MIND (Multimedia Artificial Intelligent Networking Database) Journal

MIND Journal | ISSN (p): 2528-0015 | ISSN (e): 2528-0902 | Vol. 8 | No. 2 | Halaman 173 - 187
DOI : https://doi.org/10.26760/mindjournal.v8i2.173-187 Desember 2023

Prakiraan Hujan menggunakan Metode Random Forest dan Cross Validation

GALIH ASHARI RAKHMAT, WISNU MUTOHAR

Informatics, Faculty of Industrial Technology, Institut Teknologi Nasional Bandung
Email : galihas@itenas.ac.id, wmutohar@mhs.itenas.ac.id
Received 05 Juli 2023 **|** *Revised* 28 Juli 2023 **|** *Accepted* 21 September 2023

ABSTRAK

*Cuaca hujan adalah salah satu unsur iklim yang tinggi keragamannya, membuat pola sebaran hujan disetiap daerah cenderung tidak menentu**.** Akibatnya informasi akan kondisi cuaca hujan yang diberikan tidak tepat, maka perlu dilakukan prakiraan hujan yang akurat untuk dapat mengantisipasi kondisi cuaca hujan yang tidak menentu. Dengan menggunakan data historis cuaca, serta penggunaan model machine learning dapat diterapkan untuk analisis permasalahan prakiraan hujan. Menggunakan metode random forest dengan komponen hyperparamter e_estimator dan max_depth serta teknik cross validation untuk menghasilkan kinerja model yang optimal. Lalu mengevaluasi kinerja model menggunakan matriks MSE, RMSE dan MAE. Data cuaca BMKG digunakan rentang tahun 2015-2021 dengan jumlah 2557 data yang dimana dibagi menjadi 80% data latih dan 20% data uji. Hasil pengujian model random forest dengan nilai n_estimator : 100, max_depth : None, dan cross validation : 3 menghasilkan kinerja paling optimal. Dengan menghasilkan matriks evaluasi nilai MSE : 0,086, RMSE : 0,290 dan MAE : 0,186. Serta dalam pengujian aplikasi penentuan kondisi hujan dari 30 kasus data menghasilkan persentase 60%.*

***Kata kunci**: Prakiraan Hujan, Machine Learning, Random Forest, Cross Validation*

Rainy weather is one of the climate elements that has high diversity, making the pattern of distribution of rain in each area tends to be erratic. As a result, the information on rainy weather conditions provided is not correct, it is necessary to make accurate rain forecasts to be able to anticipate erratic rainy weather conditions. Using historical weather data, as well as the use of machine learning models can be applied to analyze rain forecasting problems. Using the random forest method with hyperparameter components e_estimator and max_depth and cross-validation techniques to produce optimal model performance. Then evaluate the performance of the model using the MSE, RMSE, and MAE matrices. BMKG weather data is used for the 2015-2021 range with a total of 2557 data which is divided into 80% training data and 20% test data. The results of testing the random forest model with a value of n_estimator: 100, max_depth: None, and cross-validation : 3 produce the most optimal performance. By producing an evaluation matrix of MSE values: 0.086, RMSE: 0.290, and MAE: 0.186. As well as in testing the application to determine rain conditions from 30 cases of data, it produced a percentage of 60%.

***Keywords**: Rain Forecasting, Machine Learning, Random Forest, Cross Validation*

1. PENDAHULUAN

Pentingnya penerapan metode *machine learning* yang dapat melakukan prediksi cuaca hujan. Hujan adalah salah satu unsur iklim yang paling tinggi, karakteristik hujan diberbagai wilayah tentu tidak sama, dikarenakan adanya kondisi yang mempengaruhi terjadinya hujan. Akibatnya pola sebaran hujan disetiap daerah cenderung berbeda dan tidak merata **(Putramulyo & Alaa, 2018)**. Besarnya hujan pada masa yang datang tidak dapat dipastikan secara tepat, tetapi dapat diperkirakan dengan memanfaatkan data cuaca dari masa lalu untuk sebagai dasar dalam memprediksi besaran hujan **(Rofiq, dkk, 2020)**.

Lembaga pemerintah Indonesia yang bernama Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) adalah sebuah instansi pemerintah yang memiliki tanggung jawab melakukan pengamatan dan pengukuran mengenai faktor gejala alam yang terjadi serta faktor yang mempengaruhi cuaca dalam terjadinya hujan. Sementara dalam membuat prakiraan cuaca, terdapat beberapa permasalahan yang perlu diatasi. Pertama, diperlukan banyak sumber data, termasuk pengamatan langsung, gambar kondisi awan dari citra satelit, dan hasil pemindaian radar. Permasalahan kedua, prakiraan cuaca sering kali bergantung pada pengetahuan dan keterampilan prakirawan cuaca, yang dimana dapat menghasilkan prakiraan yang berbeda-beda dari satu prakirawan ke prakirawan lain. Perbedaan pendapat dapat membingungkan dalam pengambilan keputusan yang pada akhirnya mengakibatkan penurunan kualitas informasi prakiraaan cuaca hujan yang disampaikan kepada masyarakat luas **(Rofiq, dkk, 2020)**.

Metode yang sering digunakan oleh peneliti untuk kasus prediksi atau klasifikasi adalah *machine learning*. *Machine learning*, didefinisikan sebagai penerapan komputer dan algoritma matematika dengan memanfaatkan pembelajaran dari data untuk menghasilkan prediksi dimasa yang akan datang. Salah satu metode pada *machine learning* yang dapat diterapkan dalam melakukan prediksi ialah *random forest* dan *cross-validation*. *Random forest* memiliki kelebihan dapat mengatasi data outlier, sedangkan teknik *cross-validation* memiliki kelebihan dapat mengatasi overfitting serta memberikan peningkatan performa pada kinerja model **(Roihan, dkk, 2020)**.

Pada tahun 2021 penelitian yang dilakukan **(Mursianto, dkk, 2021)** menggunakan metode klasifikasi *Random Forest* dan *XGBoost* serta implementasi teknik *Smote* pada kasus prediksi hujan dengan menggunakan data cuaca Australia. Lalu Pada tahun 2018 **(Primajaya & Sari, 2018)** menggunakan metode *random forest* dengan pengujian 10-*fold cross validation* dalam prediksi hujan. Pada penelitian ini enggunakan data parameter cuaca yakni suhu rata-rata (TEMP), *mean sea level pressure* (SLP), *mean station pressure* (STP), kecepatan angin rata-rata (WDSP), dan kecepatan angin maksimum (MXSPD).

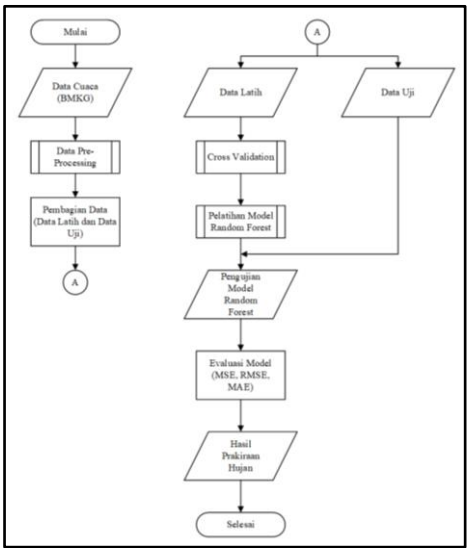
Berdasarkan penjelasan tersebut, maka penelitian ini bertujuan untuk melakukan prediksi hujan dengan pengimplentasian metode *random forest* dan teknik *cross-validation* dalam memprakirakan cuaca hujan, menggunakan parameter pengaruh terjadinya hujan dari data hasil pengamatan yang dilakukan BMKG yaitu suhu minimum, suhu maksimum, suhu rata-rata, kelembaban rata-rata, lamanya penyinaran, kecepatan angin maksimum, kecepatan angin rata-rata, arah angin kecepatan maksimum dan arah angin terbanyak. Lalu kinerja dari model *random forest* dan teknik *cross-validation* diukur dengan menggunakan *matriks* evaluasi model yaitu *mean square error* (MSE), *root mean square error* (RMSE), dan *mean absolute error* (MAE) untuk dapat melihat seberapa baik prediksi yang dilakukan. Diharapkan mendapatkan hasil yang baik guna memberikan keakuratan prakiraan cuaca hujan serta memberikan informasi yang tepat dengan penerapan model pembelajaran mesin.

2. METODE

2.1. *Flowchart* Sistem

Pada flowchart sistem mula-mula dataset cuaca diinputkan pada sistem dengan ekstensi file data yaitu CSV (Microsoft Excel). Selanjutnya dataset masuk pada tahap data *pre-processing,* dimana dataset dilakukan analisis dengan penggambaran grafik dan pengecekan data menggunakan beberapa teknik visualisasi data dan juga EDA (*Exploratory Data Analysis*). Berdasarkan hasil informasi yang didapat dari visualisasi data dan juga EDA, dataset yang akan digunakan untuk pemodelan *machine learning* terlebih dahulu dilakukan pembersihan data (*Data Cleansing*) agar data siap digunakan pada pemodelan.

Selanjutnya *dataset* yang telah melalui tahapan data *pre-processing* tersebut, kemudian dilakukan pembagian menjadi 2 bagian. Data bagian pertama sebagai data latih untuk pelatihan model, data pada bagian kedua sebagai data uji pada pengujian model yang telah dilatih sebelumnya. Selanjutnya setelah dilakukan pembagian dataset, langkah berikutnya yaitu terlebih dahulu menggunakan data latih masuk pada proses *cross validation* dimana dilakukan *splitting* data untuk dilakukan pelatihan model *random forest* dengan data *training*. Setelah dilakukan *training* maka akan menghasilkan model yang dimana akan digunakan untuk dilakukan pengujian dengan menggunakan data *testing* (data uji), lalu model diuji dengan data uji. Setelah dilakukan pengujian model maka dilakukan evaluasi dari model yang dihasilkan dengan menentukan nilai MSE, RMSE, dan MAE. Proses tersebut ditujukan pada Gambar 1.



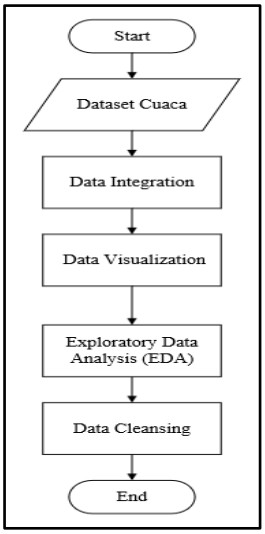
Gambar 1. Flowchart Sistem

Hasil evaluasi, model yang menghasilkan performasi paling optimal berdasarkan hasil eksperimen uji dan analisisnya akan disimpan untuk kemudian digunakan dalam aplikasi untuk melakukan prakiraan hujan berdasarkan inputan fitur cuaca.

2.2. Flowchart Data *Pre-Processing*

Dalam proses pra-pemrosesan dataset cuaca, langkah data *integratian* dilakukan untuk mengabungkan semua tabel data cuaca BMKG yang berkisar dari januari 2015 hingga Desember 2021 menjadi satu tabel data cuaca. Setelah dilakukan data *integration*, dataset akan dianalisis terlebih dahulu dengan dilakukan penggambaran grafik dan pemeriksaan data menggunakan beberapa teknik visualisasi data dan EDA. Selanjutnya, setelah dianalisis data-data yang digunakan untuk pemodelan *machine learning* perlu dilakukan data *cleansing*. Pada penelitian ini salah satu dari tahap data *cleansing* adalah pembersihan data kosong dan data

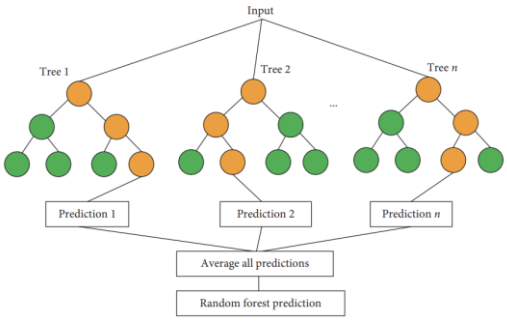
tidak terukur yang dinyatakan dengan nilai 8888 yang dibantu dengan library python bernama pandas. Data yang memiliki nilai 8888 dan data kosong (NULL) akan dilakukan penghapusan atau *drop* baris yang memiliki data kosong tersebut serta nilai 8888. Setalah proses *cleansing* akan menghasilkan data yang siap untuk pemodelan *machine learning*. Flowchart *Pre-processing* ditujukan pada Gambar 2.



Gambar 2. Flowchart Data *Pre-processing*

2.3. *Random Forest*

Pengembangan dari metode *decision tree* yaitu *random forest* yang melibatkan penggunaan beberapa *decision tree*. Setiap *decision tree* pada *random forest* dilatih dengan menggunakan sampel individu, dan dalam pembentukan pohon, setiap atribut dibagi secara acak antara atribut berbagai subset yang tersedia **(Supriyadi, dkk, 2020)**. Algoritma *random forest* (RF) beroperasi dengan menggunakan pemisah biner yangn bersifat rekursif untuk mencapai node akhir dalam struktur pohon keputusan, yang pada dasarnya adalah pohon klasifikasi dan regresi. *Random forest* menghasilkan sejumlah besar pohon independen dengan cara memilih subset acak dari data pelatihan dengan menggunakan bantuan teknik *bootstrap* dan juga mengambil subset acak dari variabel input. *Bootstrap aggregating* atau biasa disebut dengan Bagging, melibatkan penggunaan teknik Bootstrap atau pengambilan sampel sebanyak n kali dengan penggantian dari data pelatihan asli untuk membuat set pelatihan yang berbeda-beda. Dalam prosesnya, setiap data pelatihan digunakan untuk membuat pohon klasifikasi, dan hasil akhir adalah agregasi suara terbanyak dalam kasus klasfisikasi dan rata-rata dalam kasus regresi **(Religia, dkk, 2021)**. Berikut adalah struktur umum dari random forest yang ditunjukan Gambar 3.



Gambar 3. Struktur Umum Model *Random Forest*

Pohon keputusan terdiri dari tiga jenis simpul: *root node*, *internal node*, dan *leaf node*. *Root node*, yang juga disebut akar, adalah simpul paling atas dalam pohon keputusan. *Internal node* adalah simpul percabangan yang memiliki minimal dua output dan satu input. Sementara itu, *leaf node* atau terminal node adalah simpul terakhir yang memiliki satu input dan tidak menghasilkan output tambahan. Dalam pembentukan pohon keputusan, nilai *entropy* digunakan untuk mengukur tingkat ketidakmurnian, dan *information gain* digunakan sebagai kriteria untuk memilih atribut yang dijadikan sebagai simpul, baik sebagai akar (*root*) atau simpul internal. Pemilihan atribut ini didasarkan pada nilai *information gain* tertinggi **(Komunikasi, dkk, 2017)**. Rumus *entropy* dan *gain information* ditunjukan pada Persamaan (1) dan Persamaan (2).

$$Entropy\ (Y)\ =\ -\sum_i P(c|Y)log_2\ P(c|Y) \tag{1}$$

Dimana,

Y = himpunan kasus

p(c|Y) = proporsi nilai Y terhadap kelas c.

$$Information\ Gain\ (Y,a) = Entropy\ (Y) - \sum_{veValues(a)} \frac{|V_v|}{|Y_a|}\ Entropy\ (Y_v) \tag{2}$$

Dimana,

Values(a) = merupakan semua nilai yang mungkin dalam himpunan kasus a.
Y_v = subkelas dari Y dengan kelas v yang berhubungan dengan kelas a.
Y_a = semua nilai yang sesuai dengan a.

2.4. Cross Validation

Cross-validation ialah sebuah teknik yang digunakan dalam mengevaluasi dan membandingkan kinerja algoritma pembelajaran dengan membagi dataset menjadi dua bagian. bagian pertama digunakan untuk mempelajari data latih untuk melatih model, sementara yang lainnya digunakan untuk menguji atau mengvalidasi model **(Bhattacharya, 2014)**. Salah satu implementasi dari metode cross-validation ini dikenal sebagai k-fold cross-validation, dimana dataset awal dibagi secara acak menjadi K kelompok. Kemudian salah satu dari kelompok-kelompok ini diambil sebagai data uji, sedangkan sisanya data latih **(Cahyanti, dkk, 2020)**.

3. HASIL DAN PEMBAHASAN

3.1. Dataset

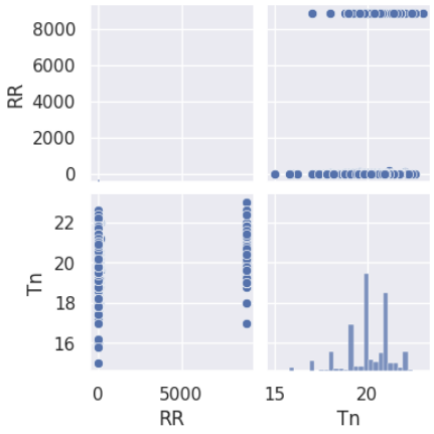
Dataset cuaca bersumber dari webiste BMKG http://dataonline.bmkg.go.id/home yang telah dilakukan pengunduhan dengan rentang data tahun 2015-2021 dengan jumlah 2557 data. Mula – mula dataset dilakukan tahap data *integratian* untuk mengabungkan data cuaca BMKG dari januari 2015 sampai Desember 2021 menjadi satu data cuaca tahun 2015-2021 dan setelah digabungkan menjadi satu tabel data cuaca disimpan dengan pengubahan file ekstasi dari excel ke csv. Dataset cuaca yang diunduh memiliki fitur-fitur yaitu Tanggal, suhu minimum (Tn), Suhu maksimum (Tx), Suhu rata-rata (Tavg), Kelembaban rata-rata (RH_avg), lamanya penyinaran (ss), Arah angin kecepatan maksimum (ddd_x), Arah angin terbanyak (ddd_car), Kecepatan angin maksimum (ff_x), dan Kecepatan angin rata-rata (ff_avg).

3.2. *Data Visualization and Exploratory Data Analysis* (EDA)

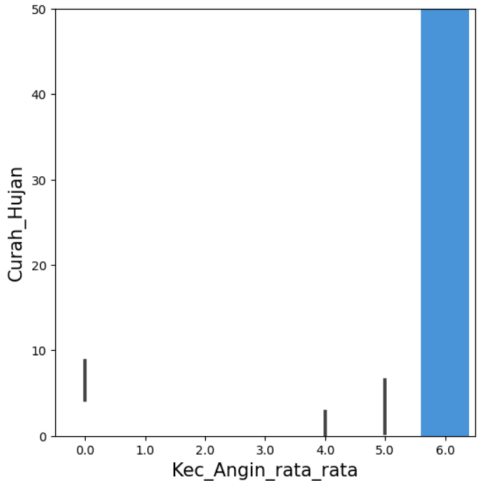
Pada tahap visualisasi data dan Exploratory Data Analysis (EDA), fitur – fitur yang ada pada dataset di analisis persebaran datanya terhadap fitur target yaitu curah hujan. Mula-mula dataset terlebih dahulu dilakukan pengecekan informasi yang terdapat pada dataset, seperti jumlah data, nama-nama kolom, dan tipe datanya serta sebagainya. Dari hasil pengecekan informasi dataset diketahui jumlah dataset yakni 2557 data dengan jumlah kolom 11 termasuk kolom tanggal, tipe data pada dataset terdapat tipe data float (numerik) dan data object (string).

Langkah selanjutnya, yaitu melakukan pengecekan terhadap nilai NULL (nilai kosong) dan nilai 8888 (tidak diukur) yang terdapat pada setiap fitur dataset. Hasil pengecekan menunjukan bahwa pada semua fitur dataset kecuali tanggal, mengandung nilai NULL pada datanya sedangkan untuk nilai 8888 (tidak diukur), fitur yang mengandung nilai 8888 yakni fitur target (Curah Hujan).

Langkah berikutnya yaitu pengecekan data outlier pada setiap fiturnya dan persebaran data dari setiap fiturnya terhadap fitur target yaitu curah hujan. Dimulai dari fitur bersifat numerikal kontinyu dahulu, dari hasil pengecekan dapat dilihat bahwa fitur suhu minimum (Tn), suhu maksimum (Tx), suhu rata-rata (Tavg), kelembaban rata-rata (RH_avg), lama penyinanar (ss), dan arah angin kecepatan maksimum (ddd_x) menunjukan nilai – nilai persebaranya stabil terhadap variabel target curah hujan dari hasil *scatter plot diagram* yang dilakukan. Namun pada dataset setiap fiturnya memiliki nilai NULL dan pada fitur target (curah hujan) juga terdapat nilai tidak terukur (8888) sehingga hasil gambaran *scatter plot diagram* yang dihasilkan cenderung lebih ke nilai – nilai yang tinggi atau kesisi samping atau atas bawah diagram yang diperlihatkan pada Gambar 4 yakni pada fitur Tn dan RR, ini disebabkan terdapat nilai 8888 (tidak terukur) pada fitur target curah hujan. Pada fitur bersifat numerikal diskrit yaitu kecepatan angin maksimum (ff_x), kecepatan angin rata-rata (ff_ avg) dan arah angin terbanyak (ddd_car) persebaran datanya tidak stabil terhadap variabel target curah hujan (RR) yang dilihat dari hasil *bar plot diagram* visualisasi datanya. Namun pada fitur kecepatan angin rata-rata (ff_ avg) menunjukan nilai semakin tinggi dapat mempengaruhi tingginya curah hujan (RR) yang ditunjukan pada Gambar 5.



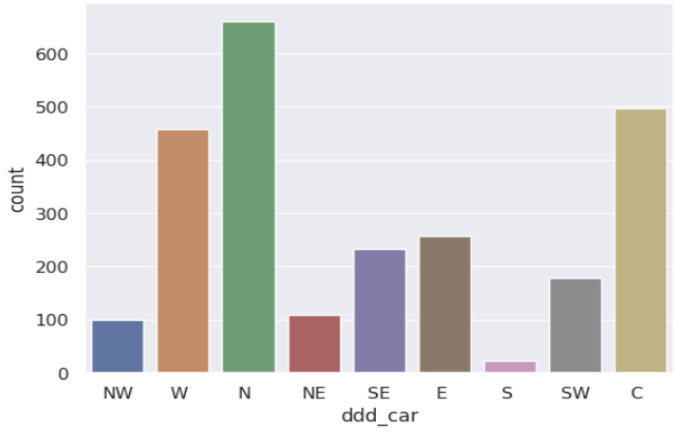
Gambar 4. Visualisasi Data RR dan Tn



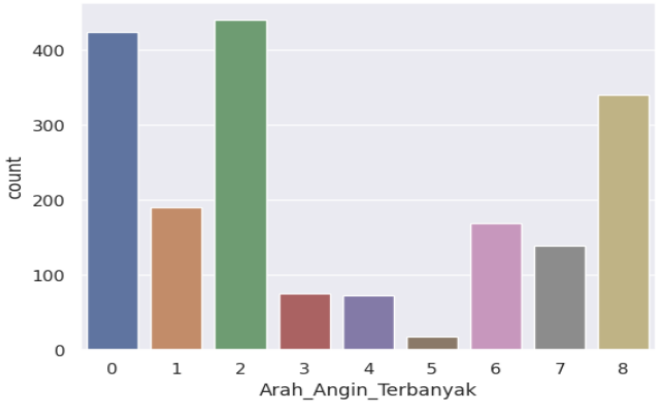
Gambar 5. Visualisasi Data Kecepatan Angin Rata-rata

3.3. Data Cleansing

Berdasarkan hasil visualisasi data dan *Exploratory Data Analysis* (EDA), selanjutnya melakukan pembersihan data (*data cleansing*) dan transformasi data. Langkah pertama yang dilakukan adalah pengubahan nama kolom pada fitur dataset, ini dilakukan agar nama kolom dapat mudah dipahami dan diingat dengan lebih jelas yang mana sebelumnya nama kolom pada dataset berupa singkatan atau kode saja, seperti curah hujan diberi kode dengan RR. Langkah kedua yaitu melakukan transformasi data pada fitur arah angin terbanyak (ddd_car), yang dimana fitur arah angin terbanyak memiliki type data huruf (*string*) yang tidak seragam dengan fitur lain yang memilki *type* data angka (*numerik*). Sehingga dilakukan transformasi dengan bantuan library scikit-learn (*sklearn*) yaitu fungsi *LabelEncoder*, yang mengubah data huruf ke *type* data angka. *LabelEncoder* mengubah setiap nilai dalam kolom fitur menjadi angka yang berurutan. Dimana data pada fitur arah angin terbanyak (ddd_car) yang berisi data antara lain NW (*northwest*), W (*west*), N (*north*), NE (*northwest*), SE (*southeast*), E (*east*), S (*south*), SW (*southwest*) dan C (*calm* / tidak ada angin / lemah) lalu data yang berupa huruf (*string*) tersebut diubah ke data angka (*numerik*) secara berurutan menjadi NW : 0, W : 1, N : 2, NE : 3, SE : 4, E : 5, S 6, SW 7 dan C : 8. Yang ditunjukan pada Gambar 5 dan Gambar 6.



Gambar 5. Sebelum Transformasi Data



Gambar 6. Sesudah Transformasi Data

Setelah melakukan transformasi data pada fitur arah angin terbanyak (ddd_car) dan menjadi *type* data *numerik*, langkah ketiga yaitu melakukan drop fitur tanggal dan pembersihan nilai data bernilai NULL dan data bernilai 8888 (tidak terukur) pada setiap fitur yang memiliki nilai NULL dan nilai 8888 dengan bantuan *library pandas*. Setelah melakukan drop fitur dan pembersihan nilai NULL serta nilai 8888 maka, didapatlah 10 fitur yang tersisa dengan jumlah data 1867 data setelah tahap *pre-processing*. 10 fitur tersebut adalah suhu minimum, suhu maksimum, suhu rata-rata, kelembaban rata-rata, lamanya penyinaran, kecepatan angin maksimum, kecepatan angin rata-rata, arah angin kecepatan maksimum dan arah angin terbanyak. Sedangkan fitur tanggal tidak digunakan karena pada penelitian ini menggunakan fitur yang mempengaruhi terjadinya hujan sehingga tanggal tidak digunakan sebab tidak termasuk mempengaruhi terjadinya hujan.

3.4. Skenario Eksperimen

Pada penelitian ini terdapat nilai *outlier* pada fitur dataset, *outlier* merupakan data yang memiliki ciri-ciri yang berbeda jauh dari pengamatan serta dalam bentuk nilai ekstrim. Sehingga data *outlier* akan dimasukan pada bagian skenario eksperimen model penelitian dengan menggunakan data *outlier* dan tanpa data *outlier*. Dimana untuk persiapan penggunaan tanpa data *outlier* dilakukan transformasi menggunakan rumus kuartil dengan menghitung ambang atas dan ambang bawah sehingga nantinya nilai akan berada pada antara nilai ambang atas dan ambang bawah. Lalu sebelum dilakukan ekperimen model random forest pada penelitian ini, terdapat beberapa percobaan yang dilakukan pada metode yaitu dilakukan percobaan eksperimen dengan menggunakan metode *random forest* saja dan menggunakan metode *random forest* dengan tambahan teknik *cross validation*. Serta dilakukan eksperimen pada parameter *random forest* yakni n_*estimator* atau jumlah pohon keputusan yang dibentuk, lalu max_*depth* atau kedalaman pohon keputusan yang dibentuk. Menurut **(Haristu & Rosa, 2019)** semakin banyak pohon (*n_estimator*), maka semakin besar akurasi yang dihasilkan serta semakin besar nilai atau tidak dibatasi kedalaman pohonnya (*max_depth*) membuatnya memiliki akurasi yang semakin baik. Berikut daftar skenario eksperimen diperlihatkan pada Tabel 1.

Tabel 1. Daftar Skenario Eksperimen

Kode Eksperimen	Data	Metode	n_estimator	max_depth	CV
A1	Dengan data outlier	Random Forest	100	none	-
A2	Tanpa data outlier		100	none	-
B1	Dengan data outlier		100	3	-
B2			100	5	-
B3			100	7	-
B4			100	10	-
B5			200	3	-
B6			300	5	-
B7			400	7	-
B8			500	10	-
C1	Tanpa data outlier		100	3	-
C2			100	5	-
C3			100	7	-
C4			100	10	-
C5			200	3	-
C6			300	5	-
C7			400	7	-
C8			500	10	-
D1	Dengan data outlier	Random Forest + Cross Validation	100	None	3
D2			100	None	4
D3			100	None	5
D4			100	None	6
D5			100	None	7
D6			100	None	8
D7			100	None	9
D8			100	None	10
E1	Tanpa data outlier		100	None	3
E2			100	None	4
E3			100	None	5
E4			100	None	6
E5			100	None	7
E6			100	None	8
E7			100	None	9
E8			100	None	10

Pada Tabel 1 diatas kedalaman pohon (*max_depth*) sama dengan none artinya kedalaman pohon tidak dibatasi sehingga kedalaman pohon keputusan dapat tumbuh secara penuh. Sedangkan CV (*cross-validation*) sama dengan (-) artinya pada eksperimen tersebut tidak menggunakan komponen teknik cross validation.

3.5. Pelatihan Model

Hasil pelatihan model dengan menggunakan data pelatihan pada dataset yang datanya belum ditransformasi outlier (Dengan data *outlier*), menghasilkan nilai evaluasi model ditujukan pada Tabel 2.

Tabel 2. Pelatihan Model (Dengan data *outlier*)

Kode Eksperimen	MSE	RMSE	MAE
A1	27.897	5.281	3.083
B1	167.761	12.952	7.663
B2	136.480	11.682	6.986
B3	100.115	10.005	6.017
B4	54.272	7.366	4.424
B5	167.193	12.930	7.638
B6	136.663	11.690	6.982
B7	96.362	9.816	5.959
B8	53.318	7.301	4.406
D1	0.125	0.351	0.068
D2	0.101	0.315	0.188
D3	0.077	0.269	0.128
D4	0.120	0.342	0.092
D5	0.104	0.309	0.123
D6	0.100	0.301	0.144
D7	0.103	0.310	0.118
D8	0.081	0.267	0.088

Sedangkan hasil pelatihan model dengan menggunakan data pelatihan pada dataset yang datanya sudah dilakukan ditransformasi *outlier* (tanpa data *outlier*), menghasilkan nilai evaluasi model seperti pada Tabel 3.

Tabel 3. Pelatihan Model (Tanpa data outlier)

Kode Eksperimen	MSE	RMSE	MAE
A2	6.778	2.603	1.920
C1	45.382	6.736	5.033
C2	38.286	6.187	4.569
C3	28.099	5.300	3.891
C4	15.258	3.906	2.839
C5	45.469	6.743	5.036
C6	38.397	6.196	4.597
C7	27.804	5.272	3.880
C8	14.646	3.827	2.795
E1	0.235	0.481	0.242
E2	0.237	0.485	0.228
E3	0.244	0.489	0.242
E4	0.249	0.494	0.246
E5	0.237	0.480	0.250
E6	0.242	0.489	0.222
E7	0.252	0.499	0.229
E8	0.239	0.486	0.248

Berdasarkan hasil pelatihan model, baik dengan menggunakan data outlier dan tanpa data outlier terlihat bahwa pada kode eksperimen B1 (random forest dengan n_estimator : 100, max_depth : 3, CV : -) menunjukan metode dengan kinerja model terburuk yang dihasilkan, hal tersebut terlihat dari matriks evaluasi MSE, RMSE, dan MAE yang dihasilkan merupakan terbesar dibandingkan dengan eksperimen model lainnya. Metode optimal yang dihasilkan

pada tahap pelatihan didapatkan dengan kode eksperimen D8 (random forest + cross validation dengan n_estimator : 100, max_depth : None, CV : 10) dengan menghailkan nilai evaluasi terkecil.

3.6. Pengujian Model

Model yang telah dihasilkan dari pelatihan sebelumnya akan diuji menggunakan data uji. Hasil pengujian model dengan menggunakan dataset belum ditransformasi *outlier* (dengan data *outlier*), menghasilkan nilai evaluasi model seperti pada Tabel 4.

Tabel 4. Pengujian Model (Dengan data outlier)

Kode Eksperimen	MSE	RMSE	MAE
A1	130.370	11.417	7.487
B1	117.615	10.845	7.079
B2	116.945	10.814	6.962
B3	118.407	10.881	7.033
B4	124.314	11.149	7.246
B5	118.950	10.906	7.105
B6	116.838	10.809	7.006
B7	119.976	10.953	7.078
B8	123.426	11.109	7.233
D1	0.086	0.290	0.186
D2	0.108	0.322	0.114
D3	0.108	0.318	0.062
D4	0.130	0.351	0.154
D5	0.236	0.401	0.285
D6	0.154	0.373	0.206
D7	0.169	0.390	0.089
D8	0.328	0.487	0.184

Sedangkan hasil pengujian model dengan menggunakan dataset yang datanya sudah dilakukan ditransformasi outlier (tanpa data outlier), menghasilkan nilai evaluasi model seperti pada Tabel 5.

Tabel 5. Pengujian Model (Tanpa data outlier)

Kode Eksperimen	MSE	RMSE	MAE
A2	42.633	6.529	5.020
C1	41.482	6.440	5.019
C2	40.192	6.339	4.881
C3	40.200	6.340	4.884
C4	41.773	6.463	4.949
C5	41.339	6.429	5.009
C6	40.463	6.361	4.910
C7	40.427	6.358	4.885
C8	41.189	6.417	4.915
E1	0.167	0.405	0.154
E2	0.175	0.418	0.156
E3	0.198	0.427	0.178
E4	0.182	0.391	0.253
E5	0.147	0.376	0.167
E6	0.204	0.442	0.127

Kode Eksperimen	MSE	RMSE	MAE
E7	0.175	0.392	0.158
E8	0.168	0.401	0.205

Berdasarkan hasil pengujian model, baik dengan menggunakan data outlier dan tanpa data outlier terlihat bahwa pada kode eksperimen A1 (random forest dengan n_estimator : 100, max_depth : None, CV : -) menghasilkan model dengan kinerja paling buruk, hal ini terlihat dari hasil nilai matriks evaluasi MSE, RMSE, dan MAE yang merupakan terbesar dari model-model eksperimen lainnya. Sedangkan untuk model dengan performansi paling optimal yaitu dengan kode eksperimen D1(*random forest* + *cross validation* dengan n_estimator : 100, max_depth : None, CV : 3) hal tersebut terlihat dari hasil matriks evaluasi yang merupakan terkecil dari model – model eksperimen lainnya.

3.7. Hasil Pengujian Implementasi

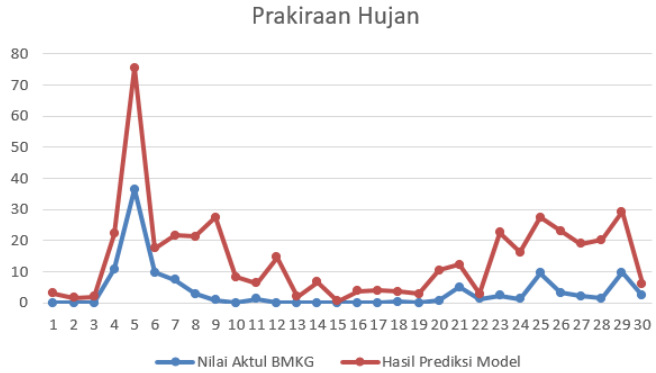
Berikut ini perbandingan nilai hasil prakiraan BMKG dengan hasil prakiraan model yang diimplementasikan pada aplikasi prakiraan hujan berbasis website, dimana pada implementasi ini menggunakan model optimal dari hasil eksperimen uji, diperliharkan pada Tabel 6.

Tabel 6. Hasil Pengujian Implementasi

Tn	Tx	Tavg	RH_avg	ss	ddd_x	ddd_car	ff_x	ff_avg	RR		Kondisi Hujan	
									BMKG	Prediksi	BMKG	Prediksi
20.8	30	24.8	70	4.1	290	W	4	2	0	3.088	Berawan	Ringan
19.8	31	24.4	70	7	270	W	4	2	0	1.69	Berawan	Ringan
20.4	30.6	24.5	74	7.5	210	W	3	2	0	2	Berawan	Ringan
20.8	31.6	24.6	79	8	280	W	3	1	10.8	11.434	Ringan	Ringan
20.6	30.6	23.6	84	5.7	230	C	3	1	36.5	39.037	Sedang	Sedang
19.5	30.2	24	77	5	260	SW	2	1	9.8	7.674	Ringan	Ringan
20.8	28.8	23.6	79	5.9	250	SW	4	2	7.4	14.177	Ringan	Ringan
20.6	28.6	23.3	80	6.2	230	SW	3	2	2.8	18.536	Ringan	Ringan
20.8	28.8	23.8	80	2.7	280	C	4	2	0.8	26.649	Ringan	Ringan
20.6	30.4	24.5	75	4	180	W	4	2	0	8.246	Berawan	Ringan
21	29.2	24.2	78	7.7	250	W	4	2	1.2	5.116	Ringan	Ringan
21.2	30.2	24	79	3	270	W	5	3	0	14.599	Berawan	Ringan
21.2	29.6	24.3	75	8.1	220	W	4	2	0	1.913	Berawan	Ringan
21.2	31	24.8	76	3.1	240	SW	4	2	0	6.601	Berawan	Ringan
22.2	30.8	25.6	70	8.1	200	W	4	2	0	0.451	Berawan	Berawan
22	30.4	25	74	7.1	230	W	6	2	0	3.835	Berawan	Ringan
22	30.2	24.8	74	6	240	SW	4	2	0	3.991	Berawan	Ringan
21.2	31.4	25.2	70	4.3	220	SW	4	2	0.3	3.373	Berawan	Ringan
21	31.2	24.4	76	9	250	C	3	1	0	2.811	Berawan	Ringan
21.1	31.4	24.8	75	4.7	260	W	3	2	0.7	9.722	Ringan	Ringan
21.2	30.8	24	70	7.2	250	C	4	2	5	7.327	Ringan	Ringan
21.7	29.4	23.5	83	7.6	240	C	3	1	1.2	1.702	Ringan	Ringan
20.2	29.6	23.2	81	2.4	280	C	3	1	2.4	20.199	Ringan	Sedang
20.4	30	23.9	78	4.6	250	C	5	2	1.3	14.867	Ringan	Ringan
21.2	27.4	23.3	80	2.1	290	W	8	3	9.6	17.822	Ringan	Ringan
21	28	23.4	81	2.9	290	W	4	2	3.2	19.822	Ringan	Ringan
20	26.2	22.6	86	0.8	230	C	3	1	2.2	16.822	Ringan	Ringan
20	29.6	23	82	0.1	270	W	4	2	1.4	18.862	Ringan	Ringan
19.8	27.5	23	80	5.5	270	SW	4	2	9.8	19.406	Ringan	Ringan
20	29.6	24.3	74	6.2	270	W	4	2	2.3	3.845	Ringan	Ringan

Pada Tabel 6 diatas, hasil perbandingan nilai prakiraan BMKG dengan prakiraan model pada aplikasi untuk memprakirakan kondisi hujan tersebut digunakan data terbaru yang tidak digunakan pada pelatihan maupun pengujian model yang dimana data digunakan adalah data tahun 2022. Berdasarkan pengujian 30 (tiga puluh) data, model berhasil melakukan prakiraan

kondisi hujan cukup baik terlihat dari 30 (tiga puluh) data tersebut model memprakirakan 18 (delapan belas) kondisi hujan dengan benar dari 30 (tiga puluh) percobaan, ini membuktikan bahwa model pada aplikasi cukup bekerja dengan baik dalam melakukan prakiraan kondisi hujan dan nilai curah hujan dengan memberikan persentase kebenaran kondisi hujan 60%. Serta hasil nilai prakiraan BMKG dengan nilai prakiraan model terlihat mendekati namun juga ada yang jauh dari prakiraan BMKG dengan hasil prakiraan model. Ditunjukan pada Gambar 6 yang merupakan visualisasi diagram hasil pengujian prakiraan BMKG dan prakiraan model pada aplikasi, menghasilkan visualisasi prakiraan hujan.



Gambar 6. Pengujian Aplikasi Prakiraan Hujan

4. KESIMPULAN

Berdasarkan hasil analisis serta eksperimen penelitian, maka disimpulkan pada poin – poin berikut :

1. Prakiraan hujan dapat dilakukan menggunakan dataset cuaca BMKG dengan mengimplementasikan metode *random forest* dan teknik *cross validation.* untuk dapat melakukan prakiraan hujan melalui beberapa proses mulai dari persiapan data yang dimana dilakukan analisis pada data untuk mengetahui informasi yang dimiliki oleh data cuaca, selanjutnya dilakukan juga pembersihan dan standarisasi data agar dapat digunakan pada model, lalu dilakukan pemisahan data menjadi data uji dan data latih, selanjutnya dilakukan teknik *cross validation* dengan pembagian lipatan K adalah 3.4.5.6.7.8.9.10, selanjutnya data latih digunakan pada model *random forest,* pada *random forest* akan dilakukan *boostrap sampling* dahulu ke setiap subset banyaknya pohon keputusan yang digunakan, setelah itu akan dibentuk struktur pohon keputusan berdasarkan fitur data cuaca disetiap subsetnya, dibentuk struktur pohon yang terdiri dari simpul akar (*root node*), simpul dalam (*internal node*), dan simpul daun (*leaf node*) hingga menghasilkan struktur pohon keputusan sempurna, pada penelitian ini terdapat eksperimen kombinasi *hyperparameter* jumlah pohon (*n_estimator*) dengan nilai 100, 200, 300,400, 500 dan kedalaman pohon (*max_depth*) dengan nilai 3,5,7,10 serta *None* (tidak dibatasi limit kedalaman pohonya). Setelah melakukan tahapan tersebut dan model telah diuji dengan data uji, maka model akan dievaluasi untuk menghasilkan model optimal yang akan digunakan dalam melakukan prakiraan hujan berdasarkan fitur cuaca yang dimana hasil keluaran prakiraan hujan pada aplikasi berbasis website ialah nilai curah hujan dan klasifikasi kondisi cuaca hujan yakni tidak ada hujan/berawan, hujan ringan, hujan sedang dan hujan lebat.

2. Model *random forest* dan teknik *cross validation* memberikan hasil performansi yang lebih optimal dari pada model *random forest* tanpa menggunakan teknik *cross validation*. Membuktikan bahwa teknik *cross validatian* berpengaruh terhadap kinerja model *random forest* sebab pada *cross validation* data dibagi sesuai nilai k yang dimana sebagian data menjadi data latih dan data uji secara bergantian. Hal ini yang memungkinkan model untuk dipelajari dari berbagai kombinasi data ynag berbeda, sehingga model dapat mengenali pola data secara lebih menyeluruh. Dengan demikian, model dapat memiliki kesempatan untuk belajar dari berbagai variasi data dan dapat meningkatan kemampuan dalam mengenali data baru. Hasil evaluasi pada model *random forest* dan *cross validation* menggunakan nilai hyperparametr n_estimator : 100, max_depth : None, dan CV (nilai k cross validation) : 3 memberikan nilai matriks evaluasi MSE : 0.086, RMSE : 0.290 dan MAE : 0.186. Nilai *matriks error* tersebut merupakan yang terkecil dan optimal dibandingkan hasil evaluasi model – model eksperimen lainnya pada penelitian ini. Hasil tersebut juga didapat pada eksperimen yang menggunakan data outlier ini artinya data outlier memiliki informasi yang penting, sehingga lebih berpengaruh dan memberikan hasil yang lebih baik pada model.

REFERENSI

Arisandi, R. R. R., Warsito, B., & Hakim, A. R. (2022). Aplikasi Naïve Bayes Classifier (Nbc) Pada Klasifikasi Status Gizi Balita Stunting Dengan Pengujian K-Fold Cross Validation. *Jurnal Gaussian*, *11*(1), 130–139. https://doi.org/10.14710/j.gauss.v11i1.33991

Cahyanti, D., Rahmayani, A., & Husniar, S. A. (2020). Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara. *Indonesian Journal of Data and Science*, *1*(2), 39–43. https://doi.org/10.33096/ijodas.v1i2.13

Haristu, R. A., & Rosa, P. H. P. (2019). Penerapan Metode Random Forest untuk Prediksi Win Ratio Pemain Player Unknown Battleground. *MEANS (Media Informasi Analisa Dan Sistem)*, *4*(2), 120–128. https://doi.org/10.54367/means.v4i2.545

Mursianto, G. A., Falih, M., Irfan, M., Sakinah, T., & Sandya, D. (2021). *Perbandingan Metode Klasifikasi Random Forest dan XGBoost Serta Implementasi Teknik SMOTE pada Kasus Prediksi Hujan*. *September*, 41–50.

Naser, M. Z., & Alavi, A. H. (2020). *Insights into Performance Fitness and Error Metrics for Machine Learning*. 1–25. https://www.ptonline.com/articles/how-to-get-better-mfi-results

Primajaya, A., & Sari, B. N. (2018). Random Forest Algorithm for Prediction of Precipitation. *Indonesian Journal of Artificial Intelligence and Data Mining*, *1*(1), 27. https://doi.org/10.24014/ijaidm.v1i1.4903

Putramulyo, S., & Alaa, S. (2018). Prediksi Curah Hujan Bulanan Di Kota Samarinda Menggunakan Persamaan Regresi Dengan Prediktor Data Suhu dan Kelembapan Udara. *Eigen Mathematics Journal*, 13–16. https://doi.org/10.29303/emj.v2i2.20

Religia, Y., Nugroho, A., & Hadikristanto, W. (2021). Klasifikasi Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(1), 187–192. https://doi.org/10.29207/resti.v5i1.2813

Rofiq, H., Pelangi, K. C., & Lasena, Y. (2020). Penerapan Data Mining Untuk Menentukan Potensi Hujan Harian Dengan Menggunakan Algoritma Naive Bayes. *Jurnal Manajemen Informatika Dan Sistem Informasi*, *3*(1), 8–15. http://mahasiswa.dinus.ac.id/docs/skripsi/jurnal/19417.pdf

Roihan, A., Sunarya, P. A., & Rafika, A. S. (2020). *Pemanfaatan Machine Learning dalam Berbagai Bidang : Review paper*. *5*(April), 75–82.

Supriyadi, R., Gata, W., Maulidah, N., & Fauzi, A. (2020). Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah. *E-Bisnis : Jurnal Ilmiah Ekonomi Dan Bisnis*, *13*(2), 67–75. https://doi.org/10.51903/e-bisnis.v13i2.247

Milanović, S., Milanović, S. D., Marković, N., Pamučar, D., Gigović, L., & Kostić, P. (2021). Forest fire probability mapping in eastern serbia: Logistic regression versus random forest method. *Forests*, *12*(1), 1–17. https://doi.org/10.3390/f12010005

Sandag, G. A. (2020). Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest. *CogITo Smart Journal*, *6*(2), 167–178. https://doi.org/10.31154/cogito.v6i2.270.167-178

Wahyuni, E. D., Arifiyanti, A. A., & Kustyani, M. (2019). Exploratory Data Analysis dalam Konteks Klasifikasi Data Mining. *Prosiding Nasional Rekayasa Teknologi Industri Dan Informasi XIV Tahun 2019 (ReTII)*, *2019*(November), (pp. 263–269). http://journal.itny.ac.id/index.php/ReTII

Yadav, D. C., & Pal, S. (2020). Prediction of heart disease using feature selection and random forest ensemble method. *International Journal of Pharmaceutical Research*, *12*(4), 56–66. https://doi.org/10.31838/ijpr/2020.12.04.013