JNATIA Volume 3, Nomor 3, Mei 2025 p-ISSN: 2986-3929
Jurnal Nasional Teknologi Informasi dan Aplikasinya e-ISSN: 3032-1948

Analisis Prediksi Ukuran Baju dengan Metode Regresi Polinomial

Desak Putu Tia Rusilia Wati[a1], I. A. GSuwiprabayanti Putra [a2]

Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Udayana
Jalan Raya Kampus UNUD, Bukit Jimbaran, Kuta Selatan, Badung, Bali, Indonesia
[1]wati.2208561143@student.unud.ac.id
[2]iagsuwiprabayantiputra@unud.ac.id (Corresponding Author)

Abstract

In the modern era, online shopping for clothing presents the challenge of determining the correct size. This research aims to predict clothing sizes using Polynomial Regression, which can capture the non-linear relationships between body metrics and clothing sizes. The study utilizes a dataset from Kaggle comprising weight, height, age, and clothing size attributes. Through data preprocessing, including feature transformation, engineering, selection, and cleaning, the dataset is prepared for analysis. Various models are evaluated, and Polynomial Regression is identified as the most effective, achieving an R^2 score of 0.70755203. Hyperparameter tuning using GridSearchCV further optimizes the model, resulting in a final R^2 score of 72.555511% with degree 5 and alpha 1. The evaluation indicates that while the model accurately predicts sizes, it sometimes struggles with adjacent sizes, particularly in medium ranges. This research demonstrates the potential of Polynomial Regression in improving the accuracy of clothing size predictions, thereby facilitating better online shopping experiences.

Keywords: *Polynomial Regression, Hyperparameter tuning, R^2 score, GridSearchCV*

1. Pendahuluan

Pada era modern ini, belanja pakaian tidak lagi memerlukan kunjungan fisik ke toko, karena dapat dilakukan secara online. Namun, salah satu tantangan yang dihadapi konsumen adalah menentukan ukuran baju yang sesuai. Berat, tinggi badan dan Indeks Massa Tubuh (IMT) merupakan beberapa faktor yang mempengaruhi ukuran baju, dan oleh karena itu diperlukan metode yang dapat memprediksi ukuran baju dengan optimal.

Penelitian sebelumnya menggunakan implementasi algoritma *Decision Tree Cart* untuk merekomendasi ukuran baju. Penelitian ini menunjukkan bahwa pendekatan matematis dan algoritma dapat meningkatkan akurasi prediksi ukuran baju dengan akurasi sebesar 67% [1]. Berdasarkan penelitian yang sudah dilakukan sebelumnya yaitu dalam memprediksi ukuran baju. Penulis akan melakukan analisa prediksi ukuran baju dengan metode Regresi Polinomial.

Regresi Polinomial menjadi salah satu metode yang dapat digunakan untuk melakukan prediksi ukuran baju. Regrasi Polinomial merupakan bentuk analisa regresi di mana hubungan antara variabel bebas dan variabel terikatnya dimodelkan di dalam orde polinomial [2]. Metode ini mampu menangkap hubungan non-linear antara variabel input dan variabel output, sehingga diharapkan mampu memberikan hasil prediksi yang lebih baik.

Dalam penelitian ini, akan dilakukan eksplorasi penggunaan Regresi Polinomial serta dilakukan *hyperparameter tuning* untuk memperoleh nilai R2 yang tinggi. Diharapkan penelitian ini dapat memberikan kontribusi signifikan dalam memudahkan menentukan ukuran baju yang tepat dengan informasi berat, tinggi badan dan IMT. Indeks massa tubuh (IMT) telah di buktikan berkorelasi kuat dengan estimasi persentase lemak tubuh [3]. Diharapkan penelitian ini dapat memberikan kontribusi signifikan dalam memudahkan menentukan ukuran baju yang tepat dengan informasi berat dan tinggi badan.

p-ISSN: 2986-3929
e-ISSN: 3032-1948

2. Metode Penelitian

2.1. Pengumpulan Data

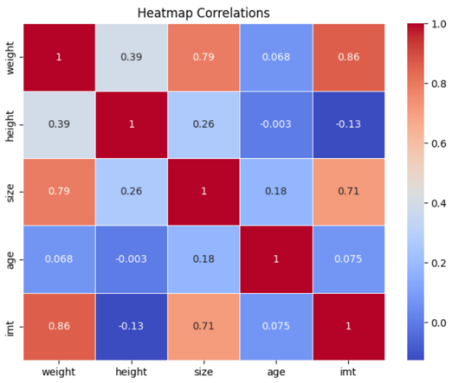
Penelitian ini menggunakan dataset yang diunduh dari platform Kaggle dengan judul "Clothes-Size-Prediction". Dataset tersebut mencakup empat atribut utama yaitu berat, tinggi, umur dan ukuran baju. Peneliti menggunakan data set ini untuk memprediksi ukuran baju dengan memanfaatkan atribut relevan yang tersedia pada data set tersebut.

2.2. Preprocessing Data

Sebelum melakukan analisis lebih lanjut, data perlu melalui beberapa tahap pra-pemrosesan untuk memastikan kualitasnya. Pra-pemrosesan data dilakukan menggunakan Google Colab untuk memanfaatkan kemudahan akses dan kapabilitas komputasi yang disediakan. Beberapa proses dilakukan pada tahap ini adalah sebagai berikut:

a. *Feature Transformation*
Kolom 'size', yang awalnya berbentuk kategorikal, diubah menjadi tipe data numerik menggunakan teknik Label Encoding. Transformasi ini dilakukan dengan memetakan setiap kategori ukuran baju menjadi angka. Langkah ini penting untuk memudahkan model dalam memahami dan memproses variabel kategorikal.

b. *Feature Engineering*



Gambar 1. Alur Metode Penelitian Analisis Sentimen

Selanjutnya, dilakukan penambahan fitur baru berupa Indeks Massa Tubuh (imt). Penambahan fitur ini dilakukan karena korelasi antara BMI dan 'size' menunjukkan nilai yang cukup tinggi (0,71), yang berarti BMI memiliki hubungan yang signifikan dengan ukuran baju.

c. *Feature Selection*
Feature selection merupakan proses mengidentifikasi dan memilih variabel-variabel input yang paling relevan dengan target variabel [4]. Pada proses ini Kolom 'age' dihapus dari dataset karena korelasi antara 'age' dengan 'size' menunjukkan nilai yang paling rendah (0,18) dibandingkan dengan variabel lain. Korelasi yang rendah ini menunjukkan bahwa usia tidak memberikan kontribusi signifikan dalam memprediksi ukuran baju, sehingga atribut ini dihilangkan untuk mengurangi kompleksitas model dan dilakukan penghapusan baris yang memiliki nilai missing value.

d. *Data Cleaning*
Penghapusan outlier pada variabel 'weight' terhadap 'size' dilakukan. Berdasarkan heatmap korelasi, 'weight' memiliki korelasi yang sangat tinggi dengan 'size' (0,79). Outlier

pada variabel ‘weight’ dapat mempengaruhi keakuratan model, sehingga penghapusan outlier diperlukan untuk mengurangi noise dan meningkatkan performa model.

2.3. Data Splitting

Splitting data adalah proses membagi dataset menjadi dua atau lebih bagian yang saling ekslusif untuk melatih serta menguji model [5]. Dataset dibagi menjadi dua subset: data latih dan data uji. Pembagian ini dilakukan untuk mengevaluasi performa model dengan data yang tidak digunakan selama pelatihan, memastikan bahwa model tidak overfitting. Dalam penelitian ini, 80% data digunakan untuk pelatihan dan 20% sisanya digunakan untuk pengujian.

2.4. Model Selection

Pada tahap ini, berbagai model prediksi dipilih dan dilatih menggunakan data latih. Model yang dipertimbangkan oleh penulis tertera pada tabel 1.

Tabel 1. Model Pertimbangan

Model Pertimbangan
K-Means Classifier
K-Nearest Neighbors
Decision Tree
Naive Bayes
Random Forest Classifier
Linear Regression
Logistic Regression
Ridge Regression
Lasso Regression
Polynomial Regression

2.5. Hyperparameter Tuning

Pada proses tuning parameter dilakukan untuk mendapatkan parameter yang paling optimal menggunakan Teknik GridSearch Cross Validation yang mana GridSearchCV adalah bagian dari modul scikit-learn yang bertujuan secara otomatis dan sistematis melakukan validasi beberapa model dan setiap hyperparameter [6].

2.6. Model Evaluation

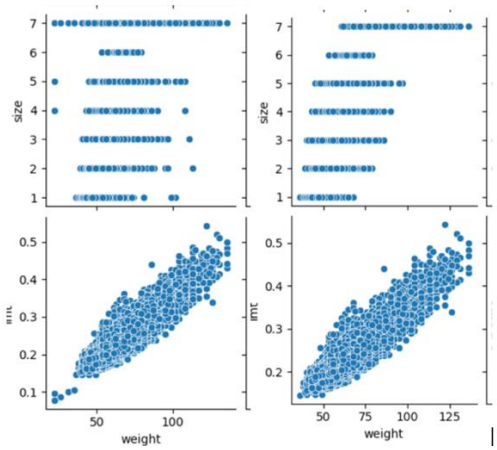
Setelah menemukan kombinasi parameter terbaik melalui hyperparameter tuning, langkah selanjutnya adalah melakukan evaluasi model untuk mengevaluasi performa model pada data uji (testing set). Proses ini dilakukan untuk memastikan bahwa model yang telah dioptimalkan tidak hanya bekerja baik pada data latih (training set) tetapi juga pada data yang belum pernah dilihat sebelumnya.

3. Hasil dan Pembahasan

Hasil dari pra-pemrosan data telah dilakukan untuk mencapai model yang optimal. Dari proses proses transformasi feature pada kolom ‘size’, penambahan feature berupa fitur indek massa tubuh (imt), seleksi fitur dengan penghilangan kolom ‘age’, pemrosesan pada nilai ‘missing value’. Selanjutnya dilakukan data cleaning dengan melakukan penghapusan outlier pada variabel ‘weight’ terhadap ‘size’.

p-ISSN: 2986-3929
e-ISSN: 3032-1948

Berikut merupakan hasil setelah dilakukan pembersihan data dapat dilihat pada gambar 2.



Gambar 2. Persebaran ‘berat’ dan ‘imt’ terhadap ‘size’

Langkah selanjutnya yaitu membagi data latih (training set) sebesar 80% dan data uji (testing set) sebesar 20%.

3.1. Hasil Pemilihan Model

Pemilihan model yang dilakukan menghasilkan nilai akurasi model klasifikasi yang tertulis pada tabel 2 dan nilai R2 pada model regresi tertulis pada tabel 3.

Tabel 2. Hasil Akurasi Model

Model	Akurasi
K-Means Classifier	9.342299
K-Nearest Neighbors	43.857953
Decision Tree	50.485396
Naive Bayes	49.144778
Random Forest Classifier	50.392940

Tabel 3. Hasil Nilai R2 Model

Model	Nilai R2
Linear Regression	0.64622219
Logistic Regression	0.50161799
Ridge Regression	0.64524896
Lasso Regression	0.63717535
Polynomial Regression	0.70755203

Berdasarkan hasil di atas, model regresi seperti regresi linier, regresi punggungan (ridge regression), regresi lasso, dan regresi polinomial menunjukkan hasil yang lebih baik. Secara khusus, regresi polinomial memiliki nilai R2 paling tinggi yaitu 0.70755203 yang menggambarkan

p-ISSN: 2986-3929
e-ISSN: 3032-1948

model ini baik dalam menangkap hubungan non-linear antara fitur dan target, dan memberikan penjelasan yang lebih baik tentang variabilitas data.

Dari penelitian ini, penulis menggunakan metode regresi polinomial untuk memprediksi ukuran baju. Selanjutnya dilakukan hyperparameter tuning dengan menggunakan GridSearchCV pada pipeline yang menggabungkan *'Polynomial Features'* dan '*Ridge regression'*. Tujuannya adalah untuk menemukan kombinasi parameter terbaik yang menghasilkan nilai R^2 tertinggi, yang menggambarkan seberapa baik model tersebut dapat menjelaskan variabilitas dalam data. Berikut merupakan hasil yang telah dilakukan tertera pada tabel 4.

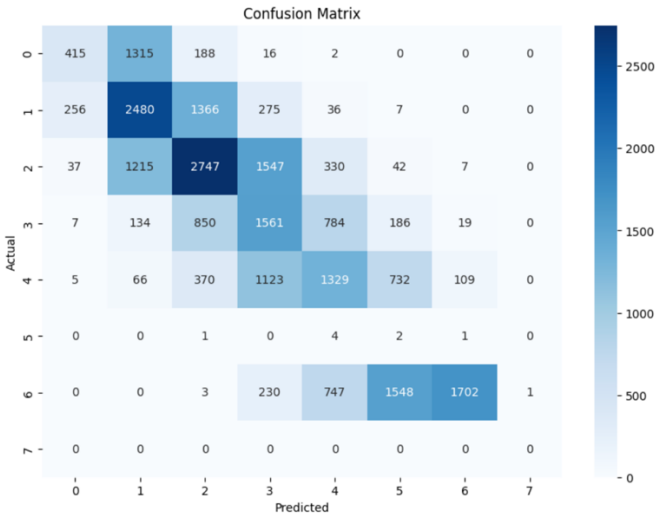
Tabel 4. Hasil Hyperparameter Tuning

Degre	Alpha	R^2 *Score* (%)
5	0.01	72.552928
5	1	72.555511
7	0.01	72.447407
7	1	72.447729

Dari hasil tuning tersebut, didapatkan kombinasi parameter terbaik yaitu *Degree* = 5 dan *Alpha* = 1 dengan nilai R^2 *Score* sebesar 72.555511%. Kombinasi ini menunjukkan performa terbaik dalam menjelaskan variabilitas data.

3.2. Hasil Evaluasi Model

Selanjutnya model akan diuji menggunakan data testing. Confusion matrix membantu dalam mengidentifikasi sejauh mana prediksi model mendekati nilai sebenarnya. Hasil dari pengujian dari model regresi polynomial ditunjukkan oleh *confusion matrix* pada Gambar 3.



Gambar 3. Confusion Matrix Hasil Pengujian

Dari confusion matrix di atas, dapat dilihat bahwa model memiliki kecenderungan untuk salah memprediksi ukuran yang berdekatan, terutama pada kelas-kelas dengan ukuran menengah seperti ukuran 1, 2, dan 3. Hal ini menunjukkan bahwa model memiliki beberapa kesulitan dalam membedakan ukuran yang berdekatan. Hal ini disebabkan karena seseorang dengan berat badan dan tinggi badan tertentu dapat mengisi data dengan ukuran baju yang kebesaran, mengingat preferensi beberapa orang untuk menggunakan baju oversize.

4. Kesimpulan

Penelitian ini menunjukkan bahwa metode regresi polinomial efektif dalam memprediksi ukuran baju berdasarkan berat, tinggi badan, dan IMT. Model regresi polinomial yang dioptimalkan melalui hyperparameter tuning mencapai nilai R^2 sebesar 0.72555511, menunjukkan kemampuan yang baik dalam menjelaskan variabilitas data. Namun, evaluasi model menunjukkan bahwa prediksi ukuran sering kali salah pada ukuran yang berdekatan, terutama untuk ukuran menengah. Hal ini mungkin disebabkan oleh preferensi individu terhadap pakaian yang lebih besar. Secara keseluruhan, penelitian ini berhasil menunjukkan bahwa regresi polinomial dapat memprediksi ukuran baju dengan cukup baik, memberikan kontribusi yang signifikan dalam memudahkan penentuan ukuran baju yang tepat dalam belanja online.

Daftar Pustaka

[1] F. A. Oktavirahani and R. Maharesi, “Implementasi Algoritma Decision Tree Cart Untuk Merekomendasikan Ukuran Baju,” JURIKOM (Jurnal Riset Komputer), vol. 9, no. 1, pp. 138–147, 2022.

[2] B. F. Susanto, S. Rostianingsih, and L. W. Santoso, “Analisa Audio Features dengan Membandingkan Metode Multiple Regression dan Polynomial Regression untuk Memprediksi Popularitas Lagu,” Jurnal Infra, vol. 9, no. 2, pp. 77–83.

[3] M. Situmorang, “Penentuan Indeks Massa Tubuh (IMT) melalui Pengukuran Berat dan Tinggi Badan Berbasis Mikrokontroler AT89S51 dan PC,” Jurnal Teori dan Aplikasi Fisika, vol. 3, no. 2, 2015.

[4] Brownlee, J. 2020. How to Perform Feature Selection for Regression Data. URI = https://machinelearningmastery.com/feature-selection-for- regression-data/

[5] Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl, K. C. (2020). Data mining for Business Analytics: Concepts, Techniques, and Applications in R (3rd ed.). John Wiley \& Sons, Inc.

[6] P. M. Kouate, “Machine Learning: Gridsearchcv & Randomizedsearchcv,” Https://Towardsdatascience.Com/Machine-Learning-Gridsearchcv-Randomizedsearchcv-D36b89231b10, Sep. 11, 2020.