Final Project

Prepared for: Dr Eric Tham

Prepared by: Ashish Adhana(Student ID :13846427)
Riya Sen (Student ID: 13106631)
Clement Seah Han Liang (Student ID: 13847751)

Due date: 28/05/2021

Subject: CP3403 Data Mining
Final Group Project

Table of contents

Abstract

The aim of this study is to implement different types of preprocessing data mining methods and to resort to the various types of methods (i.e data cleaning, integration, reduction and transforming) on a dataset to come up with a business proposal. The goal is to be able to use the dataset to gain a better understanding of the data and apply them into real world situations.

After choosing the dataset of Mall Customer Segmentation that had information of customers of a mall through membership cards. The data went through pre-processing methods such as discretisation and removing unneeded attributes. Clustering was performed using K-means and hierarchical clustering to look for patterns in the data. Classification algorithms such as OneR, Naïve Bayes and J48 were used, to look for the model with the greatest accuracy on predicting which category of customers can be focused and targeted for future promotions and sales by the client side's marketing team and planning strategists.

Introduction

The purpose of this report would be to analyse a given set of data using different algorithms and find the different patterns arising due to it. These patterns found would then be used in real life scenarios to find out the trends. The analysis would be of the spending patterns of various consumers who enter the mall.

The objective of this analysis would be to use different types of pre-processing, clustering and association methods which are crucial to data mining to process the given data to gather the best results possible. The different methods used in the report would utilize different techniques and the results would therefore be explained thoroughly and compared with each other for more clarity on the analysis. This would help discover their precision, adequacy and productivity.

The result would give a more clearer idea as to what affects the spending pattern of customers who enter the mall each day. The results derived would have the most accuracy percentage of the correctly identified instances. The results would also have less false positives and false negatives.

Justification of Dataset

The dataset used in this report is obtained from Kaggle repository. This dataset was chosen as it has a lot of information about the different characteristics of customers who enter the mall. This gives a lot of variables to be analysed and achieve more accurate results. The dataset consists of various variables :

- Customer ID
- Gender
- Age
- Annual Income
- Spending Score

Objectives

The objective of this study is to recommend a market strategy to the client based on the customer's age, gender, annual income and spending score which has been presented from the demographics. The study's scrutiny will be done using clustering methods like K-means and Agglomerative and classification methods like One-R, Naïve Bayes and J48, provide instant results using its accuracy from the total instances based on the training set and display the results via Weka's classifier output for viewing and analysing.

Data set details

The "Mall Customer Segmentation Data" from Kaggle Data Repository was selected for scrutinisation to understand which of the customers can be combined and targeted on the basis of their Age, Annual Income and Spending Score to assist the marketing team in working out their game plan. It has a total of 200 records with some data that needs to be cleaned and transformed. There are some attributes which are not necessary to be taken in record while implementing the classification and clustering algorithms.

Pre-processing method

The purpose of preprocessing is that it helps us to understand the dataset better. Preprocessing consists of: switching the format of the data, data cleaning, and data discretization which can be either from the supervised or unsupervised filter.

The dataset provided by Kaggle was already present in .csv format (Comma Separated Values), therefore the dataset was not necessary to convert to any other format. We started the process by removing the attribute CustomerID, which seemed irrelevant in this dataset as the CustomerID is unique to each customer, meaning each customer has been ranked by appointing them their respective IDs. Removing the CustomerID attribute also implied removal of noisy data, aka data cleaning. During the process of data cleaning we need to make sure that there is no data which is either blank or unique to each customer.

The next step was to apply the Discretize filter from the weka>>filters>>unsupervised >>attribute>>Discretize. The function of Discretize is used to create 'n' number of bins based on our requirement and also tries and converts the numeric attributes in the dataset to nominal values. We created four bins to segregate the Age, Annual Income and Spending Score of the customers to help us understand and mine the improved data.

Following which, in order to make the dataset more accurate and to make the dataset more interpretable, we applied the filter 'NumerictoNominal' to all attributes hence ensuring the results produced will be more precise and meticulous.

Now after the pre-processing, our data looked like this:

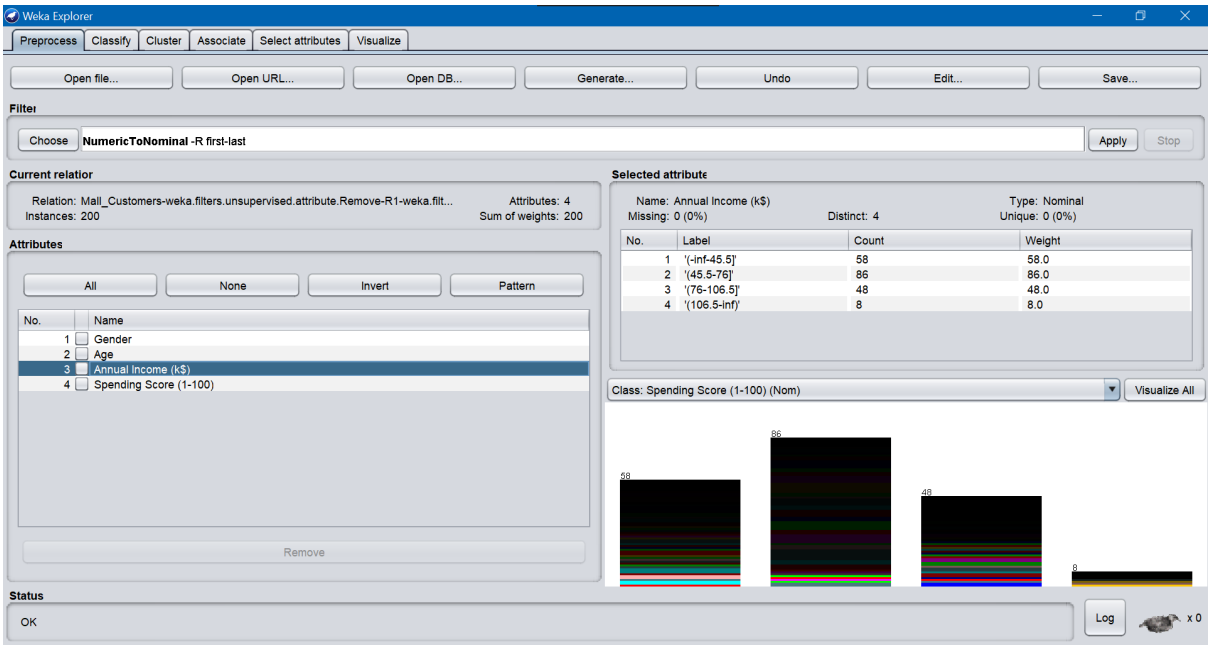


Fig 1.1

Visualized:

Fig 1.2

Methods of Machine Learning Used

SimpleK-means Clustering

K-means Clustering is a type of partitioning clustering that produces a single-level clustering result. In K-means clustering, the value for k, the number of clusters, must be predetermined. The algorithm will then randomly choose k points to be the cluster centers. The remaining instances are then assigned to their nearest cluster center one at a time, calculating a new cluster center each time a new instance is assigned to it. After all cluster centers are determined, we reassign the instances to the clusters center until the cluster centers do not change.

Hierarchical Clustering

Hierarchical clustering is a type of algorithm which groups people with similar characteristics. These groups are called clusters. The endpoint consists of a set of clusters where each cluster differs from each other and the attributes of each cluster are somewhat similar to each other. It works by treating each finding as a separate cluster. Then it repeats the following two steps on a loop: (1) identify the two clusters that are closest together, and (2) merge the two most similar clusters. This loop continues until the merging is done for all the clusters. After the merging is done we can visualise the tree and see where to split the clusters.

One-R Classification

The One-R algorithm is a part of the rules family in data mining classification. It is straightforward and to the point. This technique produces one rule for each instance in the dataset, and then proceeds with the rule which has the narrowest total errors. To further explain, the method basically would generate rules for each value in the respective attribute it is processing. Further, it would determine the absolute error rate from all the values given in the attribute. This procedure within this algorithm continues for each said attribute present in the dataset and compares the absolute error value with the previous attribute it evaluated. This method further keeps eradicating its previous absolute error rate if its value was higher than the error rate of the current attribute. Ultimately, the algorithm comes up with the best attribute with the absolute lowest error rate, and that particular attribute becomes ‘one rule’ for the entire dataset. The only drawback of this algorithm is that the rule produced are a little confusing for a simple human to understand and interpret, and the results are less precise when compared to other classification algorithms which are said to be state-of-the-art .

Naïve Bayes

Naïve Bayes Algorithm is a contingency based machine learning technique. The upside of implementing this algorithm is that it can compute better results using a small dataset, easy to implement and can handle extensive attribute statistics. Compared to decision trees like the J48, its performance is proportionate. Naive Bayes algorithm is a part of the Bayes family of the list of classifiers available on Weka. It implements the Naive Bayes theorem with stable self-reliant expectations. The drawback this algorithm faces is that the data may not be that accurate given its diminished calculated time along with attributes which have total dependencies on each other.

J48 (Decision Tree)

J48 is also known as the decision tree algorithm. This classifying algorithm helps visualize our results in a tree-like structure with measured rules maintained by the algorithm.
How j48 works is as follows: the accurate data is obtained by deducting pre breakdown of the input and the post breakdown of the input data. After which the based on the outcome branches are constructed by the algorithm and it puts forward the best branch as its best instance. As far as the best choice is determined, the said steps are repeated by the algorithm. The procedure then takes place repeatedly for each branch till the structure of the right side of the leaf is developed.

Layout of Data Mining Process

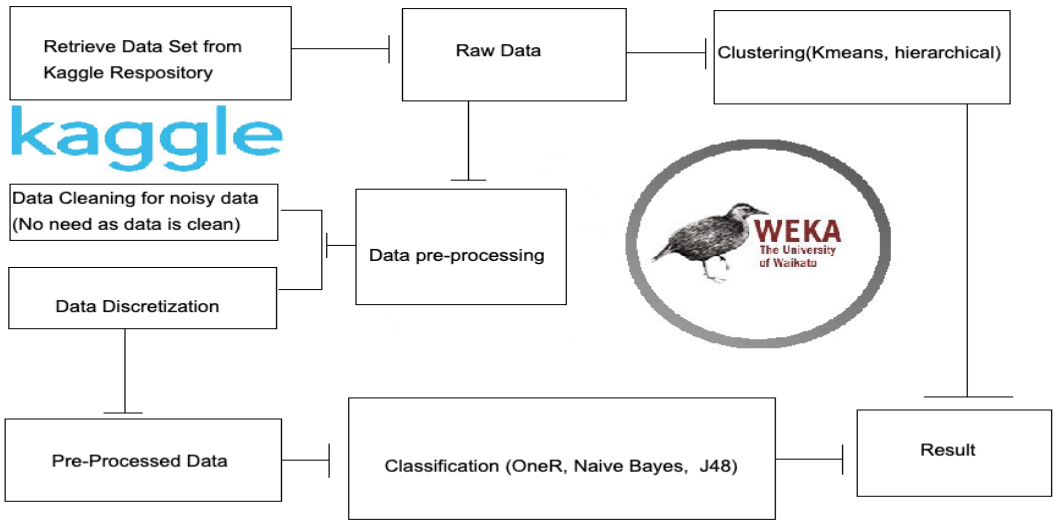


Fig 1.1

Pre-Processing Details

Raw Data Visualized:

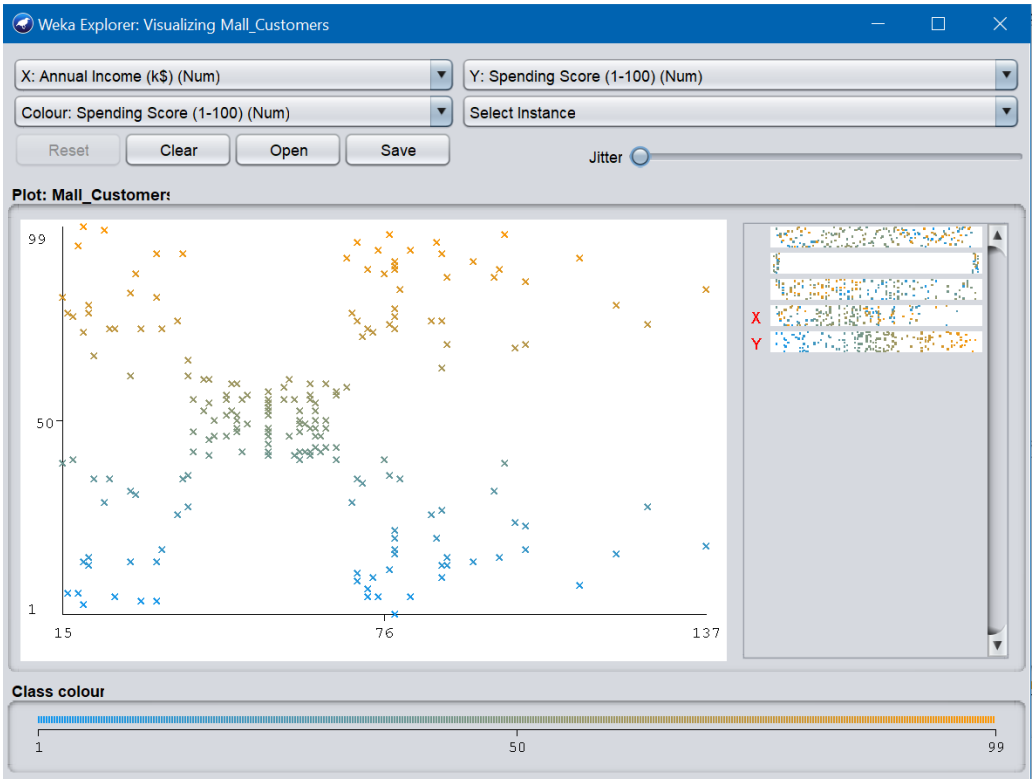


Fig 2.1

Results

K-means Clustering

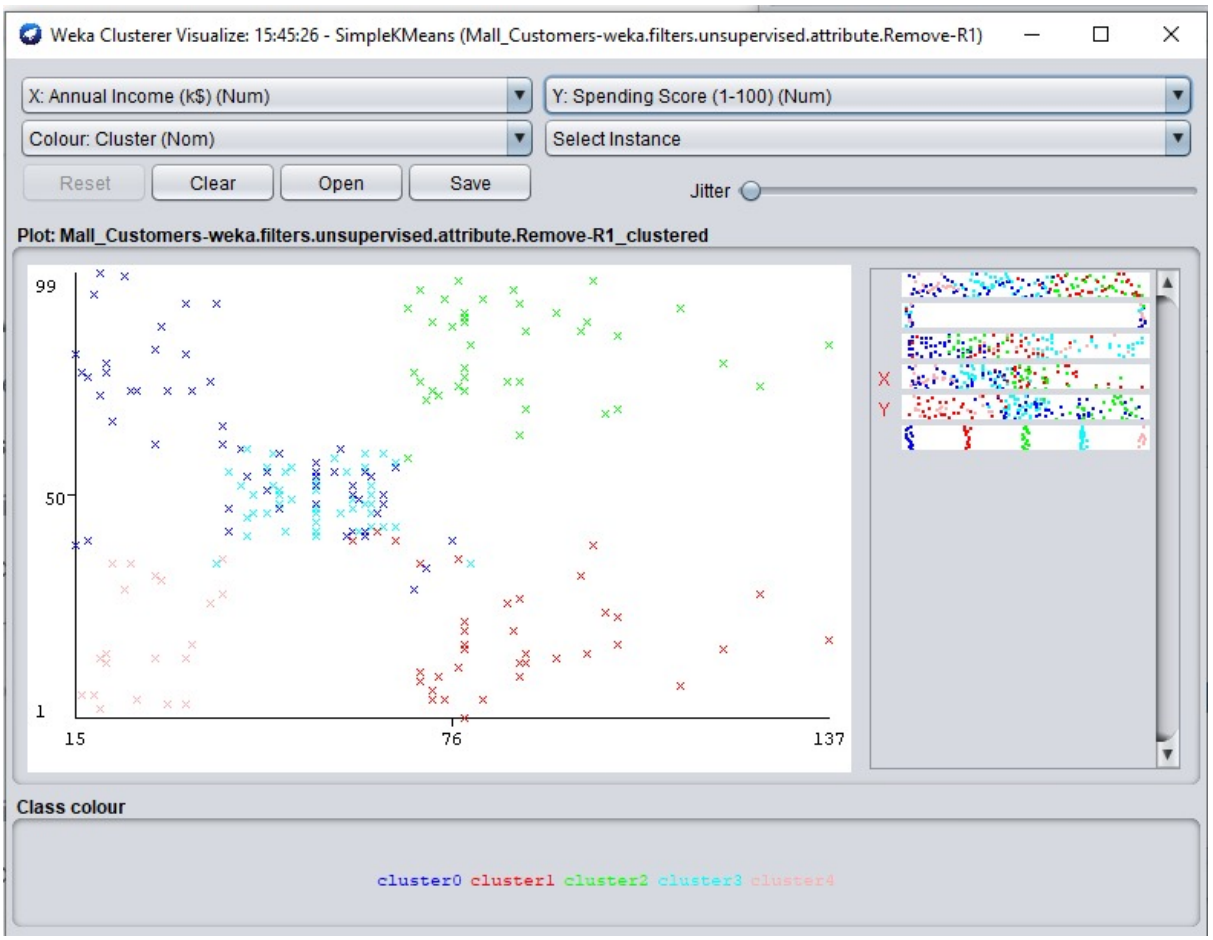


Fig 3.1

```
Number of iterations: 10
Within cluster sum of squared errors: 10.611100503926322

Initial starting points (random):

Cluster 0: 19,64,46
Cluster 1: 54,101,24
Cluster 2: 22,57,55
Cluster 3: 50,58,46
Cluster 4: 48,39,36

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2          3          4
                          (200.0)     (57.0)     (35.0)     (43.0)     (45.0)     (20.0)
=============================================================================================
Age                        38.85     25.614    41.1714    32.7674    56.3333      46.25
Annual Income (k$)         60.56    44.4737    87.8857    81.9535    54.2667      26.75
Spending Score (1-100)      50.2    57.9825    17.4286    82.5581    49.0667      18.35



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===
```

Fig 3.2

For the K-means clustering, we used the raw data set and chose K to be 5. By visualising the spending score and annual income graph we can see 5 clumps of instances. Using K-means clustering and ignoring the gender attribute, we can see that there are 5 distinct clusters in Fig 3.1.

From the figure we can see that the mall should aim their marketing efforts to cluster 1,3,4, as they are the least likely to spend in the mall. From Fig 3.2, we can see that clusters 1,3,4 are customers with the average age of over 40. This shows that regardless of annual income the customers age over 40 are less likely to hit a spending score over 75. Therefore, using K-means clustering, we can determine that the mall's marketing team should appeal more towards those over 40.

Hierarchical Clustering (Agglomerative)

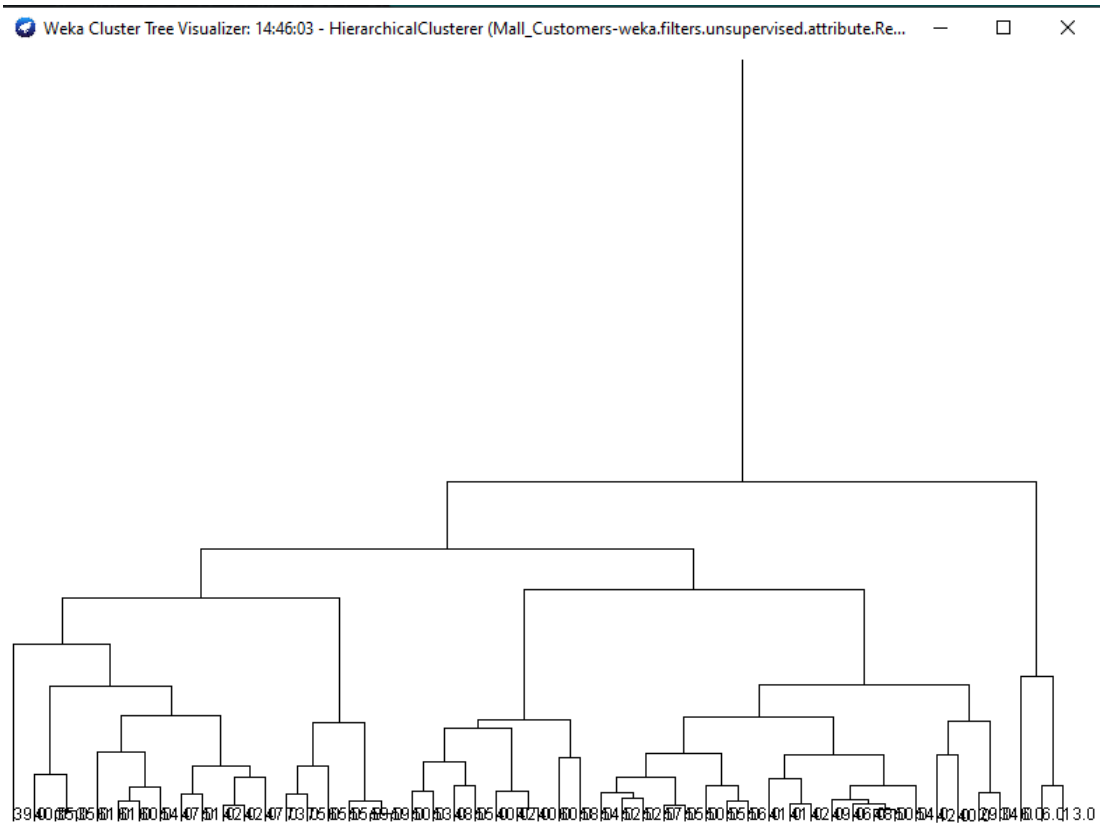


Fig 3.3

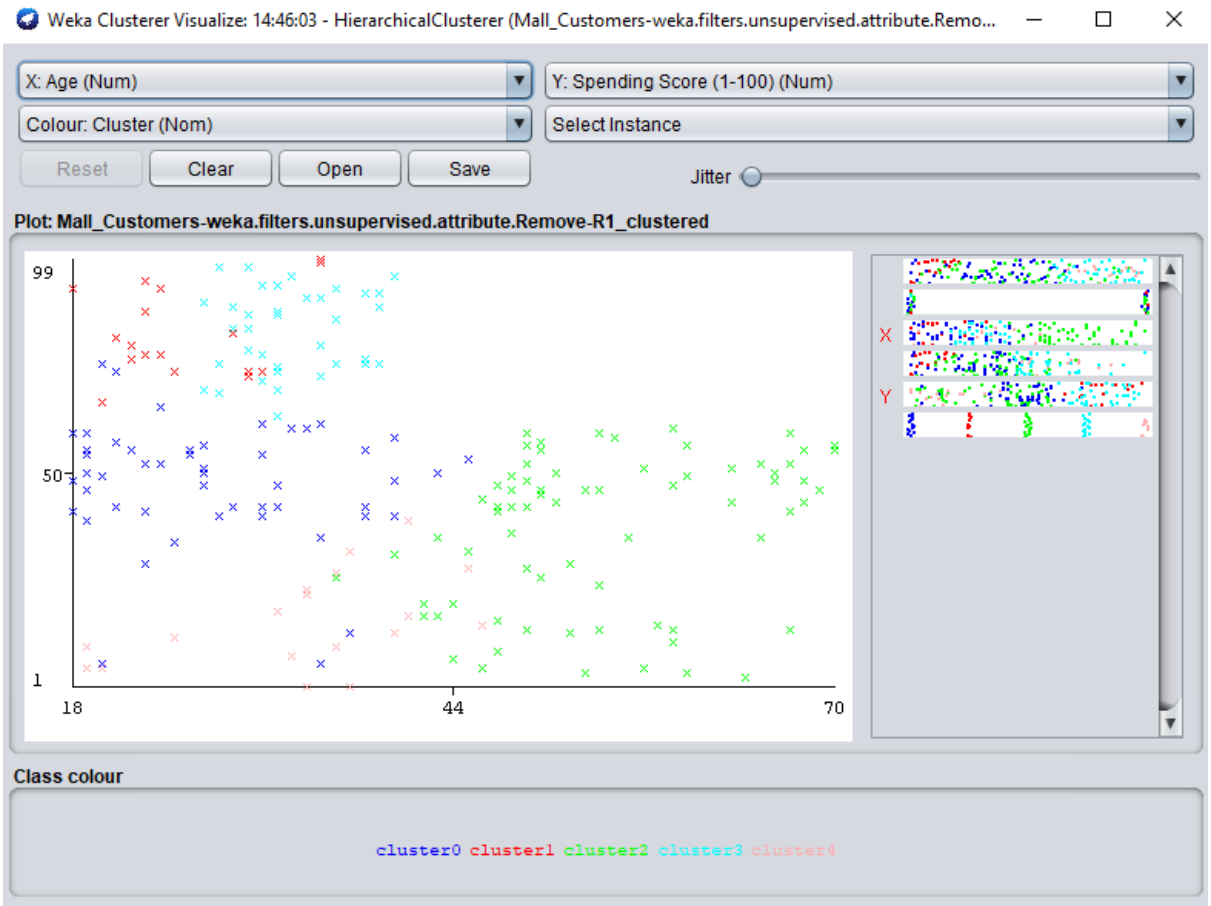

Weka Clusterer Visualize: 14:46:03 - HierarchicalClusterer (Mall_Customers-weka.filters.unsupervised.attribute.Remo...
X: Age (Num)
Y: Spending Score (1-100) (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: Mall_Customers-weka.filters.unsupervised.attribute.Remove-R1_clustered
99
50
1
18
44
70
X
Y
Class colour
cluster0 cluster1 cluster2 cluster3 cluster4

Fig 3.4

Fig 3.5

For hierarchical clustering, we used the raw data set and chose the complete link type to cluster the results. By visualising the tree in Fig 3.3, we can see that 5 clusters is the most optimal number to split the clusters.

After splitting the clusters we can see that by visualising the Age to spending score chart Fig 3.4, clusters 0,2,4 are the clusters that have the least spending score. Cluster 2 are all the customers that age are above 44.

By looking at annual income of the clusters in Fig 3.5, we can see that the annual income is not indicative of the spending score of the customers. As cluster 4 has a high annual income while still not having a high spending score.

OneR Result

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.rules.OneR -B 6
Relation:     Mall_Customers-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Discretize-B4-M-1
Instances:    200
Attributes:   4
              Gender
              Age
              Annual Income (k$)
              Spending Score (1-100)
Test mode:    evaluate on training data

=== Classifier model (full training set) ===

Spending Score (1-100):
        1       -> '(76-106.5]'
        3       -> '(-inf-45.5]'
        4       -> '(-inf-45.5]'
        5       -> '(45.5-76]'
        6       -> '(-inf-45.5]'
        7       -> '(45.5-76]'
        8       -> '(106.5-inf)'
        9       -> '(45.5-76]'
        10      -> '(45.5-76]'
```

Fig 3.6

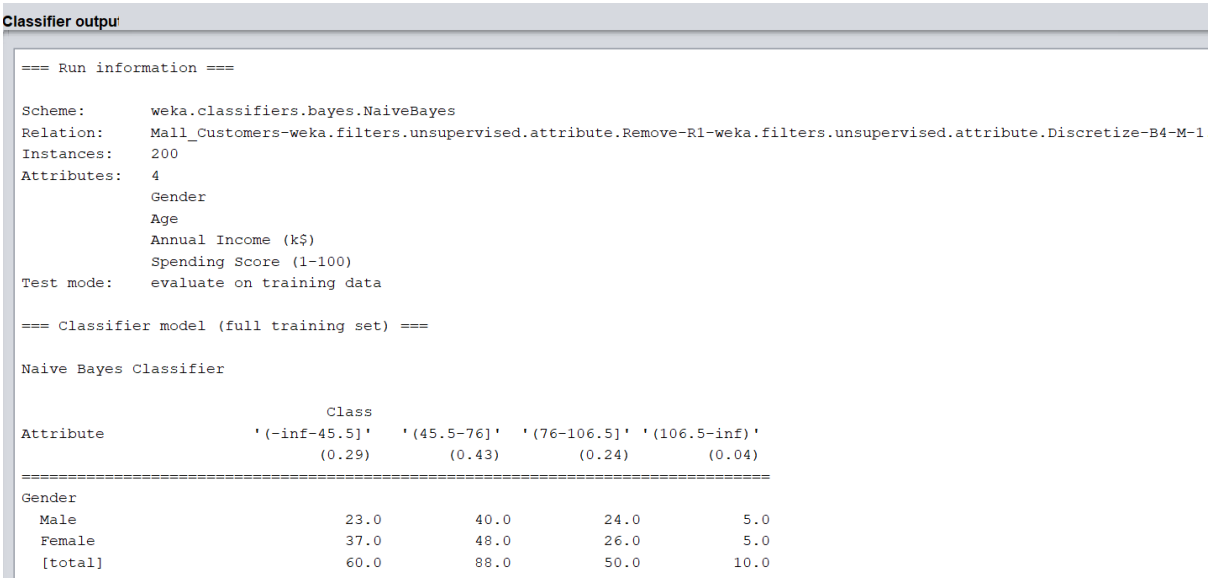
Classifier output

```
=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances         156               78      %
Incorrectly Classified Instances        44               22      %
Kappa statistic                          0.6639
Mean absolute error                      0.11
Root mean squared error                  0.3317
Relative absolute error                 32.6733 %
Root relative squared error             80.9276 %
Total Number of Instances              200

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.741    0.127    0.705      0.741    0.723      0.606    0.807     0.598     '(-inf-45.5]'
                 0.907    0.167    0.804      0.907    0.852      0.733    0.870     0.769     '(45.5-76]'
                 0.688    0.046    0.825      0.688    0.750      0.685    0.821     0.642     '(76-106.5]'
                 0.250    0.000    1.000      0.250    0.400      0.492    0.625     0.280     '(106.5-inf)'
Weighted Avg.    0.780    0.119    0.788      0.780    0.772      0.675    0.830     0.669
```

Fig 3.7

We used the training set for this algorithm. Referring to Fig 3.7, we kept the Annual Income as the main class and there were a total of 156 correctly classified instances as in 78% and 44 of incorrectly classified instances, as in 22%. This approach chose Spending Score as it's 'one rule' as seen in Fig 3.6. The TP rate is 90% as seen in the figure above and FP rate is 12%, which seems decent since the algorithm has predicted that customers with Annual Income falling under 45,500$-70,000$ have the highest spending score. Even with a 78% accuracy unfortunately, this algorithm result cannot be trusted because it relies on only one attribute and computes its results.

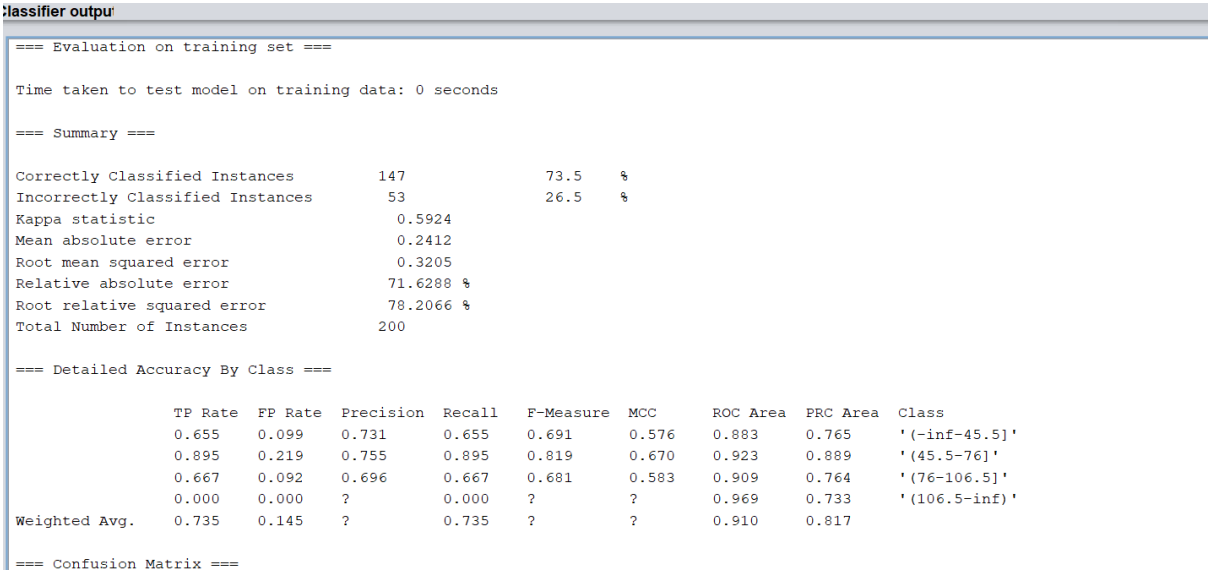
Naïve Bayes Result



```
Classifier output

=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     Mall_Customers-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Discretize-B4-M-1
Instances:    200
Attributes:   4
              Gender
              Age
              Annual Income (k$)
              Spending Score (1-100)
Test mode:    evaluate on training data

=== Classifier model (full training set) ===

Naive Bayes Classifier

                              Class
Attribute               '(-inf-45.5]'   '(45.5-76]'  '(76-106.5]' '(106.5-inf)'
                              (0.29)        (0.43)        (0.24)        (0.04)
================================================================================
Gender
  Male                         23.0          40.0          24.0           5.0
  Female                       37.0          48.0          26.0           5.0
  [total]                      60.0          88.0          50.0          10.0
```

Fig 3.8

For the Naïve Bayes algorithm Annual Income was set as the class and its columns were compared with the rest that is Gender, Age and Spending Score. As seen in Fig 3.8, there are 23 Males and 37 Females who fall under 45,500$ category, 40 Males and 48 Females who fall under the 45,500-76000 category, 24 Males and 26 Females who fall under 76000$-106,500$ category and 5 Males and Females respectively who fall under the category of 106,000$ and more.



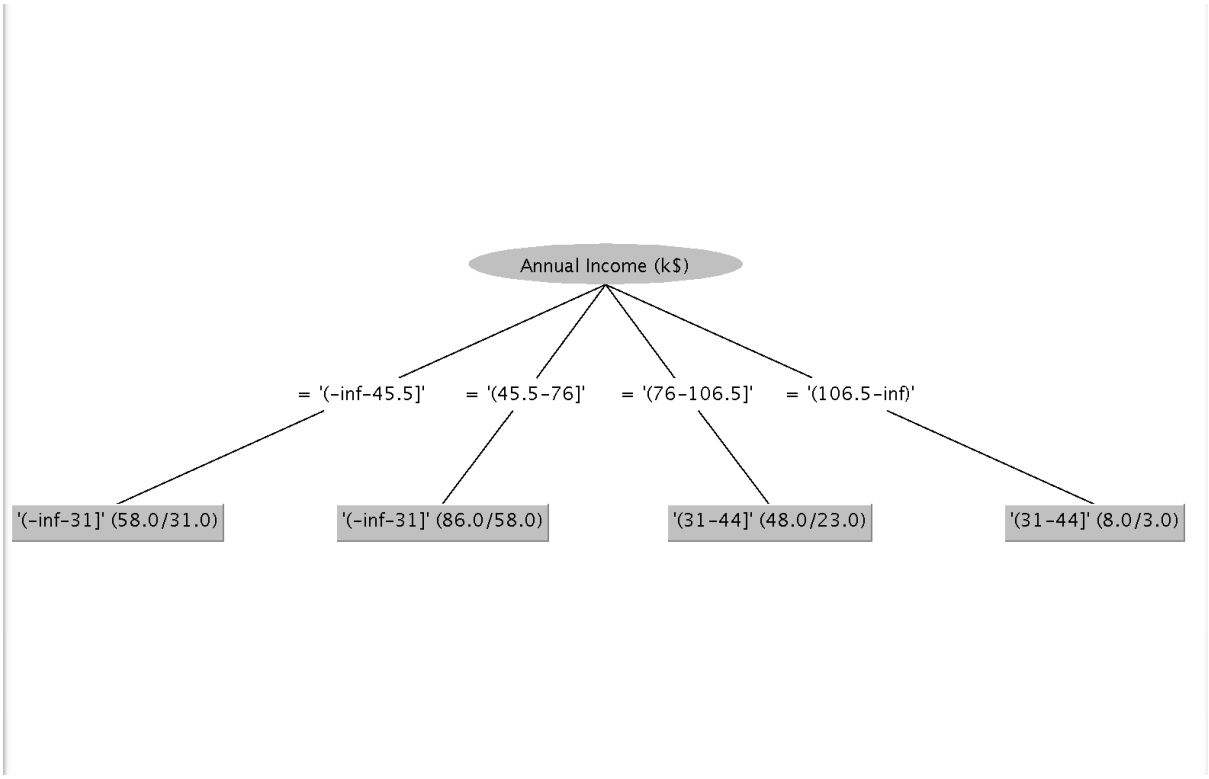
```
Classifier output

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances         147               73.5    %
Incorrectly Classified Instances        53               26.5    %
Kappa statistic                          0.5924
Mean absolute error                      0.2412
Root mean squared error                  0.3205
Relative absolute error                 71.6288 %
Root relative squared error             78.2066 %
Total Number of Instances              200

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.655     0.099     0.731      0.655     0.691     0.576     0.883     0.765     '(-inf-45.5]'
               0.895     0.219     0.755      0.895     0.819     0.670     0.923     0.889     '(45.5-76]'
               0.667     0.092     0.696      0.667     0.681     0.583     0.909     0.764     '(76-106.5]'
               0.000     0.000     ?          0.000     ?         ?         0.969     0.733     '(106.5-inf)'
Weighted Avg.  0.735     0.145     ?          0.735     ?         ?         0.910     0.817

=== Confusion Matrix ===
```

Fig 3.10

As pictured in Fig 3.10, there are 147 correctly classified instances and 53 incorrectly classified instances with an accuracy of 73.5%. The TP and FP rate of the category of the Annual Income 45,500$-76,000$ is 89% and 21% respectively.

J48 Results

```
=== Summary ===

Correctly Classified Instances          85               42.5    %
Incorrectly Classified Instances       115               57.5    %
Kappa statistic                          0.1322
Mean absolute error                      0.3434
Root mean squared error                  0.4144
Relative absolute error                 95.1446 %
Root relative squared error             97.5793 %
Total Number of Instances              200

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.786    0.685    0.382      0.786    0.514      0.107    0.592     0.409     '(-inf-31]'
                 0.492    0.187    0.536      0.492    0.513      0.312    0.667     0.441     '(31-44]'
                 0.000    0.000    ?          0.000    ?          ?        0.558     0.238     '(44-57]'
                 0.000    0.000    ?          0.000    ?          ?        0.661     0.185     '(57-inf)'
Weighted Avg.    0.425    0.297    ?          0.425    ?          ?        0.617     0.353

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 55 15  0  0 |  a = '(-inf-31]'
 31 30  0  0 |  b = '(31-44]'
 34  9  0  0 |  c = '(44-57]'
 24  2  0  0 |  d = '(57-inf)'
```



As seen in the above figure, the summary shows correctly classified instances at 42.5%. The incorrect instance is at 57.5%. The TP rate is 0.425. In the above tree diagram, it can be seen that the main attribute selected is the Annual Income. People with annual income of upto $45.5k, are normally aged upto 31and spend a little less. People with annual income from $45.5k to $76k who are also normally aged until 31 years tend to spend the most.

Issues

The issues we faced were as follows:

1. Previously, the two datasets Telco and StockXSneakers we had chosen from the Kaggle dataset repository didn't give clear results in the clustering aspect of data mining. We were hoping to study, analyze and provide our predictions in each aspect, i.e Clustering and Classification.
2. DBSCAN computation was not possible as well since the data had little or no noise points apparently and the epsilon value was not applying to the dataset given that the dataset contained only 200 total instances. The visualized data was not very clear and the formation of the clusters did not take place.
3. Apriori algorithm application was disabled in weka when selected since the dataset''s two main attributes, Annual Income and Spending Score was converted using NumerictoNominal. We also tried using the raw dataset without any pre-processing which provided us with 90% confidence level given the small dataset provided. Hence we cannot interpret any results in Apriori.
4. (J48) The results from the J48 algorithm were not reliable as there were only 42.5% of correctly classified instances when the main attribute was selected as the annual income. When the main attribute was changed to spending score, the percentage of correctly identified instances increased but the tree diagram was not formed. So we decided to go for another algorithm which would give clearer results.
5. This dataset taken from the kaggle dataset repository also had one of the purposes to interpret results for Market Basket Analysis which is the part of customer segmentation process. Since this topic hasn't been covered in our subject topics, we decided it's best not to go ahead and compute Market Basket Analysis outcome.

Comparison

Clustering method

Comparing the two clustering methods, K-means and hierarchical. Both clustering data points to customers age above 45 are unlikely to have a spending score above 75. They also both show that gender and annual income of the customer do not affect the spending score. K-means clustering split the instances into 5 clusters, predetermined by us, changing the number k will yield very different clusters. Hierarchical clustering starts from the bottom up until all the instances are in the same cluster. By interpreting the hierarchical tree, we can determine the proper number of clusters. Hierarchical clustering is a more dependable method than k-means clustering.

However, to use the clustering method, we had to ignore some attributes to get the patterns shown. Since we were not able to exploit the full data given, the clustering method is bound to not show the whole relationship between the attributes. Hence, the clustering method is not as indicative of the truth than the other tests.

Classification Method

Based on the data computed, One-R gave 78% accuracy whereas Naïve Bayes gave 73.5% accuracy. The TP rates of One-R and Naive Bayes for the category 45,500$-76,000$ is 90% and 89% respectively. Given that in both the computed results the FP rate is low, both gave calculations which are pretty decent for this dataset since it was pretty clean and had no noise points or any missing values or redundancy before pre-processing. However, since the attributes in Naïve Bayes are very much dependent on each other, the result isn't very precise. But compared to One-R the method does not rely on only one rule which makes it the better algorithm in this research.

Since we used Annual Income as the class in both the algorithms, the second category (45,500$-76,000$) showed more activity than any other categories.
Although One-R showed better accuracy than Naïve Bayes, it is unreliable since it selects only one rule according to which the data is computed. Accuracy shouldn't be the important aspect when performing machine learning algorithms on datasets rather the other values like TP, FP rates should be considered as well. Naïve Bayes is a better option as not only it handles smaller datasets, its computation time recorded is lesser compared to One-R.

Conclusion and Discussion

Clustering

The clustering method is not the best algorithm to rely on as a main model to a given dataset. Clustering methods such as K-means are heavily dependent on Euclidean distance which works well when there are a few attributes, but starts to fail as more attributes are included into the dataset. The clustering methods are also very sensitive to outliers. Having a few outliers can skew the data if not accounted for, causing a loss in information.

We believe that the clustering method can still be used as a rough estimate and point us in the right direction. The results of the clustering methods have shown that most of the customers over 45 have very low spending scores and thus the mall should heavily advertise to them in order to get them to spend more in the mall.

Classification

To summarize, One-R and Naïve Bayes showed better results than J48, which didn't comply with the given dataset. Naïve Bayes is a better choice since it is the correct approach in this project. It would be recommended to the client to focus on the Annual Income category customers between 45-500$-76,000$ to be observed and targeted for upcoming marketing strategies and future customer buildup rather than focusing on the Spending score.

This experiment gave us an in depth knowledge as to how machine learning algorithms work and how differently the results can be computed and how these said results can affect various decision making situations and projects. With the help of machine learning, any dataset is open to interpretation given that their values and variables can be experimented on with their outcomes to be very surprising and challenging. Weka software was of great assistance to us with clear visual representation in handling the tasks and experiments and making our observations.

References

Mall Customer Segmentation Data. (2018, August 11). [Dataset]. https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python

Telco Customer Churn Data. (2018, February 24). [Dataset]. https://www.kaggle.com/blastchar/telco-customer-churn

StockX Sneaker Data Contest. (2020, July 18). [Dataset]. https://www.kaggle.com/hudsonstuck/stockx-data-contest