*ISSN: 1979-2328*

# *Application of Data Mining for Rainfall Prediction Classification in Australia with Decision Tree Algorithm and C5.0 Algorithm*

Penerapan *Data Mining* Untuk Klasifikasi Prediksi Hujan di Australia dengan Algoritme *Decision Tree* dan Algoritme C5.0

**Irwansyah Saputra[1], Dinar Ajeng Kristiyanti[2]**

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

[1*]92irwansyah@apps.ipb.ac.id, [2*]dinarajengkristiyanti@apps.ipb.ac.id

*Keywords:Classification;Decision Tree; C5.0; Rstudio; Rain Australia*

***Abstract***

***Purpose:*** *This study aims to predict rain in Australia with a machine learning classification approach. Precise and accurate rain prediction is very important for planning and management of water resources, flood warning, construction activities and aviation operations, and others.*
***Design/methodology/approach:*** *The methods or stages applied in classifying rain predictions in Australia are through several stages including Data Collection, Data Pre-processing (including Missing Value handling in it), Classification Modeling by applying and comparing Decision Tree and C5.0 algorithms, Result validation using Dataset Partition and 10-Cross Fold Validation and Model Evaluation using Confusion Matrix.*
***Findings/result:*** *The methods or stages applied in classifying rain predictions in Australia are through several stages including Data Collection, Data Pre-processing (including Missing Value handling in it), Classification Modeling by applying and comparing Decision Tree and C5.0 algorithms, Result of validation using Dataset Partition and 10-Cross Fold Validation and Model Evaluation using Confusion Matrix.*
***Originality/value/state of the art:*** *In addition to the classification model used, dataset validation, either by partitioning the dataset or by 10-cross fold validation, can also affects the accuracy of the prediction results.*

*ISSN: 1979-2328*

Kata kunci:Klasifikasi; *Decision Tree*; C5.0; RStudio; Hujan Australia

**Abstrak**

**Tujuan:** Penelitian ini bertujuan untuk memprediksi hujan di Australia dengan pendekatan klasifikasi *machine learning*. Prediksi hujan yang tepat dan akurat sangat penting untuk perencanaan dan pengelolaan sumber daya air, peringatan banjir, kegiatan konstruksi dan operasi penerbangan serta yang lainnya.

**Perancangan/metode/pendekatan:** Metode atau tahapan yang diterapkan dalam melakukan klasifikasi prediksi hujan di Australia yaitu melalui beberapa tahapan diantaranya Pengumpulan Data, *Data Pre-processing* (termasuk dilakukan penanganan *Missing Value* didalamnya), Pemodelan Klasifikasi dengan menerapkan dan membandingkan algoritme *Decision Tree* dan C5.0, Validasi Hasil menggunakan Partisi *Dataset* dan *k-Cross Fold Validation* serta Evaluasi Model menggunakan *Confussion Matrix.*

**Hasil:** Berdasarkan hasil yang diperoleh, evaluasi menggunakan 10-*Cross Fold Validation* lebih unggul yang memiliki akurasi paling tinggi sebesar 87.35% untuk algoritme *Decision Tree* dan akurasi sebesar 86.85% untuk algoritme C5.0 *Rule-Based Model*, dibandingkan dengan metode Split 80:20 pada kasus prediksi hujan di Australia.

**Keaslian/*state of the art*:** Selain model klasifikasi yang digunakan, validasi *dataset* baik itu dengan partisi *dataset* atau *k-Cross Fold Validation* juga dapat mempengaruhi akurasi hasil prediksi.

## 1. Pendahuluan

Australia adalah sebuah benua dan negara yang terdiri dari beberapa negara bagian. Iklim di delapan negara bagian Australia sangat bervariasi. Sebagian besar wilayah Australia memiliki empat musim yaitu musim panas, musim dingin, musim gugur, dan musim semi [1]. Curah hujan di Australia sangat beragam, karena benua ini memiliki iklim yang beragam, dan Australia bagian utara memiliki iklim tropis. Kemudian, bagian barat daya dan pesisir selatan beriklim subtropis. Australia barat penuh dengan gurun, iklimnya kering, sangat panas di siang hari, tetapi sangat dingin di malam hari. Terakhir, pantai timur Australia memiliki iklim laut, sehingga daerah tersebut memiliki curah hujan yang cukup tinggi sepanjang tahun [2].

Prediksi hujan yang akurat merupakan salah satu tugas yang paling menantang dan penting di dunia saat ini [3], tidak terkecuali Australia karena memiliki iklim yang beragam [4]. Hal yang membuat menantang dikarenakan hujan bersifat dinamis dari fenomena iklim dan fluktuasi acak yang terlibat dalam proses fisik [5]. Biasanya prediksi hujan dibuat untuk beberapa periode waktu yang meliputi mingguan, bulanan dan prediksi musiman. Prediksi hujan yang tepat dan akurat sangat penting untuk perencanaan dan pengelolaan sumber daya air, peringatan banjir, kegiatan konstruksi dan operasi penerbangan serta yang lainnya [6]. Agar hasil prediksi hujan

menjadi optimal, berbagai kompleksitas perlu ditangani [7], seperti data statistik cuaca yang memiliki banyak fitur diantaranya kelembaban, tekanan, kecepatan angin, polutan, konsentrasi, dan lainnya.

Pendekatan *machine learning* telah banyak dilakukan dalam memprediksi hujan. Penerapannya yaitu dengan menggunakan metode komputasi dan memprediksi hujan dengan mengambil dan mengintegrasikan pengetahuan tersembunyi dari pola *linear* dan *non-linear* dari data cuaca masa lalu [8]. Teknik klasifikasi akan menghasilkan seperangkat aturan yang disebut aturan yang akan digunakan sebagai indikator untuk dapat memprediksi kelas data yang ingin diprediksi [9]. Beberapa penelitian telah banyak dilakukan untuk memprediksi hujan dengan pendekatan *machine learning* dengan menerapkan beragam algoritme klasifikasi diantaranya *Random Forest* [8][10], *Logistic Regression* [8][11] dan *Support Vector Regression* [11]. Bahkan pada penelitian [12], prediksi hujan dilakukan dengan membandingkan beragam *classifier* seperti *Naïve Bayes*, *Support Vector Machine*, *Decision Tree*, *Neural Network* dan *Random Forest*. Pada penelitian lainnya [13][14][15][16][17][18], *Decision Tree* telah banyak digunakan dalam prediksi hujan. Algoritme *Decision Tree* merupakan alat pengambilan keputusan terbaik yang bersifat sederhana dalam struktur, namun memiliki performa yang baik dalam analisis beberapa variabel [14]. Algoritme C5.0 juga telah diterapkan untuk memprediksi hujan [19]. Algoritme C5.0 adalah algoritme klasifikasi yang dapat diterapkan pada kumpulan data yang besar [19]. Algoritme C5.0 memiliki efisiensi dan memori yang lebih baik daripada algoritme C4.5 [20].

Data hujan di Australia dirilis oleh seorang ilmuan data bernama Joe Young, dan telah diunggah ke situs kaggle dengan judul *dataset Rain in Australia* [21]. Kaggle adalah komunitas *online* ilmuwan 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 [22]. Penelitian ini akan berfokus untuk memprediksi hujan di Australia dengan membandingkan akurasi kedua algoritme *Decision Tree* dan algoritme C5.0. Verifikasi hasil dibagi menjadi dua bagian, pertama, kumpulan data dibagi sebanyak 80% sebagai data latih dan 20% sebagai data uji. Kedua, *K-Fold Cross-Validation* digunakan sebanyak 10 kali untuk validasi.

## 2. Metode/Perancangan

Metode atau tahapan yang diterapkan dalam melakukan klasifikasi prediksi hujan di Australia yaitu melalui beberapa tahapan diantaranya Pengumpulan Data, *Data Pre-processing* (termasuk dilakukan penanganan *Missing Value* didalamnya), Pemodelan Klasifikasi dengan menerapkan dan membandingkan algoritme *Decision Tree* dan C5.0, Validasi Hasil menggunakan Partisi *Dataset* dan 10-*Cross Fold Validation* serta Evaluasi Model menggunakan *Confussion Matrix.*

### 2.1. Pengumpulan Data

*Dataset* yang digunakan dalam penelitian ini adalah *Rain in Australia* dari kaggle.com *Repository*. *Dataset* diambil dari situs https://www.kaggle.com/jsphyg/weather-dataset-rattle-package yang merupakan data yang berisikan pengamatan cuaca harian dari berbagai stasiun cuaca Australia. *Dataset Rain in Australia* terdiri dari satu *file rar* yang berisi satu *file dataset* dengan nama weatherAUS.csv. Bentuk format *file* dalam *dataset* adalah *csv* atau *comma separated values* yang berarti data didalamnya dipisahkan oleh koma. *Dataset* tersebut terdiri dari 24 atribut dan 142.193 *record*. Penjelasan dari setiap atribut dapat dilihat pada **Tabel 1.**

*ISSN: 1979-2328*

**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 |

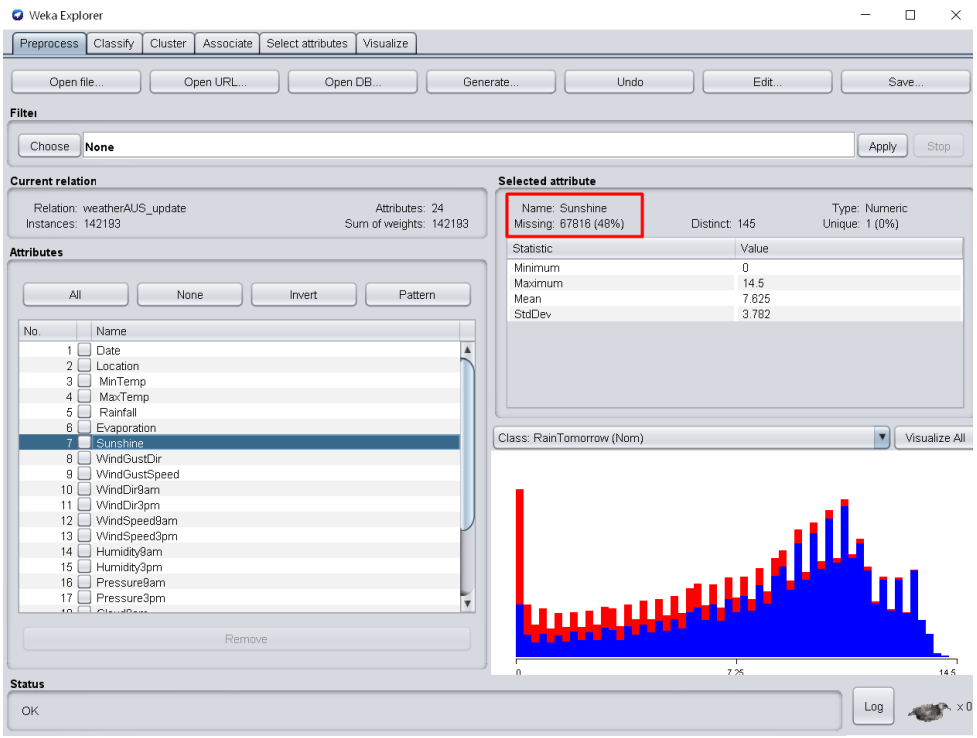
Selain itu, *dataset* tersebut memiliki missing *value* berkisar dari 1% hingga 48% yang terbesar pada beberapa atribut. Cara yang dapat dilakukan untuk menangani *missing value* adalah menghapus *tupel missing value* atau mengisinya dengan nilai konstanta berdasarkan rata-rata dari nilai dalam data tersebut [23]. *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 *record*nya 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 iniadalah *Dataset Rain in Australia* yang terdiri dari 24 atribut [24].

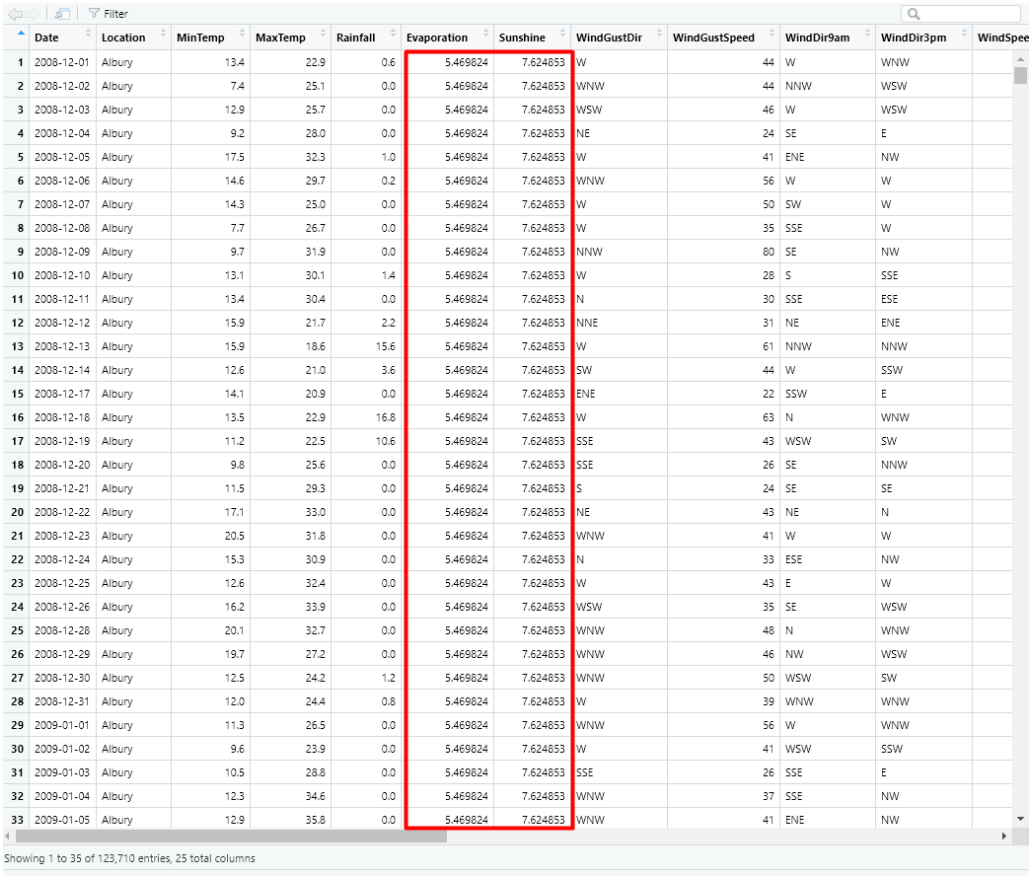
Proses pembersihan data dari *noise* dapat dilakukan sesuai kebutuhan karena *noise* data banyak ragamnya. Contohnya, pada *dataset Rain in Australia* terdapat *record* yang memiliki *missing value* di beberapa atribut sehingga tidak akan optimal saat diproses untuk diekstraksi pada

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* didalamnya 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*



Filter

| | 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. Sebelum *dataset* diolah, tipe data *chr* pada beberapa atribut akan diganti dengan atribut *factor* agar dapat bekerja dengan baik pada aplikasi *Rstudio.*

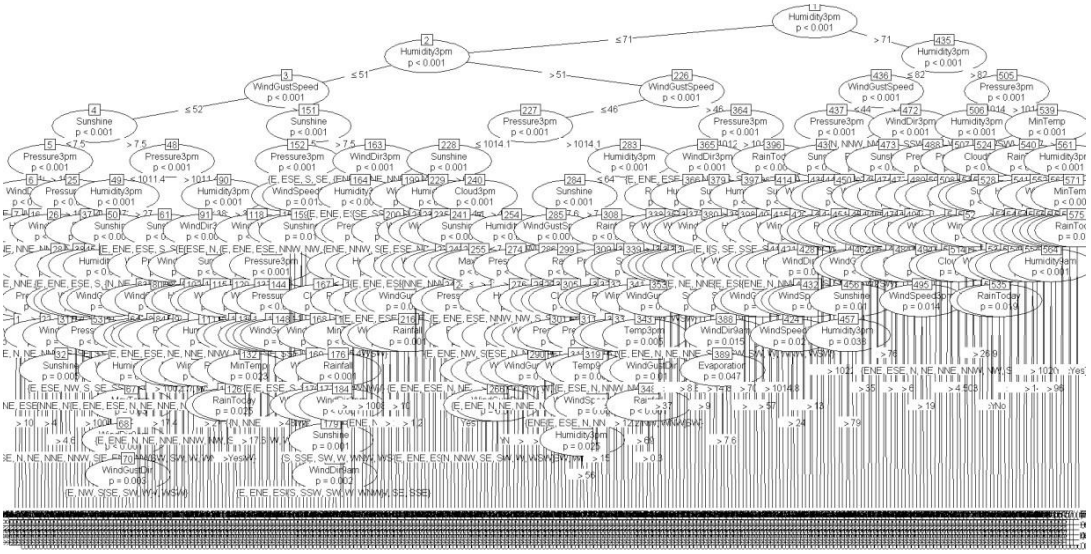
## 3. Hasil dan Pembahasan

### 3.1. *Processing and Modelling Dataset*

*Dataset* yang sudah melalui *praprocessing* kemudian diproses menggunakan aplikasi *RStudio*. Algoritme *Decision Tree* akan diterapkan pada percobaan yang pertama. Validasi pertama adalah dengan mempartisi *dataset* ke dalam dua bagian, yaitu 80% untuk data latih dan 20% untuk data uji. Validasi kedua adalah dengan menggunakan *10-Cross Fold Validation.*

#### 3.1.1. Algoritme *Decision Tree* dengan Partisi *Dataset 80:20*

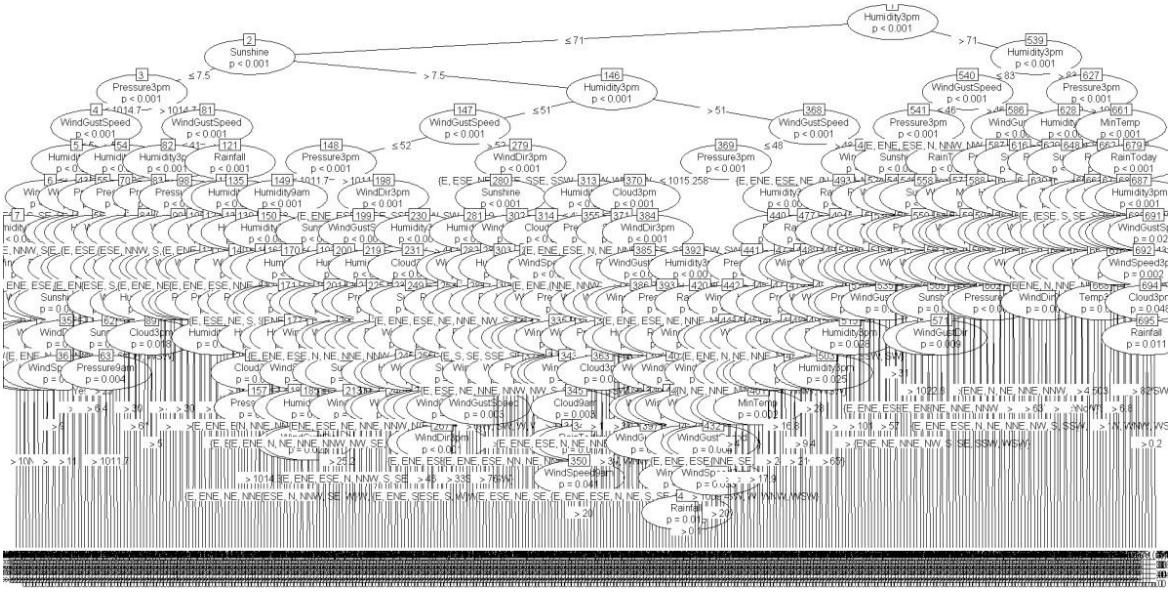
*Library Party* yang terdapat pada *RStudio* dapat digunakan untuk menyelesaikan pemodelan algoritme *Decision Tree*. Sebanyak 80% partisi data latih dimodelkan dan diujikan terhadap 20% data yang sudah disiapkan dari *dataset*. *Decision Tree* yang dihasilkan dari pemodelan dapat dilihat pada **Gambar 3.**



**Gambar 3.** Pemodelan *Decision Tree* dengan Partisi Dataset 80:20

### 3.1.2. Algoritme *Decision Tree* dengan *10-Cross Fold Validation*

Pemodelan berikutnya menggunakan algoritme *Decision Tree* dengan teknik *10-Cross Fold Validation*. Dengan *dataset* yang sama, *Decision Tree* yang dihasilkan dari pemodelan dapat dilihat pada **Gambar 4.**



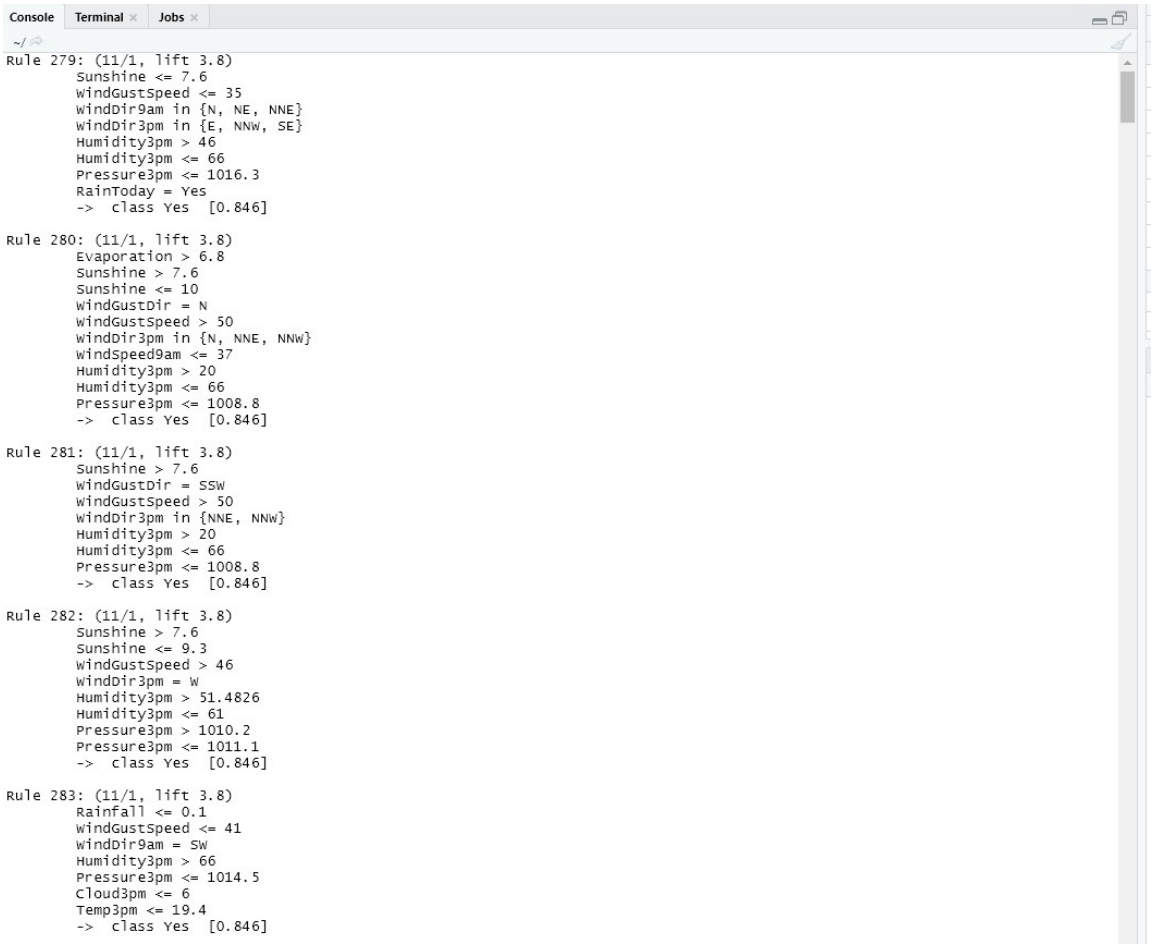
**Gambar 4.** Pemodelan *Decision Tree* dengan Teknik 10-Cross Fold Validation

### 3.1.3. Algoritme C5.0 dengan Partisi *Dataset* 80:20

*Library* C50 yang terdapat pada *RStudio* dapat digunakan untuk menyelesaikan pemodelan algoritme C5.0. Sebanyak 80% partisi data latih dimodelkan dan diujikan terhadap 20% data yang sudah disiapkan dari *dataset*.

*ISSN: 1979-2328*

Selain menghasilkan pohon keputusan, algoritme C5.0 juga menghasilkan aturan sebagai hasil dari pemodelan yang sudah dilakukan. Aturan tersebut berjumlah 373 aturan. Tampilan *Console* pada *RStudio* tidak memungkinkan untuk menampilkan hasil aturan tersebut, sehingga dapat dilihat sebagian contoh aturan tersebut pada **Gambar 5.**

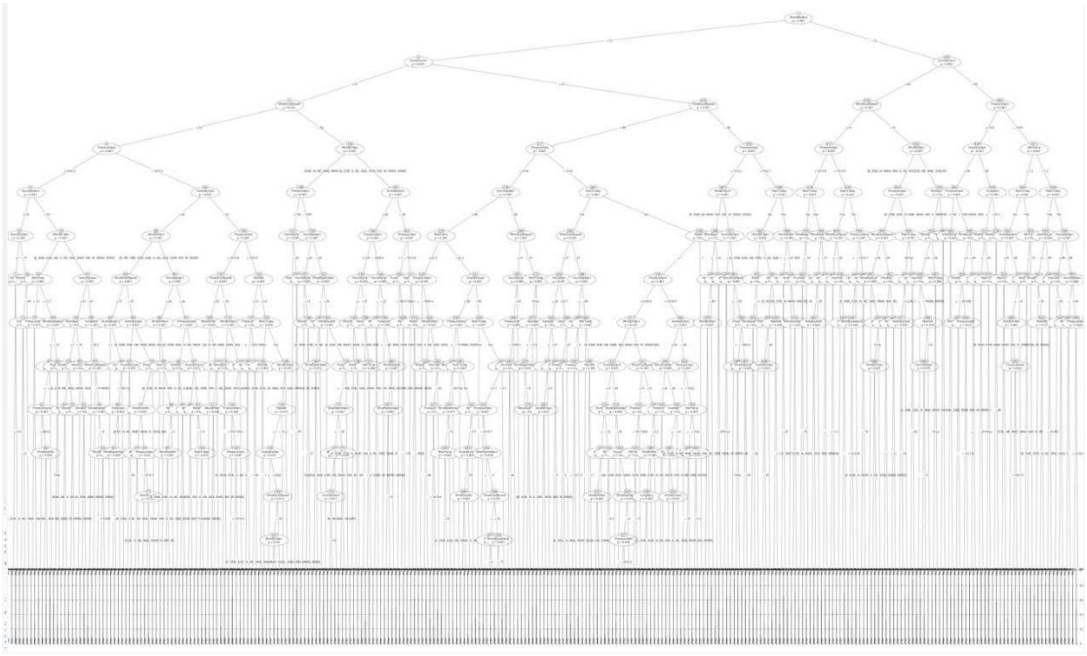



**Gambar 5.** Aturan yang Dihasilkan Algoritme C5.0 dengan Partisi *Dataset* 80:20

### 3.1.4. Algoritme C5.0 dengan 10-*Cross Fold Validation*

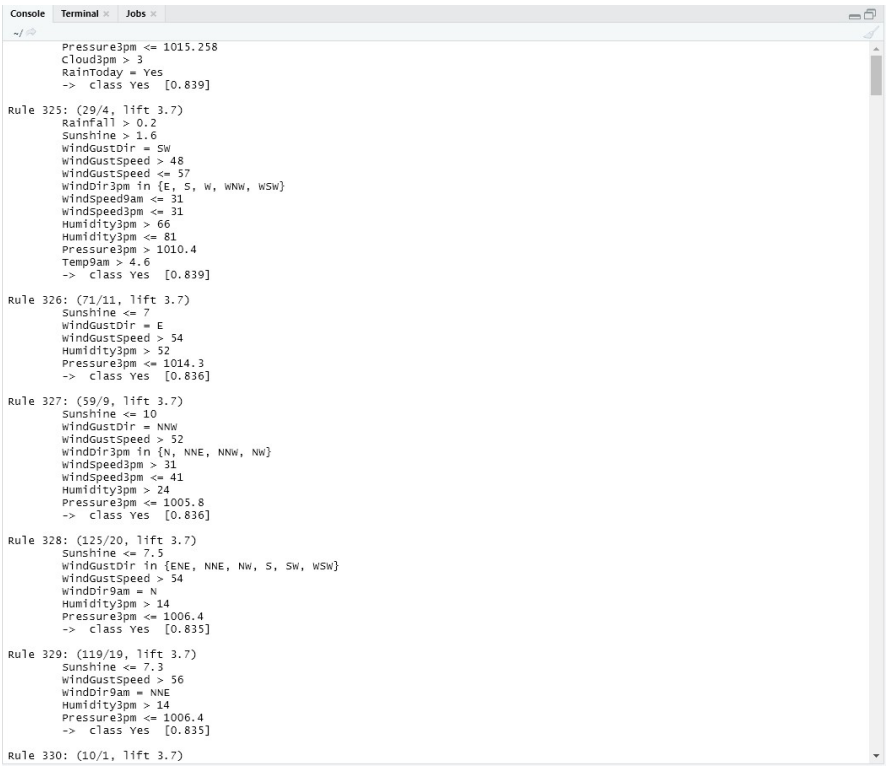
Pemodelan berikutnya menggunakan algoritme C5.0 dengan teknik 10-*Cross Fold Validation.* Dengan *dataset* yang sama, *Decision Tree* yang dihasilkan dari pemodelan algoritme C5.0 dapat dilihat pada **Gambar 6.**

*ISSN: 1979-2328*



**Gambar 6.** Algoritme *Decision Tree* dan Algoritme C5.0 dengan Teknik 10-*Cross Fold Validation*

Selain menghasilkan *Decision Tree*, algoritme C5.0 juga menghasilkan aturan sebagai hasil dari pemodelan yang sudah dilakukan. Aturan tersebut berjumlah 417 aturan. Tampilan *Console* pada *RStudio* tidak memungkinkan untuk menampilkan hasil aturan tersebut, sehingga dapat dilihat sebagian contoh aturan tersebut pada **Gambar 7.**



**Gambar 7.** Aturan yang Dihasilkan Algoritme C5.0 dengan Teknik 10-Cross Fold Validation

### 3.2. *Evaluation*

Setelah pemodelan menggunakan metode algoritme *Decision Tree* dan Algoritme C5.0 selesai, selanjutnya akan dijelaskan mengenai *Confusion Matrix* dari masing-masing algoritme. Dalam sistem klasifikasi, *Confusion matrix* berisi informasi mengenai klasifikasi aktual dan prediksi. Performa dari sistem tersebut pada umumnya dilakukan evaluasi menggunakan data dalam matriks. Terdapat dua kelas pada *Confusion matrix* yang mengklasifikasikan kelas positif dan kelas negatif [25] ditunjukan oleh **Tabel 3.**

**Tabel 3.** *Confusion Matrix*

| | | **Aktual** | |
|---|---|---|---|
| | | **Negatif** | **Positif** |
| **Prediksi** | **Negatif** | **a** | **b** |
| | **Positif** | **c** | **d** |

Keterangan:
a adalah jumlah prediksi yang benar bahwa sebuah *instance* negatif,

b adalah jumlah prediksi yang salah bahwa sebuah *instance* positif,
c adalah jumlah salah prediksi yang sebuah contoh negatif, dan
d adalah jumlah prediksi yang benar bahwa sebuah *instance* bernilai positif.

#### 3.2.1. Akurasi Algoritme *Decision Tree* dan Algoritme C5.0

*Confusion Marix* dari setiap pemodelan yang sudah dilakukan, yaitu dari algoritme *Decision Tree* dan algoritme C5.0 dapat dilihat pada **Tabel 4-9.**

**Tabel 4.** *Confusion Matrix* Algoritme *Decision Tree* Split 80:20 Terhadap Data Uji

| **Accuracy: 84,55%** | | |
|---|---|---|
| | **No** | **Yes** |
| **No** | 73206 | 10535 |
| **Yes** | 3962 | 11397 |

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{11397+73206}{73206+10535+3962+11397} = 0.853713421$$

**Tabel 5.** *Confusion Matrix* Algoritme *Decision Tree* 10-*Cross Fold Validation*

| **Accuracy: 87,35%** | | |
|---|---|---|
| | **No** | **Yes** |
| **No** | 9690 | 1111 |
| **Yes** | 454 | 1116 |

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{1116 + 9690}{9690 + 1111 + 454 + 1116} = 0.873494463$$

**Tabel 6.** *Confusion Matrix* Algoritme C5.0 Split 80:20 Terhadap Data Uji (*Tree-BasedModel*)

| **Accuracy: 83,99%** | | |
|---|---|---|
| | **No** | **Yes** |
| **No** | 17864 | 2654 |
| **Yes** | 1286 | 2806 |

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{2806 + 17864}{2806 + 17864 + 1286 + 2654} = 0.839902479$$

**Tabel 7.** *Confusion Matrix* Algoritme C5.0 Split 80:20 Terhadap Data Uji (*Rule-BasedModel*)

| **Accuracy: 84,58%** | | |
|---|---|---|
| | **No** | **Yes** |
| **No** | 18018 | 2664 |
| **Yes** | 1132 | 2796 |

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{2796 + 18018}{18018 + 2664 + 1132 + 2796} = 0.845753759$$

**Tabel 8.** *Confusion Matrix* Algoritme C5.0 10-*Cross Fold Validation* (*Tree-Based Model*)

| **Accuracy: 85,95%** | | |
|---|---|---|
| | **No** | **Yes** |
| **No** | 9475 | 1069 |
| **Yes** | 669 | 1158 |

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{1158 + 9475}{9475 + 1069 + 669 + 1158} = 0.859510145$$

**Tabel 9.** *Confusion Matrix* Algoritme C5.0 10-*Cross Fold Validation* (*Rule-Based Model*)

| **Accuracy: 86,85%** | | |
|---|---|---|
| | **No** | **Yes** |
| **No** | 9608 | 1091 |
| **Yes** | 536 | 1136 |

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{1136 + 9608}{9608 + 1091 + 536 + 1136} = 0.868482742$$

*ISSN: 1979-2328*

Semua hasil tersebut dapat digabungkan dalam satu tabel, yaitu **Tabel 10.**

**Tabel 10.** *Confusion Matrix* Algoritme *Decision Tree* dan Algoritme C5.0

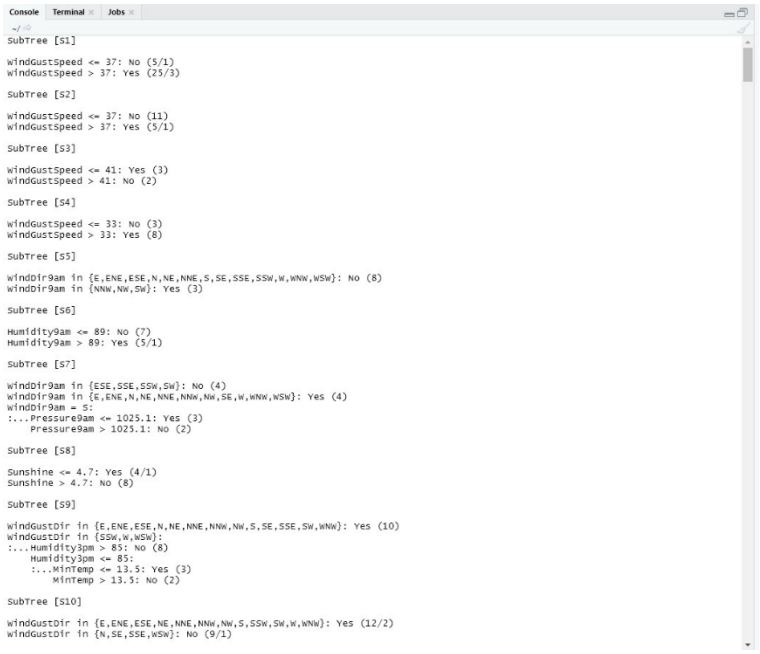
| No. | Algoritme | Akurasi |
|---|---|---|
| 1. | *Decision Tree* (*Split* 80:20) | 84.55% |
| 2. | *Decision Tree* (10-*Cross Fold Validation*) | 87.35% |
| 3. | C5.0 *Tree-Based Model* (*Split* 80:20) | 83.99% |
| 4. | C5.0 *Rule-Based Model* (*Split* 80:20) | 84.58% |
| 5. | C5.0 *Tree-Based Model* (10-*Cross Fold Validation*) | 85.95% |
| 6 | C5.0 *Rule-Based Model* (10-*Cross Fold Validation*) | 86.85% |

### 3.2.2. Evaluasi Model Split 80:20 dengan 10-Cross Fold Validation

Pada Tabel 10 dapat disimpulkan bahwa dalam kasus klasifikasi prediksi hujan di Australia, skema model menggunakan 10-*Cross Fold Validation* lebih unggul dalam semua kasus dibandingkan dengan model split 80:20. Keunggulan 10-*Cross Fold Validation* dialami oleh kedua algoritme yaitu algoritme *Decision Tree* dan algoritme C5.0.

### 3.2.3. Hasil Aturan dari Model yang Memiliki Akurasi Paling Tinggi yaitu C5.0 10-Cross Fold Validation

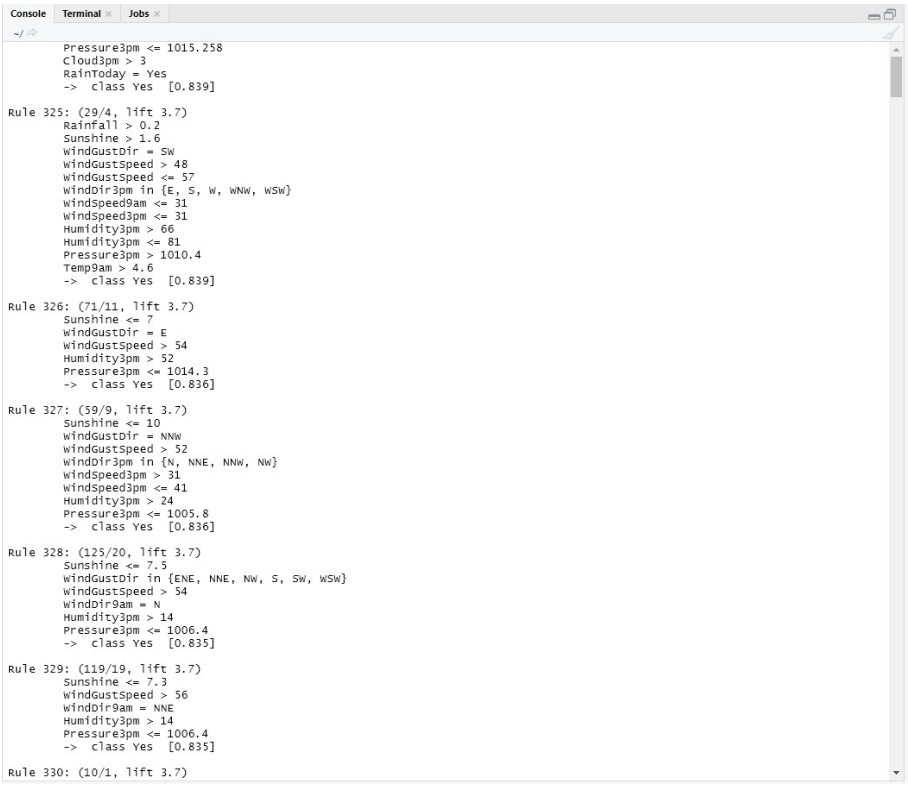
Hasil aturan dari model algoritme C5.0 dengan Teknik 10-*Cross Fold Validation* (*Tree- Based Model*) dapat dilihat pada **Gambar 8.**




**Gambar 8.** Contoh Aturan yang Dihasilkan Algoritme C5.0 dengan Teknik 10-*Cross FoldValidation* (*Tree-Based Model*)

*ISSN: 1979-2328*

Sedangkan hasil aturan dari model algoritme C5.0 dengan Teknik 10-*Cross Fold Validation* (*Rule-Based Model*) dapat dilihat pada **Gambar 9.**




**Gambar 9.** Contoh Aturan yang Dihasilkan Algoritme C5.0 dengan Teknik 10-*Cross FoldValidation* (*Rule-Based Model*)

### 3.2.4. Aturan-Aturan Berpengaruh dari Model yang Memiliki Akurasi Paling Tinggi yaitu C5.0 10-*Cross Fold Validation*

Aturan-aturan yang berpengaruh dalam menentukan label kelas berdasarkan model berbasis aturan dan model berbasis *tree* dapat dilihat pada **Gambar 10.**

| | id | LHS | RHS | support | confidence |
|---|---|---|---|---|---|
| | <int> | <chr> | <chr> | <int> | <dbl> |
| 1 | 1 | Rainfall <= 0.2 & Sunshine > 10.3 & WindDir3pm %in% c('E'~ | No | 12771 | 0.975 |
| 2 | 2 | MinTemp <= 10.6 & Rainfall <= 0.5 & Sunshine > 8.2 & Pres~ | No | 7231 | 0.974 |
| 3 | 3 | Rainfall <= 2.9 & WindGustSpeed <= 46 & WindDir9am == 'S'~ | No | 71 | 0.973 |
| 4 | 4 | WindGustSpeed <= 41 & WindDir9am == 'SE' & Cloud3pm <= 3 | No | 1438 | 0.956 |
| 5 | 5 | Rainfall <= 2.9 & WindGustSpeed <= 46 & WindDir3pm == 'NW~ | No | 428 | 0.942 |
| 6 | 6 | MinTemp <= 19.9 & WindGustDir == 'S' & WindDir3pm %in% c(~ | No | 75 | 0.935 |
| 7 | 7 | Rainfall <= 2.9 & WindGustDir %in% c('SSW', 'W', 'WSW') &~ | No | 1625 | 0.935 |
| 8 | 8 | WindDir9am == 'SW' & Humidity3pm <= 95 & Pressure3pm > 10~ | No | 146 | 0.932 |
| 9 | 9 | MinTemp <= 4.4 & WindDir9am == 'E' & Pressure3pm > 1014.9~ | No | 723 | 0.932 |
| 10 | 10 | Rainfall <= 2.2 & WindGustSpeed <= 26 & Cloud3pm <= 5 | No | 9569 | 0.926 |

**Gambar 10.** Aturan-Aturan Berpengaruh dari Model yang Memiliki Akurasi Paling Tinggi yaitu C5.0 10-*Cross Fold Validation*

## 4. Kesimpulan dan Saran

Berdasarkan hasil yang diperoleh pada penelitian, evaluasi menggunakan 10-*Cross Fold Validation* lebih unggul yaitu memiliki akurasi paling tinggi sebesar 87.35% untuk algoritme

*ISSN: 1979-2328*

*Decision Tree* dan akurasi sebesar 86.85% untuk algoritme C5.0 *Rule-Based Model*, dibandingkan dengan metode Split 80:20 pada kasus prediksi hujan di Australia. Meskipun hasil akurasi yang diperoleh untuk kedua metode tersebut rata-rata di atas 80.00%.

Untuk penelitian selanjutnya, disarankan menggunakan algoritme klasifikasi lainnya untuk ikut dibandingkan dalam prediksi hujan di Australia. Sehingga dapat diperoleh hasil akurasi prediksi yang lebih baik dari penelitian sebelumnya.

## Daftar Pustaka

[1] Anonim, "Cuaca di Australia," *Tourism Australia*, 2020. http://australiaxy.com/id-id/facts-and-planning/weather-in-australia.html.

[2] P. Yasmin, "Karakteristik, Iklim dan Daftar Negara di Benua Australia," *Detiktravel*. Detik.com, 2020, [Online]. Available: https://travel.detik.com/travel-news/d-5158330/karakteristik-iklim-dan-daftar-negara-di-benua-australia.

[3] 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-January, pp. 1114–1117, doi: 10.1109/ICOEI.2017.8300884.

[4] 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.

[5] A. M. Bagirov and A. Mahmood, "A Comparative Assessment of Models to Predict Monthly Rainfall in Australia," *Water Resour. Manag.*, vol. 32, no. 5, pp. 1777–1794, 2018, doi: 10.1007/s11269-018-1903-y.

[6] 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.

[7] M. P. Darji, V. K. Dabhi, and H. B. Prajapati, "Rainfall forecasting using neural network: A survey," in *Conference Proceeding - 2015 International Conference on Advances in Computer Engineering and Applications, ICACEA 2015*, 2015, no. December, pp. 706–713, doi: 10.1109/ICACEA.2015.7164782.

[8] 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.

[9] D. A. Kristiyanti, E. Purwaningsih, E. Nurelasari, A. Al Kaafi, and A. H. Umam, "Implementation of Neural Network Method for Air Quality Forecasting in Jakarta Region," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012037.

[10] A. Dikshit, B. Pradhan, and A. M. Alamri, "Short-term spatio-temporal drought

forecasting using random forests model at New South Wales, Australia," *Appl. Sci.*, vol. 10, no. 12, 2020, doi: 10.3390/app10124254.

[11] V. A. Vuyyuru, G. Apparao, and S. Anuradha, "Prediction Of Rainfall With A Machine Learning Approach," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 7, pp. 1762–1776, 2021.

[12] S. Zainudin, D. S. Jasim, and A. A. Bakar, "Comparative analysis of data mining techniques for malaysian rainfall prediction," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 6, no. 6, pp. 1148–1153, 2016, doi: 10.18517/ijaseit.6.6.1487.

[13] M. Marjanović, M. Krautblatter, B. Abolmasov, U. Đurić, C. Sandić, and V. Nikolić, "The rainfall-induced landsliding in Western Serbia: A temporal prediction approach using Decision Tree technique," *Eng. Geol.*, vol. 232, no. February 2017, pp. 147–159, 2018, doi: 10.1016/j.enggeo.2017.11.021.

[14] A. Geetha and G. M. Nasira, "Data mining for meteorological applications: Decision trees for modeling rainfall prediction," in *2014 IEEE International Conference on Computational Intelligence and Computing Research, IEEE ICCIC 2014*, 2015, pp. 0–3, doi: 10.1109/ICCIC.2014.7238481.

[15] R. N, S. S, and K. S, "Comparison of Decision Tree Based Rainfall Prediction Model with Data Driven Model Considering Climatic Variables," *Irrig. Drain. Syst. Eng.*, vol. 05, no. 03, 2016, doi: 10.4172/2168-9768.1000175.

[16] J. Dou *et al.*, "Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan," *Sci. Total Environ.*, vol. 662, pp. 332–346, 2019, doi: 10.1016/j.scitotenv.2019.01.221.

[17] W. Wei, Z. Yan, and P. D. Jones, "A decision-tree approach to seasonal prediction of extreme precipitation in eastern China," *Int. J. Climatol.*, vol. 40, no. 1, pp. 255–272, 2020, doi: 10.1002/joc.6207.

[18] N. Prasad, P. Kumar, and M. M. Naidu, "An approach to prediction of precipitation using Gini Index in SLIQ decision tree," *Proc. - Int. Conf. Intell. Syst. Model. Simulation, ISMS*, pp. 56–60, 2013, doi: 10.1109/ISMS.2013.27.

[19] E. Kurniawan, F. Nhita, A. Aditsania, and D. Saepudin, "C5.0 algorithm and synthetic minority oversampling technique (SMOTE) for rainfall forecasting in bandung regency," *2019 7th Int. Conf. Inf. Commun. Technol. ICoICT 2019*, vol. 4, pp. 1–5, 2019, doi: 10.1109/ICoICT.2019.8835324.

[20] N. Patil, R. Lathi, and V. Chitre, "Customer Card Classification Based on C5 . 0 & CART Algorithms," *Int. J. Eng. Res. Appl.*, vol. 66, no. 3, pp. 37–39, 2012.

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

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

*ISSN: 1979-2328*

[23] I. Aprillani, A. A. Suryani, F. T. Informatika, and U. Telkom, "Analisis Penanganan Missing Value Dengan Metode Predictive Mean Matching (Pmm)," 2012.

[24] 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.

[25] F. Provost and T. Fawcett, "Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions," in *THE THIRD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING*, 1997, pp. 43–48.