

Nama : Rizal Rezaroebajo

NIM : A11.2021.13881

Kelas : STKI A11.4701

## **Model Sistem Rekomendasi Produk E-Commerce Amazon**

### **Ringkasan :**

Perusahaan e-commerce seperti Amazon biasanya menggunakan sistem rekomendasi yang berbeda untuk memberikan saran kepada pelanggan. Amazon saat ini menggunakan *collaborative filtering items*, yang menskalakan kumpulan dataset dan menghasilkan sistem rekomendasi berkualitas tinggi secara real time. Jenis pemfilteran ini mencocokkan setiap item yang dibeli dan diberi peringkat oleh pengguna dengan item serupa, lalu menggabungkan item serupa tersebut ke dalam daftar rekomendasi untuk pengguna. Dalam proyek ini akan dibuat model rekomendasi untuk produk elektronik Amazon. Sistem ini adalah sistem penyaringan informasi yang berusaha untuk memberikan rekomendasi produk teratas yang diminati oleh pengguna.

### **Permasalahan :**

Pelanggan yang mengunjungi website tentu ingin pengalaman yang terbaik. Hal tersebut memicu sistem e-commerce untuk terus beradaptasi dan terus menyesuaikan keinginan pelanggan. Ketika pelanggan melakukan pencarian tanpa adanya rekomendasi lainnya, pelanggan akan cenderung langsung meninggalkan e-commerce dan ini tentu akan merugikan perusahaan.

Pernyataan masalah yang terdapat dalam proyek ini, yakni:

1. Bagaimana caranya agar pelanggan tidak meninggalkan website setelah barang yang diinginkan terpenuhi.
2. Bagaimana tingkat pengunjung dapat meningkat.
3. Bagaimana memberikan sistem rekomendasi yang baik.

### **Tujuan :**

1. Dapat merekomendasikan barang lain sehingga pelanggan tidak cepat pergi dari website.
2. Meningkatkan tingkat pengunjung website dengan menampilkan rekomendasi sehingga pelanggan tetap betah berada di dalam website.
3. Membuat model serta mengevaluasi model dengan metrik RMSE dan LOSS sehingga lebih cepat.

## Alur Penyelesaian :

Proyek ini diselesaikan melalui langkah-langkah berikut :

1. Mengambil dataset dari Kaggle.
2. Melakukan eksplorasi data (EDA) untuk memahami distribusi dataset.
3. Melakukan preprocessing
4. Melatih dua model: RMSE dan LOSS.
5. Mengevaluasi performa model berdasarkan akurasi, loss, dan confusion matrix.
6. Membuat kesimpulan berdasarkan hasil evaluasi.

## Penjelasan Dataset, EDA, dan Proses Features Dataset :

### Dataset :

Berikut fitur-fitur yang terdapat pada dataset dan deskripsinya:

userId = ID user yang telah melakukan pembelian pada produk di website

productId = ID produk yang telah dibeli oleh user

rating = Rating dari produk yang dinilai oleh user pada produk

timestamp = Tanggal terjadinya transaksi

Dataset ini merupakan Dataset *Amazon Product Reviews*. Data pada dataset ini berbentuk: (7824481, 4)

### Data Describe

Deskripsi data-data yang terdapat pada dataset ini ialah:

	rating	timestamp
count	7.824481e+06	7.824481e+06
mean	4.012337e+00	1.338178e+09
std	1.380910e+00	6.900426e+07
min	1.000000e+00	9.127296e+08
25%	3.000000e+00	1.315354e+09
50%	5.000000e+00	1.361059e+09
75%	5.000000e+00	1.386115e+09
max	5.000000e+00	1.406074e+09

## Data Info

Memberikan info berupa tipe data pada setiap kolom yang ada pada dataset.

Column	Dtype
userId	object
productId	object
rating	int8
timestamp	int64

## EDA :

Exploratory Data Analysis

Berikut beberapa teknik visualisasi data beserta insight yang didapatkan dari dataset ini:

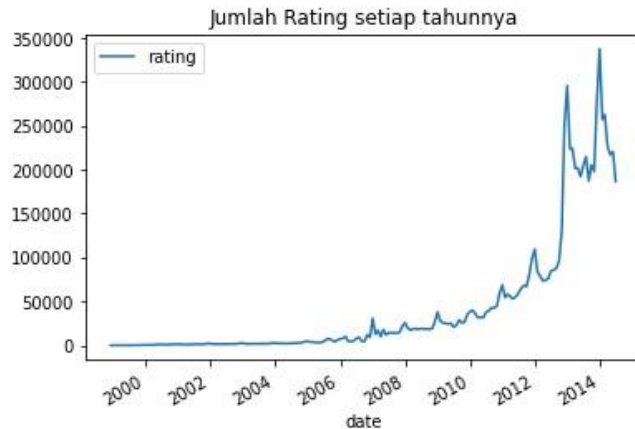
Melihat data berdasarkan waktu dan ratingnya

Melalui data ini dapat diketahui berapa banyak rating yang diterima oleh perusahaan per harinya.

Timestamp	Rating
2014-07-14	9701
2014-07-15	6892
2014-07-16	5943
2014-07-17	4781
2014-07-18	4912
2014-07-19	4183
2014-07-20	4273
2014-07-21	5458
2014-07-22	5010
2014-07-23	695

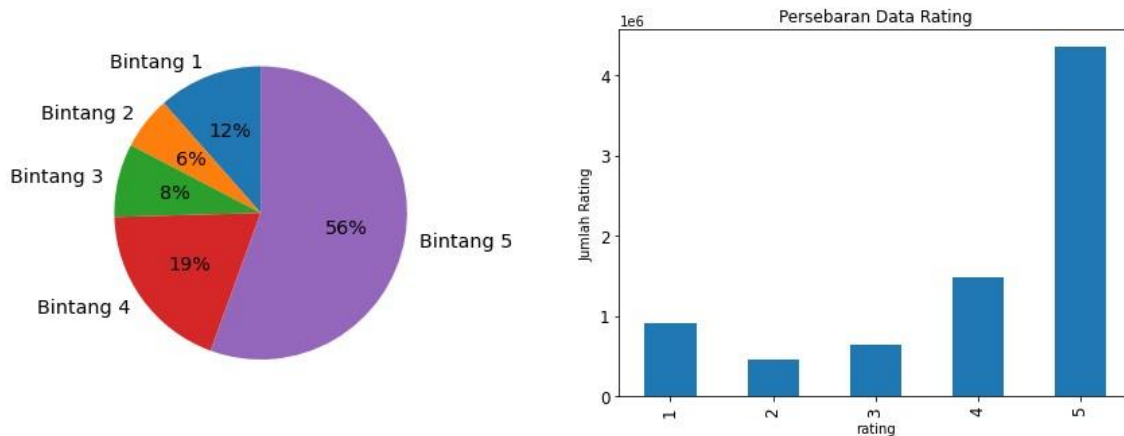
- Melihat grafik trends berdasarkan waktu dan total pemberian rating

Melalui data ini dapat diketahui berapa banyak rating yang didapatkan per tahunnya. Dan diketahui bahwa rating terbanyak berada pada tahun 2014.



- Melihat banyaknya persebaran rating yang diberikan oleh pengguna

Melalui data ini dapat diketahui bahwa terdapat > 50% rating bintang 5 yang diperoleh.



Melihat produk teratas berdasarkan jumlah rating

Kita dapat melihat produk-produk apa sajakah yang paling banyak diberikan rating oleh pengguna.

productId	Jumlah Rating	Rata-rata Rating
B0074BW614	18244	4.491504
B00DR0PDNE	16454	3.931020
B007WTAJTO	14172	4.424005
B0019EHU8G	12285	4.754497
B006GWO5WK	12226	4.314657

### Proses Learning :

Pengolahan data dilakukan dalam beberapa tahap mulai dari mengubah string menjadi list hingga menyusun algoritma. Berikut detail Data Preparation yang dilakukan:

Tahap 1: Memasukkan data, sebelumnya tabel yang ada tidak ada nama kolomnya. Tampilannya seperti di bawah ini:

<b>AKM1MP6P0OYPR</b>	<b>0132793040</b>	<b>5.0</b>	<b>1365811200</b>
A2CX7LUOHB2NDG	0321732944	5.0	1341100800
A2NWSAGRHCP8N5	0439886341	1.0	1367193600
A2WNBOD3WNDNKT	0439886341	3.0	1374451200
A1GI0U4ZRJA8WN	0439886341	1.0	1334707200
A1QGNMC6O1VW39	0511189877	5.0	139743360

Tahap 2: Dengan melihat dokumentasi sumber data, dilakukan rename kolom sehingga akan tampil seperti di bawah ini:

<b>userId</b>	<b>productId</b>	<b>rating</b>	<b>timestamp</b>
A2CX7LUOHB2NDG	0321732944	5.0	1341100800
A2NWSAGRHCP8N5	0439886341	1.0	1367193600
A2WNBOD3WNDNKT	0439886341	3.0	1374451200
A1GI0U4ZRJA8WN	0439886341	1.0	1334707200
A1QGNMC6O1VW39	0511189877	5.0	139743360

Tahap 3: Setelah data tampil, tahap selanjutnya ialah mengecek ketersediaan isi pada kolom, karena ketika ada missing kolom akan menyebabkan sistem rekomendasi kurang optimal

<b>Kolom</b>	<b>Total Kosong</b>
userId	0
productId	0
rating	0
timestamp	0

Pengolahan data sampai di tahap ini, selanjutnya data yang sudah bersih akan diolah pada modelling.

### **Performa Model :**

Modeling pada proyek ini dilakukan dengan *TensorFlow Recommenders* (TFRS). *TensorFlow Recommenders* (TFRS) adalah *library* untuk membuat model sistem pemberi rekomendasi. Model ini dibangun menggunakan *Keras* dan bertujuan untuk memiliki kurva belajar yang lembut sambil tetap memberi fleksibilitas untuk membuat model yang kompleks. Beberapa tahapan yang dilakukan, yakni:

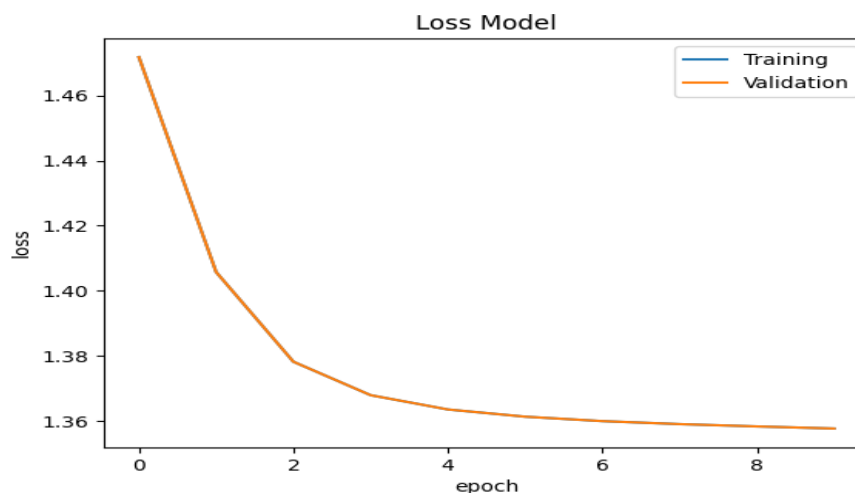
- Tahap 1: Melakukan Import Packages yaitu Import *tensorflow* dan *tensorflow\_recommenders*
- Tahap 2: Mendefinisikan RankingModel kedalam sebuah *class* menggunakan *tf.keras.Model*. Juga terdapat 3 kolom yang didefinisikan, yaitu pada kolom UserID, ProdukID, dan Rating. Yang mana pada setiap kolom terdapat *tf.keras.Sequential* yang digunakan untuk membuat dan melatih *deep learning model*. Pada *tf.keras.Sequential* terdapat beberapa *layer* yakni: *tf.keras.layers.experimental.preprocessing.StringLookup*, *tf.keras.layers.Embedding* untuk kolom UserID dan ProdukID. Sedangkan untuk kolom Rating menggunakan 3 *layer* *tf.keras.layers.Dense*.
- Tahap 3: Membuat amazonModel kedalam sebuah *class* menggunakan *tfrs.models.Model*. Model sebelumnya yakni RankingModel juga dipanggil pada amazonModel. Model ini menggunakan *layer* dari TF Keras. Juga menggunakan *tf.keras.losses.MeanSquaredError* dan *tf.keras.metrics.RootMeanSquaredError* untuk mengukur *metrics*.
- Tahap 4: Setting untuk setiap model yang dilatih, supaya saat model pelatihan tidak terdapat duplikat yang menyebabkan memori berlebih.

```
userIds = recent_prod.userId.unique()
```

```
productIds = recent_prod.productId.unique()
```

```
total_ratings= len(recent_prod.index)
```

- Tahap 5: Data dibagi menjadi dua bagian dengan perbandingan 80% untuk *training* dan 20% untuk *testing*.
- Tahap 6: Kemudian model yang telah dibuat dilatih atau *training* berdasarkan metrik *RMSE* dan *Loss*. Berdasarkan hasil pelatihan data didapat hasil plot seperti di bawah ini:



## Result

- Tahap 7: Uji coba prediksi dengan memanggil *function* model dengan parameter user id. Berikut hasil yang ditampilkan, yaitu Top 5 produk teratas yang direkomendasikan untuk pengguna.

Keterangan : untuk produk ID yang tampil

B002FFG6JC

B004ABO7QI

B006YW3DI4

B0012YJQWQ

B006ZBWV0K

## Evaluation

Pada proyek ini digunakan metrik *RMSE* untuk pelatihan model. *RMSE* merupakan singkatan dari *Root Means Squared Error*. Dalam penggunaan *machine learning* sistem rekomendasi ini, *RMSE* berperan untuk mencari bobot besar berdasarkan *loss* yang besar. *RMSE* digunakan karena tidak menginginkan *outlier* pada model. Dengan begitu, *goals* diawal untuk membuat rekomendasi berhasil dibuat dengan performa yang baik. Sehingga, ketika diterapkan di dalam model bisnis *e-commerce* pengguna tidak akan langsung meninggalkan *e-commerce* dengan cepat karena akan muncul item yang direkomendasikan. Tetapi, pengguna juga dapat melihat-lihat rekomendasi dari produk yang bersangkutan dan hal tersebut dapat menaikkan tingkat keramaian *website*.

Berdasarkan evaluasi diperoleh hasil sebagai berikut:

Metrics	Value
root_mean_squared_error	1.3099995851516724
loss	1.7815556526184082
regularization_loss	0
total_loss	1.7815556526184082

Link Github : <https://github.com/RizalReza23/STKI-A11202113881-UAS>