Machine Learning Project (5343904)

Introduction

The task is to apply Rule based approach which specifically based on the man handling of rules on Groceries dataset, Classification task using 2 different approaches and solve the same data classification using clustering. We have used different approaches for the task such as probability based Bayes, Linear approach as Logistic Regression, Decision Trees. For the Clustering we have used Birch, K- Mean and agglomerative clustering and have depicted the clusters using PCA.

Task A - Data importing and cleaning

Data is imported from the excel file and we have dropped the data using a threshold where any column which has more than 66% missing values, has been dropped. We can use most frequent values for the imputation of missing values but it is highly unlikely the case as we never know which the customer would have bought so we can use K mean imputation too.

Task B

- **Association Rule:**
    - We have used Apriori based rule mining approach where we have defined the set rules based on the Support, Confidence and Lift.
    - Support is an indication of how frequently the items appear in the data. Mathematically, support is the fraction of the total number of transactions in which the item set occurs.
    - Confidence indicates the number of times the if-then statements are found true. Confidence is the conditional probability of occurrence of consequent given the antecedent.
    - Lift can be used to compare confidence with expected confidence. This says how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is.

```
Rule: hamburger meat -> Instant food products
Support: 0.003050330452465684
Confidence: 0.379746835443038
Lift: 11.42143769597027
====================================
Rule: whipped/sour cream -> baking powder
Support: 0.004575495678698526
Confidence: 0.25862068965517243
Lift: 3.607850330154072
====================================
Rule: root vegetables -> beef
Support: 0.017386883579054397
Confidence: 0.3313953488372093
Lift: 3.0403668431100312
====================================
Rule: whipped/sour cream -> berries
Support: 0.009049313675648195
Confidence: 0.27217125382262997
Lift: 3.796885505454703
====================================
Rule: liquor -> bottled beer
Support: 0.004677173360447382
Confidence: 0.4220183486238532
Lift: 5.253861340146324
```

It shows us here that the Root Vegetables and Beef are highly likely to be bought together just like liquor and bottled beer. The inference makes sense too as we can easily know that a person who drinks liquor is highly likely to drink beer too instead of milk

- Classification
    - The base Logistic Regression and Decision Tree model gave us a perfect score of 1 so we can say that the models are good and data is perfectly separable by the models.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	842
1	1.00	1.00	1.00	783
accuracy			1.00	1625
macro avg	1.00	1.00	1.00	1625
weighted avg	1.00	1.00	1.00	1625

Base Linear Logistic Regression Model is able to classify the data perfectly

-

	precision	recall	f1-score	support
0	1.00	1.00	1.00	842
1	1.00	1.00	1.00	783
accuracy			1.00	1625
macro avg	1.00	1.00	1.00	1625
weighted avg	1.00	1.00	1.00	1625

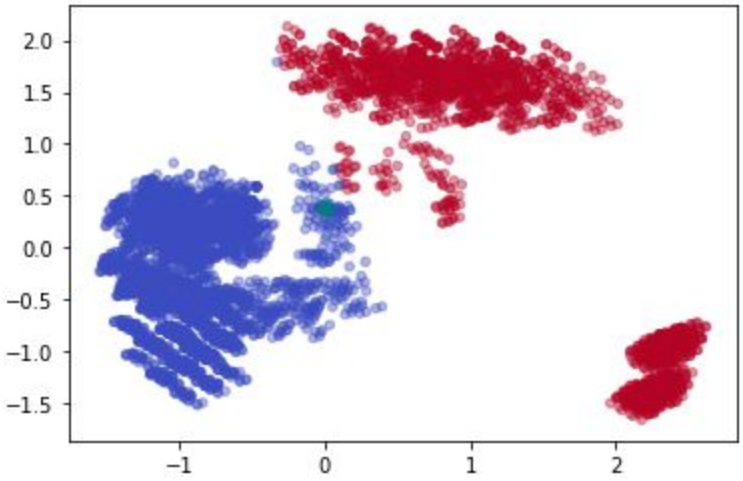
- Our classifier is able to classify the data to a complete precision, accuracy and recall of 1

- On the other hand, the Bayseian classifier is not so good with the data and gave us a score of around 0.95 f1-score on test data.

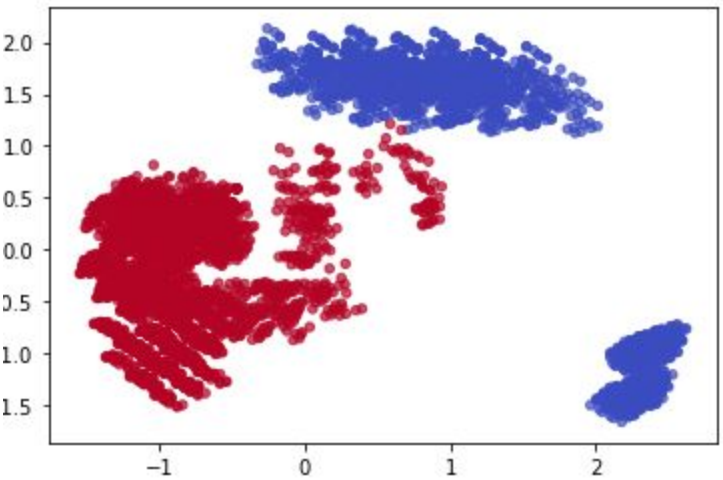
	precision	recall	f1-score	support
0	1.00	0.91	0.95	842
1	0.91	1.00	0.95	783
accuracy			0.95	1625
macro avg	0.96	0.96	0.95	1625
weighted avg	0.96	0.95	0.95	1625

Classifier is able to classify the data with an f1 score of 0.95 and precision of 0.96 and recall of 0.96. So we can say it is not a good model as it performed worse than the base model

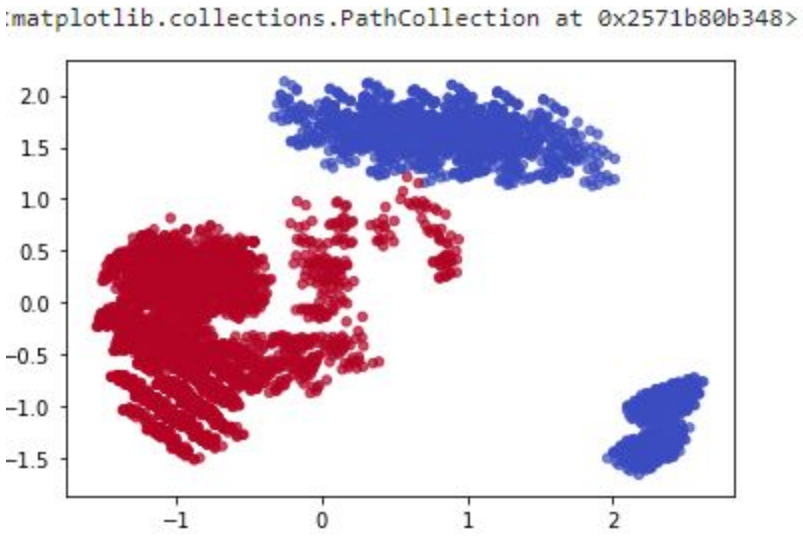
- **Clustering:**
    - We have performed 3 different techniques for clustering starting from K-Means clustering. ***K -Means*** cluster the data in groups but is not as effective as we can see by the plots below as there are inconsistencies with the groupings.



    - BIRCH is giving us perfect clusters but is heavy in terms of computations. We have asked BIRCH to form 2 clusters just like K-Means.



- Agglomerative with 2 clusters is the same as BIRCH in terms of cluster formations but is worse in terms of time and computation complexity.



Discussion

Logistic Regression is the simplest and the best model for our dataset as it is linear, simple and puts less strain on the processors. Decision Trees produce the same results but are far more heavy than the Logistic Regression. We should not use probability based methods as we do not have the info about the probability distributions of data.

K-Mean turns out to be a bad estimator of clusters for the data as it is mingling the data from different clusters. Agglomerative and BIRCH produce the same results and we can use their performances by using adjusted Index scores.