

***A Critical Evaluation of the
Strengths of using Excel for Pre-
Processing Data. – Visual
Representation, and Evaluation of
Data Analysis using Weka***

MAY 2019

MODULE NAME:	DATA HANDLING & BUSINESS INTELLIGENCE
NAME:	HEATHER JACKSON
STUDENT NUMBER:	1704024

CONTENTS

TABLE OF CONTENTS

1. INTRODUCTION	3
2. EXCEL	4
3. WEKA.....	7
4. CONCLUSION	10
5. REFERENCES	11
6. APPENDICES	12

APPENDICES

APPENDIX A.....	IF STATEMENT
APPENDIX B	PIVOT TABLE & CHART
APPENDIX C.....	CLUSTER ANALYSIS
APPENDIX D	CLUSTER ANALYSIS
APPENDIX E.....	PREDICTION ANALYSIS
APPENDIX F.....	ASSOCIATION ANALYSIS
APPENDIX G.....	ASSOCIATION ANALYSIS

1. INTRODUCTION

This Report will be split into two parts Part 1 will look at the benefits, drawbacks, and types of spreadsheet software's such as excel in place for the pre-processing, analysis, and the visualisation of data available to an organisation. This section will have reference to, and use examples from Sarn Helen Stationary, providing insight into how, and why it would benefit an organisation such as Sarn Helen Stationary to pre-process, analyse and visualise their data in this way. Part 2 will investigate Data Mining and Machine Learning, including a brief outlining of both, what each involves, as well as how they are beneficial to an organisation such as Sarn Helen Stationary. Secondly part 2 of the report will look at three types of analysis available to Sarn Helen Stationary in the form of Association, Prediction and Cluster analysis, these will be then used on the Weka software to give real life examples each of these types of analysis, this will be carried out using the data provided for Sarn Helen Stationary.

2. PART 1 - EXCEL

2.1 Pre-Processing

Newman (2018) Defines Pre-processing as the process of ‘cleansing’ data, ridding it of irrelevant information, therefor creating a more reader friendly data set. Continuing that Pre-Processing creates a consistently formatted data set, as well as eliminating repetition, Nourani, Alami & Vousoughi, (2015) argue that Pre-Processing data is essential to any organisation undertaking data mining as, if redundant, irrelevant data is used it can significantly distract from the relevant knowledge that could have been gained through data mining. Continuing that the use of pre-processing to eliminate ‘noisy data’ helps an organisation to better understand and analyse data what is found in the future. Nourani, Alami & Vousoughi, (2015) and Newman (2018) can all agree that pre-processing is the first step in successful data mining, this process can be performed immediately after the collection of data, transforming raw data into relevant, readable data.

2.2 Analysing

Coles and Rowley (1996) highlight Scenario analysis otherwise known as the what- if? Analysis, this type of analysis allows for the user to change the value of one or more variables in a model. This type of analysis allows managers to answer questions such as, what would be the effect on profit if we were subject to a 5% material cost rise. The Excel worksheet attached that looks at the Sarn Helen stationary data uses the IF statement =IF(I4>I1640,"Profit","Loss") to analyse weather or not there was a profit or loss made for a certain transaction View Appendix A for further details.

Price (2013) highlights a second type of analysis in the form of a pivot table, a pivot table is a tool that can be used to summarise a table that contains an excessive volume of data, creating

MAIN FINDINGS

a more concise, relevant table. To help Sarn Helen Stationary identify the sum of their profit per region, this report has created a Pivot table and chart to help create an easy to read data set containing just the relevant information, view appendix B for further information.

Courter and Marquis (2002) identify another type of analysis in the form of excels Vlookup function, this lookup function provides three benefits to its user: convenience, a level of error prevention, and flexibility.

2.3 Visualising

Eaton (2018) states that due to the ever-growing volume of data organisations are collecting and attempting to analyse, it is crucial that visualisation tools are used to help create more readable data. Eaton (2018) continues that excel although once considered complex, is now the go to software for most organisations to visualise and analyse their data. Cunningham (2018) continues this point, stating that the visualisation of data allows the user to easily identify anomalies in the data, data visualisation tools come in a variety of forms, the most commonly used are: Charts, Graphs, and Trees. These type of visualisation tools also allow the user to identify pattern's and trends, then discover the exceptions to said trends, View Appendix B for visual representation of the profit per county of Sarn Helen Stationary.

2.4 Why Using Spreadsheet software for decision making would benefit Sarn Helen

Stationary

Aggarwal (2004) states that managers within all industries often face problems related to decision making, these problems are frequently related to data classification, particularly in relation to manufacturing. This statement can be translated to Sarn Helen Stationary, in terms of their manufacturing/ordering, if Sarn Helen Stationary are unaware as to what products are bringing them profit and what products are continually leaving them with a loss, they will be unable to make effective decisions in terms of what they manufacture/order. Coles and Rowley (1996) refer to spreadsheets that help the decision-making process as 'Decision support Systems' or 'DSS' for short, continuing that these DSS systems allow for predictions relating to 'what would happen if' a particular variable was to change. This then allows managers insight to what would and would not be detrimental to an organisation.

While analysis software can be beneficial to an organisation in the decision-making progress Mccollum (2017) highlights some of the drawbacks such systems have, these include: the misuse of data, poor data quality, breach of security, and inaccurate results. Continuing that while these factors could potentially be catastrophic to an organisation, if the data is managed correctly, the benefits should outweigh the drawbacks.

3. PART 2 - WEKA

Weka (2019) is a collection of algorithms used for machine learning, containing tools for the visualisation, association, clustering, prediction and regression of data. Weka (2019) is a data mining software that allows its user to gain knowledge from its collected, and pre-processed data. Weka (2019) has been used to analyse a limited amount of Sarn Helen Stationary's Data, in the form of Association, Prediction, and cluster analysis.

3.1 Data Mining and Machine Learning

Han and Kamber (2001) define data mining as the process of mining for knowledge from large amounts of data, continuing that large amounts of data contain information; However, this information must be searched for and found. Berson and Smith (1997) highlight the importance of data mining to be the ability to make quick informed decisions, using all available data. Continuing that without data mining, decision making would be dramatically slowed down, inevitably leading to poorer organisational performance. Data Mining, and Data warehousing are considered machine learning, the recognition of trends and the ability to recognise anomalies has allowed the knowledge gained from data to be increased significantly. Lewis (2018) highlights that machine learning has developed for both software and hardware at an expedient rate through-out the last decade, that this expansion has created artificial intelligence that can be used at the forefront of most industries.

3.2 Prediction

Han and Kamber (2001) and, Berton and Smith (1997) agree that predictive analysis is using historical information to predict future events. This can be used to predict trends based on weather conditions, time of year, demographic, and previous purchase history. Appendix F and G show a prediction tree for Sarn Helen Stationary based on Profit made per region of sale. It

is important to note that there is only a marginal 2% difference between the correct instances of prediction and the incorrect instances of prediction, meaning that the accuracy of this prediction tree is only 50%. It would not be recommended that Sarn Helen Stationary use this prediction tree when making decisions, and or predictions for the future. This type of prediction tree is limited in its uses, the tree works better if yes or no answers to each factor are available, the tree becomes complex and hard to understand if multiple variants are available.

3.3 Association

Han and Kamber (2001) state that association analysis is the process of ‘discovering association rules showing attribute-value conditions that occur frequently together in a given set of data’. Berson and Smith (1997) expand on this using the example that seventy percent of all records that contain variables D and C also contain variables E and F. The percentage in the previous statement is considered the confidence factor. View Appendix E for an example of Association for Sarn Helen Stationary. If an organisation can find patterns, and trends within their data then they can use the data to predict future events, such as creating a sales forecast.

3.4 Clustering

Han and Kamber (2001) highlight the crucial differences of Association and Prediction, from Clustering; Cluster analysis looks at data without consulting a ‘class-label’, stating that ‘objects are clustered or grouped based on maximizing the interclass similarity and minimizing the interclass similarity’. In simpler terms, objects that are clustered together have a large amount of similarities, while objects in different clusters have a large amount of differences. View Appendix C, and D, for a cluster analysis relating to the profit, loss, and returns, in relation to the regions Sarn Helen Stationary function within.

3.5 Why Use Data Analysis

Berry and Linoff (2000) highlight the benefits of data analytics, particularly the use of prediction association and clustering to be the ability for organisations to identify new opportunities, increase profit and keep their customers happy. Data analytics give an organisation insight that could not otherwise be gained, the ability to see the association between historic events allows an organisation to make an informed decision into whether to repeat said event. Previous successes can be replicated and improved, for example if data once analysed showed that Sarn Helen Stationary do not make a profit from selling office furniture in Dyfed then the organisation can make an informed decision to increase marketing activity within that region. With Data analysis systems allowing for data from all aspects of the organisation to be analysed, aspects including but not limited to: Transaction History, customer demographic, supplier information, and the organisations internal data. This allows for a well-rounded understanding of an organisation to be created.

Mela and Moorman (2018) highlight some of the downsides of data mining and analysis to be, the increased need of security for customer data as well as the increased cost in training for the position of a data analysis, highlighting that if the individual tasked with the job is not up to task, the knowledge gained from the data could be inaccurate leading to uninformed bad decisions.

4. CONCLUSION

To conclude, excel is still the go to software for most organisations, offering the ability to pre-process, analyse, and visualise data, this for most small to medium sized organisations is sufficient. However, if an organisation is looking for a more sophisticated data mining software allowing a deeper insight into consumer trend and patterns, while creating predictions for the organisation, a software such as Weka should be looked at. Both software allows an insight into current business performance based on historic events, permitting for an organisation to make informed decisions based on this ‘mined knowledge’. It is important to note that there are still downsides to using data mining software, the main downsides highlighted were the cost and security of data mining. Due to the limitations of this report the benefits and drawbacks of data analysis could not be looked at in enough detail, it would be recommended that future studies focus on this aspect of analysis.

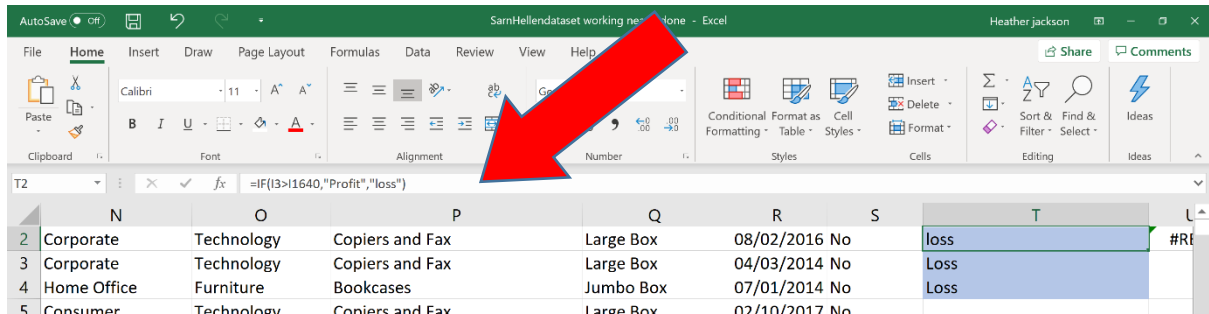
5. REFERENCES

- Aggarwal, K. (2004). An EXCEL-Based Decision-Making Tool. *Problems and Perspectives in Management*, 2(2)
- Berry, M.J.A. & Linoff, G., (2000). *Mastering data mining : the art and science of customer relationship management*, New York: Wiley Computer Publishing.
- Berson, A. and Smith, S. (1997). *Data Warehousing, Data Mining, & OLAP*. London: Mcgraw-Hill.
- Brown, Meta S. & Brown, Meta S, (2014). *Data mining for dummies*,
- Coles, S. and Rowley, J. (1996). Spreadsheet modelling for management decision making. *Industrial Management & Data Systems*, 96(7), pp.17-23.
- Cunningham, m. (2018). Using Visualization Software in the Audit of Revenue Transactions to Identify Anomalies. *Issues in Accounting Education*, 33(4), pp.33–47.
- Eaton, (2018). Data Visualization Software. *CPA Journal*, 88(6), pp.50–54.
- Han, J. and Kamber, M. (2001). *Data mining*. 1st ed. London: Accademic Press.
- Lewis, T. (2018). Learning machine learning. *Communications of the ACM*, 61(12), pp.24–28
- Mccollum, T. (2017). THE DOWNSIDE OF ANALYTICS: Data analysis technology could negatively impact internal audit independence. *Internal Auditor*, 74(5), p.13.
- Mela, C. and Moorman, C. (2018) *Why Marketing Analytics Hasn't Lived Up to Its Promise*
- Newman, D. (2018). *In A World Where Data Rules, All Companies Are Tech Companies*. [online] Forbes.com. Available at: <https://www.forbes.com/sites/danielnewman/2018/06/06/in-a-world-where-data-rules-all-companies-are-tech-companies/> [Accessed 10 May 2019].
- Nourani, Alami & Vousoughi, (2015). Wavelet-entropy data pre-processing approach for ANN-based groundwater level modeling. *Journal of Hydrology*, 524, pp.255–269.
- Price, M. (2013). *Excel 2013*. Leamington Spa, Warwickshire, U.K.: In Easy Steps.
- Weka (2019). *Weka 3 - Data Mining with Open Source Machine Learning Software in Java*. [online] Cs.waikato.ac.nz. Available at: <https://www.cs.waikato.ac.nz/ml/weka/> [Accessed 13 May 2019].

APPENDICIES

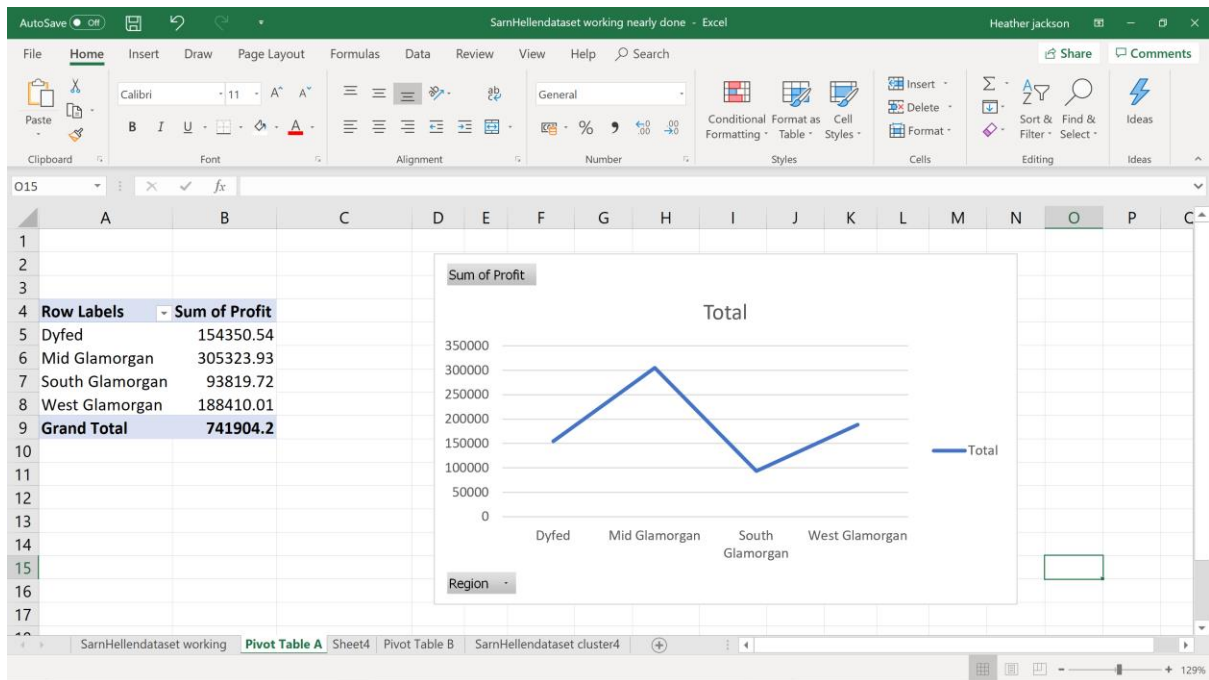
6. APPENDICIES

Appendix A

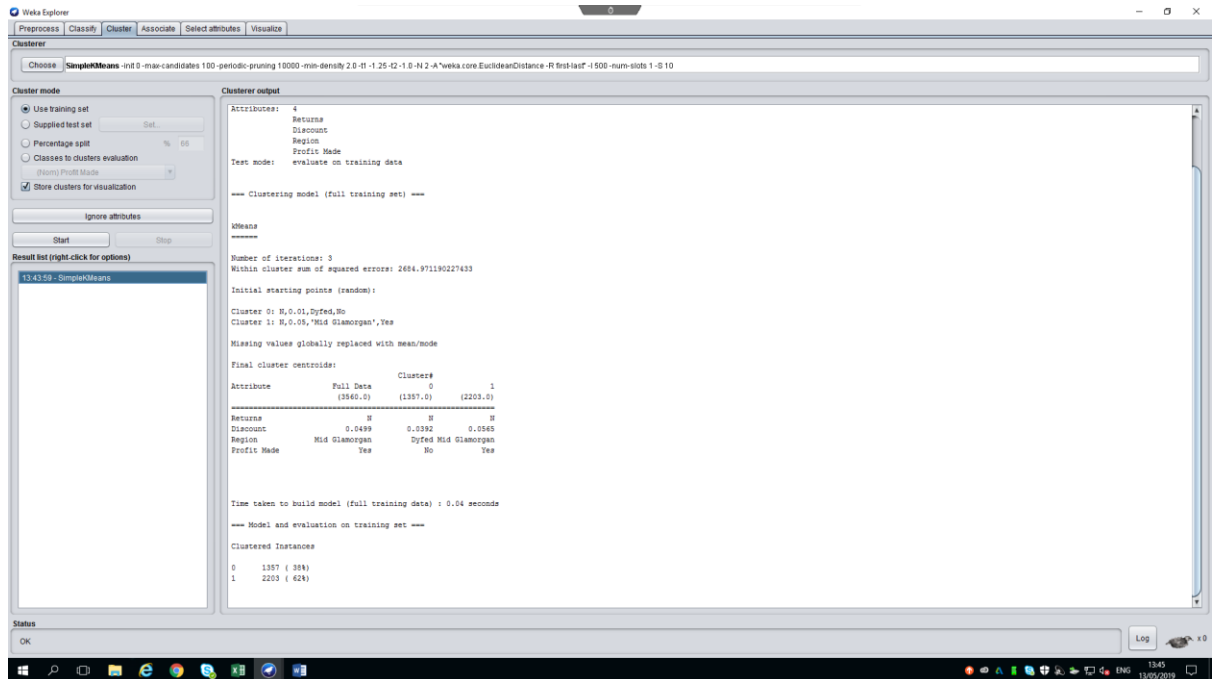


	N	O	P	Q	R	S	T	U
2	Corporate	Technology	Copiers and Fax	Large Box	08/02/2016	No	loss	#R
3	Corporate	Technology	Copiers and Fax	Large Box	04/03/2014	No	Loss	
4	Home Office	Furniture	Bookcases	Jumbo Box	07/01/2014	No	Loss	
5	Consumer	Technology	Copiers and Fax	Large Box	02/10/2017	No		

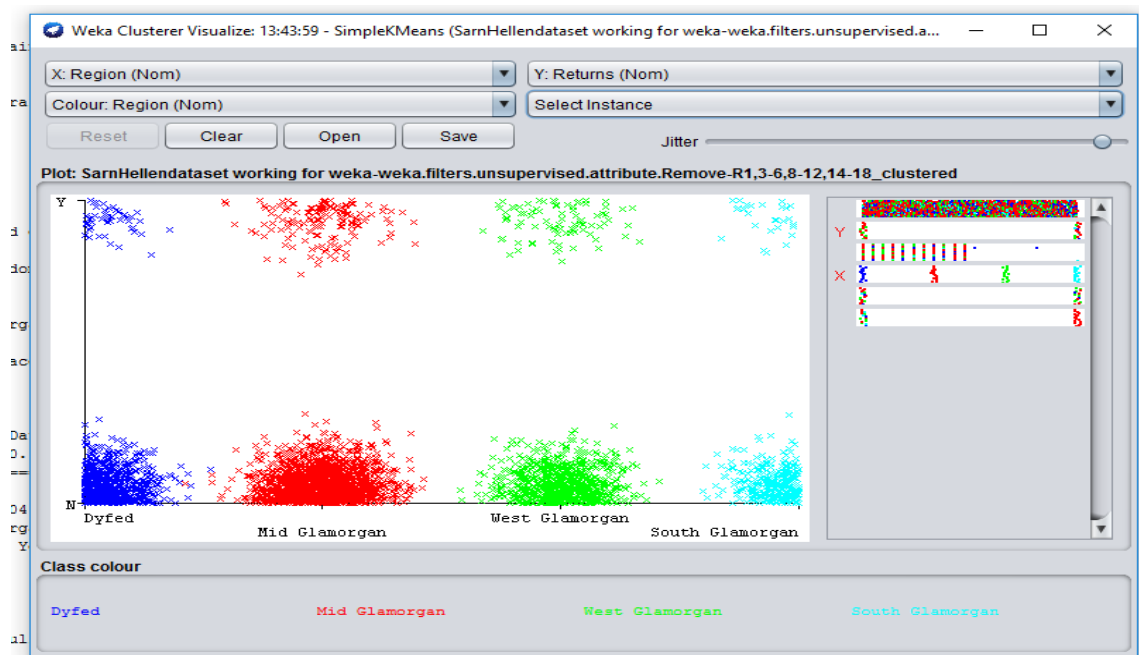
Appendix B



Appendix C



Appendix D



APPENDICIES

Appendix E

The screenshot shows the Weka Explorer interface with the 'FilteredAssociator' filter applied. The 'Result list' on the left shows three entries, with the third one selected. The 'Output' pane on the right displays the following information:

```
==== Run information ====
Scheme: weka.associations.FilteredAssociator -F "weka.filters.Multifilter -F "weka.filters.unsupervised.attribute.ReplaceMissingValues \" -c -1 -W weka.associations.Apriori -- -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S 1.0 -c -1
Relation: SarnHelledataset working for weka-weka.filters.unsupervised.attribute.Remove-R1,3-6,8-12,14-18-weka.filters.unsupervised.attribute.Remove-R2
Instances: 3460
Attributes: 3
Returns
Region
Profit Made

==== Associator model (full training set) ====

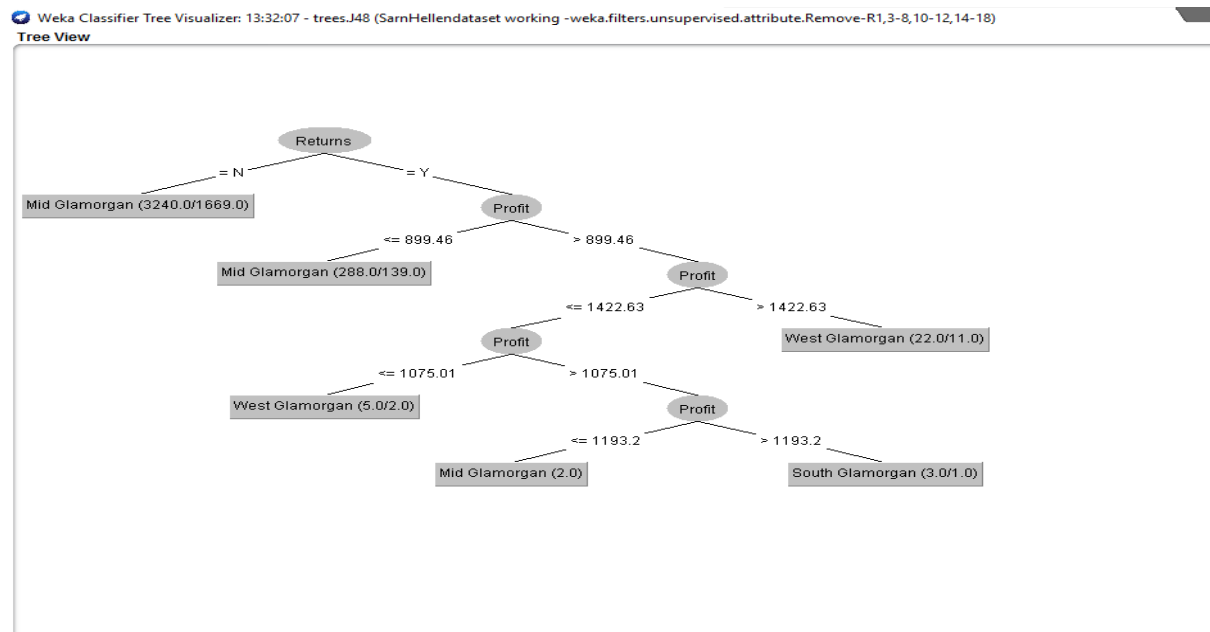
FilteredAssociator using weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S 1.0 -c -1 on data filtered through weka.filters.Multifilter -F "weka.filters.unsupervised.attribute.ReplaceMissingValues "

Filtered Header
Relation 'SarnHelledataset working for weka-weka.filters.unsupervised.attribute.Remove-R1,3-6,8-12,14-18-weka.filters.unsupervised.attribute.Remove-R2-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.Multifilter-Pve
Attribute Returns (R,Y)
Attribute Region (Dyfed,'Mid Glamorgan','West Glamorgan','South Glamorgan')
Attribute 'Profit Made' (R0,Yes)

@data

Associator Model
Apriori
=====
Minimum support: 0.1 (356 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18
Generated sets of large itemsets:
Size of set of large itemsets L(1): 6
Size of set of large itemsets L(2): 10
Size of set of large itemsets L(3): 3
Best rules found:
1. Region=Dyfed 681 ==> Returns=R 631 <conf:(0.93)> lift:(1.02) lev:(0) [11] conv:(1.2)
2. Region=Mid Glamorgan Profit Made=Yes 951 ==> Returns=R 870 <conf:(0.91)> lift:(1.01) lev:(0) [4] conv:(1.04)
3. Profit Made=Yes 1922 ==> Returns=R 1704 <conf:(0.91)> lift:(1) lev:(0) [4] conv:(1.02)
4. Region=Mid Glamorgan 1727 ==> Returns=R 1371 <conf:(0.91)> lift:(1) lev:(-0) [0] conv:(0.99)
```

Appendix F



Appendix G

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose J48 C 0.25-M 2

Test options

☐ Use training set

☐ Supplied test set

☒ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Item) Region

Start Stop

Result list (right-click for options)

13.32.07 - best_J48

Classifier output

```
Recurrence - Y
| Profit <= 899.46: Mid Glamorgan (288.0/139.0)
| Profit > 899.46
| | Profit <= 1422.63
| | | Profit <= 1075.01: West Glamorgan (5.0/2.0)
| | | Profit > 1075.01
| | | Profit <= 1189.2: Mid Glamorgan (2.0)
| | | Profit > 1189.2: South Glamorgan (3.0/1.0)
| | Profit > 1422.63: West Glamorgan (22.0/11.0)
```

Number of Leaves : 6

Size of the tree : 11

Time taken to build model: 0.03 seconds

==== Stratified cross-validation ====

==== Summary ====

Correctly Classified Instances	1725	48.4551 %
Incorrectly Classified Instances	1835	51.5449 %
Kappa statistic	-0.0007	
Mean absolute error	0.3329	
Root mean squared error	0.4081	
Relative absolute error	99.9797 %	
Root relative squared error	100.0279 %	
Total Number of Instances	3560	

==== Detailed Accuracy By Class ====

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.000	0.000	?	0.000	?	?	0.502	0.192	Dyfed
	0.999	0.999	0.485	0.999	0.653	-0.511	0.495	0.482	Mid Glamorgan
	0.000	0.001	0.000	0.000	0.000	-0.016	0.505	0.234	West Glamorgan
	0.000	0.000	?	0.000	?	?	0.497	0.091	South Glamorgan
Weighted Avg.	0.485	0.493	?	0.485	?	?	0.499	0.333	

==== Confusion Matrix ====

a	b	c	d	<-- Classified as
0	491	0	0	a = Dyfed
0	1725	2	0	b = Mid Glamorgan
0	826	0	0	c = West Glamorgan
0	325	1	0	d = South Glamorgan

Status

OK

Log

13.32 13/05/2019