

PREDIKSI KEMUNGKINAN HUJAN DENGAN MENGUNAKAN ALGORITMA MACHINE LEARNING

(2308207010001) Hafizha Dini Giandra

(2308207010004) Fadhilah Syafa

(2308207010007) Mutaqqin



TABLE OF CONTENTS

- 01 Latar belakang
- 02 Referensi
- 03 Metode yang digunakan
- 04 Alur eksperimen
- 05 Hasil
- 06 Kesimpulan






LATAR BELAKANG



Kemungkinan terjadinya hujan sangat mempengaruhi alur kerja di berbagai bidang seperti perencanaan produksi pangan, pengelolaan sumber daya air, prediksi longsor dan/atau banjir, bahkan hal-hal trivial seperti perencanaan kegiatan sehari-hari terutama kegiatan outdoor.

Prediksi hujan yang akurat dan diharapkan dapat memberikan tahap intervensi baru kepada sektor-sektor yang terkena dampak dari hujan, namun saat ini prediksi hujan secara akurat masih menjadi tantangan besar bagi para ahli iklim.



REFERENSI

🔍 Anwar et al. (2020)

Membandingkan metode J48, Random Forest, Naïve Bayes, dan Multilayer Perceptron (MLP). Data diperoleh dari BMKG untuk wilayah Tanjung Mas, Semarang, sejak tahun 2013 hingga 2019.

Hasil klasifikasi algoritma MLP dan J48 menghasilkan akurasi terbaik (hingga 78,4%) dibandingkan dengan algoritma lain, walaupun perbedaannya kecil.

REFERENSI

🔍 Sanie et al. (2020)

Menggunakan ensemble learning dengan menggabungkan beberapa pengklasifikasi machine learning yaitu NBC, DT, SVM, RF, dan NN. Data diperoleh dari Drainage and Irrigation Department dan Malaysian Meteorological Department.

Hasil menunjukkan bahwa metode ensemble (gabungan C4.5, SVM, dan ANN) dengan majority voting menghasilkan nilai precision, recall, dan F-measures yang paling baik, yaitu masing-masing sebesar 76%, 77%, dan 76%.

REFERENSI

🔍 Hudnurkar dan Rayavarapu (2022)

Melakukan prediksi kemungkinan hujan menggunakan SVM dan ANN. Data diperoleh dari National Data Center of Indian Meteorological Department dari tahun 2000 hingga 2018. Tiga dataset dari sumber ini yaitu dataset stasiun Shivajinagar, stasiun Nashik, dan stasiun Chikalthana.

Hasil akurasi klasifikasi yang diperoleh adalah 82,1% dengan SVM dan 82,8 dengan ANN untuk dataset stasiun Shivajinagar. Hasil klasifikasi dataset stasiun Nashik adalah 74,4% oleh SVM dan 71,6% oleh ANN. Hasil klasifikasi dataset stasiun Chikalthana adalah 77,9% oleh SVM dan 75,5% oleh ANN.

METODE YANG DIGUNAKAN

⚙ Decision Tree (DT)

Decision Tree (DT) merupakan salah satu metode machine learning sederhana untuk klasifikasi dan regresi. Secara kasar proses pengklasifikasian dengan decision tree terlihat seperti kumpulan syntax if-then. Keuntungan metode ini adalah alur algoritmanya mudah dibaca dan dimengerti, serta kecepatan klasifikasi yang cepat.

METODE YANG DIGUNAKAN

⚙ Random Forest (RF)

Random forest (RF) adalah metode yang dihasilkan dari pengembangan metode decision tree. Random forest merupakan kumpulan decision tree yang dibangun dengan nilai parameter, seleksi fitur, dan jumlah sampel yang acak dan independen antar pohon. Pembangunan pohon dengan cara ini mengecilkan kemungkinan terjadinya overfitting terhadap model yang dihasilkan, dimana overfitting merupakan salah satu kekurangan dari metode decision tree.

METODE YANG DIGUNAKAN

⚙️ Logistic Regression

Logistic regression adalah metode statistik yang biasanya digunakan untuk klasifikasi biner, yaitu klasifikasi yang variable datanya bersifat kategorikal atau diskrit dan hanya memiliki dua label kelas. Kelebihan dari logistic regression adalah metode implementasinya yang sederhana dibandingkan dengan metode lain dan waktu pelatihannya lebih sedikit.

METODE YANG DIGUNAKAN

⚙ Long Short-Term Memory (LSTM)

LSTM merupakan pengembangan dari algoritma Recurrent Neural Network (RNN) yang dirancang untuk mengatasi keterbatasan RNN tradisional dalam memahami dan mengingat ketergantungan jangka panjang pada data berurutan. LSTM mampu menyimpan informasi untuk jangka waktu yang lama, sehingga dapat digunakan untuk memproses, memprediksi, dan mengklasifikasikan informasi berdasarkan data konteks deret waktu.

ALUR EKSPERIMEN

- Pengambilan dataset**
- Pra-pengolahan dataset**
- Klasifikasi dengan DT**
- Klasifikasi dengan RF**
- Klasifikasi dengan logistic regression**
- Klasifikasi dengan LSTM**
- Evaluasi hasil klasifikasi**

ALUR EKSPERIMEN

→ Pengambilan dataset

Dataset yang digunakan berasal dari situs Kaggle, yaitu Rain in Australia dataset yang memiliki 145.640 baris data. Dataset ini berisi hasil observasi cuaca harian dari berbagai lokasi di Australia selama 10 tahun (2007-2017). Terdapat 23 fitur pada dataset ini tujuh diantaranya, termasuk label kelas, adalah data nominal dan sisanya adalah data numerik.



ALUR EKSPERIMEN

→ Pra-pengolahan dataset

Tahapan pra-pengolahan:

- Penghilangan atribut-atribut yang tidak relevan;
- Pengambilan baris-baris data yang tidak memiliki nilai null atau NaN di semua kolomnya;
- Penyamaan jumlah sampel kedua label kelas;
- Normalisasi dataset; dan
- Pemisahan training dan testing dataset dengan rasio 70:30.

ALUR EKSPERIMEN

→ Klasifikasi dengan DT

Klasifikasi dengan metode decision tree dilakukan dengan menggunakan fungsi `DecisionTreeClassifier` dalam library Sklearn.



ALUR EKSPERIMEN

→ Klasifikasi dengan RF

Klasifikasi dengan metode decision tree dilakukan dengan menggunakan fungsi RandomForestClassifier dalam library Sklearn.



ALUR EKSPERIMEN

→ Klasifikasi dengan logistic regression

Dalam proyek ini klasifikasi dengan metode logistic regression dilakukan dengan dua cara, yaitu dengan menggunakan fungsi LogisticRegression dalam library Sklearn dan menggunakan code yang dibuat tanpa bantuan library Sklearn yang didapat dari pembelajaran mata kuliah PDSAI.



ALUR EKSPERIMEN

→ Klasifikasi dengan LSTM

Klasifikasi dengan metode LSTM dilakukan dengan menggunakan ensemble fungsi LSTM dan fungsi Dense dalam library TensorFlow. Fungsi Dense digunakan untuk memutuskan kelas sampel dengan menggunakan fungsi sigmoid.



ALUR EKSPERIMEN

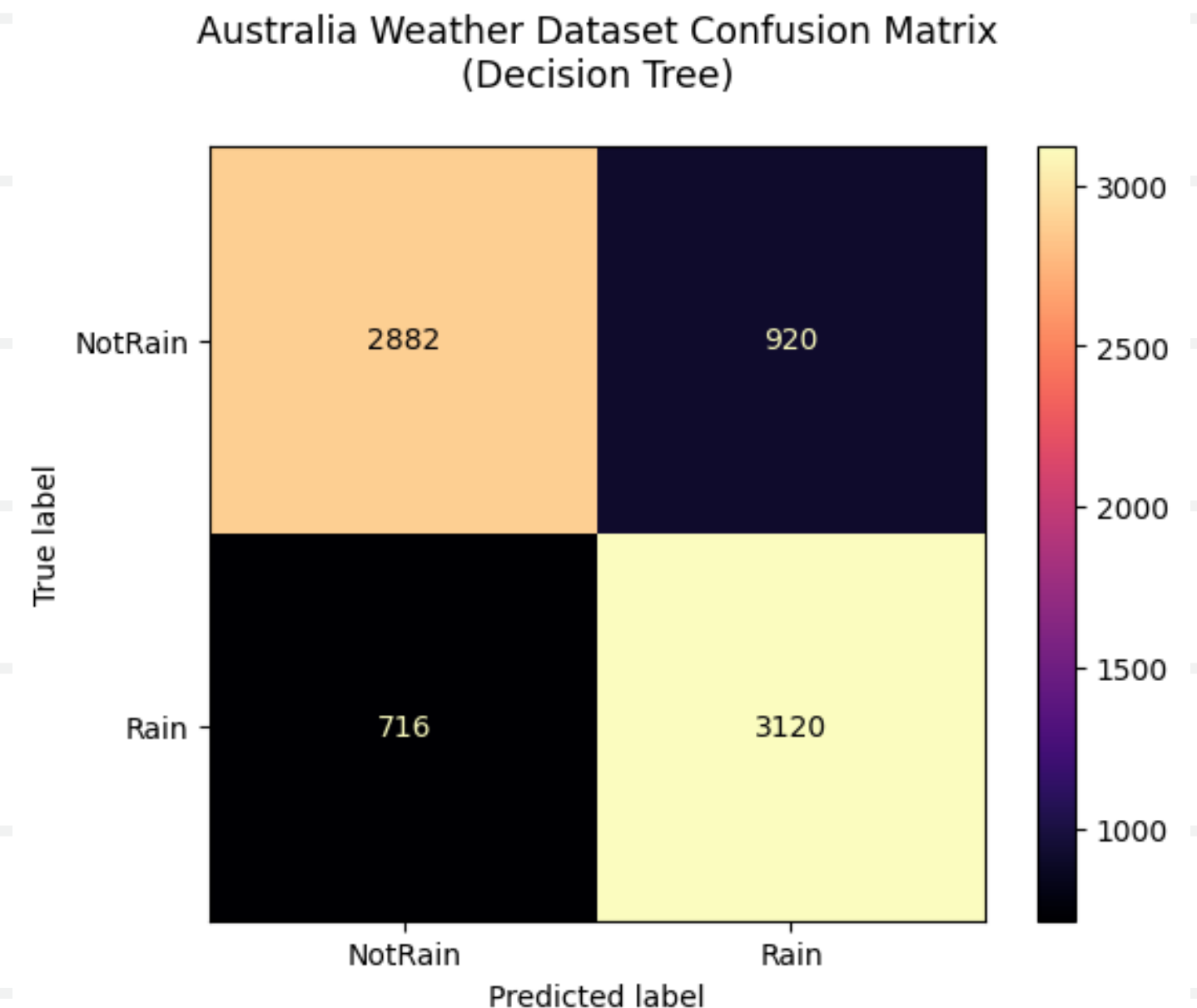
→ Evaluasi hasil klasifikasi

Evaluasi hasil klasifikasi dilakukan dengan menggunakan beberapa metrik pengukuran yaitu accuracy, precision, recall, dan F1-score. Semua perhitungan ini akan dilakukan dengan bantuan library Sklearn.



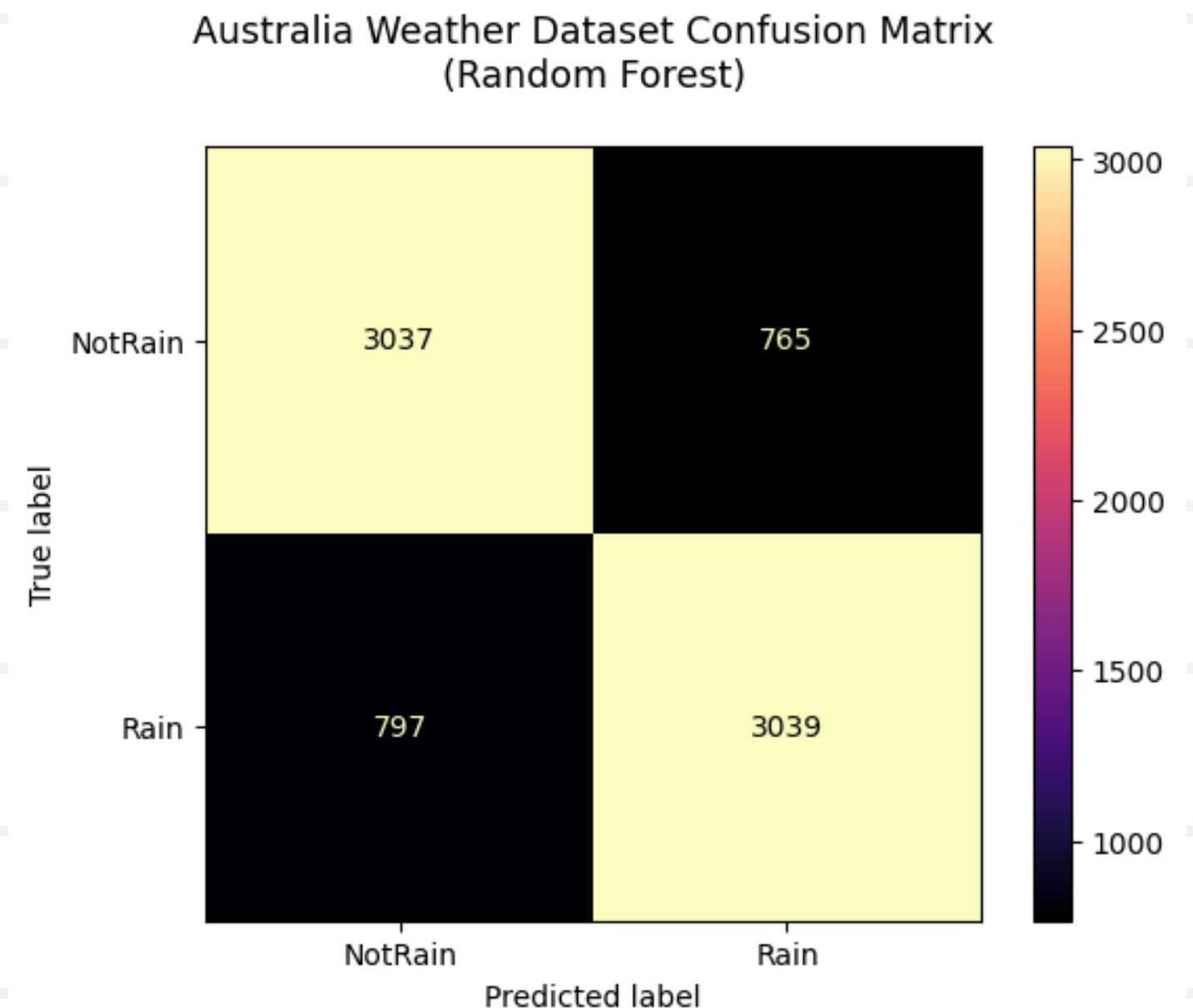
HASIL

Parameter-parameter yang disesuaikan pada klasifikasi dengan metode DT adalah criterion, max_depth, min_samples_leaf, dan max_features. Hasil klasifikasi terbaik didapat dengan nilai parameter criterion = gini, max_depth = 6, min_samples_leaf = 150, dan max_features = 13, dengan nilai akurasi sebesar 78,581%.



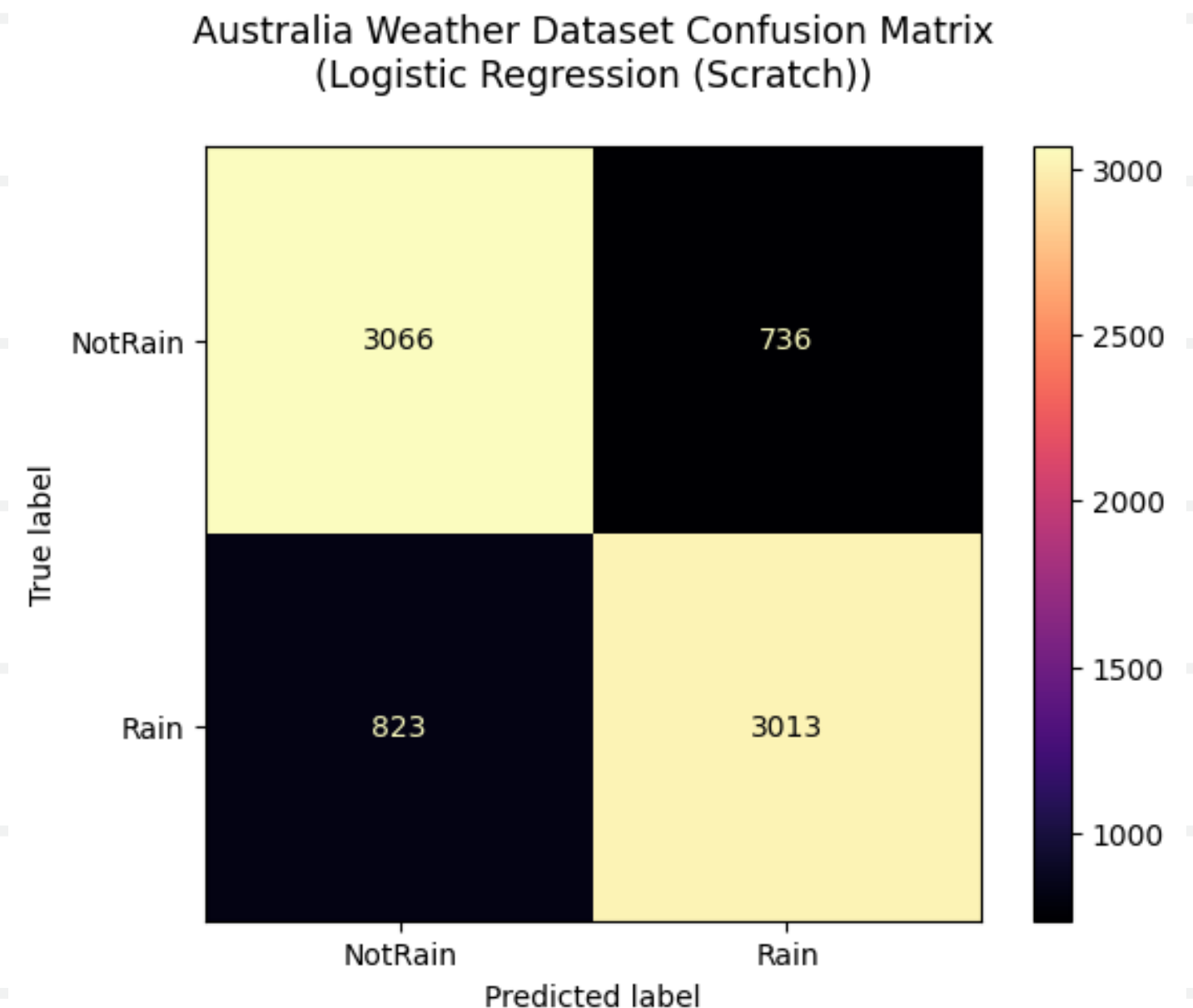
HASIL

Parameter-parameter yang disesuaikan pada klasifikasi dengan metode RF adalah `n_estimators`, `criterion`, `max_depth`, `min_samples_leaf`, dan `max_features`. Hasil klasifikasi terbaik didapat dengan nilai parameter `n_estimators = 250`, `criterion = entropy`, `max_depth = 7`, `min_samples_leaf = 100`, dan `max_features = 11`, dengan nilai akurasi sebesar 79,55%.



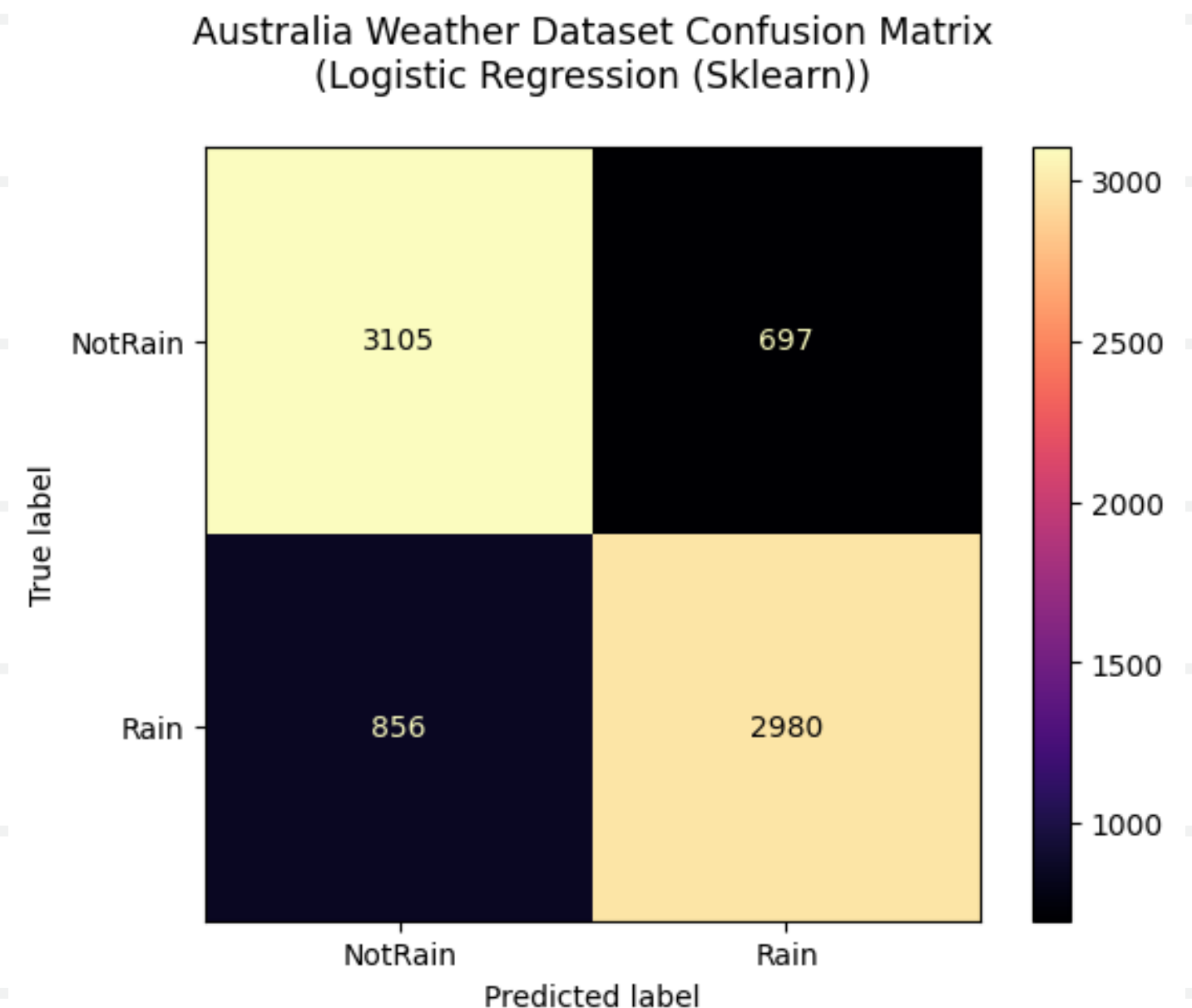
HASIL

Parameter-parameter yang disesuaikan pada klasifikasi dengan metode logistic regression menggunakan code from scratch adalah alpha, epochs, dan batch_size. Hasil klasifikasi terbaik didapat dengan nilai parameter alpha = 0,0005, epochs = 10000, dan batch_size = 10%, dengan nilai akurasi 79,589%.



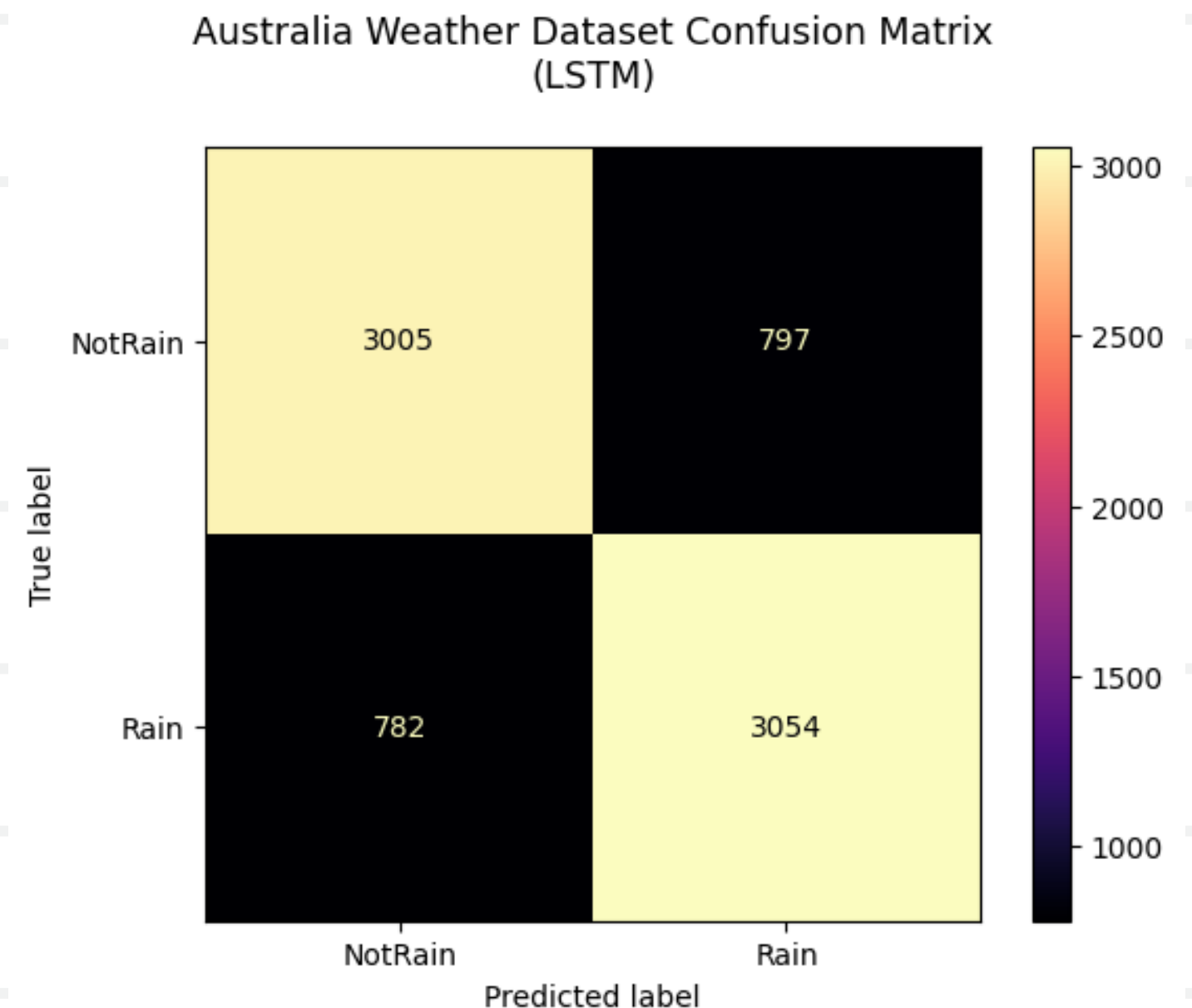
HASIL

Parameter-parameter yang disesuaikan pada klasifikasi dengan metode logistic regression menggunakan fungsi LogisticRegression adalah max_iter, solver, dan tol. Hasil klasifikasi terbaik didapat dengan nilai parameter max_iter = 250, solver = sag, dan tol = 0,005, dengan nilai akurasi 79,667%.



HASIL

Parameter-parameter yang disesuaikan pada klasifikasi dengan metode ensemble LSTM adalah units, activation, optimizer, epochs, dan batch_size. Hasil klasifikasi terbaik didapat dengan nilai parameter units = 64, activation = tanh, optimizer = adam, epochs = 20, dan batch size = 16, dengan nilai akurasi sebesar 79,327%.



HASIL



Metode	Recall	Precision	F1-score	Accuracy
Decision Tree (DT)	0,81335	0,77228	0,79228	78,581%
Random Forest (RF)	0,79223	0,7989	0,79555	79,55%
Logistic Regression (Sklearn)	0,77685	0,81044	0,79329	79,667%
Logistic Regression (Scratch)	0,78545	0,80368	0,79446	79,589%
Long Short-Term Memory (LSTM)	0,79614	0,79304	0,79459	79,327%

KESIMPULAN

- Eksperimen dilakukan dengan menggunakan empat metode klasifikasi yaitu decision tree, random forest, logistic regression, dan LSTM.
- Nilai akurasi tertinggi diperoleh dari metode logistic regression, yaitu sebesar 79,667%.
- Hasil eksperimen tidak lebih baik dari eksperimen sebelumnya, yaitu eksperimen oleh Zhao et al. (2022)



~ TERIMA KASIH ~

