INFO 5810

DATA ANALYSIS & KNOWLEDGE DISCOVERY

TERM PROJECT-TEXT MINING

WOMEN’S E\_COMMERCE CLOTHING REVIEWS DATA ANALYSIS

AMRUTH JALIGAMA

STUDENT ID: 11396245

**Introduction**

The main reason why I chose to work on this topic is because there is a rapid increase in number of e-commerce websites emerging day by day especially when it comes to the clothing and apparel products and many people feel comfortable buying apparel from e-commerce websites because of the flexibility it provides. Most of the time we purchase products based on our liking, the rating and reviews we find online. I do a lot of online shopping when it comes to apparel but I do read the reviews and see the rating before any purchase. Through the analysis I perform, I analyze how the reviews are based on the rating and the relation between different numerical measures.

**About The Dataset**

The dataset I want to work on is Women’s e-commerce clothing reviews dataset. The cleaned dataset I use for my analysis consists of 10 columns and 10,000 rows. The columns are as follows ClothingID, Age, Title, Review Text, Rating, Rating, Recommended IND, Positive Feedback Count, Division Name, Department Name and Class Name. The dataset has been downloaded from kaggle website.

Dataset Link: [Women's E-Commerce Clothing Reviews | Kaggle](https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews)

**Data Cleaning & Data Filtering Steps**

1. The original dataset contains 23,468 records with 11 columns.
2. I have removed the first column from the original dataset because it simply represents the row number of each record.
3. After the first step performed above I remove all the rows which have Title column values as blanks, Review Text column values as blank, Division Name values as blanks.
4. After the above two steps we are left with 19,663 rows where one row is the name of the columns.
5. So we have 19,662 records in our dataset and to make it suit our requirement for the project we delete top 4,831 rows and bottom 4,831 rows which makes it total of 10,000 records in our dataset.
6. After we import the cleaned dataset into Rapidminer Studio because of some syntax errors in the data the number of records in the dataset is reduced to 9,497 rows and there are some null values which form in some rows. The rows which have missing values are filtered out using the filter example operator in Rapidminer Studio.

**Analysis Goals**

1. As the Positive Feedback Count increases does the rating increase and is there any relation between Rating and Recommended IND?
2. How the reviews are for dresses whose rating is above average?
3. Are there any reasons as to why some dresses have rating below average?

**Data Mining Techniques Used**

1. Correlation Analysis- Through this technique we see the relationship between two variables i.e., if two variables when comparing to each other are directly proportional to each other or inversely proportional to each other. The scale of correlation analysis ranges from -1 to +1 where ‘-‘ symbol represents negative relationship i.e., there is an inverse relationship between two variables and no symbol means there is a positive relationship between two variables i.e., there is a direct relationship between two variables.
2. Association Analysis- Association Analysis is a data mining technique through which we can try to find the pattern between items in a group. Simply put the real time applications of this is when we purchase an item online through amazon website we get other recommended products this is done through association Amazon analyses the data of people who have purchased first item and purchased the second item and the data of how frequently was the second item purchased when the first item was purchased is analyzed. This is nothing but Association Analysis. In this case of Amazon the term used to refer these associations is called as market basket analysis.
3. Cluster Analysis- It is a data mining technique through which we group objects based on their similarities. Similar items are grouped into same group or clusters. There are different algorithms through which clustering can be done some of them are K-means Clustering, Hierarchical Clustering, Density Based Spatial Clustering, Mean shift Clustering and so on.

**Final Result Diagram**

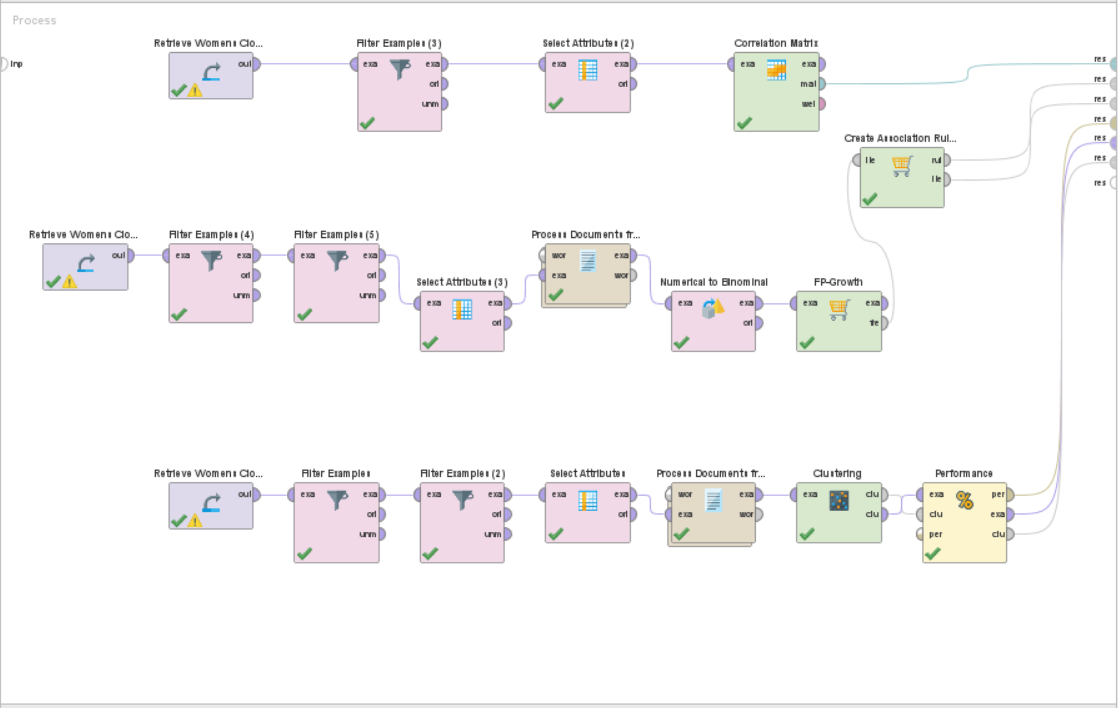


Figure 1 Final Result Diagram of Correlation Analysis, Association Analysis and Cluster Analysis

**Conclusion**

From the analysis I performed on the data I conclude that there is no relation between positive feedback count and rating as the correlation coefficient value in the correlation matrix is -0.070 it implies there is no correlation. The measures Rating and Recommended IND are moderately positively correlated as the correlation coefficient value in the correlation matrix is 0.793 so we can conclude that as the Rating increases the product recommendation increases. From the Association analysis results with a support of 5% and minimum confidence value of 40 % for the dresses whose rating is above average we can infer that the customers love the fabric and color of the product, feel that the looks are great, length is perfect, the fit is as per the ordered size. From the Cluster Analysis results using K-means clustering with number of clusters being 7 i.e., K value being 7 we can infer that the reviews for dresses whose rating is below average is as follows the customers are disappointed with the dress, fabric, they want to return the product, it is wrinkled, it looks cheap, felt baggy, the dresses are huge.