PERBANDINGAN METODE LINEAR REGRESSION, RANDOM FOREST & K-NEAREST NEIGHBOR UNTUK PREDIKSI PRODUKSI HASIL PANEN PADI DI PROVINSI JAWA BARAT

Jonatan Hutahaean, Dadang Yusup, Purwantoro
Informatika, Universitas Singaperbangsa Karawang
Jalan HS. Ronggo Waluyo, Karawang, 41361; (0267) 641177
2010631170149@student.unsika.ac.id

ABSTRAK

Padi merupakan bahan utama beras yang termasuk salah satu bahan pokok terbesar di dunia. Indonesia berada di peringkat ketiga sebagai produsen beras terbesar di dunia, setelah Tiongkok dan India. Provinsi Jawa Barat secara konsisten berada dalam tiga besar penghasil padi nasional. Meskipun selalu berada dalam tiga besar, produksi padi di Jawa Barat mengalami penurunan setiap tahun. Produksi padi sangat berpengaruh terhadap pemenuhan kebutuhan pangan pokok berupa beras, karena permintaan beras meningkat setiap tahun. Produksi padi dapat meningkat atau menurun karena beberapa faktor, seperti luas areal tanam, luas panen, dan produktivitas padi. Pada penelitian ini, digunakan teknik regresi untuk memprediksi produksi hasil panen padi di Provinsi Jawa Barat menggunakan metode *linear regression, random forest, dan k-nearest neighbor*. Perbandingan yang dihasilkan dilihat dari *R2-score, mean absolute error,* dan *mean squared error.* Pada perbandingan metode, digunakan skenario yang sama, yaitu *cross-validation 10-fold*. Hasil uji coba dengan menggunakan skenario yang sama menunjukkan bahwa ketiga metode dapat digunakan untuk memprediksi produksi hasil panen padi di Provinsi Jawa Barat. Kesimpulan dari penelitian ini menujukkan bahwa metode *linear regression* memiliki performa lebih baik daripada metode *random forest* dan *k-nearest neighbor* dengan *R2-score* 98,33, *MAE* 27746,46, dan *MSE* 1688264771,87.

Kata kunci : *linear regression, random forest, k-nearest neighbor, panen padi*

1. PENDAHULUAN

Padi merupakan bahan utama beras yang termasuk salah satu bahan pokok terbesar di dunia [1]. Saat ini, produksi padi dunia menempati peringkat ketiga setelah jagung dan gandum. Hal ini berlaku di Benua Asia, yang menjadi tempat tinggal petani yang menghasilkan sekitar 90% dari total produksi beras dunia [2]. Indonesia berada di peringkat ketiga sebagai produsen beras terbesar di dunia, setelah Tiongkok dan India. Provinsi Jawa Barat secara konsisten berada dalam tiga besar penghasil padi nasional. Pada tahun 2021, Provinsi Jawa Barat mencatat produksi padi sebesar 9.113 ton [3]. Meskipun selalu berada dalam tiga besar, produksi padi di Jawa Barat menalami penurunan setiap tahun. Menurut data Open Data Jabar, pada tahun 2015 Jawa Barat menghasilkan 10.856.438 ton padi, tahun 2016 sebanyak 12.031.508 ton, tahun 2017 sebanyak 11.849.636 ton, tahun 2018 sebanyak 11.073.462 ton, tahun 2019 sebanyak 13.200.114 ton, dan pada tahun 2020 sebesar 10.156.939 ton.

Produksi Padi sangat berpengaruh terhadap pemenuhan kebutuhan pangan pokok berupa beras, karena permintaan beras meningkat setiap tahun [1]. Produksi padi dapat meningkat atau menurun karena beberapa faktor, seperti luas areal tanam, luas panen, dan produktivitas padi. Dengan adanya prediksi panen padi di Provinsi Jawa Barat, diharapkan dapat mengetahui gambaran mengenai hasil panen pada tahun-tahun mendatang.

Penelitian sebelumnya telah mengungkapkan berbagai faktor yang mempengaruhi produktivitas padi, seperti hasil produksi, luas panen, luas areal tanam, curah hujan, dan jumlah hari hujan [4].

Penelitian lain mengungkapkan bahwa hasil uji simultan model regresi menunjukkan tiga variabel utama, yaitu tanggal, komoditas, dan pasar, dengan variabel yang diprediksi adalah harga [5].

Penelitian sebelumnya menguji metode regresi liner dengan menggunakan MAPE untuk memprediksi jumlah produksi lima jenis sayuran, yaitu cabai rawit, kangkung, bawang merah, terong, dan tomat [6].

Penlitian lain berfokus pada penggunaan regresi linear dengan delapan variabel independen dan delapan variabel dependen [7].

Penelitian sebelumnya membandingkan algoritma *CART* dan *k-nearest neighbor* untuk memprediksi luas lahan panen padi di Kabupaten Karawang, menggunakan evaluasi dengan *correlation coefficient, mean absolute error,* dan *root mean squared error* [8].

Penelitian ini diharapkan dapat memberikan rekomendasi metode untuk menangani data dengan target numerik atau regresi, terutama untuk prediksi. Metode yang direkomendasikan telah melalui seleksi pengujian dengan teknik klasifikasi dan regresi.

2. TINJAUAN PUSTAKA

2.1. Prediksi

Prediksi adalah upaya untuk mengantisipasi atau memproyeksikan kejadian yang akan terjadi di masa yang akan datang dengan mempertimbangkan informasi yang relevan dari masa lampau melalui pendekatan ilmiah. Tujuannya adalah untuk memperoleh pemahaman tentang apa yang mungkin

terjadi di masa depan dengan tingkat kepastian tertentu. Proses prediksi bisa bersifat kualitatif, melibatkan pendapat para ahli, atau bersifat kuantitatif, menggunakan metode perhitungan matematis. Salah satu metode prediksi kuantitatif yang umum adalah melalui analisis deret waktu [9].

2.2. Produksi

Produksi merupakan kegiatan yang bertujuan untuk menghasilkan nilai manfaat saat ini dan di masa depan. Selain itu, produksi diartikan sebagai proses konversi dari *input* menjadi *output*. Dengan demikian, semua jenis *input* yang terlibat dalam proses produksi untuk menciptakan *output* dikenal sebagai faktor-faktor produksi [10].

2.3. Padi

Padi adalah tanaman pangan primer di Indonesia, utamanya karena mayoritas penduduknya menjadikan beras sebagai makanan pokok sehari-hari. Sebagai komoditas pangan, padi memainkan peran vital dalam perekonomian masyarakat indonesia dengan memasok kebutuhan karbohidrat yang mengenyangkan untuk konsumsi sehari-hari [11].

2.4. *Linear Regression*

Metode prediksi *linear regression* adalah teknik prediksi yang memanfaatkan garis lurus untuk mengekspresikan hubungan antara dua variabel atau lebih [12]. *Linear regression* mengadopsi pendekatan yang sederhana dengan mengasumsikan bahwa hubungan antara dua variabel dapat dijelaskan oleh sebuah rais lurus dengan rumus [13].

2.5. *Random Forest*

Random forest adalah algoritma machine learning yang mengikuti konsep supervised dalam pembentukan kelas pengklasifikasi. Algoritma ini menggabungkan hasil prediksi dari beberapa pohon keputusan [14].

2.6. *K-Nearest Neighbor*

K-nearest Neighbor (KNN) merupakan bagian dari kategori pembelajaran *instance-based*. Algoritma ini juga merupakan salah satu teknik *lazy learning*. *KNN* dilakukan dengan mencari kelompok k objek dalam data pelatihan yang paling mirip dengan objek dalam data baru atau pengujian. Untuk menjalankan proses ini, diperlukan sebuah sistem klasifikasi yang dapat mencari informasi [15].

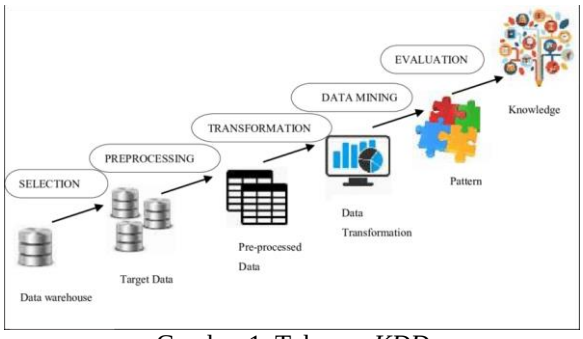
2.7. Google Colab

Google colab merupakan aplikasi yang dikeluarkan oleh Google Research. Google colab memudahkan penulisan dan eksekusi kode Python melalui *browser* web tanpa batasan, memberikan keuntungan yang signifikan dalam bidang *machine learning*, analisis data, dan pendidikan [16].

Google colab, yang dikenal pula sebagai Google Colaboratoru, merupakan lingkungan pengembangan terpadu (IDE) untuk bahasa pemrograman Python. Platform ini memanfaatkan server Google dengan spesifikasi perangkat keras yang canggih untuk melakukan komputasi [17].

3. METODE PENELITIAN

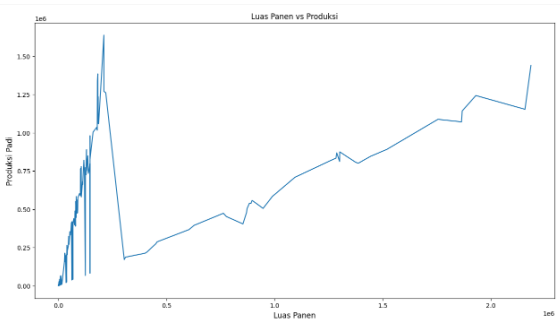
Metode yang dipakai dalam penelitian yaitu *Knowledge Discovery in Database (KDD)*. *KDD* merupakan suatu proses yang bertujuan untuk mengungkap dan menganalisis informasi yang terkandung dalam data yang sangat besar. Tindakan ini dilakukan dengan tujuan memperoleh pengetahuan yang bernilai. Langkah-langkah penelitian yang menerapkan metode *KDD* tersebut terlihat dalam Gambar 1 di bawah ini.



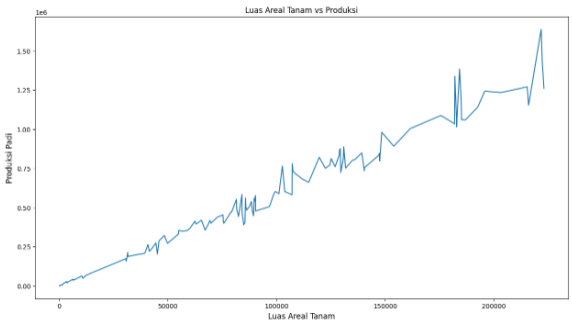
Gambar 1. Tahapan *KDD*

3.1. *Dataset*

Data untuk penelitian ini diperoleh dari Open Data Jabar. Data yang digunakan dalam penelitian ini mencakup jumlah produksi padi, luas areal tanam, luas panen, dan produktivitas yang diambil dari rentang tahun 2015 hingga 2020. sebanyak 162 data berhasil dikumpulkan dari sumber tersebut. Gambar 2 hingga Gambar 6 menjelaskan tentang hubungan antara fitur-fitur yang digunakan. Dalam Gambar 2 dan Gambar 3 terlihat bahwa kedua fitur tersebut memiliki hubungan yang linier.

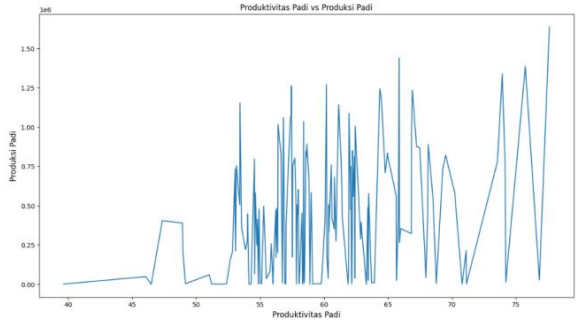


Gambar 2. Hubungan antara produksi padi dengan luas panen

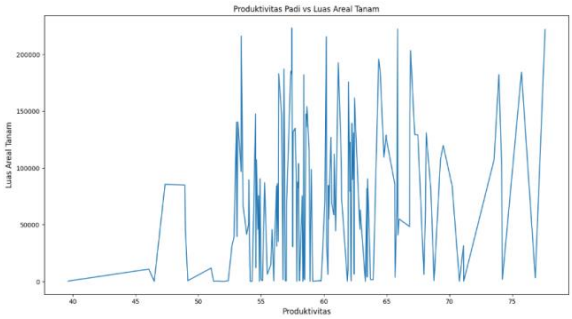


Gambar 3. Hubungan antara produksi padi dengan luas areal tanam

Dalam Gambar 4 dan Gambar 5 terlihat hubungan antara jumlah produktivitas dengan produksi padi dan luas areal tanam.

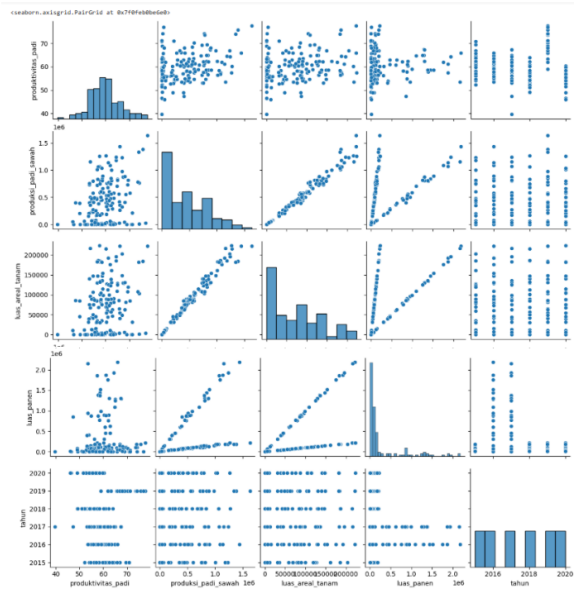


Gambar 4. Hubungan antara produktivitas padi dengan produksi padi



Gambar 5. Hubungan antara produktivitas dengan luas areal tanam

Dalam Gambar 6 menampilkan hubungan antara produksi padi, produktivitas, luas panen, dan luas areal tanam.



Gambar 6. Hubungan antara produksi padi, produktivitas, luas panen, dan luas areal tanam

3.2. *Selection*

Dalam tahap seleksi, penulis melakukan pemilihan data yang relevan. Kemudian, data tersebut diadaptasi dengan keperluan penelitian. Berikut ini adalah contoh data yang dipilih dan terdokumentasikan dalam Gambar 7.

	produktivitas_padi	produksi_padi_sawah	luas_areal_tanam	luas_panen
0	63.42	488926	81723	77088
1	60.56	760668	126991	125611
2	61.34	772705	124583	125971
3	60.36	472912	79392	78345
4	62.09	749960	122509	120789
...	...	...	...	...
157	51.52	2278	470	442
158	58.32	494	81	85
159	54.13	1139	208	210
160	46.06	48676	10928	10569
161	55.49	35762	6450	6444

Gambar 7. Data yang sudah diseleksi

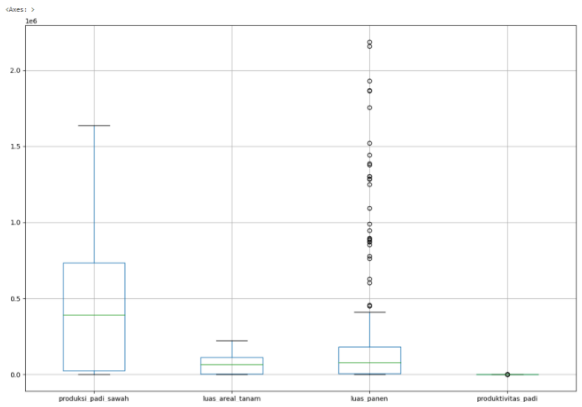
3.3. *Preprocessing*

Dalam tahap *preprocessing*, dilakukan pengecekan terhadap nilai yang hilang dalam data. Setelah dilakukan pengecekan, sebagaimana yang terlihat dalam Gambar 8, tidak ditemukan adanya data yang mengandung nilai yang hilang, sehingga tidak perlu dilakukan pengisian nilai pada data yang hilang.

```
produktivitas_padi     0
produksi_padi_sawah    0
luas_areal_tanam       0
luas_panen             0
dtype: int64
```

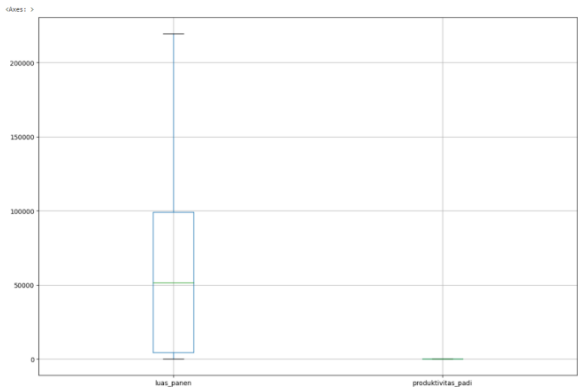
Gambar 8. Pengecekan *missing value*

Langkah selanjutnya yaitu proses pencarian *outlier* memakai *boxplot.* Pada Gambar 9 bisa dilihat terdapat *outlier* pada luas panen dan produktivitas padi.



Gambar 9. *Boxplot* data *outlier*

Pada Gambar 10, terlihat bahwa *outlier* pada luas panen dan produktivitas padi telah ditangani menggunakan metode *IQR*. *IQR (Interquartile Range)* berfungsi untuk mengukur tentang rentang antara kuartil pertama dan kuartil ketiga dalam himpunan data.



Gambar 10. *Boxplot* data tanpa *outlier*

3.4. *Transformation*

Langkah berikutnya adalah mengubah fitur dengan tipe data kategorikal seperti nama_kabupaten_kota. Fitur nama_kabupaten_kota, yang merupakan data dengan tipe data kategorikal, akan diubah menggunakan metode *one-hot encoding. One-hot encoding* merupakan salah satu teknik sederhana yang digunakan pada data kategorikal. Dalam metode ini, data kategorikal dalam bentuk himpunan akan diubah menjadi variabel acak. Proses mengubah fitur dapat dilihat dalam Gambar 11.

```
[ ]  #one hot encoding
     from sklearn.preprocessing import OneHotEncoder
     string_feat = ['nama_kabupaten_kota']
     ohe = OneHotEncoder()
     ohe.fit(new_df[string_feat])
     data_ohe_res = pd.DataFrame(ohe.transform(new_df[string_feat]).toarray(),
                                 columns=ohe.get_feature_names_out())
     new_df = pd.concat([new_df, data_ohe_res], axis=1)
     new_df = df.drop(columns=string_feat)
     new_df.head()
```

Gambar 11. Proses mengubah fitur

3.5. *Data Mining*

Pada tahap ini, dilakukan proses pemodelan dengan menggunakan metode *linear regression, random forest, dan k-nearest neighbor*, dengan hasil seperti yang ditunjukkan pada Gambar 12 hingga Gambar 14.

```
from sklearn.linear_model import LinearRegression
LinReg_model = LinearRegression(copy_X= True, fit_intercept= False, n_jobs= 1, positive= True)
LinReg_model.fit(x_train, y_train)
```

```
LinearRegression
LinearRegression(fit_intercept=False, n_jobs=1, positive=True)
```

Gambar 12. Pemodelan metode *linear regression*

```
from sklearn.ensemble import RandomForestRegressor
RFReg_model = RandomForestRegressor(n_estimators=227, max_depth=9, min_samples_split=2, min_samples_leaf=1, random_state=0)
RFReg_model.fit(x_train, y_train.ravel())
```

```
RandomForestRegressor
RandomForestRegressor(max_depth=9, n_estimators=227, random_state=0)
```

Gambar 13. Pemodelan metode *random forest*

```
from sklearn.neighbors import KNeighborsRegressor
KNNReg_model = KNeighborsRegressor(n_neighbors= 3)
KNNReg_model.fit(x_train, y_train)
```

```
KNeighborsRegressor
KNeighborsRegressor(n_neighbors=3)
```

Gambar 14. Pemodelan metode *k-nearest neighbor*

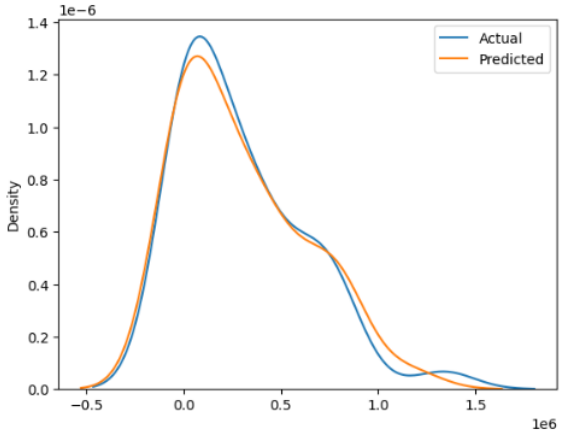
3.6. *Evaluation*

Dalam melakukan evaluasi terhadap model yang telah dibuat kami menggunakan beberapa metode seperti *R2-Score, Mean Absolute Error (MAE),* dan *Mean Squared Error (MSE).*

4. HASIL DAN PEMBAHASAN

4.1. *Linear Regression*

Linear regression adalah algoritma pemodelan regresi yang digunakan untuk memprediksi nilai suatu variabel berdasarkan nilai variabel lain. Visualisasi prediksi hasil pemodelan *linear regression* ditampilkan dalam Gambar 15.



Gambar 15. Visualisasi prediksi hasil pemodelan *linear regression*

Pada evaluasi metode *linear regression* dilakukan satu kali dengan menggunakan *cross*

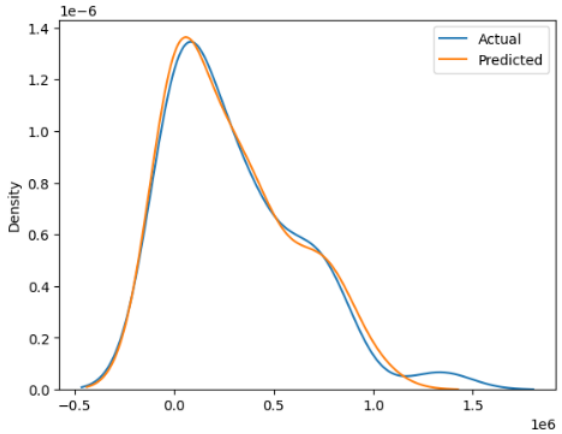
validation folds 10. Hasil dari evaluasi tersebut tersaji dalam Tabel 1.

Tabel 1. Hasil evaluasi metode *linear regression*

R2-score	MAE	MSE
98,33	27746,86	1688264771,87

4.2. *Random Forest*

Random forest adalah algoritma pemodelan regresi yang digunakan untuk memprediksi nilai-nilai berkelanjutan. Visualisasi prediksi hasil pemodelan *random forest* ditampilkan dalam Gambar 16.



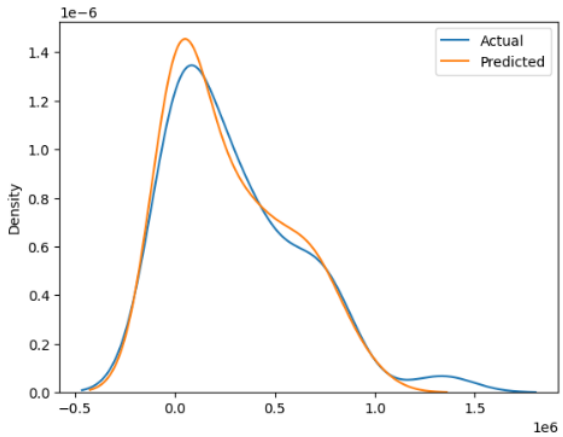
Gambar 16. Visualisasi prediksi hasil pemodelan *random forest*

Untuk evaluasi metode *random forest hanya* dilakukan satu kali dengan menggunakan *cross validation folds* 10. Hasil dari evaluasi tersebut tersaji dalam Tabel 2.

Tabel 2. Hasil evaluasi metode *random forest*

R2-score	MAE	MSE
96,323	29505,09	3775244869,43

4.3. *K-Nearest Neighbor*



Gambar 17. Visualisasi prediksi hasil pemodelan *KNN*

K-Nearest Neighbor (KNN) adalah algoritma yang membangun model regresi dengan memanfaatkan rata-rata atau median dari k tetangga terdekat untuk memprediksi nilai target. Visualisasi prediksi hasil pemodelan *KNN* ditampilkan dalam Gambar 17.

Pada evaluasi metode *K-Nearest Neighbor (KNN)* dilakukan satu kali dengan *cross validation folds* 10. hasil dari evaluasi tersebut tersaji dalam Tabel 3.

Tabel 3. Hasil evaluasi metode *KNN*

R2-score	MAE	MSE
94,01	41549,62	6052106475,86

5. KESIMPULAN DAN SARAN

Kesimpulan yang diperoleh dari peneitian ini adalah dari hasil perbandingan metode *linear regression, random forest,* dan *k-nearest neighbor* untuk prediksi produksi hasil panen padi di Provinsi Jawa Barat menyatakan bahwa metode *linear regression* dengan *cross validation folds* 10 memiliki performa lebih baik dari metode *random forest dan k-nearest neighbor* dilihat dari hasil *R2-score* sebesar 98,33, *mean absolute error* sebesar 27746,86, dan *mean squared error* sebesar 1688264771,87.

Adapun saran untuk penelitian selanjutnya adalah dengan menggunakan metode selain *linear regression, random forest,* dan *k-nearest neighbor*, seperti *support vector machines, neural networks,* dan lainnya. Dengan demikian, dapat dilakukan perbandingan hasil untuk mendapatkan evaluasi yang lebih optimal.

DAFTAR PUSTAKA

[1] H. W. Herwanto, T. Widiyaningtyas and P. Indriana, "Penerapan Algoritme *Linear Regression* untuk Prediksi Hasil Panen Tanaman Padi," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 8, no. 4, 2023.

[2] E. E. Pratiwi, A. W. Widodo and W. F. Mahmudy, "Penerapan Algoritme Genetika pada Kasus Optimasi Penentuan Bibit dan Pemerataan Subsidi Pupuk (Studi Kasus: Desa Pandansari, Kabupaten Kediri)," *Jurnal Pengembang Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 5, 2017.

[3] S. Setiawati, "10 Lumbung Padi Nasional, Jawa Sing Ada Lawan," *Cnbcindonesia.com*, Mei. 9, 2023.

[4] T. N. Padilah and R. I. Adam, "Analisis Regresi Linier Berganda Dalam Estimasi Produktivitas Tanaman Padi Di Kabupaten Karawang," *FIBONACCI Jurnal Pendidikan Matematika dan Matematika*, vol. 5, no. 2, 2019.

[5] K. Puteri and A. Silvanie, "Machine Learning Untuk Model Prediksi Harga Sembako Dengan Metode Regresi Linier Berganda," *Jurnal Nasional Informatika*, vol. 1, no. 2, pp. 82-94, 2020.

[6] A. A. Basahona, R. Ishak and A. Husna, "Penerapan Metode Linier Regresi Untuk Prediksi Produksi Sayur-Sayuran," *Jurnal Nasional Cosphi,* vol. 3, no. 2, pp. 54-57, 2019.

[7] N. Ariyani and A. Z. Arifin, "Prediksi Tingkat Pengangguran di Kabupaten Tuban Tahun 2020 Menggunakan Metode Regresi Linier Sederhana," *MathVision: Jurnal Matematika,* vol. 3, no. 1, pp. 6-13, 2021.

[8] M. F. Aziz, S. Defiyanti and B. N. Sari, "Perbandingan Algoritma CART dan K-Nearest Neighbor Untuk Prediksi Luas Lahan Panen Tanaman Padi di Kabupaten Karawang," *Jurnal TAM (Technology Acceptance Model),* vol. 9, no. 2, 2018.

[9] N. Nurmahaludin, "Analisis Perbandingan Metode Jaringan Syaraf Tiruan Dan Regresi Linear Berganda Pada Prakiraan Cuaca," *Jurnal INTEKNA,* vol. 14, no. 2, 2014.

[10] M. E. Nasution, "Pengenalan Eksklusif Ekonomi Islam," *Kencana Prenada Media Grup,* Jakarta, 2006.

[11] H. W. Herwanto, T. Widiyaningtyas and P. Indriana, "Penerapan Algoritme Linear Regression untuk Prediksi Hasil Panen Tanaman Padi," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi,* vol. 8, no. 4, 2019.

[12] P. Sulardi, T. Hendro and F. R. Umbara, "Prediksi Kebutuhan Obat Menggunakan Regresi Linier," *Prosiding Seminar Nasional Teknologi dan Informatika,* 2017.

[13] R. Zunaidhi, W. S. J. Saputra and N. K. Sari, "Aplikasi Peramalan Penjualan Menggunakan Metode Regresi Linier," *SCAN,* vol. 7, no. 3, 2012.

[14] S. Saadah and H. Salsabila, "Prediksi Harga Bitcoin Menggunakan Metode Random Forest (Studi Kasus: Data Acak Pada Awal Masa Pandemic Covid-19)," *Jurnal Komputer Terapan,* vol. 7, no. 1, pp. 24-32, 2021.

[15] A. M. Argina, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Diabetes," *Indonesian Journal of Data and Science,* vol. 1, no. 2, pp. 29-33, 2020.

[16] R. G. Guntara, "Pemanfaatan Google Colab Untuk Aplikasi Pendeteksian Masker Wajah Menggunakan Algoritma Deep Learning YOLOv7," *Jurnal Teknologi dan Sistem Informasi Bisnis,* vol. 5, no. 1, pp. 55-60, 2023.

[17] G. I. E. Soen, M. and R. , "Implementasi Cloud Computing dengan Google Colaboratory Pada Aplikasi Pengolah Data Zoom Participants," *JITU : Journal Informatic Technology And Communicatione,* vol. 6, no. 1, pp. 24-30, 2022.