Prediksi Kemungkinan Hujan dengan menggunakan Decision Tree, Logistic Regression, dan LSTM

(2308207010001) Hafizha Dini Giandra
(2308207010004) Fadhilah Syafa
(2308207010007) Mutaqqin

Table of contents

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

Zhao et al. (2022)

Melakukan prediksi hujan menggunakan DT, Logistic Regression, Long Short Term Memory (LSTM), AdaBoost, Bagging Algorithm, dan kNN. Data yang digunakan adalah dataset Rain in Australia dari situs Kaggle yang juga digunakan dalam projek ini.

Hasil akurasi tertinggi diperoleh dari metode Logistic Regression dan LSTM yaitu sebesar 85%.

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)

RF merupakan metode yang dikembangkan dari DT. RF terdiri dari banyak DT yang dibangun dengan nilai parameter, seleksi fitur dan jumlah sampel yang acak dan independen. Pembangunan pohon dengan cara ini mengecilkan kemungkinan overfitting terhadap dataset.

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 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 deret waktu.

Alur eksperimen

01 Pengambilan dataset

02 Preprocessing dataset

03 Klasifikasi dengan DT dan RF

04 Klasifikasi dengan Logistic Regression

05 Klasifikasi dengan LSTM

06 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

Preprocessing dataset

Tahapan preprocessing:

- 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
- Pemisahan training dan testing dataset dengan rasio 70:30

Alur eksperimen

Klasifikasi dengan DT dan RF

Klasifikasi dengan metode DT dilakukan dengan menggunakan fungsi DecisionTreeClassifier, sedangkan metode RF dilakukan dengan menggunakan fungsi RandomForestClassifier. Kedua fungsi tersebut berada di dalam *library* Sklearn.

Alur eksperimen

Klasifikasi dengan Logistic Regression

Dalam projek 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

Dalam projek ini klasifikasi LSTM dilakukan dengan menggunakan fungsi LSTM dalam *library* TensorFlow.

Alur eksperimen

Evaluasi hasil klasifikasi

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

~ Terima Kasih ~