Project Overview

Perkembangan teknologi telah membuat pergeseran perilaku pelanggan dari pembelian melalui offline shop ke online shop. Pembelian secara online membuat pelanggan melakukan transaksi terhadap barang yang diinginkan darimana saja. Hal tersebut banyak mempengaruhi persepi pelanggan jika ingin melakukan pembelian antara lain harga, produksi, promosi dan tempat [1]. Berdasarkan persepi tersebut banyak strategi yang dilakukan dari sisi penjual untuk meningkatkan penjualannya diantaranya adalah Online Customer Rating.

Online Customer Rating adalah bagian review yang menggunakan bentuk symbol bintang daripada bentuk teks dalam mengekspresikan pendapat pelanggan [1]. Rating dapat diartikan sebagai penilaian dari pengguna pada refensi pelanggan yang mengacu pada keadaa n psikologis dan emosional terhadap suatu barang [2]. Tingkat rating yang kuantitasnya lebih dari 25% dengan rating yang diberikan buruk, maka pelanggan akan 2 kali berpikir untuk membeli produk tersebut [3]. Pengaruh rating yang cukup tinggi memberikan penjual memikirkan strategi yang tepat untuk meningkatkan penjualan.

Pada projek ini akan mencoba memprediksi menggunakan pendekatan machine learning tingkat rata-rata rating pada suatu produk berdasarkan harga, kategori dan total reviews yang didapatkan pada produk tersebut. Dataset didapatkan dari website Kaggle dengan nama dataset Lazada review yang berisi informasi seputar produk. Tahapan yang akan dilakukan adalah EDA, Data Preprocessing, Model dan Evaluasi. Model yang digunakan pada projek ini ialah Support Vector Regressor, Linear Regression dan Boosting Model.

Business Understanding

Pada pernyataan yang telah dijelaskan, masalah yang diangkat adalah

1. Bagaimana mengetahui korelasi fitur lain terhadap label (rata-rata rating)?
2. Bagaimana kemampuan model yang digunakan dalam melakukan prediksi pada nilai label?

Tujuan dari masalah yang diangkat adalah

1. Mengetahui fitur-fitur yang memiliki pengaruh kuat terhadap label
2. Mengetahui model machine learning dalam melakukan prediksi terhadap dataset yang digunakan

Solusi yang digunakan adalah

1. Melihat pengecekan outliers dan distribusi data
2. Menggunakan fungsi korelasi pandas dengan visualisasi dari seaborn (heatmap)
3. Melakukan Univariate dan Multivariate fitur-fitur independent (fitur yang dinyatakan bukan sebagai label)
4. Melakukan Evaluasi dengan Mean Squared Error (MAE)

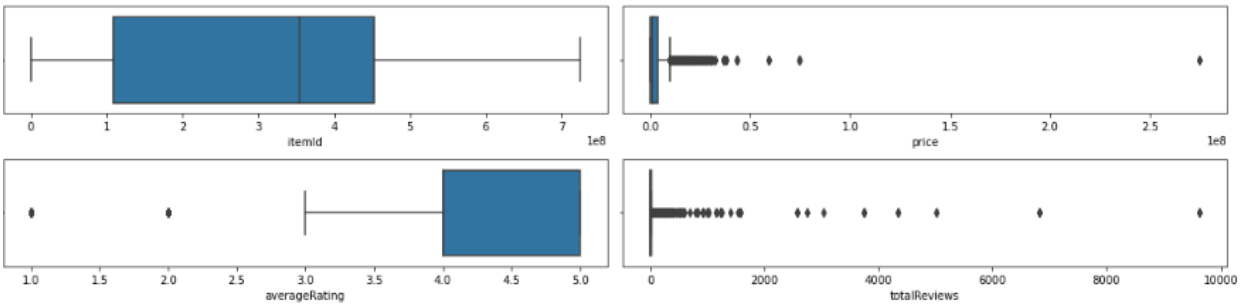
Data Understanding

Dataset yang digunakan pada projek ini adalah dataset Lazada Review yang didapatkan dari Kaggle yang terdiri dari 10942 bari dan 9 fitur. Fitur yang terdapat antara lain.

1. itemId: Nomor identitas dari tiap-tiap produk (Tipe numerik)
2. category (Tipe objek): Jenis kategori tiap produk. Kategori pada dataset ini terdiri dari
    a. beli-hardisk-eksternal
    b. jual-flash-drives
    c. beli-smart-tv
    d. shop-televisi-digital
    e. beli-laptop
3. name: Nama dari tiap-tiap produk (Tipe objek)
4. brandName: Nama brand yang mengeluarkan produk tersebut (Tipe objek)
5. url: alamat website dari produk tersebut (Tipe objek)
6. price: Harga dari produk tersebut (Tipe numerik)
7. averageRating: rata-rata tingkat rating produk tersebut (Tipe numerik)
8. totalReviews: Total reviews yang didapatkan pada produk tersebut (Tipe numerik)
9. retrievedDate: Tanggal terakhir diakses produk tersebut (Tipe objek)

Pada projek ini fitur yang digunakan sebagai fitur dependent/label adalah fitur **averageRating** dan tahapan yang digunakan adalah Exploratory Data Analisis (EDA) yang mencakup.

1. **Cek data null dan menghapus outlier menggunakan rumus IQR (Interquartile Range).**




```
itemId            0
price           750
averageRating   859
totalReviews   1612
```

Jika dilihat, bahwa nilai outlier price, averageRating, dan total reviews memiliki total outlier yang tinggi. Projek ini langsung menghapus data outlier tersebut dengan rumus matematis IQR dikarenakan jumlah data yang dimiliki terbilang besar.

```
Q1 = df.quantile(0.25)
Q3 = df.quantile(0.75)

IQR = Q3-Q1
```

```
df = df[~((df < (Q1 - 1.5 * IQR)) | (df > (Q3 + 1.5 * IQR))).any(axis = 1)]
df.shape
```

```
(7796, 9)
```

Setelah data outlier dihapus, data yang tersisa 7796 baris dan 9 kolom.

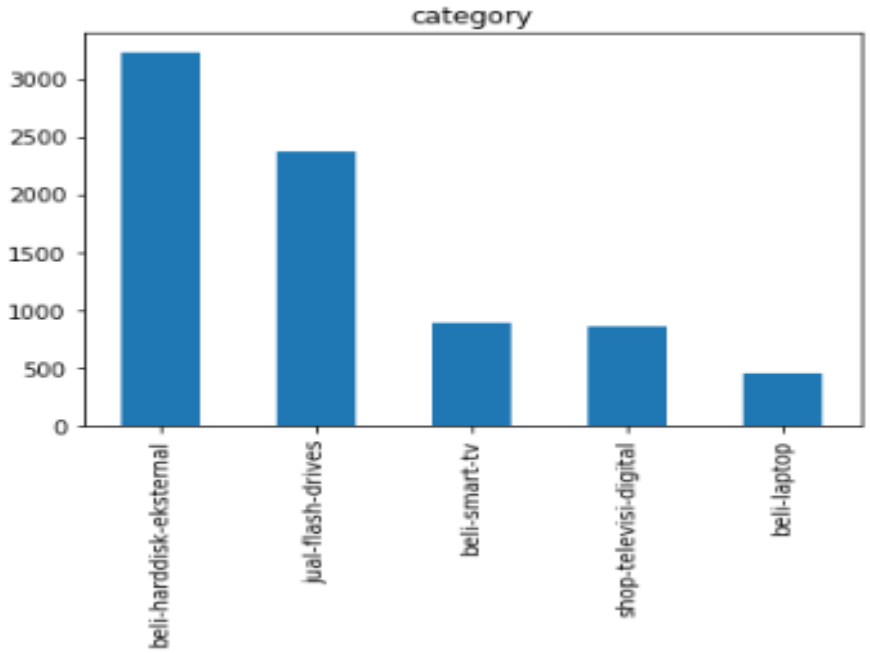
2. Univariate Analysis terhadap fitur kategori dan fitur numerik.

a. Kategorik

Fitur yang digunakan pada projek ini adalah fitur category. Hal ini untuk mencegah dimensi/fitur yang banyak saat melakukan One-Hot Encoder. Dimensi yang terlalu banyak dapat menurunkan tingkat keakuratan model yang digunakan pada projek ini dalam melakukan prediksi. Hal lainnya adalah penyimpanan data yang lebih sedikit, mengurangi waktu komputasi, dan kualitas data yang meningkat [4]. Berikut nilai unik pada tiap-tiap fitur.

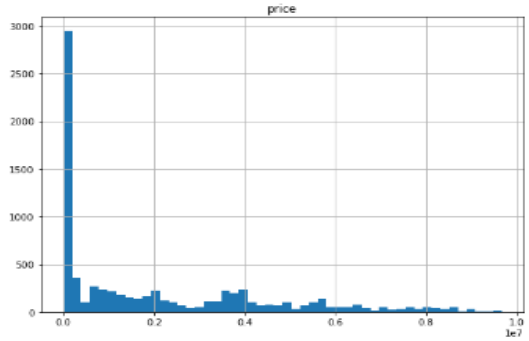
```
itemId          4422
category           5
name            4286
brandName        235
url             4422
price           1861
averageRating      5
totalReviews     217
retrievedDate      1
```

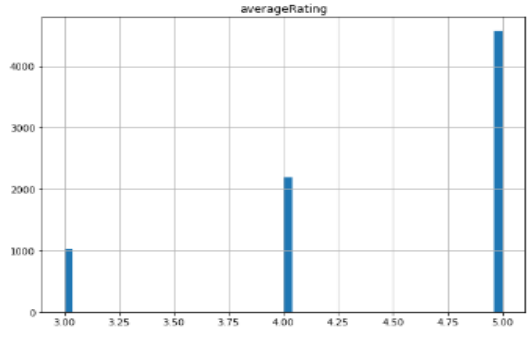
Fitur category yang memiliki data unik paling sedikit merupakan alasan fitur category akan digunakan pada tahap selanjutnya. Berikut frekuensi dan presentasi data unik pada fitur category.

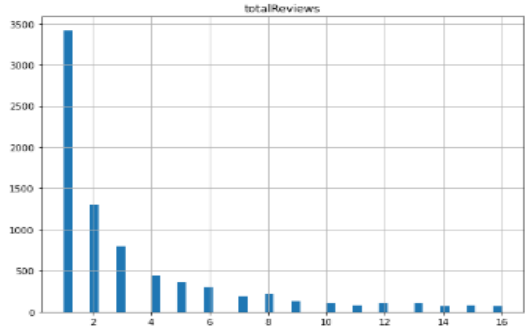


b. Numerik

Fitur numerik yang digunakan adalah fitur price, averageRating, totalReviews. ItemId tidak digunakan karena ialah identitas dari produk tersebut.





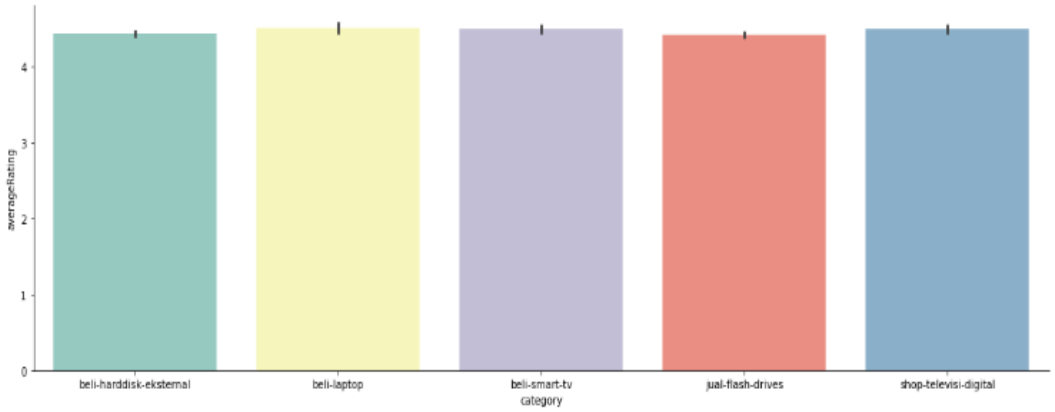


- Pada distribusi nilai price, semakin tinggi nilai yang dihasilkan akan menurunkan hasil frequency nilai tersebut
- Total reviews bergerak berbanding lurus dengan nilai price dimana semakin tinggi nilai price. Maka total reviews yang didapatkan semakin tinggi dan sebaliknya.
- Tingginya total reviews dengan nilai price yang kecil menghasilkan nilai rata-rata rating (averageRating) yang menurun.
- Kesimpulan, user lebih menyukai pembelian dengan harga yang lebih kecil namun dengan rata-rata rating yang didapatkan cukup rendah

3. Multivariate Analysis terhadap fitur kategori dan numerik terhadap label.

a. Kategorik

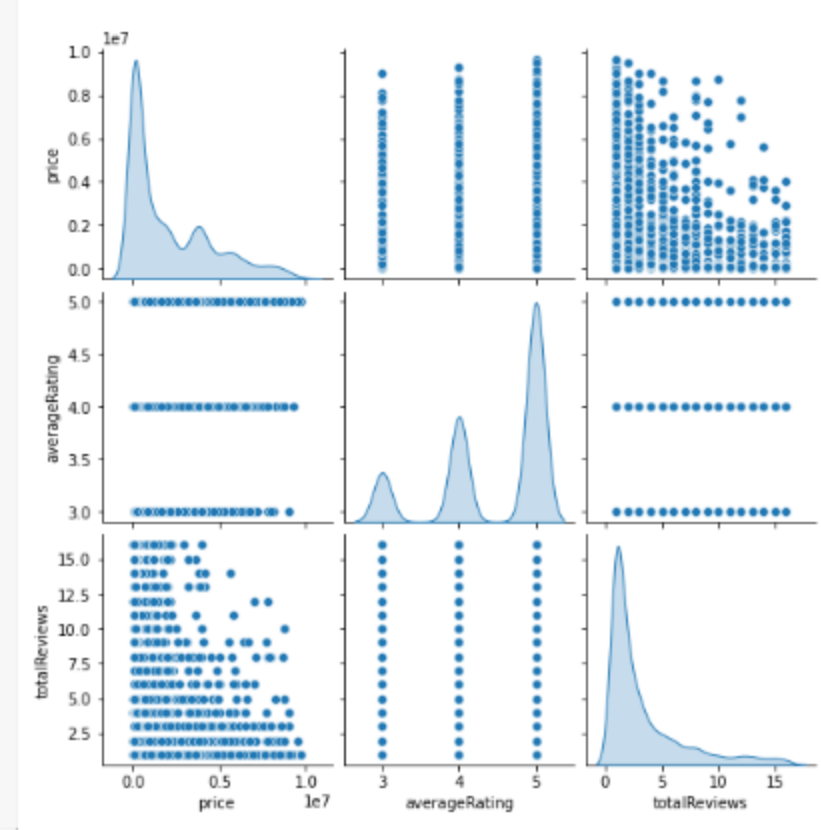
Tahapan ini melihat pengaruh nilai pada fitur category terhadap label (fitur averageRating).



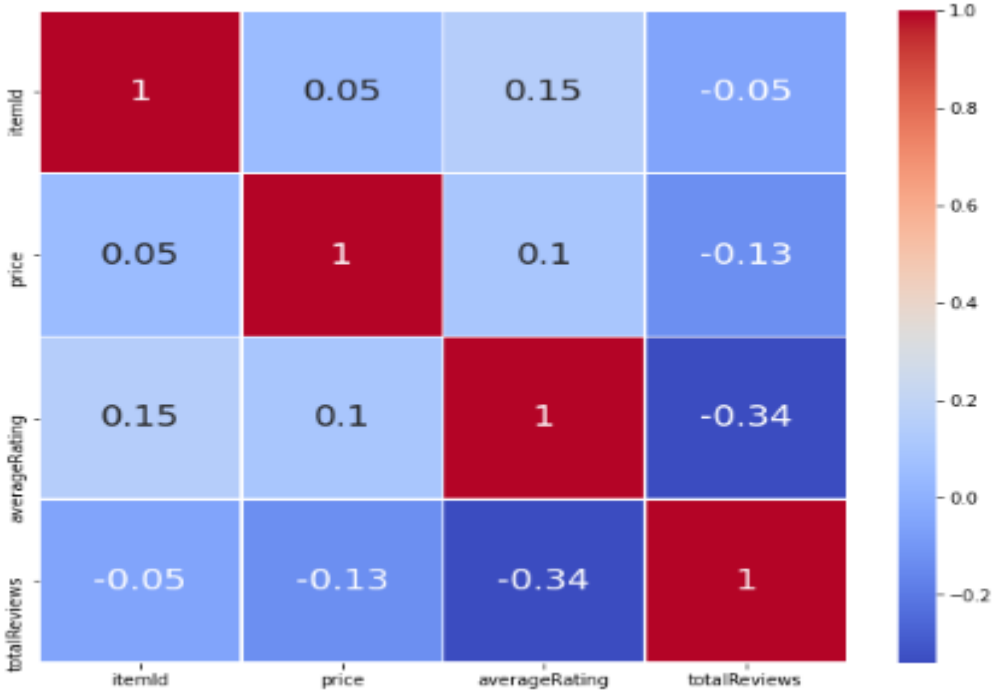
Nilai pada tiap-tiap nilai kategori memiliki nilai yang sama. Sehingga bisa disumpulkan bahwa fitur category memiliki pengaruh yang lemah.

b. Numerik

Pada fitur numerik, akan menggunakan visualisasi distplot untuk mengetahui persebaran data dan tingkat korelasinya.



Ternyata dapat diketahui bahwa nilai price terhadap averageRating tingkat persebaran datanya tidak linear dan hal itu berbanding terbalik pada totalReviews terhadap averageRating. Untuk mengetahui angka korelasi akan digunakan sns.heatmap.



Ternyata korelasi yang didapatkan antara price dengan averageRating di bawah 0.5 dan korelasi antara totalReviews dan averageRating besifat minum yang artinya bahwa jumlah reviews akan berbanding dengan nilai averageRating yang didapatkan. Nilai rentang korelasi dimulai dari -1 hingga +1 [5].

Data Preparation

Tahapan yang dilakukan pada data preparation projek ini adalah

1. **One-Hot Encoder.**

   One-Hot Encoder merupakan metode binary converter dimana nilai '1' dilambangkan untuk tiap kondisi yang sesuai dan nilai '0' sebagai kondisi yang tidak sesuai [6]. Projek ini menggunakan dummies milik pandas untuk melakukan one-hot encoder pada fitur category.

	price	totalReviews	averageRating	cut_beli-harddisk-eksternal	cut_beli-laptop	cut_beli-smart-tv	cut_jual-flash-drives	cut_shop-televisi-digital
0	2499000	8	4	1	0	0	0	0
1	3788000	3	3	1	0	0	0	0
2	3850000	2	3	1	0	0	0	0
3	1275000	11	3	1	0	0	0	0
4	3984100	1	5	1	0	0	0	0

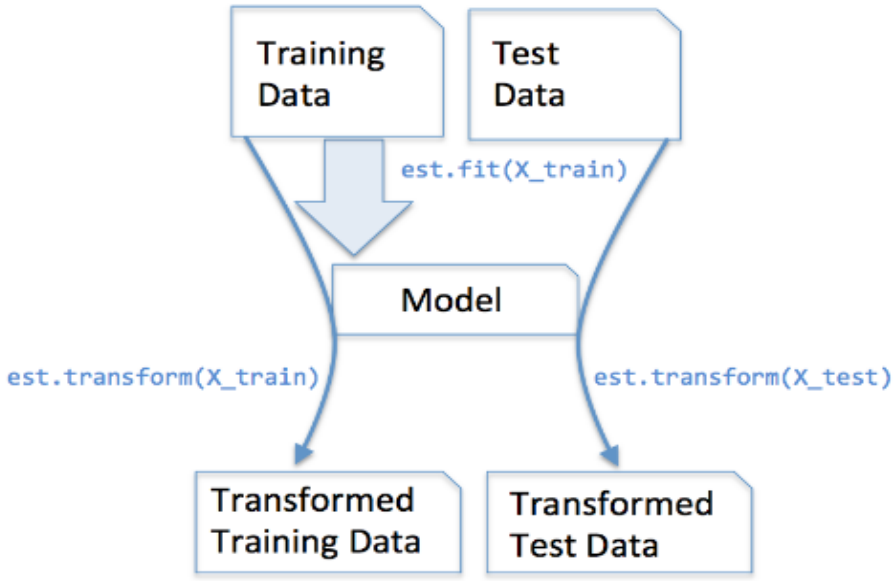
2. **Split dataset dengan presentasi 90% (Train dataset) : 10% (Testing dataset).**

   Setelah melakukan one-hot encoder, akan dilakukan pembagian dataset dengan presentasi 90:10. Pembagian data ini bersifat subjektif tergantung proporsi data yang dimiliki. Semakin besar dataset yang dimiliki maka proposi data train lebih banyak [7].

```
Total of Dataset : 7796
Total of Train Dataset : 7016
Total of Test Dataset : 780
```

3. **Transformasi data menggunakan fungsi Standard Scaler.**

   Standardization adalah metode yang mengubah rentang nilai yang besar menjadi rentang nilai yang ragam. Rentang nilai yang terlalu besar dapat menyebabkan perhitungan jarak menjadi bias yang berpengaruh terhadap model [8]. Transformasi dilakukan setelah melakukan pemisahan data train dan data testing. Hal tersebut bertujuan untuk mencegah bedanya rentang nilai setelah normalisasi [9]. Data training perlu dilakukan fit dan transform, namun data testing hanya perlu melakukan transform. Hal tersebut dikarenakan data testing digunakan untuk menguji model dari hasil standardization data train yang telah dilatih model.



Gambar diatas merupakan gambaran bagaimana normalisasi bekerja pada data train dan testing.

```
scaler.fit(X_train[numerical_kolom])
X_train[numerical_kolom] = scaler.transform(X_train.loc[:, numerical_kolom])
X_train.head()
```

```
X_test[numerical_kolom] = scaler.transform(X_test.loc[:, numerical_kolom])
X_test.head()
```

Modeling

Model yang digunakan pada projek ini ialah *Support Vector Regressor* (SVR), *Boosting* Algorithm, *Linear Regression*.

1. **Support Vector Regressor (SVR)**
   SVR merupakan bagian dari *support vector machine* (SVM). Tipe kernel yang digunakan pada projek ini adalah *radial basis function* (RBF). Hal tersebut dikarenakan data yang tidak terpisah secara linear.

```
from sklearn.svm import SVR

model_svr = SVR(kernel = 'rbf')
```

```
model_svr.fit(X_train, y_train)
```

```
SVR
SVR()
```

2. **Boosting Algorithm**
   Algoritma ini bertujuan untuk meningkatkan performa atau akurasi prediksi dengan cara menggabungkan beberapa model sederhana. Boosting sendiri terbagi menjadi adaptive boosting dan gradient boosting. Pada projek ini akan menggunakan adaptive boosting.

```
from sklearn.ensemble import AdaBoostRegressor
boosting = AdaBoostRegressor(learning_rate = 0.5, random_state = 100)
```

```
boosting.fit(X_train, y_train)
```

```
AdaBoostRegressor(learning_rate=0.5, random_state=100)
```

3. **Linear Regression**
   Linear regression merupakan model yang sering digunakan dalam permasalahan prediksi tipe data numerik.

```
from sklearn.linear_model import LinearRegression
linear = LinearRegression()
```

```
linear.fit(X_train, y_train)
```

```
LinearRegression()
```

Evaluation

Tahapan ini akan mengevaluasi model yang telah terbentuk dengan pendekatan matematis yaitu Mean Squared Error (MSE) dan melihat hasil prediksi dengan data actual.

1. **Mean Squared Error (MSE)**
   MSE merupakan salah satu model evaluasi terbaik pada masalah regresi. MSE sendiri bekerja melakukan perhitungan error antara nilai hasil prediksi dengan nilai sebesarnya. Berikut formula dari MSE [10]. MSE sendiri bekerja melakukan perhitungan error antara nilai hasil prediksi dengan nilai sebesarnya. Berikut formula dari MSE.

$$\mathbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

   Formula diatasla yang digunakan untuk mengevaluasi hasil prediksi model. Semakin besar nilai MSE, maka semakin buruk kinerja model dan sebaliknya [11].

	Model	Train	Testing
0	SVR	0.48	0.39
1	Boosting	0.43	0.39
2	Linear	0.46	0.40

   Terjanya nilai MSE yang dihasilkan tiap model dibawah 0.5. Hal ini menyatakan bahwa kinerja model cukup baik dengan model Boosting memiliki MAE lebih kecil dari sisi train dan testing. Selanjutnya akan menguji data actual yang diambil dari 3 data dari data testing.

	Nilai Rill	SVR	Boosting	Linear
0	4.0	4.303432	3.872449	4.352629
1	5.0	3.913296	4.097528	4.245346
2	5.0	4.899930	4.665452	4.621308
3	4.0	4.336413	4.245389	4.390146
4	4.0	3.881615	4.097528	3.801673

Pada hasil tersebut menyatakan kinerja model baik di averageRating dengan index 0, 2, 3, 4. Hal tersebut searah dengan tingkat MAE yang dihasilkan.

[1] A. Farki, I. Baihaqi, and M. Wibawa, “Pengaruh online customer review rating terhadap kepercayaan place di indonesia,” vol. 5, no. 2, 2016.

[2] N. Li and P. Zhang, “Consumer online shopping attitudes and behavior: An assessment of research,” *Eighth Am. Conf. Inf. Syst.*, no. October 2002, pp. 508–517, 2002.

[3] A. Farhan Hasrul, Suharyati, and R. Sembiring, “Analisis Pengaruh Online Customer Review dan Rating Terhadap Minat Beli Produk Elektronik di Tokopedia,” *KORELASI. Konf. Ris. Nas. Ekon. Manajemen, dan Akuntansi. Fak.*, vol. 2, no. 1, pp. 1352–1365, 2021, [Online]. Available: https://conference.upnvj.ac.id/index.php/korelasi/article/view/1155/857

[4] D. Hediyati and I. M. Suartana, “Penerapan Principal Component Analysis (PCA) Untuk Reduksi Dimensi Pada Proses Clustering Data Produksi Pertanian Di Kabupaten Bojonegoro,” *J. Inf. Eng. Educ. Technol.*, vol. 5, no. 2, pp. 49–54, 2021, doi: 10.26740/jieet.v5n2.p49-54.

[5] R. A. Wibowo, A. A. Kurniawan, T. Elektro, and U. Tidar, “Theta Omega : Journal o f Electrical Engineering , Computer a nd Information Technology,” *J. Electr. Eng. Comput. Inf. Technol.*, vol. 1, no. 2, pp. 1–6, 2020, [Online]. Available: https://jurnal.untidar.ac.id/index.php/thetaomega/article/view/3552

[6] M. Safitri, S. Priansya, R. P. Wibowo, and K. I. T. S. Sukolilo, “Ekstraksi Fitur Modular Kebutuhan Fungsional Ruang Baca Menggunakan Hierarchical Clustering & Pattern Recognition,” *Sesindo 2016*, 2016.

[7] M. S. Baladina, “Perbandingan Nilai Akurasi Algoritma Klasifikasi Data Mining pada Mammographic Mass Dataset UCI Machine Learning,” no. 06211540000120, 2013.

[8] N. Aini, A. Lestari, M. N. Hayati, F. Deny, and T. Amijaya, “Analisis cluster pada data kategorik dan numerik dengan pendekatan Cluster Ensemble (Studi kasus : puskesmas di Provinsi Kalimantan Timur kondisi Desember 2017),” *J. EKSPONENSIAL Vol. 11*, vol. 11, pp. 117–126, 2020.

[9] B. Vrigazova, “The Proportion for Splitting Data into Training and Test Set for the Bootstrap in Classification Problems,” *Bus. Syst. Res.*, vol. 12, no. 1, pp. 228–242, 2021, doi: 10.2478/bsrj-2021-0015.

[10] I. Sungkawa and R. T. Megasari, “Nilai Ramalan Data Deret Waktu dalam Seleksi Model Peramalan Volume Penjualan PT Satria Mandiri Citra Mulia,” *ComTech*, vol. 2, no. 2, pp. 636–645, 2011.

[11] I. Suprayogi, Trimaijon, and Mahyudin, “Model Prediksi Liku Kalibrasi Menggunakan Pendekatan Jaringan Saraf Tiruan (ZST) (Studi Kasus : Sub DAS Siak Hulu),” *J. Online Mhs. Fak. Tek. Univ.*

Riau, vol. 1, no. 1, pp. 1–18, 2014.