

PREDIKSI CURAH HUJAN MENGGUNAKAN METODE CATEGORICAL BOOSTING (CATBOOST)

Andrian Febriansyah Istianto, Asep Id Hadiana, Fajri Rakhmat Umbara

Teknik Informatika, Universitas Jenderal Achmad Yani

Jl. Terusan Jend. Sudirman, Cibeber, Kec. Cimahi Sel., Kota Cimahi, Jawa Barat 40531

Andrian.febriansyah@student.unjani.ac.id

ABSTRAK

Curah hujan memiliki peran krusial dalam banyak aspek kehidupan manusia dan lingkungan. Curah hujan memiliki dampak sosial dan ekonomi yang besar, termasuk kerusakan infrastruktur, banjir, dan gangguan transportasi. Banjir dan peristiwa ekstrem serupa dapat mengakibatkan kerugian jiwa dan kerusakan properti yang parah setiap tahun, dan ini dihubungkan dengan perubahan iklim yang diperkirakan akan semakin sering dan merusak di masa yang akan datang. Dalam upaya untuk memantau, mengantisipasi dan memproyeksikan terjadinya hujan, maka penelitian ini berfokus dalam memprediksi curah hujan dengan menggunakan data mining. Dalam penelitian ini, berfokus dalam penerapan metode *Categorical Boosting (CatBoost)* yang dapat mengelola data kategorikal secara efisien yang merupakan pengembangan dari *Gradient Boosting Decision Tree (GBDT)*. Berdasarkan data yang diperoleh dari kaggle yang terdiri dari 23 atribut, penelitian ini akan memprediksi curah hujan. Pra-pemrosesan data melibatkan langkah-langkah seperti pembersihan data, penyeimbangan kelas, pemilihan fitur, dan pengkodean data. Pengujian menggunakan metode *Categorical Boosting* menghasilkan tingkat akurasi prediksi curah hujan sebesar 94,22% berdasarkan evaluasi *Confusion Matrix*. Penelitian ini bertujuan untuk memprediksi curah hujan dengan menggunakan metode *Categorical Boosting (CatBoost)*.

Kata kunci : Curah Hujan, Prediksi, Categorical Boosting, Confusion Matrix

1. PENDAHULUAN

Iklim merupakan hal sangat penting dalam kehidupan kita sehari-hari[1]. Salah satu faktor yang memicu perubahan iklim adalah curah hujan, maka dari itu curah hujan memiliki peran penting[2]. Kategorisasi iklim dapat dilakukan dengan menggunakan data curah hujan dengan membandingkan jumlah rata-rata bulan kering dengan jumlah rata-rata bulan basah. Keakuratan dalam memprediksi curah hujan secara akurat sangat penting karena dapat digunakan untuk berbagai tujuan. Suhu, kelembaban udara, kecepatan angin, dan tekanan udara semuanya berdampak pada curah hujan[3].

Curah hujan merupakan salah satu parameter iklim yang paling berpengaruh dalam banyak aspek dalam kehidupan kita sehari-hari. Dampak sosial ekonomi dari fenomena curah hujan cukup signifikan, mulai dari kerusakan infrastruktur, banjir hingga gangguan transportasi. Banjir dan peristiwa ekstrem serupa yang dapat menyebabkan banyak korban jiwa dan kerusakan properti benda setiap tahunnya merupakan akibat dari perubahan iklim dan diperkirakan akan semakin sering terjadi dan bersifat menghancurkan di tahun-tahun mendatang[4] [5].

CatBoost merupakan metode yang berfokus pada karakteristik sebuah kategori. *CatBoost* juga merupakan bagian dari metode *Gradient Boosting Decision Tree (GBDT)*, yang memanfaatkan decision trees sebagai prediktor dasar[6]. *CatBoost* ini diperkenalkan oleh Prokhorenkova et al[7], dan Dorogush et al[8], yang dimana ini digunakan untuk mendapatkan kerugian terendah ketika berbicara tentang fitur kategorikal dan kumpulan data yang banyak dengan tingkat kehilangan informasi yang rendah[9]. Secara singkat, *CatBoost* dirancang khusus

untuk menangani data kategorikal dan mampu mengatasi nilai yang hilang serta volume data yang besar[10]. *CatBoost* juga mampu menangani jenis data lain seperti *time series*, teks, *medical data*, dan keuangan guna meningkatkan akurasi prediksi. Selain itu, metode ini dapat menghadapi volume data yang besar dan mendukung kemampuan pemrosesan paralel[11].

Penelitian lainnya berdasarkan Usman C, dan Sudibyo U, menerapkan prediksi terhadap curah hujan dengan metode *machine learning*, yaitu Regresi Logistik, *Random Forest*, dan *Gradient Boosting* dengan 9 variabel yang mempengaruhi curah hujan, yaitu suhu minimum, suhu maksimum, suhu rata-rata, kelembaban, lama penyinaran matahari, kecepatan angin maximum, kecepatan angin rata-rata, kecepatan arah angin dan arah angin maksimum serta memiliki tujuan, yaitu berupa kategori curah hujan. Pada penelitian tersebut berfokus terhadap menyatakan akurasi dari masing masing model, mengukur keseluruhan pengklasifikasi biner[12].

Berdasarkan penelitian terdahulu, prediksi curah hujan dilakukan dengan metode berbasis jaringan saraf tiruan seperti *K-Nearest Neighbor* dengan akurasi 83%, *Decision Tree* dengan akurasi 83%, *Random Forest* dengan akurasi 83%, dan *Neural Networks* dengan akurasi 84%, yang dimana pada penelitian tersebut menunjukan bahwa metode *Neural Networks* bekerja cukup baik untuk memodelkan prediksi curah hujan[13]. Penelitian selanjutnya yaitu prediksi curah hujan dengan menggunakan metode *AdaBoost* yang dimana merupakan metode *boosting* yang menghasilkan akurasi sebesar 82%, penelitian ini menggunakan parameter seperti iterasi sebesar 250

dengan *learning rate* 0.02 untuk mendapatkan hasil akurasi sebesar 82%[14].

Berdasarkan penjelasan di atas, penelitian ini telah menetapkan penggunaan metode *CatBoost* untuk melakukan prediksi terhadap curah hujan yang dimana pada penelitian ini akan menggunakan variabel kelembaban, sinar matahari, tekanan udara, temperatur minimum, temperatur maksimum, kecepatan angin, dan pengupaan dengan fokus penelitian yaitu bagaimana metode *CatBoost* menangani data prediksi curah hujan yang bersifat kategorikal, serta mengevaluasi nilai akurasi yang dihasilkan dari metode *CatBoost* dengan menggunakan data yang diperoleh dari kaggle.com yang berjudul "*Rain in Australia – Predict next-day in Australia*".

2. TINJAUAN PUSTAKA

Tinjauan pustaka di bawah ini bertujuan untuk menguraikan teori pendukung dan penelitian terdahulu mengenai penelitian yang dilakukan sebagai berikut.

2.1. Prediksi

Kebanyakan orang menganggap prediksi sebagai tindakan menggambarkan masa depan. Dengan mempertimbangkan pengalaman, pandangan, dan fakta lain saat melakukan peramalan, maka peramalan berbeda dengan menebak-nebak. Peramalan adalah kata yang sering digunakan untuk merujuk pada "prediksi". Meskipun banyak orang yang keliru mengira bahwa kedua istilah tersebut dapat dipertukarkan, namun ada perbedaan kecil namun krusial di antara keduanya. Sementara "peramalan" biasanya didasarkan pada data dan model, "prediksi" biasanya didasarkan pada opini dan pengalaman[15].

Karena ada begitu banyak variabel yang dapat memengaruhi prediksi, terkadang mustahil untuk mendapatkan data atau, sebaliknya, sangat sulit untuk mengumpulkan data yang akurat dengan menggunakan sensor yang sangat sensitif. Model probabilistik digunakan untuk membuat prakiraan, dan kesalahan sering terjadi karena ketidakpastian yang cukup besar dari informasi yang terkadang tersedia[13].

2.1. Machine Learning

Disiplin ilmu *machine learning* berkaitan dengan bagaimana mesin mengambil pengetahuan melalui pengalaman. Mengingat bahwa kemampuan untuk belajar adalah kualitas utama dari entitas yang disebut cerdas dalam arti yang paling luas, bagi banyak ilmuwan, istilah "*machine learning*" dan "*artificial intelligence*" dapat dipertukarkan. Membangun sistem komputer yang dapat beradaptasi dan belajar dari pengalaman mereka adalah tujuan dari *machine learning*[16]. *Machine Learning* memungkinkan pendeteksian pola tersembunyi dalam data dan memodifikasi algoritma untuk meningkatkan ketahanan pola[17].

Sekumpulan instruksi yang dikenal sebagai algoritma *machine learning* memungkinkan komputer untuk secara otomatis belajar dari masa lalu dan

menjadi lebih baik tanpa perlu pemrograman yang rumit[18].

2.2. Data Mining

Data Mining adalah proses menemukan pola yang berpotensi signifikan dalam sebuah kumpulan data yang berjumlah besar. *Machine Learning*, Statistika, dan Kecerdasan Buatan (AI) digunakan dalam proses ini, hal tersebut digunakan dalam melakukan proses pengestrakan informasi tentang kemungkinan kejadian di masa yang akan datang yang mungkin terjadi. Berbagai komponen *data mining* antara lain kategorisasi data, integrasi data, transformasi data, diskritisasi data, dan penilaian pola. Teknik data mining digunakan dalam pemasaran, penjualan, penemuan ilmiah, pengembangan produk, prediksi data untuk mengidentifikasi hubungan tersembunyi dan tak terduga di antara data[19].

2.3. Gradient Boosting

Gradient Boosting adalah metode *Machine Learning* tambahan untuk klasifikasi dan regresi. Metode ini membuat ramalan menggunakan ansambel model prediksi *Decision Tree* yang lemah. Gagasan bahwa boosting dapat dilihat sebagai metode optimasi untuk fungsi biaya tertentu pertama kali dikemukakan. Sebagai hasilnya, telah dicatat bahwa banyak algoritma boosting yang berkembang di bidang-bidang seperti statistik dan kecerdasan buatan[20].

Gradient Boosting menggunakan pengklasifikasi pembelajar yang kuat yang berkorelasi baik dengan klasifikasi sebenarnya dan dibangun dengan menggabungkan berbagai pengklasifikasi pembelajar yang lemah yang hanya berkorelasi sedikit dengan klasifikasi sebenarnya untuk membuat prediksi. Strategi ini digunakan secara berulang[21].

2.4. CatBoost

CatBoost merupakan algoritma *boosting* yang terurut, dan diimplementasikan berdasarkan algoritma peningkatan gradien dengan menggunakan *decision tree* yang digunakan untuk regresi, dan dari masing-masing *tree* sesuai dengan partisi sebuah *feature space* dan *output value*[22].

CatBoost ini menangani data kategorikal secara otomatis menggunakan metode statistik. *CatBoost* ini dapat menghindari *overfitting* data dengan mengoptimalkan berbagai parameter masukan, hal tersebut tidak berurusan dengan karakteristik kategori selama pemrosesan[23]. *CatBoost* ini akan melakukan eksekusi permutasi secara acak pada data kategori alih-alih penggantian biner dan tiap-tiap label dihitung rata-ratanya[24].

Untuk mengatasi masalah kebocoran target, pertama, dengan mengimplementasikan *ordered boosting*, sebuah versi yang efektif dari metode *gradient boosting*[8] [25]. Selanjutnya, pendekatan ini bekerja dengan baik dengan data set yang kecil. Yang terakhir, fitur kategorikal dapat ditangani melalui peningkatan kategorikal. Penanganan ini biasanya dilakukan selama tahap pra-processing dan hanya

mengganti variabel kategori asli dengan satu atau beberapa nilai numerik[9].

2.5. Evaluasi Model

Kinerja model dapat diukur dengan membandingkan hasil prediksi dengan data aktual menggunakan *confusion matrix*. *Confusion matrix* merupakan sebuah tabel yang digunakan untuk menilai performa model klasifikasi. Tabel ini membandingkan hasil prediksi model dengan nilai aktual dari data yang diamati. *Confusion matrix* terdiri dari empat komponen, yakni *True Positive (TP)*, *True Negative (TN)*, *False Positive (FP)*, dan *False Negative (FN)*. Untuk mengetahui kinerja model dapat dilakukan dengan beberapa metrik seperti *accuracy*, *precision*, *recall*, *f1-score*[26].

1. Accuracy

Accuracy adalah metrik evaluasi yang paling umum digunakan untuk klasifikasi.

$$Accuracy = \frac{(TP+TN)}{Total\ sample} \times 100 \quad (1)$$

2. Precision

Precision mengindikasikan sejauh mana hasil prediksi model sesuai dengan data yang diminta. Rumus untuk menghitung *Precision*:

$$Precision = \frac{TP}{(TP+FP)} \times 100\% \quad (2)$$

3. Recall digunakan untuk mengukur sejauh mana sistem dapat menghasilkan nilai yang benar dengan target positif. Rumus untuk menghitung Recall:

$$Recall = \frac{TP}{(TP+FN)} \times 100\% \quad (3)$$

4. F1-Score

F1-Score dapat digunakan sebagai *matrix* evaluasi yang menggabungkan kedua nilai *Precision* dan *Recall*. *F1-Score* merupakan alternatif yang digunakan untuk mengukur akurasi sistem.

$$F - 1 = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (4)$$

3. METODE PENELITIAN

Pada penelitian ini telah melakukan beberapa tahapan metode penelitian seperti pada Gambar 1.

3.1. Dataset

Data yang digunakan pada penelitian ini diperoleh dari website Kaggle yaitu dengan judul “Rain in Australia – Predict next-day in Australia” yang berisikan data sekitar 10 tahun pengamatan cuaca harian dari banyak lokasi di seluruh Australia dengan jumlah data sebanyak 145460.

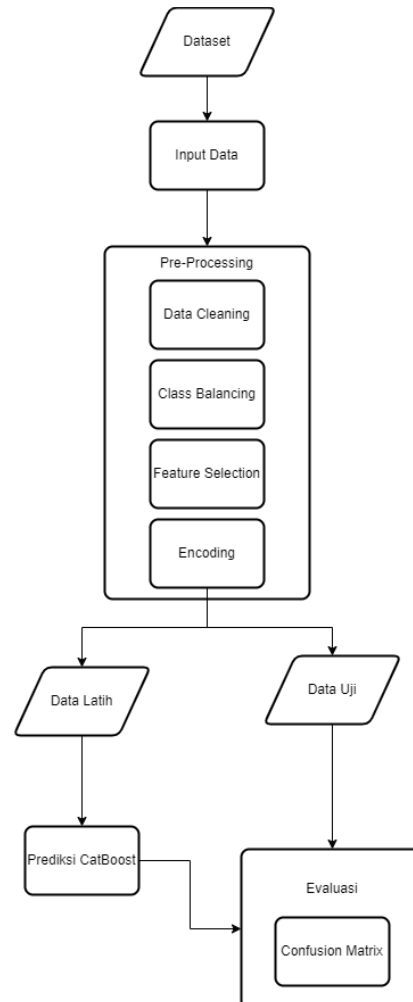
3.2. Preprocessing

Tahap ini melibatkan beberapa proses penting untuk mempersiapkan data yang nantinya digunakan dalam proses analisis dan pemodelan. Proses yang dilakukan pada penelitian ini termasuk pembersihan data, penyeimbangan kelas, seleksi fitur, dan *encoding*.

3.3. Pembagian Data

Pembagian data adalah tahap kritis dalam pembelajaran mesin yang melibatkan memisahkan

dataset Anda menjadi *subset* yang berbeda untuk melatih, mengvalidasi, dan menguji model. Tujuannya adalah untuk memastikan bahwa model yang dihasilkan dapat diuji dan dievaluasi dengan benar serta memiliki kemampuan umum yang baik untuk melakukan prediksi pada data baru.



Gambar 1. Tahapan Penelitian

3.4. Data Mining

Pusat dari seluruh proses ini adalah *data mining*. Pada tahap ini, diterapkan metode dan algoritma berdasarkan informasi yang diinginkan. Metode dan algoritma ini dapat berupa klasifikasi, prediksi, klastering, estimasi, dan lainnya. Dalam penelitian ini, algoritma yang digunakan adalah CatBoost.

3.5. Evaluasi

Evaluasi merupakan langkah terakhir dengan tujuan untuk menguraikan atau menginterpretasikan pola yang dihasilkan pada proses *data mining*. Pada tahap ini, dilakukan pemeriksaan terhadap akurasi.

4. HASIL DAN PEMBAHASAN

Penelitian ini melakukan prediksi curah hujan dengan menggunakan metode *CatBoost*. Tujuan penelitian ini untuk menghasilkan nilai akurasi pada metode prediksi yang dipakai untuk memprediksi

curah hujan. Pada Tabel 1 menunjukkan contoh data dari *dataset* yang digunakan pada penelitian ini.

Tabel 1. Contoh dataset Rain in Australia

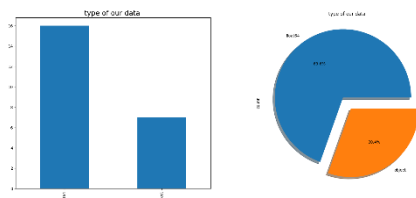
Date	Location	...	Rain Today	Rain Tomorrow
12/1/2008	Albury	...	No	No
12/2/2008	Albury	...	No	No
...	...	...	...	...
6/24/2017	Uluru	...	No	No
6/25/2017	Uluru	...	Yes	NA

Pada Tabel 2 menunjukkan atribut yang ada pada data pada *dataset* yang digunakan pada penelitian ini.

Tabel 2. Atribut data curah hujan

No	Atribut Data
1	Date
2	Location
3	MinTemp
4	MaxTemp
5	Rainfall
6	Evaporation
7	Sunshine
8	WindGustDir
9	WindGustSpeed
10	WindDir9am
11	WindDir3pm
12	WindSpeed9am
13	WindSpeed3pm
14	Humidity9pm
15	Humidity3pm
16	Pressure9am
17	Pressure3pm
18	Cloud9am
19	Cloud3pm
20	Temp9am
21	Temp3pm
22	RainToday
23	RainTomorrow

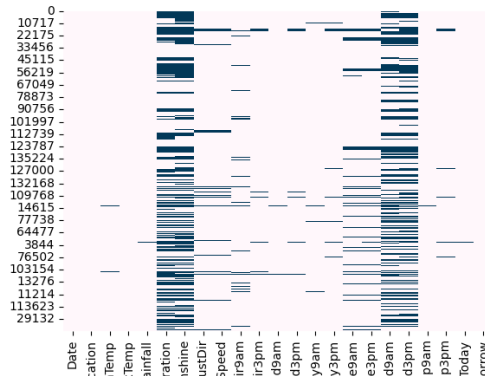
Berdasarkan *dataset* yang diperoleh dapat disimpulkan bahwa data curah hujan terdiri dari 23 atribut yang dimana atribut tersebut diantaranya terdapat 7 atribut yang bertipe kategorikal, dan 16 atribut lainnya bersifat numerik. Untuk perbandingannya dapat dilihat pada Gambar 2.



Gambar 2. Perbandingan tipe data

Tahapan selanjutnya yaitu dengan melakukan preprocessing data. Tahap ini melibatkan beberapa proses utama, termasuk pembersihan data, penyeimbangan kelas, seleksi fitur, dan *encoding*.

Tahap pertama yang dilakukan adalah dengan melakukan pembersihan data yang dimana pada tahapan ini data yang diperoleh akan dilakukan pembersihan untuk menganalisis informasi dari data yang tidak lengkap seperti data curah hujan yang tidak memiliki kelengkapan data, dengan tujuan untuk mempermudah proses pencarian dan pengolahan data.



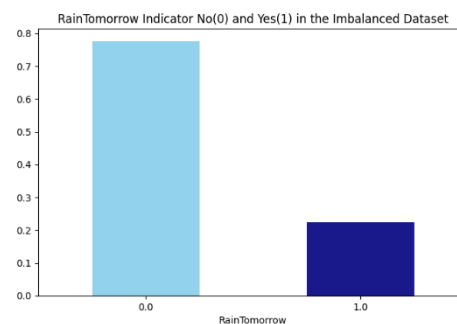
Gambar 3. Pola data yang hilang

Berdasarkan Gambar 3, dapat dinyatakan bahwa atribut “*Evaporation*”, “*Sunshine*”, “*Cloud9am*” dan “*Cloud3pm*” merupakan atribut yang memiliki persentase kehilangan yang tinggi. Maka dari hasil tersebut akan dilakukan pembersihan data dengan teknik *oversampling* terhadap data yang memiliki persentase kehilangan yang tinggi sehingga menghasilkan persentase kehilangan yang sedikit seperti pada Gambar 4.

	Total	Percent
Sunshine	104831	0.475140
Evaporation	95411	0.432444
Cloud3pm	85614	0.388040
Cloud9am	81339	0.368664

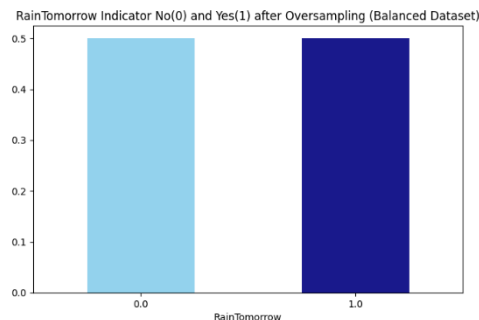
Gambar 4. Hasil pembersihan data

Tahapan berikutnya adalah penyeimbangan kelas, pada data yang diperoleh terdapat kelas yang tidak seimbang seperti pada Gambar 5.



Gambar 5. Kelas sebelum penyeimbangan

Pada Gambar 5 terdapat dua kelas utama yaitu kelas “0.0” dan kelas “1.0” yang dimana kelas tersebut memiliki ketidak seimbangan yang signifikan. Untuk mengatasi ketidakseimbangan terhadap kelas seperti pada Gambar 5 maka akan dilakukan *oversampling* kelas minoritas sehingga data yang sebelumnya tidak seimbang seperti pada Gambar 5 akan menjadi seimbang seperti pada Gambar 6.



Gambar 6. Kelas setelah penyeimbangan

Setelah dilakukan penyeimbangan kelas, tahapan berikutnya adalah dengan melakukan seleksi fitur. Pada tahapan ini data yang diperoleh akan dilakukan seleksi fitur untuk melatih model prediksi curah hujan. Sebelum melakukan seleksi fitur, harus dilakukan standarisasi data terlebih dahulu dengan menggunakan *MinMaxScaler* untuk menghindari nilai negatif. Kemudian setelah proses standarisasi data selesai dilakukan, maka akan dilakukan seleksi fitur dengan fitur yang akan digunakan dapat dilihat pada Gambar 7.

```
features = WindGustDir, Location, MinTemp, MaxTemp, Rainfall, Evaporation, Sunshine, WindGustDir,
WindGustSpeed, WindDir9am, WindDir3pm, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm,
Pressure9am, Pressure3pm, Cloud9am, Cloud3pm, Temp9am, Temp3pm,
target = RainTomorrow
```

Gambar 7. Seleksi fitur

Pada Gambar 7 dapat disimpulkan bahwa atribut yang akan digunakan sebagai fitur adalah *Location*, *MinTemp*, *MaxTemp*, *Rainfall*, *Evaporation*, *Sunshine*, *WindGustDir*, *WindGustSpeed*, *WindDir9am*, *WindDir3pm*, *WindSpeed9am*, *WindSpeed3pm*, *Humidity9am*, *Humidity3pm*, *Pressure9am*, *Pressure3pm*, *Cloud9am*, *Cloud3pm*, *Temp9am*, *Temp3pm*, dan *RainToday*. Sedangkan untuk target atau kelas yang akan digunakan adalah atribut *RainTomorrow*.

Setelah melakukan penyeimbangan kelas dilanjutkan dengan melakukan proses *encoding*. Pada tahapan ini akan melakukan *encoding* yang dimana akan mengubah nilai kategorikal menjadi nilai numerik dengan menggunakan *LabelEncoder* dari modul *sklearn.preprocessing*. pada Gambar 8, Gambar 9, Gambar 10, dan Gambar 11 dapat dilihat bahwa terdapat perubahan terhadap atribut yang bernilai kategorikal telah berubah menjadi atribut dengan nilai numerik.

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	\
0	396	2	13.4	22.9	0.6	NaN	NaN	
1	397	2	7.4	25.1	0.0	NaN	NaN	
2	398	2	12.9	25.7	0.0	NaN	NaN	
3	399	2	9.2	28.0	0.0	NaN	NaN	
4	400	2	17.5	32.3	1.0	NaN	NaN	
...	...	...	...	...	...	...	...	...
133585	950	17	1.1	11.8	0.4	NaN	NaN	
117307	3114	29	9.6	14.8	11.8	NaN	4.3	
87307	503	8	24.8	31.5	0.0	5.4	3.4	
56427	1721	5	4.2	7.9	11.4	NaN	NaN	
27220	480	34	19.5	31.6	0.2	10.4	NaN	

Gambar 8. Hasil encoding 1

	WindGustDir	WindGustSpeed	WindDir9am	...	Humidity9am	Humidity3pm	\
0	13	44.0	13	...	71.0	22.0	
1	14	44.0	6	...	44.0	25.0	
2	15	46.0	13	...	38.0	30.0	
3	4	24.0	9	...	45.0	16.0	
4	13	41.0	1	...	82.0	33.0	
...	...	...	...	...	...	...	...
133585	14	13.0	3	...	97.0	98.0	
117307	13	69.0	14	...	85.0	71.0	
87307	9	48.0	10	...	63.0	60.0	
56427	12	63.0	6	...	99.0	93.0	
27220	6	54.0	5	...	76.0	56.0	

Gambar 9. Hasil encoding 2

	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	\
0	1007.7	1007.1	8.0	NaN	16.9	21.8	
1	1010.6	1007.8	NaN	NaN	17.2	24.3	
2	1007.6	1008.7	NaN	2.0	21.0	23.2	
3	1017.6	1012.8	NaN	NaN	18.1	26.5	
4	1010.8	1006.0	7.0	8.0	17.8	29.7	
...	...	...	...	...	...	...	...
133585	1020.1	1013.3	NaN	NaN	4.0	9.5	
117307	1015.4	1015.7	8.0	8.0	11.4	12.6	
87307	1008.6	1006.2	5.0	6.0	28.7	30.3	
56427	997.2	998.9	8.0	8.0	4.6	6.8	
27220	1016.9	1011.6	NaN	NaN	22.5	31.1	

Gambar 10. Hasil encoding 3

	RainToday	RainTomorrow	\
0	0.0	0.0	
1	0.0	0.0	
2	0.0	0.0	
3	0.0	0.0	
4	0.0	0.0	
...	...	...	...
133585	0.0	1.0	
117307	1.0	1.0	
87307	0.0	1.0	
56427	1.0	1.0	
27220	0.0	1.0	

Gambar 11. Hasil encoding 4

	Location	MinTemp	...	Temp3pm	RainToday
63514	35	-0.400000	...	13.700000	0.0
17722	24	8.142953	...	18.191141	0.0
39812	45	8.900000	...	20.800000	0.0
13397	21	7.000000	...	26.100000	0.0
35653	38	19.600000	...	23.400000	0.0
...	...	...	...	...	...
70365	18	7.700000	...	14.700000	1.0
100539	22	9.700000	...	23.300000	0.0
4298	4	6.300000	...	13.400000	0.0
20166	26	16.900000	...	28.500000	1.0
140968	13	24.700000	...	30.300000	0.0

96812 rows × 21 columns

Gambar 12. Contoh data latih

Setelah dilakukan pemrosesan terhadap data tahapan selanjutnya akan melakukan pembagian data. Pada penelitian ini menggunakan skema pengujian dengan proporsi data latih sebesar 70% dan data uji sebesar 30% dengan tujuan data latih digunakan untuk

proses oleh model prediksi dan data uji digunakan untuk menguji model prediksi. Contoh data untuk data latih dapat dilihat pada Gambar 12, dan contoh data untuk data uji dapat dilihat pada Gambar 13.

	Location	MinTemp	...	Temp3pm	RainToday
119864	32	15.0	...	25.8	0.0
89309	8	24.2	...	30.5	0.0
87669	8	23.2	...	28.8	1.0
92549	14	14.4	...	19.7	1.0
121874	31	13.3	...	24.8	0.0
...	...	...	...	...	...
144044	41	6.7	...	29.6	0.0
134857	17	7.2	...	22.6	0.0
13575	21	6.3	...	18.3	0.0
67312	18	11.1	...	15.9	0.0
85199	7	18.6	...	24.5	0.0

41492 rows x 21 columns

Gambar 13. Contoh data uji

Tahapan selanjutnya dengan proses *data mining*. Penelitian ini akan melakukan prediksi dengan menggunakan metode *CatBoost*. Diawali dengan pembuatan model *CatBoost*. Dalam *CatBoost* memerlukan beberapa parameter pengukuran yang telah ditentukan setelah melakukan beberapa kali eksperimen sehingga dapat diperoleh hasil prediksi yang baik. Tahapan ini juga akan menggunakan parameter pada model yang telah dibuat. Terdapat beberapa parameter yang akan digunakan pada model *CatBoost* ini, yang dapat dilihat pada Tabel 3.

Tabel 3. Parameter catboost

Parameter	Kegunaan
<i>iteration</i>	Parameter <i>iteration</i> mengacu pada jumlah iterasi atau langkah <i>boosting</i> yang akan dilakukan oleh <i>CatBoost</i> .
<i>max_depth</i>	Parameter <i>max_depth</i> mengontrol kedalaman maksimum dari pohon dalam ansambel. Kedalaman pohon mempengaruhi kompleksitas model.

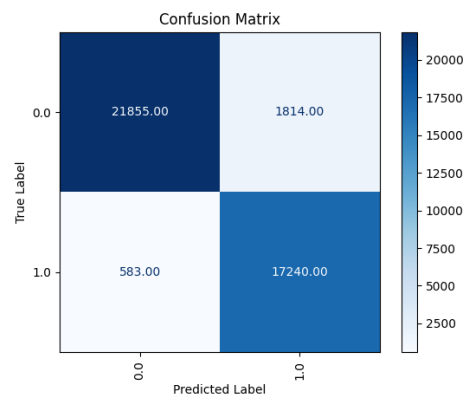
Setelah melakukan beberapa tahapan sebelumnya, maka akan dilakukan beberapa percobaan untuk menemukan hasil prediksi terbaik berdasarkan data curah hujan dengan melakukan percobaan penyetelan parameter. Hasil eksperimen dengan melakukan penyetelan parameter ini dapat meningkatkan kinerja model *CatBoost*, dengan hasil yang dapat dilihat pada Tabel 4.

Tabel 4. Hasil eksperimen parameter

Parameter	Percobaan 1	Percobaan 2	Percobaan 3
<i>iteration</i>	300	100	50
<i>max_depth</i>	8	8	16
Akurasi	89,7%	87,1%	94,2%

Pada Tabel 4 dapat disimpulkan bahwa terdapat akurasi yang tidak jauh berbeda, tiap eksperimen yang dihasilkan telah didasarkan pada penelitian sebelumnya yang telah melakukan hal tersebut.

Untuk hasil eksperimen kedua yaitu berfokus kepada penerapan model *CatBoost* dengan melihat akurasi yang diperoleh oleh model tersebut. Untuk menguji performa model, pada penelitian ini menggunakan evaluasi model dengan *confusion matrix*. Perhitungan dengan menggunakan *confusion matrix* digunakan untuk menentukan persentase data yang sesuai dengan kenyataan dan dibandingkan dengan data yang tidak sesuai atau tidak ada. Untuk hasil *confusion matrix* dapat dilihat pada Gambar 12.



Gambar 14. Confusion matrix

Pada gambar 14 terdapat beberapa nilai yang dapat dihitung untuk perhitungan akurasi model. Hasil dari *confusion matrix* tersebut dapat disimpulkan bahwa terdapat nilai TP sebesar 21855.00, FP sebesar 1814.00, FN sebesar 583.00, dan TN sebesar 17240.00 yang dapat dihitung untuk mendapatkan nilai *accuracy*, *precision*, *recall*, *f1-score* dengan hasil seperti pada Tabel 5.

Tabel 5. Matriks evaluasi

Matriks Evaluasi	Skor
<i>Accuracy</i>	94,22%
<i>Precision</i>	92,33%
<i>Recall</i>	97,40%
<i>F1-Score</i>	94,80%

5. KESIMPULAN DAN SARAN

Dalam penelitian ini, menggunakan metode *CatBoost*, data yang digunakan terdiri dari 145460 data yang dibagi menjadi 70% data latih dan 30% data uji, kemudian data tersebut mengimplementasikan beberapa tahap *pre-processing* yaitu *data cleaning*, *class balancing*, *feature selection* dan *encoding*. Dari hasil akurasi yang diperoleh dari evaluasi menggunakan *confusion matrix* cukup tinggi dalam memprediksi curah hujan yaitu sebesar 94.2% berdasarkan atribut yang digunakan pada data. Kesimpulan dari penelitian ini adalah bahwa metode *CatBoost* adalah sebuah metode yang dapat menangani data yang bersifat kategorikal dengan jumlah data yang banyak sebanyak 145460 data. Serta untuk

penelitian selanjutnya disarankan untuk mempertimbangkan penggunaan metode yang lebih beragam, terutama metode yang dapat menangani data yang bersifat kategorikal seperti data yang digunakan pada penelitian ini, dan juga jumlah data yang lebih banyak untuk meningkatkan tingkat akurasi yang lebih tinggi.

DAFTAR PUSTAKA

- [1] U. Shah, S. Garg, N. Sisodiya, N. Dube, and S. Sharma, "Rainfall prediction: Accuracy enhancement using machine learning and forecasting techniques," *PDGC 2018 - 2018 5th Int. Conf. Parallel, Distrib. Grid Comput.*, pp. 776–782, Dec. 2018, doi: 10.1109/PDGC.2018.8745763.
- [2] N. Nyaupane, S. R. Mote, M. Bhandari, A. Kalra, and S. Ahmad, "Rainfall-Runoff Simulation Using Climate Change Based Precipitation Prediction in HEC-HMS Model for Irwin Creek, Charlotte, North Carolina," *World Environ. Water Resour. Congr. 2018 Watershed Manag. Irrig. Drainage, Water Resour. Plan. Manag. - Sel. Pap. from World Environ. Water Resour. Congr. 2018*, pp. 352–363, 2018, doi: 10.1061/9780784481400.033.
- [3] N. Ritha, M. Bettiza, and A. Dufan, "Prediksi Curah Hujan dengan Menggunakan Algoritma Levenberg Marquardt dan Backpropagation," Oct. 2016, Accessed: Dec. 01, 2022. [Online]. Available: <https://media.neliti.com/media/publications/266810-prediksi-curah-hujan-dengan-menggunakan-d6a27956.pdf>
- [4] A. Y. Barrera-Animas, L. O. Oyedele, M. Bilal, T. D. Akinosho, J. M. D. Delgado, and L. A. Akanbi, "Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting," *Mach. Learn. with Appl.*, vol. 7, p. 100204, Mar. 2022, doi: 10.1016/J.MLWA.2021.100204.
- [5] J. Sun, Y. Zhang, J. Ban, J. S. Hong, and C. Y. Lin, "Impact of Combined Assimilation of Radar and Rainfall Data on Short-Term Heavy Rainfall Prediction: A Case Study," *Mon. Weather Rev.*, vol. 148, no. 5, pp. 2211–2232, May 2020, doi: 10.1175/MWR-D-19-0337.1.
- [6] A. A. Ibrahim, R. L. Ridwan, M. M. Muhammed, R. O. Abdulaziz, and G. A. Saheed, "Comparison of the CatBoost Classifier with other Machine Learning Methods," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 11, pp. 738–748, 2020, doi: 10.14569/IJACSA.2020.0111190.
- [7] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *Adv. Neural Inf. Process. Syst.*, vol. 2018-December, pp. 6638–6648, Jun. 2017, doi: 10.48550/arxiv.1706.09516.
- [8] A. V. Dorogush, V. Ershov, and A. G. Yandex, "CatBoost: gradient boosting with categorical features support," Oct. 2018, doi: 10.48550/arxiv.1810.11363.
- [9] S. Ben Jabeur, C. Gharib, S. Mefteh-Wali, and W. Ben Arfi, "CatBoost model and artificial intelligence techniques for corporate failure prediction," *Technol. Forecast. Soc. Change*, vol. 166, p. 120658, May 2021, doi: 10.1016/J.TECHFORE.2021.120658.
- [10] T. Wu, J. Li, W. Zhang, W. Guo, and X. Jiao, "A Novel Reference Evapotranspiration Forecasting Model Based on Categorical Feature Encoding Methods," *Yingyong Jichu yu Gongcheng Kexue Xuebao/Journal Basic Sci. Eng.*, vol. 30, no. 6, pp. 1402–1419, Dec. 2022, doi: 10.16058/J.ISSN.1005-0930.2022.06.007.
- [11] S. Lee, T. P. Vo, H. T. Thai, J. Lee, and V. Patel, "Strength prediction of concrete-filled steel tubular columns using Categorical Gradient Boosting algorithm," *Eng. Struct.*, vol. 238, Jul. 2021, doi: 10.1016/j.engstruct.2021.112109.
- [12] C. D. Usman and U. Sudibyo, "Klasifikasi Curah Hujan di Kota Semarang Menggunakan Machine Learning," *Pros. Sains dan Teknol.*, vol. 1, no. 1, pp. 1–5, Aug. 2022, Accessed: Dec. 07, 2022. [Online]. Available: <https://www.jurnal.pelitabangsa.ac.id/index.php/SAINTEK/article/view/1091>
- [13] A. Sarasa-Cabezuelo, "Prediction of Rainfall in Australia Using Machine Learning," *Inf. 2022, Vol. 13, Page 163*, vol. 13, no. 4, p. 163, Mar. 2022, doi: 10.3390/INFO13040163.
- [14] Y. Zhao, H. Shi, Y. Ma, M. He, H. Deng, and Z. Tong, "Rain Prediction Based on Machine Learning," 2022.
- [15] "Cara Kerja Data Mining - Seri Data Mining for Business Intelligence (3) | BINUS UNIVERSITY MALANG | Pilihan Universitas Terbaik di Malang," <https://binus.ac.id/malang/2019/01/cara-kerja-data-mining-seri-data-mining-for-business-intelligence-3/> (accessed Feb. 13, 2023).
- [16] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, Jan. 2017, doi: 10.1016/J.CSBJ.2016.12.005.
- [17] A. Yosipof, R. C. Guedes, and A. T. García-Sosa, "Data mining and machine learning models for predicting drug likeness and their disease or organ category," *Front. Chem.*, vol. 6, no. MAY, May 2018, doi: 10.3389/fchem.2018.00162.
- [18] N. Khan, D. A. Sachindra, S. Shahid, K. Ahmed, M. S. Shiru, and N. Nawaz, "Prediction of droughts over Pakistan using machine learning algorithms," *Adv. Water Resour.*, vol. 139, p. 103562, May 2020, doi: 10.1016/J.ADVWATRES.2020.103562.
- [19] D. Papakyriakou and I. S. Barbounakis, "Data

- Mining Methods: A Review,” *Int. J. Comput. Appl.*, vol. 183, no. 48, pp. 5–19, Jan. 2022, doi: 10.5120/ijca2022921884.
- [20] M. R. Machado, S. Karray, and I. T. De Sousa, “LightGBM: An effective decision tree gradient boosting method to predict customer loyalty in the finance industry,” *14th Int. Conf. Comput. Sci. Educ. ICCSE 2019*, pp. 1111–1116, Aug. 2019, doi: 10.1109/ICCSE.2019.8845529.
- [21] C. Li, “A Gentle Introduction to Gradient Boosting”, Accessed: Feb. 13, 2023. [Online]. Available: <https://github.com/cheng-li/pyramid>
- [22] P. Kang, Z. Lin, S. Teng, G. Zhang, L. Guo, and W. Zhang, “Catboost-based framework with additional user information for social media popularity prediction,” *MM 2019 - Proc. 27th ACM Int. Conf. Multimed.*, pp. 2677–2681, Oct. 2019, doi: 10.1145/3343031.3356060.
- [23] M. Saber *et al.*, “Examining LightGBM and CatBoost models for wadi flash flood susceptibility prediction,” <https://doi.org/10.1080/10106049.2021.1974959>, 2021, doi: 10.1080/10106049.2021.1974959.
- [24] P. S. Kumar, K. Anisha Kumari, S. Mohapatra, B. Naik, J. Nayak, and M. Mishra, “CatBoost ensemble approach for diabetes risk prediction at early stages,” *1st Odisha Int. Conf. Electr. Power Eng. Commun. Comput. Technol. ODICON 2021*, Jan. 2021, doi: 10.1109/ODICON50556.2021.9428943.
- [25] H. Wang and L. Cheng, “CATBOOST MODEL WITH SYNTHETIC FEATURES IN APPLICATION TO LOAN RISK ASSESSMENT OF SMALL BUSINESSES A PREPRINT,” 2021.
- [26] G. S. P. Ghantasala, N. V. Kumari, and R. Patan, *Cancer prediction and diagnosis hinged on HCML in IOMT environment*. Elsevier, 2021. doi: 10.1016/B978-0-12-821229-5.00004-5.