Perbandingan Model Regresi Untuk Memprediksi Harga Jual Cabai Rawit Berdasarkan Iklim Harian

Miko Ardian*[1], Siti Khomsah[2], Ridwan Pandiya[3]
[1,2]Program Studi Sains Data; Institut Teknologi Telkom Purwokerto,
[3]Program Studi Teknik Informatika; Institut Teknologi Telkom Purwokerto;
Jl.DI Panjaitan No.128, Kec. Purwokerto Selatan, Kab. Banyumas, JawaTengah.
e-mail: ***[1]20110011@ittelkom-pwt.ac.id**, [2]siti@ittelkom-pwt.ac.id,
[3]ridwanpandiya@ittelkom-pwt.ac.id

Abstrak

Cabai memberikan sumbangan inflasi sebesar 0,15% pada Juli 2022. Inflasi cabai disebabkan oleh kenaikan harga, kenaikan tersebut diakibatkan fluktuasi harga jual. Fluktuasi harga cabai rawit disebabkan beberapa faktor, seperti kondisi iklim. Iklim ekstrem mengakibatkan penurunan produksi sehingga menyebabkan perubahan harga. Perlu dilakukan prediksi untuk memperkirakan harga jual cabai rawit berdasarkan kondisi iklim harian yang terdiri dari variabel suhu, kelembaban, curah hujan, waktu pemaparan, dan kecepatan angin. Teknik regresi umumnya digunakan untuk memprediksi masa depan. Pemilihan algoritma regresi harus disesuaikan dengan karakteristik data dan uji asumsi klasik. Data yang digunakan dalam penelitian ini merupakan data yang tidak memiliki distribusi normal dan linearitas. Pada penelitian terdahulu, algoritma XGBoost Regression, KNN Regression, dan Random Forest Regression digunakan untuk menangani data dengan karakteristik tersebut. Evaluasi ketiga algoritma menghasilkan bahwa XGBoost Regression menjadi model terbaik dengan nilai MAE terkecil yaitu 3388, nilai MAPE terkecil yaitu 9,96% yang masuk dalam kategori sangat baik, dan R2-Score terbesar yaitu 0,91. Dengan menggunakan metode SHAP diketahui bahwa suhu merupakan variabel yang mempunyai kontribusi paling signifikan dengan rata-rata nilai SHAP sebesar +7003,8 yang menunjukkan bahwa variabel tersebut mempunyai pengaruh positif terhadap prediksi harga jual. Variabel lain yaitu kecepatan angin dan waktu pemaparan juga mempunyai kontribusi yang besar terhadap harga jual.

***Kata kunci**: Cabai Rawit, Iklim, Harga Jual, Regresi, Prediksi, Perbandingan*

Abstract

Chili contributes to inflation of 0.15% in July 2022. Inflation is caused by an increase in selling prices, this increase is caused by fluctuations in selling prices. Several factors, such as climate conditions cause fluctuations in the price of cayenne pepper. Predictions need to be made to estimate the selling price based on daily climate conditions. Regression techniques are generally used to predict the future. The data used in this research is data that doesn't have a normal distribution and does not have linearity. The XGBoost Regression, KNN Regression, and Random Forest Regression algorithms handle data with these characteristics. Evaluation of the three algorithms resulted in XGBoost Regression being the best model with the smallest MAE=3388, the smallest MAPE=9,96%, which is in the very good category, and the largest R2-Score=0,91. Using the SHAP method, temperature is the variable with the most significant contribution with an average SHAP value of +7003,8 which shows that this variable positively influences selling price predictions.

***Keywords:** Cayenne Pepper, Climate, Selling Price, Regression, Prediction, Comparasion*

1. PENDAHULUAN

Cabai rawit adalah jenis tanaman hortikultura yang populer di Indonesia karena digunakan sebagai salah satu bumbu masakan [1]. Cabai sebagai bagian dari *volatile foods* menyumbang inflasi sebesar 0,15% pada Juli 2022 [2]. Inflasi cabai disebabkan karena naiknya harga jual, kenaikan tersebut diakibatkan oleh harga jual yang mengalami fluktuasi [3]. Melalui platform Hargapangan.com yang merupakan pusat informasi fluktuasi harga pangan nasional mencatat pada salah satu daerah yaitu kota Semarang, harga cabai rawit sering naik dan turun setiap bulannya. Pada Maret 2021, terjadi kenaikan harga tertinggi yaitu Rp. 110.000/kg, hal tersebut disebabkan oleh faktor cuaca dan peningkatan kebutuhan industri kuliner [4].

Iklim yang ekstrim seperti kekeringan atau banjir dapat berdampak buruk pada lahan, menghambat proses penanaman dan berdampak buruk pada produksi tanaman sehingga tidak mencukupi jumlah permintaan pasar [1]. Pada 2022, BBC Indonesia melaporkan bahwa hasil panen cabai yang semula 200 kg menurun menjadi 8-10kg pada musim hujan [2]. Penelitian [5] menyimpulkan bahwa perlu dilakukan prediksi untuk memperkirakan harga jual cabai rawit berdasarkan keadaan iklim yang terdiri dari variabel suhu, kelembapan, curah hujan, lama penyinaran dan kecepatan angin [6] agar petani dan produsen dapat menyesuaikan pola tanam dan produksi sesuai dengan prediksi harga yang diharapkan, sehingga dapat mengoptimalkan keuntungan.

Prediksi umumnya dibuat secara otomatis melalui penggunaan metode regresi. Regresi merupakan sebuah metode untuk mengestimasi nilai y berdasarkan nilai x yang telah diberikan [7][8]. Pemilihan algoritma regresi harus disesuaikan dengan karakteristik serta uji asumsi klasik terhadap data yang digunakan [9]. Data yang digunakan pada penelitian ini merupakan data yang tidak berdistribusi normal, terjadi autokorelasi serta tidak adanya linearitas. Data keadaan iklim memiliki beberapa variabel seperti suhu, kelembapan, curah hujan dan kecepatan angina [12][13]. Variabel-variabel tersebut memiliki tipe data kontinu dan berbentuk desimal. Distribusi data juga memiliki banyak *outlier*s dikarenakan data harga jual cabai rawit yang digunakan mengalami fluktuasi.

Beberapa algoritma regresi yang banyak digunakan untuk melakukan prediksi pada data yang non-linear antara lain *XGBoost Regression* [12], *KNN Regression* [13] dan *Random Forest Regression* [14]. Dari ketiga algoritma tersebut memiliki kelebihan dan kekurangan yang berkaitan antara satu sama lain, maka dari itu dilakukan perbandingan sehingga dapat diketahui algoritma apa yang memiliki performa terbaik dalam memprediksi [15].

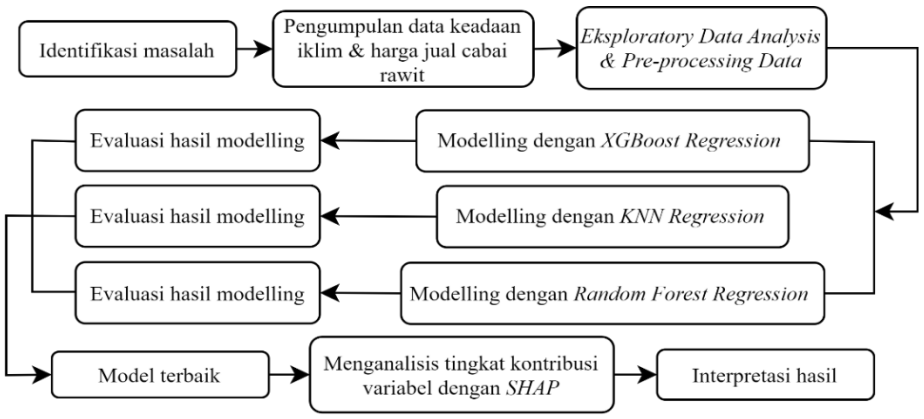
Penelitian terdahulu menerapkan *Random Forest Regression* pada data *non*-linier untuk memprediksi harga rumah, hasilnya *Random Forest Regression* dapat memprediksi dengan akurasi 81,5% berdasarkan luas tanah, luas bangunan dan sebagainya [16]. Menurut penelitian *KNN Regression* memiliki nilai *error* terendah dalam memprediksi daya turbin berdasarkan kecepatan angin, arah angin, dan suhu udara [17]. Penelitian sebelumnya menggunakan data iklim dari BMKG dan harga jual cabai di daerah kabupaten Bandung untuk memprediksi harga cabai. Harga cabai yang telah diprediksi kemudian diklasifikasi menggunakan algoritma *SVM* dengan optimasi *ANFIS* dengan rata-rata akurasi sebesar 92,68%. Penelitian berikutnya bertujuan untuk membandingkan 6 model regresi yang berbeda sehingga ditemukan model terbaik dalam memprediksi rating restoran serta mengidentifikasi variabel yang mempengaruhi rating sebuah restoran, hasilnya yaitu model *Random Forest Regression* dengan akurasi 92% memiliki performa terbaik dibanding dengan model lainnya [18].

Berdasarkan uraian yang telah dipaparkan, penelitian ini membandingkan beberapa algoritma regresi seperti *XGBoost Regression, KNN Regression* dan *Random Forest Regression* untuk mendapatkan model yang akurat dalam memprediksi harga jual cabai rawit berdasarkan keadaan iklim. Perbandingan performa algoritma dilakukan dengan menggunakan matriks evaluasi seperti *MAE, MAPE* dan *R2-Score* untuk menilai seberapa baik serta seberapa kecil *error* yang dihasilkan model dalam memprediksi harga berdasarkan variabel input yang telah ditentukan [19]. Penelitian ini bukan hanya sekedar mendapatkan model yang akurat dalam

memprediksi saja, tetapi diharapkan juga dapat mengetahui variabel-variabel iklim yang berpengaruh signifikan terhadap prediksi harga jual cabai rawit.

2. METODE PENELITIAN

Alur penelitian dari awal hingga mendapatkan hasil yang diharapkan ditunjukkan pada Gambar 1. Penelitian ini menggunakan bahasa pemrograman *Python* dengan platform *google colaboratory* untuk membantu dalam menganalisisi dan *modelling.* Penelitian diawali dengan pengumpulan data dari *website* resmi sesuai data yang dimaksud. Dilanjutkan dengan EDA untuk melihat detail mengenai data dan melihat indikasi anomali pada data. Apabila ternyata data memiliki indikasi yang mencurigakan, maka dilakukan proses *pre-processing* untuk menangani dan menyesuaikan data dengan algoritma yang digunakan. Masuk ke tahap utama yaitu data fitting dengan menggunakan algoritma *XGBoost Regression*, *KNN Regression*, dan *Random Forest Regression*. Hasil pemodelan tersebut diuji dan dievaluasi kinerjanya dengan menggunakan *MAE*, *MAPE*, dan *R2-Score*. Model terbaik selanjutnya digunakan untuk menganalisis variabel-variabel yang berkontribusi terhadap harga jual dengan menggunakan metode *SHAP*. Setelah proses analisis data selesai, tahap terakhir yaitu menginterpretasi hasil yang diperoleh dan menarik kesimpulan.



Gambar 1. Diagram Alur Penelitian

2.1 Pengumpulan Data

Data yang digunakan pada penelitian ini yaitu data sekunder yang diperoleh dari *website* resmi. Data keadaan iklim yang terdiri dari variabel Suhu, Kelembapan, Curah Hujan, Lama Penyinaran dan Kecepatan Angin didapat dari *website* resmi BMKG yaitu https://dataonline.bmkg.go.id/data_iklim. Data harga jual cabai rawit didapat dari *website* SiHaTi (Sistem Informasi Harga dan Produksi Komoditi) yaitu https://hargajateng.org/tabel-harga-komoditi. Data-data tersebut merupakan data harian yang dikumpulkan dari Januari 2016-Desember 2023. Kemudian data tersebut diintegrasikan dalam satu tabel untuk memudahkan dalam melakukan *modelling.*

2.2 EDA dan Pre-processing Data

EDA (*Exploratory Data Analysis*) merupakan langkah awal yang penting dalam proses analisis data yang membantu untuk mendapatkan intuisi tentang data sebelum menerapkan metode analisis lebih lanjut atau membangun model. Pada tahap ini dibagi menjadi beberapa bagian yaitu understanding data dan visualisasi. *Pre-processing* data dilakukan untuk mempersiapkan data sebelum masuk ke dalam pemodelan dengan tujuan untuk menyesuaikan data sesuai dengan karakteristik dari masing-masing algoritma yang digunakan dalam

pemodelan. Pada tahap ini dibagi menjadi data *cleaning*, penanganan *missing value* dan transformasi data. Transformasi menggunakan metode *BoxCox* yang digunakan untuk mengubah distribusi data yang tidak normal menjadi distribusi normal. Adapun perhitungan transformasi ini menggunakan Persamaan (1)

$$Y^{(\lambda)} = \begin{cases} \frac{Y^{(\lambda)}-1}{\lambda} \; if \; \lambda \neq 0 \\ \log Y \, , if \; \lambda = 0, \end{cases} \quad (1)$$

dimana Y merupakan nilai yang ingin ditransformasi dan λ merupakan parameter dalam *BoxCox* [20]-[21]. Setelah transformasi data, maka tahap selanjutnya yaitu standarisasi data dan *splitting* data kedalam data latih dan data uji. *Splitting* data dilakukan sebanyak lima kali untuk memastikan model yang dibuat stabil dengan data yang di acak dan mengurangi resiko *overfitting* pada model.

2.3 Pemodelan

Dalam tahap ini dilakukan pada algoritma *XGBoost Regression*, *KNN Regression* dan *Random Forest Regression*.

2.3.1 XGBoost Regression

XGBoost adalah sebuah algoritma yang digunakan dalam *gradient tree boosting* untuk membangun model *ensemble* yang kuat dan efisien. *XGBoost Regression* adalah metode regresi yang menggunakan algoritma *XGBoost* untuk memprediksi nilai kontinu. Setiap pohon keputusan memprediksi variabel target dan menghitung residual yang dihasilkan oleh pohon-pohon sebelumnya. Dalam kasus regresi, umumnya digunakan fungsi objektif yang berhubungan dengan mengurangi kesalahan prediksi, seperti *Mean Squared Error (MSE)* atau fungsi objektif yang serupa untuk meminimalkan *loss function*, formulanya pada Persamaan (2)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad (2)$$

dimana n adalah jumlah data point dalam dataset, $y_i$adalah nilai target yang sebenarnya dari data point ke-i dan $\hat{y}_i$ adalah prediksi dari model untuk data point ke-I [22].

2.3.2 KNN Regression

K-Nearest Neighbor (*KNN*) adalah metode pengenalan pola yang tidak bergantung pada parameter, berguna untuk klasifikasi dan regresi. Algoritma ini memanfaatkan perhitungan jarak antar titik data dan menentukan tetangga terdekat untuk setiap titik data tertentu. Ada berbagai metode untuk menghitung jarak ini, metode yang paling umum dikenal pada data kontinyu adalah *Euclidean*. Persamaan (3) merupakan perhitungan dari *Euclidean.*

$$d = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}, \quad (3)$$

Euclidean dihitung sebagai akar kuadrat dari jumlah selisih kuadrat antara titik baru (x) dan titik yang sudah ada (y) [23].

2.3.3 Random Forest Regression

Konsep dari algoritma *Random Forest* adalah menggabungkan prediksi dari beberapa pohon keputusan *(decision trees*) yang dibangun secara acak untuk mencapai akurasi prediksi yang lebih tinggi. *Random Forests* juga dapat digunakan untuk masalah regresi. Dalam metode *Random Forests* untuk regresi, pohon-pohon keputusan dibangun menggunakan vektor acak, di mana prediktor pohon menghasilkan nilai numerik sebagai hasilnya. Prediksi akhir diperoleh dengan mengambil mayoritas suara atau rata-rata dari prediksi pohon-pohon tersebut *Random Forests Regression* menghasilkan nilai prediksi berupa kontinyu dan menggunakan *MSE* sebagai metode dalam membagi node. Adapun perhitungan *MSE* menggunakan Persamaan (2) [24].

2.4 Matriks Evaluasi

Setelah pemodelan dilakukan, maka untuk mengetahui performa dari masing-masing algoritma, dilakukan pengukuran matriks evaluasi menggunakan *MAE*, *MAPE* dan *R2-Score*. Identifikasi model terbaik jika memenuhi persyaratan yaitu memiliki *MAE* terendah, *MAPE* terendah dan *R2-Score* terbesar.

2.4.1 Mean Absolute Error (MAE)

MAE mengukur rata-rata kesalahan yang bersifat mutlak dari prediksi tersebut. Semakin kecil nilai *MAE*, semakin baik kualitas model yang dibuat. Nilai *MAE* dihitung dengan Persamaan (4)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{4}$$

dimana n merupakan jumlah dari observasi, y_i merupakan nilai aktual dan $\hat{y}_i$ merupakan nilai hasil prediksi [14].

2.4.2 Mean Absolute Percentage Error (MAPE)

MAPE menggambarkan persentase kesalahan prediksi terhadap data aktual selama periode tertentu. Semakin kecil nilai *MAPE*, semakin baik kualitas dari model yang digunakan. *MAPE* didapatkan dengan Persamaan (5)

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| x\ 100 \tag{5}$$

Adapun kriteria nilai *MAPE* yaitu apabila *error* <10% maka masuk dalam kategori sangat baik, *error* antara 10%-20% masuk dalam kategori baik, *error* antara 20%-50% masuk dalam kategori cukup *error* >50% masuk dalam kategori buruk [14].

2.4.3 Goodness-of-fit (R2-Score)

R2-Score adalah ukuran mengenai seberapa dekat nilai prediksi dari suatu model cocok dengan nilai yang diamati. Nilai R2 ideal suatu model adalah 1 yang menunjukkan bahwa model tersebut dapat menjelaskan seluruh variabilitas pada kelas sasaran. Nilai R2 dihitung dengan Persamaan (6)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|^2}{\sum_{i=1}^{n}|y_i - \bar{y}_i|,^2} \tag{6}$$

dimana $\bar{y}_i$ mewakili rata-rata dari nilai target yang sebenarnya. Ketika nilai prediksi mendekati nilai sebenarnya, maka *MAE* mendekati nol, sehingga *R2-Score* yang mendekati 1 menunjukkan kecocokan yang baik antara nilai hasil prediksi dan nilai sebenarnya [14].

2.5 SHapley Additive explanation (SHAP)

Setelah model terbaik didapat, analisis selanjutnya yaitu pengaruh signifikan dan kontribusi antar variabel x terhadap variabel y. Pengaruh signifikan variabel x dihasilkan dari model yang memiliki performa terbaik. Untuk mengetahui kontribusi setiap variabel x terhadap variabel Harga Jual menggunakan metode *SHAP* dengan menghitung nilai *Shapely* pada masing-masing variabel. *SHAP* merupakan salah satu metode yang membantu dalam melakukan interpretasi dari hasil prediksi model pada *machine learning*. Tujuan dari *SHAP* adalah untuk menjelaskan prediksi dari sebuah *instance* x dengan menghitung kontribusi dari setiap fitur untuk prediksi. Nilai fitur dari *instance* data bertindak sebagai *players* dalam suatu koalisi. Nilai *Shapley* memberi tahu cara mendistribusikan "prediksi" secara adil di antara fitur. Persamaan (7) untuk menghitung kontribusi masing-masing variabel menggunakan nilai *Shapely* [25].

$$\varphi_j = \sum_{S \subseteq N\{j\}} \frac{|S|!(N-|S|-1)!}{N!}\big(v(S \cup \{j\} - v(S)\big), \tag{7}$$

dimana S adalah subset fitur yang digunakan dalam model, $v(S)$ adalah nilai dari fungsi karakteristik setiap prediktor dan N adalah jumlah prediktor. Kontribusi dari setiap fitur didefinisikan juga sebagai nilai rata-rata dari marginal kontribusinya terhadap seluruh permutasi yang mungkin dari set fitur [26].

3. HASIL DAN PEMBAHASAN

3.1 EDA and Pre-processing Data

Data penelitian yang diambil dari *website* BMKG dan SiHaTi yang diintegrasikan menghasilkan cuplikan data pada Tabel 1.

Tabel 1. Integrasi Data Keadaan Iklim dan Harga Jual

Tanggal	Suhu	Kelembapan	Curah Hujan	Lama Penyinaran	Kecepatan Angin	Harga Jual
01/01/2016	27	88	9,6	3,6	4	Rp 49600
02/01/2016	27,4	87	16,7	3,4	3	Rp 49600
03/01/2016	28,4	84	2	2,5	3	Rp 44400
……..	……	……..	….	……..	…….	……..
29/12/2023	28,9	84	5,6	7,9	2	Rp 78000
30/12/2023	28,9	84	2	9	1	Rp 75600
31/12/2023	28,9	84	0	7,7	2	Rp 65400

Pada penelitian ini menggunakan data sebanyak 2922 baris dan 7 kolom yang diambil dari Januari 2016-Desember 2023. Setelah ditelusuri, tidak terdapat adanya duplikasi sehingga dataset bebas dari duplikat. Selajutnya yaitu mengidentifikasi jumlah dan persentase missing value di setiap variabel pada Tabel 2.

Tabel 2. Distribusi *Missing Value* Pada Dataset

Variabel	***Missing Value***	**Persentase**
Curah Hujan	173	5,92%
Lama Penyinaran	129	4,14%
Kelembapan	20	0,68%
Suhu	12	0,41%
Kecepatan Angin	3	0,10%
Tanggal	0	0,000
Harga Jual	0	0,000

Diketahui data yang diambil dari *website* BMKG memiliki data 8888 yang mengindikasikan bahwa data pada *record* ini tidak terukur oleh alat. Data ini dihilangkan dan diubah menjadi data *null*. Dari proses ini hanya variabel Curah Hujan saja yang dihilangkan, data 8888 menjadi data *null* sehingga terjadi penambahan data *null*. Awalnya data *null* untuk variabel ini pada Tabel 2 berjumlah 173, kemudian bertambah menjadi 302.

Tabel 3. Cuplikan Data Sebelum dan Setelah Penanganan *Missing Value*

Sebelum Penanganan *Missing Value*			**Setelah Penanganan *Missing Value***		
Suhu	**Kelembapan**	**Curah Hujan**	**Suhu**	**Kelembapan**	**Curah Hujan**
28,80	75,13	**NaN**	28,80	75,13	**10,60**
NaN	**NaN**	0,2	**28,57**	**76,56**	0,2
28,44	77,53	0,0	28,44	77,53	0,0

Pada Tabel 3, nilai-nilai yang hilang yang tersebar di seluruh dataset ditangani dengan menggunakan imputasi nilai rata-rata. Teknik yang digunakan dalam tahap ini adalah mengisi

data yang hilang sesuai dengan bulan di mana data tersebut *null* sehingga pengisian ini memiliki ketentuan berdasarkan masing-masing waktu. Hal pertama yang harus dilakukan adalah mengekstrak fitur dari kolom Tanggal dan kemudian memisahkannya menjadi kolom Bulan dan kolom Tahun. Setelah itu, variabel-variabel yang memiliki nilai yang hilang di setiap bulan dan tahun dikelompokkan dan kemudian dimasukkan dalam proses perhitungan nilai rata-rata. Setelah mendapatkan nilai rata-rata setiap variabel untuk setiap bulan dan tahun, nilai-nilai ini diisi sebagai nilai yang hilang.

Tabel 4. Cuplikan Data Sebelum dan Setelah Transformasi Data

Sebelum Transformasi Data			**Setelah Transformasi Data**		
Suhu	**Kelembapan**	**Curah Hujan**	**Suhu _transformed**	**Kelembapan _transformed**	**Curah Hujan _transformed**
28,80	75,13	10,60	2,5105e+07	19825,06	1,39
28,57	76,56	0,2	2,4080e+07	20772,30	0,17
28,44	77,53	0,0	2,3475e+07	21421,30	0,00

Identifikasi selanjutnya adalah jumlah data *outlier* pada masing-masing variabel. Diketahui bahwa Curah Hujan merupakan variabel penyumbang data *outlier* sebesar 13,7%, diikuti oleh variabel Harga Jual sebesar 2,29%, dan variabel lainnya yaitu Suhu, Kecepatan Angin, Kelembaban dan Lama Paparan yang masing-masing memiliki data *outlier* dengan presentase di bawah 0% untuk semua data. Tabel 4 merupakan cuplikan data sebelum dan sesudah transformasi data menggunakan *BoxCox*. Setelah dilakukan transformasi data diketahui bahwa terjadi penurunan sebaran *outlier* pada masing-masing variabel. Penurunan sebaran *outlier* yang signifikan terjadi pada variabel Curah Hujan menjadi 0, diikuti oleh Suhu, Kelembaban dan Lama Paparan.

```
array([[ 1.0139457 , -0.56510663,  1.21179396,  0.20704295, -0.40624339],
       [ 0.5241065 , -0.17215877, -0.65772257, -0.06140115,  0.92451274],
       [ 0.2352512 ,  0.09706881, -0.9250806 , -1.08096719, -0.40624339],
       ...,
```

Gambar 2. Cuplikan Data Setelah *StandarScaler*

Setelah proses transformasi, dilakukan standarisasi data dengan menggunakan metode *StandardScaler*. Hasil dari standarisasi data ini dapat dilihat pada Gambar 2. Proses terakhir pada tahap pra-proses ini adalah pemisahan data. Pembagian data menggunakan proporsi 80:20, yaitu 80% atau sekitar 2337 baris merupakan data latih dan 20% atau sekitar 585 baris merupakan data uji. Setelah semua proses pada tahap ini dilakukan, maka data siap digunakan untuk pemodelan.

3.2 Modelling dan Evaluasi

Tabel 5, 6 dan 7 menunjukkan hasil evaluasi masing-masing model dalam memprediksi data uji.

Tabel 5. Evaluasi Model *XGBoost Regression*

Evaluasi	***Test Set***	***Train Set***
MAE	3388,03	3014,17
MAPE	8,96%	8,61%
R2-Score	0,91	0,93

Pada Tabel 4 dapat dilihat bahwa *error* nilai Harga Jual yang dihasilkan pada data uji mencapai Rp. 3388 dengan selisih Rp. 374 terhadap data latih. *Error* dalam melakukan prediksi menggunakan algoritma *XGBoost Regression* mencapai 8,96%, dengan hasil tersebut *error* prediksi menggunakan algoritma *XGBoost Regression* masuk dalam kriteria sangat baik, dengan selisih antara data uji dan data latih sebesar 0,3%. Nilai *R2-Score* sebesar 0,91 menunjukkan bahwa 91% variabilitas pada variabel dependen dapat dijelaskan oleh variabel independen dalam

model, hal ini menunjukkan bahwa algoritma *XGBoost Regression* sangat kuat dalam memprediksi variabel Harga Jual berdasarkan variabel iklim.

Tabel 6. Evaluasi Model *KNN Regression*

Evaluasi	***Test Set***	***Train Set***
MAE	5581,99	4130,41
MAPE	14,32%	11,28%
R2-Score	0,79	0,86

Pada Tabel 6 dapat dilihat bahwa *error* nilai Harga Jual yang dihasilkan pada data uji mencapai Rp 5581 dengan selisih sebesar Rp 1451 dibandingkan dengan data latih. *Error* dalam melakukan prediksi menggunakan algoritma *KNN Regression* mencapai 14,32%, dengan hasil tersebut *error* prediksi menggunakan algoritma *KNN Regression* masuk dalam kriteria baik, dengan selisih antara data uji dan data latih sebesar 3,04%. Nilai *R2-Score* sebesar 0,79 menunjukkan bahwa 79% variabilitas pada variabel dependen dapat dijelaskan oleh variabel independen dalam model, hal ini menunjukkan bahwa algoritma *KNN Regression* cukup efektif dalam memprediksi variabel Harga Jual berdasarkan variabel iklim.

Tabel 7. Evaluasi Model *Random Forest Regression*

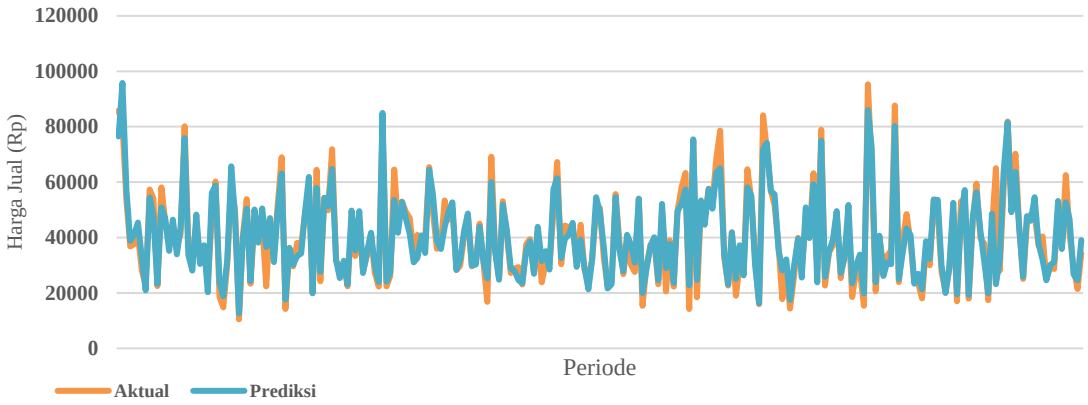
Evaluasi	***Test Set***	***Train Set***
MAE	5964,96	5683
MAPE	16,71%	16,55%
R2-Score	0,76	0,75

Pada Tabel 7 dapat dilihat bahwa *error* nilai Harga Jual yang dihasilkan pada data uji mencapai Rp 5964 dengan selisih Rp 281 terhadap data latih. *Error* dalam melakukan prediksi menggunakan algoritma *Random Forest Regression* mencapai 16,71%, dengan hasil tersebut *error* prediksi menggunakan algoritma *Random Forest Regression* masuk dalam kriteria baik, dengan selisih antara data uji dan data latih sebesar 0,16%. Nilai *R2-Score* sebesar 0,76 menunjukkan bahwa 76% variabilitas variabel dependen dapat dijelaskan oleh variabel independen dalam model, hal ini menunjukkan bahwa algoritma *Random Forest Regression* cukup efektif dalam memprediksi variabel Harga Jual berdasarkan variabel iklim.

Tabel 8. Perbandingan Performa dari Setiap Model

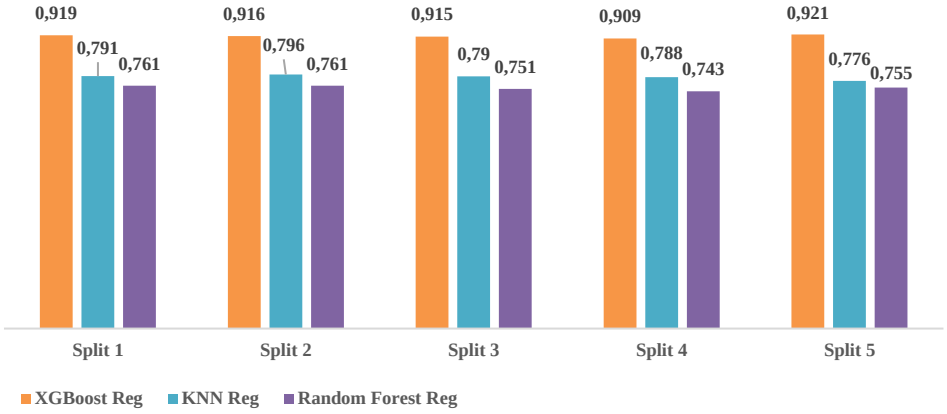
Rank	**Model**	**Matriks Evaluasi**		
		MAE	***MAPE***	***R2-Score***
1	XGBoost Regresison	3388	8,96%	0,91
2	*KNN Regression*	5581	14,32%	0,79
3	*Random Forest Regression*	5964	16,71%	0,76

Tabel 8 merupakan perbandingan kinerja masing-masing model. Dapat dilihat bahwa algoritma *XGBoost Regression* merupakan model terbaik dalam memprediksi harga jual cabai rawit berdasarkan kondisi iklim harian di Kota Semarang dengan *error* prediksi berada pada kategori sangat baik. *R2-Score* juga membuktikan bahwa algoritma *XGBoost Regression* tertinggal jauh dengan selisih 0,12 di atas model lainnya. Dari evaluasi *MAE* juga dapat dilihat bahwa *error absolute* dalam memprediksi Harga Jual sangat kecil dengan selisih Rp 2193 dibandingkan model lainnya.



Gambar 3. Diagram Garis Nilai Aktual vs Prediksi

Pada Gambar 3 terlihat bahwa hasil prediksi menggunakan model terbaik pada data uji yaitu algoritma *XGBoost Regression* menghasilkan grafik perbandingan antara nilai aktual dengan nilai prediksi yang saling berdekatan. Garis berwarna biru merepresentasikan data aktual dan garis berwarna oranye merepresentasikan data prediksi. Diketahui bahwa kedua garis tersebut saling berdekatan, dimana terdapat selisih garis yang tidak terlalu signifikan.

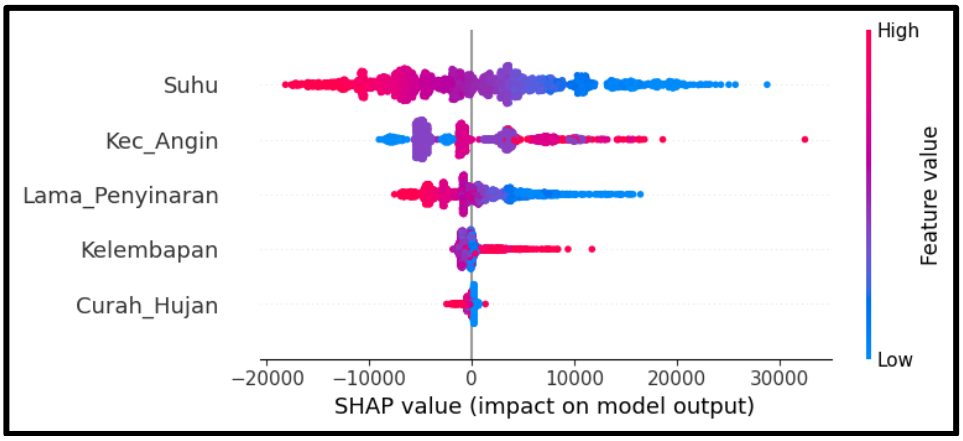


Gambar 4. Perbandingan *R2-Score* dengan 5 *Split* Berbeda

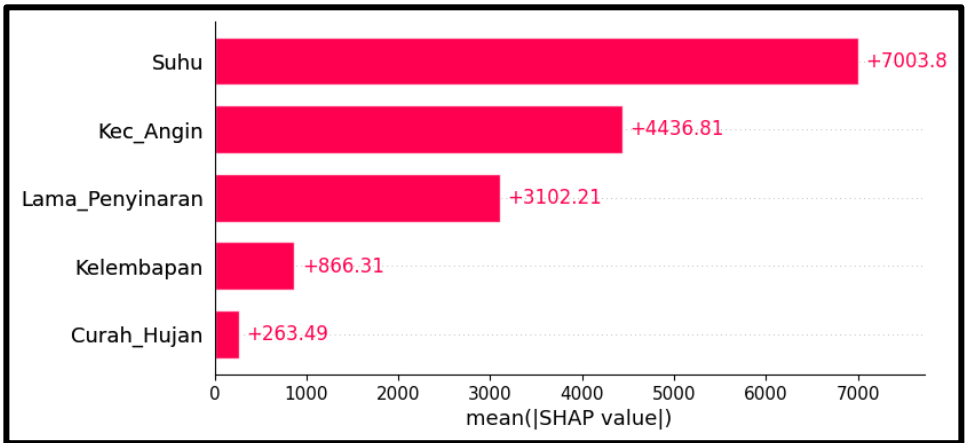
Pada Gambar 4 menampilkan hasil *R2-Score* dari lima *splitting* data yang berbeda. Terlihat bahwa model-model yang telah di *fitting* tidak menunjukkan perubahan *R2-Score* yang signifikan. Selisih nilai di setiap *split* hanya berkisar 0,001-0,01 saja. Hal ini mengartikan bahwa model-model yang telah di *fitting* dapat melakukan generalisasi yang baik pada data dan tidak adanya indikasi *overfitting*.

3.3 Variabel Contributin to Selling Prices

Kontribusi masing-masing variabel menggunakan metode *SHAP*. Kontribusi variabel secara umum ditunjukkan pada Gambar 5, sedangkan kontribusi per-variabel terhadap hasil prediksi harga jual ditunjukkan pada Gambar 6.



Gambar 5. Diagarm *SHAP*



Gambar 6. Diagram Nilai *Shapely*

Pada Gambar 5, semakin merah warna yang ditampilkan, semakin tinggi nilai variabel tersebut. Variabel suhu yang rendah memiliki dampak positif terhadap hasil prediksi harga jual, dan sebaliknya. Variabel Kecepatan Angin yang rendah memiliki dampak negatif terhadap hasil prediksi Harga Jual, dan sebaliknya. Gambar 6 menampilkan kontribusi masing-masing variabel terhadap pengaruh hasil prediksi Harga Jual. Suhu merupakan variabel yang memiliki kontribusi terbesar terhadap harga jual dengan nilai *SHAP* rata-rata +7003,8 yang menunjukkan bahwa variabel ini memiliki dampak positif terhadap harga jual.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dipaparkan dalam pembahasan, berikut merupakan beberapa kesimpulan yang dapat ditarik.

1. *XGBoost Regression* merupakan model terbaik dibandingkan dengan model lainnya dalam melakukan prediksi dengan nilai *MAE* terkecil yaitu 3388, nilai *MAPE* terkecil yaitu 9,96% yang masuk dalam kategori sangat baik dan *R2-Score* terbesar yaitu 0,91. Menggunakan lima *splitting* data dengan nilai *randomstate* yang berbeda menunjukkan bahwa selisih *R2-Score* di setiap *split* hanya berkisar 0,001-0,01 saja. Hal ini mengartikan bahwa model-model yang telah difitting dapat melakukan generalisasi yang baik pada data dan tidak adanya indikasi *overfitting*.
2. Variabel-variabel iklim yang berkontribusi terhadap nilai prediksi ditunjukkan oleh diagram *SHAP*. Diketahui bahwa Suhu merupakan variabel yang memiliki kontribusi paling

signifikan dengan nilai rata-rata *SHAP* yaitu +7003,8 yang menandakan bahwa variabel ini memberikan dampak positif terhadap prediksi Harga Jual Variabel lainnya yaitu Keceparan Angin dan Lama Penyinaran yang juga memiliki kontribusi cukup besar terhadap prediksi Harga Jual. Dari nilai kontribusi yang diberikan oleh variabel-variabel tersebut, dapat disimpulkan bahwa kenaikan atau penurunan Harga Jual cabai rawit di Kota Semarang lebih banyak dipengaruhi oleh Suhu. Kecepatan Angin dan Lama Penyinaran matahari.

Terdapat beberapa keterbatasan dalam penelitian ini yaitu lokasi penelitian hanya terfokus pada Kota Semarang. Temuan ini bergantung pada kumpulan data spesifik yang digunakan untuk analisis dan hanya menggunakan tiga algoritma, sehingga berpotensi mengabaikan algoritma lain yang lebih efektif dalam memprediksi harga dibandingkan dengan algoritma dalam penelitian ini.

5. SARAN

Penelitian ini memberikan peluang untuk eksplorasi dan modifikasi di masa mendatang, diantaranya yaitu; Pertama, cakupan wilayah penelitian dapat diperluas sehingga harga yang diprediksi dan dianalisis mencakup wilayah secara nasional. Kedua, menggunakan teknik hyperparameter-tuning untuk mengetahui parameter yang optimal dalam melakukan prediksi. Ketiga, hasil prediksi dapat ditampilkan di situs web menggunakan tampilan dasbor, sehingga memudahkan orang untuk mengakses dan memahami hasil prediksi.

DAFTAR PUSTAKA

[1] M. N. Ridho and N. E. Suminarti, “The Effect of The Climate Change on Cayenne Pepper (Capsicum frutescens L.) Productivities In Malang Regency,” *J. Produksi Tanam.*, vol. 8, no. 3, pp. 304–314, 2020.

[2] T. Purwanti, “Harga Cabai Makin ‘Pedas’, Inflasi RI Melejit 0,61%,” *CNBC Indonesia*, 2022. .

[3] Diskominfo Indramayu, “Inflasi Harga Cabe Rawit Merangkak Naik,” *Diskominfo Indramayu.go.id*, 2023. https://diskominfo.indramayukab.go.id/rakor-inflasi-harga-cabe-rawit-merangkak-naik/.

[4] P. Nasional, “PUSAT INFORMASI HARGA PANGAN STRATEGIS NASIONAL-Harga Jual Cabai Rawit Kota Semarang,” *bi.go.id*, 2023. https://www.bi.go.id/hargapangan.

[5] A. H. Nurcahyono, F. Nhita, D. Saepudin, and A. Aditsania, “Price prediction of chili in bandung regency using support vector machine (SVM) optimized with an adaptive neuro-fuzzy inference system (ANFIS),” *2019 7th Int. Conf. Inf. Commun. Technol. ICoICT 2019*, pp. 1–6, 2019, doi: 10.1109/ICoICT.2019.8835367.

[6] Y. Apriyana, E. Susanti, and F. Ramadhani, “Analysis of Climate Change Impacts on Food Crops Production in Dry Land and Design of Information System,” *Inform. Pertan.*, vol. 25, no. 1, pp. 69–80, 2016.

[7] S. Hasmita, F. Nhita, D. Saepudin, and A. Aditsania, “Chili commodity price forecasting in bandung regency using the adaptive synthetic sampling (ADASYN) and K-Nearest neighbor (KNN) algorithms,” *2019 Int. Conf. Inf. Commun. Technol. ICOIACT 2019*, pp. 434–438, 2019, doi: 10.1109/ICOIACT46704.2019.8938525.

[8] D. Novianty, N. D. Palasara, and M. Qomaruddin, “Algoritma Regresi Linear pada Prediksi Permohonan Paten yang Terdaftar di Indonesia,” *J. Sist. dan Teknol. Inf.*, vol. 9, no. 2, p. 81, 2021, doi: 10.26418/justin.v9i2.43664.

[9] G.- MARDIATMOKO, “Pentingnya Uji Asumsi Klasik Pada Analisis Regresi Linier Berganda,” *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 14, no. 3, pp. 333–342, 2020, doi: 10.30598/barekengvol14iss3pp333-342.

[10] S. D. and K. P. R. M. Satish, Prakash, S. M. Babu, P. P. Kumar, "Artificial Intelligence (AI) and the Prediction of Climate Change Impacts," *2023 IEEE 5th Int. Conf. Cybern. Cogn. Mach. Learn. Appl.*, pp. 660–664, doi: 10.1109/ICCCMLA58983.2023.10346636.

[11] X. L. and N. T. Y. Li, "Climate change prediction by combination prediction method based on deep learning models," *2023 5th Int. Acad. Exch. Conf. Sci. Technol. Innov.*, pp. 1508–1512, 2023, doi: 10.1109/IAECST60924.2023.10502965.

[12] Zhagparov, Z. Buribayev, S. Joldasbayev, A. Yerkosova, and M. Zhassuzak, "Building a System for Predicting the Yield of Grain Crops Based on Machine Learning Using the XGBRegressor Algorithm," *SIST 2021 - 2021 IEEE Int. Conf. Smart Inf. Syst. Technol.*, pp. 28–30, 2021, doi: 10.1109/SIST50301.2021.9465938.

[13] I. D. Oktaviani and A. G. Putrada, "KNN imputation to missing values of regression-based rain duration prediction on BMKG data," *J. Infotel*, vol. 14, no. 4, pp. 249–254, 2022, doi: 10.20895/infotel.v14i4.840.

[14] Y. Li *et al.*, "Random forest regression for online capacity estimation of lithium-ion batteries," *Appl. Energy*, vol. 232, no. September, pp. 197–210, 2018, doi: 10.1016/j.apenergy.2018.09.182.

[15] P. Kulkarni et al., "Comparison of Regression and Classification Models for Prediction of the Retail Sales," *2023 Int. Conf. Integr. Comput. Intell. Syst.*, pp. 1–7, 2023, doi: 10.1109/ICICIS56802.2023.10430292.

[16] E. Fitri, "Analisis Perbandingan Metode Regresi Linier, Random Forest Regression dan Gradient Boosted Trees Regression Method untuk Prediksi Harga Rumah," *J. Appl. Comput. Sci. Technol.*, vol. 4, no. 1, pp. 58–64, 2023, doi: 10.52158/jacost.v4i1.491.

[17] O. Eyecioglu, B. Hangun, K. Kayisli, and M. Yesilbudak, "Performance comparison of different machine learning algorithms on the prediction of wind turbine power generation," *8th Int. Conf. Renew. Energy Res. Appl. ICRERA 2019*, pp. 922–926, 2019, doi: 10.1109/ICRERA47325.2019.8996541.

[18] J. Priya, "Predicting Restaurant Rating using Machine Learning and comparison of Regression Models," *Int. Conf. Emerg. Trends Inf. Technol. Eng. ic-ETITE 2020*, pp. 1–5, 2020, doi: 10.1109/ic-ETITE47903.2020.238.

[19] M. S. Acharya, "A Comparison of Regression Models for Prediction of Graduate Admissions," *2019 Int. Conf. Comput. Intell. Data Sci.*, pp. 1–5, 2019.

[20] M. M. Ahsan, M. A. P. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique, "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," *Technologies*, vol. 9, no. 3, pp. 5–9, 2021, doi: 10.3390/technologies9030052.

[21] M. Saputra, J. P. Sidabuke, R. P. Sinulingga, and R. B. Tamba, "Analisis Metode Algoritma K-Nearest Neighbor (KNN) dan Naive Bayes Untuk Klasifikasi Diabetes Mellitus," *J. TEKINKOM*, vol. 6, no. 2, pp. 723–729, 2023, doi: 10.37600/tekinkom.v6i2.942.

[22] Y. Wang, Z. Pan, J. Zheng, L. Qian, and M. Li, "A hybrid ensemble method for pulsar candidate classification," *Astrophys. Space Sci.*, vol. 364, no. 8, pp. 1–15, 2019, doi: 10.1007/s10509-019-3602-4.

[23] M. M. Baharuddin, H. Azis, and T. Hasanuddin, "Analisis Performa Metode K-Nearest Neighbor Untuk Identifikasi Jenis Kaca," *Ilk. J. Ilm.*, vol. 11, no. 3, pp. 269–274, 2019, doi: 10.33096/ilkom.v11i3.489.269-274.

[24] Leo Breiman, *Random Forest*, vol. 45. Kluwer Academic Publishers. Manufactured in The Netherlands, 2001.

[25] S. Ben Jabeur, S. Mefteh-Wali, and J. L. Viviani, "Forecasting gold price with the XGBoost algorithm and SHAP interaction values," *Ann. Oper. Res.*, vol. 334, no. 1–3, pp. 679–699, 2024, doi: 10.1007/s10479-021-04187-w.

[26] L. S. . Lundberg. S.M, "A Unified Approach to Interpreting Model Predictions Scott," *31st Conf. Neural Inf. Process. Syst.*, vol. 16, no. 3, pp. 426–430, 2017.