

**CITS3401 Data Warehousing  
Project 2: Pattern discovery &  
Predictive analytics report**

**Chamika Kariyawasam(22508087) & Shathish Nagulan(22485727)**

# Table of Contents

<b>1. Association Rules</b>	<b>3</b>
1.1 Definition	3
1.2 Justification	3
1.2.1 Age	3
1.2.2 Work Class	4
1.2.3 Education Level	4
1.2.4 Hours Worked	4
1.2.5 Marital Status	4
1.2.6 Country of Origin	4
1.2.7 Gender and Race	4
1.2.8 Occupation	4
1.3 Data Set Manipulation	5
1.4 Chosen Rules	5
1.5 Recommendation	6
<b>2. Attribute Selection</b>	<b>6</b>
2.1 Information Gain	6
2.2 Using information gain in Weka	6
2.3 Decision Trees	7
2.3.1 How decision trees are read and formed	7
2.3.2 Decision Tree of Adult Training Data Set with attribute selection	8
2.3.3 Performance of decision tree with attribute selection	8
2.3.4 Decision tree with custom chosen attributes	8
2.3.5 Decision tree with non-information gain attributes	9
2.3.6 Interpreting the trees	10
2.3.7 Performance of 10 fold cross validation for IG tree	11
2.3.8 Performance of 10 fold cross validation for non IG tree	11
<b>3. Clustering</b>	<b>12</b>
3.1 Correlation between Marital Status and Income	12
3.2 Correlation between Education and Income	13
<b>4. Data reduction</b>	<b>15-19</b>
<b>5. Putting it all together</b>	<b>20</b>
<b>6. Individual contribution and references</b>	<b>20</b>

# 1. Association Rules

## 1.1 Definition

Association rule mining is an approach to discovering patterns of co-occurrence in a large dataset by identifying entities that frequently appear together in a group. This can give us an insight into the correlation between two entities; for example, products bought together in each transaction or characteristics possessed by each individual.

## 1.2 Justification

From the adult-training dataset, we were assigned to explore the correlation between a subset of attributes and the subsequent income- bracket through mining interesting patterns. The attributes which comprise this subset are discussed below.

### 1.2.1 Age

Age is an important criterion to consider when discussing the influence on income. Figure 1 demonstrates the impact of age and education experience on annual earnings.

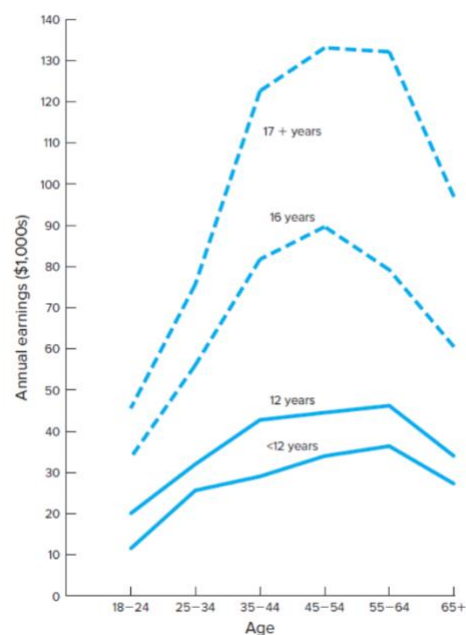


Figure 1: Age-earnings profile by Years of education [1]

A prevalent characteristic throughout Figure 1 is the rise in earnings with an increase in education levels at any age.

### **1.2.2 Work Class**

Different sectors in the economy require differing attributes in terms of the skills required for the profession. These skills are compensated accordingly and thus a range of incomes exist across professions.

### **1.2.3 Education Level**

Similar to work class, the skills required for certain occupations can only be obtained through progressing education levels. Education level can be attributed to income levels as employers seek those with the requisite educational understanding to be able to match the expectation of the salary package associated with the occupation.

### **1.2.4 Work Hours**

Most employers pay individuals on a fixed rate hourly basis, with incentive to work longer hours (e.g. double time, time and a half). Therefore, the number of hours an individual works can be directly correlated to their income bracket.

### **1.2.5 Marital Status**

Most employers pay individuals on a fixed rate hourly basis, with incentive to work longer hours (e.g. double time, time and a half). Therefore, the number of hours an individual works can be directly correlated to their income bracket.

We believe that these attributes provide the requisite scope to analyse any existing correlations with an individual's income bracket.

**The following attributes were deemed inappropriate to consider:**

### **1.2.6 Country of Origin**

The country of origin data is heavily skewed towards being a US native. An individual's country of origin does not necessarily reflect the country in which they currently reside and work in. Therefore, the country of origin has little bearing on the individual's income bracket.

### **1.2.7 Gender and Race**

These are attributes for analysing income disparities and deeper systemic issues and thus not applicable.

### **1.2.8 Occupation**

The occupation attribute is a finer grain of the work class attribute, thus can be considered under the work class umbrella,

## **1.3 Data Set Manipulation**

For the subset of attributes, the age attribute was discretised into 3 bins: [0-41], [42-66], [67-max]. The education attribute was discretised into 2 bins: [school education] (pre-school - high-school graduate) and [tertiary education] (some college – doctorate degree). Work hours was also discretised into 3 bins: [0-34], [35-66], [67-max].

## **1.4 Chosen Rules**

To construct the association rules with the Apriori algorithm, we chose the lowerBoundMinSupport to be 0.1, upper bound for minimum support to be 1.0, minMetric to be 1.1 with the metric Type being lift. This ranks the top association rules with the highest lift values. The reason why minMetric was chosen to be at 1.1 is that it allows for a larger array of values that have a positive correlation with whatever is on the left hand side and income on the right hand side.

The top 5 rules that have income on the right hand side with lift taking into accordance are:

**1. Age=42\_66 EducationID=Tertiary\_education MaritalStatus= Married-civ-spouse WorkHours=35\_66 ==> Income= >50K. conf:(0.6) < lift:(2.43)>**

Individuals who are aged 42 - 66 with a tertiary education who are married to a civilian spouse and work 35 - 66 hours earn over 50k. This is supported by a high lift factor of 2.43

**2. Age=42\_66 EducationID=Tertiary\_education MaritalStatus=Married-civ-spouse ==> Income= >50K. conf:(0.57) < lift:(2.33)>**

Individuals aged 42 - 66 with a tertiary education and are married to a civilian spouse earn over 50k.

**3. Age=42\_66 MaritalStatus= Married-civ-spouse WorkHours=35\_66 ==> Income= >50K. 1592 conf:(0.55) < lift:(2.23)>**

Individuals who are 42 - 66 who are married to a civilian spouse and work 35 - 66 hours a week earn over 50k

**4. Age=42\_66 MaritalStatus= Married-civ-spouse WorkHours=35\_66 ==> EducationID=Tertiary\_education Income= >50K. conf:(0.53) < lift:(2.23)>**

Individuals who are aged 42 - 66 who are married to a civilian spouse, and work 35 - 66 hours have a tertiary education and earn over 50k

**5. Age=42\_66 MaritalStatus= Married-civ-spouse 3272 ==> WorkHours=35\_66 Income= >50K. 1592 conf:(0.49) < lift:(2.15)>**

Individuals who are aged 42 - 66 and are married tend to work 35 - 66 hours a week and tend to earn over 50k.

All the above rules have quite high lift values. The reason as to why we have chosen lift over confidence is that some association rules do not make much sense in terms of confidence. Confidence is a form of conditional probability. For example confidence calculates the probability of a customer who buys X, how likely will they also buy Y.

lift on the other hand is a form of "unconditional". It calculates how many people bought X and Y together, therefore the confidence will be high. However if also many customers bought Y (but not together with X), the lift will be low.

Lift normalizes the confidence with the independence assumption. A lift of 1.0 means occurrence of an event X is independent to the occurrence of an event Y. A lift of <1 indicates a negative correlation between X and Y. A lift of >1 means the occurrence of X is positively correlated with Y.

## **1.5 Recommendation**

Based on these top 5 rules, our recommendation is that in order to earn 50k, an individual should obtain a tertiary education and get married by the age of 42. As a result of acquiring a tertiary education and getting married, they will tend to be allocated more hours at a workplace and in turn earn over 50k.

## **2. Attribute Selection**

Due to the irrelevancy and redundantly of certain attributes, attribute selection is crucial for decision trees. By using attribute selection, we can find the minimum number of attributes that will provide the maximum output of relevant information to build a decision tree. Mining on a reduced data set also makes the discovered pattern easier to understand.

## 2.1 Information gain

Decision trees will always try to maximize information gain (IG). An attribute with highest IG will be tested/split first. To rank the attributes with the highest gain, we must use the function InfoGainAttributeEval. The manual calculation of info gain is shown in the formula below:

$$Gain(A) = Info(D) - Info_A(D)$$

where:

- D is a tuple that belongs to a class C.

## 2.2 Utilising information gain in Weka

The top 5 attributes selected from Weka using information gain are shown below:

```
=== Run information ===
Evaluator:   weka.attributeSelection.InfoGainAttributeEval
Search:     weka.attributeSelection.Ranker -T -1.79769313486
Relation:   Kanye-weka.filters.unsupervised.attribute.Discrete
Instances:  15060
Attributes:  10
            Age
            WorkClass
            EducationID
            MaritalStatus
            Job
            Relationship
            Race
            Gender
            WorkHours
            Income
Evaluation mode: evaluate on all training data

=== Attribute Selection on all input data ===
Search Method:
  Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 10 Income):
  Information Gain Ranking Filter

Ranked attributes:
0.1655  6 Relationship
0.1585  4 MaritalStatus
0.0877  5 Job
0.0364  8 Gender
0.0324  1 Age

Selected attributes: 6,4,5,8,1 : 5
```

Figure 2: Attributes selected from Weka

## 2.3 Decision Trees

A decision tree is a decision support tool that utilises a tree-graph formation to model possible outcomes of decisions through the implantation of conditional control statements. It is represented in a flowchart – like structure in which each internal node represents a “test” on an attribute.

### 2.3.1 How decision trees are read and formed

Decision trees function by learning answers to a hierarchy of if/else questions which lead to a decision. When using our data set as an example, the target variable we are using is income. The income output takes two values; greater than 50k and less than or equal to 50k. The decision tree begins at the root node has the highest level of information gain. It then proceeds to look for the attribute with the next highest information gain.

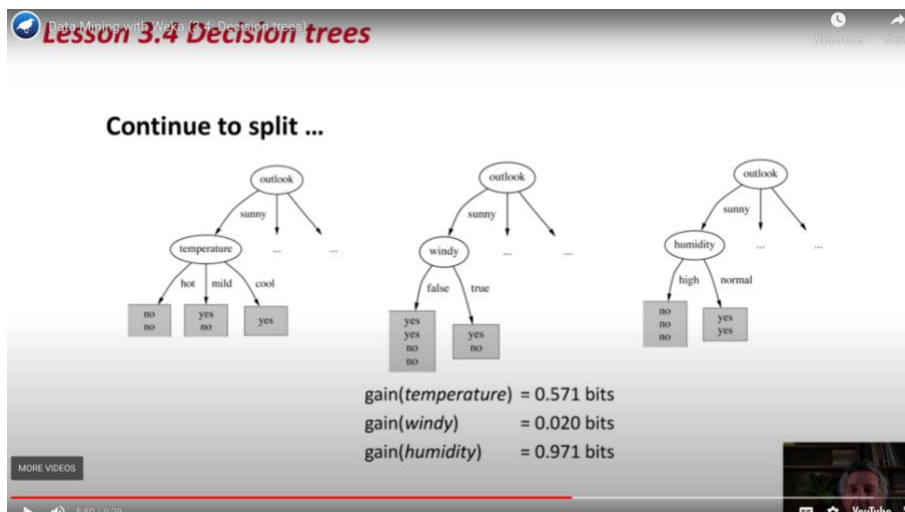


Figure 3: Decision Tree Visualisation

In reference to figure 3, once the root node has been chosen, it sees which attribute chosen next will have the highest information gain. Once the next attribute is chosen, this process continues to repeat until we get to the bottom most node, known as the leaf node.

### 2.3.2 Decision Tree of Adult Training Data set with attribute selection

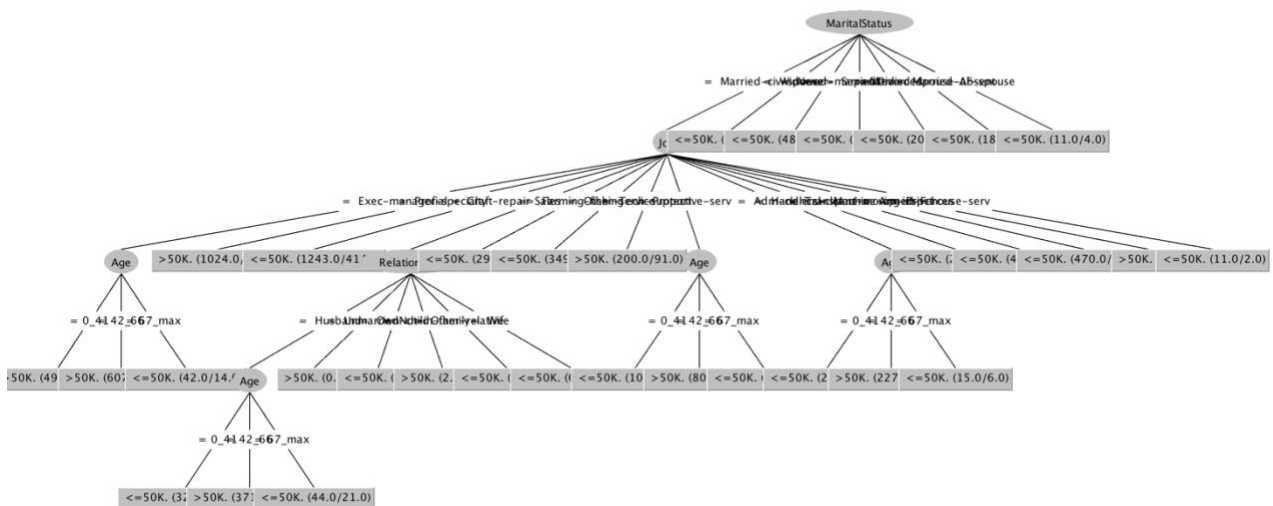


Figure 4: Decision tree with attribute selection



### 2.3.3 Performance of decision tree with attribute selection

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.525	0.089	0.658	0.525	0.584	0.472	0.833	0.596	>50K.
	0.911	0.475	0.855	0.911	0.882	0.472	0.833	0.922	<=50K.
Weighted Avg.	0.816	0.380	0.806	0.816	0.809	0.472	0.833	0.842	

=== Confusion Matrix ===

```
      a      b  <-- classified as
1941 1759 |      a = >50K.
1011 10349 |     b = <=50K.
```

Figure 5: Performance of decision tree with attribute selection using Weka

### 2.3.4 Decision tree with custom chosen attributes

The chosen attributes are age, education level, marital status, work class, work hours and income. This is because these are crucial for earning a good level of income and not just for earning over >50k.

The performance of this is much worse than when the decision tree was done using information gain. The precision and recall are both lowered. This is expected due to the fact that we expect a tree with the best attributes using information gain to have higher attributes. However, it is important to note there have been instances where when we were playing around with this, the results of the non-information gain tree have been very close to the tree with information gain.

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.440	0.094	0.604	0.440	0.509	0.389	0.817	0.543	>50K.
	0.906	0.560	0.832	0.906	0.868	0.389	0.817	0.915	<=50K.
Weighted Avg.	0.792	0.445	0.776	0.792	0.780	0.389	0.817	0.823	

=== Confusion Matrix ===

```
      a      b  <-- classified as
1629 2071 |      a = >50K.
1068 10292 |     b = <=50K.
```

Figure 6: Performance of decision tree with attribute selection using Weka

### 2.3.5 Decision tree with non-information gain attributes

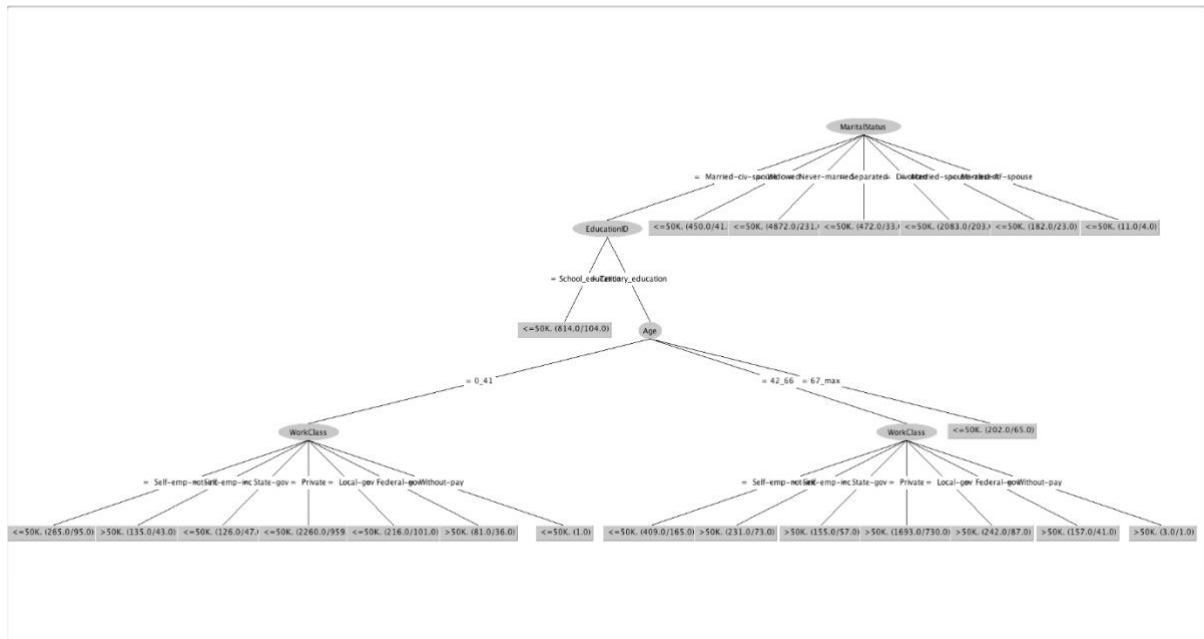


Figure 7: Decision tree with non-information gain attributes

### 2.3.6 Interpreting the trees

```
MaritalStatus = Married-civ-spouse
| Job = Exec-managerial
| | Age = 0_41: >50K. (494.0/168.0)
| | Age = 42_66: >50K. (607.0/170.0)
| | Age = 67_max: <=50K. (42.0/14.0)
| Job = Prof-specialty: >50K. (1024.0/304.0)
| Job = Craft-repair: <=50K. (1243.0/414.0)
| Job = Sales
| | Relationship = Husband
| | | Age = 0_41: <=50K. (328.0/157.0)
| | | Age = 42_66: >50K. (371.0/154.0)
| | | Age = 67_max: <=50K. (44.0/21.0)
| | Relationship = Unmarried: >50K. (0.0)
| | Relationship = Own-child: <=50K. (5.0)
| | Relationship = Not-in-family: >50K. (2.0/1.0)
| | Relationship = Other-relative: <=50K. (4.0/1.0)
| | Relationship = Wife: <=50K. (62.0/18.0)
| Job = Farming-fishing: <=50K. (294.0/54.0)
| Job = Other-service: <=50K. (349.0/49.0)
| Job = Tech-support: >50K. (200.0/91.0)
| Job = Protective-serv
| | Age = 0_41: <=50K. (106.0/43.0)
| | Age = 42_66: >50K. (80.0/34.0)
| | Age = 67_max: <=50K. (13.0/1.0)
| Job = Adm-clerical
| | Age = 0_41: <=50K. (259.0/95.0)
| | Age = 42_66: >50K. (227.0/105.0)
| | Age = 67_max: <=50K. (15.0/6.0)
| Job = Handlers-cleaners: <=50K. (252.0/50.0)
| Job = Transport-moving: <=50K. (485.0/149.0)
| Job = Machine-op-inspct: <=50K. (470.0/110.0)
| Job = Armed-Forces: >50K. (3.0)
| Job = Priv-house-serv: <=50K. (11.0/2.0)
MaritalStatus = Widowed: <=50K. (450.0/41.0)
MaritalStatus = Never-married: <=50K. (4872.0/231.0)
MaritalStatus = Separated: <=50K. (472.0/33.0)
MaritalStatus = Divorced: <=50K. (2083.0/203.0)
MaritalStatus = Married-spouse-absent: <=50K. (182.0/23.0)
MaritalStatus = Married-AF-spouse: <=50K. (11.0/4.0)
```

*Figure 8: Tree with attribute selection*

Interpreting these trees is quite difficult because if you increase the numObj, the actual decision tree process gets lost. Hence it is quite difficult to see from the above diagrams how the tree works its way from the root to the bottom leaf node. The above information demonstrates the actual process the tree follows. The tree started at its root at MaritalStatus = Married-civ-spouse. Once it reached that if condition, it moved on to its next condition which is Job = Exec-managerial. From there, it moved onto its next condition which is Age = 0\_41: >50K. (494.0/168.0). You can see that the tree branching stopped here, and it outputted a prediction. The prediction is that the income is greater than 50k. The two numbers 494 and 168 mean that 494 observations in the training set ended up in this path and 168 were incorrectly classified. This is similar to the whole tree and similar to the other tree for non-information gain tree.

### 2.3.7 Performance of 10 fold cross validation for IG(Infogain) tree

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.525	0.090	0.655	0.525	0.583	0.471	0.829	0.596	>50K.
	0.910	0.475	0.855	0.910	0.882	0.471	0.829	0.922	<=50K.
Weighted Avg.	0.815	0.380	0.806	0.815	0.808	0.471	0.829	0.842	

*Figure 9: Performance of 10 fold cross for IG tree*

### 2.3.8 Performance of 10 fold cross validation for non IG(infogain) tree

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.440	0.094	0.604	0.440	0.509	0.389	0.812	0.547	>50K.
	0.906	0.560	0.832	0.906	0.868	0.389	0.812	0.917	<=50K.
Weighted Avg.	0.792	0.445	0.776	0.792	0.780	0.389	0.812	0.826	

=== Confusion Matrix ===

a	b	<-- classified as
1629	2071	a = >50K.
1068	10292	b = <=50K.

*Figure 10: Performance of 10 fold cross for non IG tree*

When we compare the performance of the 2 models using 10 fold cross validation, we get a very similar performance comparison. The non info gain tree again has much worse precision, recall and f-measure compared to the model with info gain. This is highly expected as the attributed that were personally selected could not be as good as the attributes that were selected through weka.

### 3. Clustering

The clustering was done with simpleKmeans.

Final cluster centroids:

Attribute	Cluster#		
	0 (15060.0)	0 (7891.0)	1 (7169.0)
Age	0_41	0_41	42_66
WorkClass	Private	Private	Private
EducationID	Tertiary_education	Tertiary_education	Tertiary_education
MaritalStatus	Married-civ-spouse	Never-married	Married-civ-spouse
Job	Exec-managerial	Adm-clerical	Craft-repair
Relationship	Husband	Not-in-family	Husband
Race	White	White	White
Gender	Male	Female	Male
WorkHours	35_66	35_66	35_66

Cluster 0 <-- <=50K.  
Cluster 1 <-- >50K.

Incorrectly clustered instances : 4771.0 31.6799 %

Figure 11: Final cluster table

The clustering data reveals how cluster 0 is centred around a worker who is 0\_41 in a private work class with tertiary education and all the above attributes. The similar could also be said about cluster 1 which is centred around a different worker with different attributes. We can see cluster 0 is assigned to earning less than equal to 50k and cluster 1 is assigned with earning more than 50k. There are many ways in which the results can be interpreted. Below are 2 instances of how the clustering can be analysed.

#### 3.1 Correlation between Marital Status and Income

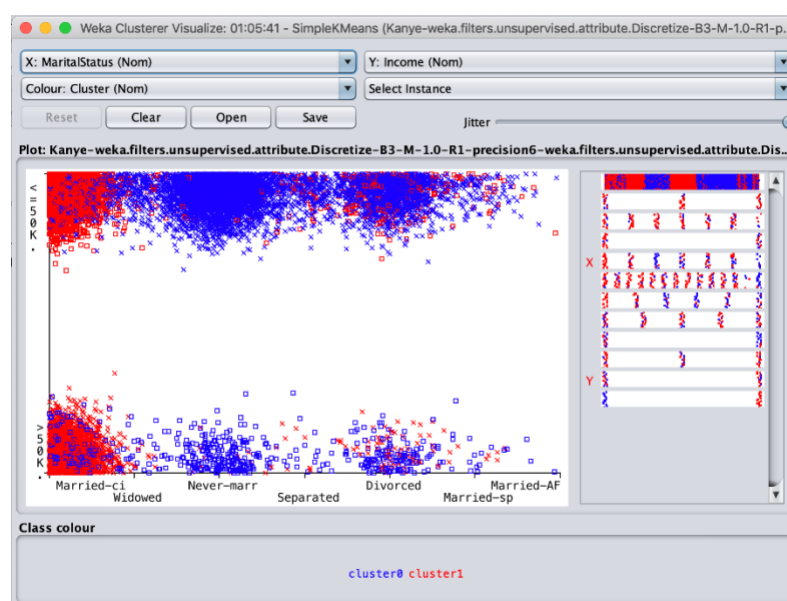


Figure 12: Marital status on income

From figure 12, we can see that the y-axis is income. The clustering confirms the association rules that we have investigated. We can see the spot where income >50k and married civilian spouse is filled with red. This confirms how being married has a strong relationship with earning over 50k. Cluster 0 is centered around earning less than 50k while cluster 1 is centered around earning more than 50k.

We see that the spot where never married and earns less than 50k is heavily populated with blue dots. From the association rules, we observed that never married individuals do not earn over 50k. The cluster diagram confirms the relationship between never being married and earn more than 50k. However, as we know in society, you do not need to be married to earn more than 50k. You can never have been married and still earn high salary. Therefore, this explains the blue dots in the never married and earning over 50k. For all the other categories of marital status, we can observe that the data is clustered in the less than or equal to earning 50k.

### 3.2 Correlation between Education and Income

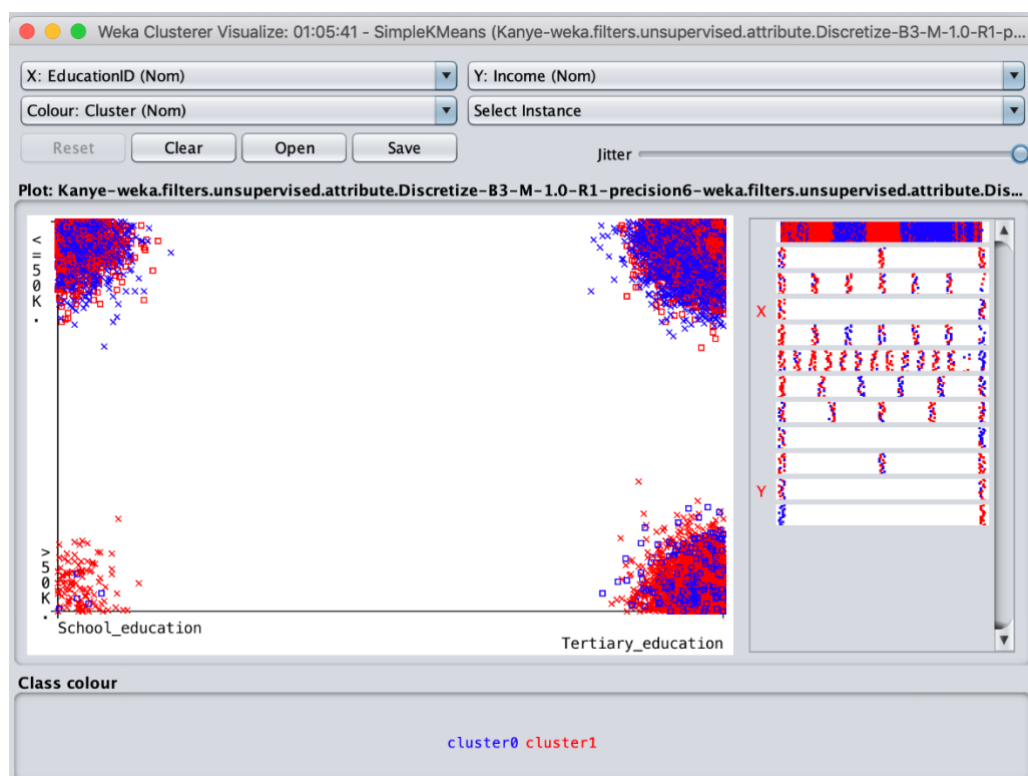


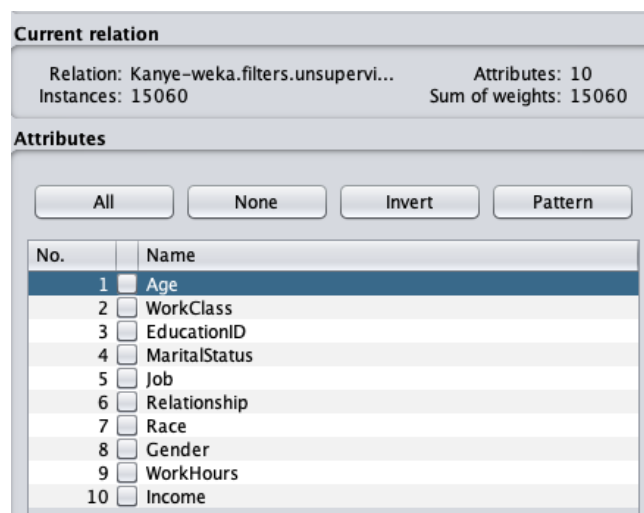
Figure 13: Education on income

We can see that school education and earning 50k is very lightly clustered. This confirms that more education leads to earning 50k. If we look at school education and earning less than or equal to 50k, we can see it is heavily clustered. Conversely, if we move to tertiary education, we can see the spot for earning over 50k and tertiary education is very highly clustered. Hence this suggests that higher education (tertiary level) can improve earnings. However, we can still see tertiary education resulting in earning less than 50k. This is due to the physical nature of the job market

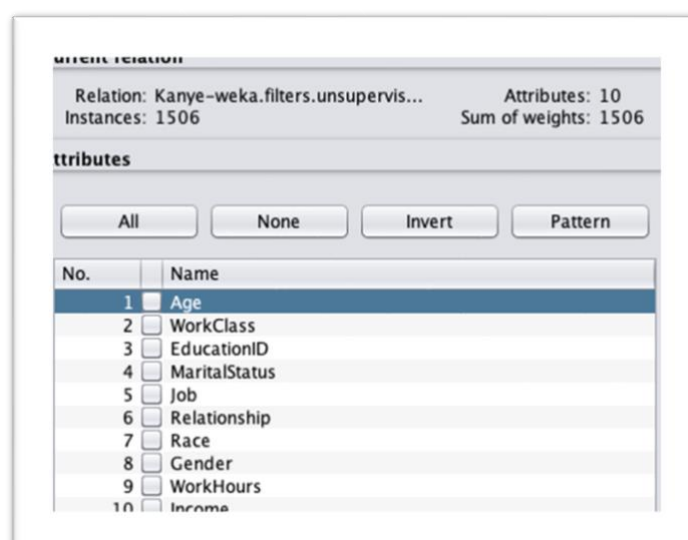
and the fact that the people who are earning less than 50k with tertiary education are individuals centered around never being married. We can see how solely having a tertiary education is not enough to earn over 50k. One must also settle down with a partner and have a family to get a full advantage of 50k.

#### **4.DATA REDUCTION**

First in order to perform numerosity reduction using sampling, we have to use a filter in pre-processing of weka called stratifiedRemovefolds. Essentially what this does is, it samples the total data set which has 15060 instances and it picks out a ratio of those instances to equally represent the whole data set. This is similar to how sampling works in real life. We collect data that equally represents different demographics of people and we use that to represent the whole population. In the context of this data set, the number of instances gets reduced from 15060 to 1506 while the number of attributes remains the same.



*Figure 14: Instances before numerosity reduction*



*Figure 15: Instances after numerosity reduction*

Once numerosity reduction has been performed, we moved on to applying feature reduction with principle component analysis.

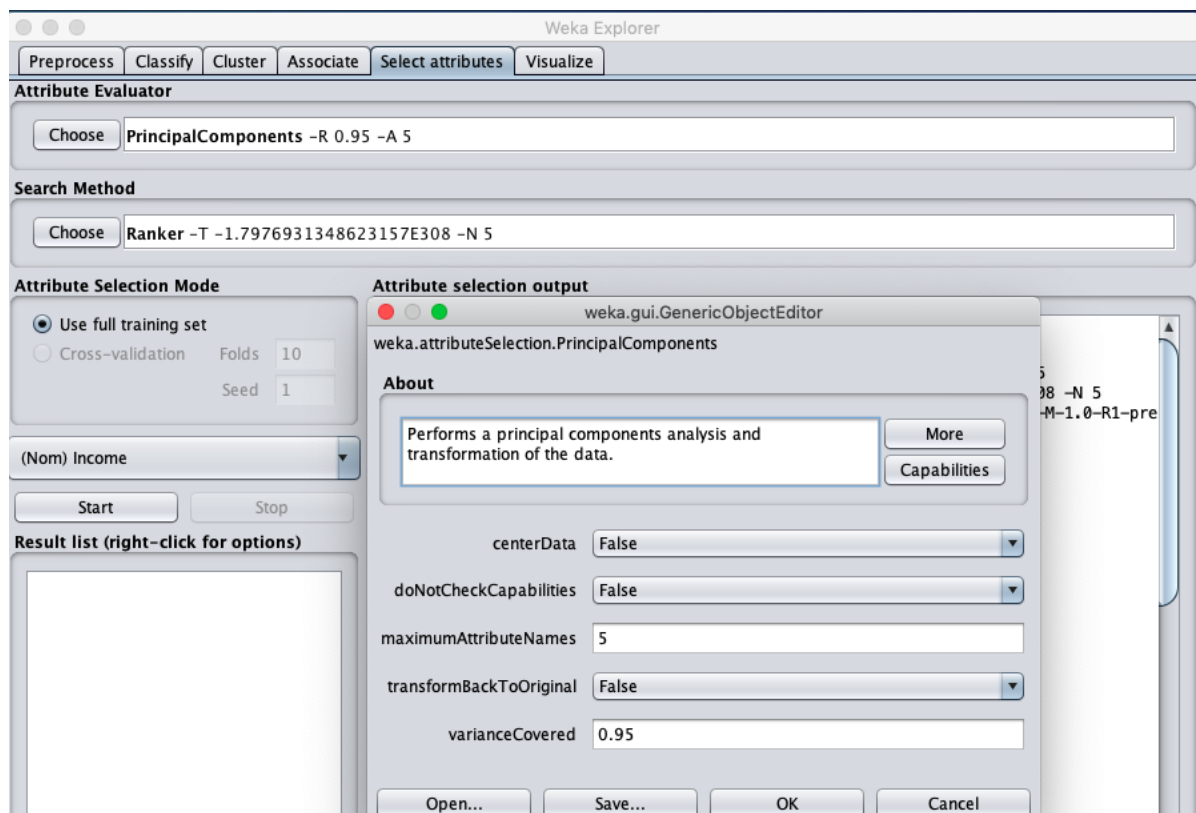


Figure 16: Principle component analysis settings

The way principle component analysis works is that it creates a correlation matrix with the features of the data set. The correlation matrix is far too big to be included in here but below is a small component of it.

#### Correlation matrix

1	-0.93	-0.24	-0.1	-0.06	-0.02	0.15	-0.04	-0.06	-0.03	-0.01
-0.93	1	-0.14	0.07	0.05	0.01	-0.13	0.04	0.08	0.04	0.04
-0.24	-0.14	1	0.09	0.03	0.01	-0.07	0.01	-0.04	-0	-0.07
-0.1	0.07	0.09	1	-0.06	-0.06	-0.48	-0.09	-0.05	-0.01	-0.01
-0.06	0.05	0.03	-0.06	1	-0.04	-0.34	-0.06	-0.04	-0.01	0.05
-0.02	0.01	0.01	-0.06	-0.04	1	-0.34	-0.06	-0.04	-0.01	0.07
0.15	-0.13	-0.07	-0.48	-0.34	-0.34	1	-0.47	-0.3	-0.04	-0.11
-0.04	0.04	0.01	-0.09	-0.06	-0.06	-0.47	1	-0.05	-0.01	0.07
-0.06	0.08	-0.04	-0.05	-0.04	-0.04	-0.3	-0.05	1	-0	0.04
-0.03	0.04	-0	-0.01	-0.01	-0.01	-0.04	-0.01	-0	1	0.01
-0.01	0.04	-0.07	-0.01	0.05	0.07	-0.11	0.07	0.04	0.01	1
-0.23	0.22	0.04	0.14	0.1	0.01	-0.17	0.04	0.02	0.03	0.01

Figure 17: correlation matrix



The way to interpret this correlation matrix is that each column represents a feature while each row also represents a feature. It is similar to a shape of a square.

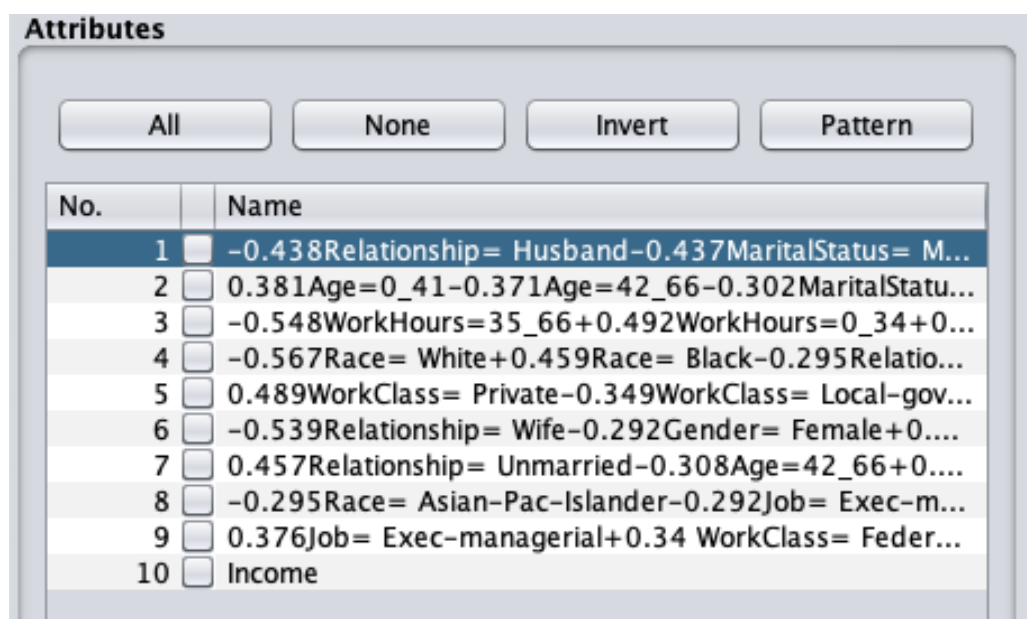
1	2	3	
1	0.5	0.65	1
0.5	1	0.554	2
0.65	0.554	1	3

Figure 18: Outline of correlation matrix

The correlation matrix from weka follows a similar shape as the above example matrix. However, we don't get to see the feature numbers outside the matrix and we only see the correlation inside. If we want to see the correlation between feature 3 and feature 2, we can see the value of 0.554. The diagonal 1s in the weka correlation matrix represent when the feature on the column is the same as the feature on the row. E.g. feature 1 on column is the same as feature 1 on the row hence they have a correlation of 100%.

The principle component analysis was done on the original data set which had 10 columns of information including income. Therefore, when doing the PCA, we limited the maximumAttributes to 9. Therefore, we have 10 attributes including income on both the original and data reduced training set. If we did not limit the maximum attributes to 9, weka actually creates 35 new attributes. However it is not wise to compare 2 models with different amounts of attributes.

These are the 9 best ranked attributes chosen from PCA



No.	Name
1	<input checked="" type="checkbox"/> -0.438Relationship= Husband-0.437MaritalStatus= M...
2	<input checked="" type="checkbox"/> 0.381Age=0_41-0.371Age=42_66-0.302MaritalStatu...
3	<input checked="" type="checkbox"/> -0.548WorkHours=35_66+0.492WorkHours=0_34+0...
4	<input checked="" type="checkbox"/> -0.567Race= White+0.459Race= Black-0.295Relatio...
5	<input checked="" type="checkbox"/> 0.489WorkClass= Private-0.349WorkClass= Local-gov...
6	<input checked="" type="checkbox"/> -0.539Relationship= Wife-0.292Gender= Female+0....
7	<input checked="" type="checkbox"/> 0.457Relationship= Unmarried-0.308Age=42_66+0....
8	<input checked="" type="checkbox"/> -0.295Race= Asian-Pac-Islander-0.292Job= Exec-m...
9	<input checked="" type="checkbox"/> 0.376Job= Exec-managerial+0.34 WorkClass= Feder...
10	<input type="checkbox"/> Income

Figure 19: 9 best attributes from PCA filter

## Decision tree performance with PCA attributes.

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.503	0.092	0.641	0.503	0.564	0.449	0.818	0.546	>50K.
	0.908	0.497	0.849	0.908	0.878	0.449	0.818	0.912	<=50K.
Weighted Avg.	0.809	0.398	0.798	0.809	0.800	0.449	0.818	0.822	

=== Confusion Matrix ===

```
a    b    <-- classified as
186 184 |    a = >50K.
104 1032 |   b = <=50K.
```

## Decision tree performance with original data set with no reduction.

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.492	0.071	0.692	0.492	0.575	0.477	0.839	0.617	>50K.
	0.929	0.508	0.849	0.929	0.887	0.477	0.839	0.924	<=50K.
Weighted Avg.	0.822	0.400	0.810	0.822	0.810	0.477	0.839	0.849	

=== Confusion Matrix ===

```
a    b    <-- classified as
1822 1878 |    a = >50K.
810 10550 |   b = <=50K.
```

As we can see from the performance, the precision of the original data set is slightly better than the decision tree with data reduction. However, it is only slightly better by 0.012. In terms of recall, the recall in non-reduced tree is better by 0.013. The F-measure is also higher in non-reduced tree. Therefore, the original model is slightly more precise than the reduced model.

We can conclude that the two trees are quite similar and performing numerosity reduction and feature reduction seems to have a negligible effect. This is as expected because data reduction can lead to decreased accuracy.

## Comparing the two models using 10 fold cross-validation.

### Model with data reduction

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

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.529	0.092	0.651	0.529	0.584	0.470	0.831	0.602	>50K.
	0.908	0.471	0.855	0.908	0.881	0.470	0.831	0.920	<=50K.
Weighted Avg.	0.815	0.378	0.805	0.815	0.808	0.470	0.831	0.842	

#### === Confusion Matrix ===

```
  a      b  <-- classified as
1957 1743 |      a = >50K.
1049 10311 |     b = <=50K.
```

### Original model with 10 fold cross validation

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

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.526	0.080	0.682	0.526	0.594	0.490	0.835	0.612	>50K.
	0.920	0.474	0.856	0.920	0.887	0.490	0.835	0.921	<=50K.
Weighted Avg.	0.823	0.377	0.814	0.823	0.815	0.490	0.835	0.845	

#### === Confusion Matrix ===

```
  a      b  <-- classified as
1946 1754 |      a = >50K.
 907 10453 |     b = <=50K.
```

Again when we run the 10 fold cross validation, we can see that the original model has better precision, recall and F-measure compared to the model with data reduction. The difference is not significant; however, we expect to see this sort of difference when data is reduced. When we applied numerosity reduction through sampling, the number of instances was reduced drastically. Therefore, the data isn't complete and isn't 100 accurate.

## **5.PUTTING ALL TOGETHER**

Unsupervised learning is a form of learning that looks for previously undetected patterns in a data set with no pre-existing labels and with minimum of human supervision. Therefore, most the time association rule mining and clustering are a part of unsupervised learning. Usually unsupervised learning does not have a target variable, however in the context of the dataset we are using, we have a target variable of income. We set the target variable to be income in the association rule mining section of this project because we wanted to see top 5 rules with income on right hand side. Therefore, we can't say the association rule mining we conducted falls under the category of unsupervised because we chose the target variable we wanted.

The clustering on the other hand is unsupervised because the output of the clustering was independent of the target variable. The 2 cluster diagrams shown in this report have income on y-axis but in reality, that y-axis could've been any other attribute.

Supervised learning is dependent on a target variable. The US adult income dataset uses income as a target variable when creating the J48 decision trees. The tree works its way from the root node to the lead node till the values of the target variable (income is revealed).

## **6.INDIVIDUAL CONTRIBUTIONS AND REFERENCES**

**Association rule mining :** Shathish & Chamika evenly. We spent a while trying to get the income on the RHS with the value being over >50k

**Attribute selection:** Mostly Chamika

**Interpreting trees with attribute and non attribute selection:** Chamika

**Clustering:** Shathish

**Data reduction:** Shathish

**Bonus Marks section:** Evenly both of us

**Editing and formatting:** Shathish

**References:**

<https://media.cheggcdn.com/study/75e/75ee6976-4af1-40f6-8f7f-55c82c9ff76d/719882-4-4QSSEI1.png>

