CIS 435
Sync Session #9

Atef Bader, PhD

Agenda

- Financial Market Technical Analysis – Discussion Board Thread for this week
  - How to manage diversified fund?
- Project Deliverable

Financial Market Technical Analysis

- Warning/Disclaimer:
  - Please note that the Financial Market Technical Analysis data sets/commentary/rules we discuss and experiment with in this class are ONLY meant to be used for educational purposes for CIS435 class
  - Please do NOT apply these concepts/theories/rules on the financial assets you manage or own
  - Please talk to your financial advisor to get the opinion for investment or trading strategies.

Financial Market Technical Analysis

- Imagine yourself a fund manager that manages a fund that has one million dollars distributed in value initially (when you started your fund) as follows:
  - 30% of your fund is allocated to Bonds (TLT)
  - 30% of your fund is allocated to Stocks (SPY)
  - 20% of your fund is allocated to Gold (GLD)
  - 20% of your fund is allocated to Dollar (UUP)

Financial Market Technical Analysis

- Consider the Death Cross and Golden Cross for MA(50) MA(200)
- For any of these financial instruments that you have in your fund:
  - If the Golden Cross occurs and the current security price is above its MA(200) by 10% you reduce your position (sell recommendation) in that security by 5%
  - If the Death Cross occurs and the current security price is below its MA(200) by 7.5% you increase your position (buy recommendation) in that security by 5%
  - Is it visible to consider the correlation between those securities when moving the money from one security to another? For example, when you sell 5% of TLT positions, buy 5% position of SPY. When you sell 5% position of GLD buy 5% of UUP

Financial Market Technical Analysis

Financial Market Technical Analysis

Financial Market Technical Analysis

- Considering the composite rules (association rules??) for our fund management (for example, when you sell 5% of TLT positions, buy 5% position of SPY)

  1. Which DM/ML Algorithm will be the best to deal with these rules?

  2. Use Weka and capture/show your experimental results to support/prove your opinion/claims

Project

- What is the final project about?
  - Use Weka and apply different DM/ML algorithms/methods you learned in the class on the dataset files provided.

  - Please note that you need to be detailed in your documentation for the final project.

  - For every step you perform, document your step, capture screen-shot for Weka results, and then comment on your findings.

  - Three metrics to keep in mind while you are writing your report:
    1. Structure and organization of the report
    1. Quality of material in the report
    2. Quality of presentation in the report

Project

- Project Definition:
  - For your final project, you will synthesize the different analytical methods you learned in this class. You could use an algorithm that we didn’t cover in class.
  - Using the Weka Explorer module, combine different Explorer methodologies to input into Weka Experimenter module.
  - Tune parameters of different methods to get the best results.
    Then, decompose the analytical process into Knowledge Flow module.
  - Finally, load the data, and test the Knowledge Flow result. You will submit a detailed report in Microsoft Word that contains the following:
    - A comparison of the analytical results from the different methods
    - An explanation of the differences among the methods

Project

- Where you should start?
  - Under Session #9 course content tab, there are two exercises that you should start with first:
    1. Experimenter Exercise (S9) MSIS 435.pdf
    2. Knowledge Flow Exercise (S9) MSIS 435.pdf

Project

- What data files you should use in your experiments?
  - Under the assignment tab/project there are 6 arff files that you will use:
    1. WineCorre.arff
    2. WineDim1.arff
    3. WineNoCorre.arff
    4. WineDim2.arff
    5. WineAll3Dim.arff
    6. WineAllData.arff

Project

- Study the input data variables
  - Types
  - Ranges
  - Variance
  - Correlation
  - Perform a visual inspection

- Experiment with preprocessing and variable selection to improve model performance

- Comparative analysis

- Understand the strengths and weaknesses of each DM/ML Algorithm

Project

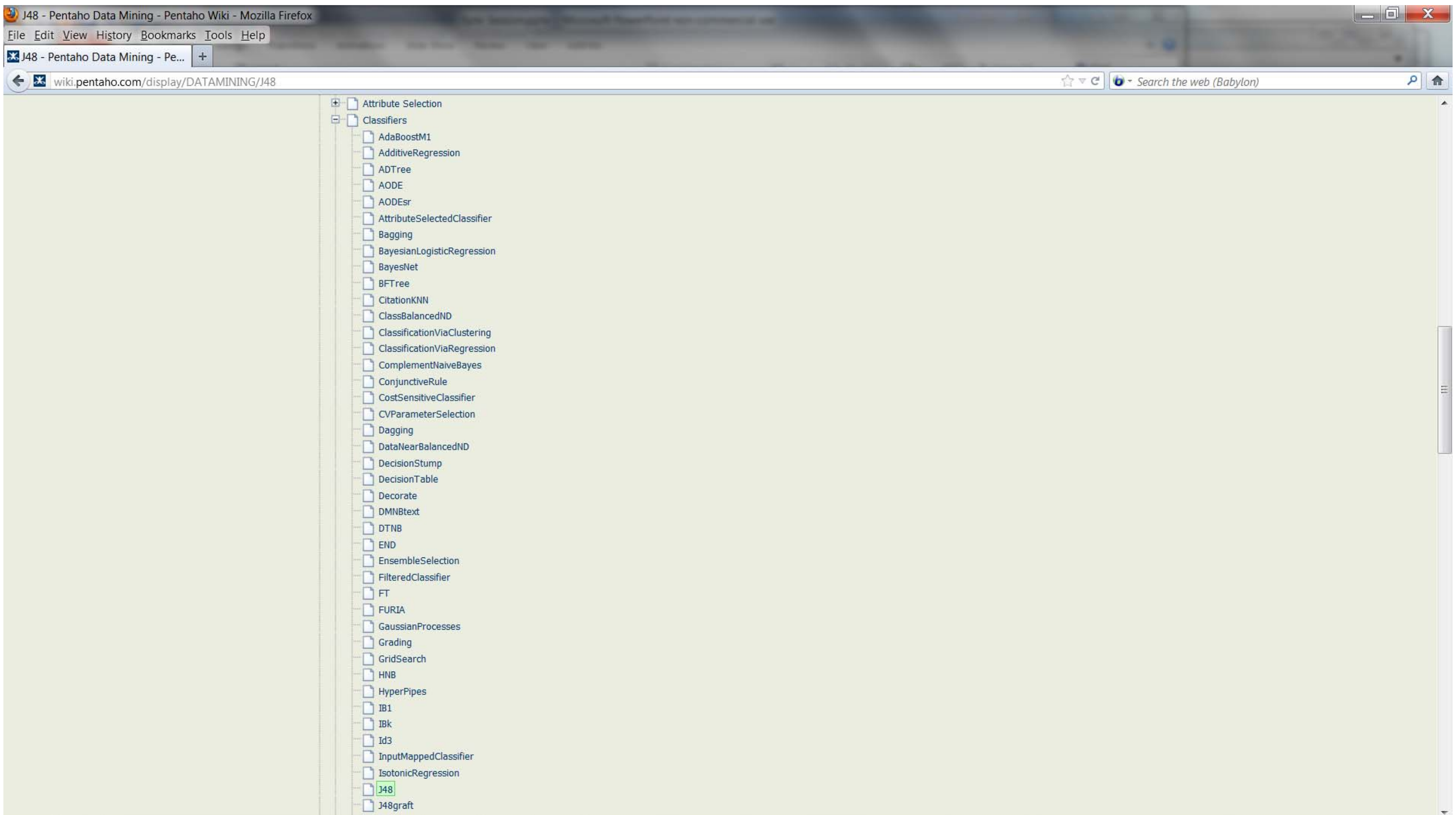
- How you should structure your final project report?
- Create Five major sections as follows
    1. Overview/Introduction
    2. Methodology/process used
        - Methods/Algorithms used
        - Experimental output/results for every method
    3. Comparative analysis for experimental results presented in section 2
    4. Conclusion and final remarks
    5. References

1. Overview

- For the final project, I reviewed the data files provided, apply a sample of data mining algorithms on that data and conduct the comparative analysis and finally provide my concluding remarks. There were six data files: one with the complete set of variables and five with subsets of the data.
- To analyze this data set I chose five different classification algorithms:
    1. J48
    2. JRip
    3. Naïve Bayes
    4. Artificial Neural Network
    5. Logistic Regression

1. Overview

- Which algorithms to choose? Pick your favorites (at least 5 to be analyzed)



2. Methodology/process used

- There are six data sets each with 178 rows of data. The WineAllData.arff file contains the superset of all variables.
- All of the data are continuous, numeric values except for the last variable which are nominal and consist of the different cultivars (A, B, C).

2. Methodology/process used

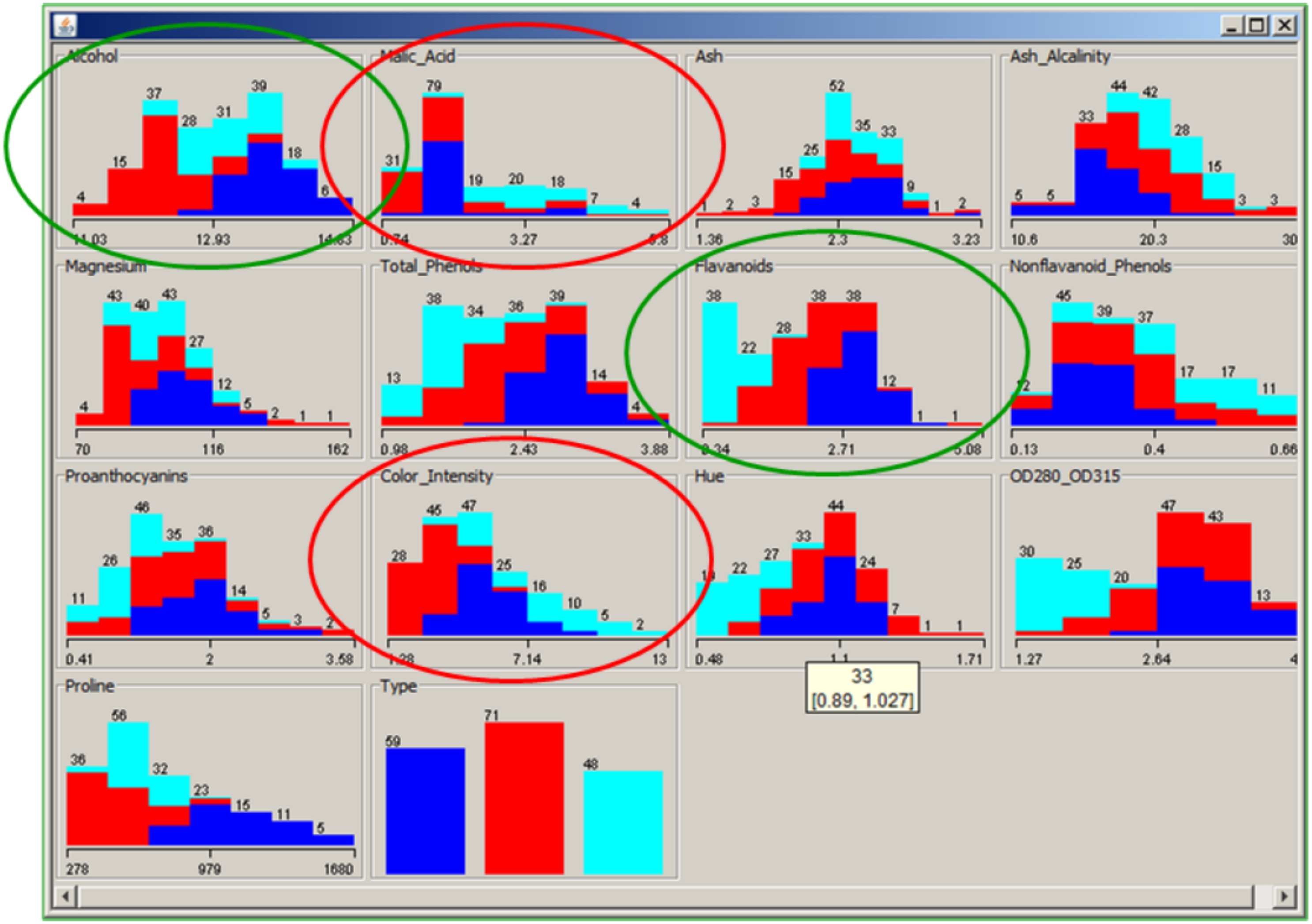
visual representation of the six data sets and the variables that are in each data set

	Alcohol numeric	Malic_Acid numeric	Ash numeric	Ash_Alcalinity numeric	Magnesium numeric	Total_Phenols numeric	Flavanoids numeric	Nonflavanoid_Phenols numeric	Proanthocyanins numeric	Color_Intensity numeric	Hue numeric	OD280_OD315 numeric	Proline numeric
WineAllData	X	X	X	X	X	X	X	X	X	X	X	X	X
WineAll3Dim	X	X	X	X	X	X	X			X	X	X	X
WineCorre	X					X	X					X	X
WineNoCorre		X	X	X	X			X		X	X		
WineDim1						X	X				X	X	
WineDim2	X				X					X			X

2. Methodology/process used

- From the following figure, it appears that there are two variables that might be important because they seem to differentiate between three possible classes. Alcohol and Flavenoids (circled in green) seemed to be a good way of differentiating the three classes because their distributions seem distinct. On the other hand, Malic Acid and Color Intensity (circled in red) seemed to have very similar distributions one on top of the other. Perhaps those are less important in classification. Let's test these observations using the Weka Explorer.

2. Methodology/process used

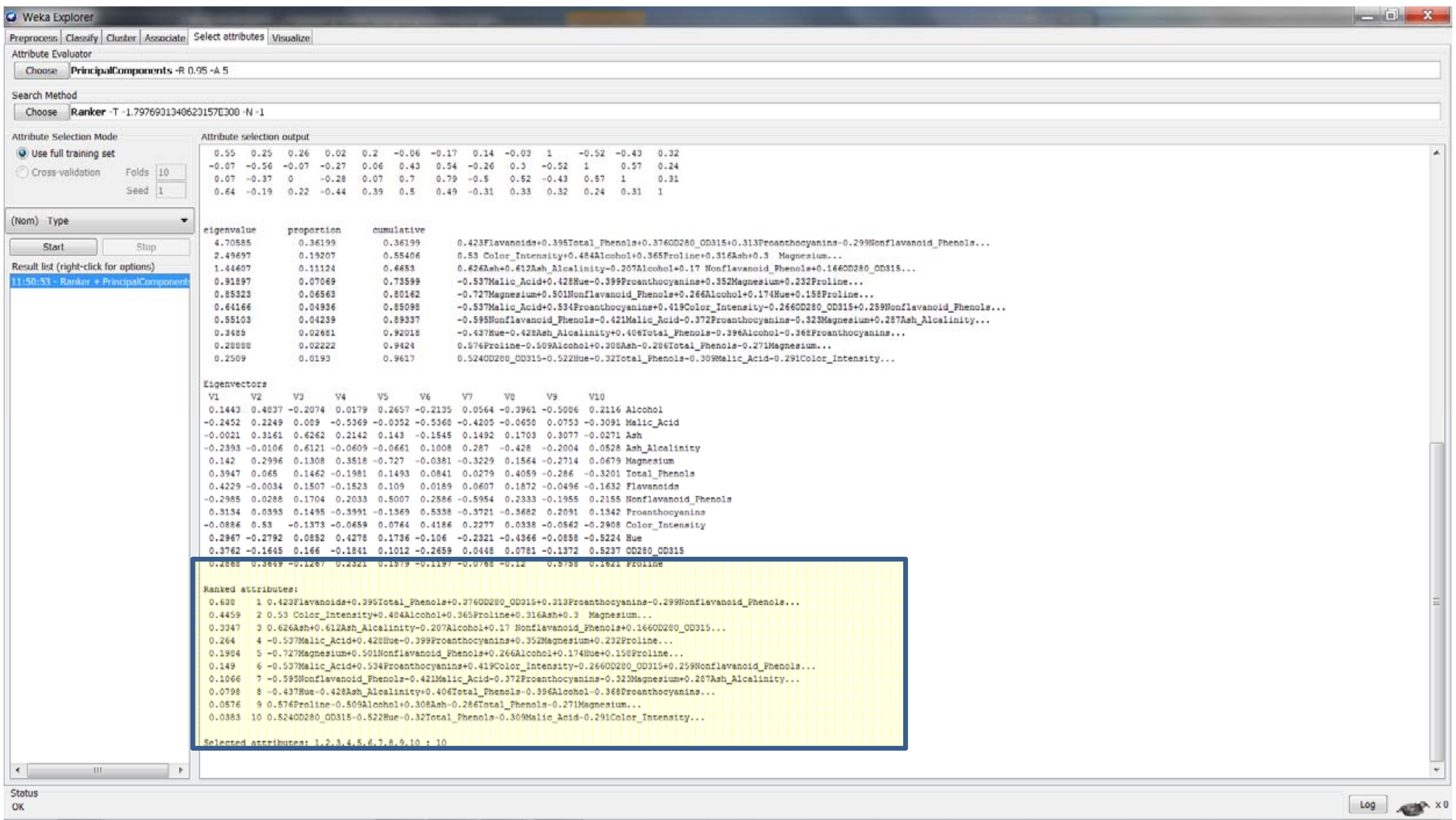


2. Methodology/process used

- Attribute Selection and Manipulation to Improve Algorithm Results
- In order to learn more about the datasets and Weka, I experimented with Weka's preprocessing and attribute selection functionality. For each version of the resulting dataset, I ran Weka's classification functionality, to see how changes to the dataset affect the algorithm.
- List of dataset manipulation:
  1. Preprocessing:
     - Supervised Discretize (First-Last and Kononeko).
     - Unsupervised Discretize (First-Last and 10 bins).
     - Unsupervised Discretize (First-Last and Find Number of Bins).
     - Supervised NominalToBinary.
     - Unsupervised Normalize.
  2. Attribute Selection:
     - WrapperSubsetEval + BestFirst
     - WrapperSubsetEval + GreedyStepwise
     - WrapperSubsetEval + RandomSearch
     - InfoGainAttributeEval + Ranker

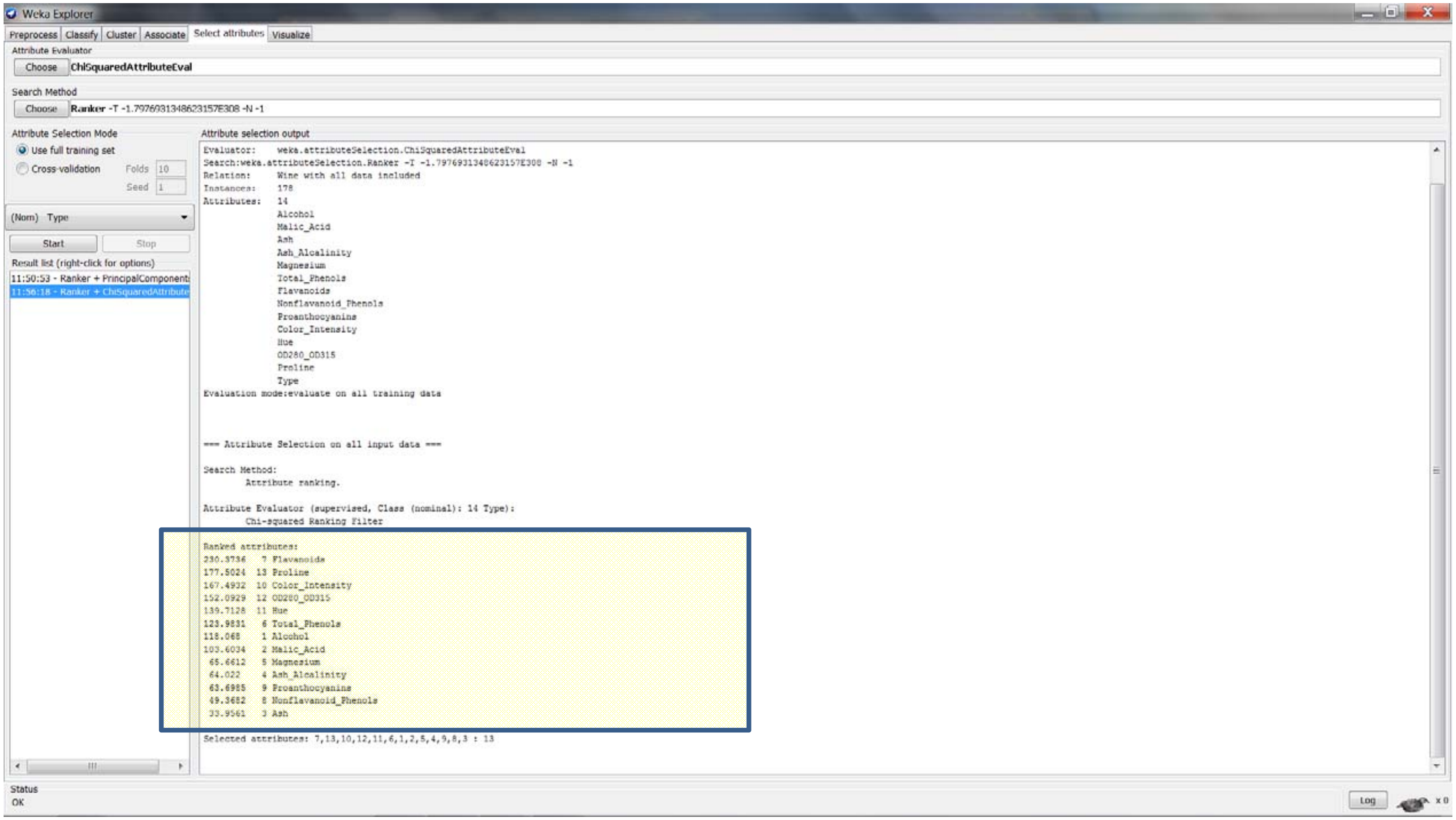
2. Methodology/process used

- As an exploratory step, I performed a Principal Components Analysis and a Chi-Squared Analysis to improve my understanding of the key explanatory variables. The results are as follows:



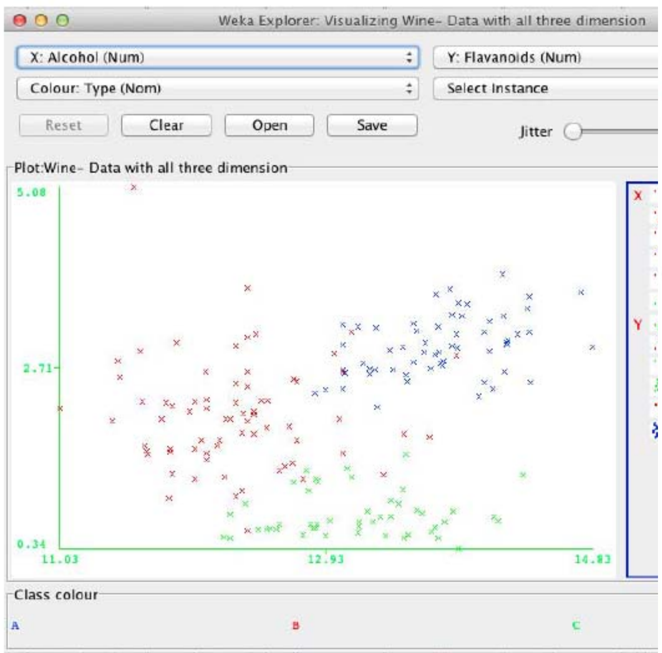
2. Methodology/process used

- As an exploratory step, I performed a Principal Components Analysis and a Chi-Squared Analysis to improve my understanding of the key explanatory variables. The results are as follows:



2. Methodology/process used

- Note how well that the class variable separates when the two analysis variables are Flavonoids and Proline. This suggests that these two variables may play a key role in the classification analysis. We will want to test this hypothesis when we perform our classification analysis.



2. Methodology/process used

- I started out with testing each dataset using the following six methods in Weka's Explorer module, tweaking various parameters in order to achieve the highest percentage of correctly classified instances. The best performing parameters for each dataset and methods are as follows:

Dataset/Method Parameter	J48	NaiveBayes Simple	IBK	JRiP	MultiLayer Perceptron	Logistic Regression
WineAll3Dim	minNumObj=1, numFolds=10	Default	KNN=20	minNumObj=3, folds=3	Default	Default
WineAllData	minNumObj=1, numFolds=10	Default	KNN=20	minNumObj=3, folds=3	Default	Default
WineCorre	minNumObj=1, numFolds=10	Default	KNN=10	minNumObj=2, folds=2	Default	Default
WineDim1	minNumObj=2, numFolds=10	Default	KNN=10	minNumObj=2, folds=6	Default	Default
WineDim2	minNumObj=1, numFolds=10	Default	KNN=25	minNumObj=3, folds=3	Default	Default
WineNoCorre	minNumObj=1, numFolds=6	Default	Default KNN=1	minNumObj=2, folds=2	Default	Default

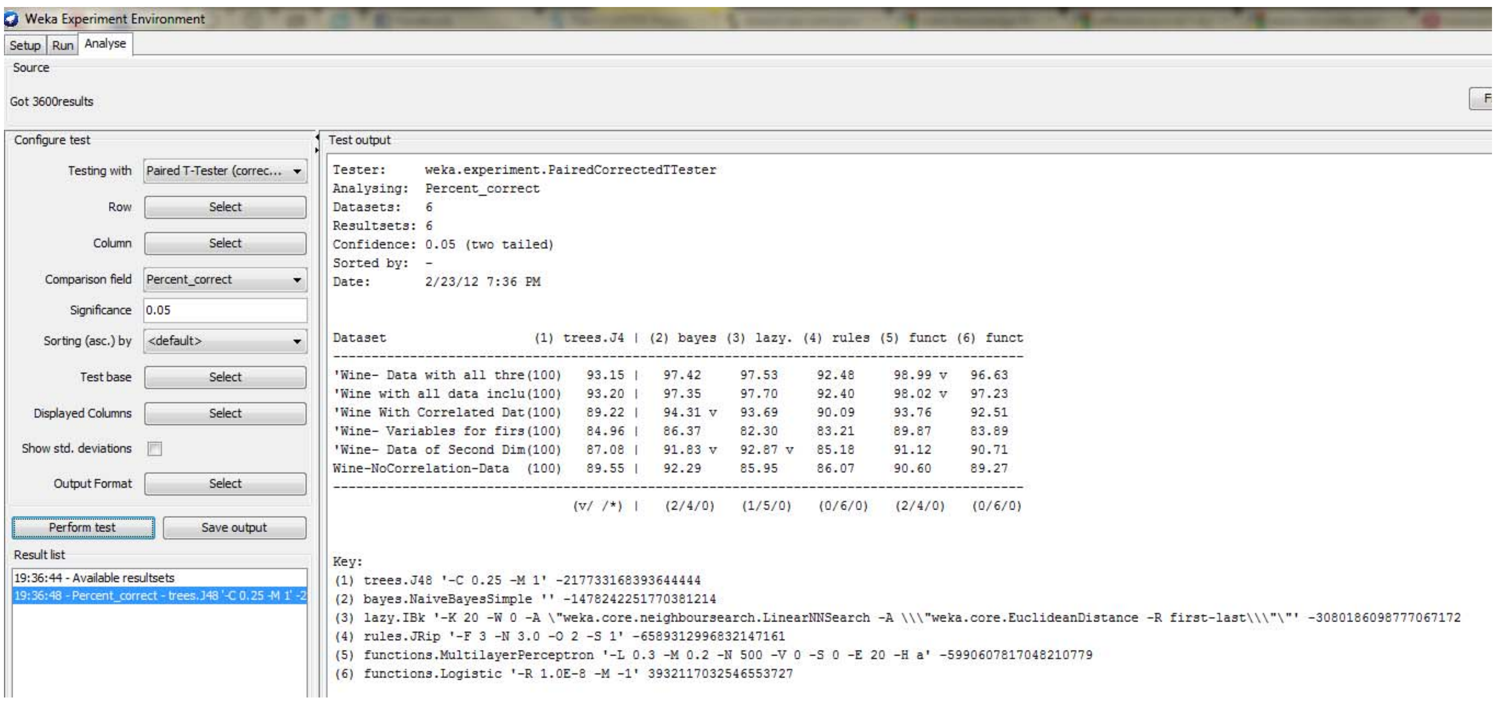
2. Methodology/process used

- After aggregating the results from the Weka Explorer module, I have decided to use the following six methods and corresponding parameters for the six datasets

Method	J48	NaiveBayes Simple	IBK	JRiP	MultiLayer Perceptron	Logistic Regression
Parameters	minNumObj=1, numFolds=10	Default	KNN=20	minNumObj=3, folds=3	Default	Default

2. Methodology/process used

- Using the above methods and parameters, I then ran the Weka Experimenter module. The results are as follows:



2. Methodology/process used

- Then I used the dataset “WineAllData” and decomposed the analytical process using Weka’s Knowledge Flow module. Using the same parameters from above, the results for the six different methods are as follows:

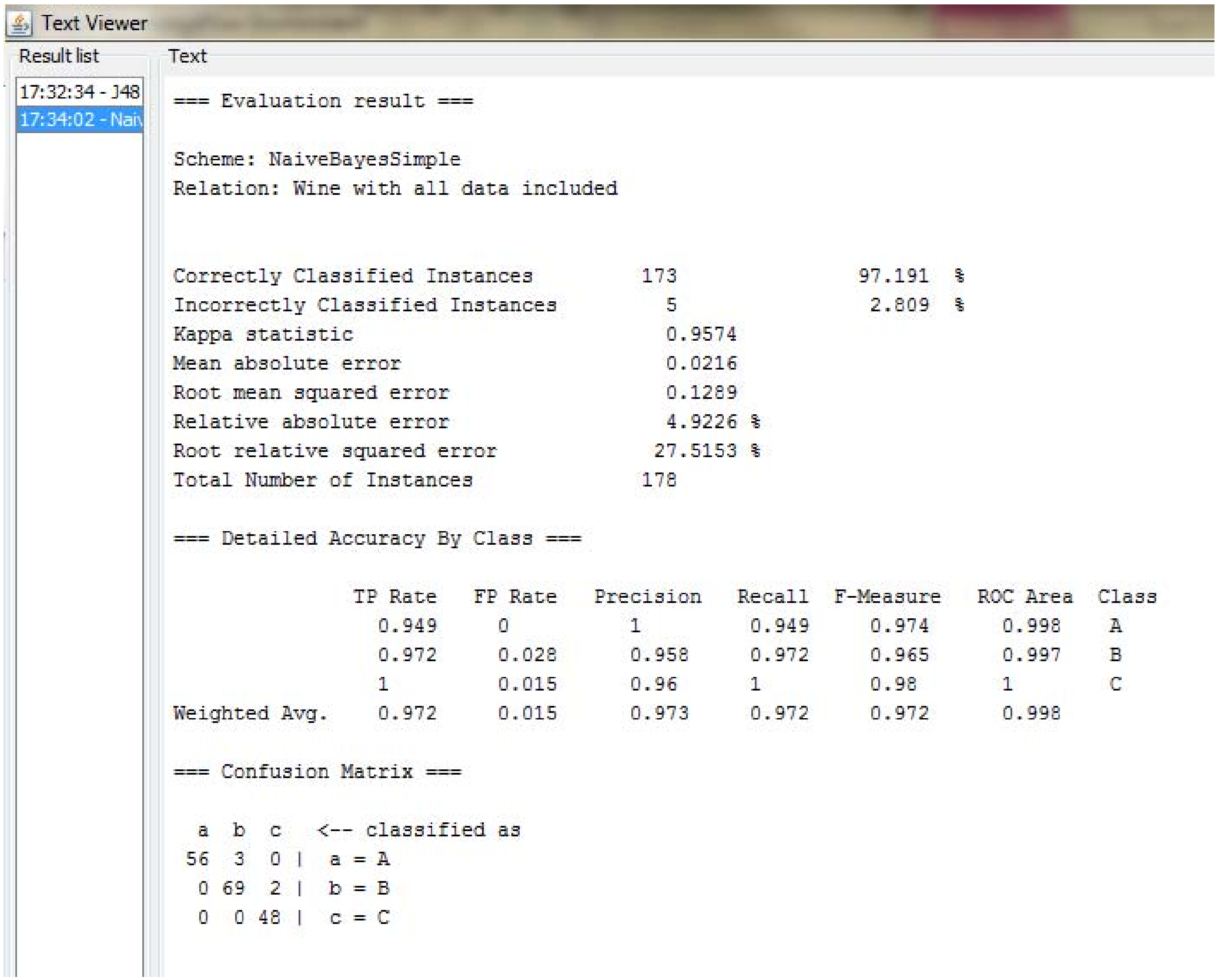
Text Viewer

Result list: 17:32:34 - J48

Text

```
=== Evaluation result ===

Scheme: J48
Options: -C 0.25 -M 1
Relation: Wine with all data included


Correctly Classified Instances         168               94.382  %
Incorrectly Classified Instances        10                5.618  %
Kappa statistic                          0.9144
Mean absolute error                      0.043
Root mean squared error                  0.1926
Relative absolute error                  9.7927 %
Root relative squared error             41.1165 %
Total Number of Instances              178

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.983     0.034      0.935     0.983     0.959      0.977    A
                 0.944     0.047      0.931     0.944     0.937      0.938    B
                 0.896     0.008      0.977     0.896     0.935      0.948    C
Weighted Avg.    0.944     0.032      0.945     0.944     0.944      0.954

=== Confusion Matrix ===

  a  b  c   <-- classified as
 58  1  0 |  a = A
  3 67  1 |  b = B
  1  4 43 |  c = C
```

2. Methodology/process used




```
=== Evaluation result ===

Scheme: NaiveBayesSimple
Relation: Wine with all data included


Correctly Classified Instances         173               97.191  %
Incorrectly Classified Instances         5                2.809  %
Kappa statistic                          0.9574
Mean absolute error                      0.0216
Root mean squared error                  0.1289
Relative absolute error                  4.9226 %
Root relative squared error             27.5153 %
Total Number of Instances              178

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.949     0           1         0.949     0.974      0.998    A
                 0.972     0.028       0.958     0.972     0.965      0.997    B
                 1         0.015       0.96      1         0.98       1        C
Weighted Avg.    0.972     0.015       0.973     0.972     0.972      0.998

=== Confusion Matrix ===

  a  b  c   <-- classified as
 56  3  0 |  a = A
  0 69  2 |  b = B
  0  0 48 |  c = C
```

2. Methodology/process used

- **J48 Algorithm Results:**
- The first algorithm that I tested was the J48 Decision Tree. A Decision Tree is a classification algorithm used for predicting group membership according to a set of rules.
- Decision Trees are simple to understand but are also simplistic in their classification because an object to be classified either falls into a class or not; there are no shades of gray (Tan, 2006). Some strengths and weaknesses of Decision Trees include:
  - Strengths
    - Easy to understand
    - Fast to learn
  - Weaknesses
    - Cannot give “partial credit” for belonging to more than one class
    - Prone to overfitting
    - Can need a lot of data to make and accurate prediction

2. Methodology/process used

- **Please note for every algorithm you use, provide a complete description for the algorithm you use:**
  - **You need to do through research and analysis for every algorithm**
    - **Its description and structure**
    - **Advantages and Disadvantages of the algorithm**
    - **What are the parameters you changed and why?**
    - **Its time complexity and space complexity**
    - **Efficient for large or small data sets (or neither)?**
    - **Document your research findings**
    - **Capture and document Weka screen-shots, experimental results**

2. Methodology/process used

- The first thing I did was to run the Weka experimenter with all data sets using the default values of the J48 algorithm. I will be looking to see if my predictions that Alcohol and Flavonoids would be good differentiator while Malic Acid and Color Intensity would not.

Dataset	(1) trees.J48
'Wine- Data with all thre(100)	93.14
'Wine with all data inclu(100)	**93.20**
'Wine With Correlated Dat(100)	88.60
'Wine- Variables for firs(100)	84.84
'Wine- Data of Second Dim(100)	87.25
Wine-NoCorrelation-Data (100)	89.55

2. Methodology/process used

- **Impact of Preprocessing on MLP Performance with the WineAllData Dataset**
  - I investigated the impact of preprocessing filters on MLP Performance with the WineAllData dataset and got the following results:
    - Baseline performance: 97.2% Correct
    - Supervised Attribute Discretize RFirst Last: 98.9%
    - Supervised Attribute Discretize RFirst Last Kononeko: 98.9%
  - The preprocessing filter did have a significant impact on the MLP performance.

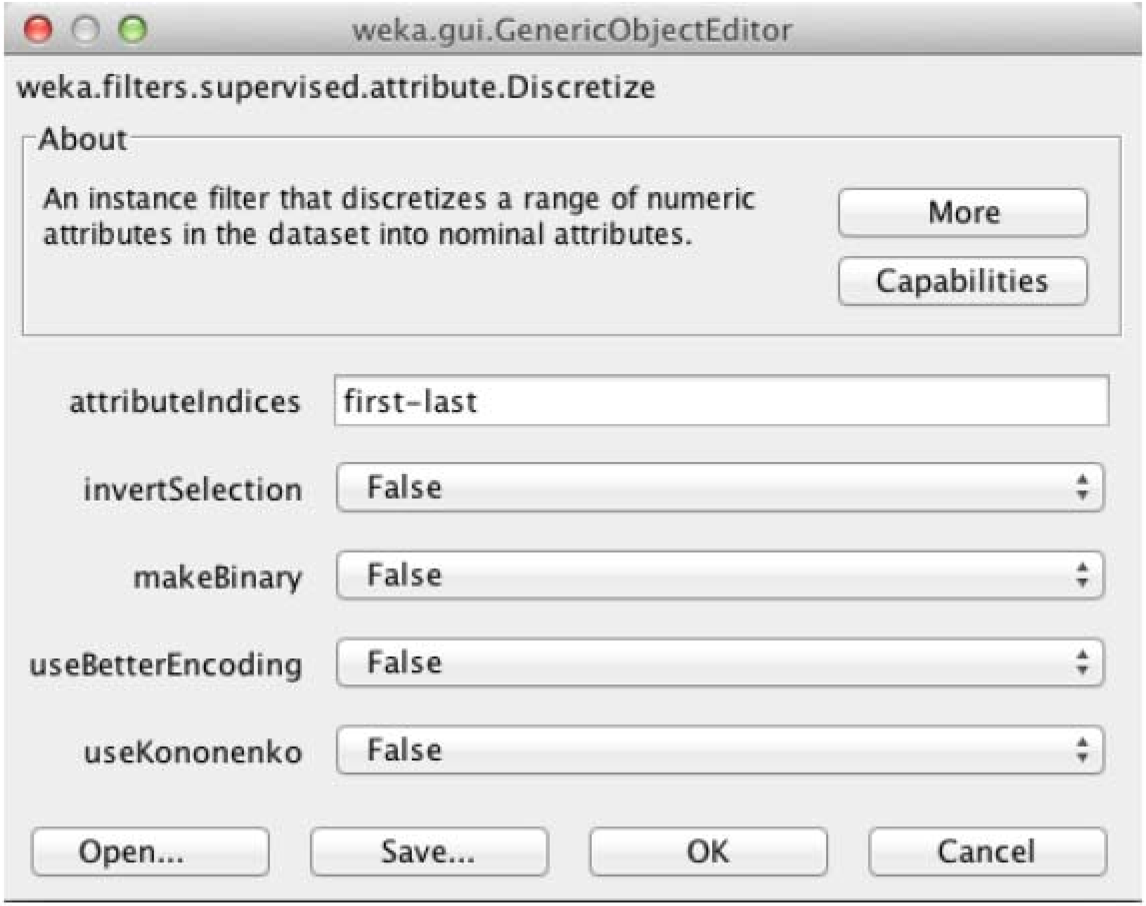
2. Methodology/process used

Baseline Run – No Preprocessing

```
=== Run information ===

Scheme:weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
Relation:     Wine with all data included
Instances:    178
Attributes:   14
              Alcohol
              Malic_Acid
              Ash
              Ash_Alcalinity
              Magnesium
              Total_Phenols
              Flavanoids
              Nonflavanoid_Phenols
              Proanthocyanins
              Color_Intensity
              Hue
              OD280_OD315
              Proline
              Type
Test mode:10-fold cross-validation

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

Correctly Classified Instances         173               97.191  %
Incorrectly Classified Instances         5                2.809  %
Kappa statistic                          0.9574
Mean absolute error                      0.0247
Root mean squared error                  0.1172
Relative absolute error                  5.6355 %
Root relative squared error             25.0058 %
Total Number of Instances              178
```

2. Methodology/process used

Preprocessing Filter Applied to WineAllData Dataset Used With MLP Classifier



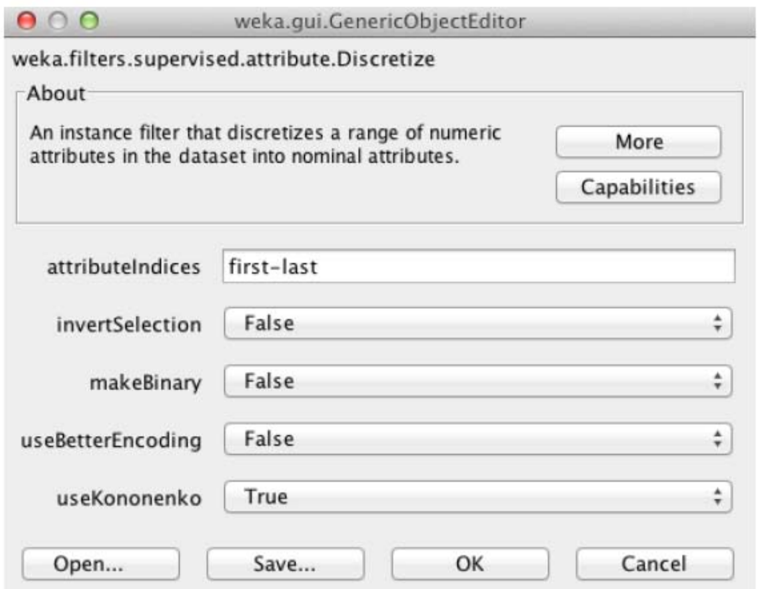
2. Methodology/process used

Preprocessing Filter Applied to WineAllData Dataset Used With MLP Classifier

```
=== Run information ===

Scheme:weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
Relation:     Wine with all data included-weka.filters.supervised.attribute.Discretize-Rfirst-
last
Instances:    178
Attributes:   14
              Alcohol
              Malic_Acid
              Ash
              Ash_Alcalinity
              Magnesium
              Total_Phenols
              Flavanoids
              Nonflavanoid_Phenols
              Proanthocyanins
              Color_Intensity
              Hue
              OD280_OD315
              Proline
              Type
Test mode:10-fold cross-validation


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

Correctly Classified Instances         176               98.8764 %
Incorrectly Classified Instances         2                1.1236 %
Kappa statistic                          0.983
Mean absolute error                      0.0131
Root mean squared error                  0.0824
Relative absolute error                  2.9838 %
Root relative squared error             17.5862 %
Total Number of Instances              178
```

2. Methodology/process used

Preprocessing Filter Applied to WineAllData Dataset Used With MLP Classifier



2. Methodology/process used

```
=== Run information ===

Scheme:weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
Relation:     Wine with all data included-weka.filters.supervised.attribute.Discretize-Rfirst-last-
weka.filters.supervised.attribute.Discretize-K-Rfirst-last
Instances:    178
Attributes:   14
              Alcohol
              Malic_Acid
              Ash
              Ash_Alcalinity
              Magnesium
              Total_Phenols
              Flavanoids
              Nonflavanoid_Phenols
              Proanthocyanins
              Color_Intensity
              Hue
              OD280_OD315
              Proline
              Type
Test mode:10-fold cross-validation


Time taken to build model: 1.54 seconds

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

Correctly Classified Instances         176               98.8764 %
Incorrectly Classified Instances         2                1.1236 %
Kappa statistic                          0.983
Mean absolute error                      0.0131
Root mean squared error                  0.0824
Relative absolute error                  2.9838 %
Root relative squared error             17.5862 %
Total Number of Instances              178

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.008      0.983     1         0.992      1        A
                 0.972     0          1         0.972     0.986      1        B
                 1         0.008      0.98      1         0.99       1        C
Weighted Avg.    0.989     0.005      0.989     0.989     0.989      1

=== Confusion Matrix ===
```

3. Comparative Analysis

- I analyzed the results to see what factors seem to drive the solutions offered. Color Intensity, Flavanoids and Proline were identified as key predictors in 4 out of 5 algorithms.

- Most influential attributes for the different algorithms:

J48	JRip	Naïve Bayes	MLP	Simple Log Regression
Flavanoids	Flavanoids	Proline	Color Intensity	Hue
Color Intensity	Color Intensity	Magnesium	Proline	Flavanoids
Proline	OD280 OD315	Ash Alcalinity	Flavanoids	Ash
	Proline	Color Intensity	Ash Alcalinity	OD280 OD315
			Hue	Nonflavanid Phenols
			OD280 OD315	Alcohol

3. Comparative Analysis

Dataset	Result			Algorithm	
	Best	Worst	Spread	Best	Worst
WineAllData	98.02	93.14	4.88	MLP	JRip
WineAll3Dim	98.99	92.82	6.17	MLP	JRip
WineCorre	94.42	89.22	5.20	Naïve Bayes	J48
WineNoCorre	92.29	85.68	6.61	Naïve Bayes	JRip
WineDim1	89.87	83.32	6.55	MLP	JRip
WineDim2	91.89	84.92	6.97	Naïve Bayes	JRip

3. Comparative Analysis

- I then ranked the performance, based on Percent Correct, of the classification methods on each datasets in the following table:

Dataset	Performance Rank				
	1st	2nd	3rd	4th	5th
WineAllData	MLP	NB	LR	J48	JRIP
WineAll3Dim	MLP	NB	LR	J48	JRIP
WineCorre	NB	MLP	LR	JRIP	J48
WineDim1	MLP	NB	J48	LR	JRIP
WineDim2	NB	MLP	LR	J48	JRIP
WineNoCorre	NB	MLP	J48	LR	JRIP

3. Comparative Analysis

- From the experimental results for the different algorithms that I analyzed, I found the following:
  1. Larger data sets seemed to perform better. Although there is the danger of overtraining with data sets that are too large, it seems that is the case with under 200 lines of data, more is better.
  2. With correlated data, algorithms designed to work with nonlinear data, such as the ANN and Logistic Regression outperformed the other algorithms. The Naïve Bayes algorithm, once I removed some of the correlation, did very well too even though it relies on independence of data.

4. Conclusion

- In conclusion, I found that data mining and machine learning algorithms are complex and unpredictable.
  - Sometimes I can do some simple analysis, such as looking at each individual attribute, and make some good predictions. Other times, my predictions will turn out to be wrong.
- Based on my initial visual analysis, I had predicted the following:
  - Flavanoids and Proline would be good predictor variables
  - Malic Acid and Color Intensity would be poor predictor variables
- Once I had completed my analysis on influential variables, I discovered the following:
    - Flavanoids, Proline and Color Intensity were good predictor variables
    - Malic Acid was a poor predictor variable
- Sometimes algorithms that should perform poorly perform well. But you can experiment with different configuration and find the best solution to solve your problem. What I found challenging is to be able to select the right algorithm for the right data. You can’t expect algorithms designed for independent data to work well with data that shows correlations.

5. References

- Tan P-N, Steinbach M, Kumar V (2006) Introduction to data mining. Pearson Addison-Wesley
- http://www.cs.waikato.ac.nz/ml/weka/
- http://wiki.pentaho.com/display/DATAMINING/JRip
- http://wiki.pentaho.com/display/DATAMINING/J48
- http://wiki.pentaho.com/display/DATAMINING/MultilayerPerceptron
- http://wiki.pentaho.com/display/DATAMINING/IBk