Multidimensional Clothing Review Analysis

Introduction To Big Data and Data Analytics (20B12CS333)

Department of CSE/IT

Jaypee Institute of Information Technology University, Noida

Team Members:

Herumb Shandilya	**(18103164)**
Dini Jain	**(18103130)**
Karan Malhotra	**(17102121)**
Ayush Kumar	**(18104031)**

Submitted to:

Prof. Megha Rathi

ABSTRACT

Fashion brands of all sizes and specialties are using technology to understand customers better than ever before. As those data collection efforts grow more sophisticated, artificial intelligence will reshape brands' approach to product design and development, with a focus on predicting what customers will want to wear next.

Understanding customer sentiments is of paramount importance in marketing strategies today. Not only will it give companies an insight as to how customers perceive their products and/or services, but it will also give them an idea on how to improve their offers. The is to understand the correlation of different variables in customer reviews on a women clothing e-commerce, and to classify each review whether it recommends the reviewed product or not. To achieve these goals, we employed univariate and multivariate analyses on dataset features except for review titles and review texts and we used Random Forest Classifier for sentiment classification.

The results reflected the link between consumer behavior and the traditional clothing market, and provided guidelines for fashion store managers to improve their marketing strategies. Sentiment analysis through collection of responses (likes, shares, comments, re-tweets) helps the fashion store managers to analyze every aspect of consumers demand from the most loved colour to the most acceptable fits. Using these tools, the paper analyzed several designs to derive visual insight, producing a first-of-its-kind analysis of per-city fashion choices and spatio-temporal trends of modern civilization.

Keywords: *sentiment analysis, machine learning, Random Forest*

INTRODUCTION

This is a Women's Clothing E-Commerce dataset revolving around the reviews written by customers. Its nine supportive features offer a great environment to parse out the text through its multiple dimensions. In this project, we attempt to analyze the customer reviews on women clothing e-commerce by employing statistical analysis and sentiment classification. We first analyze the non-text review features (e.g. Age groups that are likely to recommend the clothing line. , The probability of a clothing line with high ratings getting recommended class of dress purchased, etc.) found in the dataset, as an attempt to unravel any connection between them and customer recommendation on the product. Then, we implement a random forest model for classifying whether a review text recommends the purchased product or not.

According to Gul Kaner and Aykut Coskun (2017), Experts agreed that collaborative work between fashion and technology is essential to design fashionable, desirable and functional wearable technologies. They stated that they were willing to participate in such a collaboration. They shared their insights about stakeholders that should be actively involved in the collaboration, description of the collaborative product development process, the characteristics of the collaboration environment and barriers for a successful collaboration. In the remainder of this section, we present these insights.

BIG DATA ANALYSIS IN FASHION PERSPECTIVE

The ultimate big data and decision support application can do:

I. Data extraction

II. Provide data quality correction

III. Data cleansing, transformation, and preprocessing

IV. Provide multi-dimensional data visualization

V. Generate comprehensive report

VI. Deployment strategy

BACKGROUND STUDY

Sales, marketing and advertising have become very important today in the modern society. We can collect data on human choice and behaviour and hence the popularity of the brand in that area whereas other civilized societies and cultures are using Big Data, generated to improve its presence and Customer Engagement all over.

SENTIMENTAL ANALYSIS:

Organisations are innovating new methods to collect information about customers and use it to provide a personalized overall experience. While e-commerce itself is expanding everyday, customers are still somewhat hesitant and data speaks about their engagement. There are aspects of it that can't be changed. Customers indeed have got the opportunity to buy clothes from the convenience of home, but there is still a very high probability of their satisfaction that comes with the product and comfort that accompanies the product. Fitting remains among the top concerns of the people when they shop online. And when not satisfied well, it's a concern for additional cost of retailers as the orders with incorrect sizes from customers get spent into support, and return orders. Social Media is increasing luxury and fashion since customers are empowered with all the tools of data analysis.

CUSTOMER ENGAGEMENT :

Today, almost all organisations outlay a huge volume of data about their customers and their orders or requests or behaviour. This information can be used to monitor their possible next purchase used for marketing and forecasting about their purchases. The focus of marketing applications in clothing Industries and outlets have gotten bigger from not only identifying new customers but also to measuring the value of the customer and then taking steps to return the profitable customers. This has happened because it is really expensive to acquire new customers compared to retaining the customers who have already purchased some clothes from a company or brand. A numerous analysis on data generated by them is done using various methods to generate the customer value (this can be expected to increase the value of a firm to a huge margin) for

Fashion Industries and customers. Different Data Analysis tools are used to model customer experience.

1.3 RESEARCH ORIENTATION

The purpose is also to:

I. To highlight the role of big data analysis in the fashion industry.

II. To find data analysis methods that can also be brought to the fashion sector depending on the customers.

III. To solve the challenges in Data analysis/pre-processing as applied to the choices and practices of customers.

IV. To improve the chances of investment for the company

SYSTEM ARCHITECTURE :

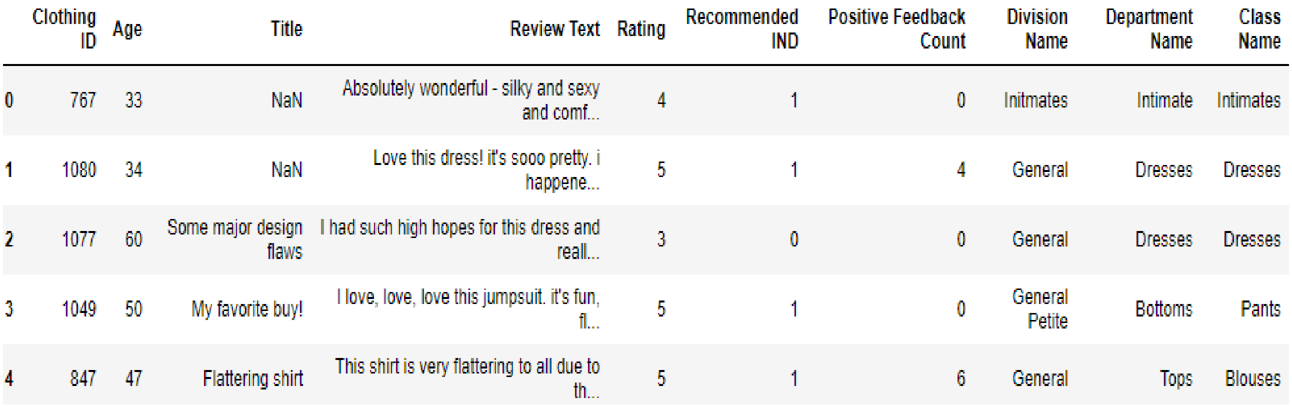
Description:

This dataset includes 23486 rows and 10 feature variables. Each row corresponds to a customer review, and includes the variables:

- **Clothing ID**: Integer Categorical variable that refers to the specific piece being reviewed.
- **Age**: Positive Integer variable of the reviewers age.
- **Title**: String variable for the title of the review.
- **Review Text**: String variable for the review body.
- **Rating**: Positive Ordinal Integer variable for the product score granted by the customer from 1 Worst, to 5 Best.
- **Recommended IND**: Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended.
- **Positive Feedback Count**: Positive Integer documenting the number of other customers who found this review positive.
- **Division Name**: Categorical name of the product high level division.

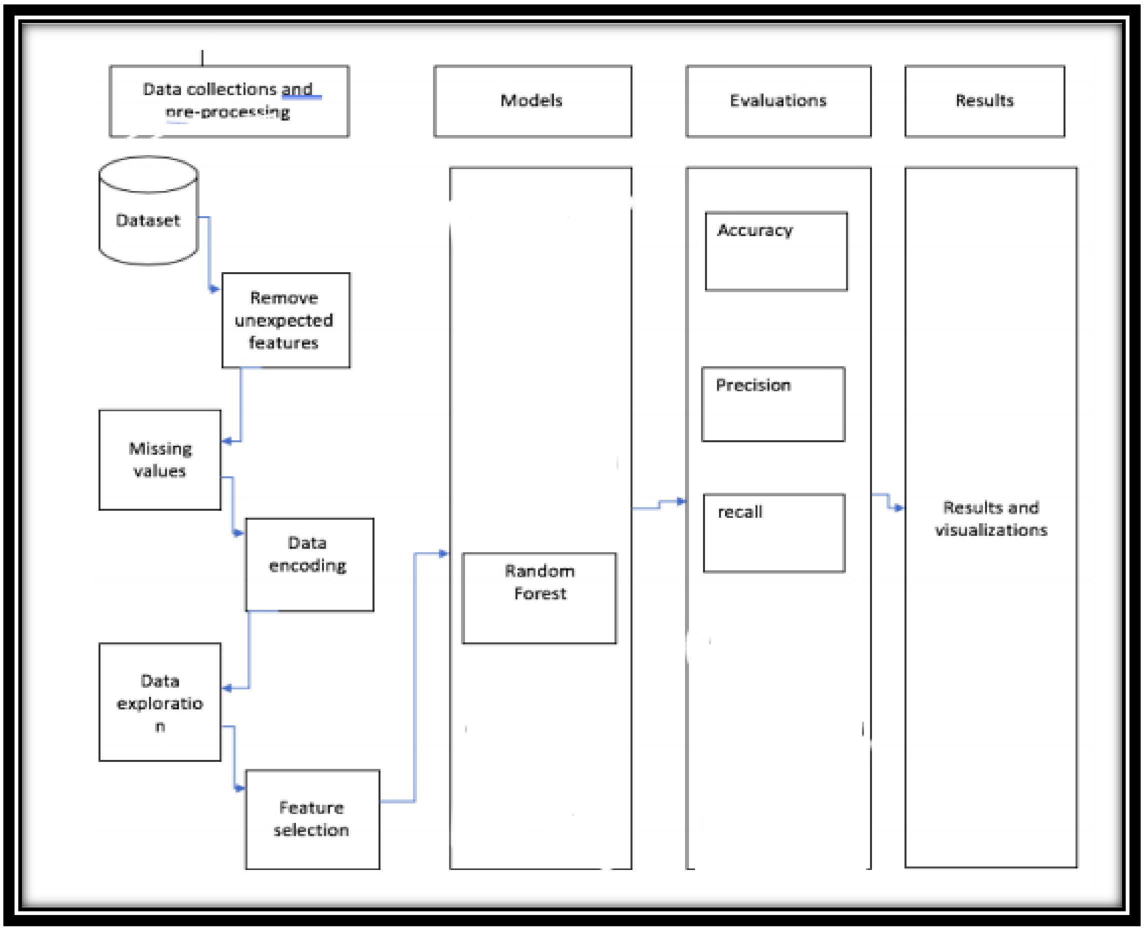
- **Department Name**: Categorical name of the product department name.
- **Class Name**: Categorical name of the product class name.

Dataset



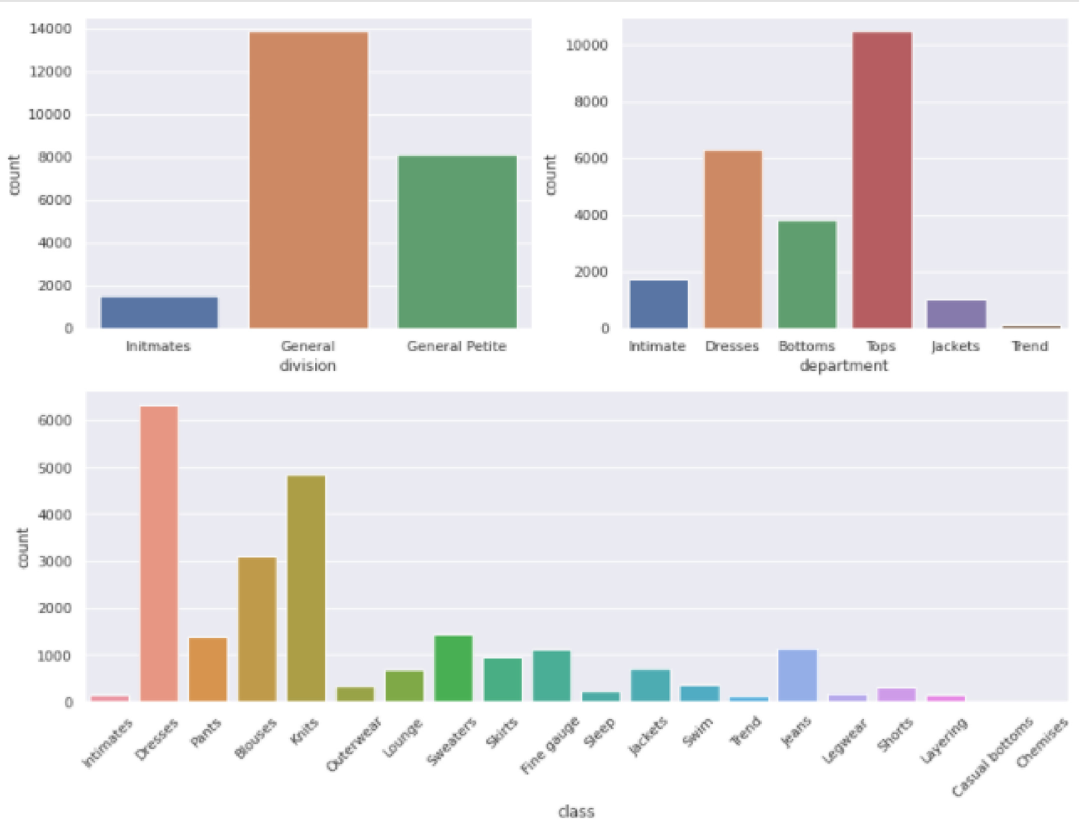
	Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
0	767	33	NaN	Absolutely wonderful - silky and sexy and comf...	4	1	0	Initmates	Intimate	Intimates
1	1080	34	NaN	Love this dress! it's sooo pretty. i happene...	5	1	4	General	Dresses	Dresses
2	1077	60	Some major design flaws	I had such high hopes for this dress and reall...	3	0	0	General	Dresses	Dresses
3	1049	50	My favorite buy!	I love, love, love this jumpsuit. it's fun, fl...	5	1	0	General Petite	Bottoms	Pants
4	847	47	Flattering shirt	This shirt is very flattering to all due to th...	5	1	6	General	Tops	Blouses

Flowgraph



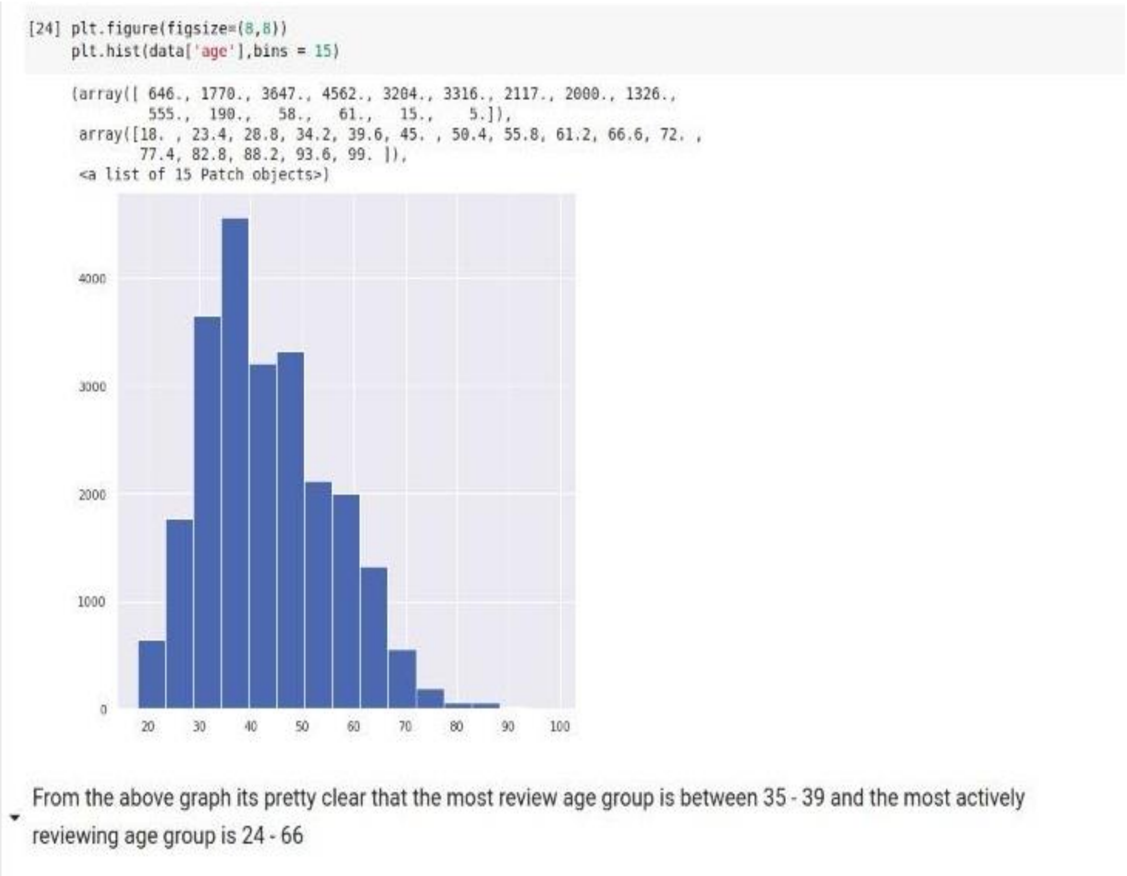
RESULTS

Fig 1. Bias Analysis of Our Dataset:



The dataset is not biased towards any division, department or class.

Fig 2. Comparison of age with number of reviews:



```
[24] plt.figure(figsize=(8,8))
     plt.hist(data['age'],bins = 15)
```

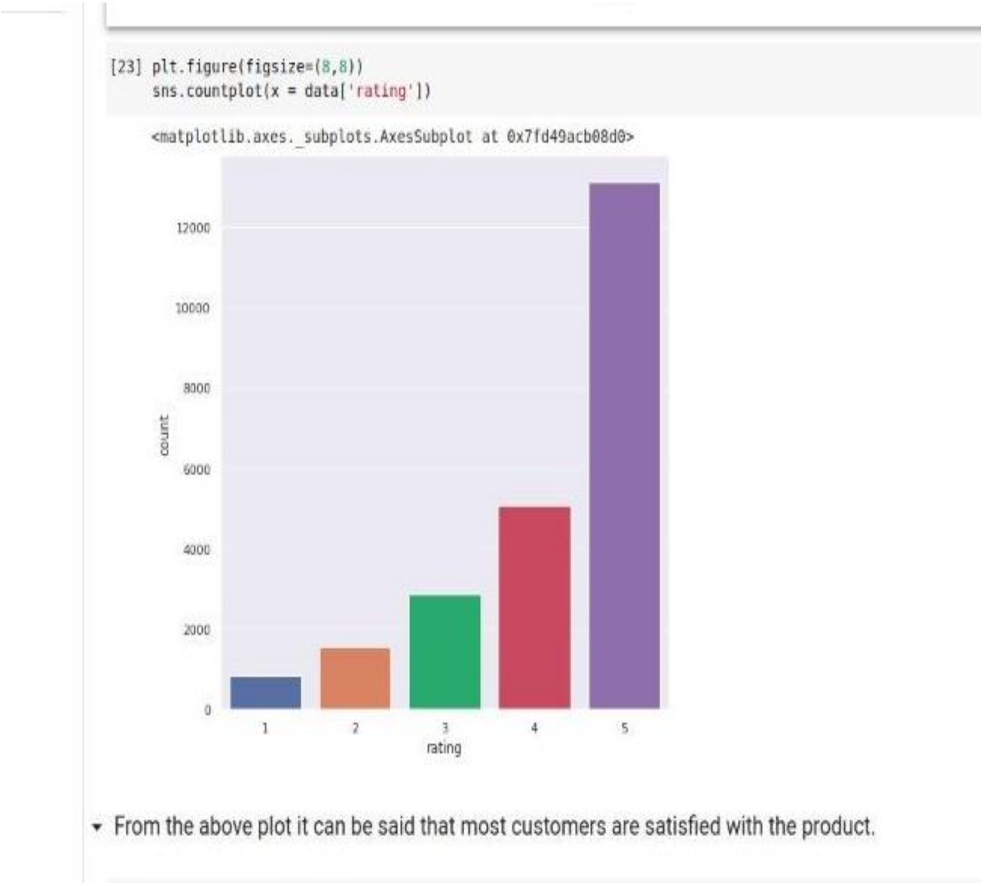
```
(array([ 646., 1770., 3647., 4562., 3204., 3316., 2117., 2000., 1326.,
          555.,  190.,   58.,   61.,   15.,    5.]),
 array([18. , 23.4, 28.8, 34.2, 39.6, 45. , 50.4, 55.8, 61.2, 66.6, 72. ,
        77.4, 82.8, 88.2, 93.6, 99. ]),
 <a list of 15 Patch objects>)
```

From the above graph its pretty clear that the most review age group is between 35 - 39 and the most actively reviewing age group is 24 - 66

(i) Youngsters aged 24-66 are more interested in providing the review.

(ii) After 50, the review has constantly declined with the age

Fig 3. Plot between Rating and Count:



```
[23] plt.figure(figsize=(8,8))
     sns.countplot(x = data['rating'])
```

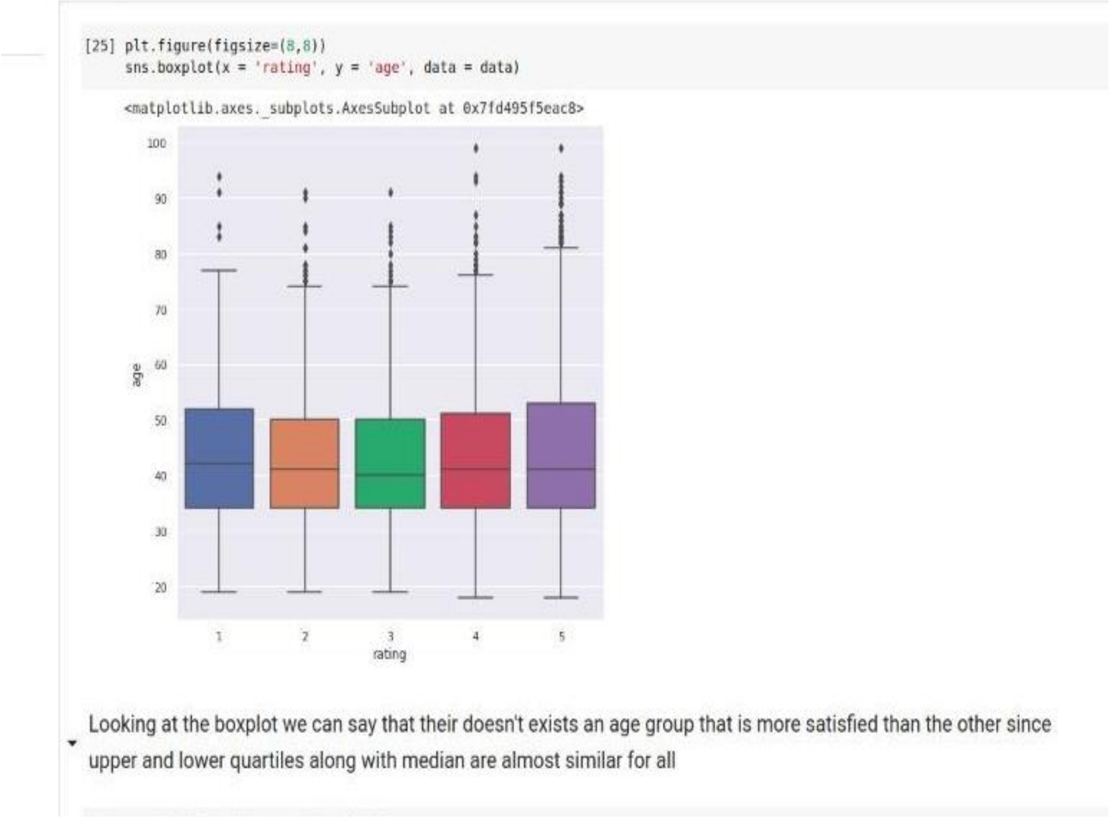
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd49acb08d0>
```

▾ From the above plot it can be said that most customers are satisfied with the product.

(i) From the matplot above, the number of people fully satisfied (5 star), is nearly the combined sum of people from 1 star to 4 star.

(ii) Most of the customers are satisfied.

Fig 4. Plot between Rating and Age:



```
[25] plt.figure(figsize=(8,8))
     sns.boxplot(x = 'rating', y = 'age', data = data)
```

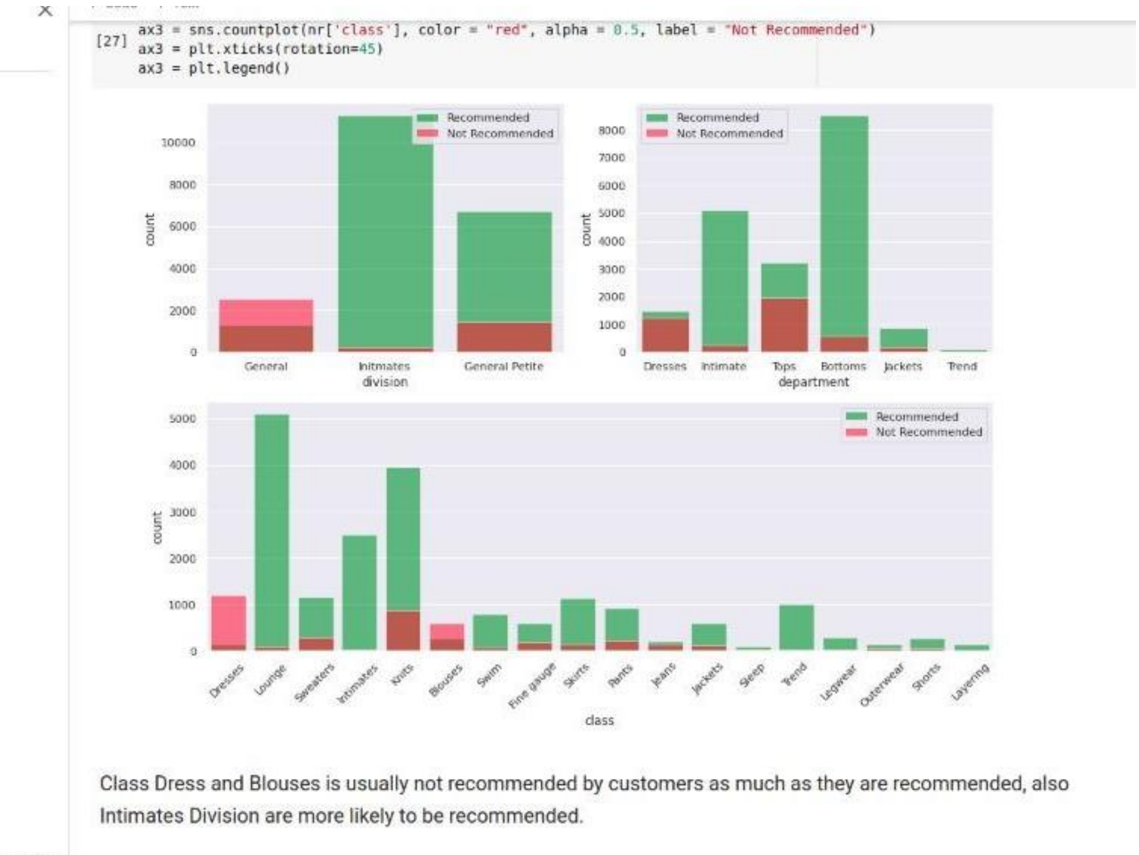
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd495f5eac8>
```

Looking at the boxplot we can say that their doesn't exists an age group that is more satisfied than the other since upper and lower quartiles along with median are almost similar for all

(i) There isn’t a significant difference in the box-plots across various age-groups.

(ii) Basically, all the age groups are satisfied to the same extent.

Fig 5. Comparison of product Class, Division and Department with Count :



```
[27] ax3 = sns.countplot(nr['class'], color = "red", alpha = 0.5, label = "Not Recommended")
     ax3 = plt.xticks(rotation=45)
     ax3 = plt.legend()
```

Class Dress and Blouses is usually not recommended by customers as much as they are recommended, also Intimates Division are more likely to be recommended.

(i) In the first one, intimate division has a very high probability of getting recommended. General ones are least recommended.

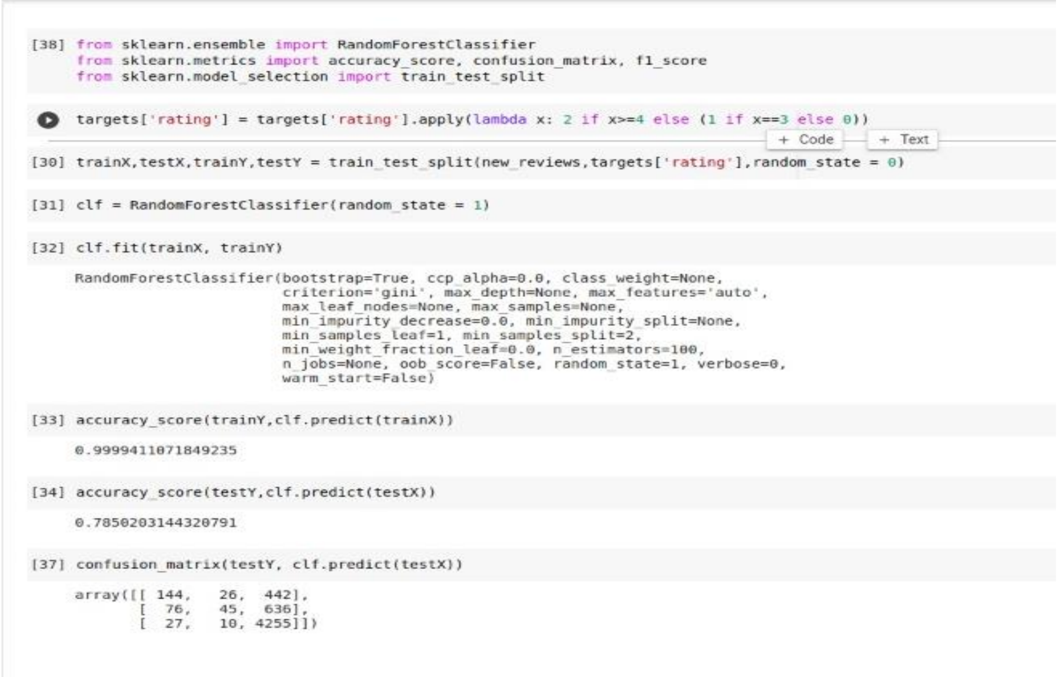
(ii) Similarly in the second one, Bottoms are the first recommendation of people followed by Tops. Trend ones are highly unlikely.

(iii) In the third one, Lounge and knits are top choices and highly likely to be recommended. Jeans, Legwear, Outerwear, Shorts and Layering are among the least recommended.

CONCLUSION

The **random forest** is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated **forest** of trees whose prediction by committee is more accurate than that of any individual tree. We received an accuracy of 99% on the Train data and 78% on the Test data respectively. This model was fitting appropriate and results and accuracy are shown :

Code Snippet with Confusion Matrix:



```
[38] from sklearn.ensemble import RandomForestClassifier
     from sklearn.metrics import accuracy_score, confusion_matrix, f1_score
     from sklearn.model_selection import train_test_split

     targets['rating'] = targets['rating'].apply(lambda x: 2 if x>=4 else (1 if x==3 else 0))

[30] trainX,testX,trainY,testY = train_test_split(new_reviews,targets['rating'],random_state = 0)

[31] clf = RandomForestClassifier(random_state = 1)

[32] clf.fit(trainX, trainY)

     RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                            criterion='gini', max_depth=None, max_features='auto',
                            max_leaf_nodes=None, max_samples=None,
                            min_impurity_decrease=0.0, min_impurity_split=None,
                            min_samples_leaf=1, min_samples_split=2,
                            min_weight_fraction_leaf=0.0, n_estimators=100,
                            n_jobs=None, oob_score=False, random_state=1, verbose=0,
                            warm_start=False)

[33] accuracy_score(trainY,clf.predict(trainX))

     0.9999411071849235

[34] accuracy_score(testY,clf.predict(testX))

     0.7850203144320791

[37] confusion_matrix(testY, clf.predict(testX))

     array([[ 144,   26,  442],
            [  76,   45,  636],
            [  27,   10, 4255]])
```

REFERENCES :

I. “Statistical Analysis on E-Commerce Reviews, with Sentiment Classification using Bidirectional Recurrent Neural Network” by Abien Fred M. Agarap

II. “Clothes reviews analysis with NLP — Part 2 Predicting items’ rating from text reviews analysis” by Valentina Alto

III. Alrehili, A. and Albalawi, K. (2019). Sentiment analysis of customer reviews using ensemble method, 2019 International Conference on Computer and Information Sciences (ICCIS) .

IV. Sentiment Analysis using machine learning algorithms: online women clothing reviews. Shuangyin Xie

Thank You!