang merupakan satuan delapan. Ini mencatat berapa banyak awan
20.	*Temp9am*	*numeric*	Suhu (derajat C) pada jam 9 pagi
21.	*Temp3pm*	*numeric*	Suhu (derajat C) pada jam 3 sore
22.	*RainToday*	*chr*	Prediksi hujan hari ini

23.	*RISK MM*	*numeric*	Jumlah hujan hari berikutnya dalam mm. Digunakan untuk membuat variabel respons RainTomorrow. Semacam ukuran "risiko".
24.	*RainTomorrow*	*chr*	Variabel Target. Prediksi hujan besok

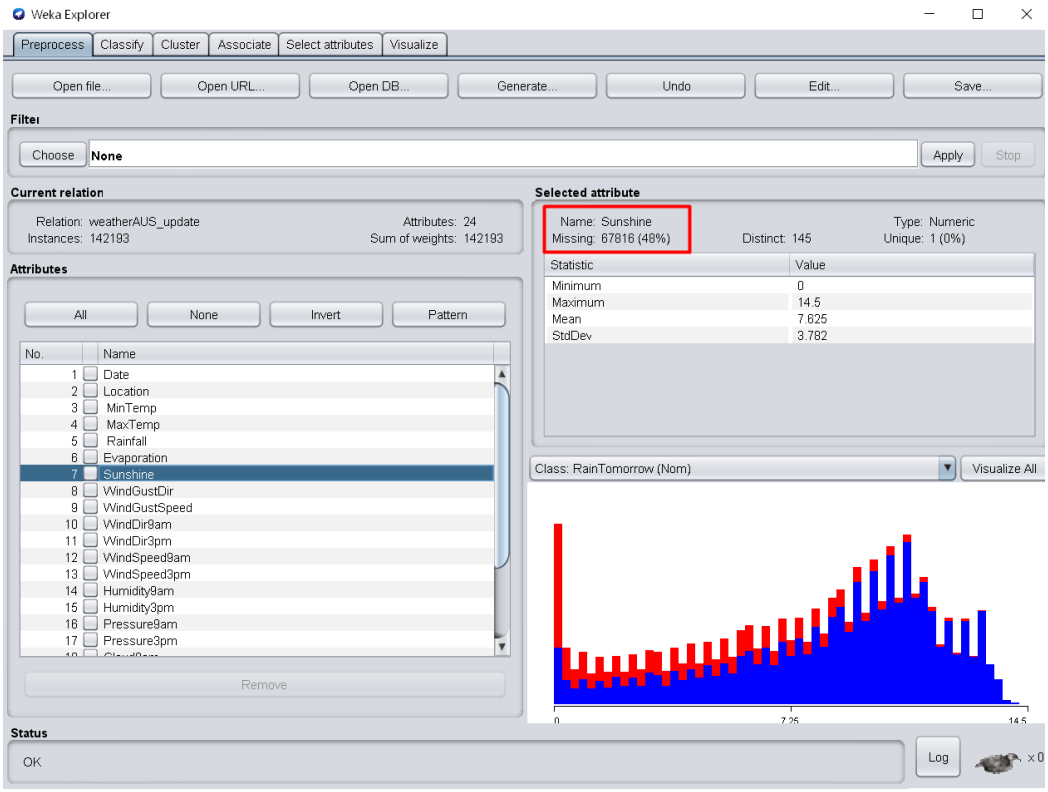
Selain itu, *dataset* tersebut memiliki missing *value* berkisar dari 1% hingga 48% yang terbesar pada beberapa atribut. Cara yang dapat dilakukan untuk menangani *missing value* adalah menghapus *tupel missing value* atau mengisinya dengan nilai konstanta berdasarkan rata-rata dari nilai dalam data tersebut [6]. *Missing value* dalam dataset *Rain in Australia* dapat dilihat pada **Tabel 2.**

Tabel 2. Total Missing Value Pada Tiap Atribut

No.	***Feature***	**Jumlah *Missing Value***	**Persentase Terhadap Keseluruhan Data**
1.	*Date*	0 *Record*	0%
2.	*Location*	0 *Record*	0%
3.	*MinTemp*	637 *Records*	>0%
4.	*MaxTemp*	322 *Records*	>0%
5.	*Rainfall*	1406 *Records*	1%
6.	*Evaporation*	60843 *Records*	43%
7.	*Sunshine*	67816 *Records*	48%
8.	*WindGustDir*	9330 *Records*	7%
9.	*WindGustSpeed*	9270 *Records*	7%
10.	*WindDir9am*	10013 *Records*	7%
11.	*WindDir3pm*	3778 *Records*	3%
12.	*WindSpeed9am*	1348 *Records*	1%
13.	*WindSpeed3pm*	2630 *Records*	2%
14.	*Humidity9am*	1774 *Records*	1%
15.	*Humidity3pm*	3610 *Records*	3%
16.	*Pressure9am*	14014 *Records*	10%
17.	*Pressure3pm*	13981 *Records*	10%
18.	*Cloud9am*	53657 *Records*	38%
19.	*Cloud3pm*	57094 *Records*	40%
20.	*Temp9am*	904 *Records*	1%
21.	*Temp3pm*	2726 *Records*	2%
22.	*RainToday*	1406 *Records*	1%
23.	*RISK MM*	0 *Record*	0%
24.	*RainTomorrow*	0 *Record*	0%

Saat menggunakan *WEKA*, terlihat atribut yang memiliki *missing value* di dalam recordnya seperti pada salah satu contoh atribut "*Sunshine*" seperti diperlihatkan **Gambar 1.**

ISSN: 1979-2328



Gambar 1. *Missing Value* pada Atribut *Sunshine* Menggunakan *WEKA*

2.2. Data Pre Processing

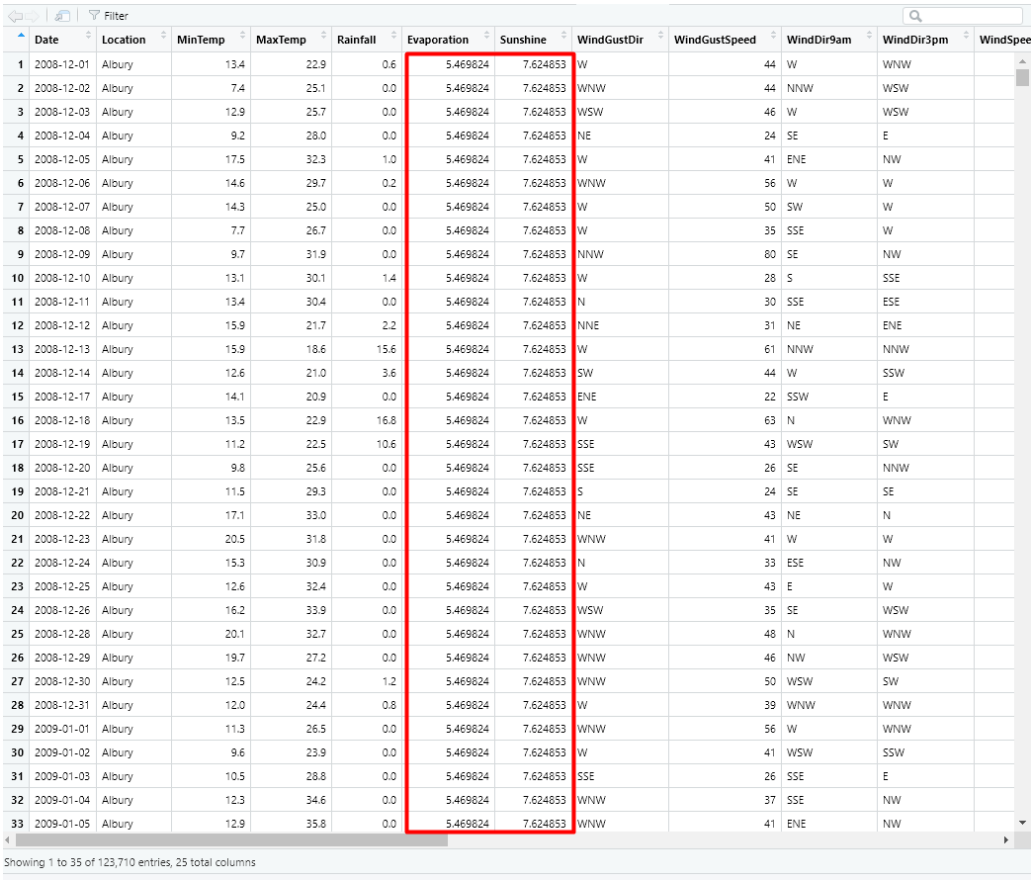
Dataset berekstensi *csv* yang terdapat pada arsip akan dilakukan tahap *pre processing* data agar lebih mudah dianalisis dan divisualisasikan. *Dataset* yang akan diproses pada tahapan ini adalah *Dataset Rain in Australia* yang terdiri dari 24 atribut [18].

Proses pembersihan data dari *noise* dapat dilakukan sesuai kebutuhan karena *noise* data banyak ragamnya. Contohnya, pada *dataset* Rain in Australia terdapat *record* yang memiliki *missing value* di beberapa atribut sehingga tidak akan optimal saat diproses untuk diekstraksi pada tahapan selanjutnya. Penanganan *missing value* akan dilakukan menggunakan aplikasi *RStudio*.

2.2.1. Penanganan *Missing Value* Menggunakan *RStudio*

Penanganan *missing value* menggunakan *RStudio* melalui kode manual, dituliskan di *GUI* yang sudah disediakan. Tabel 2 menunjukkan tipe data pada setiap atribut dalam *dataset.* Terlihat atribut yang memiliki *missing value* di dalamnya adalah atribut-atribut yang bertipe data numerik dan nominal, sehingga butuh perbedaan metode untuk menangani *missing value* pada atribut tersebut. Untuk menangani *missing value* pada tipe data numerik adalah dengan cara mengganti nilai yang hilang dengan nilai rataan dari seluruh data yang ada di dalam atributnya. Sedangkan pada tipe data nominal, *missing value* akan dihapuskan karena *record* dalam atribut tersebut terlalu beragam. Hasil penanganan *missing value* menggunakan *RStudio* dapat dilihat pada **Gambar 2**.

ISSN: 1979-2328



	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpee
1	2008-12-01	Albury	13.4	22.9	0.6	5.469824	7.624853	W	44	W	WNW	
2	2008-12-02	Albury	7.4	25.1	0.0	5.469824	7.624853	WNW	44	NNW	WSW	
3	2008-12-03	Albury	12.9	25.7	0.0	5.469824	7.624853	WSW	46	W	WSW	
4	2008-12-04	Albury	9.2	28.0	0.0	5.469824	7.624853	NE	24	SE	E	
5	2008-12-05	Albury	17.5	32.3	1.0	5.469824	7.624853	W	41	ENE	NW	
6	2008-12-06	Albury	14.6	29.7	0.2	5.469824	7.624853	WNW	56	W	W	
7	2008-12-07	Albury	14.3	25.0	0.0	5.469824	7.624853	W	50	SW	W	
8	2008-12-08	Albury	7.7	26.7	0.0	5.469824	7.624853	W	35	SSE	W	
9	2008-12-09	Albury	9.7	31.9	0.0	5.469824	7.624853	NNW	80	SE	NW	
10	2008-12-10	Albury	13.1	30.1	1.4	5.469824	7.624853	W	28	S	SSE	
11	2008-12-11	Albury	13.4	30.4	0.0	5.469824	7.624853	N	30	SSE	ESE	
12	2008-12-12	Albury	15.9	21.7	2.2	5.469824	7.624853	NNE	31	NE	ENE	
13	2008-12-13	Albury	15.9	18.6	15.6	5.469824	7.624853	W	61	NNW	NNW	
14	2008-12-14	Albury	12.6	21.0	3.6	5.469824	7.624853	SW	44	W	SSW	
15	2008-12-17	Albury	14.1	20.9	0.0	5.469824	7.624853	ENE	22	SSW	E	
16	2008-12-18	Albury	13.5	22.9	16.8	5.469824	7.624853	W	63	N	WNW	
17	2008-12-19	Albury	11.2	22.5	10.6	5.469824	7.624853	SSE	43	WSW	SW	
18	2008-12-20	Albury	9.8	25.6	0.0	5.469824	7.624853	SSE	26	SE	NNW	
19	2008-12-21	Albury	11.5	29.3	0.0	5.469824	7.624853	S	24	SE	SE	
20	2008-12-22	Albury	17.1	33.0	0.0	5.469824	7.624853	NE	43	NE	N	
21	2008-12-23	Albury	20.5	31.8	0.0	5.469824	7.624853	WNW	41	W	W	
22	2008-12-24	Albury	15.3	30.9	0.0	5.469824	7.624853	N	33	ESE	NW	
23	2008-12-25	Albury	12.6	32.4	0.0	5.469824	7.624853	W	43	E	W	
24	2008-12-26	Albury	16.2	33.9	0.0	5.469824	7.624853	WSW	35	SE	WSW	
25	2008-12-28	Albury	20.1	32.7	0.0	5.469824	7.624853	WNW	48	N	WNW	
26	2008-12-29	Albury	19.7	27.2	0.0	5.469824	7.624853	WNW	46	NW	WSW	
27	2008-12-30	Albury	12.5	24.2	1.2	5.469824	7.624853	WNW	50	WSW	SW	
28	2008-12-31	Albury	12.0	24.4	0.8	5.469824	7.624853	W	39	WNW	WNW	
29	2009-01-01	Albury	11.3	26.5	0.0	5.469824	7.624853	WNW	56	W	WNW	
30	2009-01-02	Albury	9.6	23.9	0.0	5.469824	7.624853	W	41	WSW	SSW	
31	2009-01-03	Albury	10.5	28.8	0.0	5.469824	7.624853	SSE	26	SSE	E	
32	2009-01-04	Albury	12.3	34.6	0.0	5.469824	7.624853	WNW	37	SSE	NW	
33	2009-01-05	Albury	12.9	35.8	0.0	5.469824	7.624853	WNW	41	ENE	NW	

Showing 1 to 35 of 123,710 entries, 25 total columns

Gambar 2. Penanganan *Missing Value* pada Atribut Menggunakan *RStudio*

Setelah dilakukan proses penanganan *missing value* pada seluruh atribut tersebut, *dataset* berkurang sebanyak 13% dari *dataset* awal. Ada 10 atribut tertinggi yang akan digunakan dalam kasus ini. Atribut-atribut tersebut memiliki tipe data numerik. Pemilihan atribut ini menggunakan teknik *Information Gain* untuk mendapatkan peringkat atribut yang paling berpengaruh hingga yang tidak berpengaruh. Jika diantara 10 atribut tersebut terdapat atribut yang memiliki tipe data nominal, maka akan digantikan dengan atribut setelahnya yang memiliki tipe data numerik. Hal ini dilakukan karena pengolahan pada kasus klaster menggunakan atribut yang memiliki tipe data numerik. Atribut-atribut yang akan digunakan pada penelitian ini seperti yang diperlihatkan pada **Tabel 3.**

Tabel 3. Atribut-Atribut yang Akan Digunakan

No.	*Feature*	Tipe Data	Deskripsi
1.	*MinTemp*	*numeric*	Suhu minimum dalam derajat celsius
2.	*Rainfall*	*numeric*	Jumlah curah hujan yang tercatat untuk hari itu dalam mm
3.	*Evaporation*	*numeric*	Disebut juga penguapan kelas A (mm) dalam 24 jam hingga 9 pagi
4.	*Sunshine*	*numeric*	Jumlah jam sinar matahari cerah dalam satu hari
5.	*WindGustSpeed*	*numeric*	Kecepatan (km/jam) hembusan angin terkuat dalam 24 jam hingga tengah malam

6.	*Humidity3pm*	*numeric*	Kelembapan (persen) pada jam 3 sore
7.	*Pressure3pm*	*numeric*	Tekanan atmosfer (hpa) berkurang hingga rata-rata permukaan laut pada pukul 3 sore
8.	*Cloud9am*	*numeric*	Bagian langit yang tertutup awan pada jam 9 pagi. Ini diukur dalam "oktas", yang merupakan satuan delapan. Ini mencatat berapa banyak awan
9.	*Cloud3pm*	*numeric*	Bagian langit yang tertutup awan pada jam 3 sore. Ini diukur dalam "oktas", yang merupakan satuan delapan. Ini mencatat berapa banyak awan
10.	*Temp9am*	*numeric*	Suhu (derajat C) pada jam 9 pagi

3. Hasil dan Pembahasan

3.1. Processing and Modeling Dataset

Dataset yang sudah melalui *praprocessing* kemudian diproses menggunakan aplikasi *WEKA* dan *RStudio*.

3.1.1. Clustering Menggunakan *WEKA*

Hasil pengolahan model *clustering* menggunakan *WEKA* didapatkan sebanyak 71 iterasi dengan SSE 26684.04. Kelengkapan hasil dari aplikasi *WEKA* dapat dilihat pada **Gambar 3.**

```
Final cluster centroids:
                            Cluster#
Attribute       Full Data          0          1
                (142193.0) (94104.0) (48089.0)
=============================================
MinTemp            0.4879    0.4762    0.5107
Rainfall           0.0063    0.0087    0.0016
Evaporation        0.0377    0.0344    0.0441
Sunshine           0.5259     0.453    0.6683
WindGustSpeed      0.2634    0.2627    0.2649
Humidity3pm        0.5148    0.5966    0.3548
Pressure3pm        0.6105     0.615    0.6018
Cloud9am            0.493    0.6142     0.256
Cloud3pm           0.5004    0.5989    0.3074
Temp9am            0.5103    0.4831    0.5635
```

ISSN: 1979-2328

```
Clustered Instances

0       94104 ( 66%)
1       48089 ( 34%)


Class attribute: RainTomorrow
Classes to Clusters:

     0     1  <-- assigned to cluster
 65327 44989 | No
 28777  3100 | Yes

Cluster 0 <-- Yes
Cluster 1 <-- No

Incorrectly clustered instances :       68427.0   48.1226 %
```

Gambar 3. Kelengkapan Hasil dari Aplikasi *WEKA*

3.1.2. *Clustering* Menggunakan RStudio

Sebelum *dataset* diolah, jarak antar data harus diseragamkan terlebih dahulu agar hasil perhitungan menjadi stabil. Penyeragaman jarak data disebut dengan normalisasi. Hasil dari normalisasi adalah jarak data yang berkisar 0 sampai 1. Hal ini dilakukan pada semua atribut yang akan digunakan pada kasus klasterisasi.

Setelah itu, yang selanjutnya dilakukan adalah pemilihan jumlah *cluster* yang ideal. Berikut daftar jumlah *cluster* setelah diuji coba dengan *RStudio* yang ditunjukan pada **Tabel 4.**

Tabel 4. Mencari *Cluster* Ideal

Jumlah *Cluster*	*Sum of Squares By Cluster*
8 *Clusters*	59.3%
7 *Clusters*	57.1%
6 *Clusters*	53.4%
5 *Clusters*	52.1%
4 *Clusters*	48.1%
3 *Clusters*	41.1%
2 *Clusters*	28.0%

Pada proses pemilihan jumlah *cluster* pada Tabel 4, terlihat bahwa semakin kecil *cluster* yang digunakan, maka semakin kecil pula persentasi SSE pada *cluster* tersebut. Dari Tabel 4 juga dapat dilihat bahwa *cluster* 2 merupakan *cluster* ideal dengan SSE sebesar 28.0%.

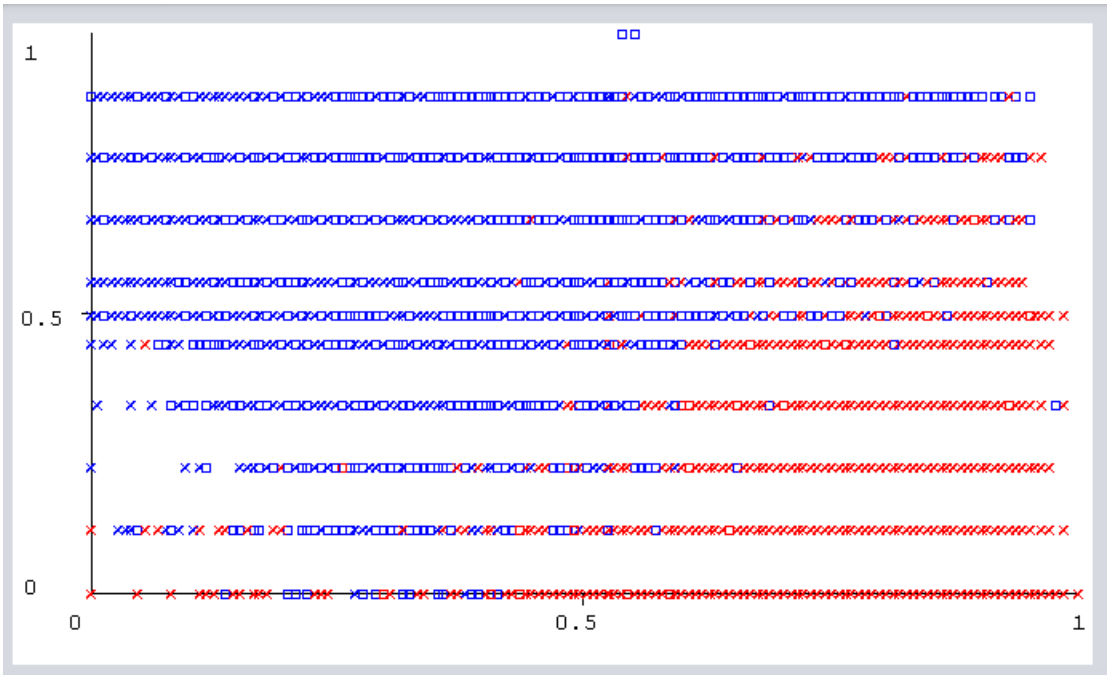
3.1.3. Visualisasi Data

Setelah proses pemilihan jumlah *cluster* ideal selesai, data yang sudah diolah tersebut akan ditampilkan menggunakan visualisasi *plot* sesuai dengan jumlah *cluster*nya. Pada *software* WEKA *cluster* hujan diwakili oleh *node* berwarna biru dan *cluster* tidak hujan diwakili oleh

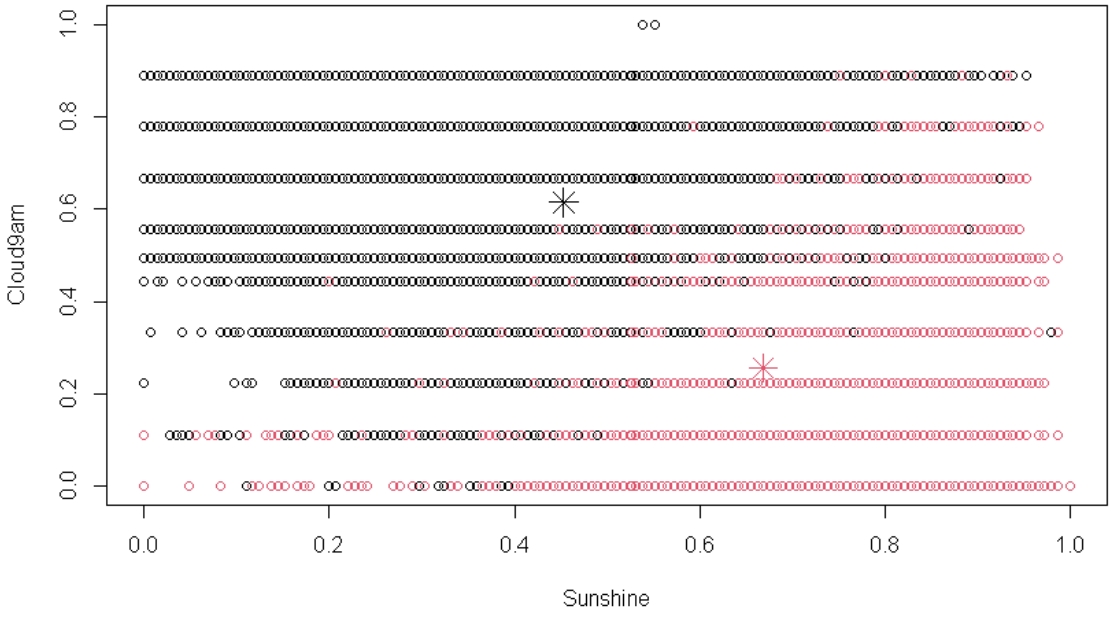
ISSN: 1979-2328

node berwarna merah. Sedangkan pada *software* RStudio *cluster* hujan diwakili oleh *node* berwarna hitam dan *cluster* tidak hujan diwakili oleh *node* berwarna merah.

Analisis pertama yang akan dilakukan adalah mengetahui pengaruh dari 2 atribut yaitu *Sunshine* dan *Cloud9am* dapat dilihat di **Gambar 4.**



Gambar 4a. Visualisasi Data dari Dua Atribut Yaitu *Sunshine dan Cloud9am* di *WEKA*

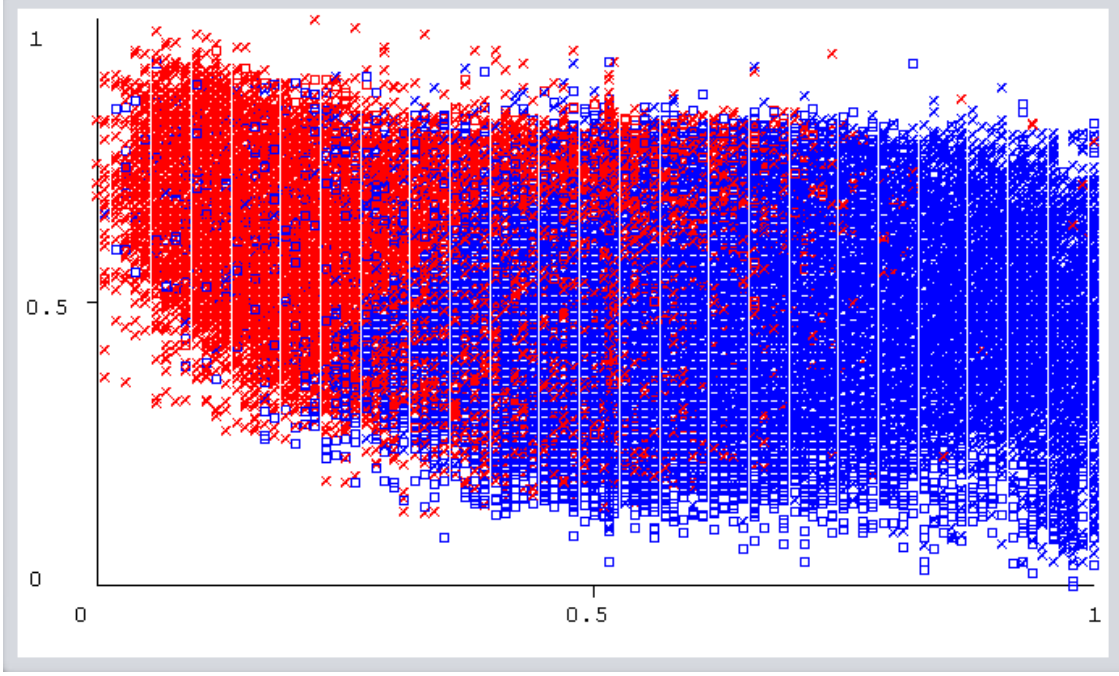


Gambar 4b. Visualisasi Data dari Dua Atribut Yaitu *Sunshine* dan *Cloud9am* di *RStudio*

ISSN: 1979-2328

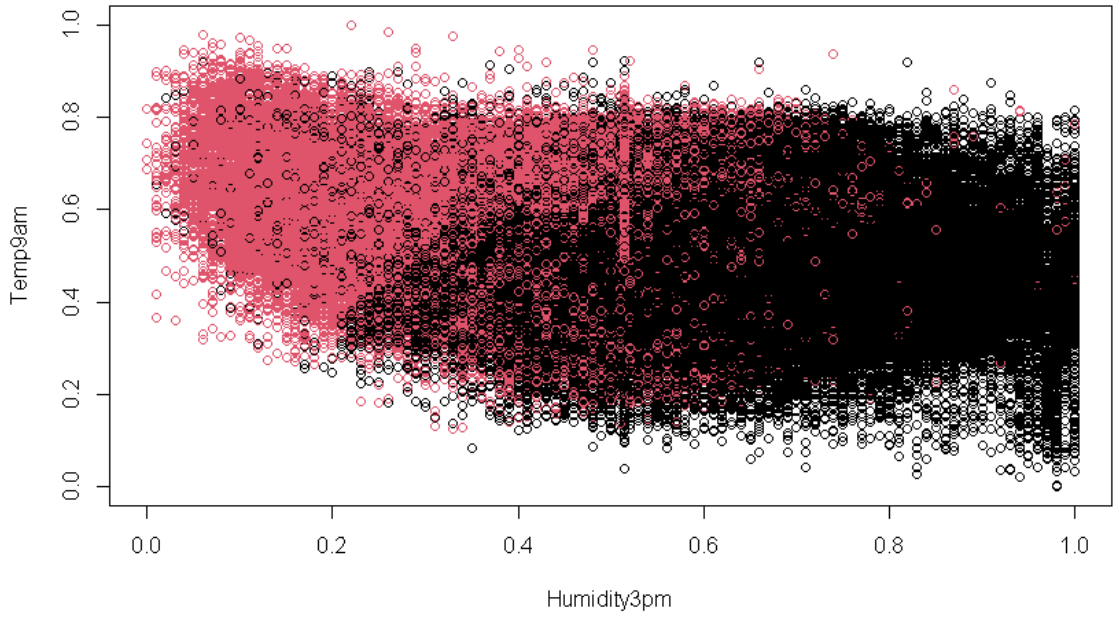
Berdasarkan Gambar 4, dapat disimpulkan bahwa semakin tinggi sinar matahari dan semakin kurangnya berawan pada hari tersebut, maka kemungkinan hujan akan terjadi semakin kecil. Begitu pula sebaliknya, semakin sedikit sinar matahari dan semakin banyaknya berawan pada hari tersebut, maka kemungkinan hujan akan terjadi semakin tinggi.

Selain menggunakan dua atribut sebelumnya, dilakukan juga pengujian menggunakan dua atribut lainnya yaitu *Humidity3pm* dan *Temp9am*. Pada *software* WEKA *cluster* hujan diwakili oleh *node* berwarna merah dan *cluster* tidak hujan diwakili oleh *node* berwarna biru. Sedangkan pada *software* RStudio *cluster* hujan diwakili oleh *node* berwarna merah dan *cluster* tidak hujan diwakili oleh *node* berwarna hitam. Pada Gambar 5 dapat disimpulkan bahwa kelembaban sangat berpengaruh pada prediksi hujan dibandingkan temperatur. Beda halnya dengan kelembaban, semakin kecil kelembaban di sore hari maka akan semakin kecil juga kemungkinan turunnya hujan pada hari itu, seperti yang terlihat di **Gambar 5.**



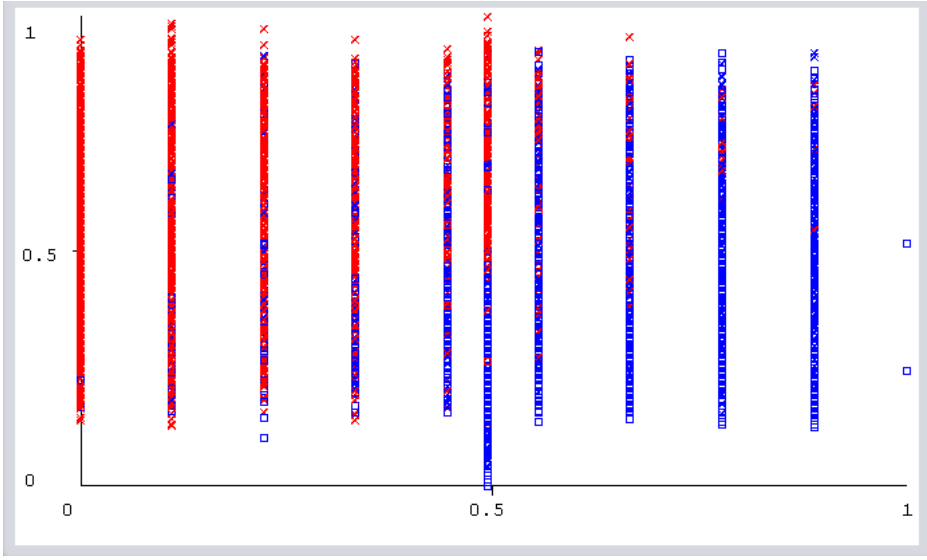
Gambar 5a. Visualisasi Data dari Dua Atribut *Humidity3pm* dan *Temp9am* di *WEKA*

ISSN: 1979-2328



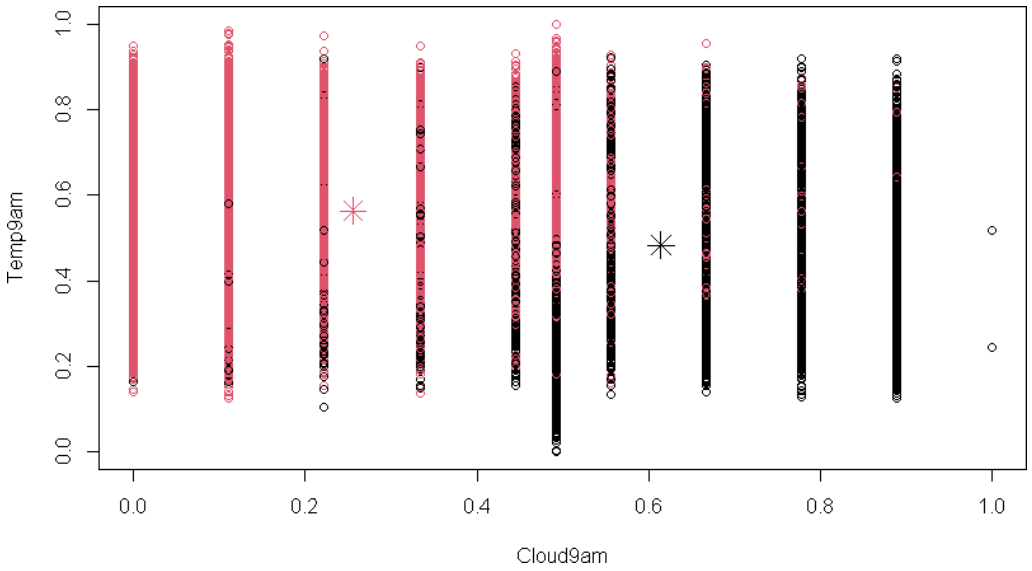
Gambar 5b. Visualisasi Data dari Dua Atribut *Humidity3pm* dan *Temp9am* di *RStudio*

Berdasarkan Gambar 5, jika pada hari tersebut cuaca sangat berawan di pagi hari, maka kemungkinan besar akan turun hujan. Hal ini sebagaimana hasil visualisasi pada **Gambar 6.**



Gambar 6a. Visualisasi Data dari Dua Atribut Yaitu *Cloud9am* dan *Temp9am* di *WEKA*

ISSN: 1979-2328



Gambar 6b. Visualisasi Data dari Dua Atribut Yaitu *Cloud9am* dan *Temp9am* di *RStudio*

Pada Gambar 6 dapat dijadikan bahan penguat dari penjelasan Gambar 5. Jika pada hari tersebut cuaca sangat berawan di pagi hari, maka kemungkinan besar akan turun hujan.

4. Kesimpulan dan Saran

Berdasarkan hasil yang diperoleh, *cluster* yang berjumlah 2 dengan SSE 28.0% merupakan *cluster* ideal untuk memprediksi hujan di Australia. Pada *software* WEKA *cluster* hujan diwakili oleh *node* berwarna biru dan *cluster* tidak hujan diwakili oleh *node* berwarna merah. Sedangkan pada *software* RStudio *cluster* hujan diwakili oleh *node* berwarna hitam dan *cluster* tidak hujan diwakili oleh *node* berwarna merah.

Untuk penelitian selanjutnya, disarankan menggunakan algoritme pengelompokan lainnya untuk ikut dibandingkan dalam prediksi hujan di Australia agar didapatkan hasil akurasi prediksi yang lebih baik dari penelitian sebelumnya. Selain itu hasil *cluster* ini juga dapat dijadikan sebagai anotasi atau label untuk penelitian selanjutnya menggunakan pendekatan *supervised learning*.

ISSN: 1979-2328

Daftar Pustaka

[1] G. Sethupathi M, Y. S. Ganesh, and M. M. Ali, "Efficient Rainfall Prediction and Analysis using Machine Learning Techniques," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 6, pp. 3467–3474, 2021.

[2] C. Thirumalai, K. S. Harsha, M. L. Deepak, and K. C. Krishna, "Heuristic prediction of rainfall using machine learning techniques," in *Proceedings - International Conference on Trends in Electronics and Informatics, ICEI 2017*, 2018, vol. 2018-Janua, pp. 1114–1117, doi: 10.1109/ICOEI.2017.8300884.

[3] S. Aftab, M. Ahmad, N. Hameed, M. S. Bashir, I. Ali, and Z. Nawaz, "Rainfall prediction in Lahore City using data mining techniques," in *International Journal of Advanced Computer Science and Applications*, 2018, vol. 9, no. 4, pp. 254–260, doi: 10.14569/IJACSA.2018.090439.

[4] A. Y. Felix, G. S. S. Vinay, and G. Akhik, "K-Means cluster using rainfall and storm prediction in machine learning technique," *J. Comput. Theor. Nanosci.*, vol. 16, no. 8, pp. 3265–3269, 2019, doi: 10.1166/jctn.2019.8174.

[5] A. M. Bagirov, A. Mahmood, and A. Barton, "Prediction of monthly rainfall in Victoria, Australia: Clusterwise linear regression approach," *Atmos. Res.*, vol. 188, pp. 20–29, 2017, doi: 10.1016/j.atmosres.2017.01.003.

[6] J. Young, "Rain in Australia," *Kagle.com*, 2018. https://www.kaggle.com/jsphyg/weather-dataset-rattle-package.

[7] J. M. Frederic Lardinois, Matthew Lynley, "Google is acquiring data science community Kaggle," 2017. .

[8] M. Nasution, "Implementasi Data Mining K-Means Untuk Mengukur Kemampuan Logika Mahasiswa (Studi Kasus : Amik Labuhan Batu)," *J. Inform.*, vol. 5, no. 1, pp. 32–37, 2019, doi: 10.36987/informatika.v5i1.667.

[9] P. Cichosz, *Data mining algorithms : explained using R*. John Wiley & Sons, Inc., 2015.

[10] X. Wu *et al.*, *Top 10 algorithms in data mining*, vol. 14, no. 1. 2008.

[11] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and a Y. Wu, "An efficient k-means clustering algorithm: analysis and implementation," *Pattern Anal. Mach. Intell. IEEE Trans.*, vol. 24, no. 7, pp. 881–892, 2002, doi: 10.1109/TPAMI.2002.1017616.

[12] A. R. Dasari, "Prediction Of Rainfall In India To Increase Agricultural Productivity Implemented In Hadoop Prediction Of Rainfall In India To Increase Agricultural Productivity Implemented In Hadoop," no. March, 2021.

[13] N. Salehnia, N. Salehnia, H. Ansari, S. Kolsoumi, and M. Bannayan, "Climate data clustering effects on arid and semi-arid rainfed wheat yield: a comparison of artificial intelligence and K-means approaches," *Int. J. Biometeorol.*, vol. 63, no. 7, pp. 861–872, 2019, doi: 10.1007/s00484-019-01699-w.

ISSN: 1979-2328

[14] Y. Cho, H. Lee, B. Lim, and S. Kim, "Classification of Weather Patterns in the East Asia Region using the K-means Clustering Analysis," *Atmosphere (Basel).*, vol. 29, no. 4, pp. 451–461, 2019, doi: 10.14191/ATMOS.2019.29.4.451.

[15] U. Kumar, "Open Access Design and Analysis of Multiclass Classification Models for Rainfall Prediction," *Res. J. Comput. Sci. Inf. Technol.*, vol. 1, no. 1, pp. 23–34, 2018.

[16] Noname, "Weka - Clustering," *tutorialspoint.com*, 2021. https://www.tutorialspoint.com/weka/weka_clustering.htm (accessed Sep. 20, 2021).

[17] W. KENTON, "Sum of Squares," 2020. .

[18] State of New York, "New York State Index Crimes," *Kagle.com*, 2019. https://www.kaggle.com/new-york-state/new-york-state-index-crimes/metadata.