

LAPORAN KECERDASAN KOMPUTASIONAL – [IF184503]



**ANALISIS PREDIKSI TERJADINYA HUJAN PADA
DATASET RAIN IN AUSTRALIA**

Gema Adi Perwira - 05111840000138
Vieri Fath Ayuba - 05111840000153
M. Samsu Dhuha - 05111840000155

ABSTRAK

Pada kehidupan sehari-hari, seringkali menemui prediksi terjadinya hujan di berbagai media massa. Kebutuhan akan keadaan cuaca esok hari sangat dibutuhkan untuk menyusun berbagai rencana. Untuk masa lampau, perkiraan terjadinya hujan sangat bergantung dengan bulannya, ada musim kemarau dan musim penghujan. Namun saat ini, curah hujan semakin sulit untuk diprediksi sehingga diperlukan model atau sistem yang dapat memprediksi curah hujan dengan akurat.

Pada penelitian tugas akhir ini dijelaskan tentang prediksi terjadinya hujan menggunakan algoritma Decision Tree, SVM, NN-Sklearn, dan juga EDA sebagai penunjang dalam pengerjaan penelitian ini.

Algoritma Decision Tree, SVM, Neural Network-Sklearn digunakan untuk mencari keakuratan terjadinya hujan pada esok hari. yang digunakan seperti heatmap, matplotlib, digunakan untuk mencari data yang akan dibuat untuk hipotesis kita.

Kata Kunci: prediksi, hujan, akurasi.

DAFTAR ISI

ABSTRAK	ii
DAFTAR ISI	iii
Bab 1.	1
PENDAHULUAN	1
1.1.	1
1.2.	1
Bab 2.	2
DESAIN DAN IMPLEMENTASI	2
2.1.	2
2.2.	33
Bab 3.	4
HASIL UJI COBA DAN DISKUSI	4
3.1.	44
3.2.	1110
DAFTAR PUSTAKA	11

Bab 1.

PENDAHULUAN

1.1. Latar Belakang

Di seluruh dunia, pastinya akan mengalami hujan. Musim hujan merupakan musim yang mutlak ada di sebagian belahan benua dunia. Curah hujan pasti memiliki intensitas yang berbeda. Faktor penyebabnya dapat terjadi karena suhu udara pada suatu wilayah atau juga karena kadar volume air yang dibawa oleh awan. Perkiraan curah hujan yang tepat dan akurat amatlah dibutuhkan oleh semua pihak.

Namun, kebanyakan orang sulit untuk memprediksi mendapatkan data perkiraan yang tepat dan akurat untuk mengukur apakah besok akan terjadi hujan atau tidak di suatu daerah. Hal itu menyebabkan aktivitas sehari-hari akan terganggu karena sulit memprediksi hujan. Oleh karena itu, dibutuhkan analisis tentang prediksi terjadinya hujan esok hari di Australia. Penelitian ini menggunakan dataset Rain in Australia untuk dianalisis. Diperlukan juga penghitungan akurasi prediksi terjadinya hujan esok hari dan analisa penyebab terbesar terjadinya hujan esok hari.

1.2. Perumusan Masalah

Berdasarkan penjelasan dari hasil latar belakang yang telah dijelaskan sebelumnya, maka diperlukan perumusan permasalahan sebagai berikut :

1. Apa fitur yang paling berpengaruh terhadap terjadinya hujan ?
2. Daerah apa yang memiliki prediksi RainTomorrow yang tertinggi?
3. Arah tiupan angin yang paling berpengaruh terhadap prediksi RainTomorrow?
4. Algoritma apa yang memiliki akurasi terbaik dalam memprediksi hujan?

Bab 2.

DESAIN DAN IMPLEMENTASI

2.1. Persiapan Data

Kami menggunakan dataset Rain in Australia. Dataset ini berisi tentang pengamatan cuaca harian pada daerah - daerah di seluruh Australia yang berisi 145460 data. Dataset ini memiliki variabel target yaitu RainTomorrow yaitu prediksi hujan pada esok hari. Berikut adalah fitur-fitur yang ada pada dataset ini beserta keterangannya :

1. Date : tanggal
2. Location : Tempat
3. MinTemp : Suhu minimal (derajat C)
4. MaxTemp : Suhu maximal (derajat C)
5. Rainfall : Curah hujan
6. Evaporation : Penguapan
7. Sunshine : total waktu panas matahari dalam sehari (jam)
8. WindGustDir : arah tiupan angin terkuat dalam 24 jam hingga tengah malam
9. WindGustSpeed : kecepatan (km/h) angin terkuat
10. WindDir9am : arah angin jam 9 pagi
11. WindDir3pm : arah angin jam 3 sore
12. WindSpeed9am : Kecepatan angin (km/h) rata2 lebih dari 10 menit di jam 9 pagi
13. WindSpeed3pm : Kecepatan angin (km/h) rata2 lebih dari 10 menit di jam 3 sore
14. Humidity9am : Kelembapan (persen) di jam 9 pagi
15. Humidity3pm : Kelembapan (persen) di jam 3 sore
16. Pressure9am : Tekanan atmosfer jam 9 pagi
17. Pressure3pm : Tekanan atmosfer jam 3 sore
18. Cloud9am : Bagian langit yg tertutup awan (0 = cerah, 8 = mendung) jam 9 pagi

19. Cloud3pm : Bagian langit yg tertutup awan (0 = cerah, 8 = mendung) jam 3 sore
20. Temp9am : Suhu jam 9 pagi
21. Temp3pm : Suhu jam 3 sore
22. RainToday : hujan hari ini (nilai 1 untuk iya, nilai 0 untuk tidak)
23. RainTomorrow : Prediksi hujan besok (nilai 1 untuk iya, nilai 0 untuk tidak)

2.2. Skenario Uji Coba

Untuk melakukan uji coba hipotesis terhadap data rain untuk menjawab tiap rumusan masalah yang perlu dilakukan dengan beberapa pertimbangan, antara lain:

- Melakukan import library yang diperlukan dalam pengerjaan analisis pada awal pengerjaan.
- Import dataset yang ingin diuji dan menangani nilai data yang kosong atau hilang menggunakan imputer.
- Data Preprocessing. Proses pemilihan data yang akan dilakukan proses pelatihan dan pengujian. Data akan disortir dan dipilih berdasarkan batasan masalah dan parameter yang penting dalam proses pelatihan nantinya.
- Melakukan uji coba dengan beberapa algoritma agar menjawab rumusan masalah yang telah dibuat.

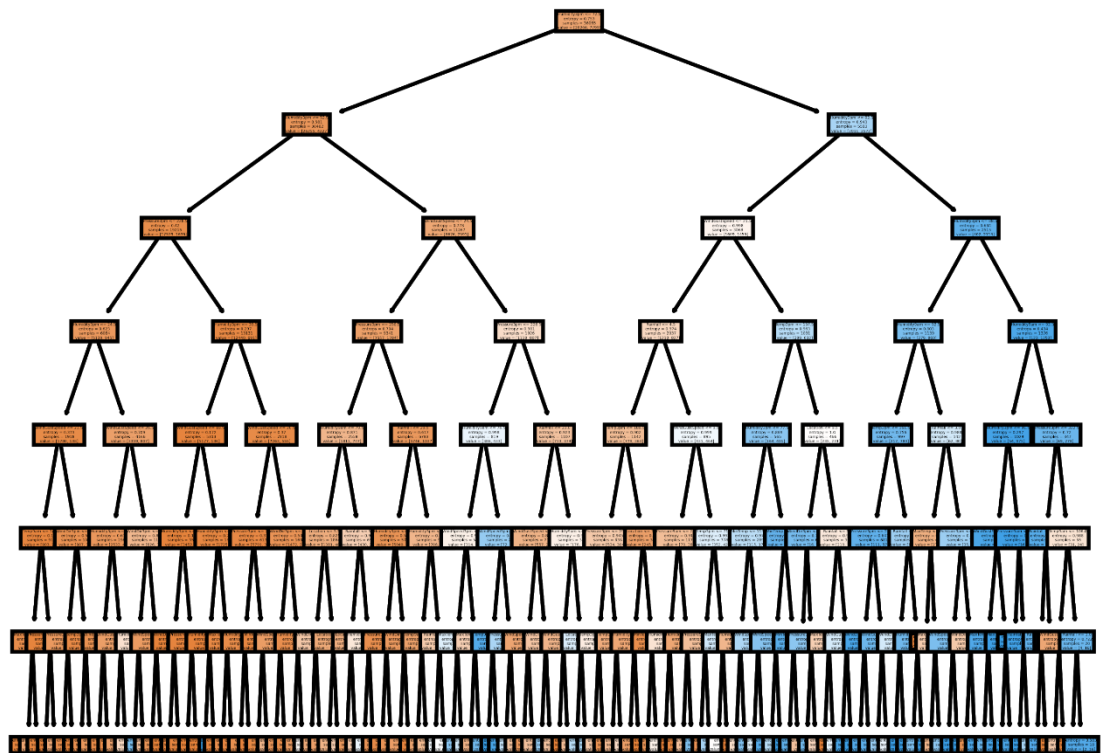
Bab 3.

HASIL UJI COBA DAN DISKUSI

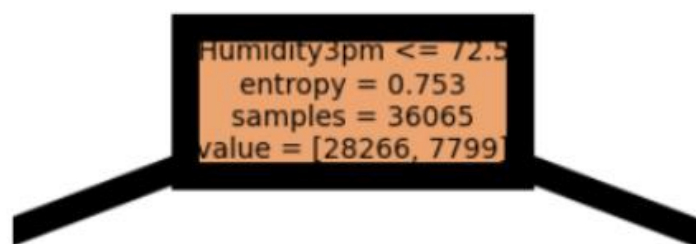
3.1. Hasil Uji Coba Skenario

3.1.1. Apa fitur yang paling berpengaruh terhadap terjadinya hujan ?

Eksplorasi pada hipotesis yang ke 1 bertujuan untuk menganalisis fitur yang paling berpengaruh terhadap terjadinya hujan. Eksplorasi ini dilakukan dengan menggunakan algoritma Decision Tree dan K - Nearest Neighbour. Dengan menggunakan eksplorasi Decision Tree didapatkan struktur pohon, dimana setiap node merepresentasikan nilai dari atribut. Untuk node teratas dari decision tree fitur yang paling berpengaruh adalah Humidity3pm.



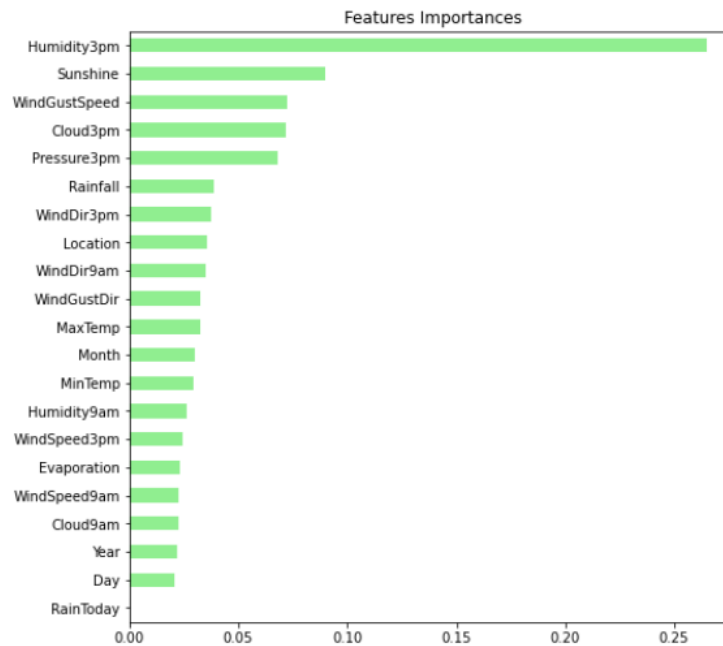
Gambar 1. Hasil Decision Tree



Gambar 2. Root dari hasil Decision Tree

Analisis yang kedua menggunakan eksplorasi dengan K- Nearest Neighbour didapatkan visualisasi hasil fitur yang paling berpengaruh

adalah Humidity3pm.



Gambar 3. Hasil KNN

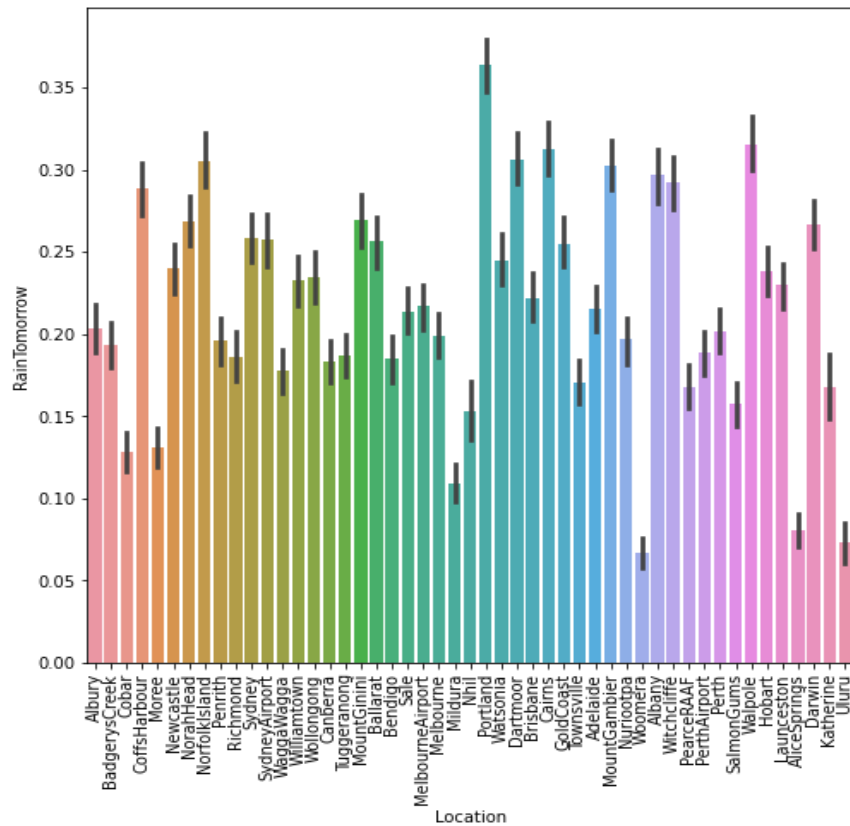
3.1.2. Daerah apa yang memiliki prediksi rainTomorrow yang tertinggi?

Eksplorasi pada hipotesis yang ke-2 didasari karena atribut location memiliki banyak atribut yang berbeda pada setiap lokasi. Atribut – atribut tersebut dapat mempengaruhi terjadinya hujan seperti humidity, wind direction, windgust speed, temperature dan lain – lain. Eksplorasi pada hipotesis yang ke-2 ini juga dilakukan dengan dasar eksplorasi menggunakan Decision Tree, dimana fitur location memiliki nilai entropy yang besar.

Dengan melakukan eksplorasi data dengan menggunakan fitur matplotlib.pyplot didapatkan visualisasi fitur Location dengan data target RainTomorrow. Dari hasil eksplorasi tersebut, didapatkan hasil lokasi tertinggi yaitu daerah Portland.

Berikut ini merupakan visualisasi dari eksplorasi menggunakan matplotlib dengan sumbu X adalah fitur location dan

sumbu Y adalah fitur RainTomorrow :



Gambar 4. Visualisasi fitur Location dengan RainTomorrow

3.1.3. Arah tiupan angin yang paling berpengaruh terhadap prediksi RainTomorrow?

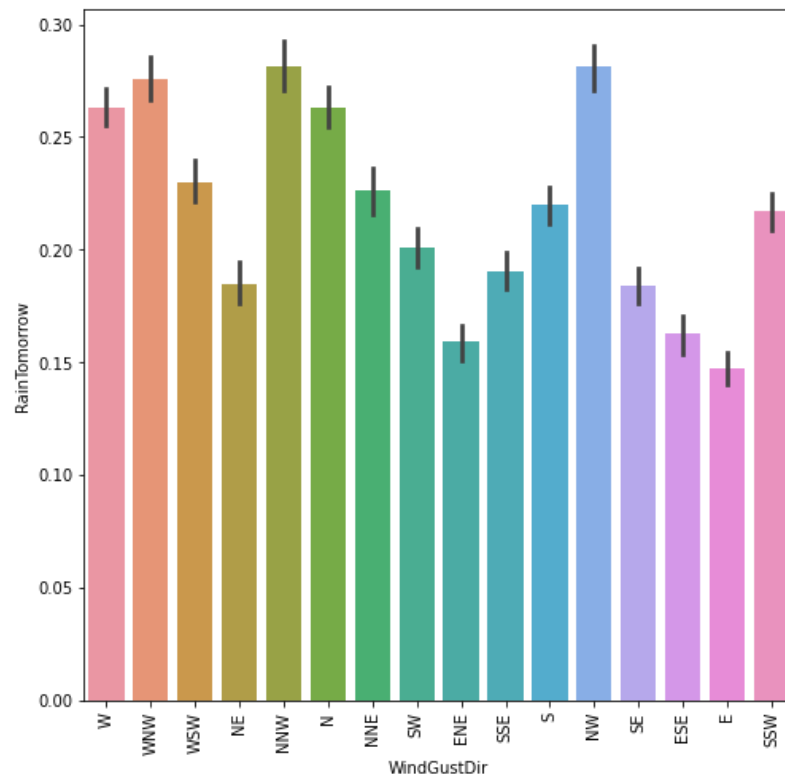
Eksplorasi pada hipotesis yang ke-3 didasari karena atribut WindGustDirection merupakan fitur yang memiliki nilai information gain yang tinggi. Hal ini didasari dengan eksplorasi menggunakan KNN.

Eksplorasi pada hipotesis yang ke-3 ini juga dilakukan dengan dasar eksplorasi menggunakan Decision Tree, dimana fitur WindGustDirection memiliki nilai entropy yang besar.

Dengan melakukan eksplorasi data dengan menggunakan matplotlib.pyplot didapatkan visualisasi fitur WindGustDir dengan

data target RainTomorrow. Dari visualisasi tersebut didapatkan hasil tertinggi yaitu arah tiupan angin North Northwest dan Northwest.

Berikut ini merupakan visualisasi dari eksplorasi menggunakan matplotlib dengan sumbu X adalah fitur WindGustDirection dan sumbu Y adalah fitur RainTomorrow :



Gambar 5. Visualisasi fitur WindGustDir dengan RainTomorrow

Dari hasil tersebut dapat disimpulkan bahwa arah angin terkuat cenderung mengarah ke Northwest dan North.

3.1.4. Algoritma apa yang memiliki akurasi terbaik dalam memprediksi hujan?

Eksplorasi pada hipotesis yang ke-4 bertujuan untuk mencari algoritma yang memiliki akurasi terbaik dalam memprediksi hujan. Pada eksplorasi ini dilakukan perbandingan akurasi dengan menggunakan algoritma Neural Network with ScikitLearn, Decision Tree, dan KNN.

Dari hasil perhitungan menggunakan NN SKLearn diatas, kami menggunakan berbagai variasi fungsi aktivasi relu dan solver adam serta iterasi. Dari eksplorasi tersebut didapatkan **hasil akurasi sebesar 0.85** dengan pameter hidden layer size (1024,512,256), learning_rate_init = 0.001, activation = 'relu', solver= 'adam', max_iter=500

[[653 354 138]					
[128 33014 785]					
[149 4950 4591]]					
		precision	recall	f1-score	support
	0	0.70	0.57	0.63	1145
	1	0.86	0.97	0.91	33927
	2	0.83	0.47	0.60	9690
	accuracy				0.85 44762
	macro avg				0.80 0.67 0.72 44762
	weighted avg				0.85 0.85 0.84 44762

Gambar 6. Hasil akurasi NN with SKLearn

Eksplorasi berikutnya yaitu menggunakan algoritma Decision Tree. Eksplorasi tersebut dilakukan dengan variasi berikut ini :

test_size: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9],

max_depth: [3, 5, 7, 11, 13, 15, 17],
criterion: ['entropy']

Dari hasil perhitungan dengan menggunakan Decision Tree diatas, didapatkan **hasil akurasi sebesar 0.8288** dengan parameter test size sebesar 0.3 dan max depth sebesar 11 dengan menggunakan entropy.

t_size	max_depth	criterion	akurasi	f1					
0.1	3	entropy	0.806	0.806	0.4	3	entropy	0.816	0.816
0.1	5	entropy	0.820	0.820	0.4	5	entropy	0.819	0.819
0.1	7	entropy	0.826	0.826	0.4	7	entropy	0.825	0.825
0.1	11	entropy	0.826	0.826	0.4	11	entropy	0.825	0.825
0.1	13	entropy	0.823	0.823	0.4	13	entropy	0.818	0.818
0.1	15	entropy	0.813	0.813	0.4	15	entropy	0.808	0.808
0.1	17	entropy	0.802	0.802	0.4	17	entropy	0.799	0.799
0.2	3	entropy	0.806	0.806	0.5	3	entropy	0.815	0.815
0.2	5	entropy	0.817	0.817	0.5	5	entropy	0.819	0.819
0.2	7	entropy	0.822	0.822	0.5	7	entropy	0.825	0.825
0.2	11	entropy	0.826	0.826	0.5	11	entropy	0.824	0.824
0.2	13	entropy	0.820	0.820	0.5	13	entropy	0.820	0.820
0.2	15	entropy	0.813	0.813	0.5	15	entropy	0.807	0.807
0.2	17	entropy	0.799	0.799	0.5	17	entropy	0.796	0.796
0.3	3	entropy	0.816	0.816	0.6	3	entropy	0.816	0.816
0.3	5	entropy	0.820	0.820	0.6	5	entropy	0.820	0.820
0.3	7	entropy	0.826	0.826	0.6	7	entropy	0.825	0.825
0.3	11	entropy	0.829	0.829	0.6	11	entropy	0.823	0.823
0.3	13	entropy	0.825	0.825	0.6	13	entropy	0.814	0.814
0.3	15	entropy	0.812	0.812	0.6	15	entropy	0.802	0.802
0.3	17	entropy	0.803	0.803	0.6	17	entropy	0.792	0.792

Gambar 7. Hasil perhitungan menggunakan Decision Tree

Eksplorasi berikutnya yaitu menggunakan algoritma K- Nearest Neighbour. Eksplorasi tersebut dilakukan dengan variasi berikut ini :

```
knn_test = {
    test_size: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6],
```

```

n_neighbor: [3, 5, 7, 9, 11, 13, 15, 17],
weights: ['uniform', 'distance'],
}

```

Dari eksplorasi tersebut didapatkan hasil akurasi sebesar 0.8137 dengan test size sebesar 0.6 dan N-Neighbours sebesar 17.

test_size	n_neighbor	weights	akurasi	f1
0.1	3	uniform	0.812	0.812
0.1	3	distance	0.817	0.817
0.1	5	uniform	0.815	0.815
0.1	5	distance	0.820	0.820
0.1	7	uniform	0.816	0.816
0.1	7	distance	0.821	0.821
0.1	9	uniform	0.817	0.817
0.1	9	distance	0.822	0.822
0.1	11	uniform	0.813	0.813
0.1	11	distance	0.819	0.819
0.1	13	uniform	0.812	0.812
0.1	13	distance	0.818	0.818
0.1	15	uniform	0.814	0.814
0.1	15	distance	0.819	0.819
0.1	17	uniform	0.813	0.813
0.1	17	distance	0.818	0.818
0.2	3	uniform	0.810	0.810
0.2	3	distance	0.815	0.815
0.2	5	uniform	0.814	0.814
0.2	5	distance	0.818	0.818
0.2	7	uniform	0.816	0.816
0.2	7	distance	0.820	0.820
0.2	9	uniform	0.816	0.816
0.2	9	distance	0.821	0.821
0.2	11	uniform	0.815	0.815
0.2	11	distance	0.820	0.820
0.2	13	uniform	0.814	0.814
0.2	13	distance	0.819	0.819
0.2	15	uniform	0.818	0.818
0.2	15	distance	0.822	0.822
0.2	17	uniform	0.814	0.814
0.2	17	distance	0.819	0.819

Gambar 8. Hasil perhitungan menggunakan KNN

Dari ketika algoritma diatas, dapat disimpulkan bahwa algoritma yang memiliki akurasi terbaik adalah Neural Network dengan menggunakan SKLearn dengan akurasi sebesar 0.85.

3.2. Pembahasan dan Kesimpulan

Uji coba dataset rain Australia yang dilakukan mendapatkan hasil yang sudah menjawab semua rumusan masalah. Sebelum mencari akurasi terbaik, hal yang dilakukan adalah dengan mencari fitur yang paling berpengaruh. Pencarian tersebut dilakukan dengan algoritma Decision Tree dan KNN. Setelah pencarian tersebut dilakukan, didapatkan fitur Humadity3pm adalah fitur yang paling berpengaruh.

Setelah melakukan uji coba pencarian fitur yang paling berpengaruh, lakukanlah pencarian nilai akurasi terbaik dengan berbagai macam algoritma yang dilakukan mendapatkan nilai akurasi yang berbeda. Akurasi yang didapatkan dengan menggunakan algoritma NN-Sklearn mendapatkan nilai 0.85, algoritma Decision Tree mendapatkan nilai 0.8288, dan algoritma KNN mendapatkan nilai 0,81. Dari nilai akurasi yang didapat bisa disimpulkan bahwa algoritma yang memiliki akurasi terbaik adalah Decision Tree dengan akurasi sebesar 0.8137.

DAFTAR PUSTAKA

Dataset Rain in Australia. (2019). Weather dataset.
<https://www.kaggle.com/jsphyg/weather-dataset-rattle-package?select=weatherAUS.csv> .

Scikit, Learn. (2020). Neural network models. https://scikit-learn.org/stable/modules/neural_networks_supervised.html.

Singh, Deepika. (2019). Machine Learning with neural network using scikit-learn.
<https://www.pluralsight.com/guides/machine-learning-neural-networks-scikit-learn>.