Customer and Item Segmentation

(Customer Purchasing Behaviour Analysis using different Data Mining Techniques)

Abstract— **The project's goal is to investigate the customer profiles in a supermarket, explore the dataset, and perform predictive analysis using various data mining techniques such as clustering and classification. By constructing various models, customers are classified as high, low, or medium spenders. The models are fine-tuned and tested. Various visualizations are used to display the results. Market Basket Analysis is performed by employing sequential pattern mining to discover the various association rules.**

Keywords—Data Exploration, clustering, classification, market basket analysis, predictive analysis, association rule mining, data mining

I. INTRODUCTION

This project aims to the following data mining operations on a customer dataset. The clustering and classification techniques are used for predictive analysis. Using the **clustering** ttechnique, the customer data set will be divided into three clusters based on the items purchased. The following information are retrieved.

• A customer's total number of items purchased.

• The number of distinct items purchased by a customer during the observation period.

• The most items a customer can buy in a single shopping session.

• The Shannon entropy of the customer's purchasing behaviour. (Guobing Qian, 2021)

Using **classification**, we will predict each customer into one of three classes based on the amount spent on each purchase. High-spending client, Middle-spending client and Low-spending.

The items that are frequently bought by the customers and the frequent combinations of item sets for future recommendations are also found using **association rules.**

II. DATA MINING

A. *What is Data Mining*

"Data mining is the process of automatically discovering useful information in large data repositories". (Tan, 2019)

B. *Data Mining Techniques*

There are numerous data mining techniques, some of which are as follows:

a. **Association** is one of the most effective methods. A pattern is revealed here based on the relationship between an item and other items in a transaction. For example, in MBA, it is used to identify products that consumers frequently purchase together. (Dubey, 2021)

b. **Classification**: When new objects enter the market, their characteristics are examined and they are assigned to a predefined class. For example, credit applicants can be classified as low, medium, or high risk. (Dubey, 2021)

c. **Prediction**: This feature predicts missing or unknown attribute values. For example, forecasting the next week's sale using currently available data. (Dubey, 2021)

d. **Clustering** is the process of organizing data into sub-groups or clusters. (Dubey, 2021)

III. PART I – DATA EXPLORATION

The dataset consists of numerical and categorical data with 471910 entries in 8 columns. To analyse the customers on amount spend in transactions an amount column is to be added. The basket date is to be converted to date time format, since it is object type. ProdDescr and CustomerId columns have null values. The statistics of the data is

	count	mean	std	min	25%	50%	75%	max
Sale	471910.0	4.030945	83.769380	-11062.06	1.25	2.08	3.75	38970.0
Qta	471910.0	10.716533	231.355136	-80995.00	1.00	4.00	12.00	80995.0
Amount	471910.0	19.030258	400.925995	-168469.60	3.75	10.08	17.70	168469.6

There are 24627 distinct baskets in the dataset. Data cleaning was needed as there was negative values in qty, zero in sales and cost.

We extracted some features for each customer, creating a new dataset indexed by their CustomersIDs, to better describe the customer profile and also improve data quality.

I: the total number of items bought by a customer during the observation period.

Iu: the number of distinct items purchased by a customer during the observation period.

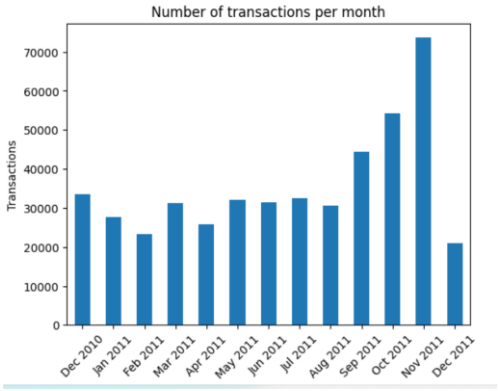
Imax: the maximum number of items purchased by a customer during a shopping session

E: the Shannon entropy on the customer's purchasing behavior

(et al., 2021)

After data cleaning some information regarding the sales were derived

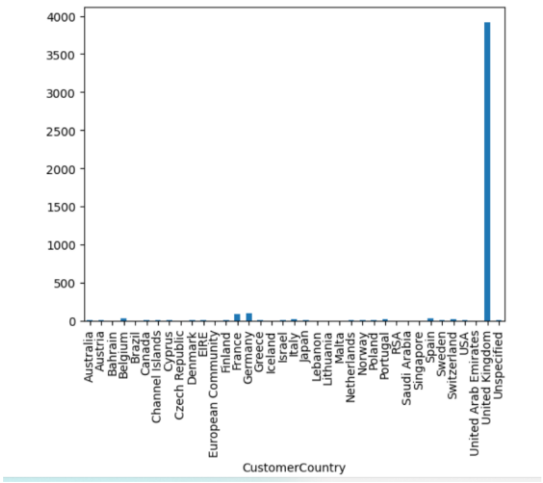
a. Transactions per month



b. Total number of transactions by each customer - 4338

count	4338.000000
mean	91.720839
std	228.785054
min	1.000000
25%	17.000000
50%	41.000000
75%	100.000000
max	7847.000000

c. Which country had the highest customers - UK



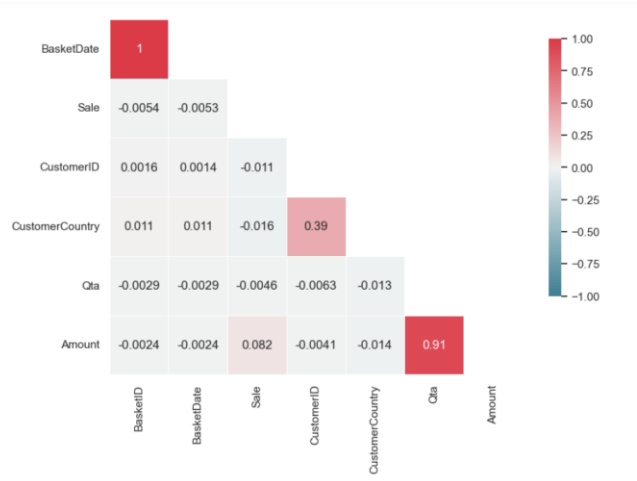
d. Best selling product

85123A (WHITE HANGING HEART T-LIGHT HOLDER)

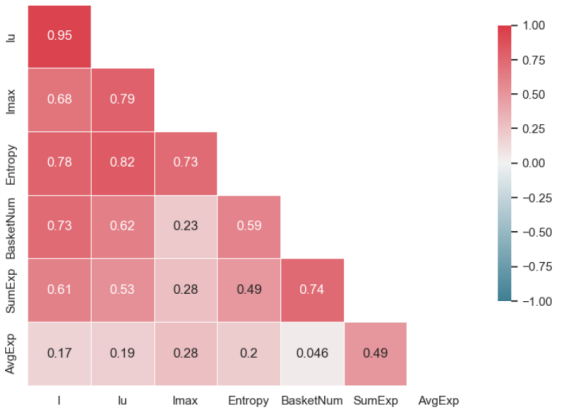
e. Worst selling product

23843 (PAPER CRAFT , LITTLE BIRDIE)

Correlation



Correlation of the new dataset



IV. PART II -CLUSTERING ANALYSIS

Normalizing data is a good practice in clustering to avoid biases. The StandardScaler (also known as Z-Score) and the Min-MaxScaler are two approaches. We tried both methods and chose the latter because the results were more good.

The different clustering algorithms were used to cluster the data.

1. K Means:

   K-Means is one of the most well-known and widely used clustering algorithms. Prototype-based clustering techniques partition data objects on a single level. K-means defines a prototype in terms of a centroid, which is typically the mean of a group of points, and is typically applied to objects in an n-dimensional continuous space. (Tan, 2019)

2. DBSCAN

   Density-based clustering finds high-density regions separated by low-density regions. DBSCAN is a simple and effective density-based clustering algorithm that demonstrates a number of key concepts for any density-based clustering approach. (Tan, 2019)

3. Hierarchical

Hierarchical clustering techniques like K-means, are relatively old in comparison to many clustering algorithms, but they are still widely used. There are two basic methods for producing a hierarchical clustering:

A. Agglomerative: Begin with the points as individual clusters and merge the closest pair of clusters at each step. This necessitates the development of a concept of cluster proximity.

B. Divisive: Begin with a single, all-inclusive cluster and split it at each step until only singleton clusters of individual points remain. In this case, we must decide which cluster to split and how to split it at each step. (Tan, 2019)

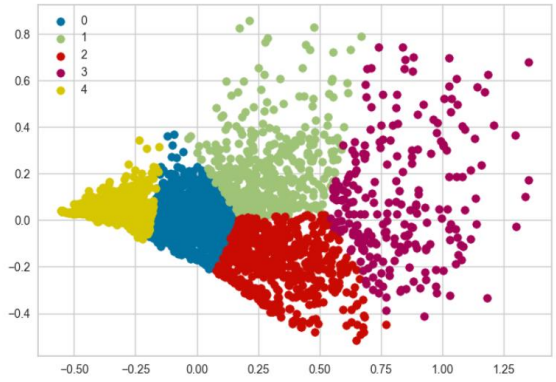
By far the most common are agglomerative hierarchical clustering techniques.

4. BIRCH

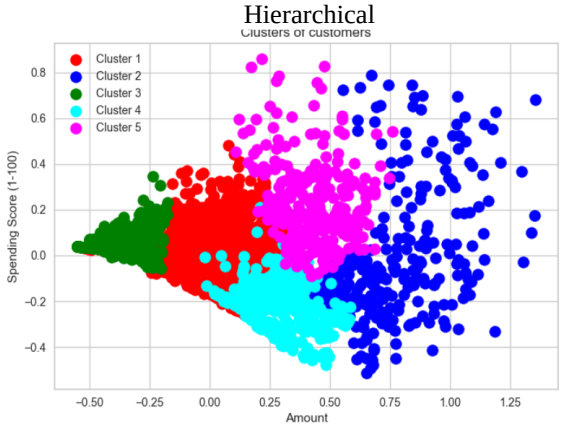
Another approach to clustering is to summarize the data, typically in a single pass, and then cluster the summarized data. The BIRCH algorithm uses a similar concept.

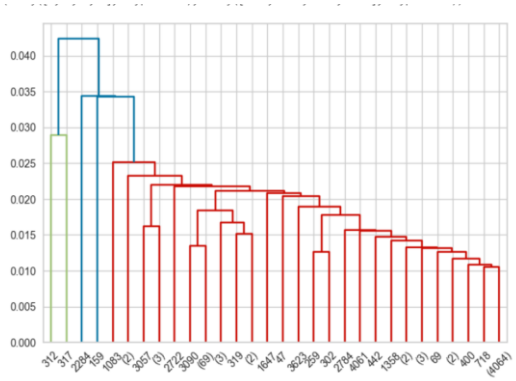
A. The different clusters obtained in the project using the clustering analysis

K-Means



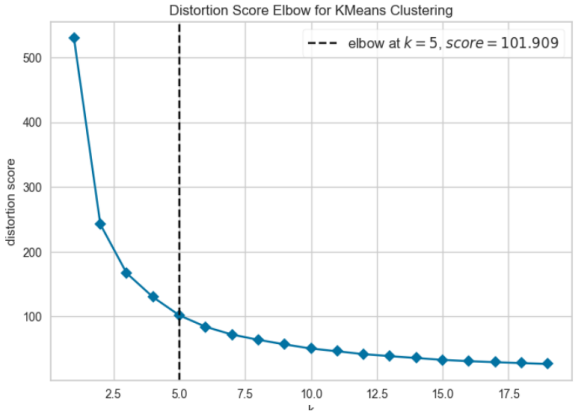
Hierarchical



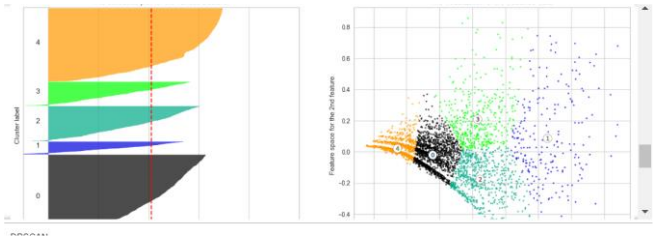


B. Finding the best parameter values

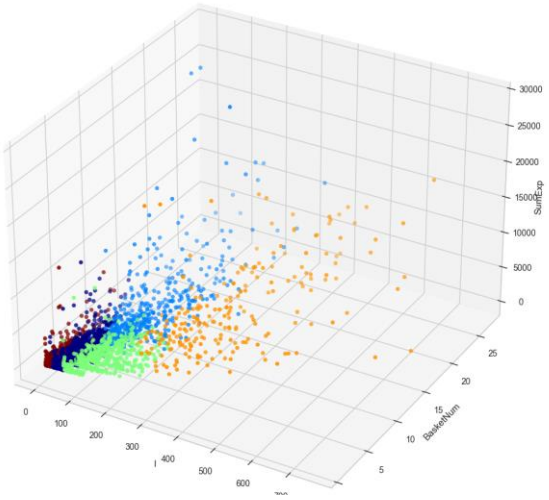
K value using elbow technique



Silhoute Analysis



3D Visualization of the clusters



CLUSTERING RESULTS

The groups are organized as follows:

• Clusters 0 and 1 refer to customers who visit the store on occasion and buy a few items: Cluster 0 refers to customers who buy a few items, while Cluster 1 refers to customers who buy more items. K-means analysis separated them, but in real-life scenarios, they're very similar: they both represent the occasional customer.

Cluster 2 is similar to the previous ones, but the difference is the greater number of purchased items. In any case, they choose low-cost items because the SumExp is nearly equal to Clusters 0 and 1.

• Cluster 3 purchases nearly the same number of items as Cluster 2, but the frequency with which they do so differs.

V. PART II -CLASSIFICATION

The objective of this project is to categorize each consumer with a label that will indicate whether they are a low, medium or high spender. Instead of focusing on individual transactions to determine whether they are expensive or not, we believed that the best approach to define this is to consider the total amount that a particular client spent. Therefore, the basic principle is that a high-spending client is someone who has spent more than average: having a low number of baskets, even with a very high overall cost, having a small quantity of baskets may not be sufficient to quantify for a certain label.

To do this, we used the previously defined parameter SumExp, which is just the sum of all purchases made by a client. Labels are provided nominally as asked and accurately denote low, medium, and high.

The dataset was sorted using the SumExp parameter, and the labels were then added with the Label attribute set too low for the first third of the data, medium for the second, and high for the last third. Given that there are three alternative labels, our issue is one of multiclass categorization. Then, in order to forecast classes, we choose to take into account the following qualities for the reasons listed below:

• I: The total amount spent depends on the number of products purchased.

• BasketNum: It stands to reason that a customer's frequency of visits to the store affects its overall cost.

• Entropy: The more this parameter increases, the more transactions there will be and the more expensive the things will be.

All the other irrelevant attributes are deleted in our project.

The complete dataset was separated into a train-set (70%) and a test-set (30%), which are obviously distinct as the Holdout method needs. This was the conclusion of our pre-processing phase.

Model Selection

In classification I used 5 classification methods

1. Naïve Bayesian
2. Decision Tree
3. Random Forest
4. KNN
5. SVM

In each of them I used hyper parameter tuning by randomly selecting parameters' combinations and tried to find the best model. Due to all of the pre-processing, on our constructed data- there is no missing values, the noise has been reduced with the outlier treatment and irrelevant attributes were already dropping at the beginning.

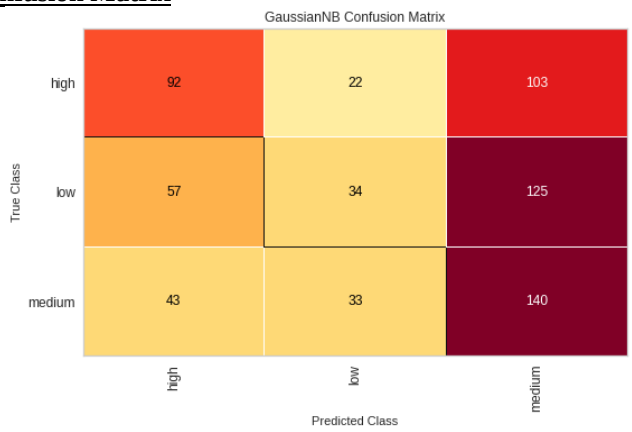
Naïve Bayesian

Train and Test accuracies are

Train score 0.4355

Test score 0.4292

Confusion Matrix



Classification Report

	precision	recall	f1-score	support
high	0.48	0.42	0.45	217
low	0.38	0.16	0.22	216
medium	0.38	0.65	0.48	216
accuracy			0.41	649
macro avg	0.41	0.41	0.38	649
weighted avg	0.41	0.41	0.38	649

Decision Tree

Using hyper parameter tuning find out the best parameters

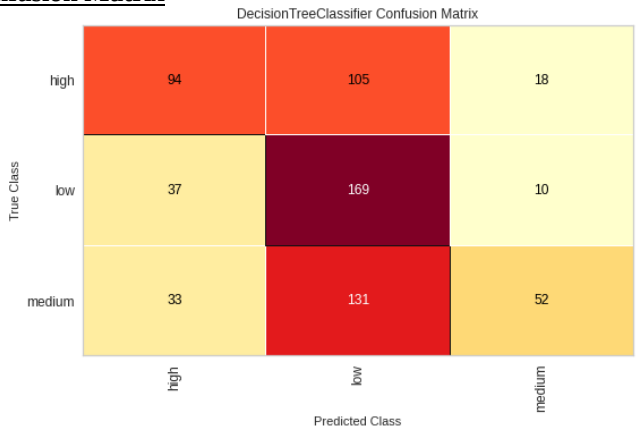
Best params: {'criterion': 'gini', 'max_depth': 5, 'max_features': 2, 'random_state': 3}

Train and Test accuracies are

Train score 0.4355

Test score 0.4292

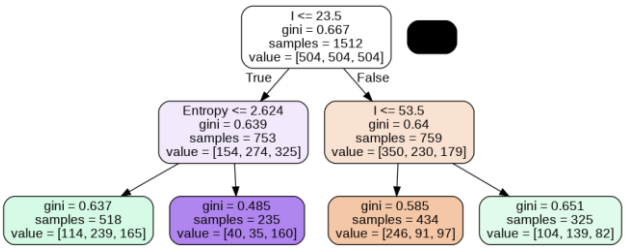
Confusion Matrix



Classification Report

	precision	recall	f1-score	support
high	0.57	0.43	0.49	217
low	0.42	0.78	0.54	216
medium	0.65	0.24	0.35	216
accuracy			0.49	649
macro avg	0.55	0.49	0.46	649
weighted avg	0.55	0.49	0.46	649

visualize the obtained decision tree till depth 3



Random Forest

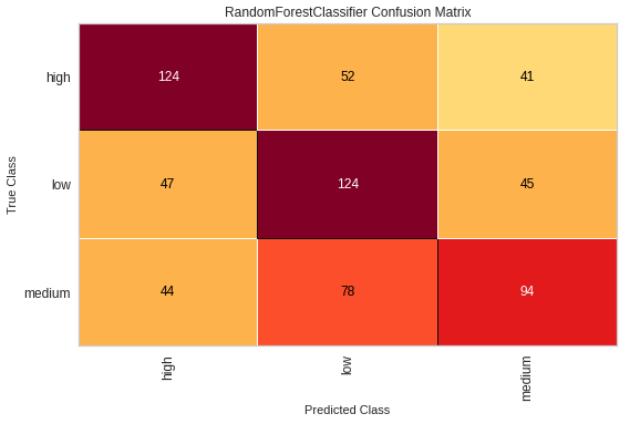
Using hyper parameter tuning find out the best parameters

Best params: {'criterion': 'entropy', 'max_depth': 7, 'max_features': 'sqrt', 'n_estimators': 300}

Train and Test accuracies are

Train score 0.6729
Test score 0.5397

Confusion Matrix

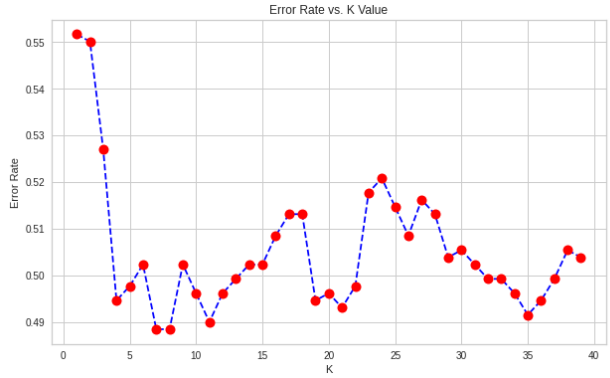


Classification Report

	precision	recall	f1-score	support
high	0.58	0.57	0.57	217
low	0.49	0.57	0.53	216
medium	0.52	0.44	0.47	216
accuracy			0.53	649
macro avg	0.53	0.53	0.53	649
weighted avg	0.53	0.53	0.53	649

KNN

To find out the k value find out the error rate vs k value graph, randomly taken maximum k value = 40

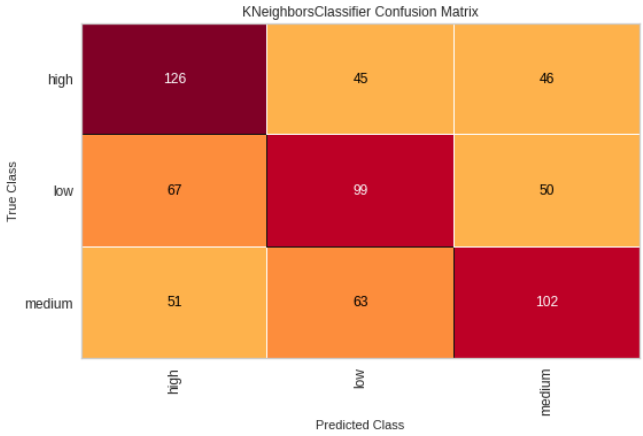


From this graph, k=12 has found out

Train and Test accuracies are

Train score 0.5847
Test score 0.5212

Confusion Matrix



Classification Report

	precision	recall	f1-score	support
high	0.52	0.58	0.55	217
low	0.48	0.46	0.47	216
medium	0.52	0.47	0.49	216
accuracy			0.50	649
macro avg	0.50	0.50	0.50	649
weighted avg	0.50	0.50	0.50	649

SVM

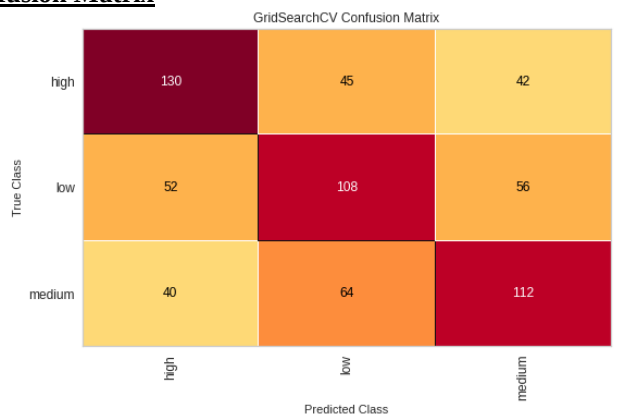
Using hyper parameter tuning find out the best parameters

Best params: {'C': 1000, 'gamma': 0.01, 'kernel': 'rbf'}

Train and Test accuracies are

Train score 0.6749
Test score 0.5271

Confusion Matrix



Classification Report

	precision	recall	f1-score	support
high	0.59	0.60	0.59	217
low	0.50	0.50	0.50	216
medium	0.53	0.52	0.53	216
accuracy			0.54	649
macro avg	0.54	0.54	0.54	649
weighted avg	0.54	0.54	0.54	649

Comparison of used predictive models

Finally, summarizing the timings and accuracy requirements on the train set and test set is presented.

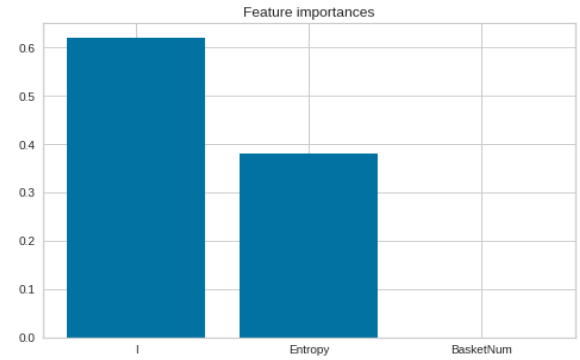
Algorithm	Fit Time (ms)	Prediction Time (ms)	Accuracy (Train Set)	Accuracy (Test Set)
Naive Bayesian	0.0049	0.0029	0.435	0.429
Decision Tree	0.52	0.0022	0.52	0.512
Random Forest	1.063	0.0841	0.673	0.539
KNN	0.004	0.016	0.585	0.521
SVM	12.546	0.029	0.675	0.527

The complexity dominant factor—the quantity of data in our training set—determines how slow SVM is. Given that we choose the rbf kernel, this can be marginally decreased by employing the rbfSVC classifier because it has a superior implementation.
The parameter n estimators (=300) in the final execution is the rationale given for Random Forest being the slowest of all the approaches. More values indicate higher performance at the expense of being slower because this parameter reflects the amount of trees that the algorithm will generate.

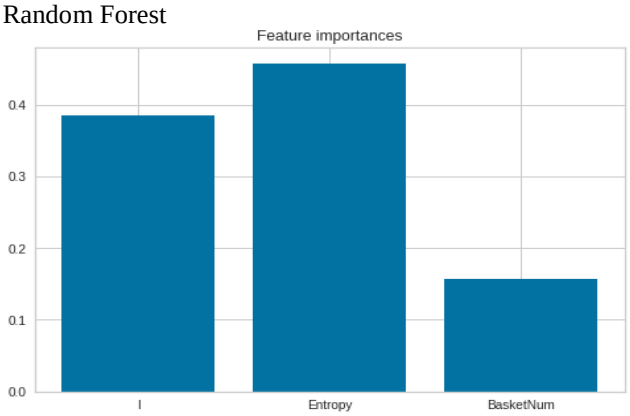
Feature Importance

Here are going to compare the feature importance in decision tree approach and Random Forest approach
On the train set and the test set, Random Forest achieves an improvement in accuracy of almost 15% and 2%, respectively. it's interesting to see how the relevance of the Feature Importance varies between the two approaches.

Decision Tree



Random Forest



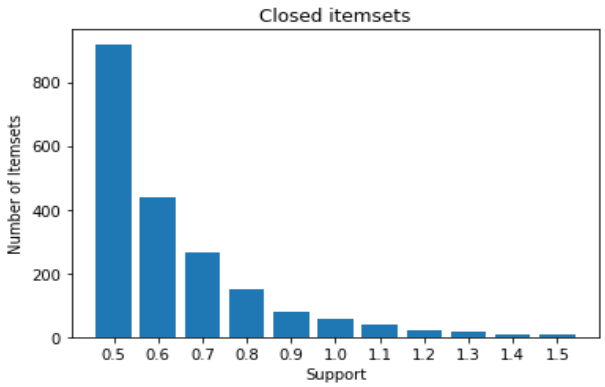
Conclusion

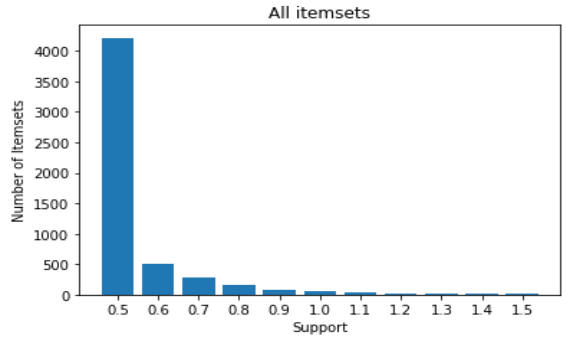
The Random Forest classification technique yields superior results, as was already said, but at the expense of time. Because of all of our data processing, the time cost is negligible, but for much larger datasets, this could be a concern.

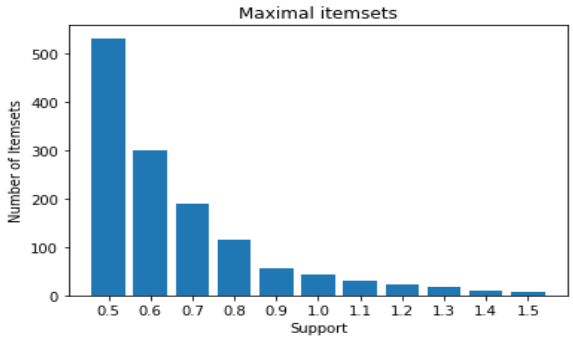
Association Rules Mining

We must first compute the frequently occurring itemsets. The Qta attribute in our initial dataset already provides the default guarantee that no item in any basket will be repeated more than once, which is a need for the construction of a set.

The comparison can be performed between various sets, always with a large number of support values. If would be expected, as the support value is increased, fewer itemsets are produced.







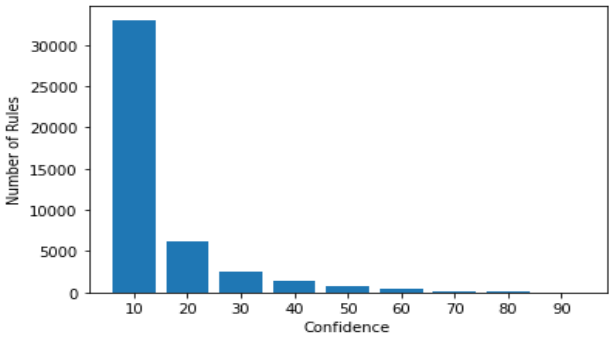
It's interesting to note that our results for closed itemsets and the most general case all itemsets are exactly the same. By definition, maximal is a subset of closed and is itself a subset of all, but because these last two collide, there isn't a superset that has the same support out of all of them.

The most significant patterns with the highest supports, which are the same for all types of itemsets, are listed in below.

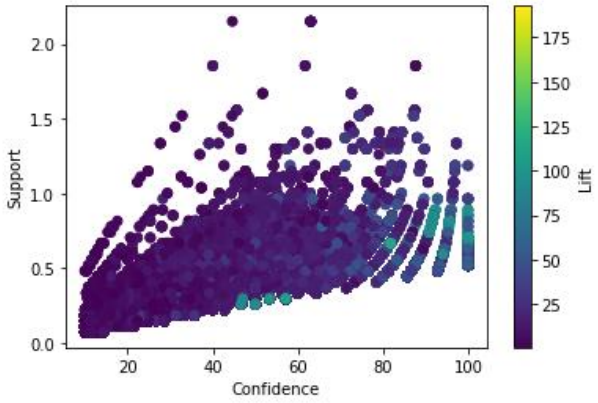
	Itemset	Count	Support
57	(HEART OF WICKER LARGE, HEART OF WICKER SMALL,...	58	2.15
40	(CANDLEHOLDER PINK HANGING HEART, RED HANGING ...	50	1.85
38	(WOODEN PICTURE FRAME WHITE FINISH, WOODEN FRA...	45	1.67
6	(ALARM CLOCK BAKELIKE ORANGE, ALARM CLOCK BAKE...	42	1.56
24	(ALARM CLOCK BAKELIKE PINK, ALARM CLOCK BAKELI...	42	1.56
55	(RED HANGING HEART T-LIGHT HOLDER, HEART OF WI...	41	1.52
23	(POPPY'S PLAYHOUSE LIVINGROOM , POPPY'S PLAYHO...	41	1.52
56	(RED HANGING HEART T-LIGHT HOLDER, HEART OF WI...	39	1.45
30	(WOOD 2 DRAWER CABINET WHITE FINISH, WOODEN PI...	38	1.41
28	(PINK 3 PIECE POLKADOT CUTLERY SET, RED 3 PIEC...	38	1.41

Association Rules

Following the creation of common patterns, we calculated the association rules using a confidence metric with a 70% threshold and illustrates the optimal trade-off between a necessary number of rules and a reasonable level of confidence in choosing the parameter.



Additionally, the scatter plot shows that our selection indicates the starting point at which the lift value becomes significant (from 20 up to nearly 70).



Association Rules

	Post	Pre	Support	Confidence	Lift
150	WHITE HANGING HEART T-LIGHT HOLDER	(CANDLEHOLDER PINK HANGING HEART, RED HANGING ...	1.853225	87.719298	5.814906
147	WOODEN PICTURE FRAME WHITE FINISH	(WOODEN FRAME ANTIQUE WHITE , WHITE HANGING HE...	1.667902	72.580645	15.665806
146	WOODEN FRAME ANTIQUE WHITE	(WOODEN PICTURE FRAME WHITE FINISH, WHITE HANG...	1.667902	72.580645	13.231255
123	ALARM CLOCK BAKELIKE GREEN	(ALARM CLOCK BAKELIKE PINK, ALARM CLOCK BAKELI...	1.556709	76.363636	15.375305
122	ALARM CLOCK BAKELIKE RED	(ALARM CLOCK BAKELIKE PINK, ALARM CLOCK BAKELI...	1.556709	76.363636	15.375305
55	ALARM CLOCK BAKELIKE GREEN	(ALARM CLOCK BAKELIKE ORANGE, ALARM CLOCK BAKE...	1.556709	87.500000	17.617537
54	ALARM CLOCK BAKELIKE RED	(ALARM CLOCK BAKELIKE ORANGE, ALARM CLOCK BAKE...	1.556709	76.363636	15.375305
166	WHITE HANGING HEART T-LIGHT HOLDER	(RED HANGING HEART T-LIGHT HOLDER, HEART OF WI...	1.519644	74.545455	4.941613
113	POPPY'S PLAYHOUSE KITCHEN	(POPPY'S PLAYHOUSE LIVINGROOM , POPPY'S PLAYHO...	1.519644	87.234043	31.805060
114	POPPY'S PLAYHOUSE BEDROOM	(POPPY'S PLAYHOUSE LIVINGROOM , POPPY'S PLAYHO...	1.519644	87.234043	34.611389
115	POPPY'S PLAYHOUSE LIVINGROOM	(POPPY'S PLAYHOUSE BEDROOM , POPPY'S PLAYHOUSE...	1.519644	74.545455	33.520606
167	WHITE HANGING HEART T-LIGHT HOLDER	(RED HANGING HEART T-LIGHT HOLDER, HEART OF WI...	1.445515	72.222222	4.787606
130	BLUE 3 PIECE POLKADOT CUTLERY SET	(PINK 3 PIECE POLKADOT CUTLERY SET, RED 3 PIEC...	1.408451	84.444444	27.449531
131	PINK 3 PIECE POLKADOT CUTLERY SET	(RED 3 PIECE RETROSPOT CUTLERY SET, BLUE 3 PIE...	1.408451	73.076923	23.471612
136	WOODEN FRAME ANTIQUE WHITE	(WOOD 2 DRAWER CABINET WHITE FINISH, WOODEN PI...	1.408451	79.166667	14.431869

ACKNOWLEDGMENT *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been

accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] *(references)*

[3] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[4] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[5] K. Elissa, “Title of paper if known,” unpublished.

[6] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[7] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[8] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.