Practical Report

January 10, 2021

Marco Ligia
Technological University Dublin
Ireland
B00139079@mydublin.ie

Part 1: Classification

Introduction Adult Mixed Dataset

This dataset consists of data that was taken from the Census bureau database by Ronny Kohavi and Barry Becker in 1994. The dataset will be used to perform a classification task in order to be able to identify if an individual earns more than $50K a year.

Dataset Description

Attribute Name	Data Type	Description
age	Numeric	Individuals Age
workclass	Nominal	Individuals work category
fnlwgt	Numeric	Final weight
education	Nominal	Individual's highest education degree
education_num	Numeric	The number assigned to Individuals level of education
marital_status	Nominal	Current Marital Status of the Individual
occupation	Nominal	Current Occupation of the Individual
relationship	Nominal	Individual's relation in a family
race	Nominal	Race of Individual
sex	Nominal	Sex of Individual
capital_gain	Numeric	Accumulated Capital Gain
capital_loss	Numeric	Accumulated Capital Loss
hours_per_week	Numeric	Number of hours per week individual work
native_country	Nominal	Individuals Native Country
label	Nominal	Does individual earn less or above $50k

Summary Statistics

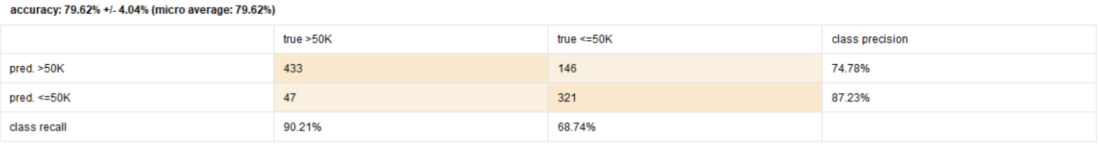
```
      age                  workclass        fnlwgt              education     education_num  
 Min.   :17   Federal-gov     : 27   Min.   : 19302   HS-grad     :257   Min.   : 2.00  
 1st Qu.:31   Local-gov       : 59   1st Qu.:113765   Some-college:211   1st Qu.: 9.00  
 Median :39   Private         :697   Median :178470   Bachelors   :187   Median :10.00  
 Mean   :40   Self-emp-inc    : 43   Mean   :189733   Masters     : 71   Mean   :10.58  
 3rd Qu.:48   Self-emp-not-inc: 86   3rd Qu.:236451   Assoc-voc   : 43   3rd Qu.:13.00  
 Max.   :90   State-gov       : 35   Max.   :750972   11th        : 36   Max.   :16.00  
                                                      (Other)     :142                  
              marital_status            occupation            relationship
 Divorced             : 94    Exec-managerial:166   Husband       :507  
 Married-AF-spouse    :  1    Prof-specialty :165   Not-in-family :189  
 Married-civ-spouse   :569    Sales          :125   Other-relative: 18  
 Married-spouse-absent: 10    Craft-repair   :105   Own-child     :110  
 Never-married        :237    Adm-clerical   : 97   Unmarried     : 65  
 Separated            : 16    Other-service  : 83   Wife          : 58  
 Widowed              : 20    (Other)        :206                       
                 race          sex        capital_gain     capital_loss  hours_per_week
 Amer-Indian-Eskimo:  5   Female:248   Min.   :    0   Min.   :   0   Min.   : 2.0  
 Asian-Pac-Islander: 29   Male  :699   1st Qu.:    0   1st Qu.:   0   1st Qu.:40.0  
 Black             : 69                Median :    0   Median :   0   Median :40.0  
 Other             :  5                Mean   : 1982   Mean   : 136   Mean   :42.1  
 White             :839                3rd Qu.:    0   3rd Qu.:   0   3rd Qu.:50.0  
                                       Max.   :99999   Max.   :2559   Max.   :99.0  

       native_country   label    
 United-States:859    <=50K:467  
 Mexico       : 11    >50K :480  
 Philippines  :  7               
 Puerto-Rico  :  7               
 Canada       :  6               
 China        :  6               
 (Other)      : 51               
```

Decision Tree

We chose to use the Decision tree model to the dataset for its flexibility in dealing with various data types and its robustness.

Results of training using default parameters.

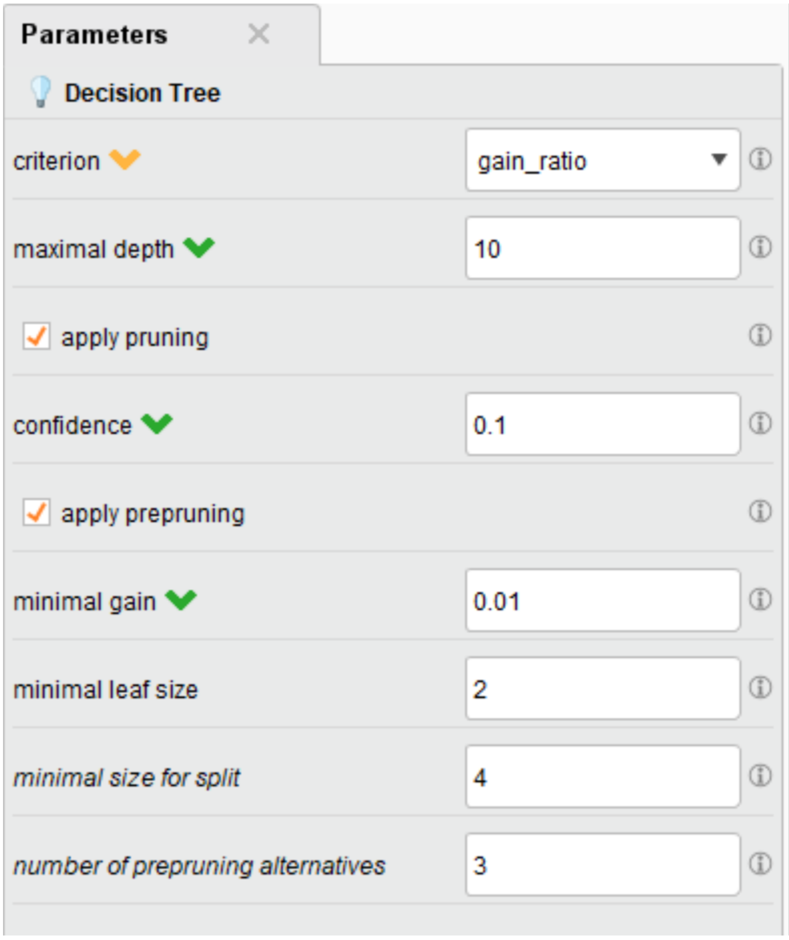


accuracy: 79.62% +/- 4.04% (micro average: 79.62%)

	true >50K	true <=50K	class precision
pred. >50K	433	146	74.78%
pred. <=50K	47	321	87.23%
class recall	90.21%	68.74%	

Figure 1 - Default Decision Tree Performance

Using the default parameters for the Decision Tree model, we achieved an accuracy of 79.62%. The accuracy was relatively strong using the default parameters, but some of the pruning parameters may be restricting the model from picking up some more detailed patterns that improve the overall accuracy.



Parameters

Decision Tree

criterion: gain_ratio

maximal depth: 10

☑ apply pruning

confidence: 0.1

☑ apply prepruning

minimal gain: 0.01

minimal leaf size: 2

minimal size for split: 4

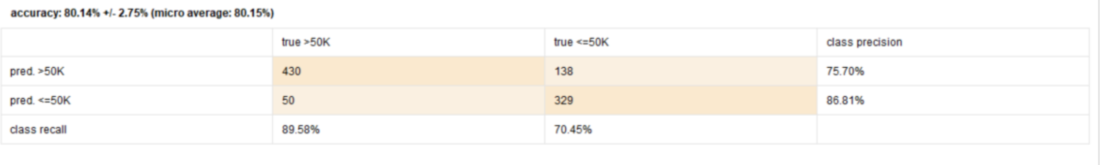
number of prepruning alternatives: 3

Figure 2 - Default Decision Tree Parameters

Figure 3 - Default Decision Tree

The default parameters can be seen in Figure 2 above, and the default decision tree can also be seen in Figure 3.

The first parameter we decided to tune was concerning pruning. We decided to disable pruning and pre-pruning of the model. We took this decision to see was pruning affecting the model from finding useful patterns in the data as the tree grew unrestricted. We also understood that not pruning the model could lead to overfitting.



accuracy: 80.14% +/- 2.75% (micro average: 80.15%)

	true >50K	true <=50K	class precision
pred. >50K	430	138	75.70%
pred. <=50K	50	329	86.81%
class recall	89.58%	70.45%	

Figure 4 - Non-Pruned Decision Tree Performance

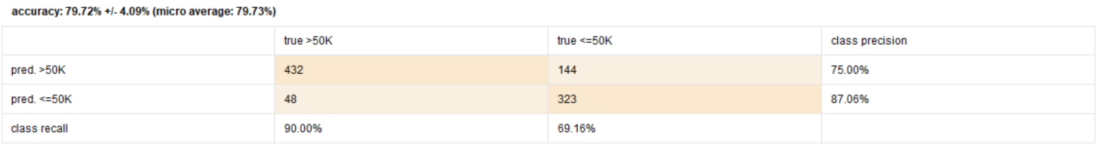
Figure 5 - Non-Pruned Decision Tree

As can be seen, the model's accuracy grew by 0.52%. to 80.14%, when pruning was not applied to the model. Additionally, the resulting Decision Tree is noticeably bigger.

Next, we decided to see the effect on the model accuracy of reducing the maximum depth of the Decision tree. We decided to test out maximum depths of 2/4/6 and 8 to see their effect.

When the maximum depth was set to 2, model accuracy dropped to 59.24%. When we set the depth to 4, it was down again to 62.41%. When the maximum depth was set to 6, the accuracy was 74.15%, and when depth was set to 8, it was down 0.53% from the default model accuracy, at 79.09%.

We then decided to set the confidence threshold for pruning to 0.4 from 0.1. This looks at how confident the model is of a pruning decision it is going to make based on a pessimistic error pruning calculation. So the default model needed a 10% confidence level in the calculation to apply pruning, whereas now it has to have a confidence of 40%. This confidence level was chosen, as it is in the range that confidence is set to by most users in Rapidminer.



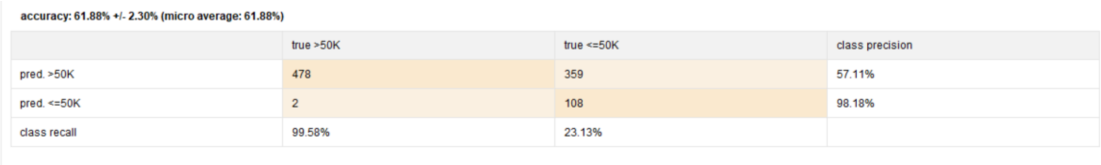
accuracy: 79.72% +/- 4.09% (micro average: 79.73%)

	true >50K	true <=50K	class precision
pred. >50K	432	144	75.00%
pred. <=50K	48	323	87.06%
class recall	90.00%	69.16%	

Figure 6 - Decision Tree Performance with a confidence level of 0.4.

As a result of the increase in pruning's confidence parameter, there was a slight increase in accuracy of 0.1% to 79.72%.

Next, we decided to adjust the minimal gain parameter for pre-pruning to be 0.2 from 0.01 to see its impact on accuracy. We wanted to test this as it would restrict the model's growth and indicate if the data did not need a big tree to be modeled accurately. This is achieved as it will prevent splits that do not achieve a 20% improvement in the gain ratio or what criterion is being used. We settled on the 0.2 setting as this is the range the majority of Decision trees are modeled on in Rapidminer.



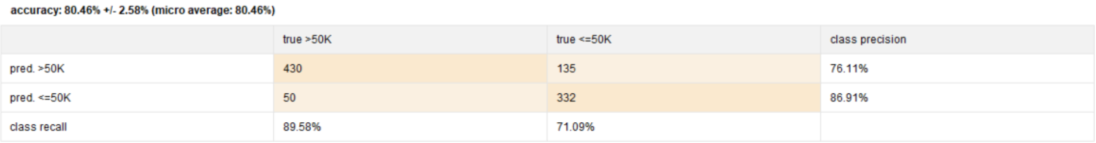
accuracy: 61.88% +/- 2.30% (micro average: 61.88%)

	true >50K	true <=50K	class precision
pred. >50K	478	359	57.11%
pred. <=50K	2	108	98.18%
class recall	99.58%	23.13%	

Figure 7 - Decision Tree with a minimal gain of 0.2

This adjustment resulted in a large decrease in an overall accuracy of just over 17% to 61.88%

The final adjustment we made was to reduce minimal leaf size from 2 to 1; this means that each leaf in the tree only needed one record to be in that criteria for a leaf to be formed, so left more detail in the model and did not require a more generic assignment of records.



accuracy: 80.46% +/- 2.58% (micro average: 80.46%)

	true >50K	true <=50K	class precision
pred. >50K	430	135	76.11%
pred. <=50K	50	332	86.91%
class recall	89.58%	71.09%	

Figure 8 - Decision Tree Min Leaf size reduced to 1

The minimum leaf size adjustment to 1 resulted in an increase in accuracy of 0.84% to 80.46%.

Patterns in the data

Based on the parameters tuned in our study, we would feel that a larger tree best models the dataset as the dataset would appear to have detailed patterns in the data. When parameters were tuned to restrict the growth of the tree, such as increasing the minimal gain, decreasing maximum depth, or having pruning and pre pruning performed, the model did not achieve as high accuracy. By not using pruning, the tree could grow large and more splits were being created. Additionally, increasing the confidence needed for pruning restricted the amount of pruning of the tree.

The individual's age told a lot when it came to classifying whether or not a person earned $50k or not. If a person had capital gains or not was the next most telling indicator, followed by the amount of capital losses accumulated, the individual's education number, their relationship in the household, and how many hours they worked.

The characteristics of an individual who earned over $50k based on the decision tree were:

- Over 24.5 and had capital gains over $6,897.50
- Under 72, with capital losses less than 2,322
- Under 72, unmarried, and working over 58 hours a week.
- Under 72, a husband and working over 14 hours a week.
- Under 72 and older than 28, not in a family with an education number above 13.5 and working over 28 hours a week

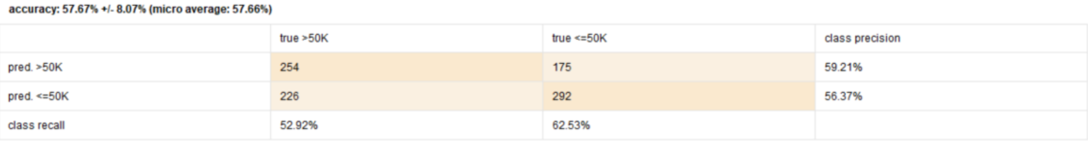
From analyzing the data's scatter plot, we could see that those earning over $50k had:

- A Masters, Doctorate or had attended a professional school.
- Over 30 years old
- Were Married
- Worked over 40 hours a week
- Had Capital gains over $5,178.
- Tend to be male.

K-Nearest Neighbor

Next, we decided to use K Nearest Neighbour (KNN) to predict who earned over $50k. We chose KNN due to its simplicity to implement and that it makes no assumptions.

Results of training using default parameters.



accuracy: 57.67% +/- 8.07% (micro average: 57.66%)

	true >50K	true <=50K	class precision
pred. >50K	254	175	59.21%
pred. <=50K	226	292	56.37%
class recall	52.92%	62.53%	

Figure 9 - KNN default results

Using the default parameters of K=5, Weighted voting being used, measure type being mixed measures, and mixed measure being mixed Euclidean distance, the default model achieved 57.67% accuracy. This low accuracy may be due to :

- Outliers in the data, which can affect KNN as all the information is obtained from the input and KNN does not generalize the data
- If there are irrelevant features in the dataset, KNN can suffer as it gives all features the same importance.

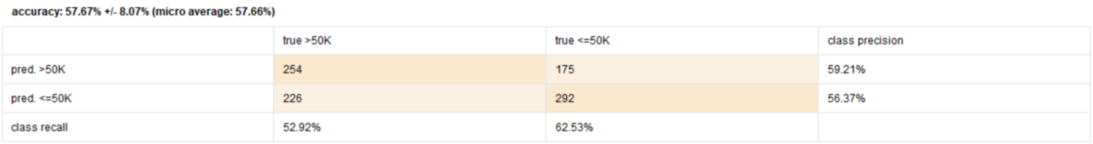
Results when training using modified parameter settings.

We first chose to tune K= 9 and k=15, to see the effect on accuracy. We chose to use 9 as opposed to 10 as it is recommended when trying to classify into two labels, an odd number is used for k to avoid situations where there is an equal amount of nearest neighbors for both.

K	Accuracy
9	56.71%
15	54.17%

As shown in the results above, as K increased, accuracy steadily decreased to 56.71& when k=9, and 54.17% when k=15.

Next decided to try KNN without weighted voting enabled.



accuracy: 57.67% +/- 8.07% (micro average: 57.66%)

	true >50K	true <=50K	class precision
pred. >50K	254	175	59.21%
pred. <=50K	226	292	56.37%
class recall	52.92%	62.53%	

Figure 10 - KNN with no weighted voting.

Turning off weighted voting resulted in identical results to the default model, meaning it has no impact on the classification power of KNN in this instance.

Patterns in the data

KNN would not be a robust model for classification on this dataset as it achieved low accuracy. As K achieved its optimal results when K =5 as opposed to a larger number, this would suggest that the records are unique overall and that there are more detailed patterns in the dataset. We would assume this as when K=9 or K=15, it would have been generically assigning records to its nearest neighbor, but the nearest neighbors may have been quite distant from the record it was trying to assign due to the higher number of K.

Random Forest

We decided to use Random forest to classify the Adult numeric dataset as it is an overall highly accurate model that merges multiple trees giving more predictive and stable results. The randomness of the trees may help to pick up on detailed patterns in the data.

Introduction Adult Numeric Dataset

This dataset consists of data that was taken from the Census bureau database by Ronny Kohavi and Barry Becker in 1994. The dataset will be used to perform a classification task in order to be to identify if an individual earns over $50K a year. The dataset has had nominal attributes removed.

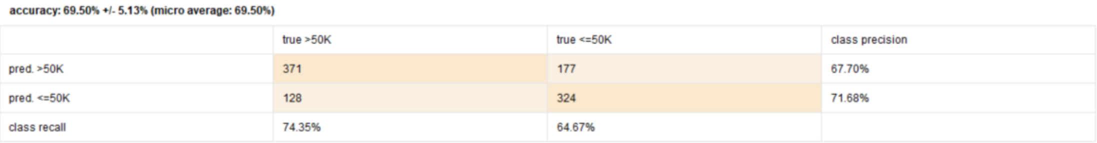
Dataset Description

Attribute Name	Data Type	Description
age	Numeric	Individuals Age
fnlwgt	Numeric	Final weight
education_num	Numeric	Number assigned to Individuals level of education
capital_gain	Numeric	Accumulated Capital Gain
capital_loss	Numeric	Accumulated Capital Loss
hours_per_week	Numeric	Number of hours per week individual work
label	Nominal	Does individual earn less or over $50k

Summary Statistics

```
      age             fnlwgt          education_num       capital_gain       capital_loss
 Min.   :-1.8112   Min.   :-1.5843   Min.   :-3.1746   Min.   :-0.2001   Min.   :-0.2676
 1st Qu.:-0.7136   1st Qu.:-0.6816   1st Qu.:-0.5737   1st Qu.:-0.2001   1st Qu.:-0.2676
 Median :-0.0864   Median :-0.1023   Median :-0.2021   Median :-0.2001   Median :-0.2676
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.: 0.6192   3rd Qu.: 0.4109   3rd Qu.: 0.9126   3rd Qu.:-0.2001   3rd Qu.:-0.2676
 Max.   : 3.9121   Max.   : 5.2416   Max.   : 2.0272   Max.   : 8.3407   Max.   : 4.9043
 hours_per_week      label
 Min.   :-3.3606   <=50K:501
 1st Qu.:-0.1513   >50K :499
 Median :-0.1513
 Mean   : 0.0000
 3rd Qu.: 0.6933
 Max.   : 4.8317
```

Results of training using default parameters.



accuracy: 69.50% +/- 5.13% (micro average: 69.50%)

	true >50K	true <=50K	class precision
pred. >50K	371	177	67.70%
pred. <=50K	128	324	71.68%
class recall	74.35%	64.67%	

Figure 11 - Random Forest Default Performance

When the default parameters were used with Random Forest, it achieved 69.5% accuracy on the dataset.

The default parameters had number of trees = 100/ Criterion was Gain Ratio/ Maximal Depth was 10/ The Subset ratio as guessed/Voting Strategy was Confidence Vote.

Results when training using modified parameter settings.

We first decided to test the effect on classification when we increased the number of trees produced by the random forest. We decided to test 175 and 250 trees being produced.

Number of Trees	Accuracy
175	71%
250	71.6%

As shown in the above results, the accuracy improved as the number of trees being produced increased. When the number of trees was 175, accuracy grew 1.5% to 71%, and when it was increased a further 75 trees, it had an increase of accuracy of 0.6% to 71.6%.

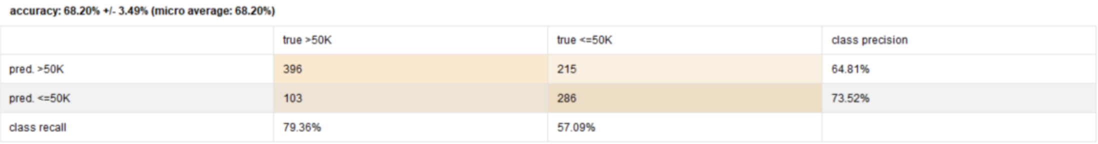
We then decided to increase the maximum depth of trees from 10 to 20, 30, and 40, in order to see do growing larger trees increase the classification of records.

Maximum Depth	Accuracy
20	72.40%
30	73.60%

40	74%

As shown in the above table, the larger the tree was allowed to grow, the better the accuracy achieved. When the maximum depth was 20, accuracy grew 2.9% from the default model to 72.4%. When maximum depth was 30, accuracy grew 4.1% from the default model to 73.6%, and when maximum depth was set to 40, accuracy grew 4.5% to 74%.

We then decided to see the effect of using majority voting for deciding the classification of records as opposed to confidence voting.



accuracy: 68.20% +/- 3.49% (micro average: 68.20%)

	true >50K	true <=50K	class precision
pred. >50K	396	215	64.81%
pred. <=50K	103	286	73.52%
class recall	79.36%	57.09%	

Figure 12- Random Forest Majority Voting

When majority voting was used, the accuracy dropped 1.3% to 68.2%.

Patterns in the data

Based on the above results from tuning the Random Forest Algorithm parameters, we would feel that there are quite detailed patterns in the dataset and so a larger tree would be best suited to modeling the classification algorithm. This is based on the fact that as we increased the tree's maximum depth, the accuracy grew noticeably. Additionally, when we grew a larger number of trees in the random forest, the results similarly improved. Both these would lead us to believe that there are some intricate patterns in the dataset, and due to the randomness of the trees being produced and the depth, these patterns were found and may not have been had the depth or number of trees been reduced. A smaller depth and amount of trees producing optimal results would indicate that the patterns in the data were generic; however, this was not the case. Also, since the majority voting mechanism resulted in inferior results to the confidence voting, it would indicate that patterns are also more detailed than generic in the dataset, as the majority voting would represent more summary patterns,

The most useful attributes in the dataset were Education Number, Age, and Capital Gains.

The characteristics of an individual who earned over $50k based on the decision tree were:

- An age over 2.51
- Capital Gain over 0.389
- A positive education number.

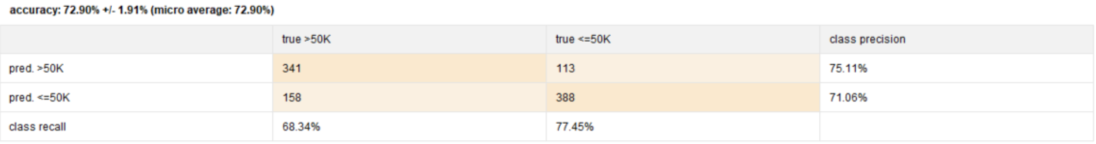
From analyzing the data's scatter plot, we could see that those earning over $50k had:

- An Education Number greater than -1, with most of those exceeding 1 earning over $50k.
- Age did not have as much of an impact, but no one with a score below 1.18 earned over $50k.
- Anyone with capital gains over 0.355 earned over $50k.
- Tended to a capital loss over 3.46.

Neural Networks

We chose to use Neural Networks for their ability to deal with complicated relationships between variables.

Results of training using default parameters.



accuracy: 72.90% +/- 1.91% (micro average: 72.90%)

	true >50K	true <=50K	class precision
pred. >50K	341	113	75.11%
pred. <=50K	158	388	71.06%
class recall	68.34%	77.45%	

Figure 13 - Neural Network Default Model Performance

The default model achieved an accuracy of 72.9%. The default parameters were:

Training Cycles = 200 Learning Rate = 0.01 Momentum = 0.9 Decay = Disabled

Normalize = Enabled Shuffle = Enabled Error epsilon = 1.0E-4

Results when training using modified parameter settings.

We first decided to increase the Training Cycles of the model. We hoped that as a result of increasing training of the model it would result in better classification of records. We decided to increase training cycles to 300, 400, and 500, respectively.

Number of Training Cycles	Accuracy
300	73.2%
400	74%
500	74.1%

As can be seen from the results, each increase in the number of training cycles increased accuracy, but accuracy improvement peaked at 400 cycles, where it achieved an accuracy of 74%, before improving slightly to 74.1% when there were 500 cycles.

We then decided to adjust the learning rate for the model. We chose to adjust the learning rate, as it impacts the weights being assigned to the model in response to estimated errors. So we hoped that by increasing these weights, the classification power of the model would improve in each run. We decided to change the learning rate from 0.01 to 0.02, 0.03, and 0.04.

Learning Rate	Accuracy
0.02	73%
0.03	73.3%
0.04	74.3%

As a result of increasing the learning rate, we saw improvements each time for each 0.01 increase. When the learning rate was increased to 0.02, the accuracy improved marginally to

73%, a 0.1% improvement. When the rate was increased further to 0.03, the accuracy improved to 73.3%, a 0.4% increase from the default model. We saw the most significant improvement in accuracy when we increased the learning rate to 0.4%. This resulted in an increase of accuracy to 74.3%, an increase of 1.4% overall.

We then decided to vary the decay and shuffle parameters. The shuffle parameter shuffles the data before learning commences. The decay parameter reduces the learning rate during the learning phase of the model.

Parameter	Accuracy
Shuffle - Disabled	55.4%
Decay - Enabled	70.3%

As expected, the disabling of the shuffle parameter greatly affected the accuracy of the model. When the shuffle parameter is enabled, it reduces variance and ensures the model does not overfit as it ensures each learning batch is representative of the overall dataset and not just a subset of the data.

By enabling the decay parameter, the performance of the model also suffered but not significantly. The model accuracy fell to 70.3%. This can be blamed on the reduced learning rate as the model went through the learning phase, meaning the rates applied to estimated errors fell as the model grew.

Patterns in the data

As shown in the above results, the dataset would appear to have detailed as opposed to generic patterns and so benefitted from having additional training cycles, as it could uncover hidden relationships in the data. Also, increasing the learning rate of the model and penalizing errors at a higher cost improved the accuracy of the model, which would support that the dataset is best by an increased amount of training cycles and an increased learning rate. The significant drop in accuracy, as a result of the shuffle parameter being disabled, supports the idea that there are detailed patterns in the dataset, as while this parameter was disabled, the model was only modeled on a subset of the data. This would show that the records in the dataset are not all generic.

Part 2: Regression

Introduction

The dataset that was used for the regression was a skeletal dataset taken from the UCI Repository.

Dataset Description

Attribute Name	Data Type	Description
biacromial	Numeric	The horizontal distance across the shoulders measured between the acromia
pelvicBreath	Numeric	The broadest measure of the pelvis between the outer edges of the upper iliac bones
bitrochanteric	Numeric	Pertains to the two trochanters
chestDepth	Numeric	The maximum horizontal distance from the vertical reference plane to the front of the chest in men or breast in women
chestDiam	Numeric	Length of the body from the neck to the abdomen
elbowDiam	Numeric	Length of the visible joint between the upper and lower parts of the arm
wristDiam	Numeric	Length of the joint that bridges the hand to the forearm
kneeDiam	Numeric	Length of the joint from femur to the tibia
ankleDiam	Numeric	Length of the region where the foot and the leg meet

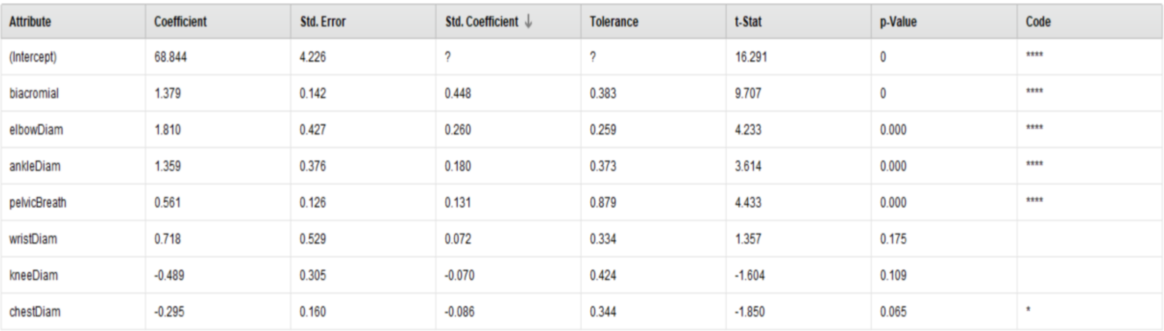
Summary Statistics

```
  biacromial     pelvicBreath   bitrochanteric    chestDepth
Min.   :32.40   Min.   :18.70   Min.   :24.70   Min.   :14.30
1st Qu.:36.20   1st Qu.:26.50   1st Qu.:30.60   1st Qu.:17.30
Median :38.70   Median :28.00   Median :32.00   Median :19.00
Mean   :38.81   Mean   :27.83   Mean   :31.98   Mean   :19.23
3rd Qu.:41.15   3rd Qu.:29.25   3rd Qu.:33.35   3rd Qu.:20.90
Max.   :47.40   Max.   :34.70   Max.   :38.00   Max.   :27.50
  chestDiam       elbowDiam       wristDiam        kneeDiam
Min.   :22.20   Min.   : 9.90   Min.   : 8.10   Min.   :15.70
1st Qu.:25.65   1st Qu.:12.40   1st Qu.: 9.80   1st Qu.:17.90
Median :27.80   Median :13.30   Median :10.50   Median :18.70
Mean   :27.97   Mean   :13.39   Mean   :10.54   Mean   :18.81
3rd Qu.:29.95   3rd Qu.:14.40   3rd Qu.:11.20   3rd Qu.:19.60
Max.   :35.60   Max.   :16.70   Max.   :13.30   Max.   :24.30
  ankleDiam          age            weight           height
Min.   : 9.90   Min.   :18.00   Min.   : 42.00   Min.   :147.2
1st Qu.:13.00   1st Qu.:23.00   1st Qu.: 58.40   1st Qu.:163.8
Median :13.80   Median :27.00   Median : 68.20   Median :170.3
Mean   :13.86   Mean   :30.18   Mean   : 69.15   Mean   :171.1
3rd Qu.:14.80   3rd Qu.:36.00   3rd Qu.: 78.85   3rd Qu.:177.8
Max.   :17.20   Max.   :67.00   Max.   :116.40   Max.   :198.1
```

Linear Regression

We used the Linear Regression model in order to perform our regression of the data. We decided to set the label as the Height attribute.

Results of Linear Regression.

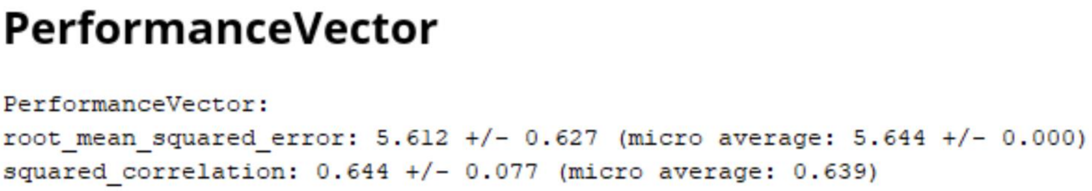


Attribute	Coefficient	Std. Error	Std. Coefficient ↓	Tolerance	t-Stat	p-Value	Code
(Intercept)	68.844	4.226	?	?	16.291	0	****
biacromial	1.379	0.142	0.448	0.383	9.707	0	****
elbowDiam	1.810	0.427	0.260	0.259	4.233	0.000	****
ankleDiam	1.359	0.376	0.180	0.373	3.614	0.000	****
pelvicBreath	0.561	0.126	0.131	0.879	4.433	0.000	****
wristDiam	0.718	0.529	0.072	0.334	1.357	0.175	
kneeDiam	-0.489	0.305	-0.070	0.424	-1.604	0.109	
chestDiam	-0.295	0.160	-0.086	0.344	-1.850	0.065	*

Figure 14 - Linear Regression Output

Based on the Linear Regression model's output, the Biacromial attribute has the most importance in predicting the height of each record. The Std Coefficient for the biacromial attribute is 0.448, indicates the number of standard deviations a dependent variable will change, per standard deviation increase in the predictor variable. So the higher the std coefficient, the higher importance it has on the dependent variable, in this case height. EelbowDiam is the next most influential attribute as it has a std coefficient score of 0.260, followed by ankleDiam with a score of 0.180. PelvicBreath has a std coefficient of 0.131.

Further reinforcing the above attributes as having the most significant importance on height are the p-Values for each attribute. The P-Value is the evidence against a null hypothesis. The smaller the p-value, the more substantial the evidence that we should reject the null hypothesis. A p-value of less than 0.05 (typically ≤ 0.05) is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct. A p-value higher than 0.05 (> 0.05) is not statistically significant and indicates strong evidence for the null hypothesis. This means we retain the null hypothesis and reject the alternative hypothesis. As shown in Figure 14 above, biacromial, elbowDiam, ankleDiam, and pelvicBreath all have p scores of 0, meaning that these attributes are statistically significant in predicting the height.



PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 5.612 +/- 0.627 (micro average: 5.644 +/- 0.000)
squared_correlation: 0.644 +/- 0.077 (micro average: 0.639)
```

Figure 15 - Performance Metrics of the Linear Regression Model.

The RMSE shows “ the standard deviation of the residuals (prediction errors). Residuals measure how far from the regression line data points are; RMSE is a measure of how to spread out these residuals are. In other words, it tells us how concentrated the data is around the line of best fit.” The above RMSE of 5.612 indicates that the model would indicate the height of a record =/- 5.612. Considering the domain and the attribute to be predicted, this is quite a good score as it is not a large gap between predicted and actual height.

The squared correlation score of 0.644 indicates that the seven attributes used in the Linear Regression model account for 64.4% of the variance. This is not satisfactory.

Part 3: Clustering

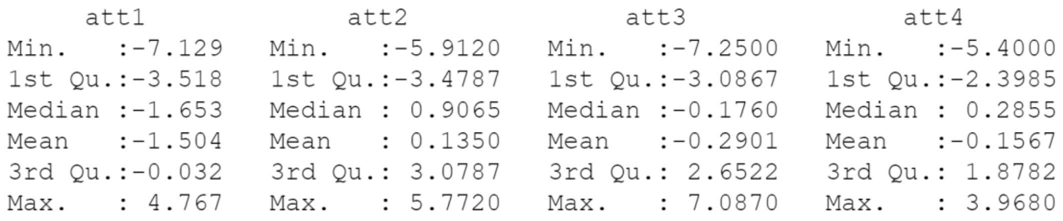
Introduction

The Cluster dataset was generated using the 'generate data' operator in Rapidminer. It contains many Gaussian clusters that will need to be identified in this section. It contains 4 numerical attributes.

Dataset Description

Attribute Name	Data Type
Att1	Numeric
Att2	Numeric
Att3	Numeric
Att4	Numeric

Summary Statistics



```
      att1              att2              att3              att4        
 Min.   :-7.129   Min.   :-5.9120   Min.   :-7.2500   Min.   :-5.4000  
 1st Qu.:-3.518   1st Qu.:-3.4787   1st Qu.:-3.0867   1st Qu.:-2.3985  
 Median :-1.653   Median : 0.9065   Median :-0.1760   Median : 0.2855  
 Mean   :-1.504   Mean   : 0.1350   Mean   :-0.2901   Mean   :-0.1567  
 3rd Qu.:-0.032   3rd Qu.: 3.0787   3rd Qu.: 2.6522   3rd Qu.: 1.8782  
 Max.   : 4.767   Max.   : 5.7720   Max.   : 7.0870   Max.   : 3.9680  
```

Figure 16 - Summary Statistics of Cluster Dataset

K Means Clustering

We decided to use K-Means to perform the clustering of the dataset due to its ease of implementation.

Overview

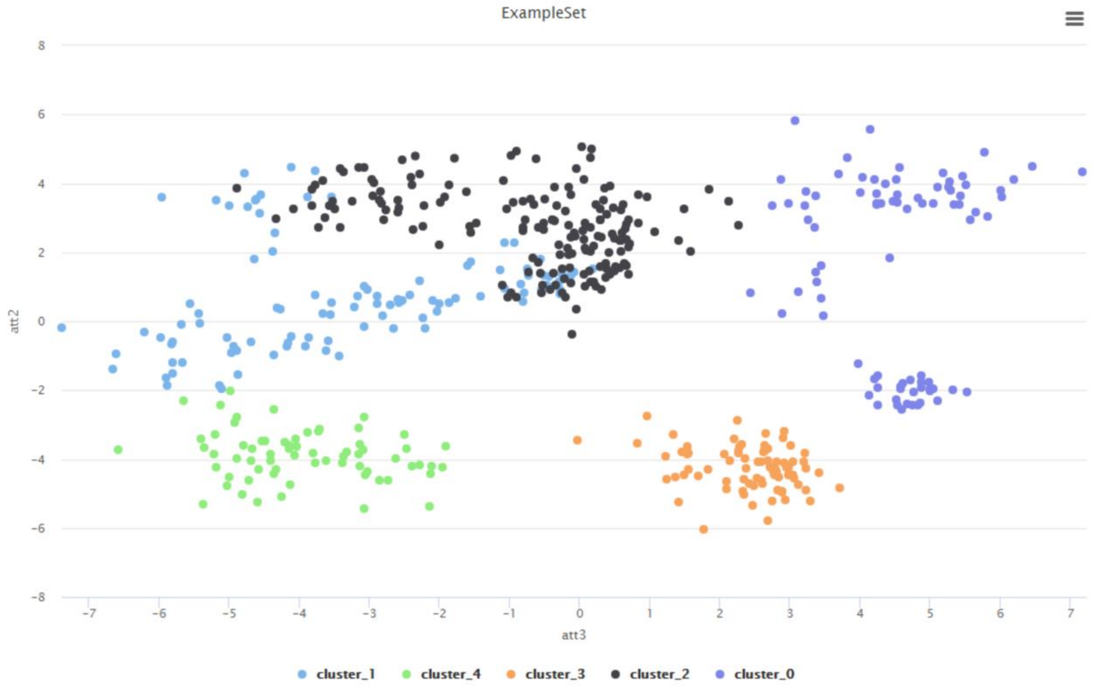


Figure 17 - K Means Default Scatter Plot - K=5

As can be seen from the above scatter plot of att3 on the X-Axis and att2 on the Y-Axis, when the default settings are used for the K Means algorithm and K=5, the clustering of the records is poor. Cluster 3 is relatively well clustered with a reasonable inter-cluster distance; however, the intracluster distance is not as good, as there are several records assigned to the cluster, skewing slightly to the right. The remaining clusters have both short inter-cluster and intra-cluster distance as they are not compact, and there is no division between the other 4 clusters except for cluster 0 that has a minimal distance from the other clusters.

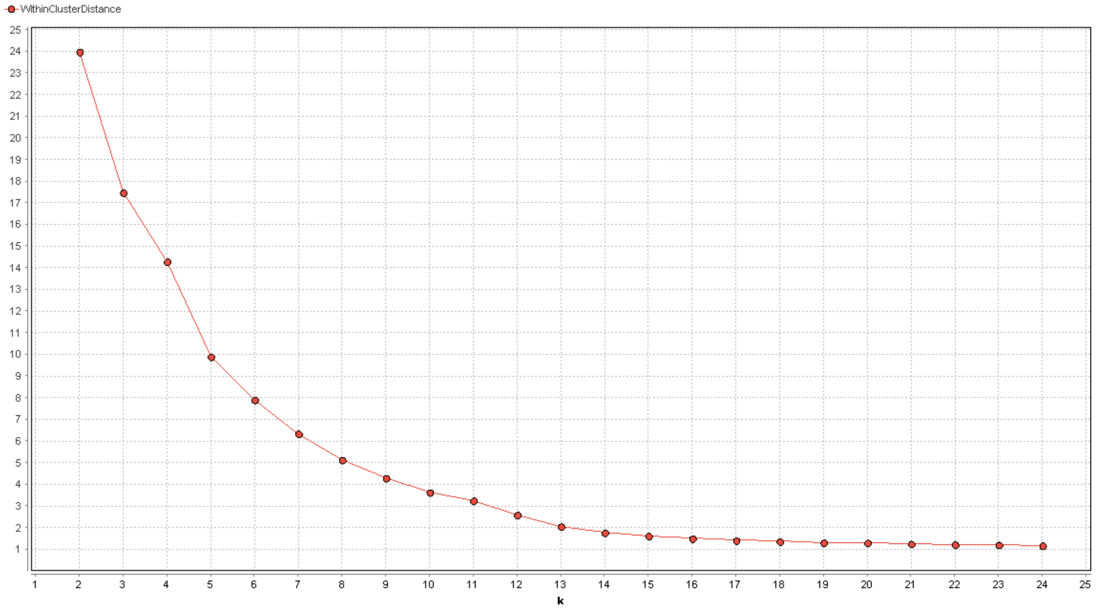


Figure 18 - Elbow Method

We completed subjective analysis on the dataset using the elbow method to try to identify the optimal range for K. As can be seen in Figure 16, the optimal range for K is relatively unambiguous, but it would appear on the above plot to be somewhere between 12-15.

However, we will perform a more robust and reliable analysis to identify the optimal value for K.

Objective measures of clusters found.

In order to get a factual basis for choosing the optimal value for K, we decided to use the Gap Statistic, Davie Bouldin Index, and the Silhouette method.

Using R, we calculated the Gap Statistic in order to try to identify the optimal value of K. This statistic equates the entire intra-cluster variation for various values of k with their anticipated values under null reference distribution of the data. The value that maximizes the gap statistic will be the value estimated to be K's optimal value.

Based on this statistic, k=12 is the optimal value, and this would fall into the range suggested earlier by the Elbow Method.

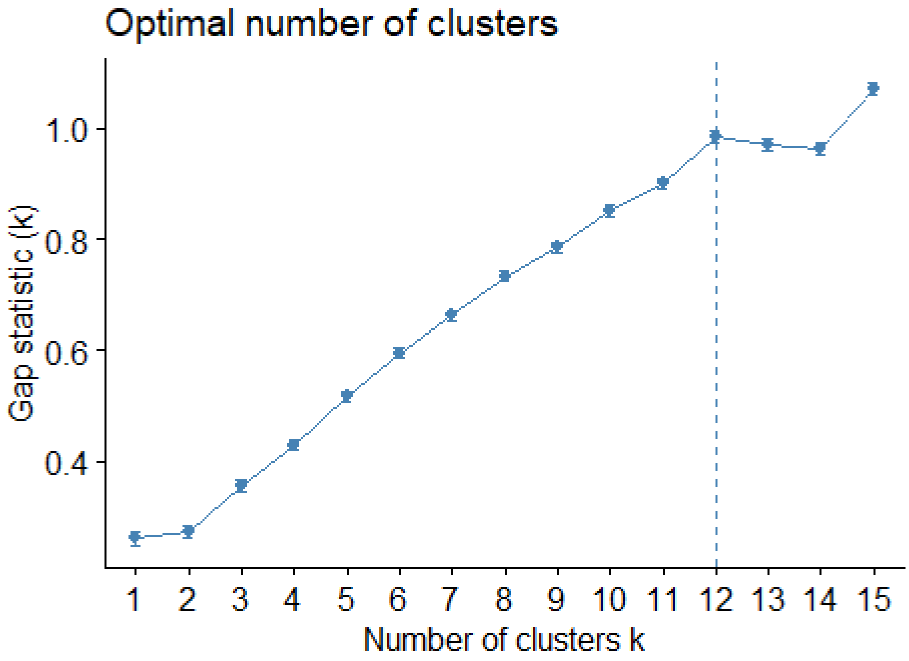
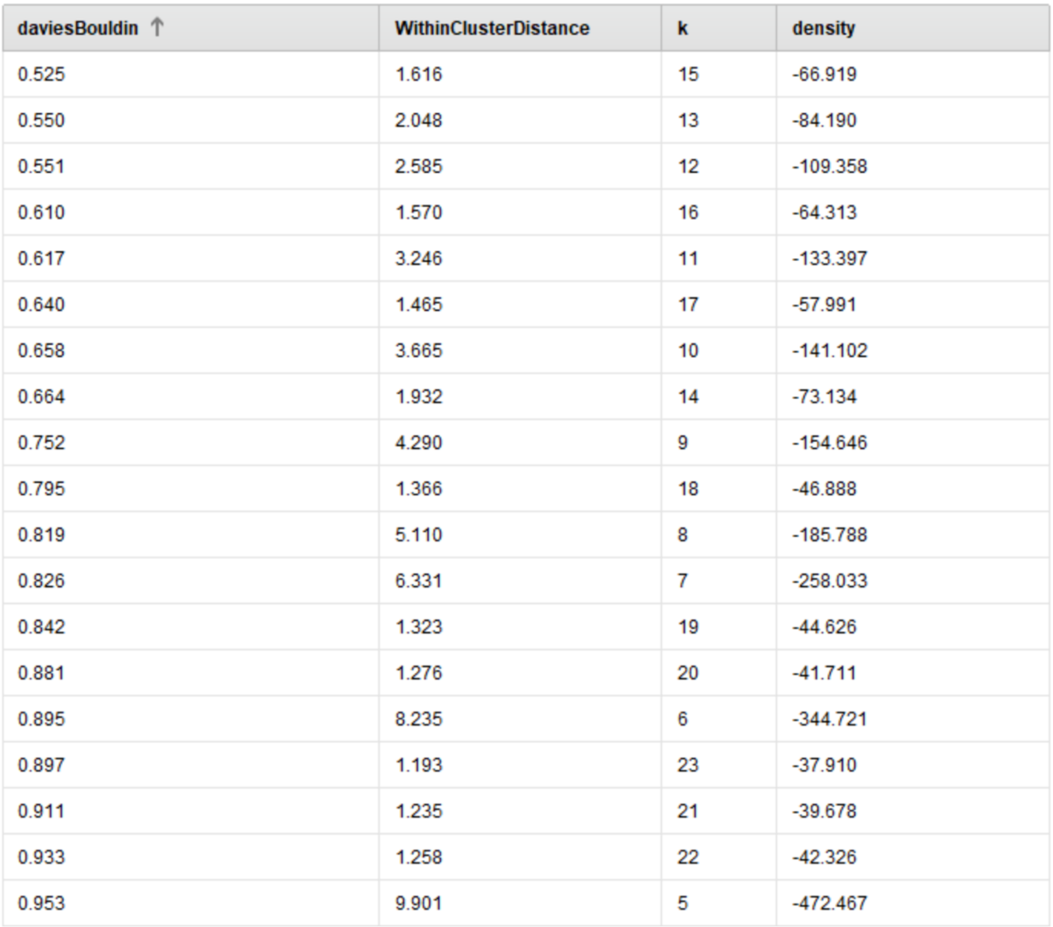


Figure 19 - Gap Statistic

Next in Rapidminer, we analyzed the Davies Bouldin Index measure for K=2 to K=24. Davies-Bouldin index is a validation tool used to decide the optimal value of K. It is a ratio between the cluster spread and the cluster's separation. The lower the Davies Bouldin Index (DBI), the better the clustering.



daviesBouldin ↑	WithinClusterDistance	k	density
0.525	1.616	15	-66.919
0.550	2.048	13	-84.190
0.551	2.585	12	-109.358
0.610	1.570	16	-64.313
0.617	3.246	11	-133.397
0.640	1.465	17	-57.991
0.658	3.665	10	-141.102
0.664	1.932	14	-73.134
0.752	4.290	9	-154.646
0.795	1.366	18	-46.888
0.819	5.110	8	-185.788
0.826	6.331	7	-258.033
0.842	1.323	19	-44.626
0.881	1.276	20	-41.711
0.895	8.235	6	-344.721
0.897	1.193	23	-37.910
0.911	1.235	21	-39.678
0.933	1.258	22	-42.326
0.953	9.901	5	-472.467

Figure 20 - Davies Bouldin Index Values

Figure 18 shows the Davies Bouldin Index score for k = 2 to k=24. As can be seen, the optimal value, according to the DBI, is k=15. K =15 had a DBI of 0.525. K=12 was the third most optimal value with a DBI of 0.551.

Finally, we decided to use the Silhouette Method to ensure that due diligence was completed before picking the optimal value for K. The Silhouette method is a metric that's values range from -1 to 1. It measures how well the data has been clustered. A higher positive score means that the clustering is well clustered, and a high negative score means the clustering is incorrect and should be reversed.

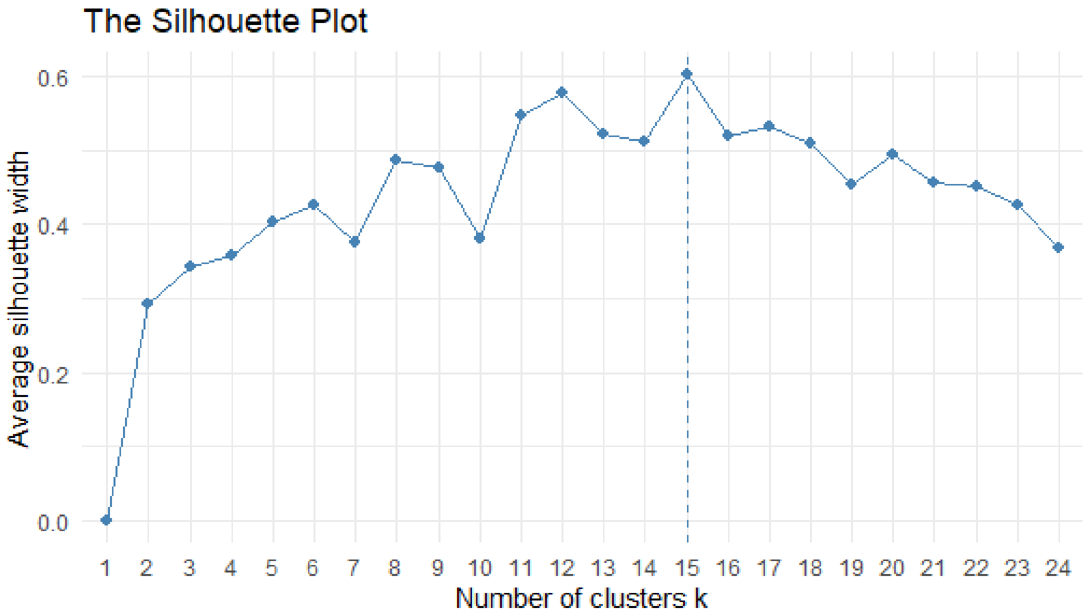


Figure 21 - Silhouette Plot

Based on the Silhouette Method, k=15 is the optimal value.

Since k=15 has been validated by the DBI score and Silhouette Method and was within the range indicated by the elbow method, we decided to make k=15.

Conclusion

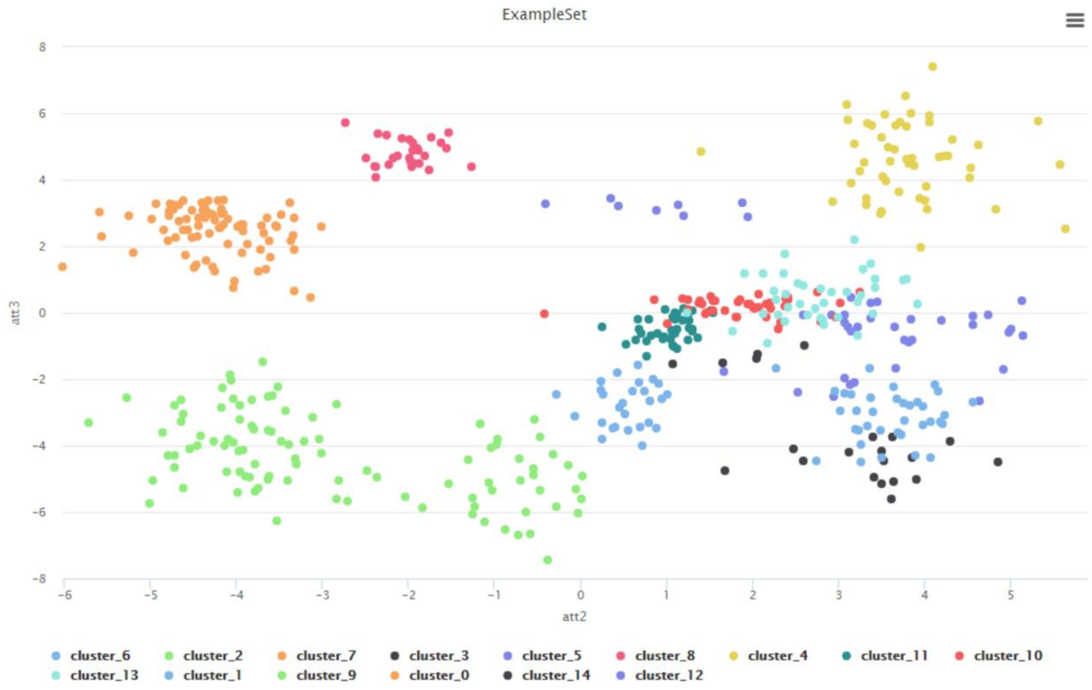


Figure 22 - K-Means Cluster K=15

Based on the analysis we have completed, k=15 would appear to be the optimal value for the cluster number. The clusters are shown in Figure 20, when K =15, the clusters are reasonably well defined.

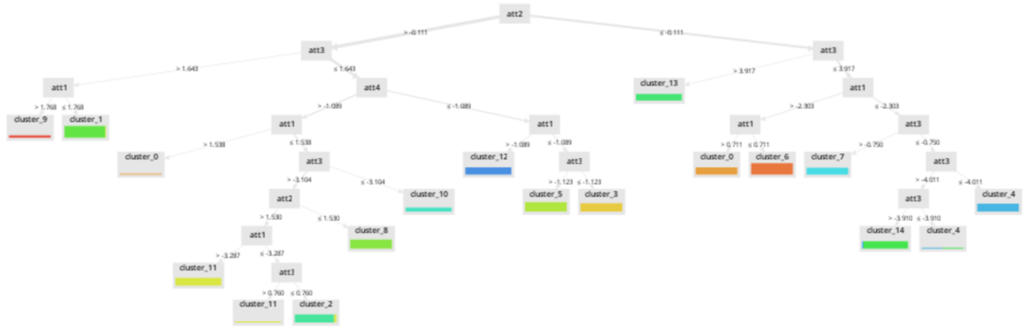


Figure 23 - Decision Tree for K-Means

Based on the Decision Tree model, we grew using the clustering data and K=15; we observed that Att1 and Att3 had the most influence on deciding the cluster membership.

Agglomerative Clustering

We then used Hierarchical agglomerative clustering to identify the optimal number of clusters according to this method.

Objective measures of clusters found.

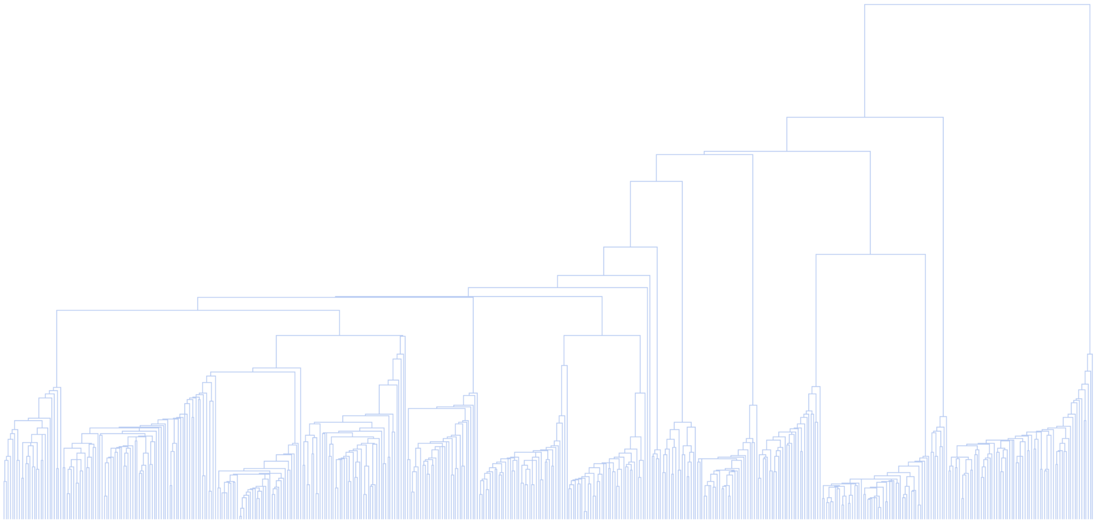


Figure 24 - Dendogram of Agglomerative Clustering

We first studied the dendrogram to get an idea of the optimal range of clusters using the agglomerative algorithm. Based on the Dendrogram, the data's largest gap would suggest that there are two or four clusters in the dataset. Furthermore, based on how these clusters are formed on the dendrogram, there is either one large cluster and one small cluster or a large cluster, two small clusters, and one tiny cluster.

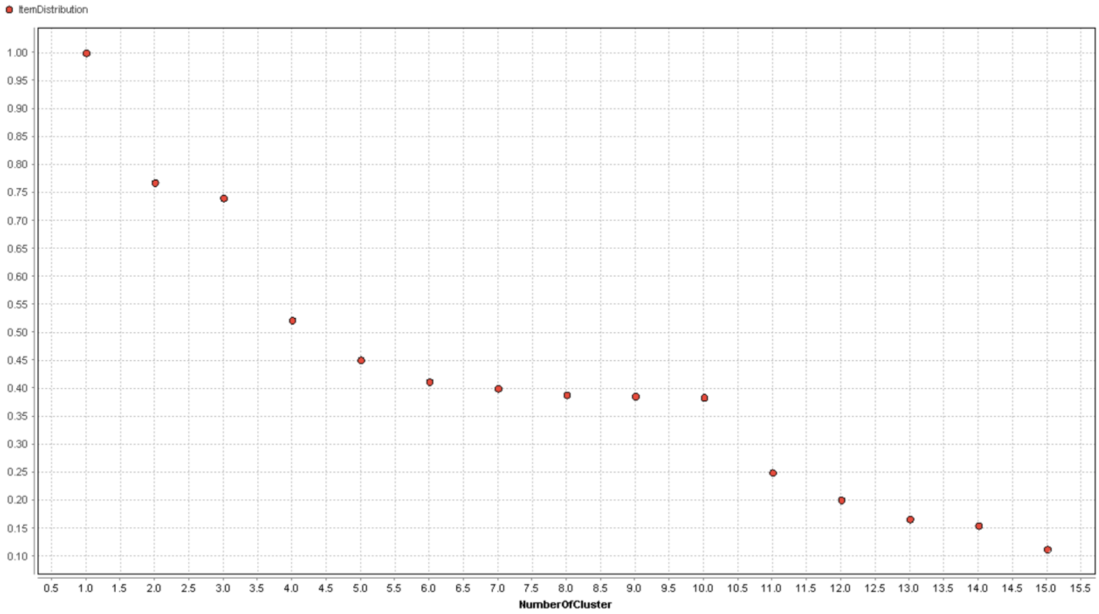


Figure 25 - Elbow Method

We then used the elbow method in order to investigate our theory further. As shown in Figure 21 above, the largest drop off in Item Distribution is when the number of clusters drops from one to two, and when the clusters drop from three to four. This would help support the analysis of the dendrogram.

Conclusion

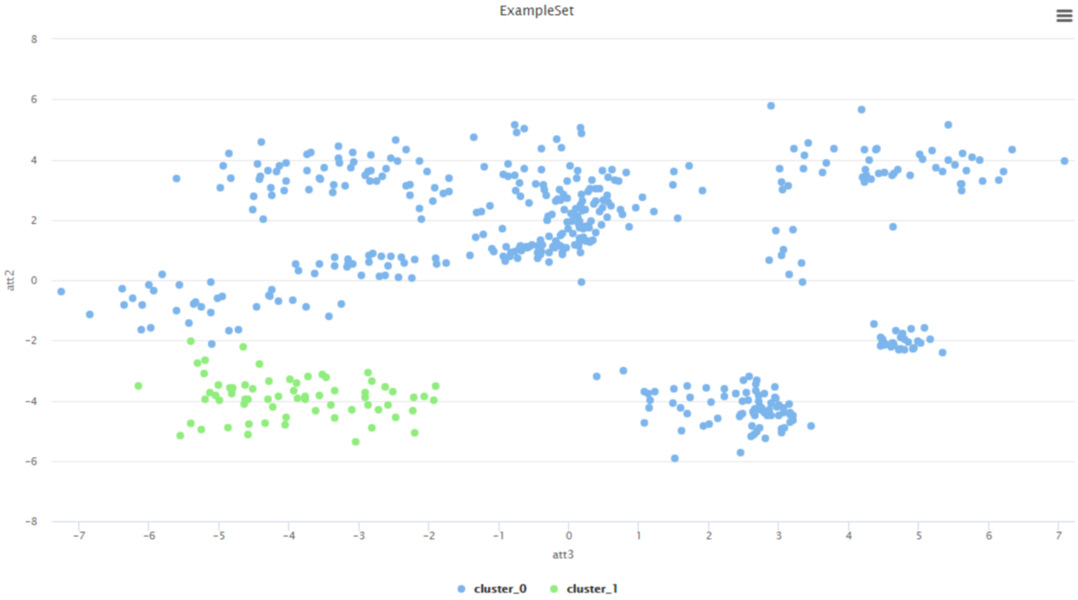


Figure 26 - Two Cluster Agglomerative Scatter Plot

As shown in the figure above, when there are two clusters, both clusters have a small distance between one another, but while cluster ones are close together, the members of cluster 0 are very dispersed.

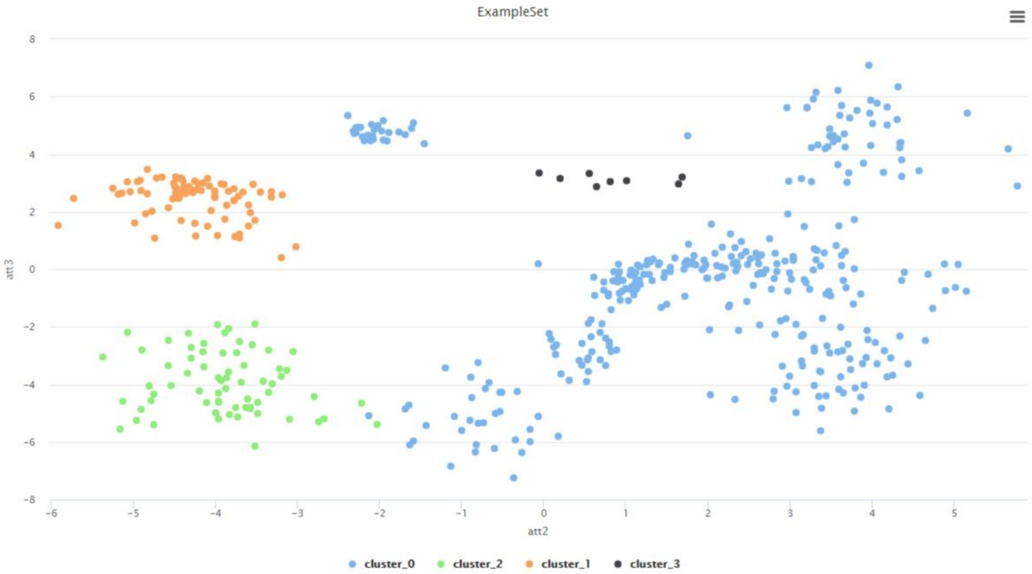


Figure 27 - Four Cluster Agglomerative Scatter Plot

When there are four clusters in the data, three of the four clusters members are arranged closely together and have adequate distance from other clusters. Cluster 0 is quite dispersed, but it has a reasonable distance from the other clusters.

In this case, four clusters would appear to be the best choice, but two clusters do not fair much worse.