ANALYSIS AND CLASSIFICATION OF RAIN DATA IN MILDURA CITY USING ORANGE DATA MINING APPLICATION

Irja' Multazamy

Prodi Teknik Informatika, Fakultas Teknik, Universitas Trunojoyo Madura

Jl. Raya Telang, Telang, Kec. Kamal, Bangkalan 69162

Email: 200411100155@student.trunojoyo.ac.id

Abstract

It is difficult for humans to predict rain without using a certain method. Rain predictions can be made by utilizing historical weather data, including parameters such as temperature, humidity, air pressure and wind patterns. This research intends to predict whether it will rain tomorrow or not using three different classification methods, namely Random Forest, Decision Tree, and Logistic Regression. This research uses the Orange Data Mining application to classify rain data in Mildura Australia with a total of 3009 data records which will be divided into two, namely 80% training data used for model training and 20% test data to test the accuracy of the model. The analysis results show a quite promising level of accuracy. Random Forest achieved an accuracy rate of 92%, Decision Tree achieved 90%, and Logistic Regression achieved 92%.

Keywords: *Rain prediction, Random Forest, Decision Tree, Logistic Regression*

ANALISIS DAN KLASIFIKASI DATA HUJAN DI KOTA MILDURA MENGGUNAKAN APLIKASI ORANGE DATA MINING

Abstrak

Sulit bagi manusia untuk melakukan prediksi hujan tanpa menggunakan suatu metode tertentu. Prediksi hujan dapat dilakukan dengan memanfaatkan data historis cuaca, termasuk parameter-parameter seperti suhu, kelembapan, tekanan udara, dan pola angin. Penelitian ini bermaksud untuk memprediksi apakah besok hujan atau tidak menggunakan tiga metode kalisifikasi yang berbeda yaitu Random Forest, Decision Tree, dan Logistic Regression. Penelitian ini menggunakan aplikasi Orange Data Mining untuk mengklasifikasikan data hujan di Mildura Australia dengan total 3009 record data yang akan dibagi menjadi dua, yaitu 80% data latih yang digunakan untuk pelatihan model dan 20% data uji untuk menguji keakuratan model. Hasil analisis menunjukkan tingkat akurasi yang cukup menjanjikan. Random Forest mencapai tingkat akurasi sebesar 92%, Decision Tree sebesar 90%, dan Logistic Regression sebesar 92%.

Kata kunci: *Prediksi hujan, Random Forest, Decision Tree, Logistic Regression*

1. Pendahuluan

Hujan adalah salah satu fenomena alam yang memiliki dampak signifikan dalam kehidupan sehari-hari manusia dalam berbagai sektor, mulai dari pertanian hingga manajemen bencana alam. Kemampuan untuk memprediksi dengan tepat memiliki nilai tak ternilai dalam perencanaan dan pengambilan keputusan.

Dalam beberapa dekade terakhir, kemajuan teknologi, terutama dalam pemantauan cuaca dan pemrosesan data, telah membuka pintu untuk pengembangan model prediksi klasifikasi hujan yang semakin canggih. Model-model ini memanfaatkan data cuaca historis dan saat ini termasuk suhu, kelembapan, tekanan udara, arah dan kecepatan angin, serta berbagai parameter lainnya. Teknik-teknik pembelajaran mesin dan kecerdasan buatan digunakan untuk menganalisis data ini dan mengidentifikasi pola-pola yang terkait dengan hujan.

Penelitian ini bertujuan untuk memprediksi hujan menggunakan dataset Rain in Australia dan mimilih lokasi di kota Mildura yang terdiri dari 23 feature dengan 1 target feature yaitu Rain Tomorrow. Penelitian ini menggunakan tiga metode kalisifikasi yang berbeda yaitu Random Forest, Decision Tree, dan Logistic Regression. Metode analisis data ini dilakukan dengan menggunakan bantuan aplikasi Orange Data Mining.

Penelitian serupa juga telah dilakukan sebelumnya oleh Irwansyah Saputra dan rekannya pada tahun 2021. Penelitian tersebut menggunakan software WEKA dan Rstudio sebagai alat bantu. Model klasifikasi yang digunakan adalah Decision Tree dan C5.0 Algorithm. Berdasarkan hasil yang diperoleh, evaluasi menggunakan 10-Cross Fold Validation lebih unggul yang memiliki akurasi paling tinggi sebesar 87.35% untuk Decision Tree dan akurasi sebesar 86.85% untuk algoritma C5.0 Rule-Based Model, dibandingkan dengan metode Split 80:20 pada kasus prediksi hujan di Australia[1].

Penelitian berikutnya berupa peramalan curah hujan menggunakan Jaringan Syaraf Tiruan oleh Prof. Mohini Darji. Model yang digunakan dalam penelitian ini diantaranya Auto Regressive Integrated Moving Average (ARIMA), Feed Forward Neural Network (FFNN), Radial Basis Function Neural Network (RBFNN), dan Time Delay Neural Network (TDNN) dengan menggunakan pendekatan Algoritma Genetika (GA) untuk mengoptimalkan bias dan bobot jaringan saraf. Dari sekian banyak model NN yang digunakan, Model GA-TDNN yang dilatih dengan algoritma Levenberg Marquardt Back Propagation mencapai akurasi prediksi curah hujan bulanan masing-masing sebesar 89,83% dan 81,38% untuk wilayah Anand dan Navsari. Dalam prediksi tahunan, model mencapai akurasi sebesar 87,87%, dan masing-masing 87,86% untuk Anand dan Navsari[2].

2. Metode

Metode yang digunakan dalam melakukan klasifikasi untuk memprediksi hujan di Australia melibatkan serangkaian tahapan yang terdiri dari pengumpulan data, proses data pre-processing yang mencakup penanganan nilai-nilai yang hilang (missing value), pembangunan model klasifikasi dengan menerapkan dan membandingkan algoritma Random Forest,

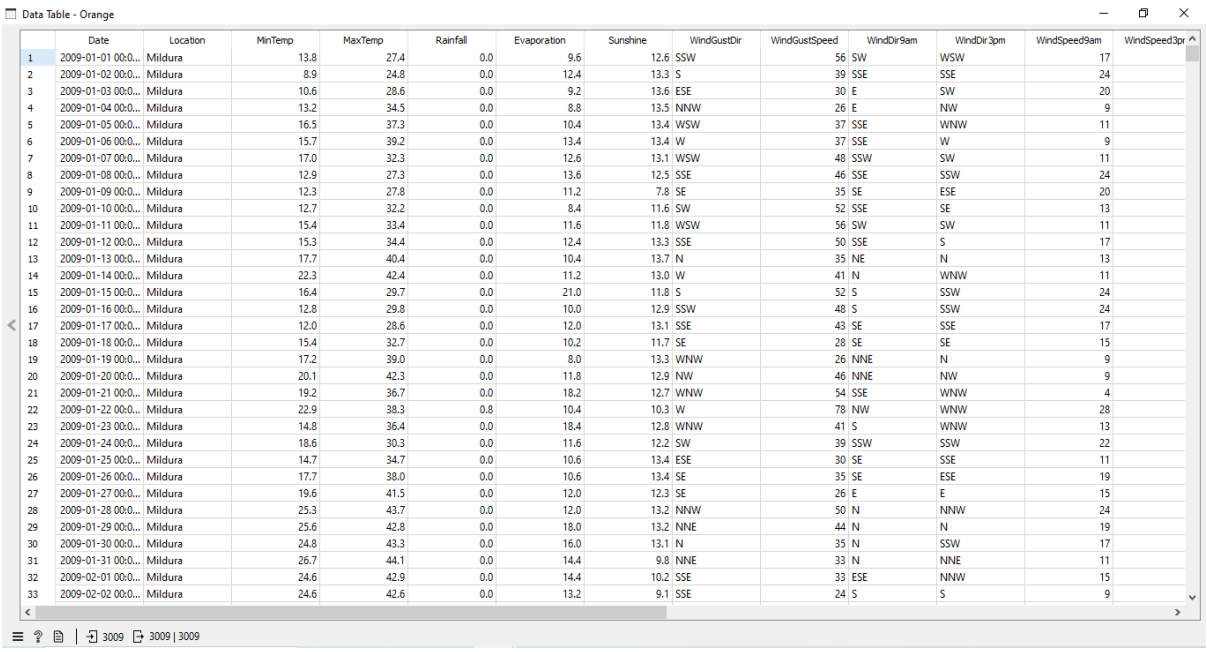
Decision Tree dan Logistic Regression, validasi hasil dengan memanfaatkan partisi dataset (80% data latih dan 20% data uji), serta penilaian kualitas model melalui Confusion Matrix.

2.1. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini adalah *Rain in Australia* yang bersumber dari repositori kaggle.com dengan link situs https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package yang berisi hasil pengamatan cuaca harian oleh stasiun cuaca di kota Mildura Australia. Dataset ini terdiri dari 3009 record data dan 23 feature dengan 1 target feature yaitu *Rain Tomorrow*.

Tabel 1. *Dataset Feature Rain in Australia (Mildura)*

No.	Fitur	Tipe Data	Deskripsi
1	Date	date	Tanggal observasi prediksi pengamatan cuaca harian
2	Location	chr	Nama 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	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, diukur dalam "oktas"
19	Cloud3pm	numeric	Bagian langit yang tertutup awan pada jam 3 sore, juga diukur dalam "oktas"
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	RainTomorrow	chr	Target Feature, Prediksi hujan besok



Data Table - Orange

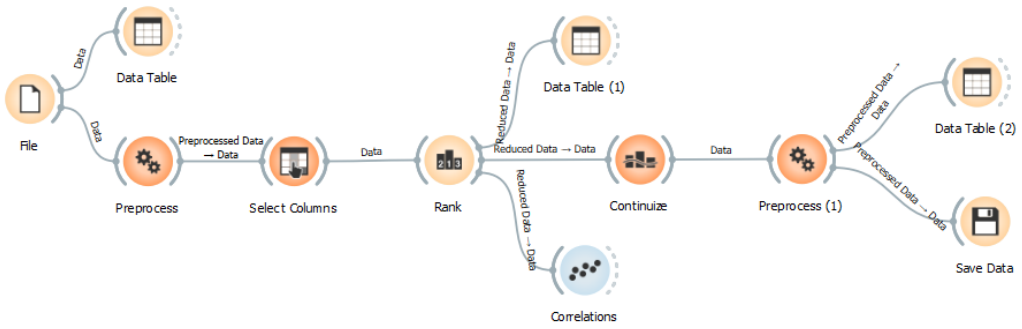
	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pr
1	2009-01-01 00:0...	Mildura	13.8	27.4	0.0	9.6	12.6	SSW	56	SW	WSW	17	
2	2009-01-02 00:0...	Mildura	8.9	24.8	0.0	12.4	13.3	S	39	SSE	SSE	24	
3	2009-01-03 00:0...	Mildura	10.6	28.6	0.0	9.2	13.6	ESE	30	E	SW	20	
4	2009-01-04 00:0...	Mildura	13.2	34.5	0.0	8.8	13.5	NNW	26	E	NW	9	
5	2009-01-05 00:0...	Mildura	16.5	37.3	0.0	10.4	13.4	WSW	37	SSE	WNW	11	
6	2009-01-06 00:0...	Mildura	15.7	39.2	0.0	13.4	13.4	W	37	SSE	W	9	
7	2009-01-07 00:0...	Mildura	17.0	32.3	0.0	12.6	13.1	WSW	48	SSW	SW	11	
8	2009-01-08 00:0...	Mildura	12.9	27.3	0.0	13.6	12.5	SSE	46	SSE	SSW	24	
9	2009-01-09 00:0...	Mildura	12.3	27.8	0.0	11.2	7.8	SE	35	SE	ESE	20	
10	2009-01-10 00:0...	Mildura	12.7	32.2	0.0	8.4	11.6	SW	52	SSE	SE	13	
11	2009-01-11 00:0...	Mildura	15.4	33.4	0.0	11.6	11.8	WSW	56	SW	SW	11	
12	2009-01-12 00:0...	Mildura	15.3	34.4	0.0	12.4	13.3	SSE	50	SSE	S	17	
13	2009-01-13 00:0...	Mildura	17.7	40.4	0.0	10.4	13.7	N	35	NE	N	13	
14	2009-01-14 00:0...	Mildura	22.3	42.4	0.0	11.2	13.0	W	41	N	WNW	11	
15	2009-01-15 00:0...	Mildura	16.4	29.7	0.0	21.0	11.8	S	52	S	SSW	24	
16	2009-01-16 00:0...	Mildura	12.8	29.8	0.0	10.0	12.9	SSW	48	S	SSW	24	
17	2009-01-17 00:0...	Mildura	12.0	28.6	0.0	12.0	13.1	SSE	43	SE	SSE	17	
18	2009-01-18 00:0...	Mildura	15.4	32.7	0.0	10.2	11.7	SE	28	SE	SE	15	
19	2009-01-19 00:0...	Mildura	17.2	39.0	0.0	8.0	13.3	WNW	26	NNE	N	9	
20	2009-01-20 00:0...	Mildura	20.1	42.3	0.0	11.8	12.9	NW	46	NNE	NW	9	
21	2009-01-21 00:0...	Mildura	19.2	36.7	0.0	18.2	12.7	WNW	54	SSE	WNW	4	
22	2009-01-22 00:0...	Mildura	22.9	38.3	0.8	10.4	10.3	W	78	NW	WNW	28	
23	2009-01-23 00:0...	Mildura	14.8	36.4	0.0	18.4	12.8	WNW	41	S	WNW	13	
24	2009-01-24 00:0...	Mildura	18.6	30.3	0.0	11.6	12.2	SW	39	SSW	SSW	22	
25	2009-01-25 00:0...	Mildura	14.7	34.7	0.0	10.6	13.4	ESE	30	SE	SSE	11	
26	2009-01-26 00:0...	Mildura	17.7	38.0	0.0	10.6	13.4	SE	35	SE	ESE	19	
27	2009-01-27 00:0...	Mildura	19.6	41.5	0.0	12.0	12.3	SE	26	E	E	15	
28	2009-01-28 00:0...	Mildura	25.3	43.7	0.0	12.0	13.2	NNW	50	N	NNW	24	
29	2009-01-29 00:0...	Mildura	25.6	42.8	0.0	18.0	13.2	NNE	44	N	N	19	
30	2009-01-30 00:0...	Mildura	24.8	43.3	0.0	16.0	13.1	N	35	N	SSW	17	
31	2009-01-31 00:0...	Mildura	26.7	44.1	0.0	14.4	9.8	NNE	33	N	NNE	11	
32	2009-02-01 00:0...	Mildura	24.6	42.9	0.0	14.4	10.2	SSE	33	ESE	NNW	15	
33	2009-02-02 00:0...	Mildura	24.6	42.6	0.0	13.2	9.1	SSE	24	S	S	9	

3009 | 3009 | 3009

Gambar 1. *Dataset Rain in Australia (Mildura)*

2.2. Data Pre-Processing

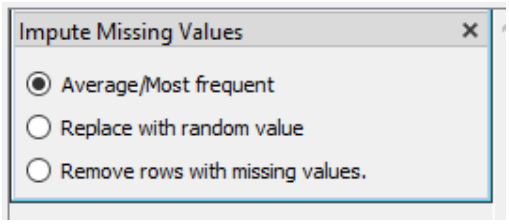
Proses pre-processing (pra-pemrosesan) merupakan tahap kritis dalam analisis data yang dilakukan sebelum data dianalisis atau dimodelkan. Dalam penelitian ini, tahapan pre-processing data dapat meliputi penanganan missing value, seleksi fitur (memilih fitur yang relevan digunakan untuk klasifikasi), label encoding (merubah tipe data string/kategori ke numerik), dan normalisasi.



Gambar 2. *Alur pre-processing*

2.2.1. Menangani Missing Value

Missing value (nilai yang hilang) harus diatasi karena keberadaan missing value dalam data dapat memiliki dampak serius pada analisis statistik, pemodelan, dan pengambilan keputusan. Salah satu cara mengatasi missing value adalah dengan melakukan imputasi nilai rata-rata.



Gambar 3. *Menangani missing value*

2.2.2. Seleksi Fitur

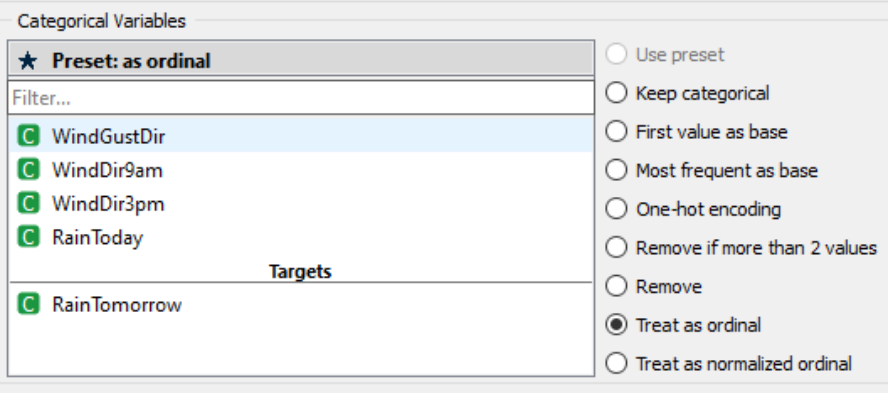
Dalam data mining, diperlukan adanya identifikasi dan pemilihan subset fitur yang paling relevan dan informatif dari dataset asli. Tujuannya adalah untuk meningkatkan kualitas analisis data, model prediktif, atau algoritma data mining dengan mengurangi dimensi data dan fokus pada fitur-fitur yang benar-benar berkontribusi terhadap hasil analisis. Dalam penelitian ini akan dilakukan perangkingan menggunakan metode Information Gain dan Information Gain Ratio, dengan memilih 14 fitur yang memiliki pengaruh paling signifikan terhadap variabel target dan date sebagai meta attribute.

		#	Info. gain	Gain ratio
1	N Sunshine		0.094	0.047
2	N Cloud3pm		0.082	0.042
3	N Humidity3pm		0.054	0.027
4	N Cloud9am		0.050	0.025
5	N Pressure3pm		0.048	0.024
6	N Pressure9am		0.044	0.022
7	C WindDir3pm	16	0.032	0.008
8	N Rainfall		0.031	0.031
9	C WindDir9am	16	0.030	0.008
10	N WindGustSpeed		0.023	0.011
11	C RainToday	2	0.022	0.045
12	C WindGustDir	16	0.019	0.005
13	N Humidity9am		0.012	0.006
14	N Temp3pm		0.011	0.006
15	N WindSpeed3pm		0.009	0.005
16	N MaxTemp		0.008	0.004
17	N MinTemp		0.006	0.003
18	T Date		0.003	0.001

Gambar 4. *Seleksi fitur*

2.2.3. Label Encoding

Label encoding adalah salah satu teknik dalam pre-processing data yang digunakan dalam konteks machine learning dan analisis data. Tujuan dari label encoding adalah mengubah variabel kategori menjadi bentuk numerik agar dapat digunakan oleh algoritma pembelajaran mesin yang membutuhkan data numerik.

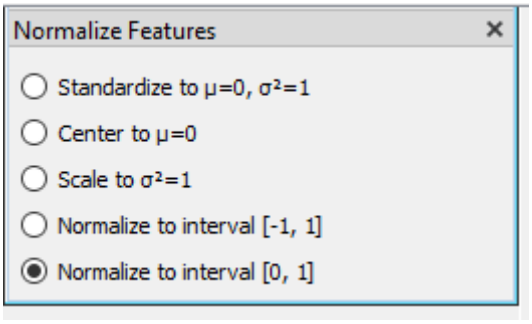


Gambar 5. *Label encoding*

Label encoding yang diperlakukan sebagai ordinal berarti bahwa label atau kategori yang diberikan angka atau nilai numerik diperlakukan sebagai data ordinal. Dalam konteks data ordinal, nilai-nilai tersebut memiliki urutan atau peringkat tertentu, tetapi selisih antar nilai tidak memiliki makna yang signifikan.

2.2.4. Normalisasi

Normalisasi merupakan proses transformasi data dalam analisis statistik dan pemodelan untuk membawa nilai-nilai dari berbagai variabel ke skala yang seragam atau relatif setara. Tujuan normalisasi adalah untuk menghilangkan perbedaan skala antara variabel, sehingga variabel dengan rentang nilai yang berbeda dapat diperlakukan dengan adil dalam analisis data atau pemodelan.

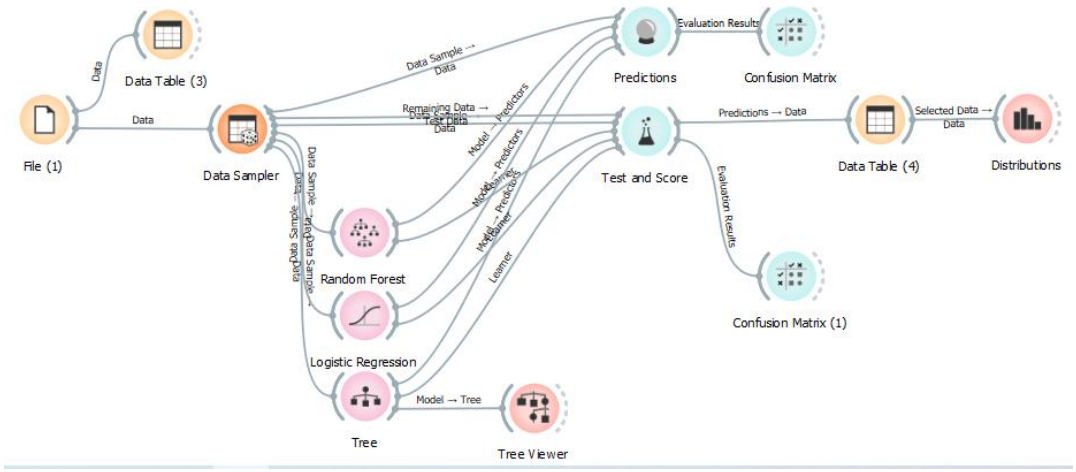


Gambar 6. *Normalisasi data*

3. Hasil dan Pembahasan

3.1. Modelling Dataset

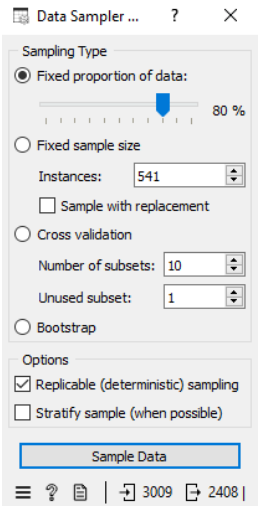
Dataset yang telah melalui proses pre-processing selanjutnya akan dilakukan modelling menggunakan 3 model klasifikasi yaitu Random Forest, Decision Tree, dan Linear Regression. Dataset akan dibagi ke dalam dua bagian, yaitu 80% untuk data latih dan 20% untuk data uji.



Gambar 7. *Modelling dataset*

3.1.1. Partisi Dataset 80:20

Dataset akan dilakukan splitting data dengan perbandingan 80% untuk data latih dan 20% untuk data uji.



Gambar 8. *Splitting data*

3.1.2. Test and Score

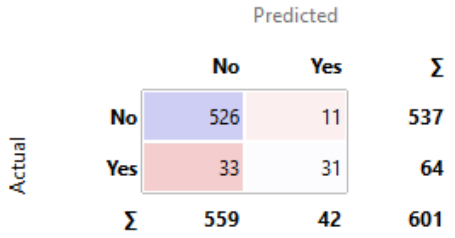
Selanjutnya akan dilakukan test dan perbandingan skor akurasi pada 3 model klasifikasi yaitu Random Forest, Decision Tree, dan Linear Regression.

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	0.888	0.927	0.920	0.919	0.927	0.561
Tree	0.709	0.908	0.905	0.903	0.908	0.489
Logistic Regression	0.921	0.922	0.916	0.914	0.922	0.537

Gambar 9. *Perbandingan hasil akurasi model*

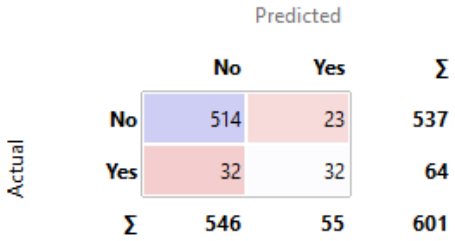
3.2. Confusion Matrix

Confusion matrix adalah tabel yang digunakan dalam evaluasi kinerja model klasifikasi dalam machine learning. Ini memberikan gambaran tentang seberapa baik model untuk memprediksi kelas target dari data yang diberikan. Berikut Confusion Marix dari setiap pemodelan yang sudah dilakukan, yaitu dari algoritma Random Forest, Decision Tree, dan Linear Regression dapat dilihat pada **Gambar 10-12.**



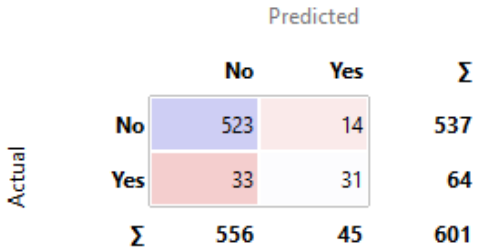
Gambar 10. *Confusion matrix algoritma Random Forest*

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} = \frac{31 + 526}{31 + 526 + 33 + 11} = 0.92$$



Gambar 11. *Confusion matrix algoritma Decision Tree*

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} = \frac{32 + 514}{32 + 514 + 32 + 23} = 0.90$$



Gambar 12. *Confusion matrix algoritma Logistic Regression*

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} = \frac{31 + 523}{31 + 523 + 33 + 14} = 0.92$$

4. Kesimpulan

Penelitian ini berhasil mengikuti serangkaian tahapan yang diperlukan untuk prediksi hujan di Mildura Australia, mulai dari pengumpulan data yang bersumber dari repositori

kaggle.com berupa dataset *Rain in Australia,* tahap pre-processing data meliputi penanganan missing value, seleksi fitur, label encoding, dan normalisasi menggunakan alat bantu Orange Data Mining.

Dalam penelitian ini, dataset dibagi ke dalam dua bagian, yaitu 80% untuk data latih dan 20% untuk data uji. Model Random Forest mencapai akurasi 92%, model Decision Tree mencapai akurasi 90%, dan model Logistic Regression mencapai akurasi 92%.

Selanjutnya, penelitian ini diharapkan dapat memberi wawasan penting untuk pengembangan alat bantu prediksi hujan yang lebih akurat dan efektif.

Daftar Pustaka

[1] Irwansyah Saputra and Dinar Ajeng Kristiyanti, "Application of Data Mining for Rainfall Prediction Classification in Australia with Decision Tree Algorithm and C5.0 Algorithm", Seminar Nasional Informatika (SEMNASIF) 2021, pp. 71-87, 2021.

[2] Prof. Mahini Darji, "Rainfall Forecasting Using Neural Networks", Jurnal Internasional Penelitian dan Tinjauan Analitik, 2019.