Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 480−488
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5930

Implementasi Hyperparameter Tuning Grid Search CV Pada Prediksi Produksi Padi Menggunakan Algoritma Linear Regresi

Fathir Jamiluddin*, Sutan Faisal, Santi Arum Puspita Lestari, Ahmad Fauzi

Fakultas Teknik Informatika, Universitas Buana Perjuangan Karawang, Karawang
Jl. HS.Ronggo Waluyo, Puseurjaya, Telukjambe Timur, Karawang, Jawa Barat, Indonesia Indonesia
Email: [1,*]if20.fathirjamiluddin@mhs.ubpkarawang.ac.id, [2]sutan.faisal@ubpkarawang.ac.id, [3]santi.arum@ubpkarawang.ac.id, [4]afauzI@ubpkarawang.ac.id
Email Penulis Korespondensi: if20.fathirjamiluddin@mhs.ubpkarawang.ac.id
Submitted: **18/09/2024;** Accepted: **26/10/2024;** Published: **27/10/2024**

Abstrak−Padi merupakan salah satu tanaman utama di Indonesia yang menghasilkan makanan pokok terebesar yaitu komoditi beras. Beras merupakan makanan pokok yang dikonsumsi oleh hampir 98% masyarakat Indonesa. Penelitian ini bertujuan untuk membandingkan Algoritma Linear Regression dan Decision Tre e dalam upaya menemukan algoritma yang paling sesuai untuk memprediksi data produksi padi. Linear Regression masih merupakan model yang berguna, terutama jika data memiliki hubungan non-linear yang tidak dapat ditangkap oleh Linear Regression. Maka dapat disimpulkan bahwa Algoritma Linear Regression dengan optimalisasi hyperparameter tunning grid search cv mampu memprediksi produksi padi lebih baik dibanding Algoritma Decision Tree dengan nilai R2-score 86.895666, MAE 261049.168107, dan MSE 160199780301.226318.

Kata kunci: Algoritma Linear Regression; Decision Tree; Padi; Grid Search CV; Hyperparameter Tuning

Abstract−Rice is one of the main crops in Indonesia that produces the largest staple food, namely rice commodities. Rice is a staple food consumed by almost 98% of Indonesian people. This study aims to compare the Linear Regression Algorithm and Decision Tree in an effort to find the most appropriate algorithm for predicting rice production data. Linear Regression is still a useful model, especially if the data has a non- linear relationship that cannot be captured by Linear Regression. So it can be concluded that the Linear Regression Algorithm with optimization of the tuning grid search cv hyperparameter is able to predict rice production better than the Decision Tree Algorithm with an R2-score value of 86.895666, MAE 261049.168107, and MSE 160199780301.226318.

Keywords: Linear Regression Algorithm; Decision Tree; Rice; Grid Search CV; Hyperparameter Tuning

1. PENDAHULUAN

Padi merupakan salah satu tanaman penghasil pangan pokok terbesar yaitu beras. Saat ini pola konsumsi beras mulai meluas ke daerah-daerah yang sebelumnya memiliki pola makanan pokok non-beras. Selain digunakan sebagai bahan pangan pokok, beras juga merupakan bahan baku industri yang strategis bagi perekonomian nasional sehingga permintaan terhadap beras semakin meningkat seiring dengan pertumbuhan jumlah penduduk, pertumbuhan ekonomi, daya beli masyarakat dan perubahan selera [1].

Kekurangan pangan dapat menimbulkan ketidakstabilan sosial, politik, dan ekonomi, sehingga dapat mengancam stabilitas nasional [2]. Untuk menjaga ketahanan pangan dan meningkatkan pendapatan serta kesejahteraan petani, produktivitas dan produksi padi harus terus ditingkatkan [3]. Setelah gandum dan jagung, padi merupakan tanaman pangan terpenting ketiga di dunia. Nasi masih menjadi makanan pokok bagi sebagian besar penduduk dunia, khususnya di Asia. Oleh karena itu, beras merupakan komoditas strategis di Indonesia karena dampaknya yang signifikan terhadap stabilitas politik dan perekonomian [4].

Untuk memahami pentingnya peningkatan produktivitas padi, diperlukan analisis data yang akurat mengenai prediksi hasil produksi padi di berbagai wilayah. Dalam penelitian ini, Pulau Sumatra sebagai salah satu provinsi dengan produksi padi yang signifikan, menjadi area penting untuk dianalisis. Data kategori tanaman pangan utama di delapan provinsi di Pulau Sumatra—Aceh, Sumatra Selatan, Sumatra Barat, Sumatra Utara, Riau, Jambi, Bengkulu, dan Lampung. dikumpulkan melalui website Kaggle. Dataset padi yang meliputi luas panen atau luas lahan dan statistik produksi tahunan digunakan untuk menyajikan data pada tahun 1993 hingga 2020. Kemudian, statistik harian curah hujan, kelembapan, dan suhu rata-rata pada tahun 1993 hingga 2020 dapat diunduh di website BMKG [5]. Pada penelitian sebelumnya, metode Regresi Linear Sederhana digunakan dalam memprediksi harga ubi kayu di CV Harum [6].

Selain itu, model regresi bertahap digunakan untuk menganalisis data tanaman sorgum, dan temuan menunjukkan bahwa variabel bebas terbaik untuk hasil biji sorgum adalah tinggi tanaman saat panen, lebar malai, dan panjang malai, dengan nilai koefisien korelasi dan determinasi sebesar 0,82 dan 0,68. Panjang malai meningkatkan produksi sebesar 0,033 ton, lebar malai meningkatkan produksi sebesar 0,5 ton, dan tinggi tanaman meningkatkan produksi sebesar 0,06 ton dengan menggunakan uji regresi linier. Proses pemilihan variabel independen yang mendominasi (xi) untuk dimasukkan ke dalam model regresi guna menghitung besarnya variabel dependen (y) untuk setiap unit (xi) dikenal sebagai analisis bertahap [7]. Sedangkan dalam penelitian dengan menggunakan metode yang sama yaitu dengan judul model prediksi kadar air media tanam menggunakan regresi linear berganda (studi kasus kebun tomat beef di Serenity farm mitra habibi garden) menghasilkan penelitian yaitu Model prediksi kadar air media tanam berhasil dikembangkan dengan menggunakan algoritma regresi linear berganda. Model dapat menghasilkan nilai prediksi kadar air media tanam 3 jam berikutnya dengan telah divalidasi

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 480−488
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5930

sebanyak dua kali, yaitu pada tahap simulasi dan pengujian model. Diperoleh akurasi model berupa nilai R2 dan nilai RMSE, dimana pada tahap simulasi model sebesar 83,80% dan 1,81%, lalu pada tahap pengujian model sebesar 72,02% dan 1,74%. Akurasi model yang dihasilkan telah memenuhi kriteria model, sehingga membuktikan bahwa algoritma regresi linear berganda relevan digunakan untuk pengembangan model prediksi kadar air media tanam tanaman tomat Beef di Greenhouse Serenity Farm [8].

Seperti pada penelitian sebelumnya, prediksi produksi padi menggunakan metode Grid Serch Cv yang kemudian dilakukan pemodelan menggunakan 6 algoritma. Menggunakan dataset rice Production Prediction on Sumatera Island, hasil pada penelitian tersebut menunjukan bahwa metode Grid Search dapat meningkatkan performa nilai R2 score dari beberapa algoritma, diantaranya Linear Regression, Gradient Boosting, Decision Tree, SVR, Random Forest, dan Bagging dimana K-neighbors memiliki tingkat akurasi terbaik [9].

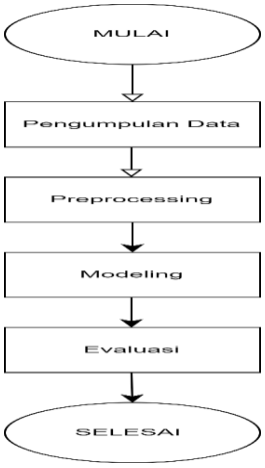
Dalam proses prediksi padi, diperlukan algoritma yang mampu memprediksi data dengan optimal.

Algoritma Linear Regression dan Decision Tree diuji serta dioptimalisasi guna meningkatkan kinerja dan akurasi model. Dengan menitikberatkan pada proses optimalisasi, penelitian ini memberikan wawasan yang lebih mendalam mengenai metode terbaik untuk menanganni masalah prediksi produksi padi

Dengan demikian penelitian ini bertujuan untuk membandingkan Algoritma Linear Regression dan Decision Tree dalam upaya menemukan algoritma yang paling sesuai untuk memprediksi data produksi padi. Sebagai tolok ukur perbandingan antara kedua algoritma tersebut, digunakan parameter R^2-Score dan untuk mendapatkan model terbaik, penelitian ini menerapkan metode Grid Search CV.

2. METODOLOGI PENELITIAN

Dalam Penelitian ini dilakukan beberapa tahapan penelitian yang dimulai dari pengumpulan data, preprocessing, modeling, dan evaluasi. Adapun alur tahapan penelitian dalam penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Alur penelitian

2.1 Pengumpulan dan Analisis Data

Data Produksi Padi Pulau Sumatra yang diperoleh dari platform Kaggle https://www.kaggle.com/datasets/ardikasatria/datasettanamanpadisumatera. Dataset ini memiliki data sebanyak 224 data yang terdiri dari 7 atribut utama yaitu: Provinsi, Tahun, Produksi, Luas Panen, Curah Hujan, Kelembapan, Suhu rata-rata.

	Provinsi	Tahun	Produksi	Luas Panen	Curah hujan	Kelembapan	Suhu rata-rata
0	Aceh	1993	1329536.00	323589.00	1627.0	82.00	26.06
1	Aceh	1994	1299699.00	329041.00	1521.0	82.12	26.92
2	Aceh	1995	1382905.00	339253.00	1476.0	82.72	26.27
3	Aceh	1996	1419128.00	348223.00	1557.0	83.00	26.08
4	Aceh	1997	1368074.00	337561.00	1339.0	82.46	26.31
...	...	...	...	...	...	...	...
219	Lampung	2016	3831923.00	390799.00	2317.6	79.40	26.45
220	Lampung	2017	4090654.00	396559.00	1825.1	77.04	26.36
221	Lampung	2018	2488641.91	511940.93	1385.8	76.05	25.50
222	Lampung	2019	2164089.33	464103.42	1706.4	78.03	27.23
223	Lampung	2020	2604913.29	545149.05	2211.3	75.80	24.58

224 rows × 7 columns

Gambar 2. Dataset

2.2 Preprocessing

Setelah memperoleh data yang akan digunakan, langkah berikutnya adalah melakukan preprocessing data, tahapan

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 480−488
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5930

preprocessing yang digunakan dalam penelitian ini yaitu : Identifikasi missing value dan outlier, seleksi fitur, one hot encoding, split data dan Feature Scalling.

a. Identifikasi Missing Value dan Outlier
   Missing value atau nilai yang hilang adalah istilah dalam analisis data untuk menyebut data yang tidak ada atau tidak diisi dalam dataset [10]. Sedangkan outlier adalah suatu nilai yang jauh berbeda dari sebagian besar data dalam satu set dataset [11].
b. Seleksi Fitur
   Seleksi fitur adalah proses dalam machine learning dan analisis data yang bertujuan untuk memilih subset fitur (variabel) yang paling relevan dan signifikan untuk digunakan dalam membangun model. Tujuan dari seleksi fitur adalah untuk meningkatkan kinerja model, mengurangi overfitting, dan mengurangi kompleksitas model dengan menghilangkan fitur yang tidak penting atau redundan [12].
c. One Hot Encoding
   One-hot encoding adalah metode yang digunakan untuk mengubah data kategori menjadi format numerik yang dapat digunakan oleh algoritma machine learning [13].
d. Featur Scaling
   Feature scaling adalah proses untuk menyesuaikan skala fitur dalam dataset agar memiliki rentang atau distribusi yang seragam, jika fitur tidak di-scale dengan benar, model bisa tidak memanfaatkan informasi dengan baik atau bahkan bekerja kurang optimal [14].

2.3 Modeling

Pada penelitian ini, Algoritma Logistic Regression dan Decision Tree digunakan dalam proses pemodelan dalam memprediksi produksi padi. Logistic Regression sebagai metode klasifikasi analisis sentimen dan menggunakan bahasa pemrograman python [15]. Linear Regresion merupakan alat perlengkapan statistik yang digunakan untuk mengetahui pengaruh antara satu atau beberapa variabel terhadap satu buah variabel. Variabel bebas, variabel independen atau variabel penjelas disebut variabel yang mempengaruhi. Sedangkan variabel terikat atau variabel dependen disebut dengan variabel yang dipengaruhi. Pada skala interval maupun ratio dapat menggunakan regresi [16]. Model ini memproyeksikan hasil ke dalam rentang probabilitas antara 0 dan 1, dan hasil tersebut digunakan untuk menentukan kelas atau kategori dari data yang diprediksi [17].

Decision Tree adalah pengolahan data dimana yang pertama melakukan pengolahan algoritma decision tree secara otomatis untuk decision tree dengan tools software Rapid Miner. Analisis data adalah melihat hasil dari algoritma decision tree, serta di analisa data dengan informasi yang berharga[18]. Decision Tree merupakan sebuah data yang terdiri dari simpul dan rusuk, simpul pada pohon dibedakan menjadi tiga yaitu simpul akar, percabangan dan daun [19].

Hyperparameter tunning digunakan dalam penelitian ini untuk menemukan set hyperparameter yang memberikan kinerja terbaik pada model [20]. Salah satu metode hyperparameter tunnning yaitu metode grid search cv, yaitu teknik mencoba semua kombinasi hyperparameter yang mungkin dalam grid yang telah ditentukan [21].

2.4 Evaluasi

Pada penelitian ini, metode evaluasi yang digunakan yaitu dengan menggunakan Mean Absolute Error, Mean Squared Error, dan R2-score. MAE (Mean Absolute Error) adalah rata-rata dari nilai absolut perbedaan antara nilai yang diprediksi dan nilai yang sebenarnya, memberi ukuran seberapa besar kesalahan dalam prediksi, dengan satuan yang sama seperti data [22]. MSE (Mean Squared Error) adalah rata-rata dari kuadrat perbedaan antara nilai yang diprediksi dan nilai yang sebenarnya, memberikan penalti yang lebih besar untuk kesalahan yang besar dibandingkan dengan MAE. R^2 score mengukur seberapa baik model menjelaskan variasi dalam data target. Ini adalah ukuran yang menunjukkan proporsi variansi dalam data target yang dapat dijelaskan oleh model [23].

3. HASIL DAN PEMBAHASAN

3.1 Preprocessing

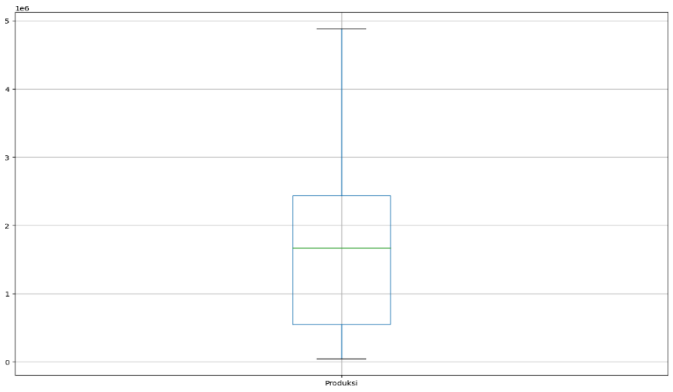
a. Identifikasi Missing Value dan Outlier
   Langkah ini dimulai dengan mendeteksi nilai yang hilang dalam dataset untuk memastikan kualitas data.

	e
Provinsi	0
Tahun	0
Produksi	0
Luas Panen	0
Curah hujan	0
Kelembapan	0
Suhu rata-rata	0

Gambar 3. Missing Value

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 480−488
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5930

Seperti pada gambar 3, pada dataset Padi di Pulau Sumatra tidak terdeteksi adanya nilai yang hilang atau missing value.

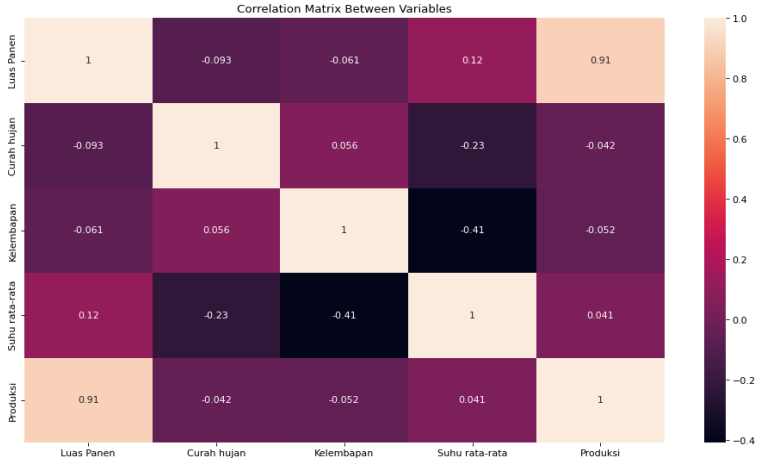


Gambar 4. Outlier

Pada Gambar 4, variabel produksi yang merupakan variabel target atau utama dalam penelitian ini tidak ditemukan adanya outlier pada dataset yang digunakan.

b. Seleksi Fitur

Matriks korelasi mengukur sejauh mana dua variabel berkorelasi atau memiliki hubungan linier satu sama lain. Korelasi dapat berkisar dari -1 hingga 1, dengan interpretasi sebagai berikut:
1: Korelasi positif sempurna. Artinya, jika satu variabel naik, variabel lainnya juga naik secara proporsional
0: Tidak ada korelasi. Variabel tidak memiliki hubungan linier satu sama lain
-1: Korelasi negatif sempurna. Artinya, jika satu variabel naik, variabel lainnya turun secara proporsional



Gambar 5. Heatmap Correlation Diagram

c. One Hot Encoding

Penelitian ini menggunakan teknik one hot encoding untuk mengubah variabel kategori menjadi format numerik agar dapat digunakan dalam proses pemodelan menggunakan algoritma machine learning. Varibel yang diubah merupakan variabel "Provinsi" yang berisi format kategori yang berupa mana beberapa daerah di Pulau Sumatra.

Kabupaten_Ciamis	Kabupaten_Cianjur	Kabupaten_Indramayu	Kabupaten_Karawang	Kabupaten_Kuningan	Kabupaten_Majalengka	Kabupaten_Subang	Kabupaten_Sumedang
0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0

Gambar 6. One Hot Encoding

d. Featur Scaling

Pada penelitian ini, feature scaling digunakan untuk menormalkan atau mengubah skala fitur- fitur dalam dataset sehingga memiliki rentang nilai yang seragam. Fitur atau variabel yang dinormalisasikan mencakup keseluruhan fitur pada dataset Hasil feature scaling menggunakan Standardization akan menghasilkan data yang distribusinya memiliki nilai rata-rata 0, dengan sebagian besar data terdistribusi di sekitar interval -1 sampai 1. Ini adalah indikasi bahwa data telah dinormalisasi berdasarkan standar deviasi fitur. seperti pada gambar 7 berikut:

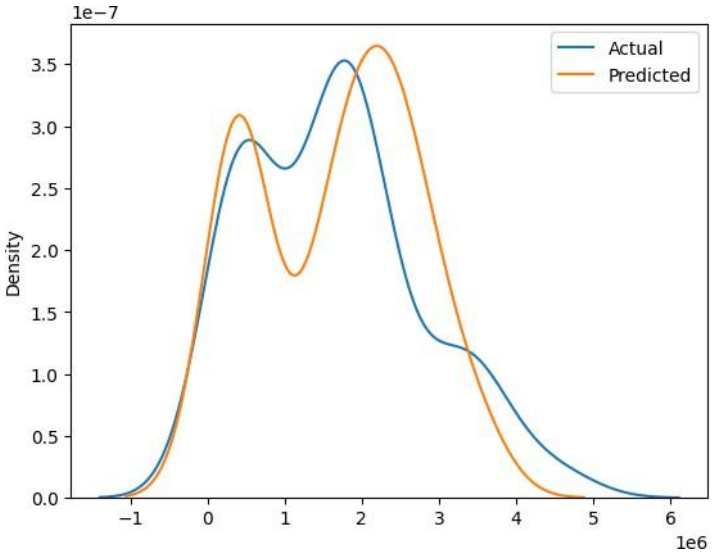
Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 480−488
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5930

```
[[-1.07170771 -0.31302577  0.53326331 -0.34207301 -0.31108551 -0.39056673
  -0.31108551 -0.4152274   2.56038192 -0.4152274  -0.4152274  -0.36514837]
 [ 0.51064077 -0.79631917  0.34793546  1.18107828 -0.31108551 -0.39056673
  -0.31108551 -0.4152274  -0.39056673 -0.4152274   2.40831892 -0.36514837]
 [-1.2012385  -0.26857953  0.77401631 -1.18826817  3.21455025 -0.39056673
  -0.31108551 -0.4152274  -0.39056673 -0.4152274  -0.4152274  -0.36514837]
 [-1.23204632 -0.12562823  0.49169445 -0.30446434 -0.31108551 -0.39056673
  -0.31108551 -0.4152274  -0.39056673 -0.4152274  -0.4152274   2.73861279]
 [-1.11782994  0.78499627  0.58002828 -0.51131204  3.21455025 -0.39056673
  -0.31108551 -0.4152274  -0.39056673 -0.4152274  -0.4152274  -0.36514837]]
[[-1.11263695]
 [ 0.26419173]
 [-1.19391204]
 [-0.88288197]
 [-1.07069477]]
```

Gambar 7. Feature Scaling

3.2 Modeling

a. Algoritma Linear Regression
Pada penelitian ini, pemodelan menggunakan Algoritma Linear Regression menghasilkan pola grafik seperti Gambar 8 di bawah ini:



Gambar 8. Grafik Prediksi Linear Regression

Analisis grafik:
1. Garis biru (Actual): Ini adalah data sebenarnya yang diukur.
2. Garis oranye (Predicted): Ini adalah data yang diprediksi oleh model linear regression.

Pola:
1. Kedua garis menunjukkan pola yang mirip: mulai dari kepadatan nol, naik ke puncak antara 0 dan 1 juta pada sumbu X, kemudian turun kembali.
2. Puncak dari data predicted sedikit lebih tinggi dibandingkan dengan data aktual, menunjukkan ada sedikit perbedaan antara prediksi dan kenyataan.

Grafik Gambar 8 menunjukkan bahwa model Linear Regression cukup baik dalam memprediksi data prediksi, meskipun ada sedikit perbedaan pada puncaknya. Ini bisa menjadi indikasi bahwa model tersebut cukup akurat, tetapi mungkin perlu beberapa penyesuaian untuk meningkatkan presisi prediksi.

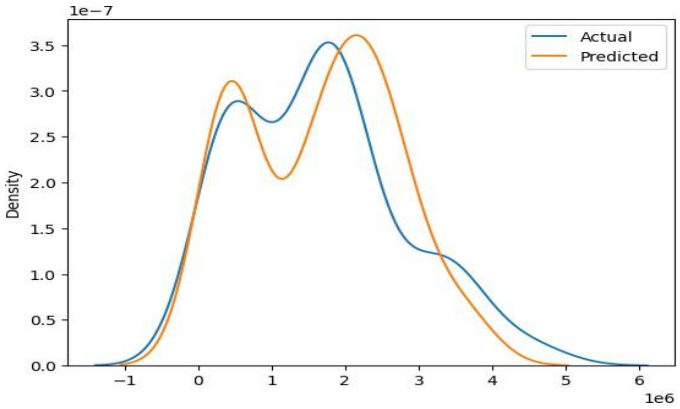
```
Linear Regression
Train :  84.68578496928782
Test  :  86.37276105524145

The Average Cross Validation Score is 82.9
```

Gambar 9. Cross Validation Score Linear Regression

Pada Gambar 9, Train Score menunjukkan seberapa baik model mempelajari data pelatihan. Skor ini cukup tinggi, menunjukkan bahwa model dapat menangkap pola dalam data pelatihan dengan baik. Test Score menunjukkan performa model pada data yang belum pernah dilihat sebelumnya (data uji). Skor ini sedikit lebih tinggi daripada skor pelatihan, yang menunjukkan bahwa model tidak overfitting dan dapat menggeneralisasi dengan baik pada data baru. Average Cross Validation Score adalah rata-rata skor dari beberapa iterasi validasi silang. Skor ini memberikan gambaran tentang seberapa konsisten performa model pada berbagai subset data. Skor ini juga cukup tinggi, menunjukkan bahwa model memiliki performa yang stabil dan dapat diandalkan.

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 480−488
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5930

b. Algoritma Linear Regression dengan Hyperparameter Tunning Grid Search CV
Selanjutnya, metode grid search cv digunakan untuk optimalisasi model linear regression dan menghasilkan grafik seperti pada Gambar 10 berikut:



Gambar 10. Linear Regression dengan Grid Search CV

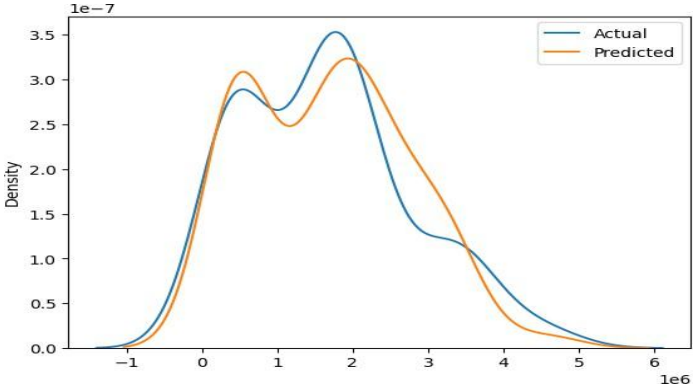
Setelah dilakukan optimalisasi menggunakan grid search cv, terdapar perbedaan utama seperti berikut:
1. Akurasi Prediksi: Setelah dilakukan hyperparameter tuning, prediksi model lebih mendekati data aktual. Ini terlihat dari garis oranye yang lebih dekat dengan garis biru di seluruh rentang grafik.
2. Puncak Data: Puncak dari data prediksi setelah tuning lebih tinggi dan lebih sesuai dengan puncak data aktual dibandingkan dengan grafik sebelumnya.
3. Keseluruhan Pola: Pola keseluruhan dari prediksi lebih halus dan lebih akurat, menunjukkan bahwa model telah dioptimalkan untuk memberikan hasil yang lebih baik.

```
Linear Regression
Train :  84.6508189494205
Test  :  86.89566591106728

The Average Cross Validation Score is 83.0
```

Gambar 11. Linear regression cross Validation Score dengan Grid Search CV

Metode grid search cv tentunya mengubah cross validation score linear regression seperti pada Gambar 11 dengan perbedaan utama dengan hasil sebelum optimalisasi sebagai berikut:
1. Akurasi Prediksi: Setelah dilakukan grid search cv, akurasi prediksi pada data uji meningkat dari 86.37 menjadi 86.90. Ini menunjukkan bahwa model lebih baik dalam memprediksi data baru setelah hyperparameter tuning.
2. Konsistensi model: Average cross validation score meningkat dari 82.9 menjadi 83.0, menunjukkan bahwa model lebih konsisten dalam performanya di berbagai subset data.
3. Performa Pelatihan: Train score sedikit menurun dari 84.69 menjadi 84.65, tetapi perbedaannya sangat kecil dan tidak signifikan. Ini menunjukkan bahwa model tidak overfitting dan tetap mampu mempelajari data pelatihan dengan baik.

c. Algoritma Decision Tree
Pemodelan menggunakan Algoritma Decision Tree menghasilkan pola grafik seperti gambar 12 di bawah ini:



Gambar 12. Grafik Prediksi Decision Tree

Kedua garis ini mengikuti pola yang mirip, dengan puncak di sekitar nilai 1 pada sumbu X dan kemudian menurun tajam. Ini menunjukkan bahwa model decision tree memiliki tingkat akurasi yang baik dalam

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 480−488
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5930

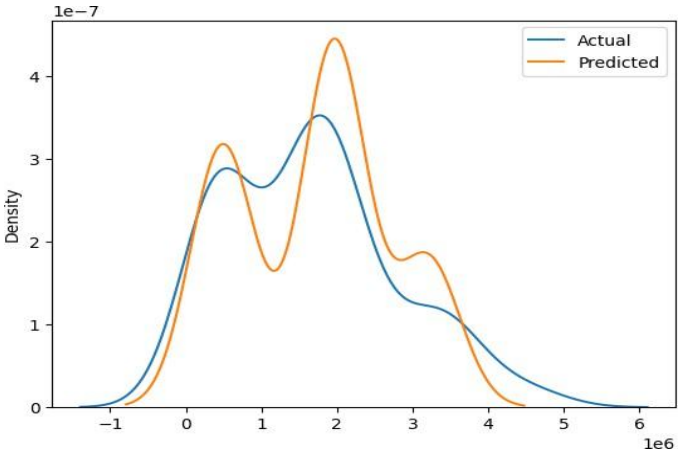
memprediksi hasil yang mendekati data aktual.

```
Decision Tree Regression
Train :  100.0
Test  :  72.48390088402934

The Average Cross Validation Score is 63.25
```

Gambar 13. Cross Validation Score Decision Tree

Pada Gambar 13, Training Score 100.0 menunjukkan bahwa model decision tree memiliki kecocokan sempurna dengan data pelatihan. Artinya, model ini mampu memprediksi data pelatihan dengan akurasi 100%. Test Score 72.48 menunjukkan seberapa baik model dapat menggeneralisasi ke data baru yang tidak terlihat selama pelatihan. Dengan skor sekitar 72.48, model ini memiliki performa yang cukup baik, meskipun tidak sempurna, dalam memprediksi data baru. Average Cross-Validation Score 63.25 memberikan estimasi akurasi yang diharapkan ketika model diterapkan pada data yang belum pernah dilihat sebelumnya. Dengan nilai 63.25, ini menunjukkan bahwa model memiliki kemampuan prediksi yang cukup baik, tetapi ada ruang untuk perbaikan.

d. Algoritma Decision Tree dengan Hyperparameter Tunning Grid Search CV
Selanjutnya, grafik hasil model decision tree dengan hyperparameter grid search cv dapat dilihat pada gambar 14 sebagai berikut:



Gambar 14. Decision Tree dengan Grid Search CV

Pada Gambar 14, berikut beberapa perbedaan dengan grafik sebelumnya:

1. Akurasi Prediksi:
   Sebelum Optimalisasi: Garis prediksi (oranye) mengikuti pola yang mirip dengan garis aktual (biru), tetapi ada beberapa perbedaan yang lebih besar di beberapa titik. Setelah Optimalisasi: Garis prediksi lebih dekat mengikuti garis aktual, menunjukkan peningkatan akurasi prediksi.
2. Puncak Densitas:
   Sebelum Optimalisasi: Puncak densitas terjadi di sekitar nilai 1 pada sumbu X. Setelah Optimalisasi: Puncak densitas terjadi di sekitar nilai 2 pada sumbu X, yang lebih sesuai dengan data aktual.
3. Rentang Densitas:
   Sebelum Optimalisasi: Rentang densitas mencapai hingga 3.5e-7. Setelah Optimalisasi: Rentang densitas mencapai lebih dari 4e-7, menunjukkan bahwa model yang dioptimalkan mampu menangkap variasi data dengan lebih baik. Optimalisasi menggunakan Grid Search CV membantu dalam menemukan kombinasi hyperparameter terbaik untuk model decision tree, yang pada akhirnya meningkatkan akurasi prediksi dan kemampuan model dalam menangkap pola data yang lebih kompleks.

```
Decision Tree Regression
Train :  84.68468800789897
Test  :  81.42164309777542

The Average Cross Validation Score is 78.52
```

Gambar 15. Decision Tree cross Validation Score dengan Grid Search CV

Gambar 15 merupakan hasil evaluasi model prediksi decision tree setelah optimalisasi menggunakan metode grid search cv, berikut beberapa perbedaannya:

1. Training Score:
   Sebelum Optimalisasi: 100.0
   Setelah Optimalisasi: 84.68

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 480−488
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5930

Penurunan: Terjadi penurunan karena model yang dioptimalkan dengan Grid Search CV lebih baik dalam menghindari overfitting. Skor 100.0 sebelumnya menunjukkan bahwa model terlalu cocok dengan data pelatihan, yang sering kali mengindikasikan overfitting.

2. Testing Score:
   Sebelum Optimalisasi: 72.48
   Setelah Optimalisasi: 81.42
   Peningkatan: Skor pengujian meningkat setelah optimalisasi, menunjukkan bahwa model yang dioptimalkan memiliki generalisasi yang lebih baik terhadap data baru.
3. Average Cross-Validation Score:
   Sebelum Optimalisasi: 63.25
   Setelah Optimalisasi: 78.52
   Peningkatan: Skor cross-validation rata-rata meningkat, menunjukkan bahwa model yang dioptimalkan lebih konsisten dalam performanya pada berbagai subset data.
4. Evaluasi
   Pada penelitian ini, evaluasi yang digunakan untuk menganalisis hasil model dari Algoritma Linear Regression dan Decision Tree yaitu dengan R2-Score, Mean Absolute Error dan Mean Squared Error seperti pada gambar 16 di bawah ini:

	R2-score	Mean Absolute Error	Mean Squared Error
Linear Regression	86.895666	261049.168107	160199780301.226318
Decision Tree	81.421643	281254.415963	227119415141.265686

Gambar 16. Evaluasi ke dua Algoritma

a) **R2-score**:
   i. **Linear Regression**: 86.895666
   ii. **Decision Tree**: 81.421643
   iii. **Penjelasan**: R2-score mengukur seberapa baik model menjelaskan variabilitas data. Nilai R2-score yang lebih tinggi menunjukkan model yang lebih baik dalam menjelaskan variasi data. Dalam hal ini, Linear Regression memiliki R2-score yang lebih tinggi dibandingkan Decision Tree, menunjukkan bahwa Linear Regression lebih baik dalam menjelaskan variabilitas data.

b) Mean Absolute Error (MAE):
   i. **Linear Regression**: 261049.168107
   ii. **Decision Tree**: 281254.415963
   iii. **Penjelasan**: MAE mengukur rata-rata kesalahan absolut antara nilai prediksi dan nilai aktual. Nilai MAE yang lebih rendah menunjukkan model yang lebih akurat. Dalam hal ini, Linear Regression memiliki MAE yang lebih rendah dibandingkan Decision Tree, menunjukkan bahwa Linear Regression memiliki kesalahan prediksi yang lebih kecil secara rata-rata.

c) **Mean Squared Error (MSE)**:
   i. **Linear Regression**: 160199780301.226318
   ii. **Decision Tree**: 227119415141.265686
   iii. **Penjelasan**: MSE mengukur rata-rata kesalahan kuadrat antara nilai prediksi dan nilai aktual. Nilai MSE yang lebih rendah menunjukkan model yang lebih akurat. Dalam hal ini, Linear Regression memiliki MSE yang lebih rendah dibandingkan Decision Tree, menunjukkan bahwa Linear Regression memiliki kesalahan prediksi yang lebih kecil secara rata-rata.

4 KESIMPULAN

Pada Algoritma Linear Regression, model yang telah dioptimalkan dengan Grid Search CV menunjukkan peningkatan akurasi yang signifikan, dengan garis prediksi yang hampir sepenuhnya mengikuti garis aktual. Ini menunjukkan bahwa optimalisasi hyperparameter dengan Grid Search CV efektif dalam meningkatkan performa model Linear Regresi. Sedangkan pada Algoritma Decision Tree, optimalisasi dengan Grid Search CV membantu dalam menemukan kombinasi hyperparameter terbaik yang meningkatkan kemampuan generalisasi model. Meskipun training score menurun, peningkatan pada testing score dan average cross-validation score menunjukkan bahwa model yang dioptimalkan lebih baik dalam memprediksi data baru dan lebih konsisten dalam performanya. Linear Regression memiliki performa yang lebih baik dalam hal R2-score, MAE, dan MSE dibandingkan dengan Decision Tree. Ini menunjukkan bahwa Linear Regression lebih baik dalam menjelaskan variabilitas data dan memiliki kesalahan prediksi yang lebih kecil. Decision Tree meskipun memiliki performa yang lebih rendah dibandingkan Linear Regression masih merupakan model yang berguna, terutama jika data memiliki hubungan non-linear yang tidak dapat ditangkap oleh Linear Regression. Maka dapat disimpulkan bahwa Algoritma Linear Regression dengan optimalisasi hyperparameter tunning grid search cv mampu memprediksi produksi padi lebih

Journal of Information System Research (JOSH)
Volume 6, No. 1, Oktober 2024, pp 480−488
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i1.5930

baik dibanding Algoritma Decision Tree dengan nilai R2-score 86.895666, MAE 261049.168107, dan MSE 160199780301.226318.

REFERENCES

[1] S. Syamsiah, R. Nurmalina, and A. Fariyanti, "Analisis Sikap Petani Terhadap Penggunaan Benih Padi Varietas Unggul Di Kabupaten Subang Jawa Barat (Attitude Analysis of Farmers Toward Using Rice Seed High Yielding Varieties in Subang Regency West Java)," J. AGRISE, vol. 16, no. 3, pp. 205–2015, 2015.

[2] B. Satria, E. M. Harahap, and Jamilah, "Peningkatan produktivitas padi sawah (Oryza sativa L.) melalui penerapan beberapa jarak tanam dan sistem tanam," J. Agroteknologi FP USU, vol. 5, no. 3, pp. 629–637, 2017, [Online]. Available: https://talenta.usu.ac.id/joa/article/view/2228

[3] Yennita Sihombing, "Kebijakan Pembangunan Pertanian Berbasis Inovasi Teknologi Sebagai Upaya Peningkatan Produksi Komoditas Pertanian Strategis Dan Pendapatan Petani Mendukung Ketahanan Pangan," Pros. Semin. Nas. Has. Penelit. Agribisnis, pp. 137–143, 2022, [Online]. Available: https://jurnal.unigal.ac.id/index.php/prosiding/article/view/7377

[4] S. Kasus and K. Bengkulu, "Analisis Perbandingan Metode Case Base Reasoning (Cbr) Dan Certainty Factor (Cf) Pada Sistem Pakar Diagnosis Hama Pengganggu Dan," vol. 10, no. 2, pp. 129–141, 2022.

[5] B. P. Statistik, "Produksi Padi Menurut Kabubpaten/Kota (Ton), 2023," jabar.bps, 2024.

[6] A. Anggara, K. Auliasari, and Y. Agus Pranoto, "Metode Regresi Linier Berganda Untuk Prediksi Omset Penyewaan Kamera Di Joe Kamera," JATI (Jurnal Mhs. Tek. Inform., vol. 7, no. 1, pp. 852–858, 2023, doi: 10.36040/jati.v7i1.6158.

[7] S. Suwardi, N. N. Andayani, F. Tabri, and M. Aqil, "Penerapan Model Regresi Bertatar dalam Penentuan Hasil Tanaman Sorgum," Agrotechnology Res. J., vol. 4, no. 2, p. 77, 2020, doi: 10.20961/agrotechresj.v4i2.42901.

[8] S. A. Listina, R. M. Sampurno, D. Ciptaningtyas, and A. Thoriq, "Model Prediksi Kadar Air Media Tanam Menggunakan Regresi Linear Berganda (Studi Kasus Kebun Tomat Beef di Serenity Farm Mitra Habibi Garden)," J. Teknotan, vol. 16, no. 3, p. 133, 2022, doi: 10.24198/jt.vol16n3.1.

[9] D. Wulandari and R. Rumini, "Pemodelan dan Prediksi Produksi Padi Menggunakan Regresi Linear," Smart Comp Jurnalnya Orang Pint. Komput., vol. 12, no. 4, 2023, doi: 10.30591/smartcomp.v12i4.5905.

[10] A. C. Darmawan, "Pengembanga Aplikasi Berbasis Web dengan Python Flask untuk Klasifikasi Data Menggunakan Metode Decision Tree C4.5," J. Pendiidikan Konseling, vol. 4, no. 5, pp. 5351–5362, 2022.

[11] Ξ. Ακρίτας, "Intrusion Detection System με αλγόριθμους μηχανικής μάθησης," pp. 1–5, 2021, [Online]. Available: http://ikee.lib.auth.gr/record/335344

[12] H. Tantyoko, D. K. Sari, and A. R. Wijaya, "Prediksi Potensial Gempa Bumi Indonesia Menggunakan Metode Random Forest Dan Feature Selection," IDEALIS Indones. J. Inf. Syst., vol. 6, no. 2, pp. 83–89, 2023, doi: 10.36080/idealis.v6i2.3036.

[13] C. Herdian, A. Kamila, and I. G. Agung Musa Budidarma, "Studi Kasus Feature Engineering Untuk Data Teks: Perbandingan Label Encoding dan One-Hot Encoding Pada Metode Linear Regresi," Technol. J. Ilm., vol. 15, no. 1, p. 93, 2024, doi: 10.31602/tji.v15i1.13457.

[14] I. Setiawan, R. Fina Antika Cahyani, and I. Sadida, "Exploring Complex Decision Trees: Unveiling Data Patterns and Optimal Predictive Power," J. Innov. Futur. Technol., vol. 5, no. 2, pp. 112–123, 2023, doi: 10.47080/iftech.v5i2.2829.

[15] M. Raja Nurhusen, J. Indra, and K. Ahmad Baihaqi, "Analisis Sentimen Pengguna Twitter Terhadap Kenaikan Harga Bahan Bakar Minyak (BBM) Menggunakan Metode Logistic Regression," J. Media Inform. Budidarma, vol. 7, no. 1, pp. 276–282, 2023, doi: 10.30865/mib.v7i1.5491.

[16] A. Hidayanti, A. M. Siregar, S. A. P. Lestari, and Y. C. Cahyana, "Model Analisis Kasus Covid-19 Di Indonesia Menggunakan Algoritma Regresi Linier Dan Random Forest," Petir, vol. 15, no. 1, pp. 91–101, 2021, doi: 10.33322/petir.v15i1.1487.

[17] A. Maulana, M. Martanto, and I. Ali, "Prediksi Hasil Produksi Panen Bawang Merah Menggunakan Metode Regresi Linier Sederhana," JATI (Jurnal Mhs. Tek. Inform., vol. 7, no. 4, pp. 2884–2888, 2024, doi: 10.36040/jati.v7i4.7281.

[18] A. M. Siregar and A. Fauzi, "Klasifikasi Kab Kota Provinsi Jawa Barat Berdasarkan Pendapatan Dari Sektor Pertanian Dengan Algoritma Decision Tree," Fakt. Exacta, vol. 13, no. 1, p. 1, 2020, doi: 10.30998/faktorexacta.v13i1.5542.

[19] S. A. Pratiwi, A. Fauzi, S. Arum, P. Lestari, and Y. Cahyana, "KLIK: Kajian Ilmiah Informatika dan Komputer Prediksi Persediaan Obat Pada Apotek Menggunakan Algoritma Decision Tree," Media Online, vol. 4, no. 4, pp. 2381–2388, 2024, doi: 10.30865/klik.v4i4.1681.

[20] D. A. Putri, S. Samijo, and ..., "Inovasi Pengembangan Produk Media Pembelajaran Interaktif BELMARU Berbasis Google Sites dalam Pembelajaran Matematika," … dan Pembelajaran, pp. 387–394, 2022, [Online]. Available: https://proceeding.unpkediri.ac.id/index.php/seinkesjar/article/view/3049%0Ahttps://proceeding.unpkediri.ac.id/index.php/seinkesjar/article/download/3049/2118

[21] A. Toha, P. Purwono, and W. Gata, "Model Prediksi Kualitas Udara dengan Support Vector Machines dengan Optimasi Hyperparameter GridSearch CV," Bul. Ilm. Sarj. Tek. Elektro, vol. 4, no. 1, pp. 12–21, 2022, doi: 10.12928/biste.v4i1.6079.

[22] T. A. E. Putri, T. Widiharih, and R. Santoso, "Penerapan Tuning Hyperparameter Randomsearchcv Pada Adaptive Boosting Untuk Prediksi Kelangsungan Hidup Pasien Gagal Jantung," J. Gaussian, vol. 11, no. 3, pp. 397–406, 2023, doi: 10.14710/j.gauss.11.3.397-406.

[23] J. Ekonomika, M. Dan, B. Jemb, E. Mariyanti, and R. Nasrah, "Burnout Sebagai Mediator Dalam Pengaruh Beban Kerja Terhadap Kinerja Karyawan," vol. 3, no. 2, pp. 224–231, 2024.