CSCI946 – BIG DATA ANALYTICS

ASSIGNMENT 1

TEAM MEMBERS

1. **AKSHIT SOOD (7409254)**
2. **VIGNESH MURUGESAN (7087597)**
3. **SAI PRIYANKA VOORADA (7040222)**
4. **SRINIVAS REDDY OBILI (6917215)**

NAME	CONTRIBUTION PERCENTAGE
AKSHIT SOOD	100%
VIGNESH MURUGESAN	100%
SAI PRIYANKA VOORADA	100%
SRINIVAS REDDY OBILI	100%

Task 1:

Problem Analysis

Goal of the assignment is out of the 9 categories, we have to choose a certain set of categories. Now from the chosen categories, we have to find the best category. We also have to find the top 10 products from the chosen categories.

First, we have explored the dataset.

Name of column	**Meaning**
category	category to which item belongs
subcategory	subcategory to which item belongs
name	name of product
current_price	price of product currently
raw_price	original price of product
currency	price of product after discount
discount	discount applied
likes_count	number of likes for product
is_new	flag to show product age
brand	brand for product
brand_url	brand url
codCountry	country where available
variation_0_color	color availability
variation_1_color	color availability
variation_0_thumbnail	thumbnail for variation 0
variation_0_image	variation 0 image
variation_1_thumbnail	thumbnail for variation 1
variation_1_image	variation 1 image
image_url	url of image
url	url
id	unique identifier for product
model	model number

Out of all the columns, we would not be using all of them. As some of these columns have values which are not useful for our use case. We need to focus on columns which are integer/decimal format. We will NOT consider id , which although is integer, as it doesn't help in finding the answers to the problem statement. There are some other columns which are urls and web links, which again, won't be useful, so we have discarded the same. Also, due to their large size, it will be computationally expensive to process them.

We have chosen to explore the jewelry, women and shoes dataset, as it is related to women buyers, and it makes more sense to combine categories which show similarities in buyer behavior.

Now to choose the top 10 products and the best category, we can take 'likes_count' into consideration. It is a feature which gives a numeric value of the number of likes a product has received. In other words, it gives a quantitative measure of the popularity of the product within the customers.

Data Preparation

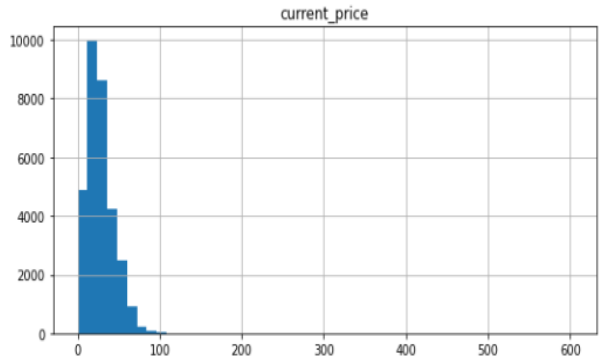
Now, we have to do 'feature selection'- Selecting a subset of all available columns which would help us to design our Classification and clustering algorithms.

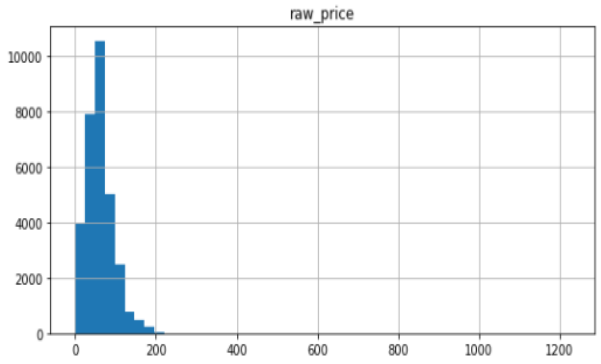
We have chosen the following columns- likes_count, current_price, raw_price and discount.

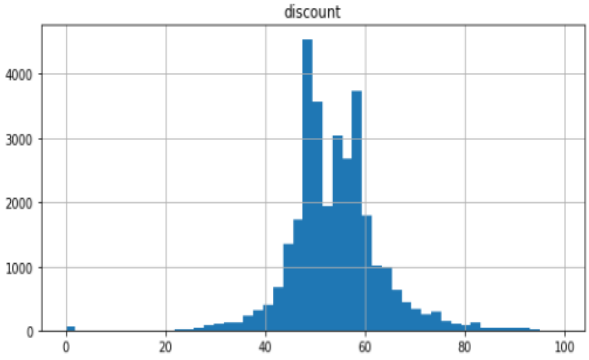
Before feeding the data into machine learning models, we have to make sure that the data does NOT have null values. If Null values are present, we can handle them by replacing them with mean, median or some other strategy.

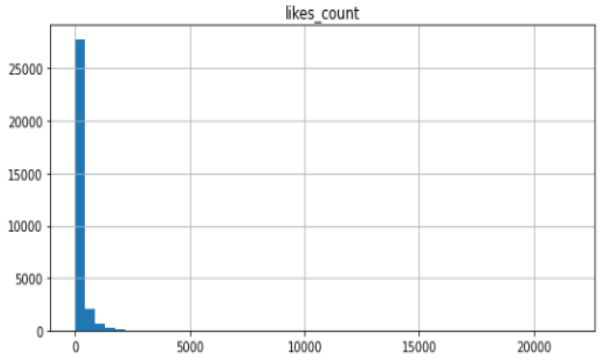
In our case, raw_price has some missing values, which does not make sense. We have replaced these values with the mean of the column.

Below graph shows some of the important features and their distribution. As we can see 0 values in the raw_price column.









We have also done normalization for the variables, so that variables now have similar scales. For example, likes_count and raw_price have different scales in the original dataset.

Most of the machine learning algorithms don't work well with categorical dataset. Before feeding the data to a machine learning algorithm, we have to convert these categorical data columns to numerical values.

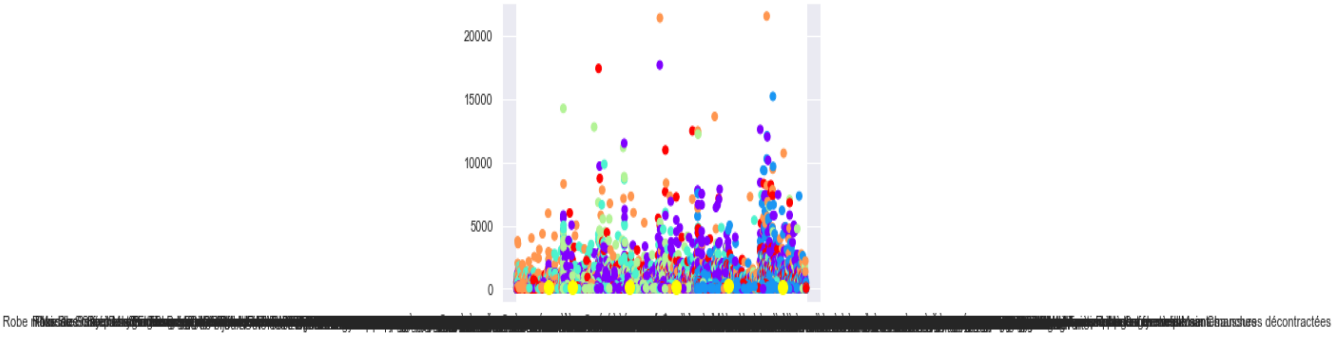
In our assignment, for classification problem, we have converted the target variable- **category**, from categorical to numerical.

We also have calculated Pearson's coefficient of correlation to find relationships between variables. We found out that current_price and raw_price are highly correlated. To ensure independence of the features selected, we can go ahead and take either of them into our algorithms.

After the data preparation has been performed, we have run clustering and classification algorithms on this data. Those have been detailed in the next sections.

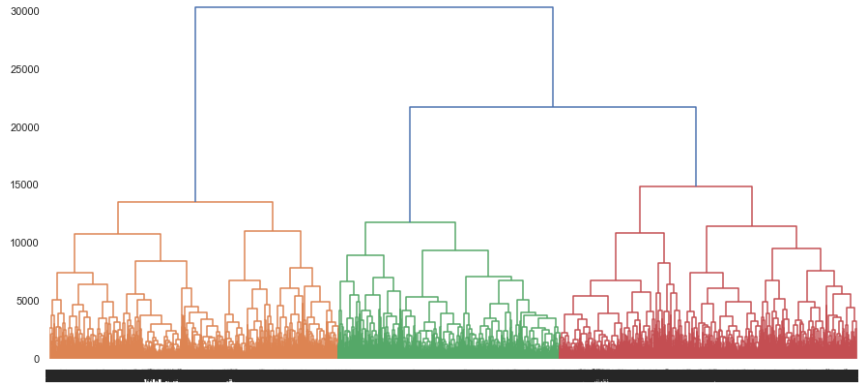
Task 2:

To form clusters in the given dataset, two columns are considered i.e, ‘likes_count’ and ‘name’. The two clustering algorithms performed in this task are K-means clustering and Hierarchical clustering. K-means clustering forms well separated clusters for most datasets and Hierarchical clustering is used to find out the optimum number of clusters that can be formed in the dataset. A total of 6 clusters are formed in both K-means clustering and Hierarchical clustering with ‘likes_count’ on Y-axis and ‘name’ on X-axis. The results are as follows.



K-means clusters

The yellow dots in the plot represent the cluster centroids.



Hierarchical clustering

The above picture is the result of hierarchical clustering in identifying the optimum number of clusters with ‘likes_count’ on Y-axis and ‘name’ on X-axis.

Task 3:

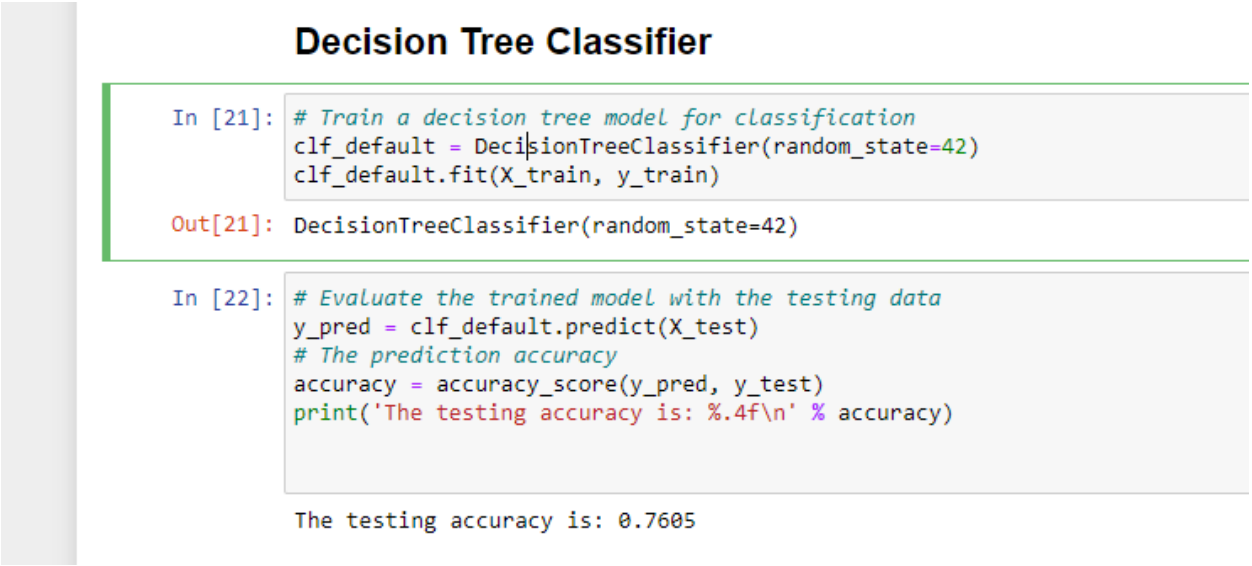
For classification, we have chosen likes_count, raw_price, current_price and discount as our independent variables. We have made sure the null values are not present in these columns. For the target/dependent variable, we have chosen 'category'. We have used the following algorithms in our use case - **Decision Tree and KNN classifier models.**

Decision Tree Classifier

In the Decision Tree Algorithm, each node will be testing an input variable. Branch of the tree will correspond to the decision made.

We have used the class **DecisionTreeClassifier** from ScikitLearn to implement the decision tree classification model in this assignment.

We have used **LabelEncoder** to transform the target variable-'Category' from categorical to numerical. Finally, we have split the data in the ratio 80:20 of training to test data, and fed it into the model. Below we can see the accuracy of the model after it has been trained.



KNN classifier

KNN Classifier model is also popularly known as K-nearest Neighbor model. It is a supervised algorithm. By calculating the distance between the test data and all of the training points, KNN tries to predict the allotted class for the test data. Then K spots are chosen that are closest to the test data.

The KNN algorithm calculates the probability that the test data fall into each of the "K" training data classes, and the class with the highest value is chosen.

Similar to the Decision Tree classifier above, we have used LabelEncoder to transform the target variable-'Category' from categorical to numerical. Finally, we have split the data in the ratio 80:20 of training to test data, and fed it into the model. Below we can see the accuracy of the model after it has been trained.

KNN Classifier model

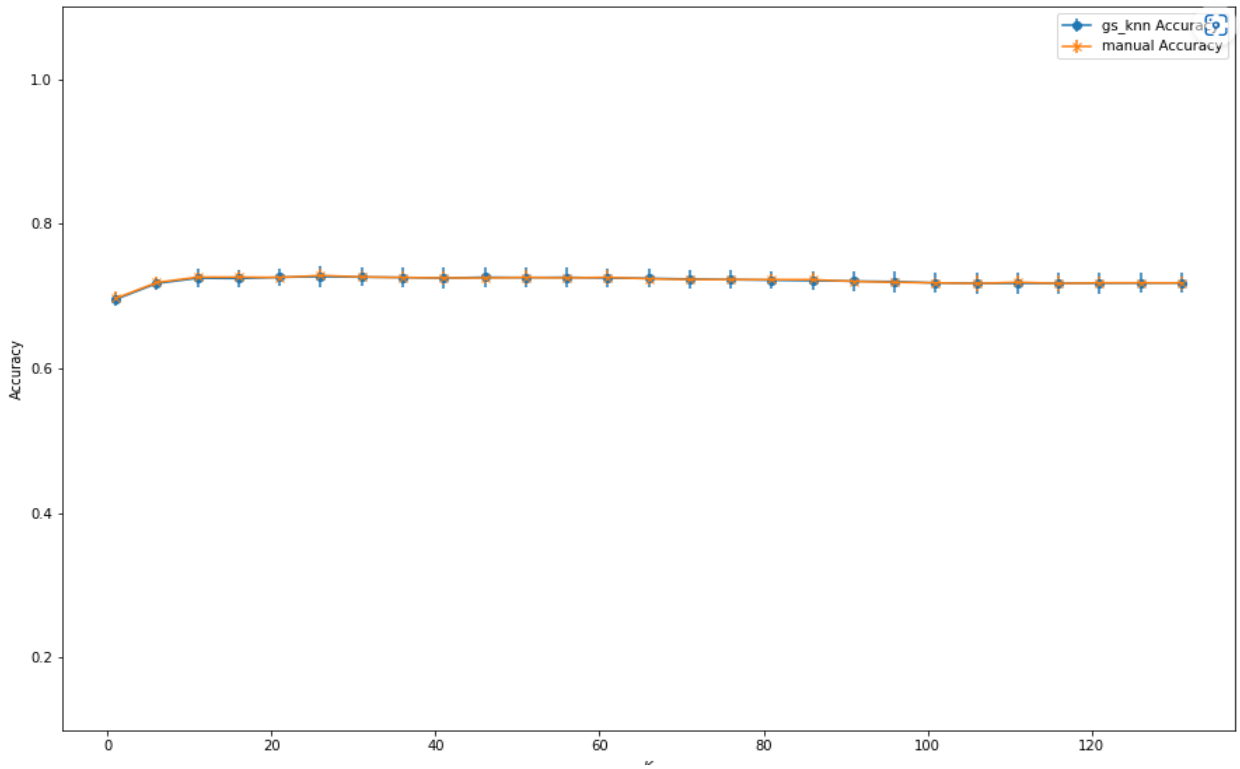
```
In [23]: # Build a KNN classifier model
         clf_knn = KNeighborsClassifier(n_neighbors=1)
         # Train the model with the training data
         clf_knn.fit(X_train, y_train)

Out[23]: KNeighborsClassifier(n_neighbors=1)

In [24]: y_pred = clf_knn.predict(X_test)
         accuracy = accuracy_score(y_test, y_pred)
         print("Accuracy is: %.4f\n" % accuracy)

         Accuracy is: 0.7141
```

For the KNN model, for optimization, we have made use of **GridSearchCV** to find the optimum value of k, to get the maximum accuracy of the model.

Below visualization shows how accuracy(y axis) changes for the classifier based on change in value of k (x axis).

We can see that for value of k=26, the accuracy of the model is 72.69%, which is the highest.

```
Best K value:  26
The accuracy: 0.7269
```



Comparison of the classifiers

Classifier	Accuracy	Comments
Decision Tree Classifier	76.05%	
KNN classifier	72.69%	Obtained with use of GridSearchCV- with k value=26

From the above table, we can see that Decision tree Classifier(76.05%) gave us better accuracy than KNN classifier(72.69%)

Task 4:

Are the clusters well separated?
The clusters formed in both K-means clustering and Hierarchical clustering are not well separated, their centroids are close to each other.

Do any of the clusters have only a few points?
No clusters have only a few points, since the dataset is large and not very much correlated to other columns, no clusters have only a few data points.

Are there meaningful/non-meaningful clusters to the analytics problems questioned in task 1?
The clusters formed in K-means clustering and Hierarchical clustering are not very much correlated to each other as the data given is not very helpful to derive meaningful analytics.

What are the advantages, shortages for clustering and in this analytics case? Which one provides results of greater value.

The advantages of clustering and classification are, clustering provides insights on hidden patterns in the dataset, whereas classification splits the data into different classes.

The disadvantages of clustering and classification in the analytics are clustering cannot split the data into meaningful clusters and classification will only help to identify the class of unknown datapoint.

To analyze the data with irrelevant columns, it is difficult to draw patterns or classes with either clustering or classification.

Are the examined algorithms suitable for Big data analytics?
The algorithms used in this assignment are suitable for big data analytics as they provide clusters and classes which can help analyze the business problems and target a particular cluster or class of customers for better business. But to draw meaningful patterns or classes, the dataset should be clear enough and each column be minimally correlated to each other.

Will data preprocess affect clustering and classification results?
Yes, the data preprocess certainly affects clustering and classification results. With data preprocessing, we remove the unnecessary columns from the dataset, use only those columns

with high correlation to the dependent variable, which improves efficiency and performance of the algorithm.

Did the classifiers separate products well from other classes?

As each product belongs to only one category after the classifier has been applied, with accuracy of almost 76%, products have been correctly predicted.

Do any classes have very few points?

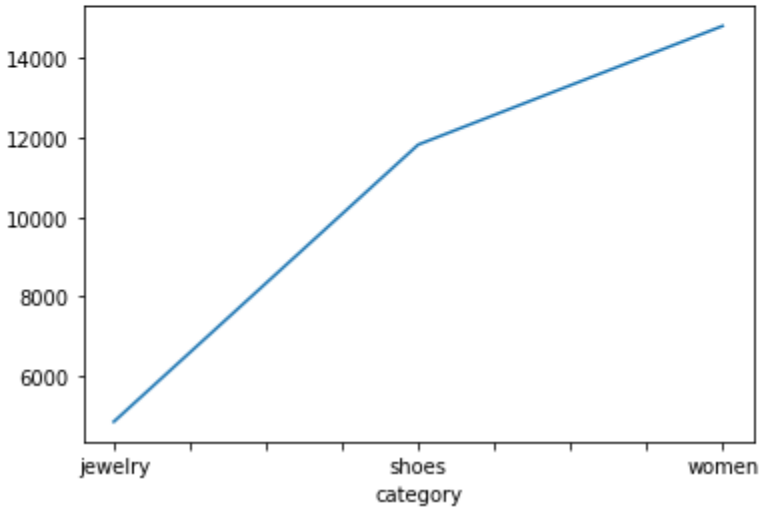
Below, we can see number of products in each class category -

```
: df.groupby('category').count()['subcategory']
```

```
: category
  jewelry     4853
  shoes      11823
  women      14809
  Name: subcategory, dtype: int64
```

```
: df.groupby('category').count()['subcategory'].plot()
```

```
: <AxesSubplot:xlabel='category'>
```



As we can see, each class has a different number of instances. ‘Jewelry’ has the least number of instances in comparison to ‘shoes’ and ‘women’, with the ‘women’ category having the largest number of instances. However, there are no classes which have very few instances.

Are there any meaningful/non meaningful classes for task 1 discussed?

For our problem statement, we have the following 3 classes- Women, Jewelry and Shoes. Although distribution is not equal in all three classes, all the 3 classes are meaningful as they would help us to better understand how customer behavior varies in all three classes.

What are the Advantages/Drawbacks of classification in our analytics use case in this assignment?

Main objective of the classifier is to label new input into the correct category.

We can label products into the correct category, if the classifier has been designed well.

One drawback in this analytics dataset is if we had more relevant features to correctly predict a product into the correct category, the model would have been more useful. Using features like price, discount have been used to label a product as more relevant features were not observed.

Having more relevant features would have increased the accuracy and performance of the model even more.

Will data pre-processing affect classification results?

Data pre-processing can affect the results greatly. It was observed that scaling(normalization) the attributes improved the accuracy of the classifiers chosen.

Feature engineering, another important factor, greatly influences the result, as it is important to ensure the features which will impact the dependent(target) variables are chosen well.

The top 10 best products based on the likes_count in the given dataset are in the following table.

category	**subcategory**	**name**	**current_price**	**raw_price**	**discount**	**likes_count**	
26971	shoes	Derbies & Mocassins	Chaussures Plats Décontractées En Suède Mocass...	14.99	54.95	73	21547
15145	women	Blouses & Chemises	Blouse Large Couleur Pure pour Femme	19.99	56.99	65	21403

15120	women	Robes vintage	Robe Longue avec Boutons Chinois	27.99	59.99	53	17684
8602	women	Robes vintage	Gracila Femme Maxi Robe Irrégulier Vêtement Vi...	29.99	60.99	51	17414
27673	shoes	Derbies & Mocassins	Chaussures De Grande Taille Semelle Souple À E...	30.08	54.95	45	15203
4863	women	Soutiens-gorge	Soutien-gorge Sexy à Décollecté Plongeant sans...	13.99	26.89	48	14252
21483	shoes	Bottes & Bottines	Bottines Plates Doublées de Fourrure	9.99	42.99	77	13615
8105	women	Brassières de sport	Soutien-gorge Sexy Antichoc Sans Armature Ling...	12.99	23.89	46	12786
26276	shoes	Sandales & Mules	SOCOFY Sandales Confortables Plates Avec Bride...	21.07	49.99	58	12591
19044	women	Vestes & Gilets	Manteau imprimé floral à feuilles à capuche	38.84	79.99	51	12482

The top product in the 3 categories is the **women** category as it is present most number of times in the top 10 products category.