Multi-Output Regression untuk Melakukan Prediksi Luas Wilayah, Kualitas Padi dan Produksi Padi pada Pulau Jawa

Felix Indra Kurniadi[1], Darmawan Satyananda[2], Elly Santika[3], Pramitha Dwi Larasati[4]

[1]School of Computer Science, Universitas Bina Nusantara
[2] Jurusan Matematika, Universitas Negeri Malang
[3]Teknik Informatika, Politeknik Negeri Jember
[4] School of Engineering and Technology, Universitas Tanri Abeng
felix.kurniadi@binus.edu[1], darmawan.satyananda.fmipa@um.ac.id [2], elly_antika@polije.ac.id[3], pramitha.dwi@tau.ac.id[4]

Diterima : 20 Februari 2022
Disetujui 27 Maret 2022

***Abstract*—Beras merupakan salah satu makanan pokok di Indonesia. Berdasarkan data dari Badan Pusat Statistik (BPS), konsumsi beras pada tahun 2015 dan 2017 sekitar 29'178.94 -ribu- ton dan 29'133.51-ribu ton. Sayangnya produksi beras pada tahun 2018 hanya mencapati 81.31 juta ton. Pada artikel ilmiah ini, kami membandingkan beberapa metode regresi multi-output seperti Regression Chain, Multi-output linear regression dan Random Forest. Data yang digunakan adalah data beras pada Pulau Jawa terutama Jawa Barat, Jawa Tengah dan Jawa Timur yang diambil dari tahun 2017-2020. Pada penelitian ini kami menggunakan 2 variabel bebas yang dikategorikan yaitu luas lahan dengan kota da, dan yang menjadi variable terikat adalah produktivitas dan Jumlah Produksi. Pemilihan kedua variable bebas ini dikarenakan kedua variable ini sangat mempengaruhi hasil produk dan produktivitas dalam mengembangkan produksi beras. Kami menggunakan pendekatan menggunakan outlier removal dan tidak menggunakan outlier removal. Hasil yang didapatkan dari kedua pendekatan ini, ditemukan bahwa outlier removal pada data yang dimiliki sangat diperlukan terutama untuk mengurangi kesalahan pada hasil tiap metode yang diusulkan.**

***Keywords*— multi-output, regressor chain, linear regression, random forest**

I. PENDAHULUAN

Nasi merupakan salah satu makanan sehari-hari yang dikonsumsi oleh negara-negara di Asia, salah satunya adalah negara Indonesia. Berdasarkan data yang didapatkan dari data BPS mengenai konsumsi beras pada tahun 2015 dan 2017, diperkirakan bahwa sekitar 29'178.94 ribu ton dan 29'133.51 ribu ton dikonsumsi [1].

Kebutuhan konsumsi ini juga diimbangi dengan produksi beras di Indonesia. Berdasarkan data yang didapatkan dari BPS dari tahun 2011 sampai 2017, terdapat peningkatan produksi beras yang awalnya berjumlah 65.75 juta ton menjadi 81.38 juta ton di tahun 2017[2]. Walaupun jumlah produksi yang besar akan tetapi Indonesia tetap melakukan import beras pada negara lain [2]. Hal ini menjadi pertimbangan dimana jika terdapat eksport dan import beras di Indonesia, maka diperlukan system yang dapat memprediksi kebutuhan lahan produksi beras untuk memenuhi kebutuhan. Sayangnya, system prediksi yang ada saat ini belum mencukupi dikarenakan output yang dikeluarkan hanyalah jumlah produksi dari berbagai fitur yang mendukung. Sedangkan, dalam kenyataan output jumlah produksi saja belum cukup dalam menyelesaikan masalah, ada factor seperti penentuan kualitas produksi yang perlu diprediksi juga. Hal inilah yang membuat

sebuah single-output regression menjadi tidak valid dan dibutuhkan sebuah system prediksi yang bernilai multi-output atau multi-target.

Konsep multi-output atau multi-target merupakan konsep dimana memiliki sebuah tujuan untuk secara bersamaan melakukan prediksi beberapa output yang memiliki real value[3]. Beberapa penelitian pada bidang prediksi panen padi berdasarkan beberapa pendekatan dengan menggunakan konsep data mining pernah dilakukan oleh [2], [4], dan [5]; Penelitian dengan pendekatan remote sensing juga pernah dilakukan oleh [6], [7] dan [8]. Sayangnya, dari beberapa penelitian yang ditela'ah, kebanyakan penelitian melihat dari multi feature dengan single output sedangkan ada beberapa factor yang perlu dipertimbangkan seperti produktivitas.

Beberapa penelitian diluar prediksi panen padi sudah melakukan pendekatan dengan mencari multi-ouputt salah satunya adalah Regressor Chain (RC) yang diusulkan oleh [9], penelitian ini berbasiskan pada pendekatan problem transformation. Konsep dari RC didasarkan dari sebuah ide untuk melakukan chaining single-target models.

Penelitian Multi-output lainnya diusulkan oleh [10], Dia mengusulkan Random Forest sebagai multi-output regression. Dengan menggabungkan konsep ensemble dari sebuah decision tree menghasilkan sebuah structured output [10]. Penelitian lainnya yang menjadi fondasi seluruh konsep multi-output regression yang diusulkan oleh [11] dengan menggunakan linear regression untuk menghasilkan multi-output. Pada penelitian ini, kami mengusulkan perbandingan beberapa metode multi-output yang sering digunakan. Metode-metode yang digunakan adalah [10], [9], dan [11].

Tujuan penelitian pada artikel ilmiah ini adalah sebagai berikut:

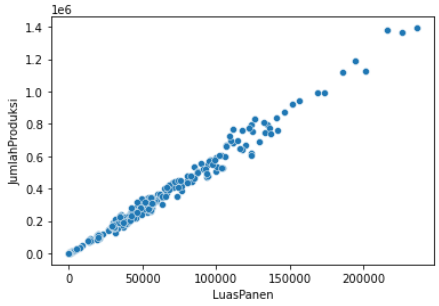
1. Mencari metode multi-output yang paling baik dari ketiga metode yang diusulkan dalam penanganan problem multi-output.
2. Mencari pengaruh proses outlier detection pada data mentah.
3. Penyiapan dataset untuk melakukan prediksi panen padi pada wilayah Pulau Jawa terutama Jawa Barat, Jawa Timur dan Jawa Tengah.

Struktur penulisan dalam artikel ini akan dibagi menjadi 5 Tahapan. Dimulai dari Pendahuluan yang menjelaskan mengenai latar belakang penelitian, Dataset yang menjelaskan mengenai dataset secara menyeluruh, Metodologi Penelitian yang menjelaskan tahapan dan Langkah-langkah yang dilakukan secara detail, Eksperimen dan Hasil pada tahapan ini akan menjelaskan hasil yang didapatkan dari metode yang digunakan dan beberapa percobaan berdasarkan pengukuran evaluasi yang sesuai. Terakhir adalah Kesimpulan yang menyimpulkan keseluruhan penelitian dengan tepat.

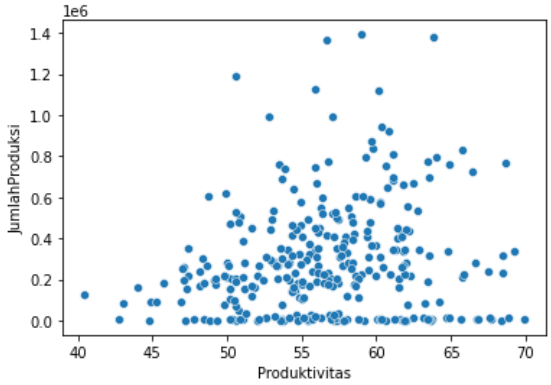
II. DATASET

Dataset yang digunakan pada penelitian ini diambil dari data Badan Pusat Statistik (BPS). Sumber data yang diambil adalah data jumlah produksi padi, luas panen dan produktivitas.Data yang diambil dimulai dari tahun 2018 sampai 2020. Untuk memberikan Batasan terhadap pengumpulan data, kami menggunakan data dari Pulau Jawa saja terutama Jawa Barat, Jawa Timur dan Jawa Tengah.

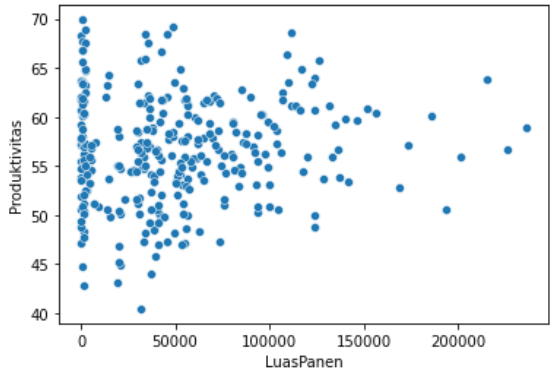
Berdasarkan data yang dikumpulkan, terdapat total sebanyak 300 data, Gambar 1 – Gambar 4 menjelaskan mengenai korelasi antara fitur yang digunakan. Dapat dilihat pada Gambar 1 dan Gambar 3 dimana kedua fitur ini memiliki korelasi secara linear. Pada Gambar 2 dan Gambar 3 dapat dilihat bahwa produktivitas tidak memiliki korelasi apapun terhadap Jumlah Produksi maupun Produktivitas.



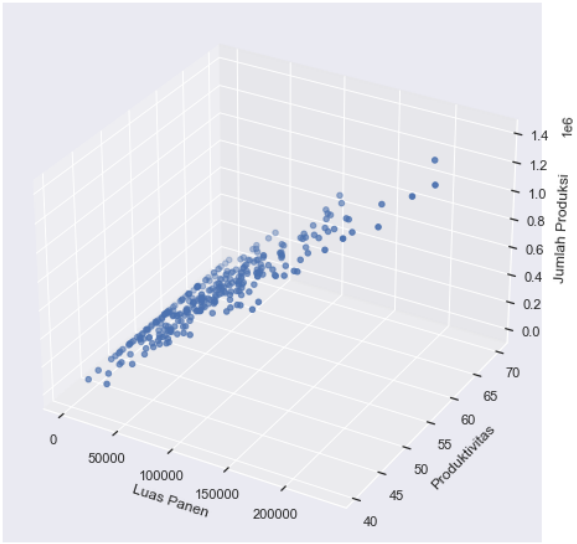
Gambar 1. Scatter Plot hubungan antara Luas Panen dengan Jumlah Produksi



Gambar 2. Scatter Plot hubungan antara Produktivitas dengan Jumlah Produksi



Gambar 3. Scatter Plot hubungan antara Luas Panen dengan Produktivitas.



Gambar 4. 3D Scatter Plot hubungan antara Luas Panen, Produktivitas dan Jumlah Produksi

III. METODOLOGI PENELITIAN

Section Metodologi Penelitian akan dibagi menjadi beberapa sub-section. Yang pertama adalah tahapan pra proses, pembuatan model menggunakan konsep statistic and machine learning dan terakhir adalah melakukan evaluasi model.

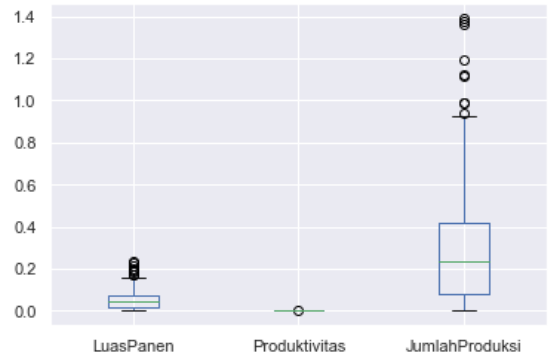
A. Praproses

Sebelum menjelaskan mengenai tahapan praproses, kami akan memulai dengan tahapan awal yaitu memisahkan antara variable bebas dan terikat. Pada penelitian ini variable bebas yang digunakan adalah Luas Panen/ Luas lahan, dan kota dimana variable terikatnya adalah produktivitas dan Jumlah produksi. Hal ini dikarenakan kota dan luas lahan sangat mempengaruhi produktivitas dan jumlah produksi.

Tahapan ini dilakukan untuk mempersiapkan data kotor menjadi data yang siap untuk digunakan dalam proses pembuatan model regresi. Langkah pertama yang kami lakukan adalah melakukan pengecekan terhadap missing value, dan penanganannya. Setelah melakukan pengecekan pada data yang kami gunakan, kami tidak menemukan data yang memiliki missing value sehingga tidak diperlukan untuk mengambil tindakan untuk melakukan filling.

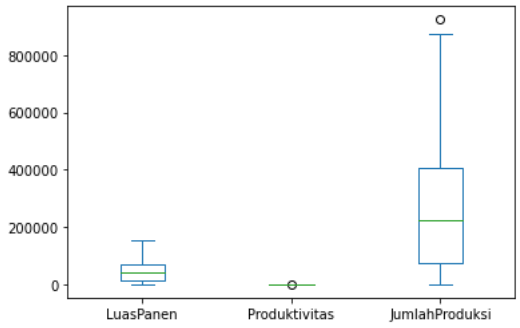
Proses selanjutnya adalah mengubah data yang Fitur dengan tipe data categorical seperti Kota. Fitur Kota yang bertipe data categorical diubah dengan metode one-hot encoding. One-hot encoding merupakan salah satu teknik yang sederhana yang dilakukan pada categorical data. Dimana sebuah data berbentuk categorical dalam bentuk himpunan (S) akan diubah menjadi nilai random variable (x). Sebagai contoh S={a,b,c} maka x1=a, x2=b, dan x3=c dengan one hot encoding menjadi x1=[1,0,0], x2=[0,1,0], x3=[0,0,1] [12].

Proses berikutnya adalah pencarian outlier dengan menggunakan boxplot. Pada gambar 5 merupakan penggambaran boxplot terhadap data awal sebelum mengalami penghapusan outlier removal. Pada tahapan ini kami tidak melihat



Gambar 5. Boxplot terhadap raw data

Dapat dilihat pada gambar diatas terdapat beberapa outlier terutama pada fitur Jumlah Produksi dan Luas Panen. Oleh karena itu kami menjadikan Fitur Jumlah Produksi untuk menjadi basis untuk menghapus outlier dari boxplot tersebut. Sehingga hasil boxplot yang didapatkan dari proses outlier detection adalah sebagai berikut.



Gambar 6. Boxplot terhadap data yang sudah dihapus outlier

Dapat dilihat dari hasil boxplot pada Gambar 6. Data outlier yang sebelumnya berada pada fitur hasil produksi sudah menghilang hanya terdapat satu data yang dianggap outlier. Data ini akan kami abaikan untuk penelitian ini.

B. Pembuatan Model

Setelah melakukan praproses dari data yang digunakan selanjutnya, kami melakukan implementasi beberapa metode yang sudah disebutkan pada Pendahuluan. Metode pertama adalah Regressor Chain (RC) yang diusulkan oleh [9]. Asumsikan bahwa full chain $(Y_1, Y_2, \dots, Y_d)$ dipilih, model pertama hanya digunakan untuk memodelkan Y_1 dan model selanjutnya digunakan untuk memodelkan Y selanjutnya. Setelah semua dilaksanakan nilai transformasi input vector diaugmentasikan kepada semua nilai asli dari target sebelumnya.

Metode kedua adalah Random Forest untuk multi output yang diusulkan oleh [10], konsep yang digunakan pertama mencari predictive clustering tree (PCT). Dimana nilai predictornya didapatkan dengan menggunakan konsep bootstrap yang bereplikasi selayaknya bagging. Secara lengkapnya setiap node dari decision trees, diambil sebuah random subset berdasarkan attributenya dan diambil yang terbaik dari subset tersebut.

Metode terakhir adalah linear regression untuk multi-response [11]. Konsep dasarnya adalah jika memiliki banyak target(d) sehingga $(y_1, y_2, \dots, y_d)^T$ dengan least square $(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_d)^T$

$$y = \overline{y_i} + \sum_{k=1}^{d} b_{ik}(\widehat{y_k} - \overline{y_k}) \quad (1)$$

C. Evaluasi

Untuk mengevaluasi model yang dibuat kami menggunakan metode R^2, *Explained Variance Score* (EVS) dan *Mean Square Error* (MSE) [3], [13].

$$EVS = 1 - \left(\frac{Var(\hat{Y} - Y)}{Var(y)}\right) \quad (2)$$

$$MSE = \frac{1}{n}\sum_{1}^{n}(Y_i - \hat{Y}_i) \quad (3)$$

$$r^2 = 1 - \frac{SS_{res}}{SS_{total}} \quad (3)$$

$$SS_{tot} = \sum_{i}(y_i - \hat{y})^2 \quad (4)$$

$$SS_{res} = \sum_{i}(y_i - f_i)^2 \quad (5)$$

IV. HASIL EKSPERIMEN

Eksperimen ini dilakukan dengan data sebanyak 300 data yang keseluruhannya dijadikan model machine learning. Untuk melihat apakah model berjalan dengan baik kami melakukan pengevaluasian dengan MSE, RRMSE dan R^2. Setelah itu hasil dari nilai evaluasi akan

dibandingkan terhadap metode-metode yang dimplementasikan pada data. Semua proses ini dilakukan dengan bantuan library scikit-learn dari python[14] untuk menyelesaikannya. Pada RC, based model yang digunakan adalah Random Forest.

Beberapa hyperparameter yang digunakan untuk Random Forest adalah penggunaan GINI sebagai criteration, n_estimator =100; untuk Regression Chain menggunakan based Random Forest dan untuk Linear Regression menggunakan *intercept random*.

Tabel 1 akan menggambarkan hasil terhadap tiap metode tanpa melakukan Outlier Removal dan Tabel 2 akan menggambarkan hasil terhadap tiap metode dengan melakukan Outlier Removal.

Tabel 1. Hasil Evaluasi dari metode Multi-output untuk melakukan prediksi terhadap produktivitas dan Jumlah Produksi Padi

Metode	EVS	MSE	R^2
Random Forest	1	0	1
Multi output Linear Regression	0.85	103’317’067.81	0.85
RC	0.65	356’024’623.06	0.65

Tabel 2. Hasil Evaluasi dari metode Multi-output untuk melakukan prediksi terhadap produktivitas dan Jumlah Produksi Padi dengan Outlier Removal

Metode	EVS	MSE	R^2
Random Forest	0.89	280’703’454.88	0.89
Multi output Linear Regression	0.88	308’560’490.59	0.88
RC	0.87	350’536’251.05	0.87

Berdasarkan hasil yang didapatkan dari penelitian ini ditemukan bahwa walaupun menggunakan seluruh data dapat menghasilkan nilai yang optimal dengan hasil EVS dan R^2 serta MSE yang sempurna akan tetapi jika dilihat pada hasil lainnya terdapat ketimpangan yang cukup signifikan sehingga sulit dikatakan model yang digunakan bernilai valid. Berbeda dengan hasil tanpa outlier removal, pada outlier removal data yang didapatkan lebih berimbang dimana nilai hasil EVS dan R^2 adalah berkisar antara 0.87-0.89, ini dapat menyatakan bahwa adanya peningkatan hasil terutama Ketika menggunakan *outlier removal*.

V. SIMPULAN

Penelitian ini ingin mencari metode terbaik untuk permasalahan prediksi multi-output dari data padi yang dimiliki. Metode yang digunakan adalah Linear Regression, Regressor Chain dan Random Forest. Penelitian ini melakukan beberapa operasi untuk menghasilkan nilai yang baik salah satu pada proses pra-proses dilakukan pencarian nilai missing value. Selanjutnya melakukan evaluasi dengan menggunakan MSE, EVS, dan R2. Berdasarkan hasil yang ditemukan dari penelitian ini, didapatkan bahwa hasil dari Random Forest memberikan hasil sempurna terutama pada yang tidak mengalami penghapusan outlier dimana MSE nya bernilai 0 dan score untuk EVS dan R2 adalah 1. Akan tetapi berdasarkan pengamatan dari data, ini bisa terjadi karena pada setiap metode memiliki nilai yang fluktuatif. Berbeda dengan penggunaan Outlier Removal dari ketiga metode algoritma menghasilkan nilai yang sama sehingga dapat diasumsikan bahwa hasil adanya peranan untuk mengurangi kesalahan yang terjadi dalam data.

Penelitian ini merupakan penelitian awal dari program prediksi multi-output dimana masih banyak yang perlu dianalisa terutama pada pengurangan nilai error yang terjadi dalam model. Pada penelitian berikutnya

DAFTAR PUSTAKA

[1] Badan Pusat Statistik, “Kajian Konsumsi Bahan Pokok 2020,” Badan Pusat Statistik, Indonesia, Aug. 2018.

[2] A. W. Sugiyarto, D. U. Wutsqa, N. Hendiyani, and A. R. Rasjava, “Optimization of Genetic Algorithms on Backpropagation Neural Network to Predict National Rice Production Levels,” in *2019 International Conference on Applied Information and Technology and Innovation (ICAITI)*, Denpasar, Indonesia, Sep. 2019, p. 5.

[3] H. Borchani, G. Varando, C. Bielza, and P. Larrañaga, “A survey on multi-output regression:

Multi-output regression survey," *WIREs Data Mining Knowl Discov*, vol. 5, no. 5, pp. 216–233, Sep. 2015, doi: 10.1002/widm.1157.

[4] N. Gandhi, O. Petkar, and L. J. Armstrong, "Rice crop yield prediction using artificial neural networks," in *2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, Chennai, India, Jul. 2016, pp. 105–110. doi: 10.1109/TIAR.2016.7801222.

[5] T.-Z. Jheng, T.-H. Li, and C.-P. Lee, "Using hybrid support vector regression to predict agricultural output," in *2018 27th Wireless and Optical Communication Conference (WOCC)*, Hualien, Apr. 2018, pp. 1–3. doi: 10.1109/WOCC.2018.8372729.

[6] F. Wang *et al.*, "Rice yield estimation at pixel scale using relative vegetation indices from unmanned aerial systems," in *2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, Istanbul, Turkey, Jul. 2019, pp. 1–4. doi: 10.1109/Agro-Geoinformatics.2019.8820226.

[7] Md. S. Alam, K. Kalpoma, Md. S. Karim, A. Al Sefat, and J. Kudoh, "Boro Rice Yield Estimation Model Using Modis Ndvi Data for Bangladesh," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, Yokohama, Japan, Jul. 2019, pp. 7330–7333. doi: 10.1109/IGARSS.2019.8899084.

[8] W. M. R. K. Wanninayaka, R. M. K. T. Rathnayaka, and E. P. N. Udayakumara, "Artificial neural network to estimate the paddy yield prediction using remote sensing, weather and non weather variable in Ampara district, Sri Lanka," in *2020 5th International Conference on Information Technology Research (ICITR)*, Moratuwa, Sri Lanka, Dec. 2020, pp. 1–6. doi: 10.1109/ICITR51448.2020.9310894.

[9] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier Chains for Multi-label Classification," *Machine Learning*, p. 16.

[10] D. Kocev, C. Vens, J. Struyf, and S. Džeroski, "Tree ensembles for predicting structured outputs," *Pattern Recognition*, vol. 46, no. 3, pp. 817–833, Mar. 2013, doi: 10.1016/j.patcog.2012.09.023.

[11] L. Breiman and J. H. Friedman, "Predicting Multivariate Responses in Multiple Linear Regression," *J Royal Statistical Soc B*, vol. 59, no. 1, pp. 3–54, Feb. 1997, doi: 10.1111/1467-9868.00054.

[12] J. T. Hancock and T. M. Khoshgoftaar, "Survey on categorical data for neural networks," *J Big Data*, vol. 7, no. 1, p. 28, Dec. 2020, doi: 10.1186/s40537-020-00305-w.

[13] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. USA: John Wiley & Sons, Inc., 1987.

[14] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2826–2830, 2011.