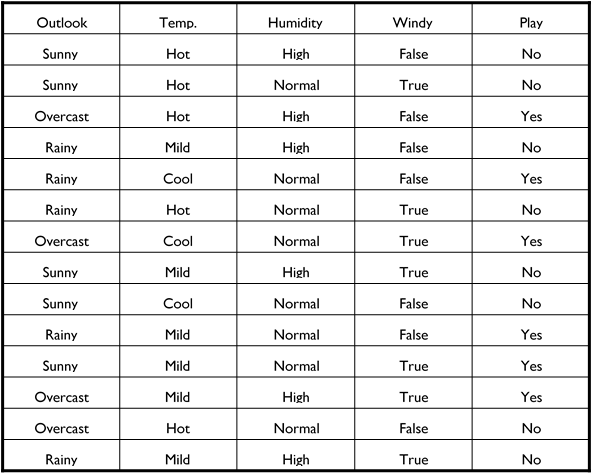
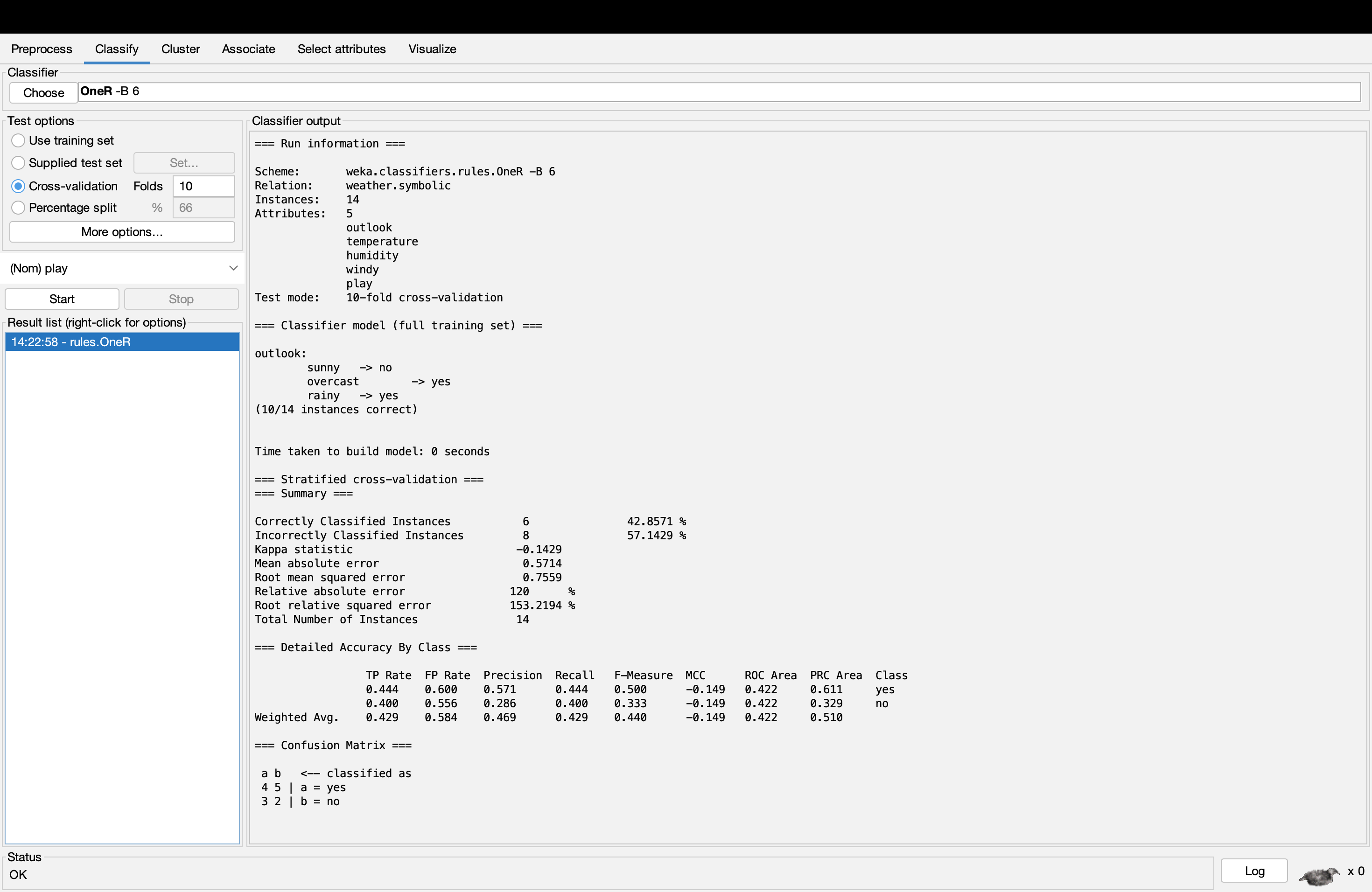
Practical 7

**Classification Review: 1R, Naïve Bayes, Decision Trees**

# Review Questions

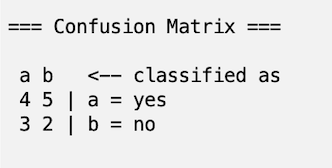
1. Compute the 1-R classification on the following dataset. Based on the classification result, derive a contingency table and compute sensitivity, specificity, recall, precision, FP-rate and TP-rate.





Based on the classification result, derive a contingency table and compute sensitivity, specificity, recall, precision, FP-rate and TP-rate.

Contingency Table



Sensitivity = TP / (TP + FN) = 4 / 4 +2 = 0.667

Specificity = TN / (TN + FP) = 5 / 3 + 5 = 0.625

Recall = 0.444, 0.400, 0.429

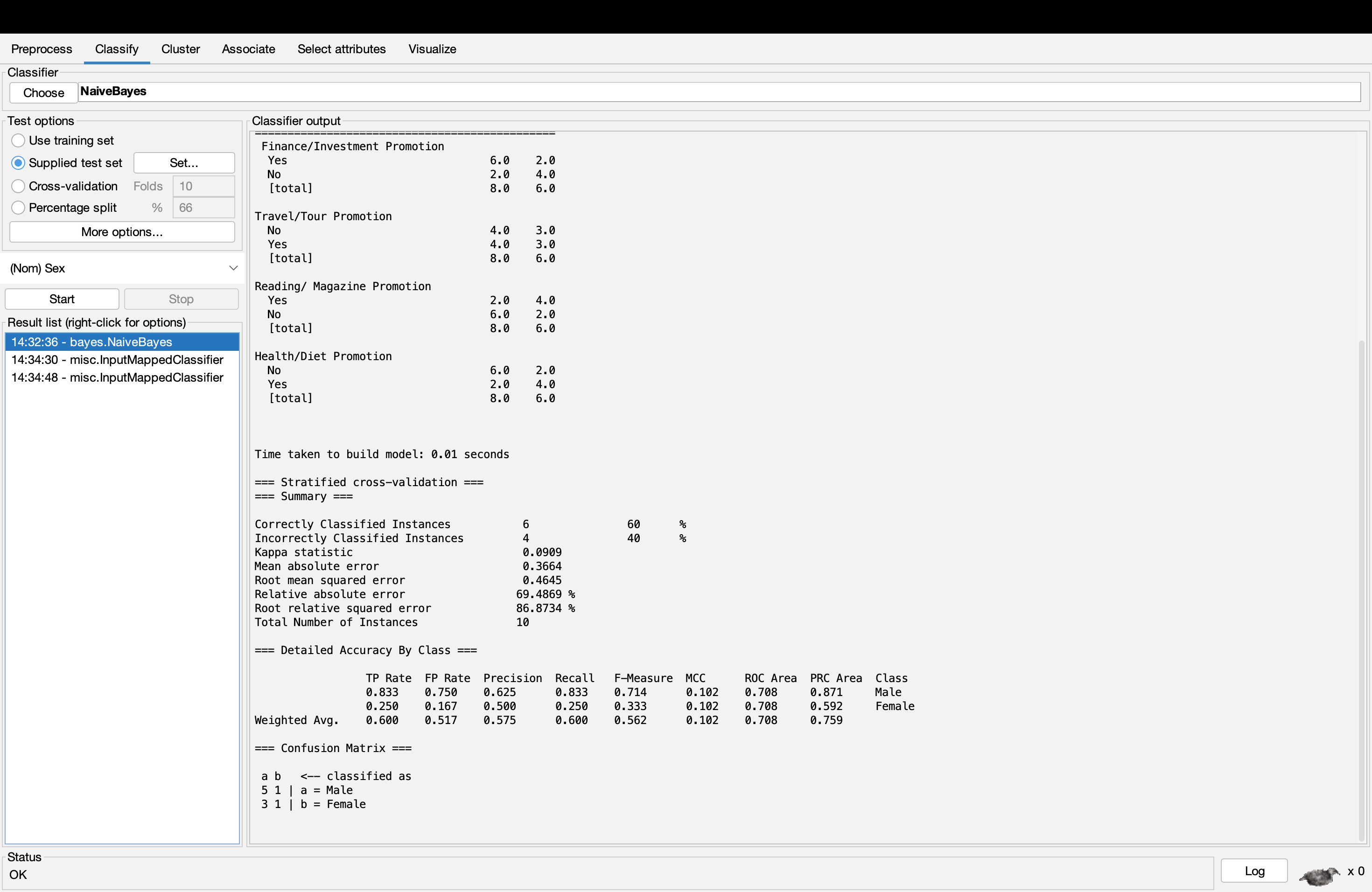
Precision = 0.571, 0.286, 0.469

FP-rate = 0.600, 0.286, 0.469

TP-rate = 0.444, 0.400, 0.429

1. Suppose we have data on a few individuals randomly surveyed. The following table gives the data on these individuals as to their response showing interest to promotional offers made in areas of Finance/Investment, Travel/Tour, Reading/Magazines, and Health/Diet. For this example, we use Sex as the output attribute whose value is to be predicted.

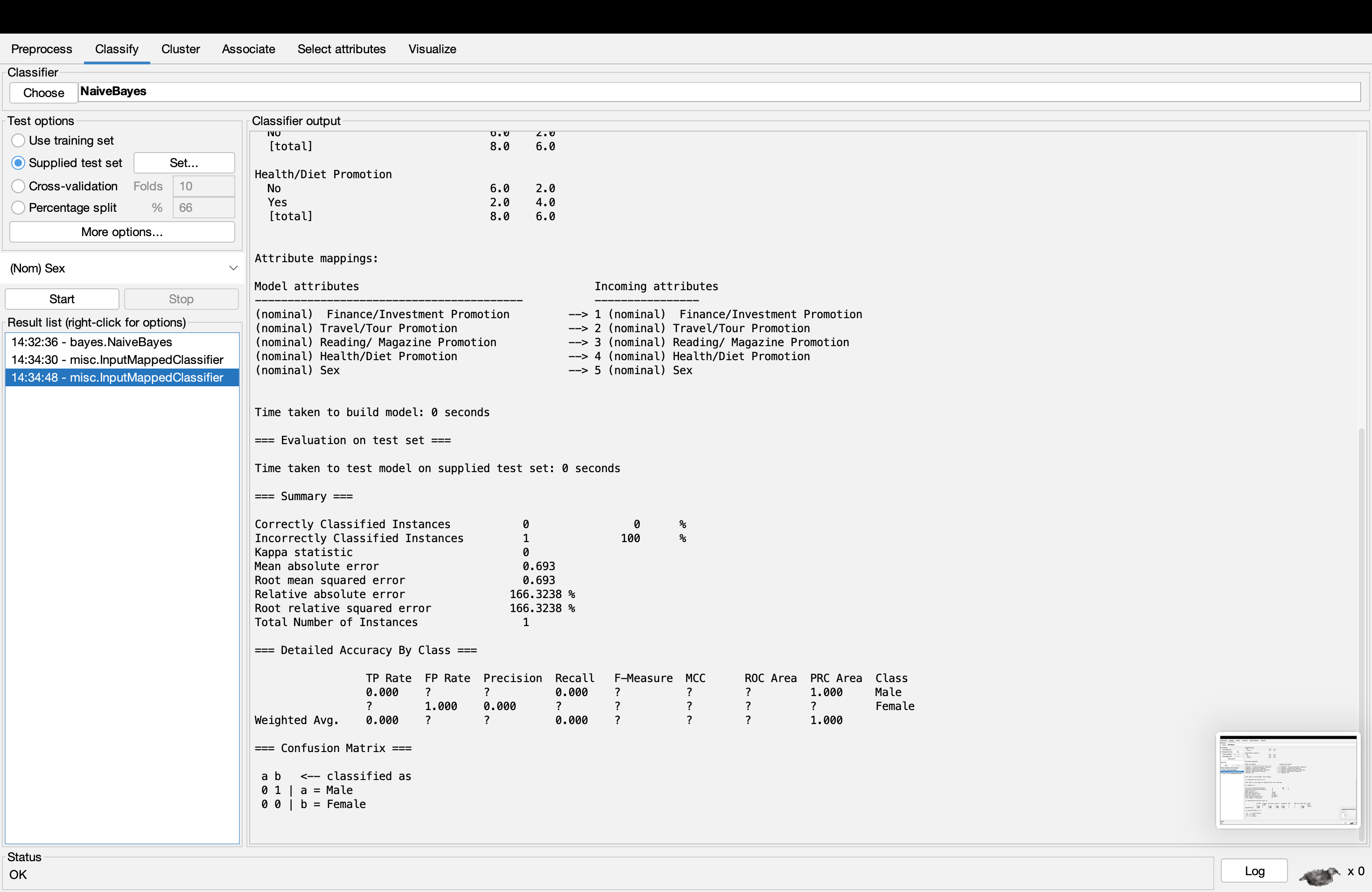
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Finance/Investment Promotion | Travel/Tour Promotion | Reading/Magazine Promotion | Health/Diet Promotion | Sex |
| Yes | No | Yes | No | Male |
| Yes | Yes | No | No | Male |
| No | Yes | Yes | Yes | Female |
| No | Yes | No | Yes | Male |
| Yes | Yes | Yes | Yes | Female |
| No | No | Yes | No | Female |
| Yes | No | No | No | Male |
| Yes | Yes | No | No | Male |
| No | No | No | Yes | Female |
| Yes | No | No | No | Male |



Use the data together with the Naïve Bayes classifier to perform a new classification for the following new instance:

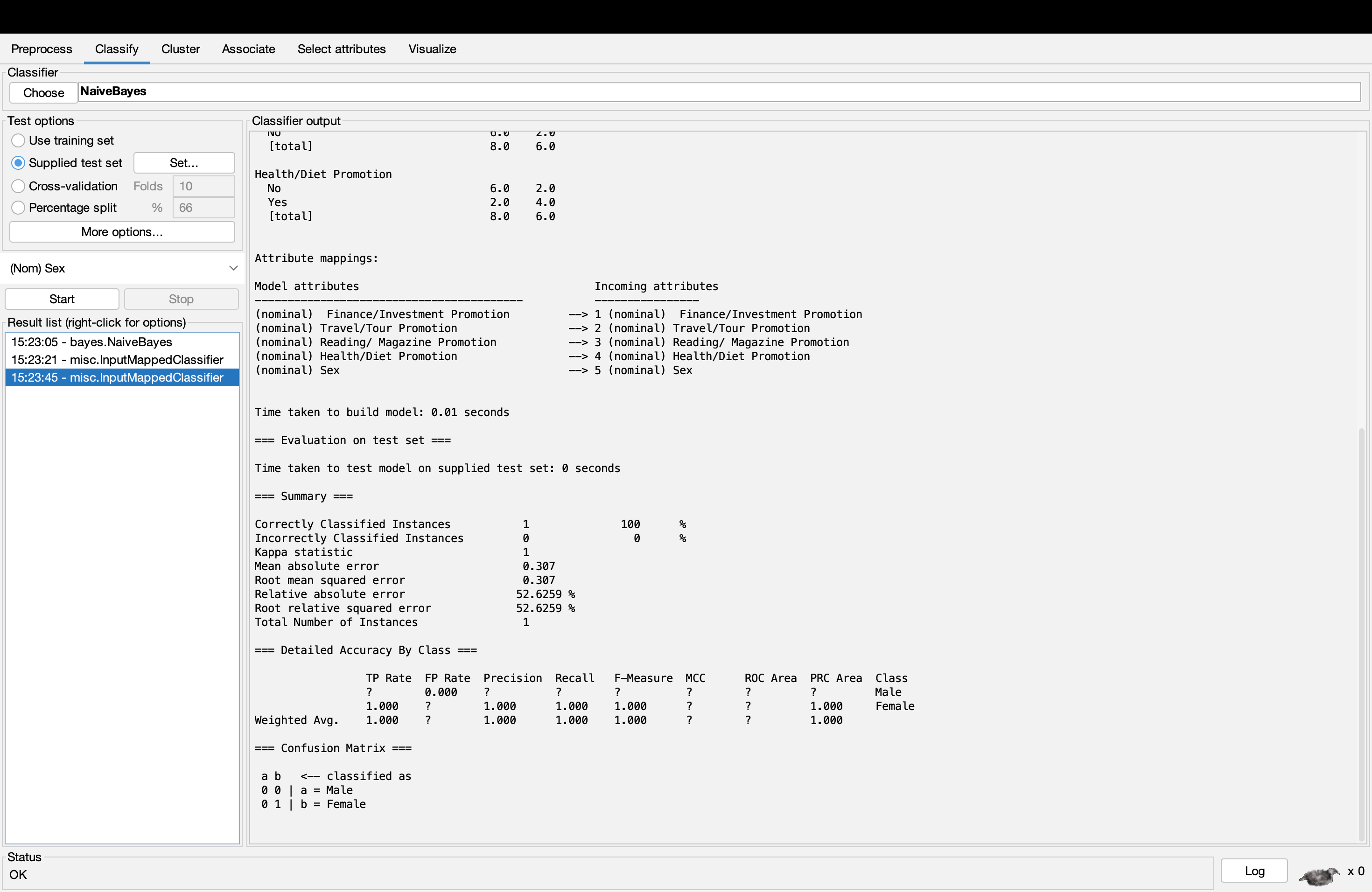
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Finance/Investment Promotion | Travel/Tour Promotion | Reading/Magazine Promotion | Health/Diet Promotion | Sex |
| **No** | **Yes** | **Yes** | **No** | **MALE** |

When trying out MALE as the sex of the individual the Precision were error and 0. Therefore, the sex of the individual cannot be Male or the amount of data is not sufficient.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Finance/Investment Promotion | Travel/Tour Promotion | Reading/Magazine Promotion | Health/Diet Promotion | Sex |
| **No** | **Yes** | **Yes** | **No** | **FEMALE** |

When trying out as the sex of the individual as FEMALE, there were still error but the Precision were 1 which means it is accurate in some places. Therefore, the sex of the individual is FEMALE. To be able to find out the real sex of the individual accurately, we will need more data than current data.



1. List and justify major strengths and weaknesses of Naïve Bayes’ Algorithm.

**Strengths of Naïve Bayes’ Algorithm**

Naïve Bayes’ Algorithm is simple algorithm which can be easily used by anyone.

This algorithm requires only minimal data and compute an accurate classification result.

The algorithm processing rate is incredibly fast even when performing with high dataset.

The algorithm performs well with even irrelevant features.

**Weaknesses of Naïve Bayes’ Algorithm**

Naïve Bayes’ Algorithm is not able to handle continuous data or features.

This Algorithm does not able to perform with missing data even though it performs well with irrelevant features.

When it comes to binary and categorical data, it cannot handle much data and limit them.

The Naïve Bayes’ Algorithm can suffer from overfitting when the data is small or it is too complex.

1. List and justify major strengths and weaknesses of decision tree algorithm.

**Strengths of Decision Tree Algorithm**

Decision tree algorithm is easy to understand and it can be easily explained.

This algorithm can handle both categorical and continuous data.

The Decision Tree Algorithm can handle missing data as well.

The algorithm can perform well when facing with noisy data and it has lesser chance when it comes to overfitting.

**Weaknesses of Decision Tree Algorithm**

The Decision Tree Algorithm does not work well with small dataset.

The production of tree may be differed when it comes to same data but different subsets.

When it comes to huge data, the algorithm may produce a very complex tree that are very difficult to understand.

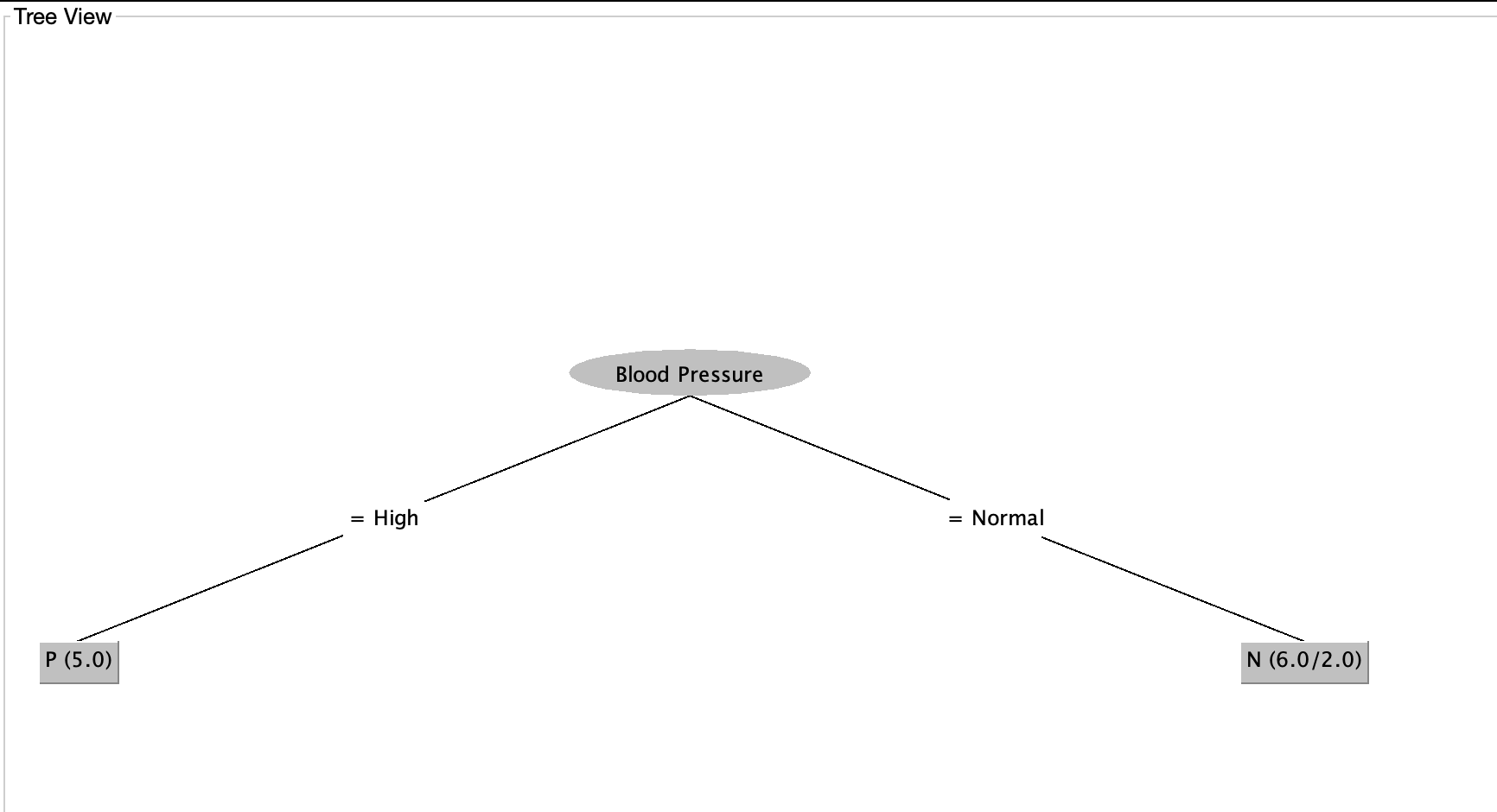
# Laboratory Questions

The data set in the table below contains data about heart disease and its conditions. The class label P means that heart disease is present and the class label N that the disease is absent.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Body  Weight | Body Height | Blood Pressure | Blood Sugar Level | Habit | Class |
| Heavy | Short | High | 3 | Smoker | P |
| Heavy | Short | High | 1 | Non-smoker | P |
| Normal | Tall | Normal | 3 | Non-smoker | N |
| Heavy | Tall | Normal | 2 | Smoker | N |
| Low | Medium | Normal | 2 | Non-smoker | N |
| Low | Tall | Normal | 1 | Non-smoker | P |
| Normal | Medium | High | 3 | Smoker | P |
| Low | Short | High | 2 | Smoker | P |
| Heavy | Tall | High | 2 | Non-smoker | P |
| Low | Medium | Normal | 3 | Smoker | P |
| Heavy | medium | Normal | 3 | Non-smoker | N |



Tree for data of present or absent of heart disease



1. Use the dataset above as a training set and perform the following tasks in WEKA:
   1. Use the J48 algorithm to construct a decision tree under each of the test options. Use the validation data set in the table shown below as a test set for the Supplied test set option.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Body  Weight | Body Height | Blood Pressure | Blood Sugar Level | Habit | Class |
| Heavy | Short | High | 2 | Smoker | P |
| Heavy | Tall | Normal | 1 | Smoker | N |
| Heavy | Medium | Normal | 3 | Smoker | N |
| Low | Short | Normal | 3 | Smoker | N |
| Low | Medium | High | 1 | Non-smoker | N |
| Low | Medium | High | 3 | Non-smoker | P |



* 1. Use the decision tree obtained from the previous task (task a) to determine the classes for the unseen data records as below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Body  Weight | Body Height | Blood Pressure | Blood Sugar Level | Habit | Class |
| Heavy | Medium | High | 1 | Smoker | P |
| Heavy | Medium | Normal | 3 | Smoker | P |

While trying out with the result of both person with the present of heart disease, the result comes out with some inaccuracy in the data.



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Body  Weight | Body Height | Blood Pressure | Blood Sugar Level | Habit | Class |
| Heavy | Medium | High | 1 | Smoker | P |
| Heavy | Medium | Normal | 3 | Smoker | N |

But when trying out with the person with high blood pressure has the heart disease but for the person with normal blood pressure does not have the heart disease, the result seems to be completely accurate.

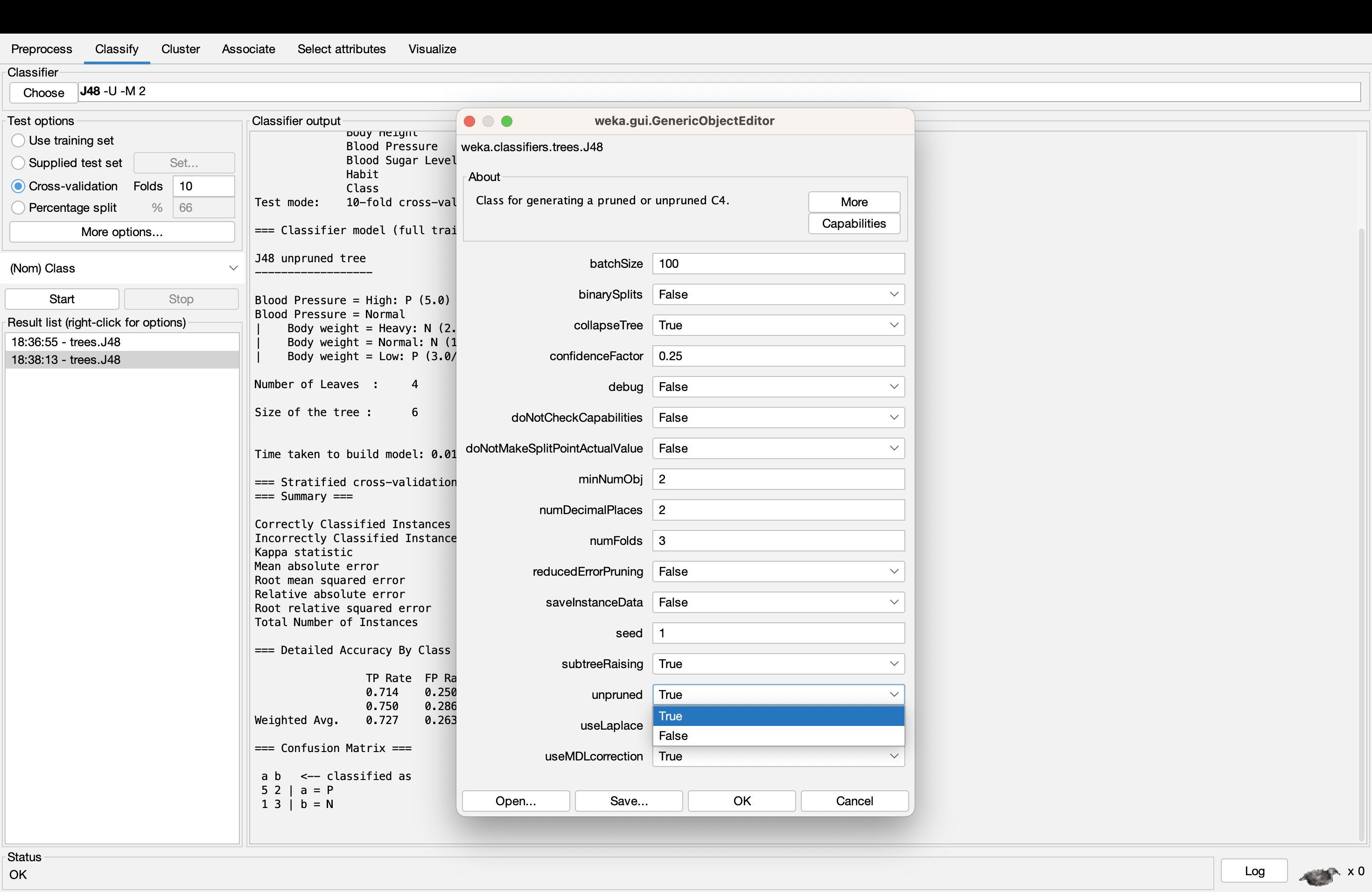


1. Perform decision tree induction in WEKA using different decision tree induction algorithms (RandomTree, J48, REPTree). Compare the resulting trees and the measures of accuracy. Switch the pruning parameter on and off to observe the unpruned and pruned trees.

J48\_unpruned = FALSE



Switching FALSE to TRUE



J48\_unpruned = TRUE



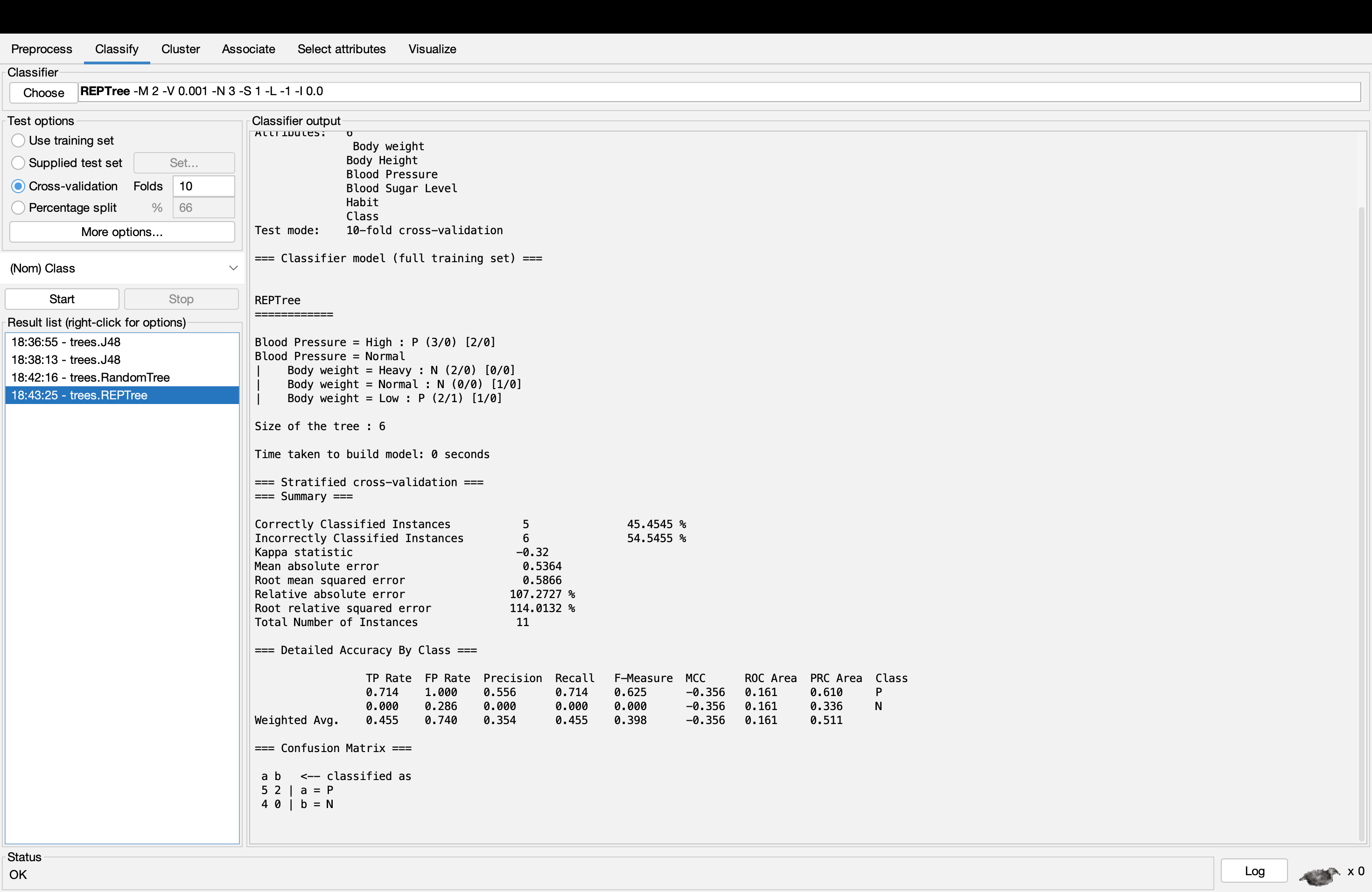
For J48 Algorithm, there is no interference in the accuracy even though pruning is True or False.

Random\_Tree

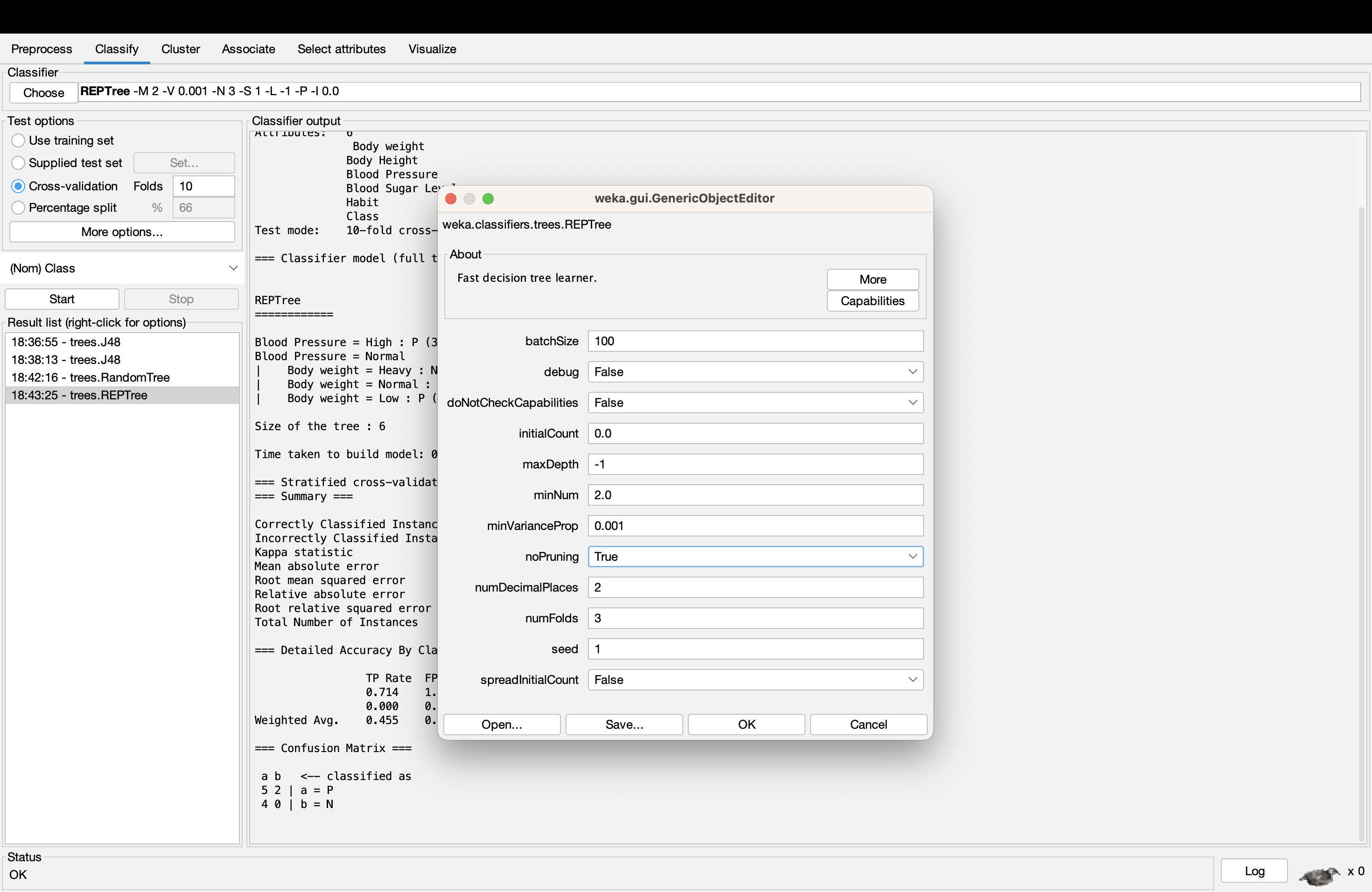


In Random Tree Algorithm, the pruning cannot be switched to False or True.

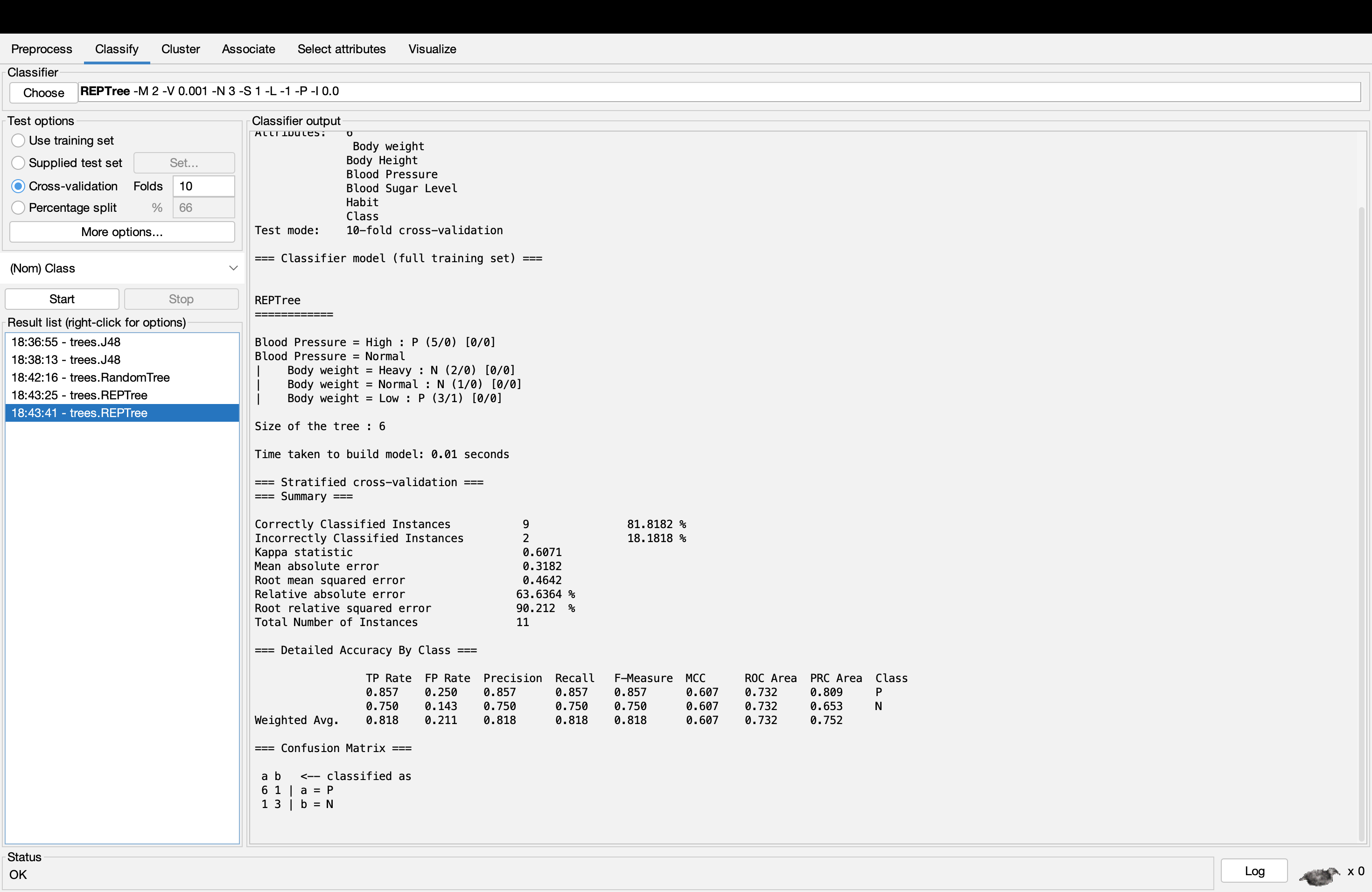
REPTree\_noPruning = False



Switching False to True



REPTree\_noPruning = True



In REPTree Algorithm, when noPruning is switched to True, the accuracy of the algorithm increases significantly.