Prosiding Nasional Rekayasa Teknologi Industri dan Informasi XIV Tahun 2019 (ReTII)
November 2019, pp. 263~269
ISSN: 1907-5995

Exploratory Data Analysis dalam Konteks Klasifikasi Data Mining

Eka Dyar Wahyuni[1], Amalia Anjani Arifiyanti[2], Mashita Kustyani[3]
[1, 2, 3] Program Studi Sistem Informasi, Universitas Pembangunan Nasional "Veteran" Jawa Timur
Korespondensi : ekawahyuni.si@upnjatim.ac.id

ABSTRAK

Setiap jenis data memiliki memiliki bentuk dan perilaku yang beragam. Oleh karenanya, proses analisis data tidak dapat dibakukan melalui suatu prosedur tertentu. EDA merupakan salah satu pendekatan analisis data yang dapat digunakan dalam memaknai informasi yang terkandung pada data. EDA bersifat fleksibel menyesuaikan perilaku data dan dapat digunakan untuk observasi data melalui berbagai sudut pandang. Strategi analisis data ini dapat digunakan untuk melengkapi hasil analisis klasifikasi data mining yang digunakan untuk mengenali pola data. Pada artikel ini akan dijelaskan bahwa EDA dapat membantu dalam memperkaya hasil analisis data dan membantu dalam tahapan praproses klasifikasi data mining. Klasifikasi data mining dilakukan dengan menggunakan algoritma *Naive Bayes Classifier dan Regresi Logistik*.
Kata kunci: *Data Mining, Exploratory Data Analysis, Klasifikasi, Naive Bayes Classifier, Regresi Logistik*

ABSTRACT

Each type of data has various forms and behaviors. Therefore, the process of data analysis cannot be standardized through a certain procedure. EDA is a data analysis approach that can be used in interpreting the information contained in data. EDA is flexible in adjusting data behavior and can be used for data observation through various points of view. This data analysis strategy can be used to complement the results of data mining classification analysis used to recognize data patterns. In this article it will be explained that EDA can help in enriching the results of data analysis and assist in the pre-processing stages of data mining classification. Data mining classification is done using the Naive Bayes Classifier algorithm and Logistic Regression.
Keyword : Data Mining, Exploratory Data Analysis, Classification, Naive Bayes Classifier, Regresi LogistikLogistic Regression

1. PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi saat ini mendukung berbagai jenis organisasi atau bisnis mengumpulkan data yang mereka butuhkan untuk diolah dan menghasilkan informasi yang dibutuhkan oleh organisasi atau bisnis tersebut. Proses analisis data dilakukan untuk menghasilkan informasi yang sebelumnya diharapkan oleh pengolah data dan bahkan dapat menghasilkan informasi atau fenomena baru yang sebelumnya belum pernah terbayangkan tersimpan pada data tersebut.

Salah satu strategi dalam analisis data adalah dengan cara Exploratory data analysis (EDA). Pada EDA, dilakukan eksplorasi data dengan berbagai cara hingga suatu informasi yang masuk akal muncul. EDA bukanlah suatu tipe model atau prosedur baku yang sudah terdefinisikan sebelumnya. Pengaplikasian EDA tergantung pada konteks dan tergantung pada rincian analisis. Dengan EDA, data dieksplorasi dengan berbagai perspektif, sehingga dapat dinyatakan bahwa EDA tidak terpaku pada suatu teknik yang baku. Hal ini dikarenakan, EDA memiliki karakteristik yang fleksibel yang diperlukan untuk melakukan identifikasi dan investigasi suatu fenomena yang muncul pada saat melakukan penelitian empiris (1). EDA melengkapi confirmatory data analysis (CDA) tradisional yang mana dilakukan dengan cara menghasilkan hipotesis, serta menemukan outlier dan asumsi yang dapat mempengaruhi validitas CDA. CDA terbatas pada paradigma bahwa semua jenis data memiliki model probabilitas yang cukup baik, formulasi pertanyaan yang sama untuk semua jenis data, dan hanya melakukan penilaian atas apa yang sudah pasti pada data tersebut. Kondisi tersebut tidak sesuai bahkan beberapa kasus bertolak belakang dengan kondisi data yang ada (2). Oleh karenanya EDA digunakan untuk melengkapi CDA dan hal ini dijelaskan melalui beberapa riset (3) (4) (5) (6) yang menjabarkan bahwa kedua hal tersebut dapat saling melengkapi untuk meningkatkan pemahaman terhadap data dan peningkatan kualitas hasil analisis data .

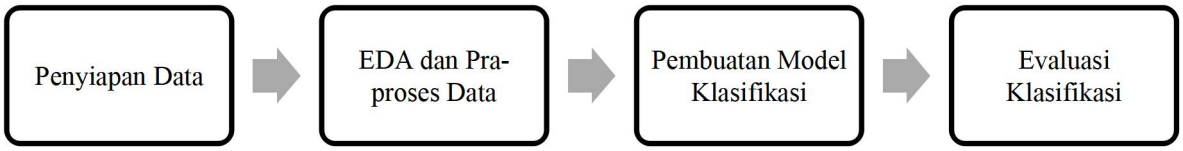
EDA tradisional dikenal terdiri dari empat hal berikut (7): 1) *Display*; 2) *Residual*; 3) *Re-expression*; dan 3) *Resistance*. *Diplay* dapat memperlihatkan pola data dan struktur analisis melalui tampilan visual. *Residual* terfokus pada sisa data hasil analisis. Fungsi matematika digunakan untuk menjelaskan tingkah laku

data dan klarifikasi analisis, hal ini merupakan bagian dari *re-expression*. Pada *resistance*, memastikan bahwa hasil analisis data tidak hanya dipengaruhi oleh segelintir data saja.

Tujuan EDA adalah mencari pola data. Hal ini sejalan dengan konsep data mining yang juga digunakan untuk mengekplorasi pola dari suatu data. Dengan adanya gempuran data dalam jumlah besar dalam era big data, eksplorasi pola data menjadi lebih rumit dikarenakan dimensi data yang terlalu besar. EDA digunakan dalam tujuan pengurangan dimensi data atau memperkaya pemahaman atas analisis data melalui visualisasi data (8). EDA juga digunakan diantaranya untuk mengoptimalkan pengetahuan mengenai data, menghasilkan variabel yang penting, mendeteksi outlier dan anomali pada data, dan menguji asumsi awal (9). Hal-hal inilah yang dapat digunakan dalam memperkaya analisis data dan membantu mengoptimalkan hasil klasifikasi dengan pendekatan data mining.

2. METODE PENELITIAN

Analisis data pada penelitian ini akan menggabungkan tahap EDA dalam eksplorasi data awal yang selanjutnya data tersebut akan dimodelkan berdasarkan metode klasifikasi data mining. Penelitian ini melalui beberapa tahap yang dapat dilihat pada gambar 1 berikut ini. Keseluruhan tahap akan diimplementasikan dengan mempergunakan bahasa Python.



Gambar 1. Tahapan Penelitian

2.1 Penyiapan Data

Dataset yang digunakan dalam penelitian ini adalah dataset "*Women Clothing E-Commerce Reviews*" yang didapatkan dari situs Kaggle.com. Dataset tersebut merupakan data review yang ditulis oleh pelanggan pada suatu *e-commerce* yang menjual pakaian wanita. Dataset ini berekstensi csv, dengan kolom sebagai berikut: *Unnamed* yang menunjukkan nomor urut baris, *Clothing ID* untuk *id clothing*. *Age* yang menunjukkan usia reviewer. *Title* yang menunjukkan judul ulasan. *Review Tex*t untuk menyimpan Teks Review. Kolom rating untuk menyimpan skor produk yang diberikan oleh pelanggan, dengan nilai antara 1 yang berarti rating produk tersebut buruk, hingga 5, yang berarti rating produk tersebut baik. *Recommended* IND yang menyatakan bahwa pelanggan merekomendasikan produk di mana 1 direkomendasikan dan 0 tidak direkomendasikan. *Positive Feedback Count* yang mendokumentasikan jumlah pelanggan lain yang menganggap ulasan ini positif. *Division Name* yang menyimpan nama divisi dari suatu produk. *Department Name*, yang mendokumentasikan nama departemen dari suatu produk. Dan *Class Name*, untuk menyimpan nama kelas dari suatu produk.

2.2 EDA dan Pra-proses Data

Pada tahap ini dilakukan eksplorasi data dengan menggunakan fungsi statistik, matematik, dan divisualisasikan dalam bentuk grafik. Hal ini akan mempermudah dalam pemahaman terhadap data dan pola dasar data. Selain itu, pendekatan tersebut juga digunakan untuk melihat dan mencari outlier yang ada pada data. Pada tahap ini juga dilakukan pra-proses data misalnya pembersihan data outlier dan penanganan data yang kosong, sehingga siap diolah/dianalisa pada tahap pembuatan model klasifikasi.

2.3 Pembuatan Model Klasifikasi

Data yang digunakan untuk klasifikasi adalah atribut *Review Tex*t. Atribut tersebut akan dipelajari oleh algoritma Naive Bayes Classifier dan Regresi Logistik untuk mengenali pola data yang bersifat review positif dan review negatif (ditentukan dari kolom Recommend IND). Seluruh data pada atribut tersebut yang telah dibersihkan dan ditransformasikan pada tahap sebelumnya, dibagi menjadi dua jenis data yaitu data latih dan data uji. Pembagian data ini menggunakan metode hold-out dengan proporsi pembagian data latih dan data uji sebesar 90%:10%.

2.4 Evaluasi

Evaluasi model klasifikasi yang dibuat pada tahap sebelumnya diuji pada tahap ini. Tahap pengujian dilakukan dengan menggunakan data uji. Evaluasi model klasifikasi dianalisis dengan membandingkan nilai akurasi secara keseluruhan dan waktu yang diperlukan untuk melakukan training model.

3. HASIL DAN ANALISIS

Hasil dari analisis data dengan EDA dan pendekatan klasifikasi dengan agoritma Naive Bayes Classifier dan Regresi Logistik dijelaskan pada bagian ini.

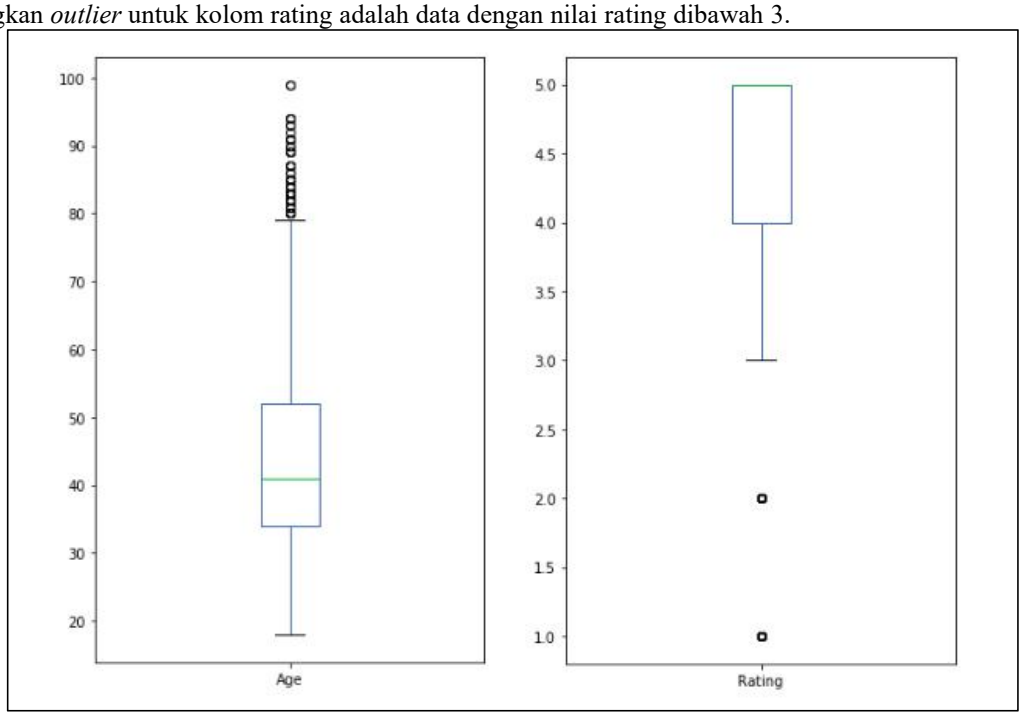
3.1 EDA & Praproses

Langkah pertama dalam mendeteksi outlier dan anomali dalam data adalah dengan mendapatkan ringkasan statistik dari data yang dimiliki, antara lain rata-rata, nilai minimum, nilai maksimum, jumlah data dll. Untuk mendapatkan informasi tersebut, dipergunakan fungsi describe() dari *library* panda. Hasil dari fungsi ini ditunjukkan dalam gambar 2.

	Unnamed: 0	Clothing ID	Age	Rating	Recommended IND	Positive Feedback Count
count	23486.000000	23486.000000	23486.000000	23486.000000	23486.000000	23486.000000
mean	11742.500000	918.118709	43.198544	4.196032	0.822362	2.535936
std	6779.968547	203.298980	12.279544	1.110031	0.382216	5.702202
min	0.000000	0.000000	18.000000	1.000000	0.000000	0.000000
25%	5871.250000	861.000000	34.000000	4.000000	1.000000	0.000000
50%	11742.500000	936.000000	41.000000	5.000000	1.000000	1.000000
75%	17613.750000	1078.000000	52.000000	5.000000	1.000000	3.000000
max	23485.000000	1205.000000	99.000000	5.000000	1.000000	122.000000

Gambar 2. ringkasan statistik dari *Women Clothing E-Commerce Reviews*

Dari gambar 2 diatas, dapat diamati bahwa dataset memiliki 6 kolom dan 23486 baris. Dari nilai min, 25%, 50%, 75% dan max, dapat diamati bahwa terdapat beberapa kolom yang terindikasi mengandung *outlier*, terutama kolom *Age* dan Rating. Cara paling mudah mendeteksi *outlier* adalah mempergunakan box-plot seperti yang ditunjukkan pada Gambar 2. *Outlier* untuk kolom *Age* adalah data dengan nilai diatas 80 tahun. Sedangkan *outlier* untuk kolom rating adalah data dengan nilai rating dibawah 3.

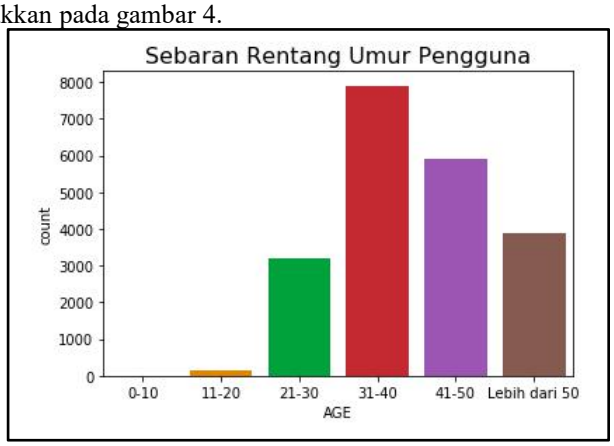


Gambar 3. box-plot untuk kolom *Age* dan Rating

Selain untuk menemukan *outlier*, EDA juga dapat dipergunakan untuk menemukan apakah suatu kolom mengandung *missing value* atau tidak, dengan mempergunakan fungsi isnull().sum(). hasil dari fungsi ini adalah kolom *Review Text, Department Name* dan juga *Class Name* memiliki *missing value*, masing-masing terdiri dari 845 baris,14 baris dan 14 baris. Setelah mengetahui bahwa dataset yang dimiliki mengandung *outlier* dan *missing value*, skenario yang bisa dilakukan ada berbagai macam cara, bisa menghapus baris yang mengandung *missing value* ataupun *outlier*, atau dengan mengisi baris yang mengandung *outlier* tersebut

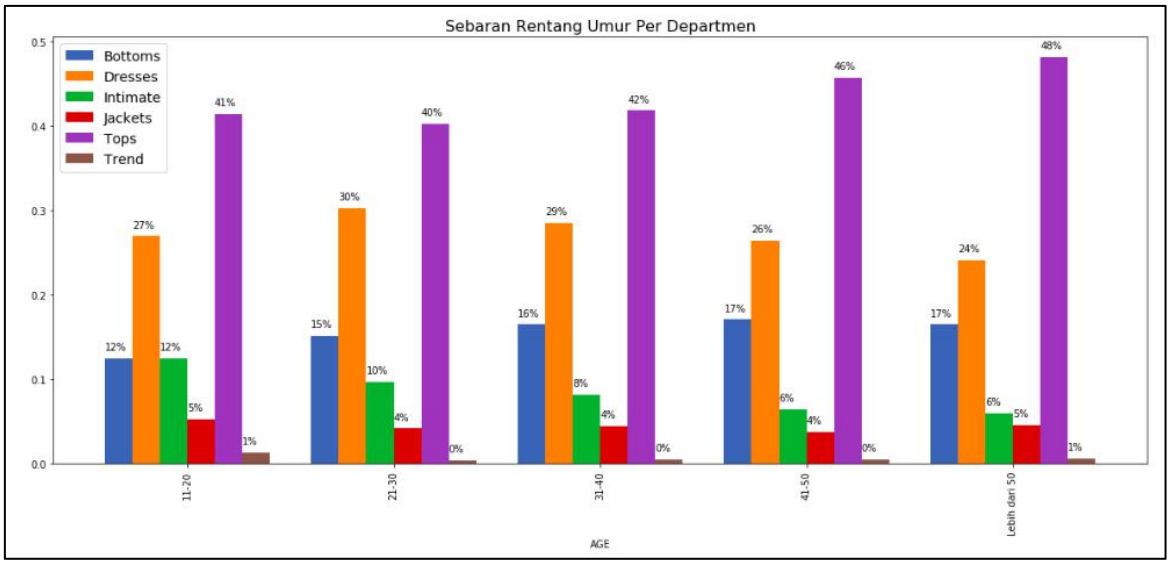
dengan suatu nilai, bisa berupa *mean*, maksimum, minimum atau bahkan dengan suatu nilai konstan. Skenario yang diambil dalam penelitian ini adalah menghapus kolom yang mengandung missing value dan juga outlier dengan mempergunakan fungsi dropna() dan drop(). Selain menghapus kolom, praproses yang dilakukan dalam penelitian ini adalah mengubah teks yang ada di kolom *Review Text* menjadi huruf kecil semua mempergunakan fungsi lowercase. Proses selanjutnya adalah tokenisasi yang akan memecah teks menjadi kata kemudian menghilangkan special karakter (simbol dan tanda baca), sehingga hanya tersisa huruf dan angka.

EDA juga berfungsi untuk mengoptimalkan pengetahuan mengenai data. Salah satu metode tradisional dalam EDA adalah visualisasi dalam bentuk grafik. Salah satu kolom yang coba diteliti lebih lanjut adalah kolom umur. Untuk memudahkan visualisasi, kolom umur dimasukkan ke dalam nilai yang berupa rentang, seperti ditunjukkan pada gambar 4.



Gambar 4. Sebaran Rentang Umur Pengguna

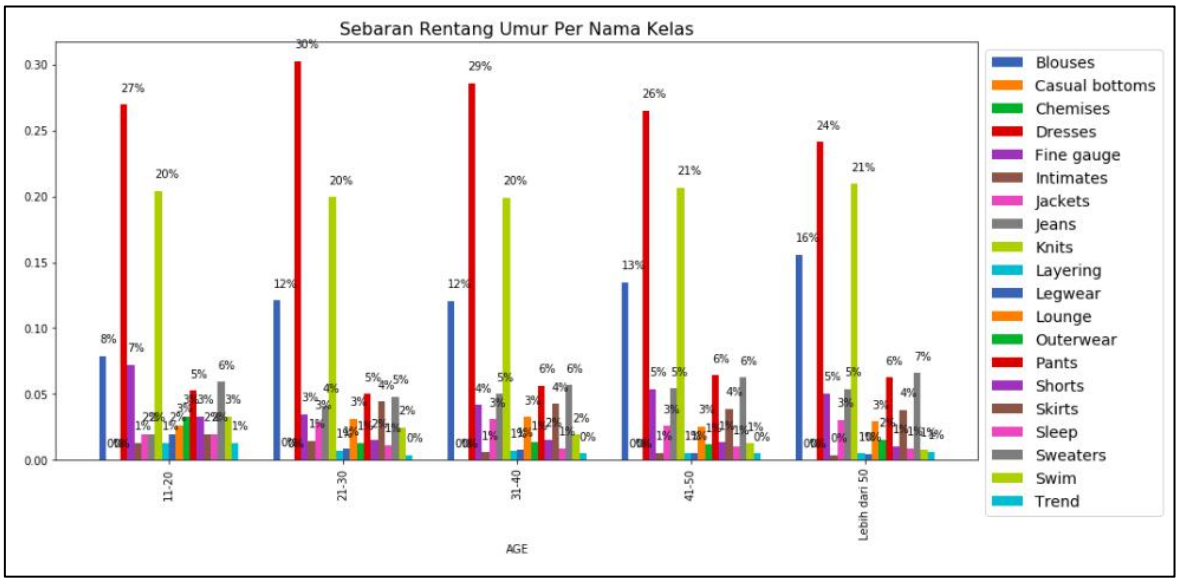
Dari gambar 4, dapat diamati bahwa pengguna terbanyak ada pada rentang umur 31-40, kemudian diikuti dengan pengguna dari rentang umur 41-50 dan umur 50 tahun keatas.



Gambar 5. Sebaran rentang umur per department

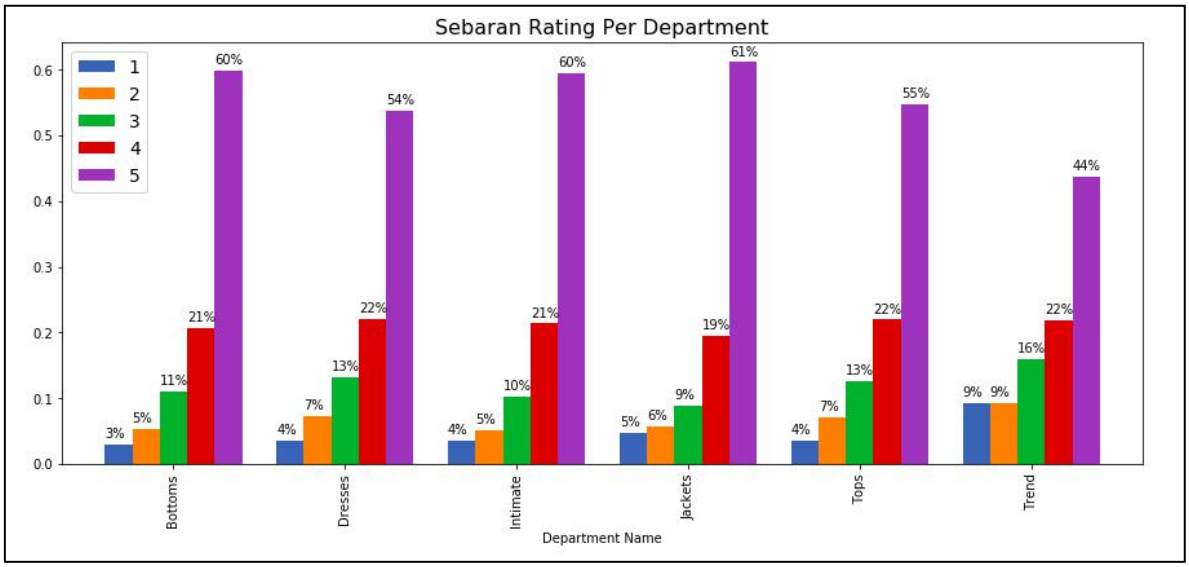
Jika diasumsikan bahwa hanya pengguna yang sudah bertransaksi (pengguna yang sudah melakukan pembelian) yang bisa memberikan komentar dan rating, dari gambar 5 dapat ditarik kesimpulan bahwa kelompok umur mayoritas (rentang 31-40, 41-50 dan lebih dari 50), lebih sering membeli barang dari department *tops*, kemudian diikuti *dress* dan *bottoms* dibandingkan barang dari department lainnya. Ketika grafik tersebut didetilkan di level bawahnya lagi, seperti yang ditunjukkan pada gambar 6, pengguna dari 3 kelompok umur ini, lebih sering membeli produk dari kelas *dress, knits* dan *blouse*. Dari 2 grafik tersebut, dapat diamati bahwa ada perubahan posisi urutan ketika data dilihat sebarannya berdasarkan departmen dan nama kelas. Ketika dibandingkan dengan departmen, *tops* ada pada urutan pertama, *dress* ada pada urutan kedua. Sedangkan ketika dibandingkan dengan nama kelas, *dress* ada pada urutan pertama, sedangkan *blouse*,

yang merupakan salah satu jenis dari *tops* ada pada urutan ke 3. Kesimpulan sederhana yang dapat ditarik adalah, department *tops* memiliki lebih dari satu kelas, dan salah satu kelasnya adalah *blouse*.

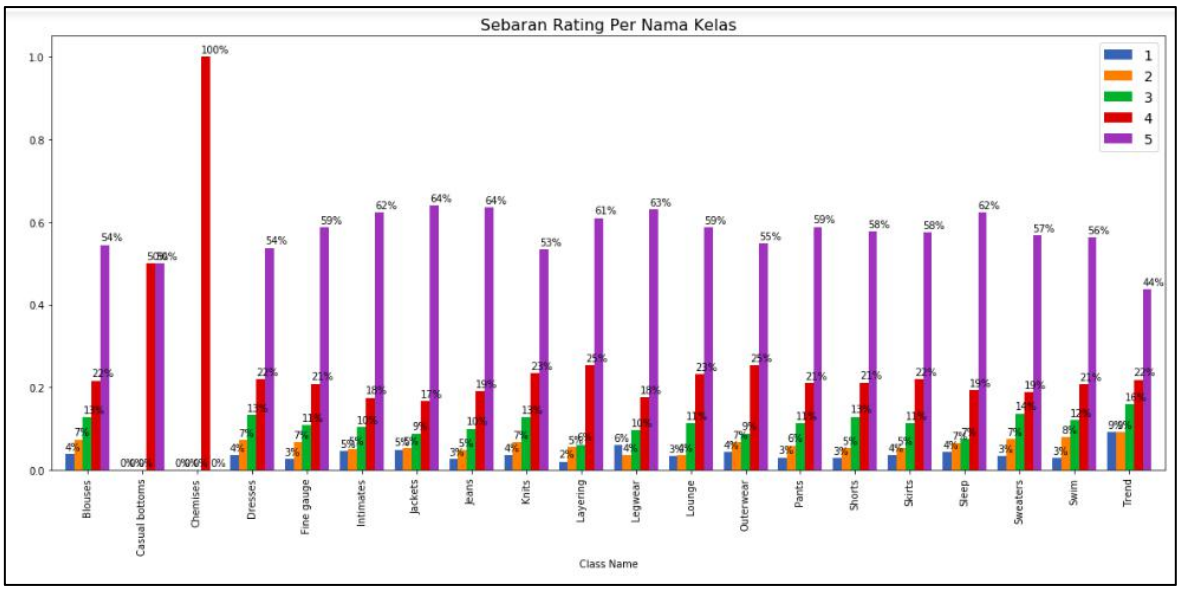


Gambar 6. Sebaran Rentang Umur Per Nama Kelas

Kolom selanjutnya yang dapat diteliti adalah rating, sebaran rating per departmen dapat dilihat pada gambar 7. Banyak pengguna yang puas (dengan memberikan rating 5) terhadap produk dari department *jacket*, *bottoms* dan *intimate*. Ketika grafik ini didetilkan per nama kelas seperti yang ditampilkan pada gambar 8, produk yang paling banyak menerima rating 5 adalah dari kelas *jacket*(64%), *jeans*(64%), *legwear*(63%), *intimate*(62%) dan *sleep*(62%). Jika diteliti lebih jauh data yang dimiliki, dapat dilihat bahwa produk dari departmen *jacket*, selain masuk ke kelas *jacket* ada yang masuk ke kelas *jeans* (mungkin karena jaket tersebut berbahan jeans).

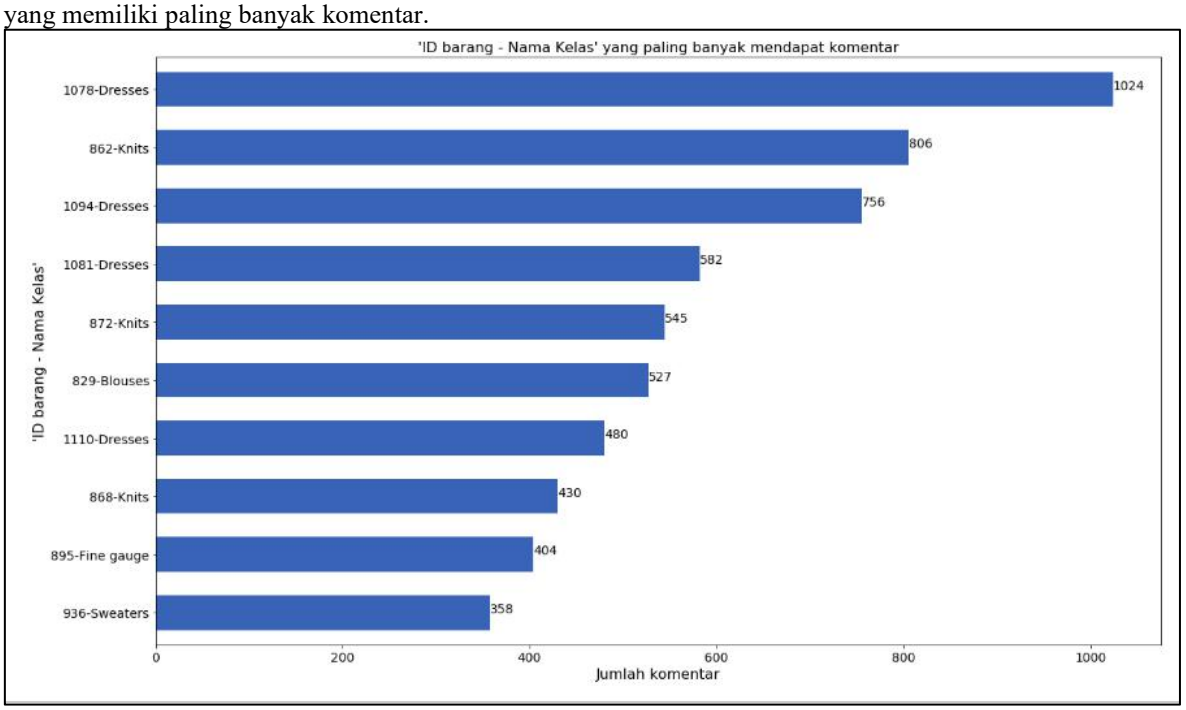


Gambar 7. Sebaran Rating Per Department



Gambar 8. Sebaran Rating Per Nama Kelas

Kolom selanjutnya yang coba diolah adalah review text dibandingkan dengan *clothing*-id. Karena dataset tidak memiliki kolom yang menyimpan nama produk, informasi terdekat yang bisa menggambarkan produk adalah *clothing* ID. Gabungan antara kolom *clothing* ID dan juga kolom *Class Name* lebih dapat memberikan informasi produk jika dibandingkan hanya dengan memakai kolom *Clothing ID* saja. Dengan asumsi hanya pengguna yang telah membeli suatu produk yang bisa memberikan komentar terhadap produk tersebut, dari gambar 9 dapat ditarik kesimpulan bahwa 3 produk yang paling banyak terjual adalah produk dengan id 1078 dari kelas *dresses,* id 862 dari kelas *knits* dan 1094 dari kelas *dresses*, karena 3 produk ini yang memiliki paling banyak komentar.



Gambar 9. 10 produk dengan komentar terbanyak

3.2 Pembuatan Model Klasifikasi

Penelitian ini akan membangun 2 model dan membandingkan kinerja dari 2 model ini. Model dibangun dengan mempergunakan algoritma Naive Bayes *Classifier* dan Regresi Logistik. Agar bisa memakai algoritma Naive Bayes *Classifier* dan Regresi Logistik, *library* LogisticRegression dan MultinomialNB perlu diimport sebelumnya.

3.3 Evaluasi

Dua model yang sudah dibangun dipergunakan untuk memprediksi data uji yang sama. hasil yang direkam adalah berupa nilai akurasi keseluruhan dan waktu yang diperlukan untuk melakukan training. Hasil evaluasi direkam dalam tabel 1.

Tabel 1. hasil evaluasi model

Model	Akurasi	Waktu untuk Training Model
I (Naive Bayes)	88,73%	0:00:00.012425
II (Regresi Logistik)	88,55%	0:00:01.448921

Dari tabel diatas, dapat disimpulkan bahwa naive bayes menghasilkan nilai akurasi yang hampir sama jika dibandingkan dengan regresi logistik, tetapi waktu training yang diperlukan naive bayes lebih cepat sepersekian detik dibandingkan dengan regresi logistik. Dengan akurasi yang hampir sama, besarnya waktu yang diperlukan akan berpengaruh signifikan, jika data yang dipakai berukuran lebih besar.

4. KESIMPULAN

EDA dapat dipergunakan untuk membantu tahapan praproses dalam data mining (klasifikasi) dengan cara menampilkan keberadaan missing value dan juga outlier. EDA juga dapat mengoptimalkan pengetahuan mengenai data, yang dapat digunakan untuk memperkaya pemahaman atas analisis data. Evaluasi antara model naive bayes ketika dipergunakan untuk mengklasifikasikan teks berbahasa inggris kedalam kelas positif dan negatif menghasilkan nilai akurasi dan juga waktu training yang lebih baik dibandingkan model yang dibangun dengan mempergunakan regresi logistik.

UCAPAN TERIMAKASIH

Ucapan terimakasih kepada Tita Ayu Rospricilia dan I Gede Okta Budi M yang membantu penulis sehingga dapat menyelesaikan paper ini.

DAFTAR PUSTAKA

[1] Behrens JT. Principles and procedures of exploratory data analysis. Psychol Methods. 1997;2(2):131–60.
[2] Cleveland WS. The Collected Works of John W. Tukey: Graphics 1965-1985. Chapman & Hall; 1988
[3] Perer A, Shneiderman B. Integrating Statistics and Visualization: Case Studies of Gaining Clarity during Exploratory Data Analysis. In: Proceedings of the SIGCHI conference on Human Factors in computing systems. Florence, Italy: ACM; 2008. p. 265–74.
[4] Theus M, Urbanek S. Interactive Graphics for Data Analysis. 2008
[5] Jebb AT, Parrigon S, Woo SE. Exploratory data analysis as a foundation of inductive research. Human Resource Management Review. 2017 Jun 1;27(2):265–76.
[6] De Mast J, Trip A. Exploratory data analysis in quality-improvement projects. Journal of Quality Technology. 2007;39(4):301–11.
[7] Velleman PF, Hoaglin DC. Applications, Basics, and Computing of Exploratory Data Analysis. Washington: The Internet-First University Press; 2014
[8] Wendy L. M, Angel R M, Solka J. Exploratory data analysis with MATLAB. Chapman and Hall/CRC; 2017.
[9] Yu CH. Exploratory data analysis in the context of data mining and resampling. International Journal of Psychological Research. 2010;3(1):9–22.