Abstract

The goal of the project is to classify a dataset of kernels into three different varieties of wheat: Kama, Rosa, and Canadian. The dataset contains 210 instances of data points. Each data point is described by 7 attributes, which are:

1. Area A
2. Perimeter P
3. Compactness C = 4\*pi\*A/P^2
4. length of kernel,
5. width of kernel,
6. asymmetry coefficient
7. length of kernel groove.

Our strategy is to first use K- Nearest Neighbor to try and classify this dataset. We will use various values of K to find the best k-value that works for this particular dataset. Moreover we will be finding optimal settings related to KNN by experimenting. Next, we will use the Multilayer perceptron algorithm to classify the dataset. For MLP, we will be varying the learning rate value to find an optimal value that works for this dataset. We will also be tweaking finer settings related to MLP and observe how well they all perform in comparison with one another.

In regards to the results, for both KNN and MLP, we will study the performance and error percentages to determine how well the algorithms performed against each other and against different settings when compared against themselves.

Dataset Description:

1. Total Instances: 210 samples of wheat kernels
2. Class Distribution: 70 instances each of Kama, Rosa, and Canadian wheat varieties
3. Attributes: 7 geometric features (all real-valued, continuous)
4. Data Quality: No missing values or anomalies found in the dataset

Methodology

Cross-validation Strategy:

We employed stratified 10-fold cross-validation for both KNN and MLP algorithms. This means:

1. The dataset (210 instances) is divided into 10 equal folds of 21 instances each
2. The stratification ensures each fold maintains the same proportion of wheat varieties
3. In each iteration:
   1. One fold (21 instances) serves as the test set
   2. Remaining nine folds (189 instances) serve as the training set
4. Process repeats 10 times with different test folds
5. Final results are averaged across all iterations

This stratified approach ensures balanced representation of all wheat varieties in both training and testing, providing more reliable performance estimates.

Data Analysis and Results

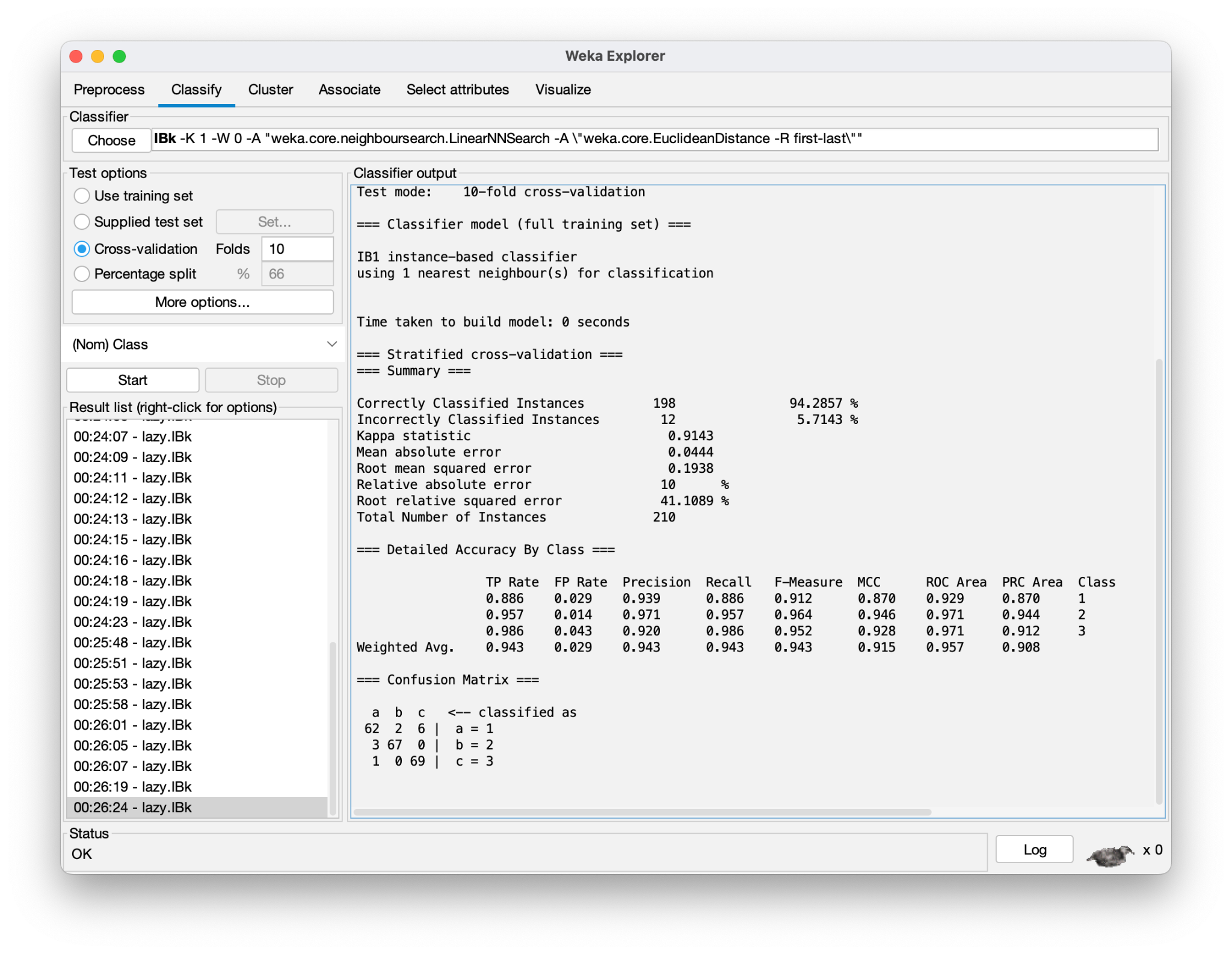
**KNN** algorithm

Since the 7 attributes are quite varied in terms of the value. I will try KNN algorithm first on standardized data (sets the mean at 0):

K = 1 Configuration:

1. Preprocessing: Standardization (zero mean)
2. Distance metric: Euclidean
3. Cross-validation: 10-fold

K = 1 Results:



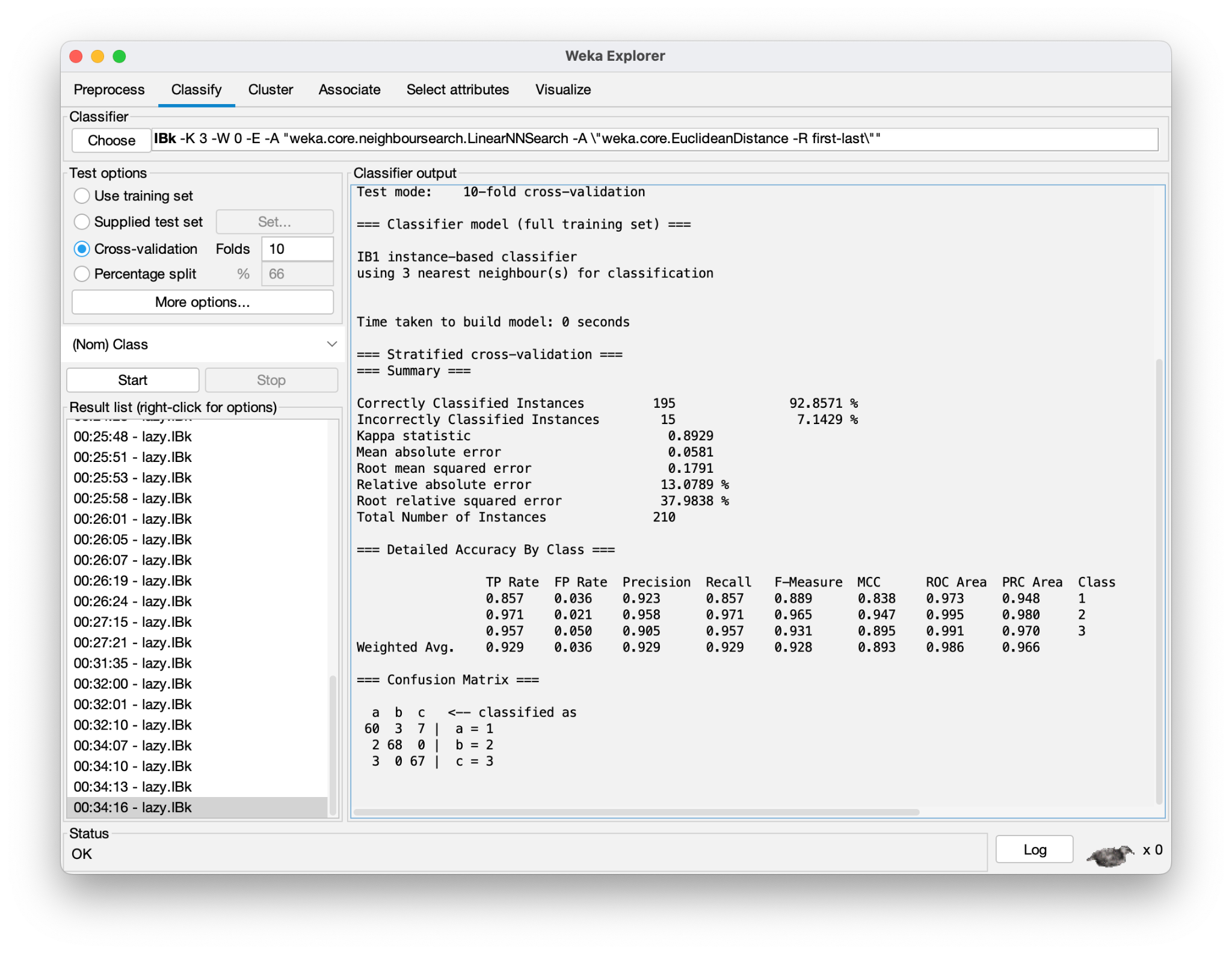
K = 1 Observation:

1. Accuracy 94.2857 % (highest achieved)
2. Confusion Matrix Analysis
   1. Strong performance of Canadian Wheat - only 1 missclassification
   2. Kama Wheat did the worst.

K = 3 Configuration:

1. Same as K=1

K = 3 Results:



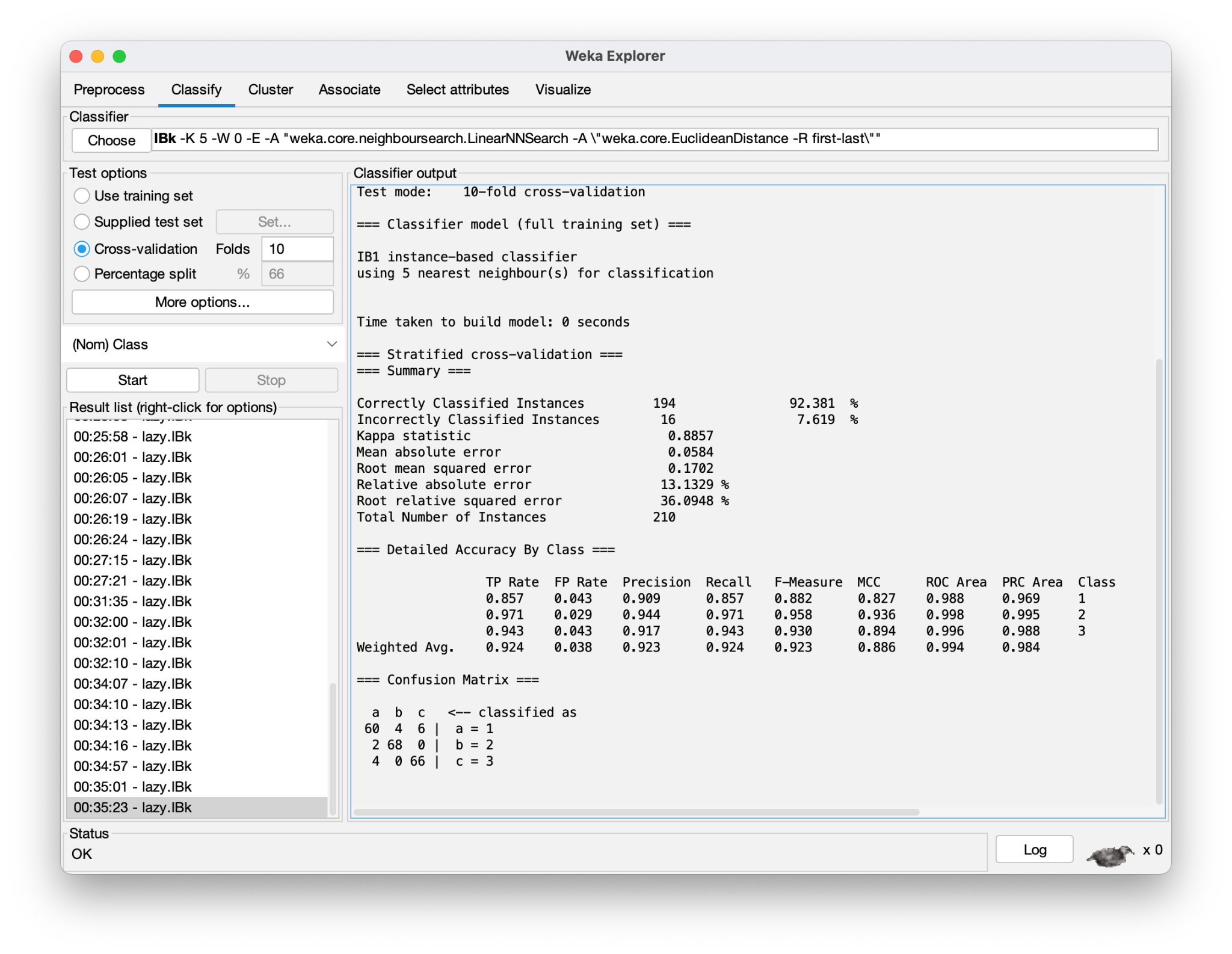
K = 3 Observations:

1. The Accuracy went down to 92.85%
2. Interestingly Rosa performed the best looking a the confusion matrix

K = 5 Configuration;

1. Same as K = 1

K= 5 Results



K = 5 Observations

1. Accuracy went down to 92.381 %

Will stop increasing the K. From my observations K = 1 generated the most accurate predictions with lowest Relative absolute error %.

KNN Key Findings:

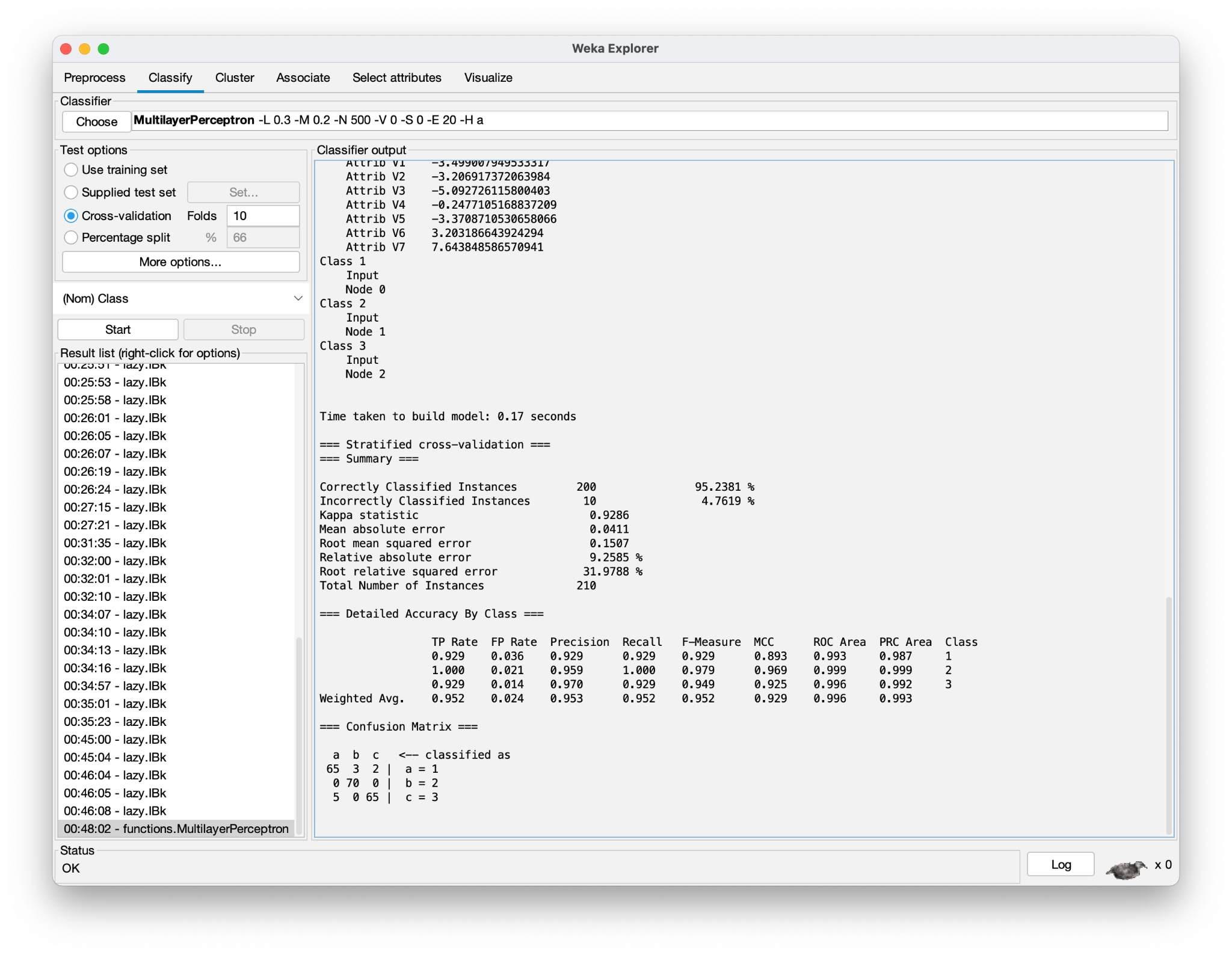
1. K=1 performed best with 94.2857% accuracy
2. Performance degraded with increasing K values
3. Distance weighting showed no significant impact
4. Rosa and Canadian consistently showed better classification than Kama
5. Using Normalization produced the same results as Standardized dataset
6. Neither Normalizing nor standardizing produced any difference when compared with un processed data

**Multilayer Perceptron (MLP) algorithm**

MLP Default Configuration:

1. Preprocessing: Standardization (zero mean)
2. learningRate: 0.3
3. momentum: 0.2
4. hiddenLayer: 'a'
5. trainingTime: 500
6. Cross-validation: 10-fold

MLP Default Config Result:



MLP Default Observations:   
Accuracy: 95.2381% (higher than KNN's best performance)

1. Build Time: 0.17 seconds
2. Confusion Matrix Analysis:
   * Perfect classification for Rosa wheat (70/70 correct)
   * Equal performance for Kama and Canadian (both 92.9% accurate)
   * Kama had 5 misclassifications (3 as Rosa, 2 as Canadian)
   * Canadian had 5 misclassifications (all as Kama)
   * Rosa showed no misclassifications in either direction

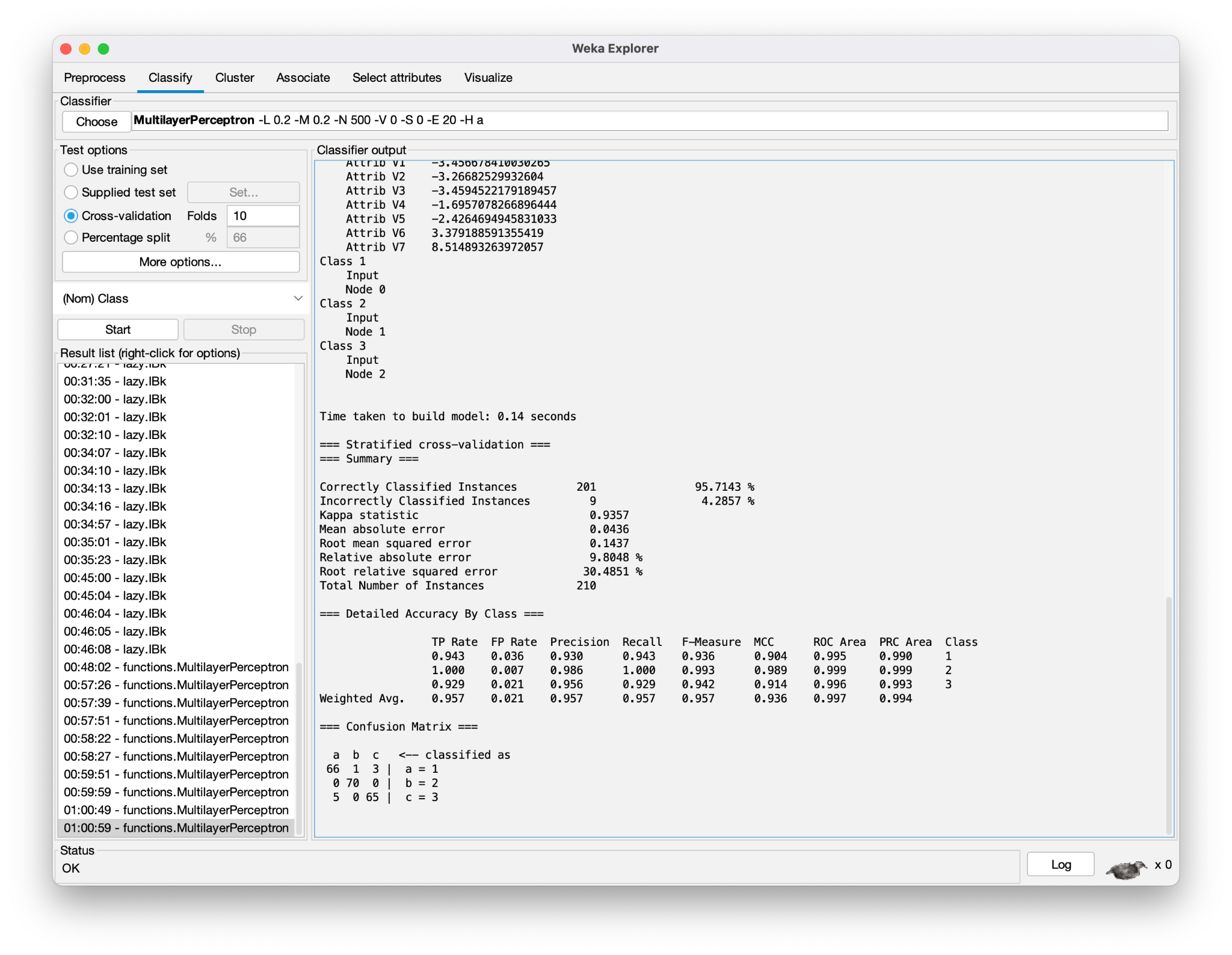
Varying Learning Rate

Keeping all the configs same as before, lets update Learning rate to be 0.1, 0.2, 0.4, and 0.5

MLP Learning rate = 0.1

Result: Accuracy = 95.231 % (same)

MLP Learning rate = 0.2

Learning rate 0.2 result  


Observation: Slight increase in accuracy when compared with 0.3. Accuracy went up to 95.7143% (0.2) compared to 95.2381% (0.3)

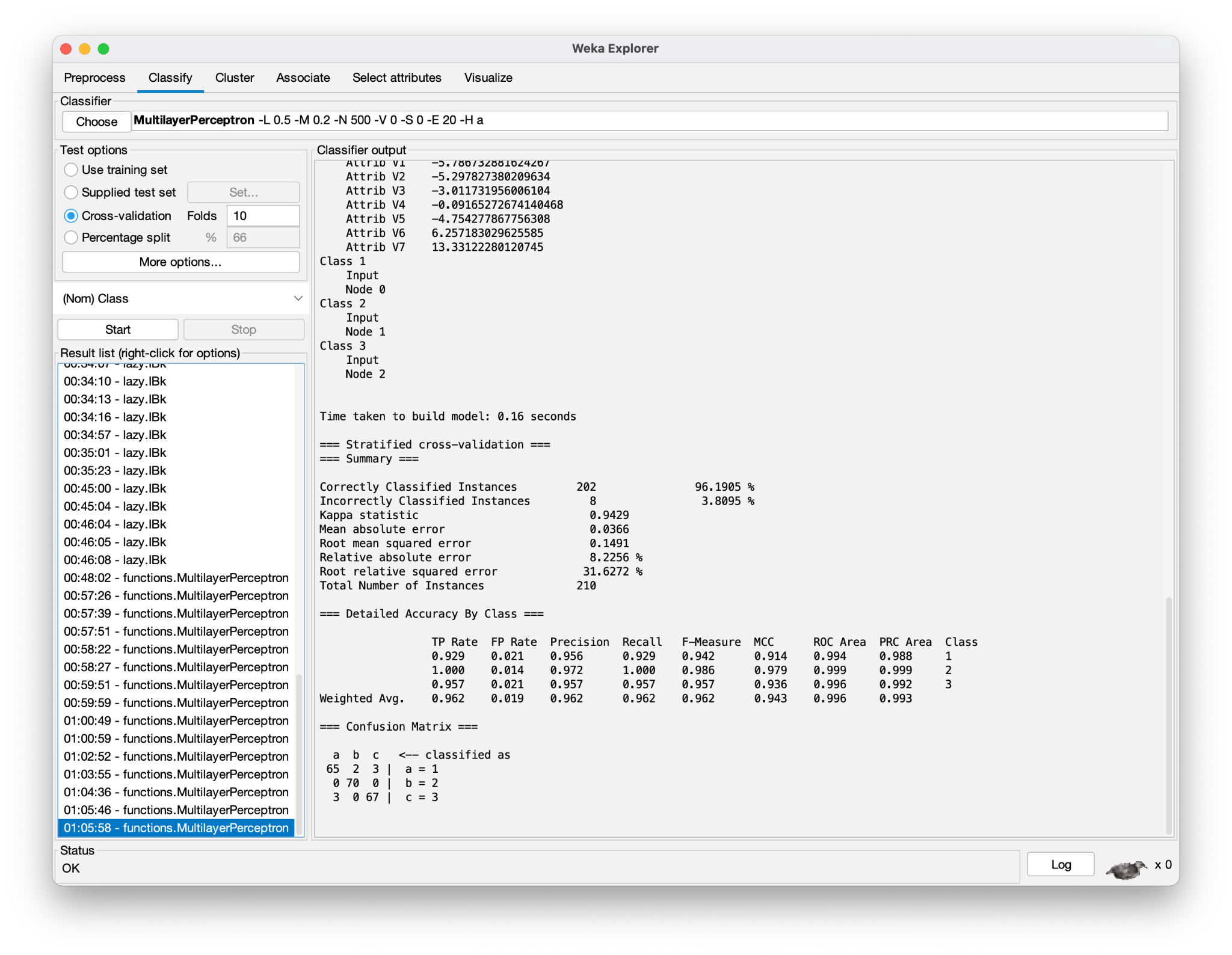
MLP Learning rate = 0.4

Accuracy remains the same as 0.2 - 95.7143% (highest yet on par with learning rate of 0.2)

MLP Learning rate = 0.5 (**BEST**)

Accuracy went up to 96.1905% (highest yet)

MLP Learning rate = 0.5 Result screenshot:



MLP Learning rate = 0.6

Accuracy went down to 95.2381%

Lets tweak **Momentum** now

Currently momentum is set at 0.2

Trying with Momentum at 0.1

Accuracy went down to 94.7619 %

Momentum = 0.3

Accuracy went up to 95.713% (still lower than momentum at 0.2)

Let’s leave momentum at 0.2. And noodle around with the Hidden layer

Currently Hidden layer is set to ‘A’ auto

Updating Hidden layer = 3

Accuracy went up to 96.6667%

Hidden layer = 4

Accuracy went down to 94.7619 %

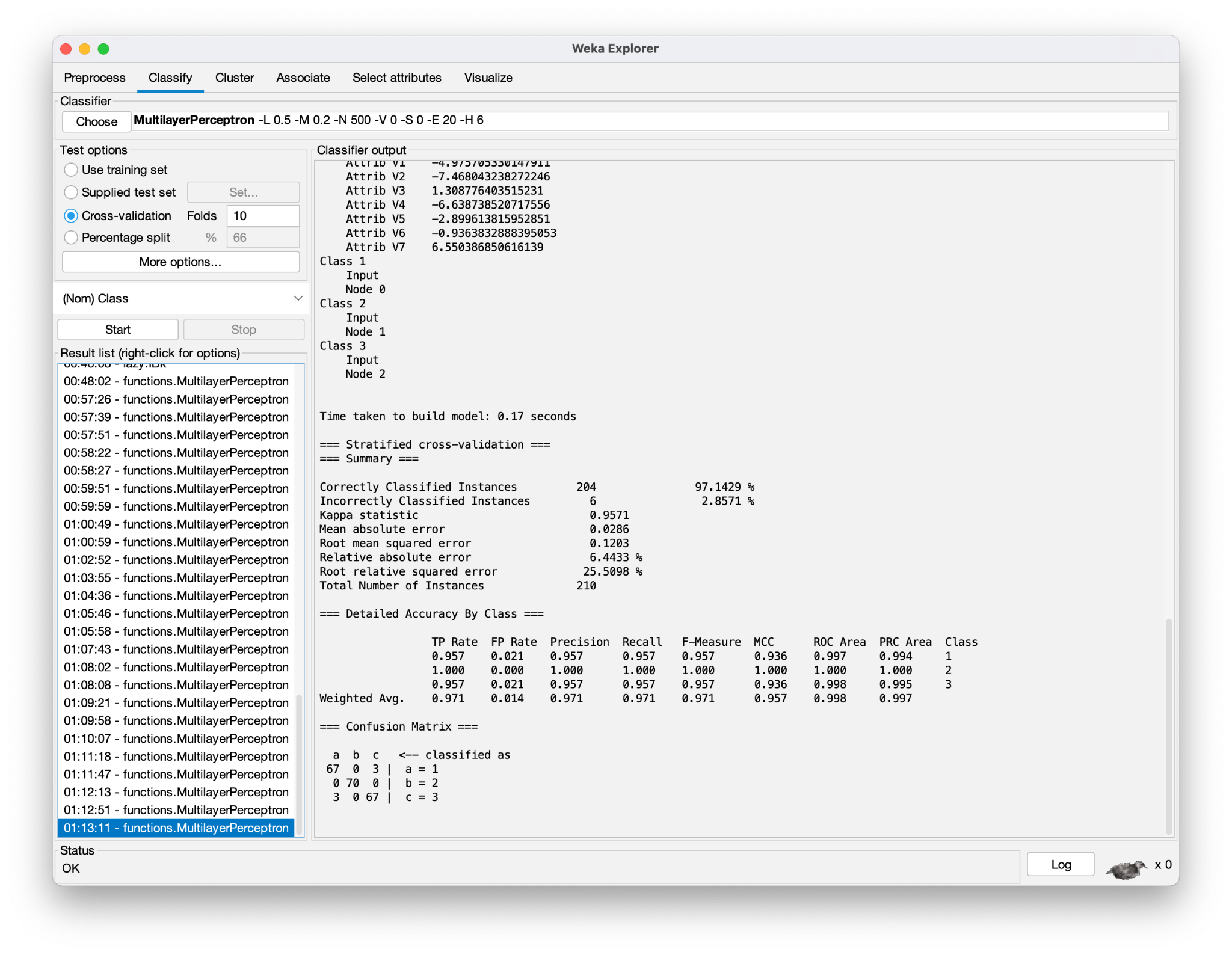
Hidden layer = 5

Accuracy is 96.1905 %

Hidden layer = 6

Accuracy went up to 97.1429 % (**highest**)

Results:

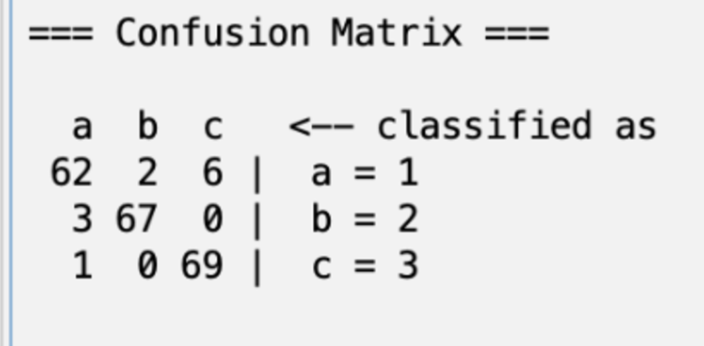


Hidden layer = 7

Accuracy went down to 96.6667 %

Confusion Matrix comparison

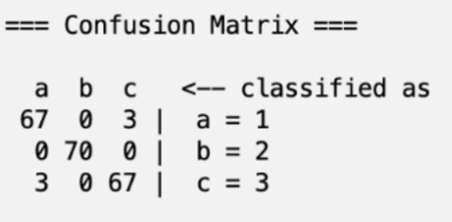
Best KNN (K=1) Confusion Matrix:



Analysis:

1. Canadian wheat had best classification (only 1 misclassification)
2. Kama wheat showed most misclassifications (8 misclassification)
3. Total accuracy: 94.2857%

Best MLP (6 hidden layers) Confusion Matrix:



Analysis:

1. Rosa wheat achieved perfect classification
2. Both Kama and Canadian had only 3 misclassifications each
3. Total accuracy: 97.1429%

Confusion matrix Key Differences:

1. MLP showed more balanced performance across classes
2. KNN struggled more with Kama wheat classification
3. MLP achieved better overall accuracy with fewer misclassifications

Error Percentage Comparison:

KNN (Best Configuration - K=1):

1. Mean absolute error: 0.0581
2. Root mean squared error: 0.1938
3. Relative absolute error: 10%
4. Root relative squared error: 41.1089%

MLP (Best Configuration - 6 hidden layers):

1. Mean absolute error: 0.0286
2. Root mean squared error: 0.1203
3. Relative absolute error: 6.4433%
4. Root relative squared error: 25.5098%

MLP vs KNN Analysis:

1. MLP demonstrates significantly lower error rates across all metrics
2. The mean absolute error is approximately halved with MLP (0.0286) compared to KNN (0.0581)
3. MLP shows a lower root mean squared error (0.1203 vs 0.1938), indicating more consistent predictions
4. The relative absolute error for MLP (6.4433%) is lower than KNN (10%)
5. Most significantly, MLP's root relative squared error (25.5098%) is considerably lower than KNN's (41.1089%), which means better predictability

MLP Conclusion

1. Optimal MLP Config
   1. Learning rate: 0.5
   2. Momentum: 0.2
   3. Hidden Layers: 6
2. Best accuracy achieved: 97.1429%
   1. Which outperformed
      1. Default MLP (95.2381%)
      2. Best KNN Result (94.2857%)
3. Tweaking the hidden layer had the most impact on improving accuracy

Overall Conclusion

* Project Goal Achievement:
  + Successfully classified three wheat varieties using both KNN and MLP algorithms
  + Systematically tested different parameters for optimal performance
  + Achieved high classification accuracy with both methods
* Best Configurations:
  + KNN: K=1 with 94.2857% accuracy
  + MLP: 6 hidden layers, learning rate 0.5, momentum 0.2 with 97.1429% accuracy
* Performance Comparison:
  + MLP achieved better classification accuracy and lower error rates
  + KNN demonstrated superior computational efficiency (~0 (in Weka it gets reported as 0, so I assume it’s a very low value < 100 ms) seconds vs 0.17 seconds)
  + More consistent predictions with MLP (lower error rates)
  + Better handling of wheat varieties with MLP
  + Error rates halved when using MLP

Recommendation:

* Choice depends on priority:
  + For highest accuracy: MLP
  + For fastest computation: KNN
  + Both algorithms achieved >94% accuracy, making either viable depending on specific needs