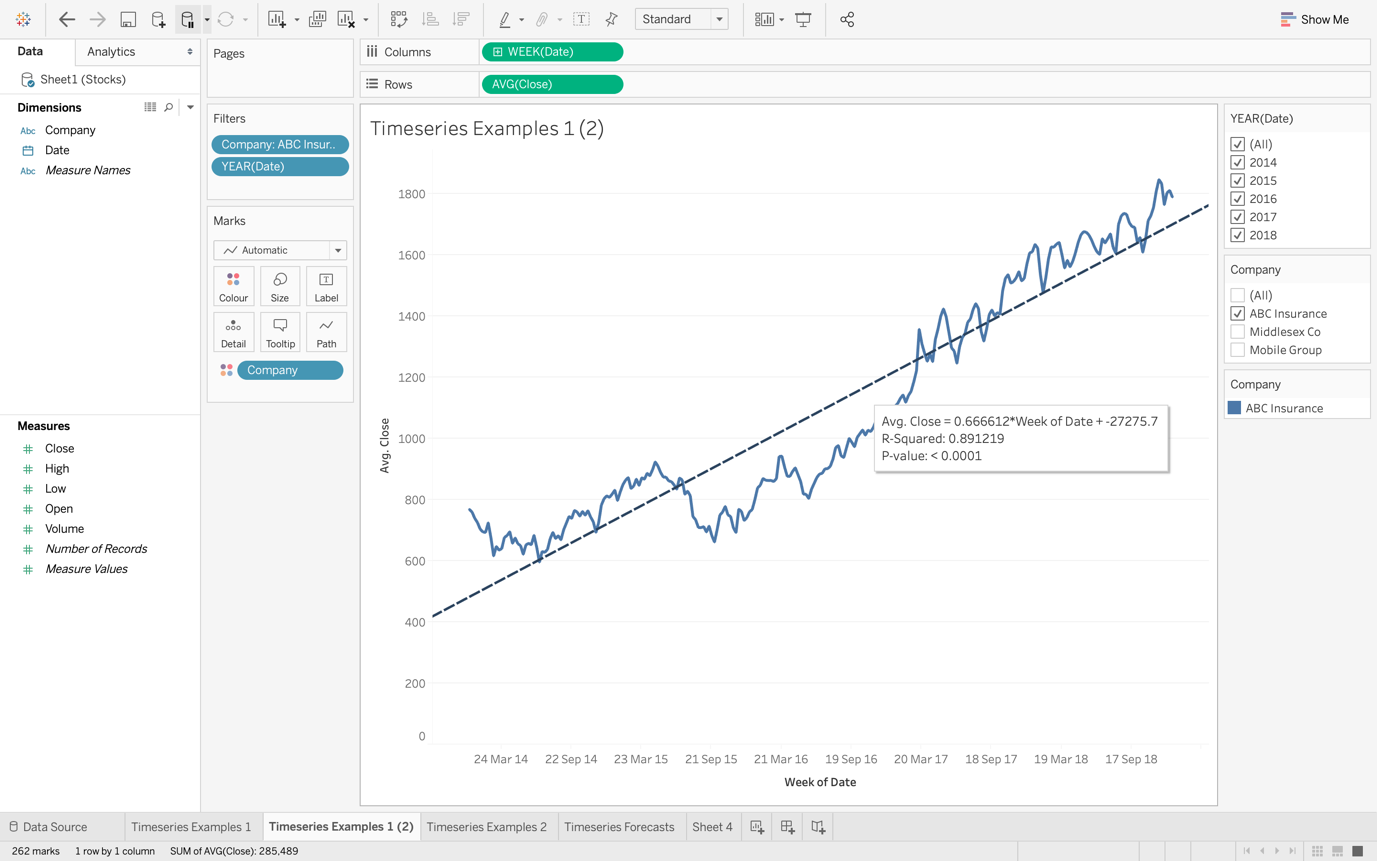
**Notes on evaluating the fit of the different data mining algorithms.**

**1. Time Series:**



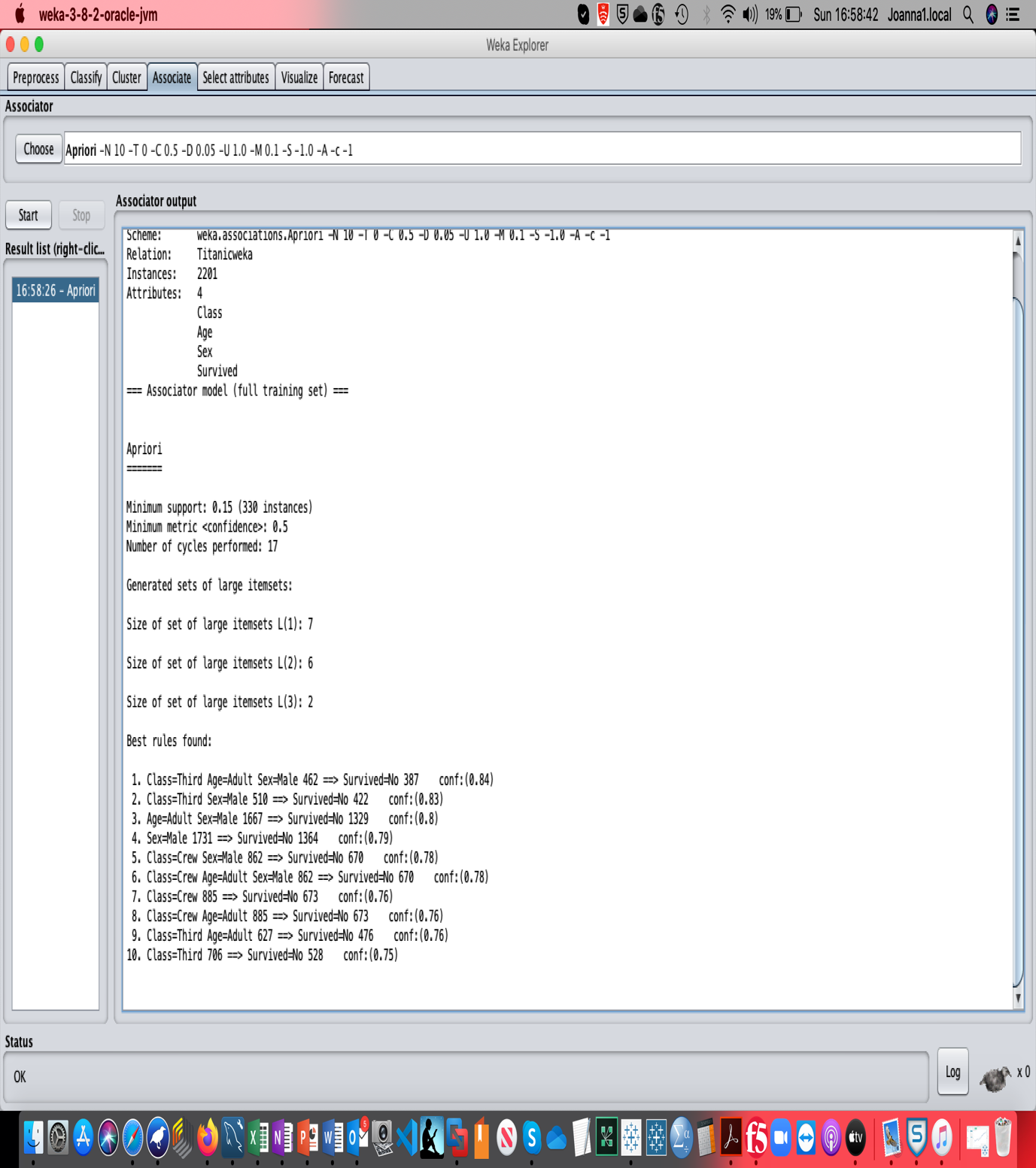
The fit of a trendline can be evaluated in two ways:

a. **Using the ‘R-Squared value’:** This is a value between 0 and 1 which is a statistical measure of how close the data values are to the fitted trendline (regression line). In other words, the closer the number is to 1, the better the trendline fits the data. A perfect fit is represented by 1 and 0 represents a trendline that does not fit the data at all. In the above screen shot the R-squared value is 0.89, which indicates that the trendline is a good fit for the data.

b. **Using the ‘R-value’:** (obtained by taking the square root of the R-squared value). This is a value between 1 and 0 which represents the correlation between two variables. This value can be positive, if the variable being investigated increases with time or negative, if the variable decreases with time. The closer the number is to 1, the stronger the correlation between the variable and time. A perfect correlation is represented by 1 and 0 represents that there is not any correlation between the variable and time. In the above screen shot the R value is 0.94, which indicates that there is a strong positive (variable close increases with time) between the shares close price and time. In other word, the share close price rose over the time period being examined.

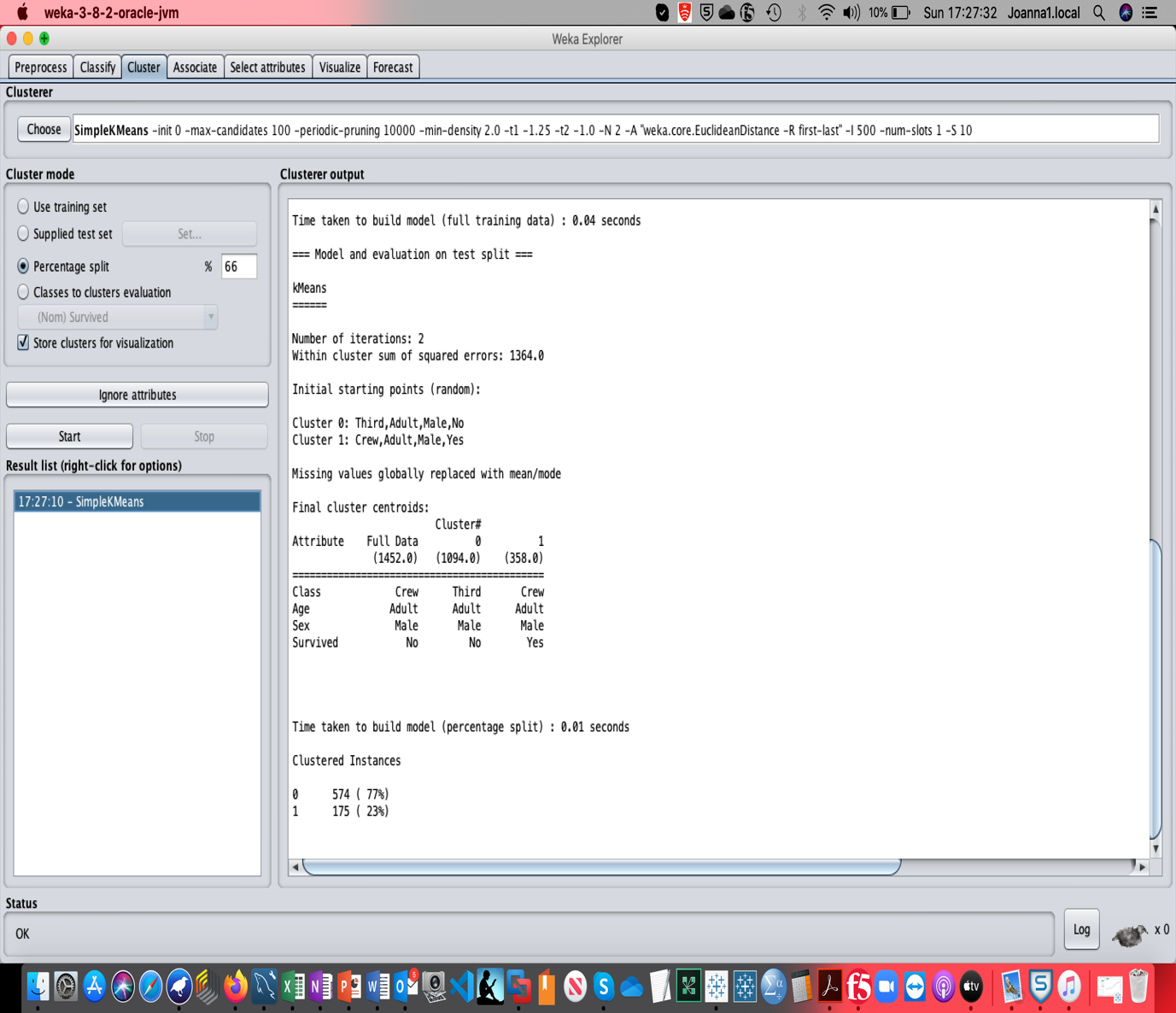
The r-squared value can be seen by hovering the mouse over the trendline.

**2. Association Rules:**



The accuracy of each association rule can be evaluated by looking at the Confidence level (conf:). This is a value between 0 and 1, which can be converted to a percentage by multiplying by 100. The percentage then represents how confident you are that the association rule is correct. For the first association rule in the above screenshot, the conf: = 0.84, which is 84%. Therefore, we are 84% sure that adult males from third class cabins did not survive.

**3. Clustering – K-means:**

****

The accuracy of the K-means clusters can be evaluated in two ways:

a. **Using the ‘within cluster sum of squared errors’:** This represents the sum of all the squared distances between each data point and the clusters centroid (mean value). The smaller the value the better. The actual value will vary, but you should use it to compare the different iterations you run. The one with the lowest “**‘within cluster sum of squared errors’”** is the one with the least errors within all the clusters, i.e. the best fit.

b. **Comparing the percentage in each cluster between the training set and test set.** This can only be done if you have used the percentage split option.

From the above screenshot, the clusters obtained from the training data have 1094 in cluster 0 and 358 in cluster 1. The was 1452 cases in the training set. Therefore, the percentage in each cluster is:

Cluster 0: 1094/1452\*100 = 75%

Cluster 1: 358/1452\*100 = 25%

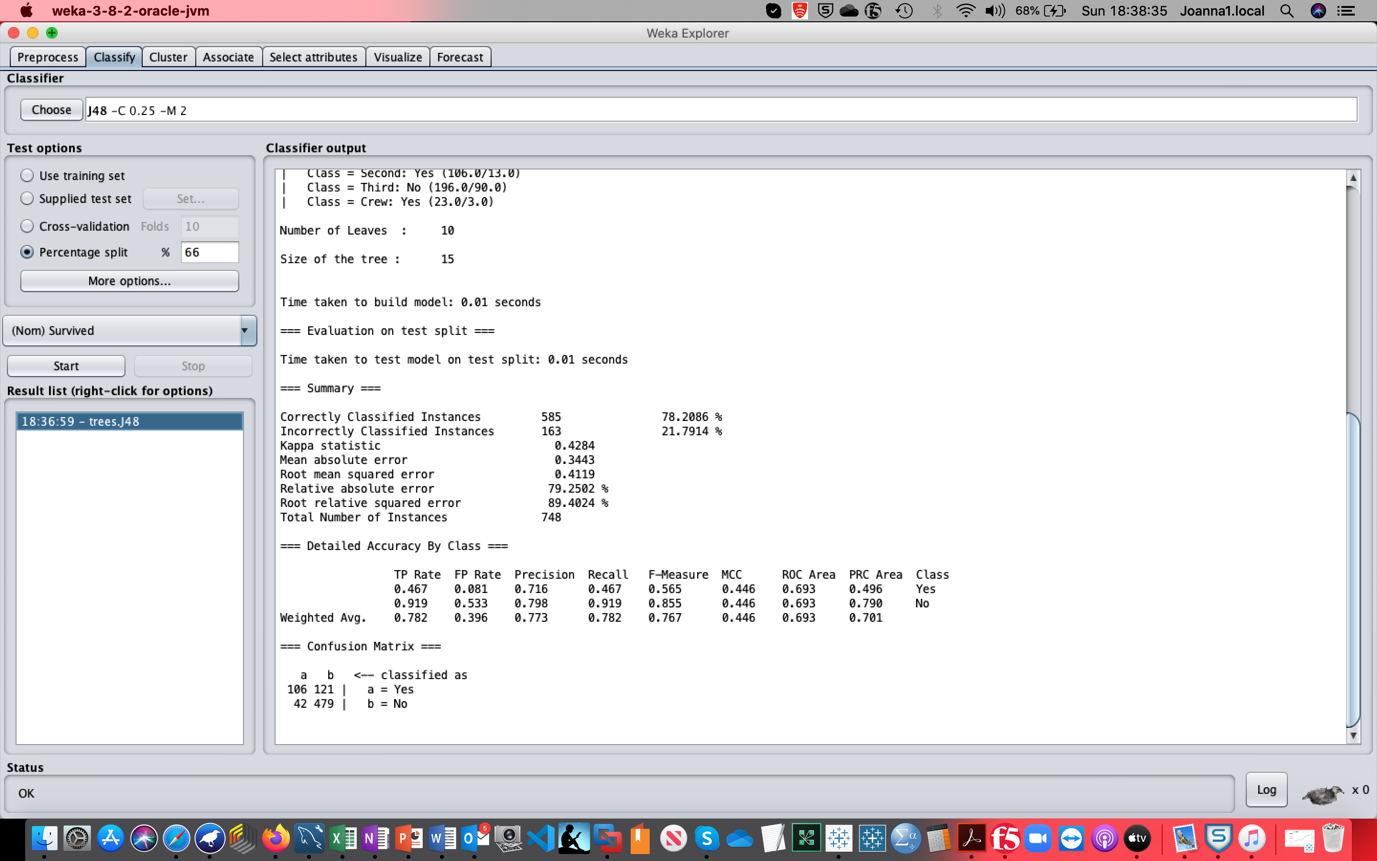
The percentages given for the testing set are shown at the bottom of the screenshot:

Cluster 0: 574 (77%)

Cluster 1: 175 (23%)

As the percentage in each cluster is similar for both the training set and the testing set, this implies that the clusters are good.

**4. Decision Trees:**

****

The accuracy of the decision tree can be evaluated in two ways:

a. **Using the Correctly Classified Instances**: This value indicated the percentage of instances that the tree correctly predicted. The higher the percentage, the better the tree has fitted the data. In the screenshot above, it is 78%, which is a good fit. A good fit is above about 75%.

b. **Using the Confusion Matrix:** This gives slightly more information about the numbers of correctly/incorrectly predicted instances. From the screenshot above the following can be discussed:

* The 106 represent the number of instances that the tree predicted the class value to be Yes, and the class variable was Yes.
* The 479 represent the number of instances that the tree predicted the class value to be No, and the class variable was No.
* The 121 represent the number of instances that the tree predicted the class value to be No, but the class variable was Yes.
* The 42 represent the number of instances that the tree predicted the class value to be Yes, but the class variable was No.

**5. Neural Networks:**

This uses the same two method as the decision trees.