**Data Analysis and Knowledge Discovery**

**Term Project Evaluation**

**TITLE: ANALYSIS TOY PRODUCTS OF AMAZON DATASET**

**Section 1: `**

I chose the Toy product from Amazon Dataset. This dataset is not able to do very large RapidMiner which is why I filter the data and I delete some columns and rows. These datasets have more stop words and I delete all of them and there are no numbers and integers, all text type and its xlsx format. I fixed this dataset no less than 6000 and no more than 10000. They have 8626 example, 0 special attributes and 12 regular attributes.

I select toy products of amazon dataset with all text type fields and I take this dataset from the Kaggle website . Most Amazon product reviews allow users to rate items based on their overall quality. The Toys and Games section allows users to provide more feedback, including product ratings. In terms of fun, durability and educational quality. All information is available in this dataset. Toy products on Amazon include a variety of electronic toys, a variety of dolls, a variety of colours and plaids, Lego, toy trains, cars and more. When people shop online, they usually check product ratings and reviews before deciding on a product because there are many brands that sell the same product. For example, when someone goes out to buy a frozen toy, they look for reviews of that product, how many have recommended it, and how many have found the review or recommendation useful. After analyse these points one can decide which one they can buy.

**Section 2 & 3:**

**Data cleaning:**

The data cleaning in this data set is performed as below:

* As there is no data available and are unnecessary, I have deleted some attributes like price, Number\_available\_in\_stock, Number\_of\_reviews, number\_of\_answer\_question, average\_reviews\_rating, etc.
* I have replaced the missing values with ‘0’ for attributes with type integer. I deleted all integer and I deleted some row these row are price, Number\_available\_in\_stock, Number\_of\_reviews, number\_of\_answer\_question, average\_reviews\_rating.

When I clean data, I removed a few columns that were not important for data analysis. The columns removed were:

* Uniq\_id
* Product\_name
* Manufacture
* Amazon\_category\_and\_sub\_category
* Customers\_who\_bought\_this\_item\_also\_bought
* Description
* Product\_information
* Product\_description
* Items\_customers\_buy\_after\_viewing\_this\_item
* Customer\_question\_and\_answers
* Customer\_reviews
* Sellers

I select data and when I input RapidMiner then I see my data is too big that is the reason I filter and delete some rows and columns then I start work. My data have more stop Words like(, ? ) ": space) and more can be filtered out and eliminated from the document because they have no actual meaning so I delate all of this.

**Goals:**

My goal is to analyse the whole dataset and provide information on the following:

* Toy product of customer review and customer made command the product good, great, love and excellent. Then I can easily choose which one I can buy with my kids and my relative's gifts.
* what customers liked and disliked about the product they purchased.
* Find the connectivity between terms they used in the review and the quality of the product they purchased.

**Data Dictionary:**

|  |  |  |
| --- | --- | --- |
| Field name | definition | type |
| Uniq\_id | Unique ID | int |
| Product\_name | Name of the product | text |
| Manufacturer | Product manufacturer name | text |
| Amazon\_category\_and\_sub\_category | Product categories | text |
| Customers\_who\_bought\_this\_item\_also\_bought | References to other similar item | text |
| Description | Helpful message | text |
| Product\_information | Product information | text |
| Product\_description | Product description | text |
| Items\_customers\_buy\_after\_viewing\_this\_item | Helpful information | text |
| Customer\_question\_and\_answers | Customer conversation | text |
| Customer\_reviews | review | text |
| sellers | Sellers information | text |

**Data Mining Techniques:**

**Select attributes**: I select a toy product from the Amazon dataset and remove a subset of the example set attribute and other features. Then I will drag and drop the selected feature from the operator panel in the process. First, I select the attribute operator to select a subset of properties.

**Process Documents from Data:** This operator processes data. By this, we can find the frequency pattern in a table. We can break the document using other filter operators / subprocesses such as tokenizing the sentence in the document into words and filtering with stop words to delete the most used words.

**Remove duplicates:** Duplicate records in a data set are difficult to find manually and deleting records is another huge task. The duplicate removal operator automatically detects and repeatedly removes records from the data.

**Replace missing values:** Missing value Replaces empty files in a record with 0 or unknown.

**Filter example:** Records are filtered based on need. Missing values, and values without any parameters can be easily filtered.

**Tokenize:** Tokens are nothing more than the lowest form of sentences, also called words. This operator converts the whole document into a series of words.

**Transform cases:** The next step in processing documents from data can be case-sensitive. Transform cases are the best operators that can help us not to double count a word because of its case (above or below). Using the transform case, we can convert either lowercase words upwards or vice versa.

**K-Means clustering (Mention the k value you used):** k- means the algorithm assigns data points (to reduce the square distance between objects in the cluster, data points, and the centre (average) of each cluster. Cluster centres start by randomly selecting data. Once the data set is found, the average position of the data points within each cluster is calculated and the cluster core is then moved to the average position ৷ we continue this cycle before all the cluster centres are in one place. Clustering is the grouping of similar items or most repeated items and the algorithm used here is called K-mean. I have used K= 4, 5.

**Cluster Distance Performance:** This operator does clustering-based performance evaluation of centroids. Two performance measures supported are Average within-cluster distance and Davies-Bouldin. Cluster distance is used to test centroid-based clustering methods with performance accuracy. The architecture of the centroid cluster contains details related to clustering. It tells us which parts of the cluster are sampled. It contains information about all the actual cluster centroids.

**FP-Growth**

I have to submit RapidMiner files that the reason I did FP-Growth, Association, Clustering, and performance altogether with Multiply rules. I attach a document of my RapidMiner result.

**Correlation analysis:**

There is positive correlation between customers reviews. I did customer Recommend and customer reviews. And negative correlation between customer reviews. Coefficient ranges from -1 to 1 . -1 being highly negative and +1 being highly positive and it explains how strong the relationship is between the attributes. There is no matrix because I delate all number and integer value in my data.

**Association Analysis:**

Association rules help us understand the relationship between words or tokens in a statement. On the right, we have all of these conclusions and they have all rules matching look like stars, good, greats, price, and more. Which information do we need we can see all of this product information. In my data, good, great, product, stars, play and quality have a high confidence level. Hence, the consumer has shown positive feedbacks in the customer's reviews. According to the association analysis, there are an intelligent number of helpful customer reviews. I took a review like this and made a tree graph around it. Most reviews say that Amazon products are easy to use, low cost, great gifts for our kids, and special day gifts offer products at much cheaper prices on Black Friday, can easily turn on parental control, and so on.

**K-Means Cluster Analysis:**

For K=4 and min confidence value=0.3

And Maximum run=10

From the K-means clustering analysis, after using a few different K values I have set the K-value to 4 with a maximum of 10 runs. Irrespective of the wording or the order of the words the terms used in most of the reviews were positive(Stars, good, great, money, happy, bought, quality, love, etc.)

**A picture containing graphical user interface

Description automatically generated**

Figure 1.1 Final result

**Graphical user interface

Description automatically generated**

Figure 1.2 Measures table

**Graphical user interface, application

Description automatically generated**

Figure 1.3 Association analysis- FP-Growth

**Graphical user interface, text, application

Description automatically generated**

Figure 1.4 Association rules

**Graphical user interface

Description automatically generated**

Figure 1.5 Association analysis Graph

**Graphical user interface, application, Word

Description automatically generated**

Figure 1.6 Cluster Model

Graphical user interface, application

Description automatically generated

Figure 1.7 Cluster

Graphical user interface

Description automatically generated

Figure 1.8 Result History

**Conclusion:**

* Association rules help us understand the relationship between words or tokens in a statement. On the right, we have all of these conclusions and they have all rules matching look like stars, good, greats, price, and more. Which information do we need we can see all of this product information. In my data, good, great, product, stars, play and quality have a high confidence level.
* In K-Meaning Clustering Analysis, by arranging the clusters one can notice words like love, purchase, gift, tablet, price, kindle read, simple, good and great. Someone noticed that the reviews were mostly positive and the customers responded positively
* Finally, I realized that a better, better and better customer item received the most reviews and was positive.
* Overall, customers seem to be satisfied with the products they buy and these reviews can be helpful for future customers.
* The conclusion here is so obvious. Reviews have greatly helped consumers make an easy decision on whether or not to buy a product Finally, I would say that user reviews, which have been verified and provided by customers, have helped buyers make a good purchase decision by providing a thorough explanation of the products, which has been particularly effective in the Toy area.