BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

2.1.1 A LightGBM based Forecasting of Dominant Wave Periods in Oceanic Waters

Artikel ini membandingkan LigthGBM dan *Extra Trees* (ET) untuk memprediksi kemunculan ombak, keduanya sama sama model kecerdasan buatan *ensemble* berbasis *decision tree* [6]. Yang membedakan adalah algoritma LightGBM yang berupa *boosting* dan algoritma ET yang berupa bagging. Ditemukan bahwa LightGBM memiliki performa yang lebih baik dengan RMSE 1.78 dibanding ET dengan RMSE sebesar 2.61.

Poin utama yang bisa di ambil dari artikel ini, yaitu:

Menarik bahwa jumlah prediksi tidak menentukan penilaian metrik akan menurun, dari artikel ini prediksi 30 hari memiliki RMSE lebih kecil dibandingkan dengan prediksi 1 hari.

2.1.2 Visibility Forecasting Using Autoregressive Integrated Moving Average (ARIMA) Models

Penelitian ini mencoba model ARIMA untuk melakukan prediksi cuaca, khususnya untuk permasalahan visibilitas atau jarak penglihatan yang dipengaruhi oleh kelembapan relatif dan temperatur [7]. Dijelaskan juga bahwa untuk menggunakan model ARIMA penting melakukan uji stasiuneritas seperti menggunakan *Auto Correlation Plot*. Selain itu, model ARIMA juga tidak begitu baik jika digunakan pada data yang memiliki efek musim, misalnya penjualan es krim yang meningkat pada musim panas.

Poin utama yang bisa di ambil dari artikel ini, yaitu:

Penting melakukan uji stasiuneritas untuk mengetahui ordo dari variabel model ARIMA.

2.1.3 Prediction of Outdoor Air Temperature and Humidity Using XGBoost

Penelitian menggunakan XGBoost untuk melakukan prediksi kondisi temperatur dan kelembapan udara [8], ada 2 poin yang bisa diambil dari artikel ini:

XGBoost memiliki performa yang baik meskipun dengan data yang sedikit, dan bisa beradaptasi sehingga performanya lebih baik dalam dataset yang besar.

Ketika melakuakan prediksi, semakin banyak data yang diprediksi akan semakin jelek performanya. Contohnya, RMSE untuk memprediksi temperatur 1 jam sebesar 3.89 dan RMSE untuk memprediksi 3 jam sebesar 6.33.

2.2 Tinjauan Teori

2.2.1 Time Series Data

Time series merupakan sebuah metode statistik dengan fokus mengolah data yang berkaitan dengan waktu [9]. Tipe data *time series* ini berbeda dengan data lainnya, karena masing-masing titik waktu memiliki datanya sendiri. Maka dari itu, kondisi pada waktu tertentu bisa dipengaruhi oleh satu sama lain, contohnya kondisi cuaca hari ini yang dipengaruhi kondisi cuaca pada satu minggu sebelumnya. Sehingga akan terdapat beberapa komponen yang terdapat pada data *time-series*, seperti pola musiman dan *trend* data.

2.2.2 *Preprocessing* Data

Agar data yang digunakan lebih layak untuk proses pelatihan model kecerdasan buatan, penting dilakukan *preprocessing* untuk mempersiapkan data. Ada beberapa langkah yang bisa dilakukan untuk mengubah data mentah menjadi data yang siap digunakan dalam pelatihan model kecerdasan buatan. *Preprocessing* yang dilakukan adalah pendeteksian dan pembersihan *data*

outlier [10], ada beberapa data yang perlu dideteksi, seperti data *null*, data yang terlalu kecil atau terlalu besar. Sederhananya, data dengan nilai yang jauh berbeda dari nilai rata-rata perlu diubah agar model kecerdasan buatan lebih representatif. Selain itu juga dilakukan *resample*, yaitu proses mengubah *timestamp* atau tanggal ke frekuensi tertentu (hari, jam, menit, detik).

2.2.3 AutoRegressive Integrated Moving Average

AutoRegressive Integrated Moving Average (ARIMA) merupakan sebuah model *time-series* yang mengabaikan variabel independen [11], dan hanya menggunakan variabel dependen dalam melakukan prediksi. Model ARIMA terdiri dari tiga bagian biasanya dituliskan dengan ARIMA (p, d, q), yaitu [12][13]:

1. AR: *autoregression*. Sebuah model regresi yang menggunakan data sebelumnya untuk melakukan prediksi (p).
2. I: *integration*. Melakukan kalkulasi fungsi diferensiasi dengan tujuan membuat data stasioner (d).
3. MA: *moving average*. Model ini menggunakan *error* dari waktu sebelumnya untuk melakukan prediksi (q).

Fungsi autokorelasi (ACF) dapat digunakan untuk membantu menentukan ordo dari algoritma MA (q), ordo disini berarti besaran nilai parameter, autokorelasi merupakan fungsi yang menghitung korelasi secara tidak langsung dari data pada waktu T dan semua data sebelumnya [14]. Misalnya dalam menghitung korelasi data T dan T-2 dihitung juga pengaruh antara T dan T-1 serta T-1 dan T-2. Sedangkan fungsi autokorelasi parsial (PACF) dapat digunakan untuk menghitung korelasi langsung dari data pada waktu T dan T-i [14]. Sehingga, untuk menghitung korelasi antara data T dan T-2, pengaruh data T-1 dianggap tidak ada sehingga langsung menghitung korelasi antara T dan T-2.

Selain itu, pada ARIMA juga terdapat parameter P, D, Q dan s. Parameter ini berfungsi untuk memberikan informasi bahwa data yang digunakan untuk pelatihan memiliki pola musiman. P, D dan Q berfungsi sama dengan p, d dan q, yang membedakan adalah interval s. Parameter s merupakan informasi dalam berapa lama data mengalami perulangan musiman, contohnya jika s = 24 berarti setiap 24 data akan mengalami perulangan. Berarti P akan memperhitungkan bahwa nilai T akan dipengaruhi oleh T-24. Sama halnya D dan Q akan menggunakan nilai pada T-24 untuk melakukan prediksi.

Stasioneritas data menjadi salah satu aspek yang perlu diperhatikan dalam pengembangan model ARIMA, stasioneritas berarti tidak terdapat perubahan yang drastis pada data, berfluktuasi disekitar nilai rata-rata dan konstan, serta tidak memiliki elemen *trend* dan *seasonality* [14]. *Augmented Dickey-Fuller Test* (*ADF Test*) dapat digunakan untuk menguji apakah data stasioner, data stasioner memiliki karakteristik yang berbeda [15] dan secara spesifik menentukan ordo d pada model ARIMA dengan menguji hipotesisnya sebagai berikut:

1. p-value ADF ≤ 0.05 ; data stasioner
2. p-value KPSS > 0.05 ; data stasioner
3. $|$ *critical value* $| <$ | statistik ADF | ; data stasioner

Namun akurasi prediksi yang diberikan oleh model ARIMA hanya akurat untuk prediksi jangka pendek [11]. Bahwa algoritma ARIMA sangat dipengaruhi oleh jumlah *steps* (prediksi), dalam prediksi jangka panjang hingga 3-step menunjukan hasil metrik pengujian yang meningkat cukup signifikan [16].

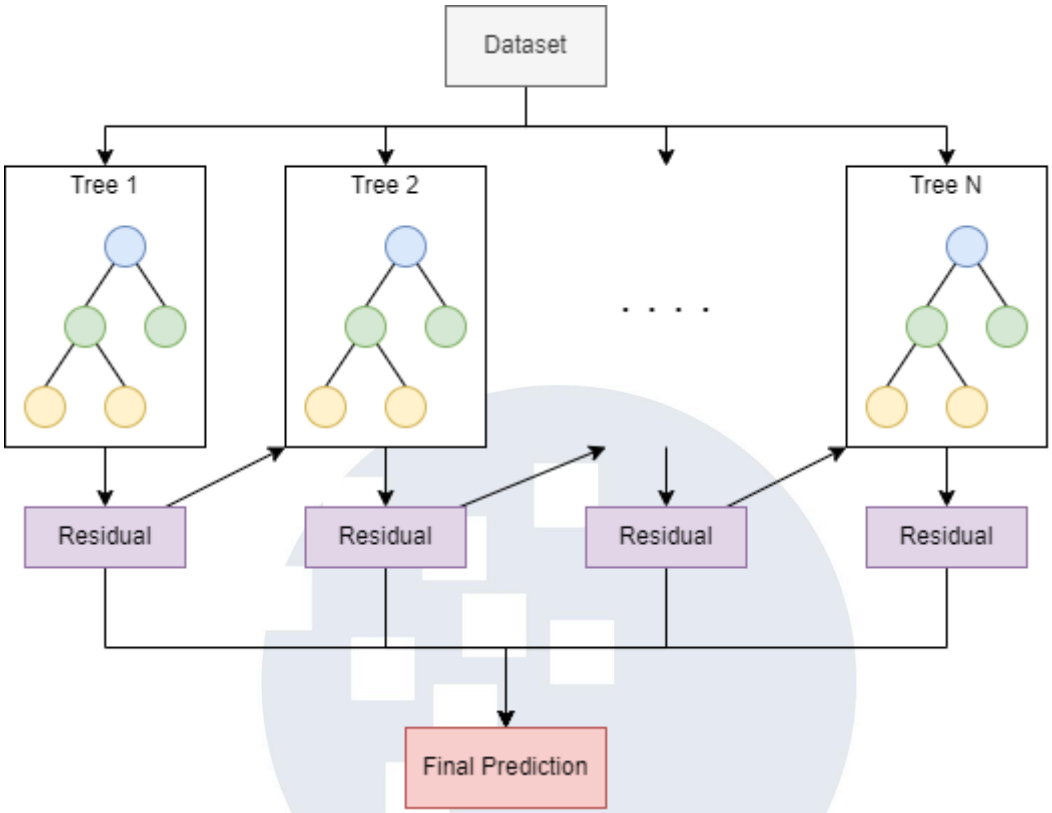
2.2.4 Gradient Boosting Machines

Ensemble merupakan sebuah teknik dalam pengembangan model kecerdasan buatan dengan menggabungkan beberapa model untuk

meningkatkan performa prediksi dari model [17]. Salah satu metode ensemble yaitu boosting, khususnya *Gradient Boosting Machines* (GBM). GBM merupakan model regresi non-parametrik [18], artinya model tidak membuat asumsi mengenai fungsi f(·) melainkan melakukan estimasi f(·) dari dataset melalui proses *trial and error method* [19].

XGBoost dan LightGBM keduanya merupakan model GBM berbasis *decision tree* [20, 21] atau dikenal juga sebagai *Gradient Boosting Decision Tree* (GBDT). Konsep boosting pada GBDT mengimplikasikan proses pelatihan yang sekuensial, masing-masing pohon akan melakukan pelatihan dengan melihat residu atau selisih hasil prediksi dari pohon sebelumnya, dapat dilihat pada Gambar 2.4 bahwa semakin banyak pohon berarti performa model akan semakin efektif. Prediksi akhir akan dilakukan dengan menghitung rata-rata dari tiap prediksi pohon.

Ada beberapa parameter yang digunakan dalam pelatihan yaitu *objective* yang menentukan tujuan prediksi dari model dan akan digunakan untuk mengatur *loss function* dengan tujuan untuk mengoptimalkan performa model, *n_estimators* atau jumlah iterasi yang akan dilakukan model, *learning_rate* yaitu parameter yang mengatur kecepatan konvergensi model menuju performa terbaik dan *booster* yang mengatur algoritma *ensemble* yang digunakan.



Gambar 2.1 *Flowchart* GBDT [22]

2.2.5 Prevalensi Kemunculan Hama

Tabel 2.1 Temperatur prevalensi hama [2]

Lalat Buah		Kutu Putih		Kumbang Girang		Tikus		Bajing	
°C	#	°C	#	°C	#	°C	#	°C	#
15	55	15	5	15	0	15	23	15	23
16	60	16	6	16	10	16	25	16	25
17	65	17	7	17	20	17	26	17	26
18	70	18	8	18	30	18	27	18	27
19	75	19	9	19	40	19	28	19	28
20	80	20	10	20	50	20	29	20	29
21	100	21	11	21	60	21	30	21	30
22	140	22	12	22	70	22	29	22	29
23	180	23	13	23	80	23	28	23	28
24	200	24	14	24	110	24	27	24	27

25	250	25	15	25	140	25	26	25	26
26	250	26	16	26	170	26	25	26	25
27	250	27	20	27	200	27	23	27	23
28	200	28	28	28	200	28	21	28	21
29	180	29	36	29	200	29	19	29	19
30	140	30	40	30	170	30	17	30	17
31	100	31	50	31	140	31	15	31	15
32	80	32	50	32	110	32	13	32	13
33	75	33	50	33	80	33	11	33	11
34	70	34	40	34	70	34	9	34	9
35	65	35	36	35	60	35	7	35	7
36	60	36	28	36	50	36	5	36	5
37	55	37	20	37	40	37	3	37	3
38	50	38	16	38	30	38	1	38	1
39	45	39	15	39	20	39	0	39	0

Tabel 2.2 Kelembapan relatif prevalensi hama [2]

Lalat Buah		**Kutu Putih**		**Kumbang Girang**		**Tikus**		**Bajing**	
RH%	**#**	**RH%**	**#**	**RH%**	**#**	**RH%**	**#**	**RH%**	**#**
50%	20	40%	7	65%	0	40%	5	40%	5
51%	25	41%	8	66%	0	41%	7	41%	7
52%	30	42%	9	67%	0	42%	9	42%	9
53%	35	43%	10	68%	10	43%	11	43%	11
54%	40	44%	11	69%	20	44%	13	44%	13
55%	45	45%	12	70%	30	45%	15	45%	15
56%	50	46%	13	71%	40	46%	17	46%	17
57%	55	47%	14	72%	50	47%	19	47%	19
58%	60	48%	15	73%	60	48%	21	48%	21
59%	65	49%	16	74%	70	49%	23	49%	23
60%	70	50%	20	75%	80	50%	25	50%	25
61%	75	51%	28	76%	110	51%	26	51%	26
62%	80	52%	36	77%	140	52%	27	52%	27
63%	100	53%	40	78%	170	53%	28	53%	28
64%	140	54%	50	79%	200	54%	29	54%	29
65%	180	55%	50	80%	200	55%	30	55%	30
66%	200	56%	50	81%	200	56%	29	56%	29
67%	250	57%	40	82%	170	57%	28	57%	28
68%	250	58%	36	83%	140	58%	27	58%	27

69%	200	59%	28	84%	110	59%	26	59%	26
70%	180	60%	20	85%	80	60%	25	60%	25
71%	140	61%	16	86%	70	61%	23	61%	23
72%	100	62%	15	87%	60	62%	21	62%	21
73%	80	63%	14	88%	50	63%	19	63%	19
74%	75	64%	13	89%	40	64%	17	64%	17
75%	70	65%	12	90%	30	65%	15	65%	15
76%	65	66%	11	91%	20	66%	13	66%	13
77%	60	67%	10	92%	10	67%	11	67%	11
78%	55	68%	9	93%	0	68%	9	68%	9
79%	50	69%	8	94%	0	69%	7	69%	7
80%	45	70%	7	95%	0	70%	5	70%	5

Tabel 2.3 Intensitas cahaya prevalensi hama [2]

Lalat Buah		**Kutu Putih**		**Kumbang Girang**		**Tikus**		**Bajing**	
Lux	**#**	**Lux**	**#**	**Lux**	**#**	**Lux**	**#**	**Lux**	**#**
0	0	0	0	0	0	0	0	0	0
100	20	100	15	100	15	100	0	100	0
200	40	200	16	200	20	200	0	200	0
300	60	300	20	300	25	300	0	300	0
400	80	400	28	400	30	400	0	400	0
500	100	500	36	500	35	500	0	500	0
600	120	600	40	600	40	600	0	600	0
700	140	700	50	700	45	700	0	700	0
800	160	800	50	800	50	800	0	800	0
900	180	900	50	900	55	900	0	900	0
1000	200	1000	40	1000	60	1000	0	1000	0
1100	250	1100	36	1100	65	1100	0	1100	0
1200	250	1200	28	1200	70	1200	0	1200	0
1300	250	1300	20	1300	70	1300	0	1300	0
1400	200	1400	16	1400	70	1400	0	1400	0
1500	180	1500	15	1500	65	1500	0	1500	0
1600	160	1600	14	1600	60	1600	0	1600	0
1700	140	1700	13	1700	55	1700	0	1700	0
1800	120	1800	12	1800	50	1800	0	1800	0
1900	100	1900	11	1900	45	1900	0	1900	0
2000	80	2000	10	2000	40	2000	0	2000	0
2100	60	2100	9	2100	35	2100	0	2100	0

2200	40	2200	8	2200	30	2200	0	2200	0
2300	20	2300	7	2300	25	2300	0	2300	0

Tabel 2.4 Curah hujan prevalensi hama [2]

Lalat Buah		**Kutu Putih**		**Kumbang Girang**		**Tikus**		**Bajing**	
RR	**#**	**RR**	**#**	**RR**	**#**	**RR**	**#**	**RR**	**#**
0	0	0	50	0	50	0	30	0	25
25	0	25	45	25	45	25	25	25	20
50	50	50	40	50	40	50	20	50	20
75	100	75	35	75	35	75	20	75	20
100	150	100	30	100	30	100	20	100	15
125	200	125	25	125	25	125	20	125	15
150	250	150	20	150	20	150	15	150	15
175	200	175	15	175	15	175	10	175	10
200	150	200	10	200	10	200	5	200	5
225	100	225	5	225	5	225	0	225	0
250	50	250	0	250	0	250	0	250	0
275	0	275	0	275	0	275	0	275	0
300	0	300	0	300	0	300	0	300	0

Tabel 2.1 menunjukan prevalensi kisaran temperatur terhadap populasi hama, tabel ini menjadi acuan dalam melihat pengaruh temperatur terhadap kemunculan hama, nilai yang diwarnai jingga menandakan jumlah (#) populasi hama yang tinggi. Tabel 2.2 menunjukan prevalensi kisaran kelembapan relatif terhadap populasi hama, tabel ini menjadi acuan dalam melihat pengaruh kelembapan terhadap kemunculan hama, nilai yang diwarnai jingga menandakan jumlah (#) populasi hama yang tinggi. Tabel 2.3 menunjukan prevalensi kisaran intensitas cahaya terhadap populasi hama, tabel ini menjadi acuan dalam melihat pengaruh intensitas terhadap kemunculan hama, nilai yang diwarnai jingga menandakan jumlah (#) populasi hama yang tinggi. Tabel 2.4 menunjukan prevalensi kisaran curah hujan terhadap populasi hama, tabel ini menjadi acuan dalam melihat pengaruh curah hujan terhadap kemunculan

hama, nilai yang diwarnai jingga menandakan jumlah (#) populasi hama yang tinggi. Data tabel diatas sangat penting sebagai indikasi kemunculan hama untuk hari kedepannya dilihat melalui hasil prediksi cuaca oleh model kecerdasan buatan, sehingga petani bisa lebih dulu mempersiapkan langkah preventif untuk menekan populasi hama.