Assignment2

Author: Jingyi Wu (jingyiw2)

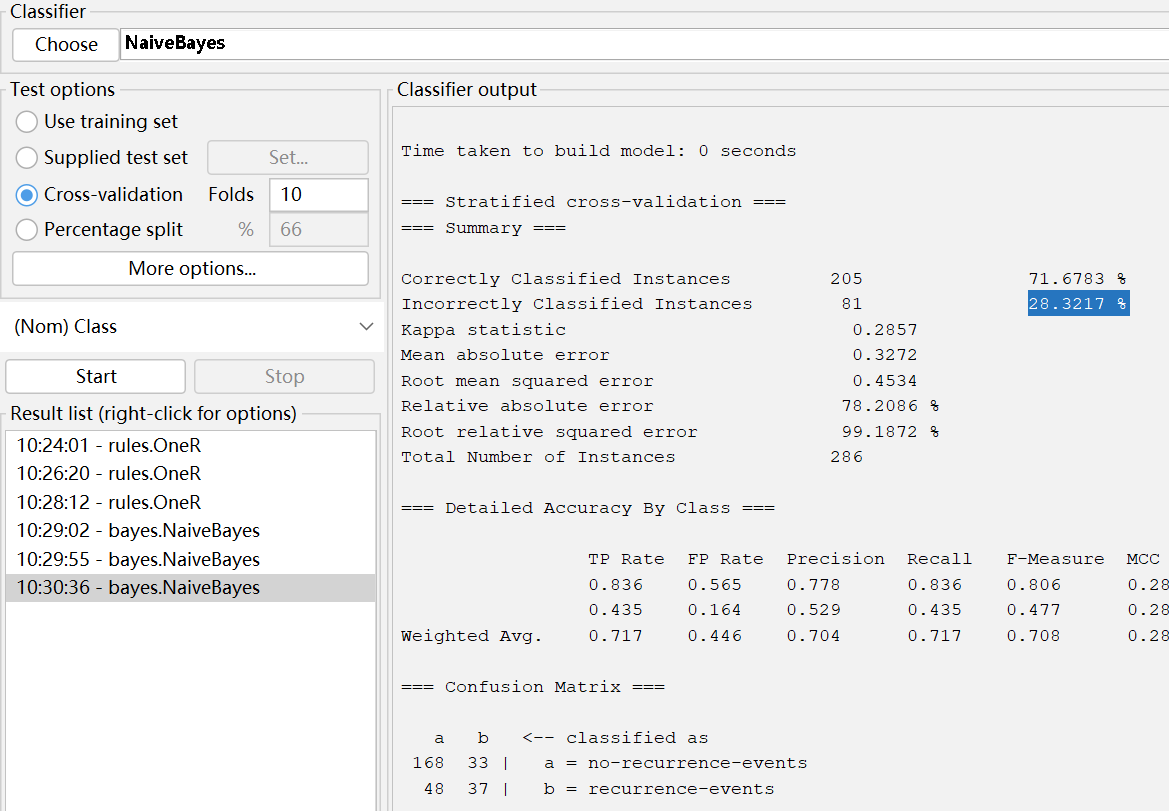
1. Classification learning

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Algorithhm | Error rate(%)=1-accuracy | Datasets | | | | |
| breast-cancer(k=4) | diabetes(k=7) | lymphography(k=8) | primary-tumor(k=4) | soybean(k=2) |
| All are nominal;  Binary class;  Only removed missing value with PRISM | After discretization;  Binary class | After discretization;  4 classes | all nominal, with missing value, 18 classes | all nominal, with missing values;  15 classes |
| OneR | Training set | 27.27% | 25.39% | 24.32% | 70.45% | 62.46% |
| 66% split | 31.96% | 23.75% | 32% | 73.33% | 69.11% |
| Cross-validation | 34.27% | 26.43% | 25% | 75.76% | 63.52% |
| Naïve Bayes | Training set | 24.83% | 22.27% | 12.84% | 37.12% | 7.30% |
| 66% split | 28.87% | 21.07% | 20% | 57.78% | 10.47% |
| Cross-validation | 28.32% | 24.61% | 16.89% | 53.03% | 8.36% |
| J4.8 | Training set | 24.13% | 20.70% | 8.78% | 35.61% | 3.38% |
| 66% split | 31.96% | 23.75% | 22% | 64.44% | 7.85% |
| Cross-validation | 24.48% | 26.17% | 20.27% | 56.82% | 8.19% |
| PRISM | Training set | 2.53% | 0.26% | 0% | 9.09% | 0.18% |
| 66% split | 28.72% | 23.37% | 36% | 73.33% | 15.18% |
| Cross-validation | 29.24% | 23.31% | 23.65% | 66.67% | 14.23% |
| Ibk | Training set | 22.73% | 25.13% | 15.54% | 45.45% | 5.69% |
| 66% split | 26.80% | 26.44% | 28% | 60% | 12.04% |
| Cross-validation | 25.52% | 30.73% | 17.57% | 56.06% | 9.43% |

1. For all 5 learning schemes, error rate with only training set is the lowest compared with 66% split and cross-validation, which implies the model learns some rules specific to train set.
2. PRISM (rules method) always has weirdly high accuracy in training set, indicating learning every rule about training set. But when new cases occur in test set, the prediction result of PRISM will become unclassified. This is also the main problem with PRISM.
3. When the classes of dependent variable increases, OneR model’s performance becomes quite unsatisfying. When the classification problem becomes quite complex, it is difficult to use only one feature to accurately get the classified outcome.
4. We can see that all model performs badly on primary tumor dataset. Maybe this could because to make the results comparable, I discretized the numeric features in advance. Then I tried to apply model like Naïve Bayes directly to the dataset. The error rate decreased to 49.85%, improving a little bit, but still not so good. I tried to adjust parameters for the model, improved a little, but still not quite satisfying. Maybe more pre-processing work is required.
5. For all discretized processing, the number of bins also affects the model.
6. Naïve Bayes and decision tree perform pretty well with multi-class classification problem. PRISM can be good when dealing with binary classification problem. Especially naïve bayes is better suited for categorical input variables than numeric ones.
7. New data can be added seamlessly using KNN algorithm, whereas this does not apply to PRISM. But note that here we discretized all variables before putting into KNN. If we put original numeric variables into the model, feature scaling is needed.
8. For the last two datasets, we removed a large percentage of instances with missing values in features.
9. If we only consider training error when evaluating models, we tend to choose the overfitted model, which learns the rule specific to train set. In this way, the model may perform badly when being applied to other dataset and lacks general applicability. But by using a single validation set and evaluating each model on multiple validation sets and averaging the validation performance, the overfitting problem can be eased.
10. Here we pick the **breast-cancer and diabetes** datasets and focus on cross validation confusion matrix.

**Breast-cancer:**

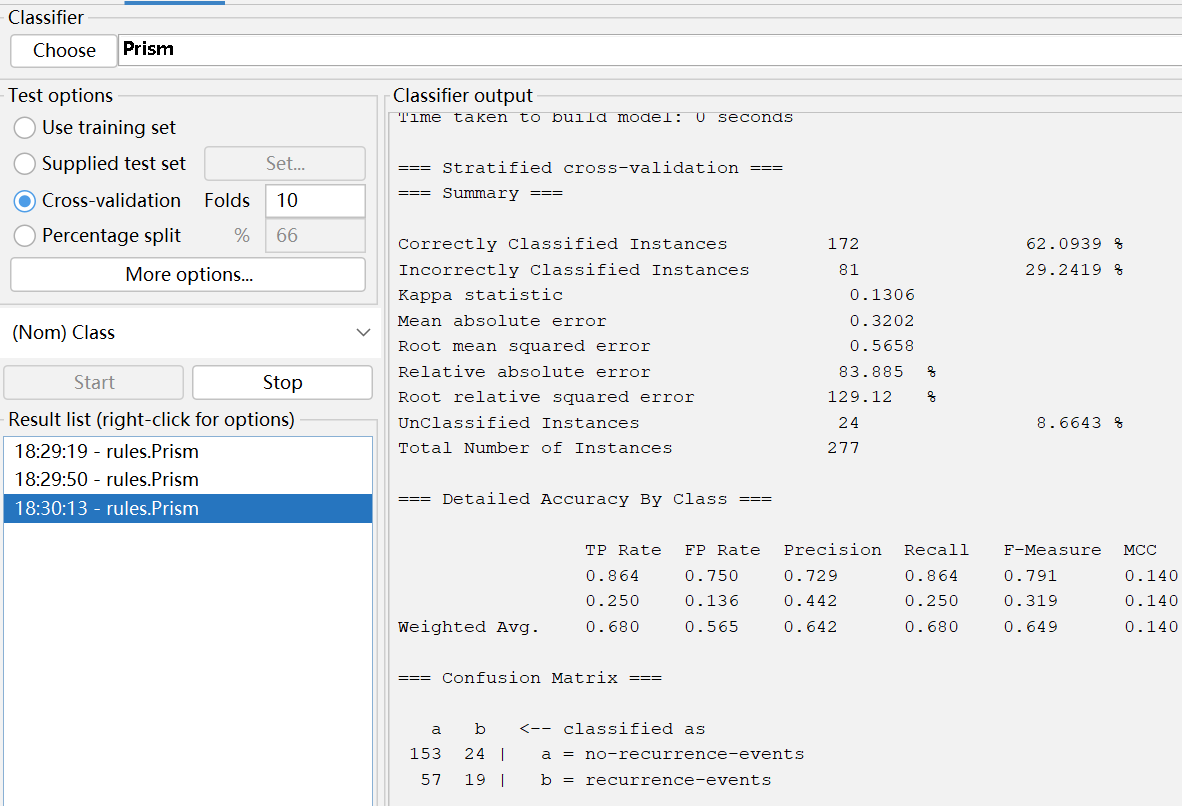
Naïve Bayes:



J4.8



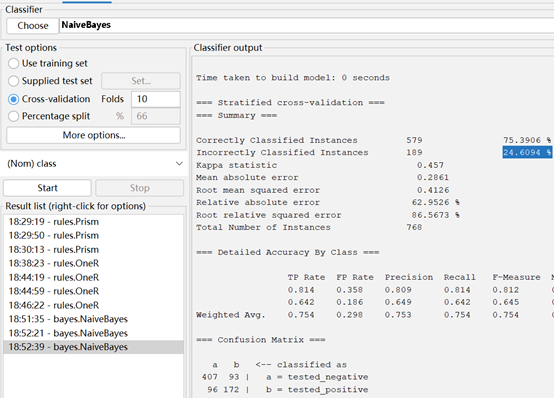
PRISM



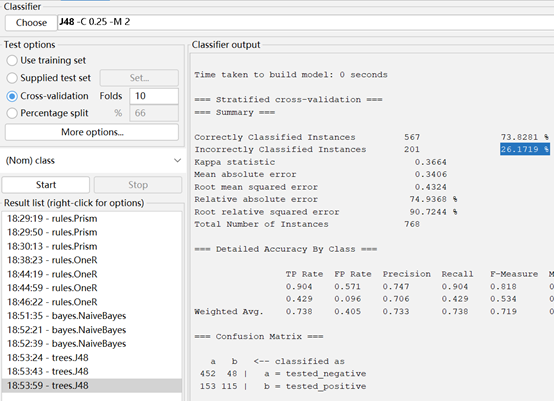
For breast cancer problem, the most important is to detect as many recurrence-events as possible. At the same time, if the precision of the classified recurrence-events is higher, then the model is preferred. From the perspective of recall rate, the naïve bayes model is the best, which means using this model can detect as many patients as possible. From the perspective of precision, decision tree J4.8 is the best, which means most people classified as recurrence are actually patients, which will not cause least anxiety in mis-classified people.

**Diabetes:**

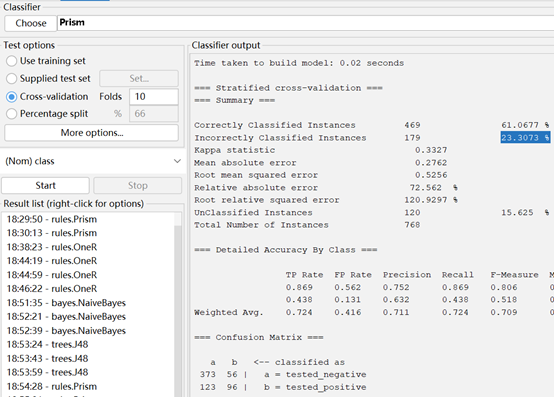
Naïve Bayes:



J4.8:



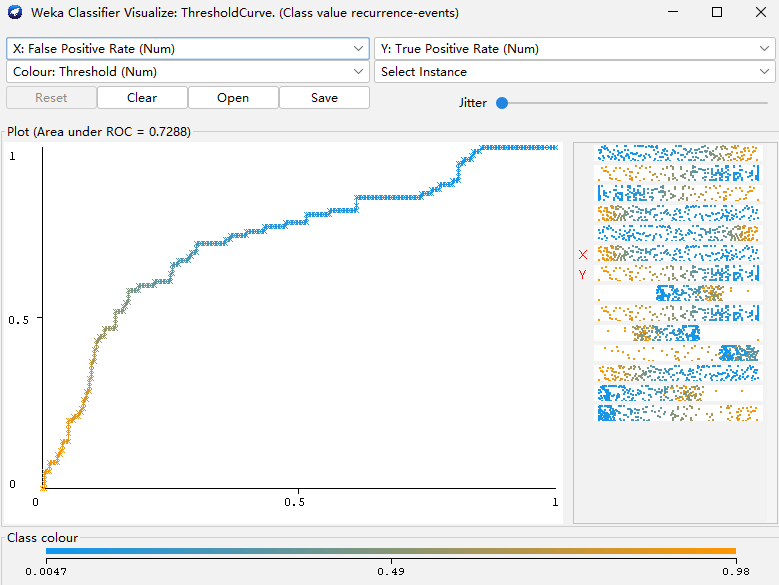
PRISM:



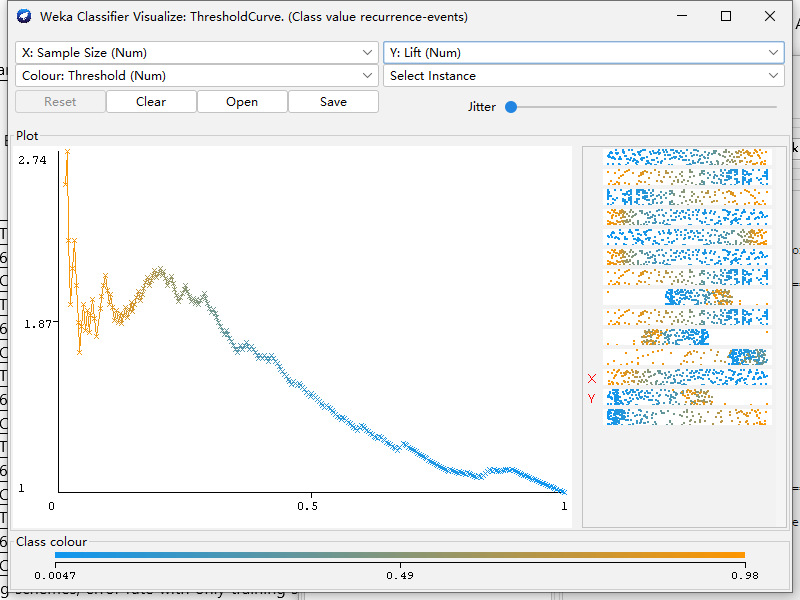
This problem is actually similar to the first one. We put emphasis on detecting as many patients with diabetes as possible and at the same time keep the precision of the classified-positive ones. Also, Naïve Bayes performs best in the recall rate. J4.8 performs best in the precision of tested-positive ones. And the difference between recall rate is larger than the precision perspective. In summary, Naïve bayes is a better option here.

1. For this question, we select the first three datasets.

**For breast-cancer**, I looked into the AUC of four models (Naïve Bayes, J4.8, PRISM, and IBK) and **Naïve bayes** has the largest AUC here, 0.7288.



Below is the lift chart on this dataset. The predicted accuracy reaches 75% after removing instances with missing values. (Note: the data of models for the first dataset are based on not removing missing values, except for PRISM)



From the lift chart, we can see that by targeting the top ranked instances, we can achieve a high lift in finding out the recurrence events.

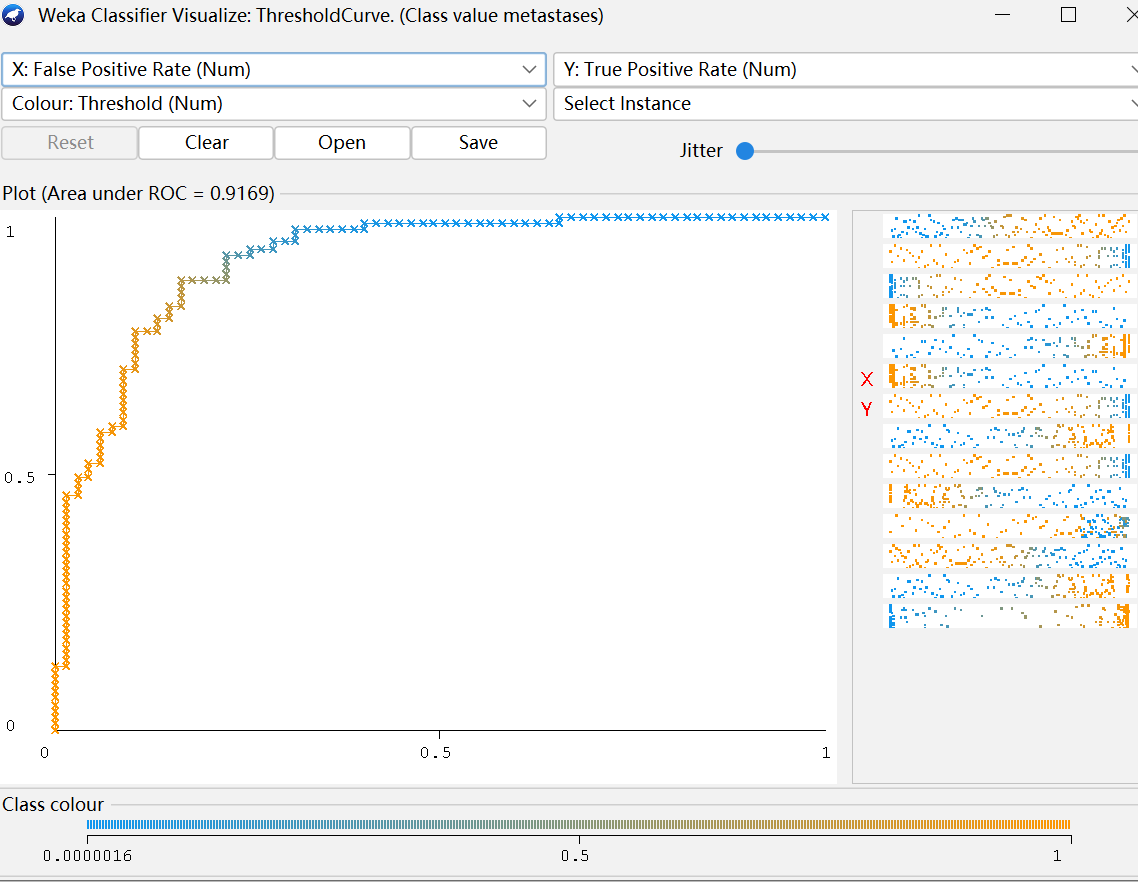
For diabetes, the AUC values for 5 models are 0.6732, 0.825, 0.7446, 0.6213, and 0.7724 correspondingly. (The order is the same as the table). Naïve Bayes generates the best performance and this also aligns with the recall rate result.



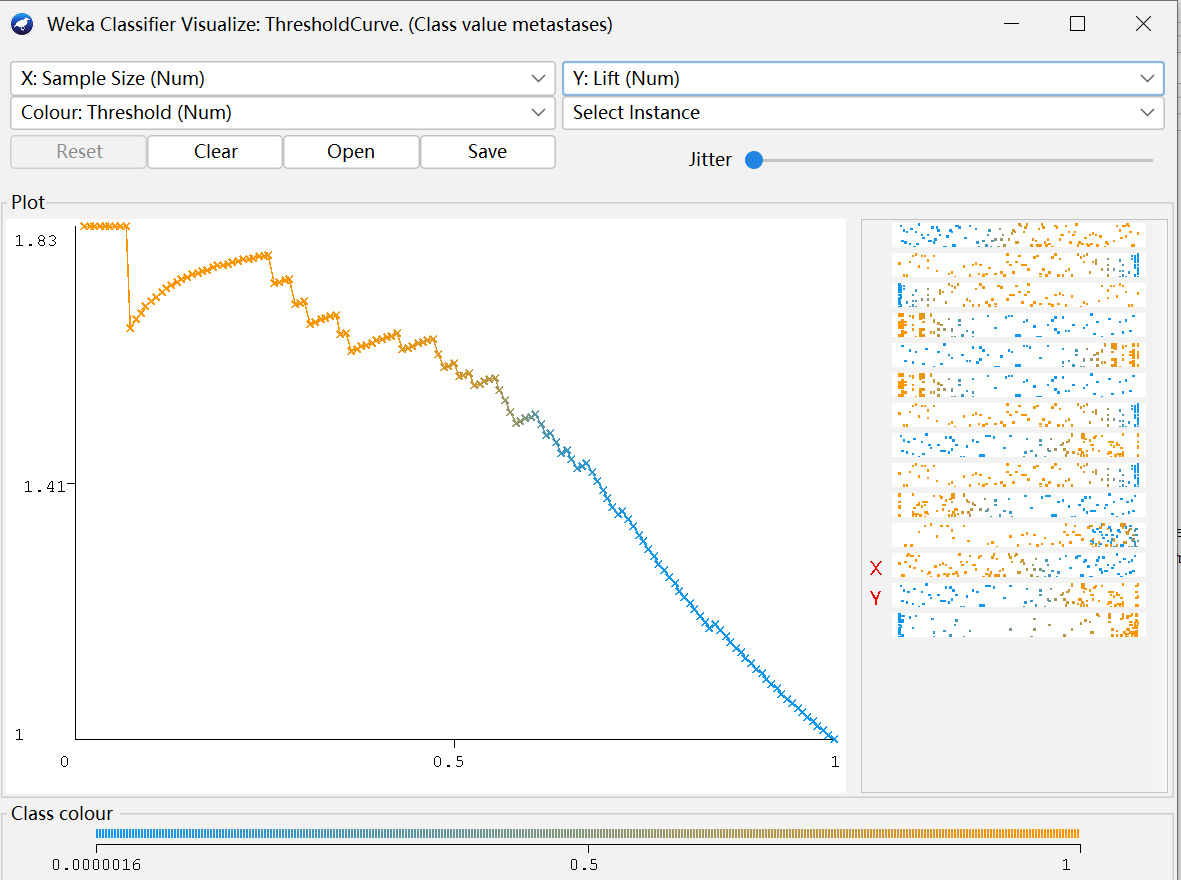


The lift chart is as above. The lift reaches pretty high when targeting the top proportion.

For lymphography, among four classes (normal, metastases, malignment, fibrosis), metastases are often paid attention, since it is quite dangerous and few patients can survive after metastases. Therefore, we looked into the AUC for this class for 5 models, the values are 0.7993, 0.9169, 0.7997, 0.8315, and 0.9081. Naïve Bayes and IBK actually perform well and Naïve Bayes performs the best.

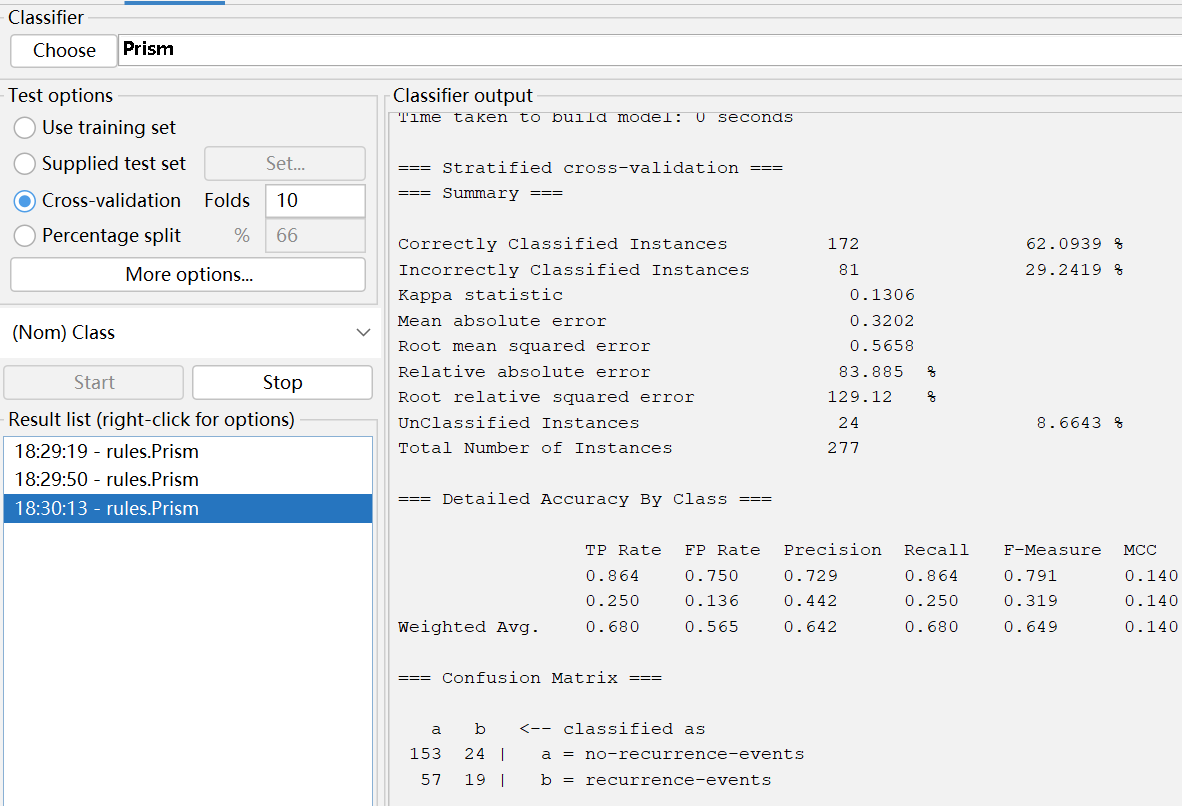
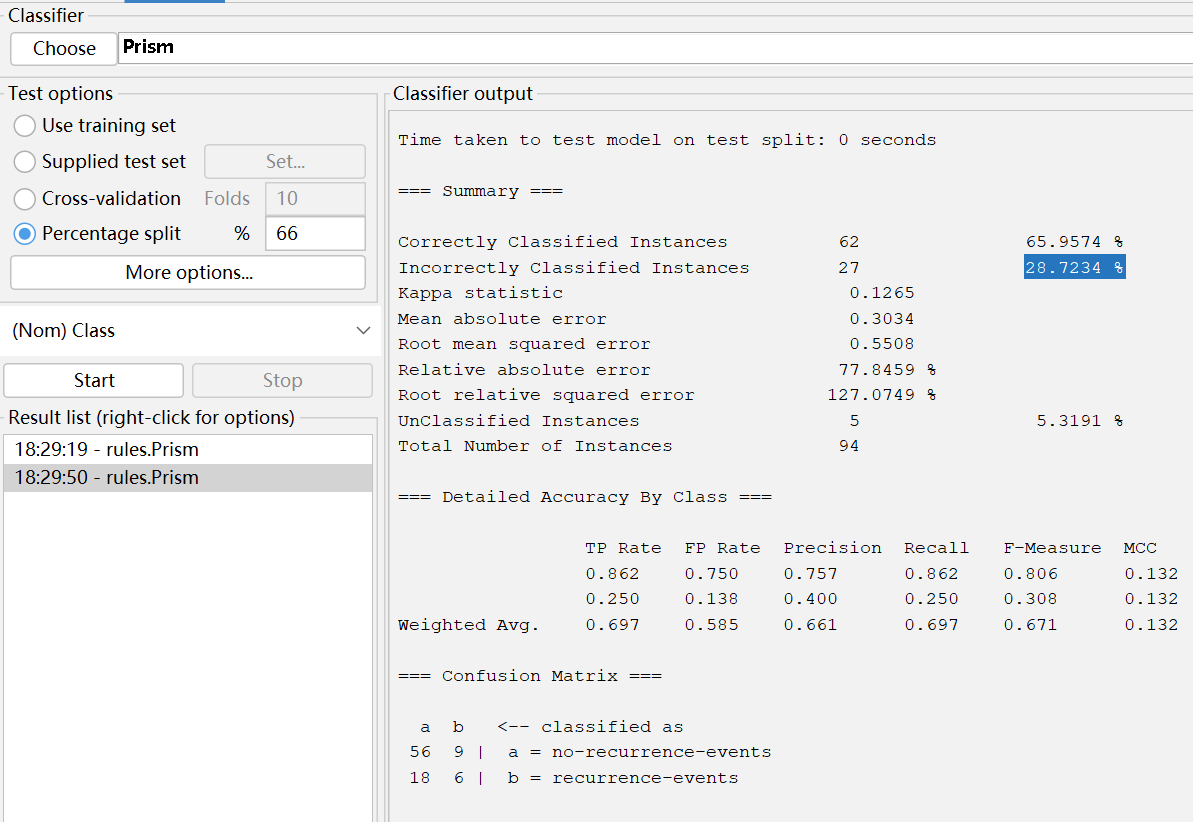
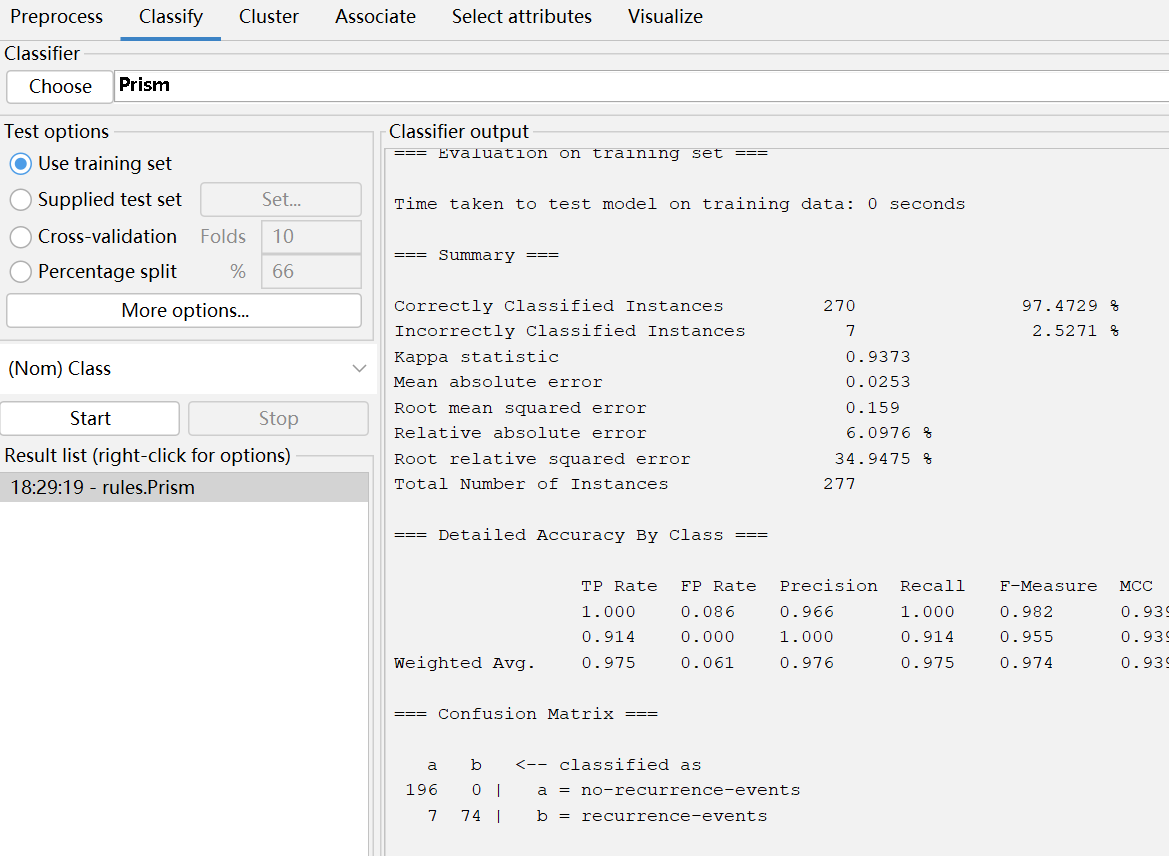
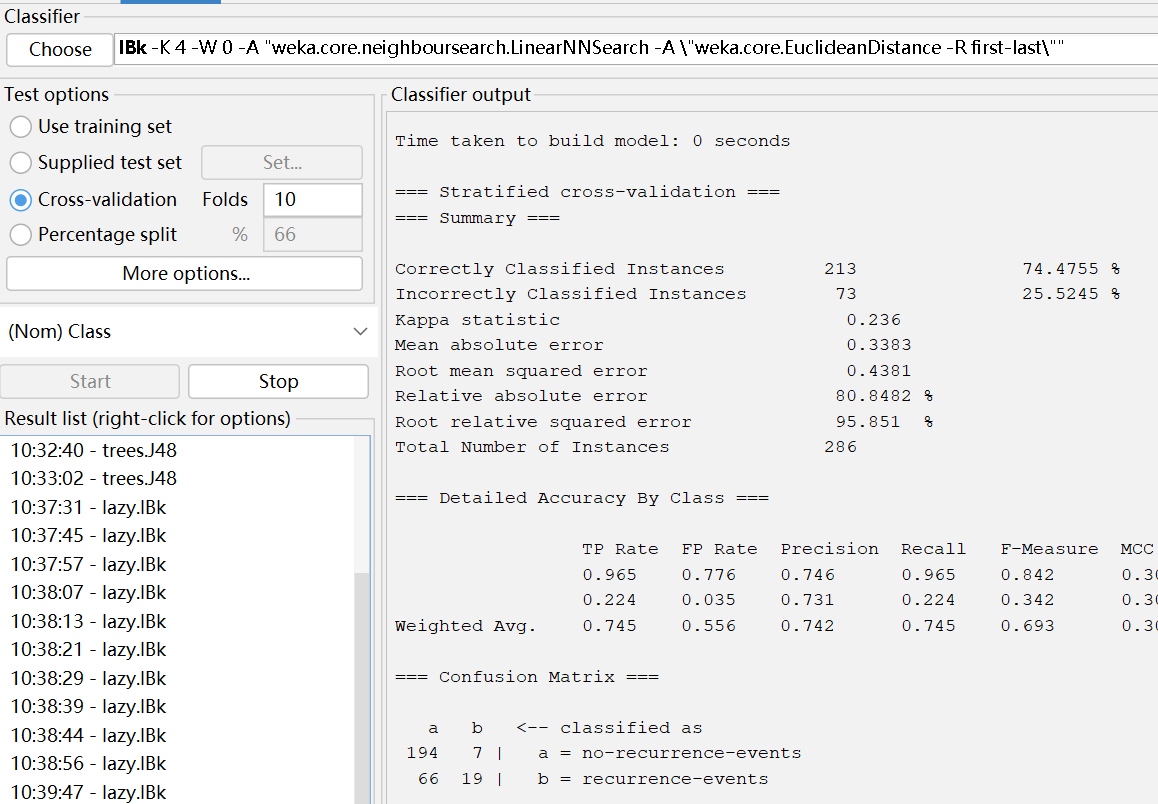
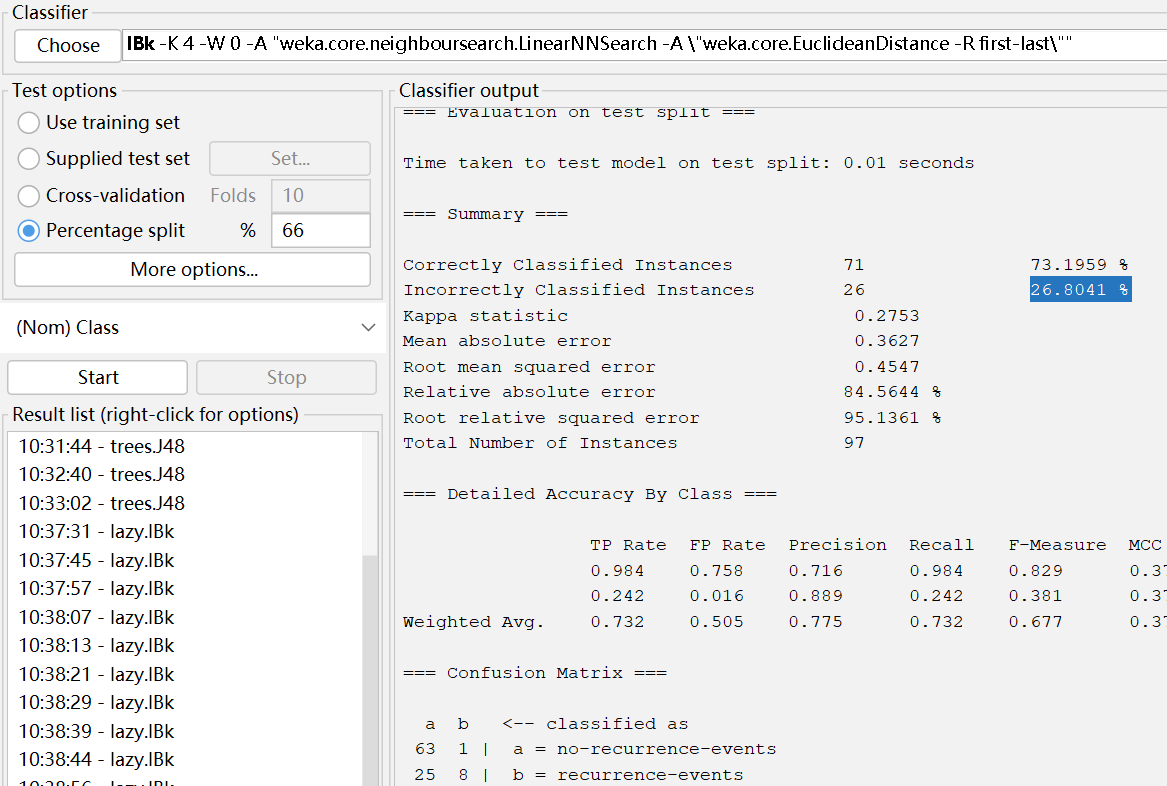
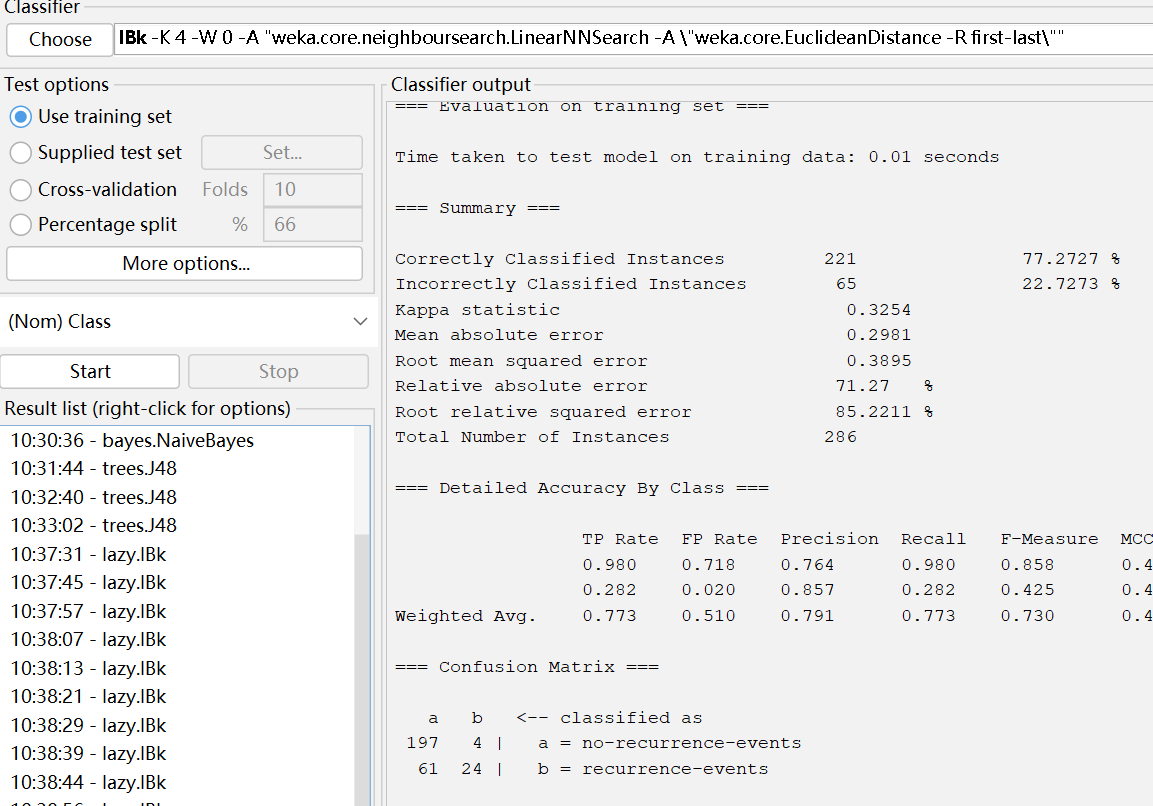
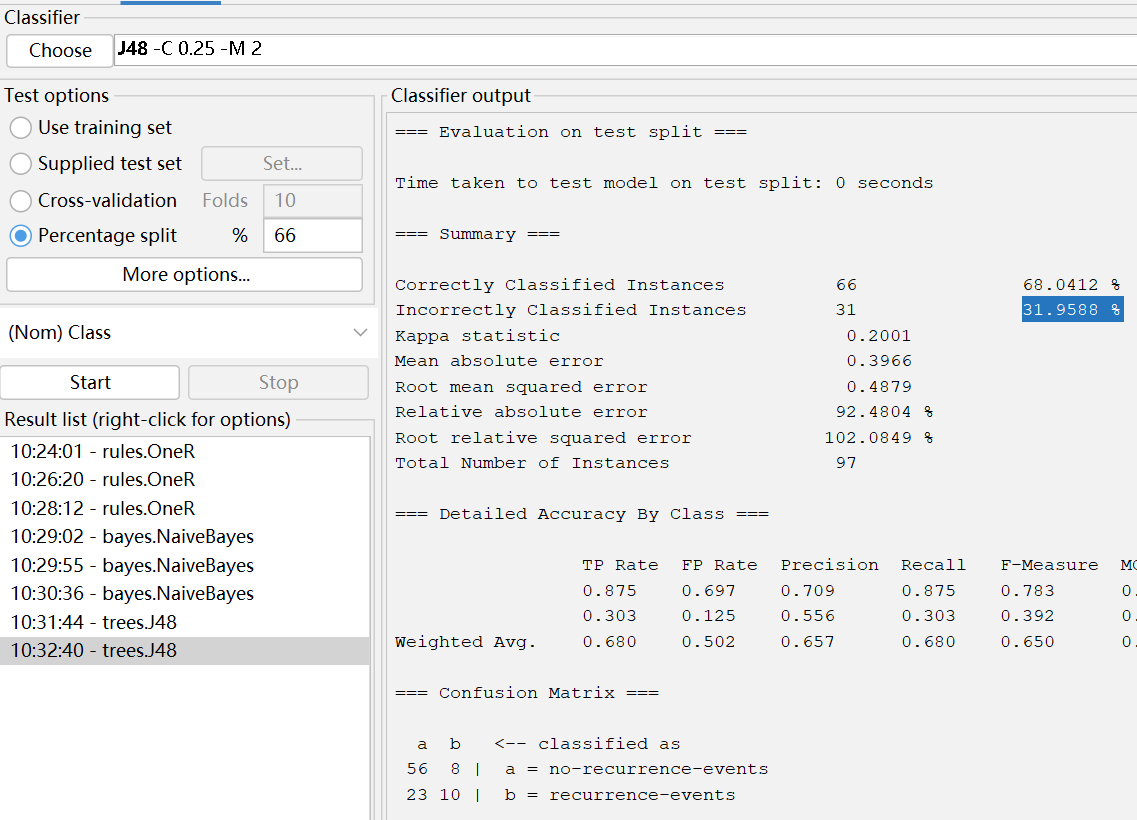
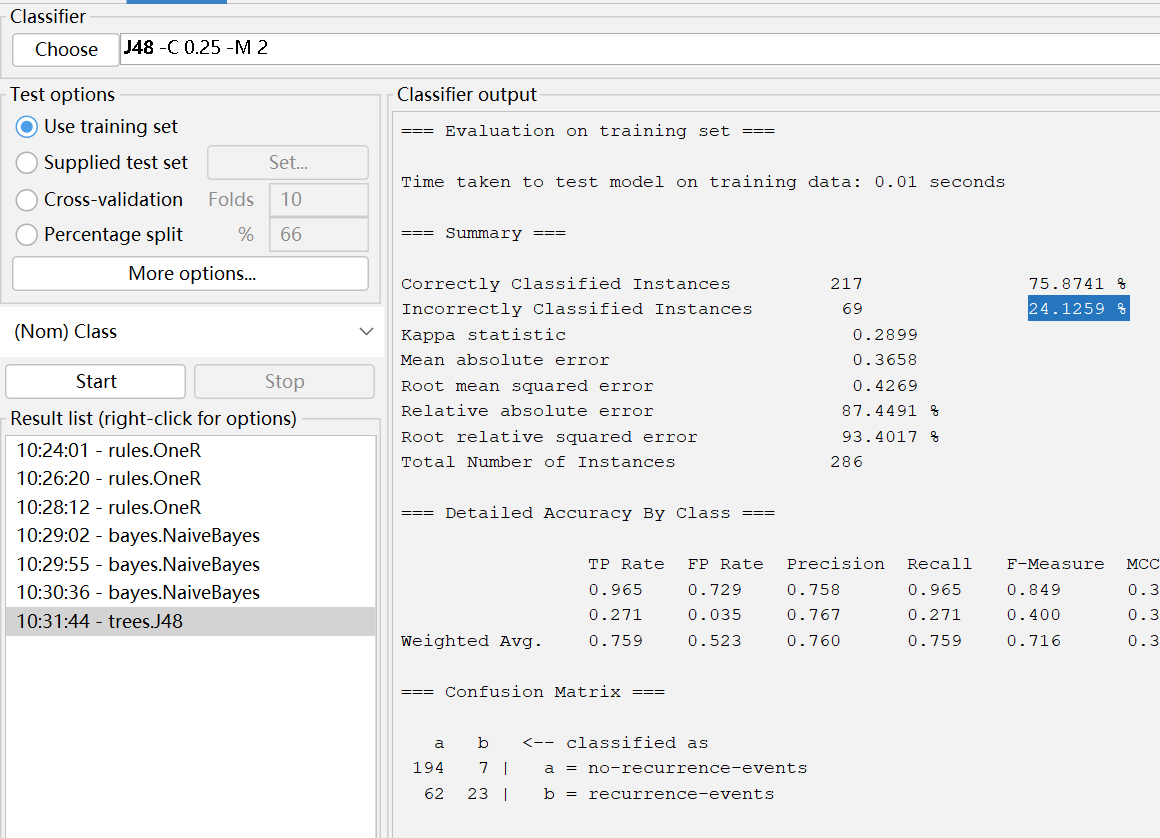
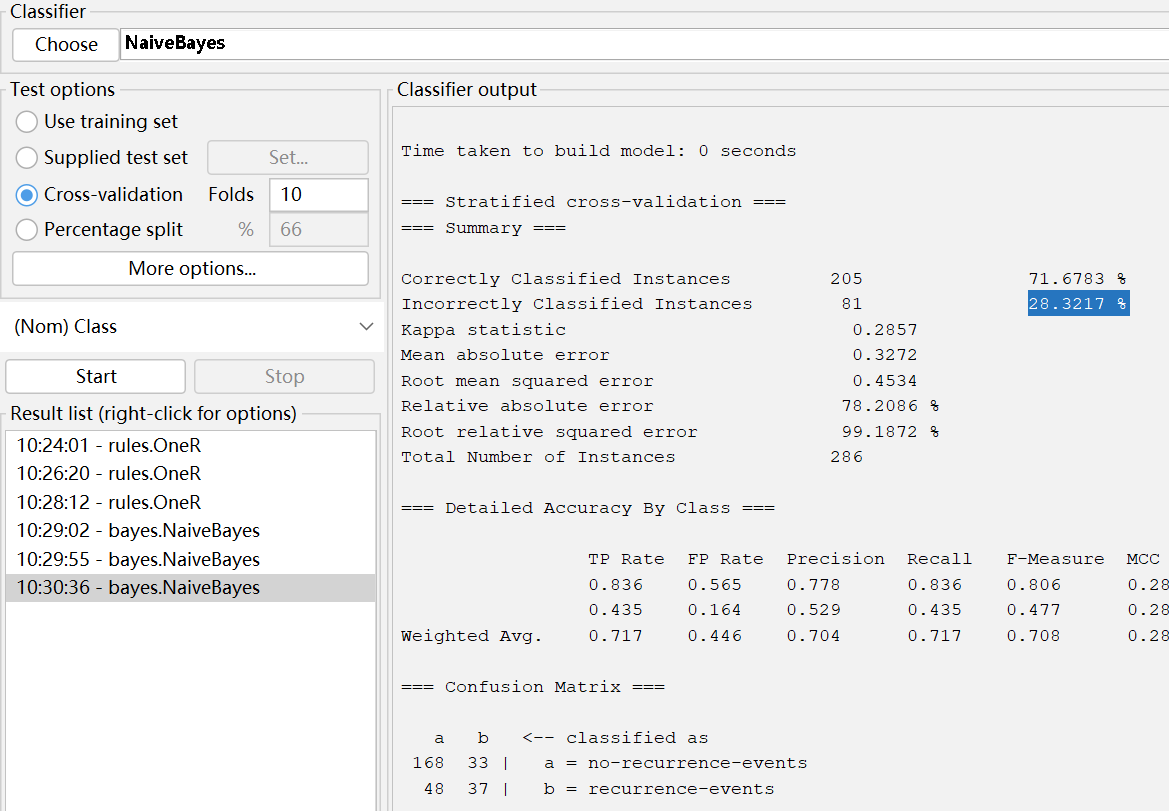
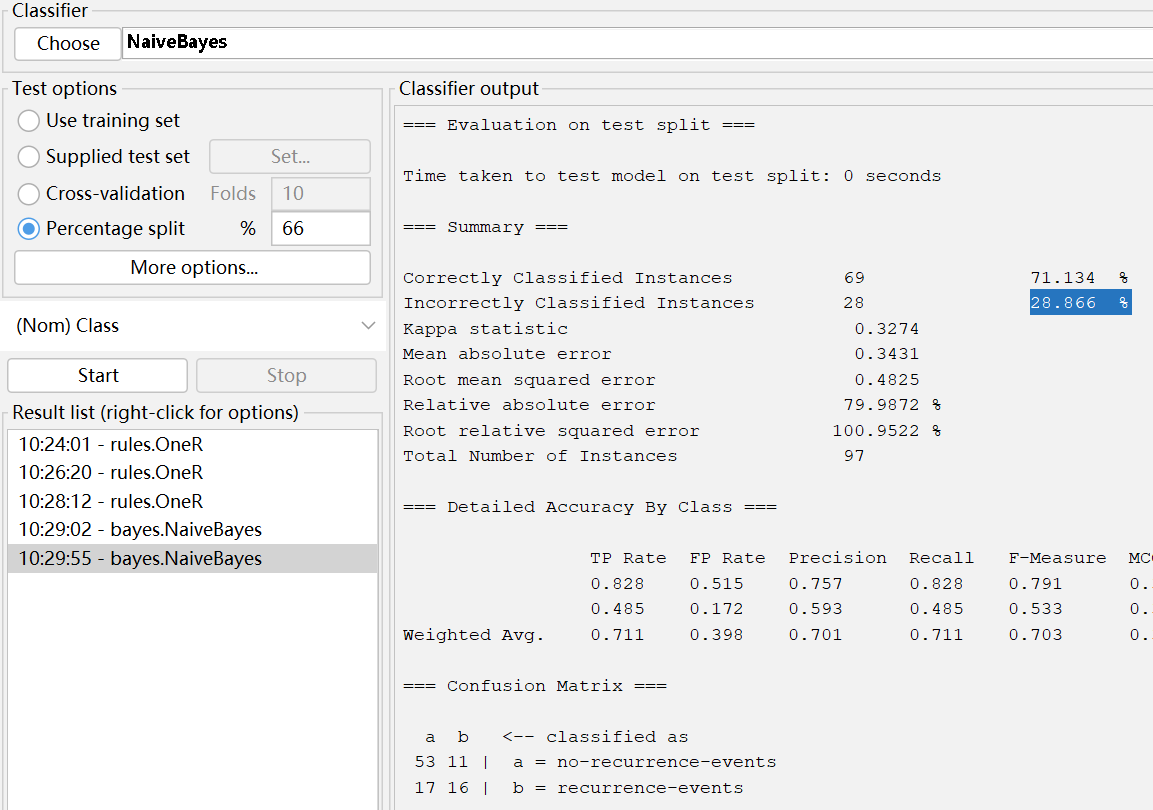
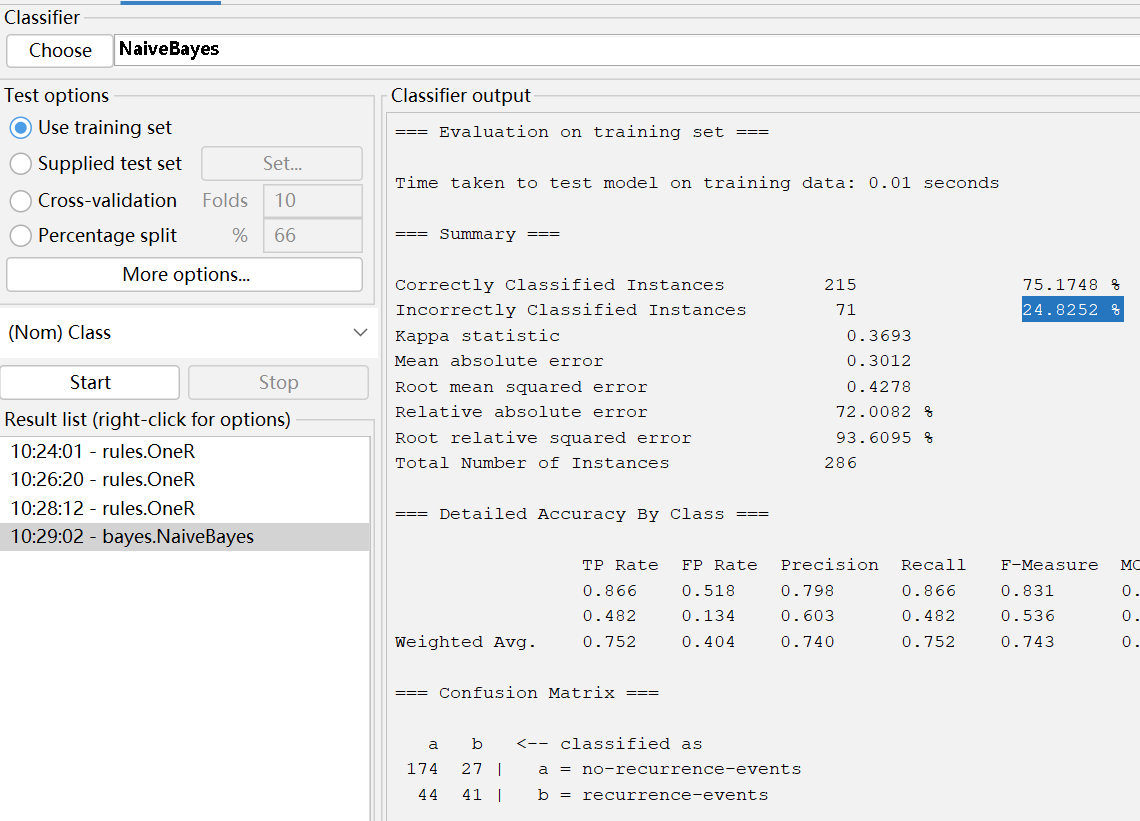
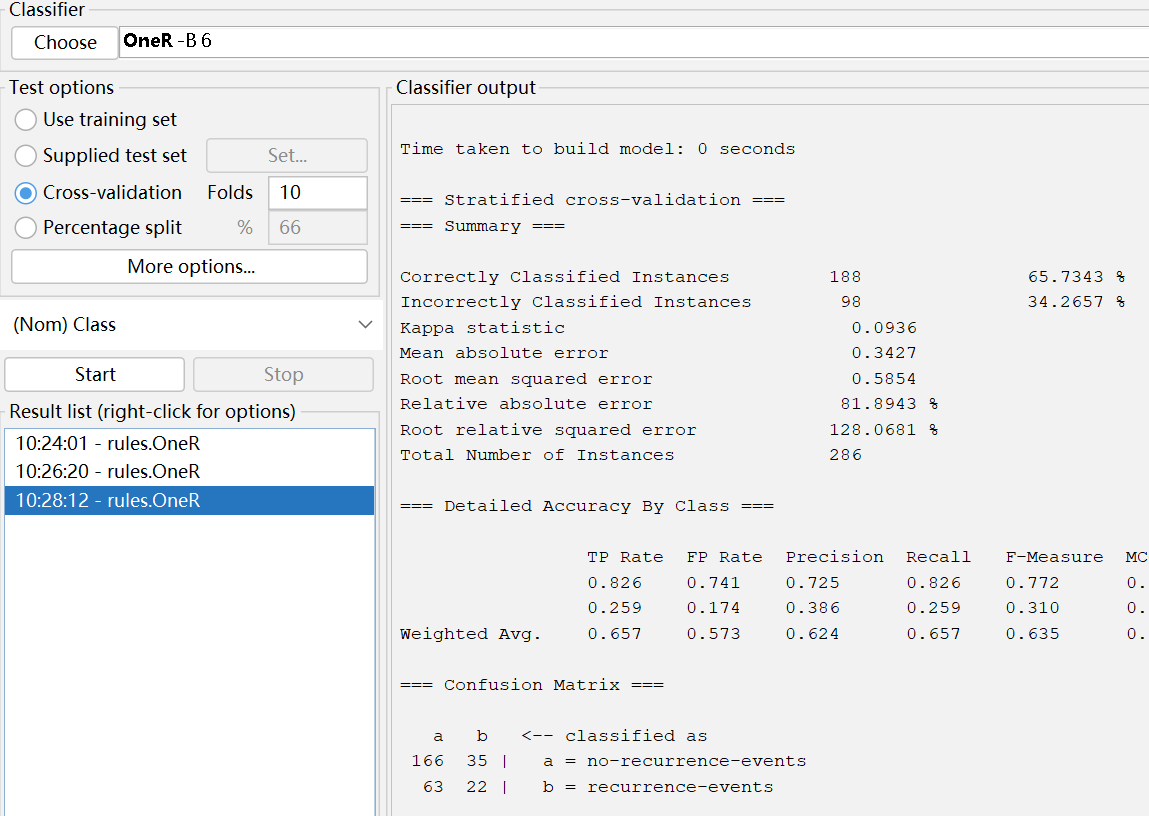
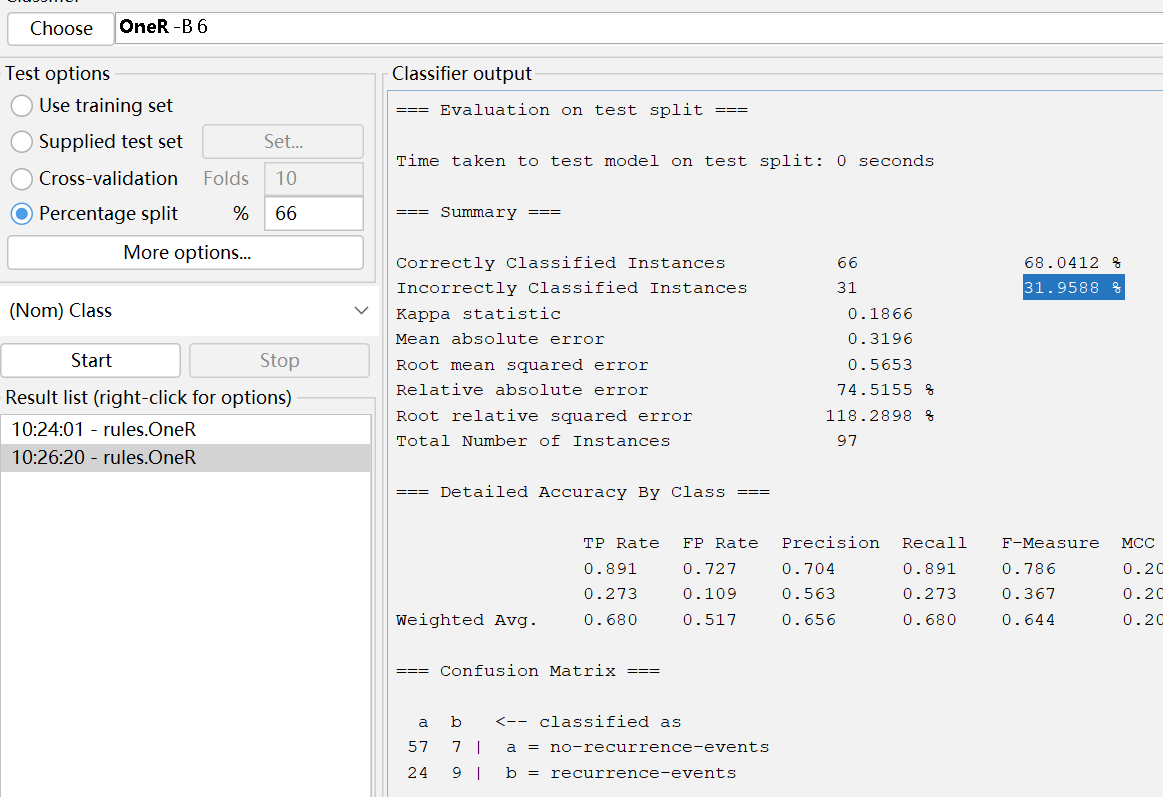
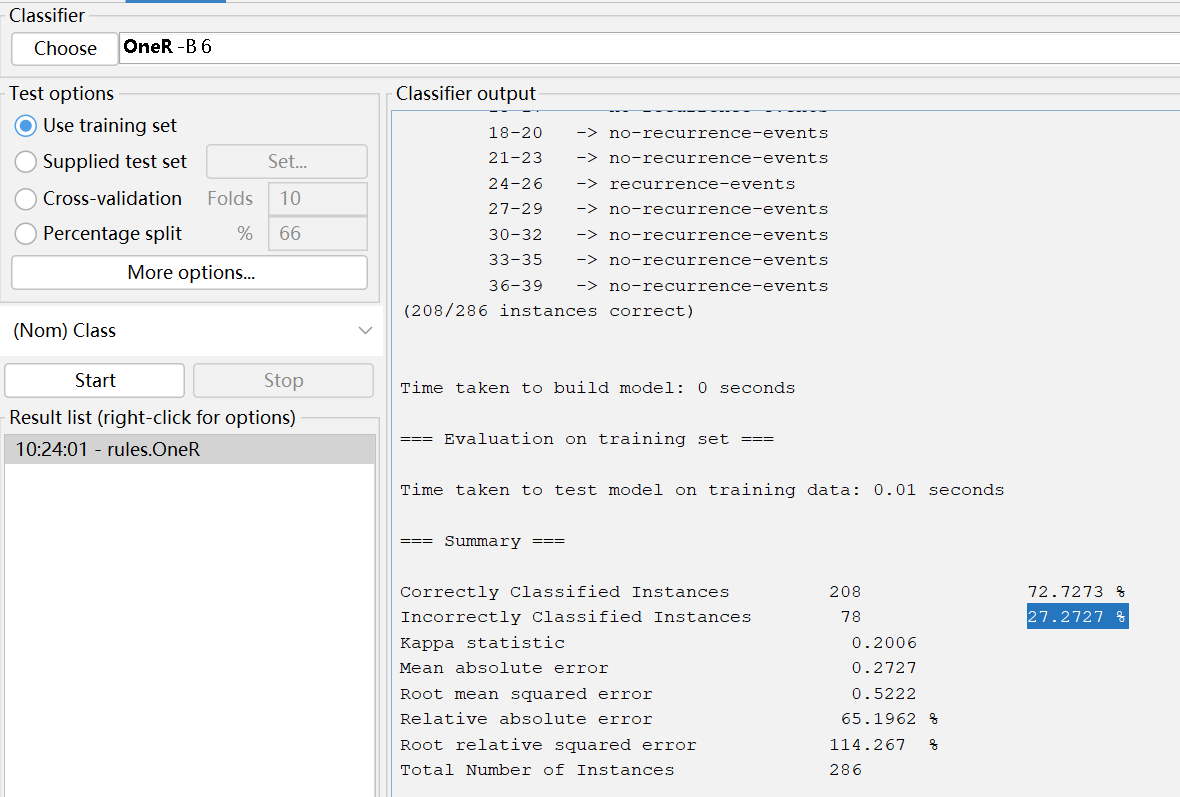


The lift chart is as below. By targeting same proportion of ordered instances, this model brings the highest lift.

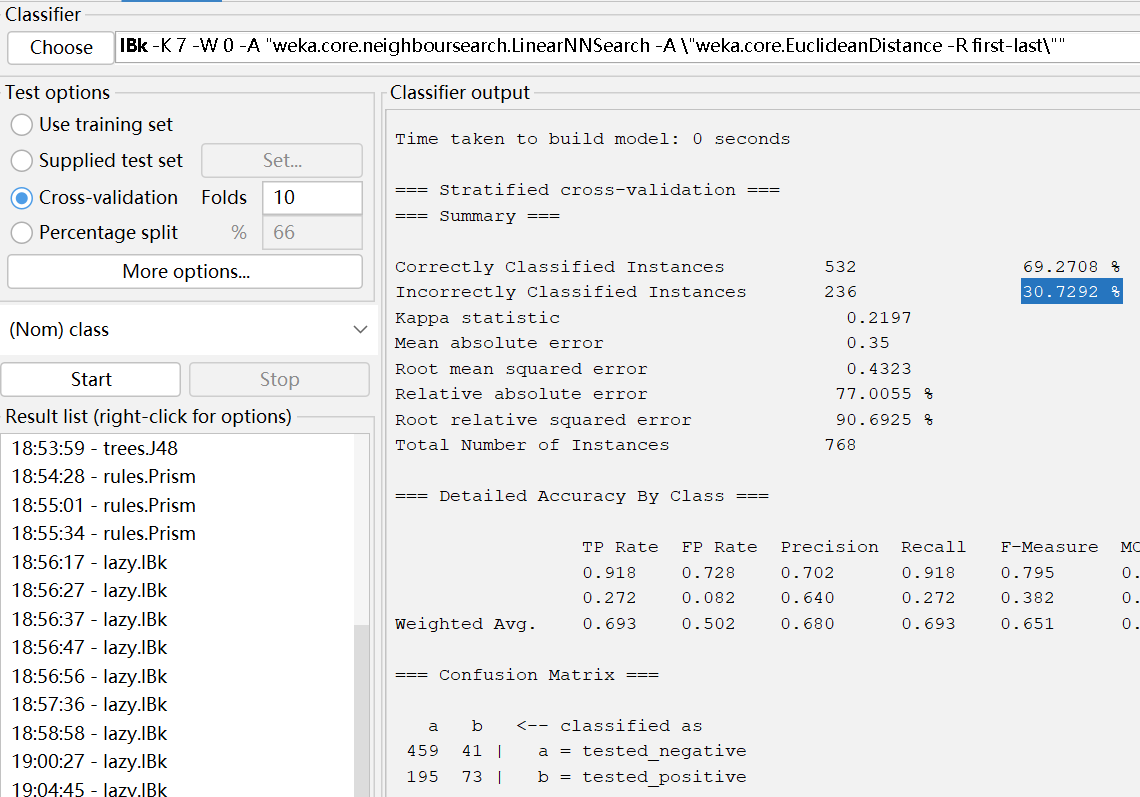
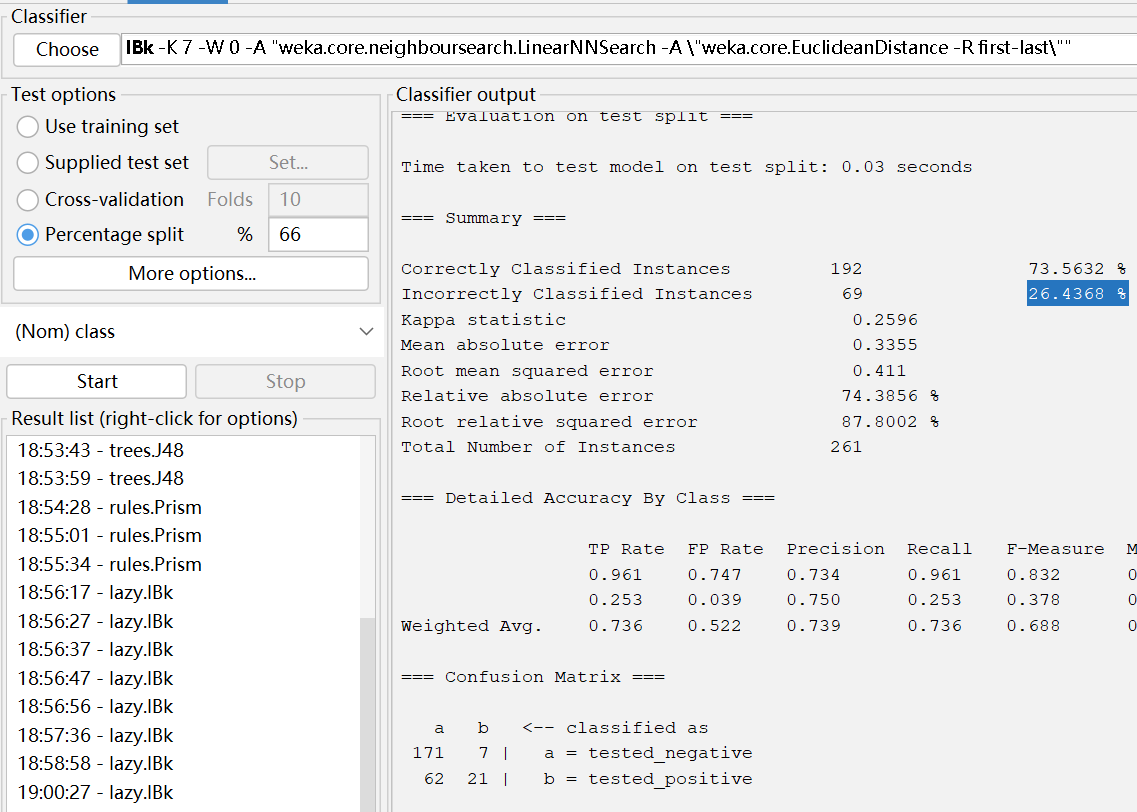
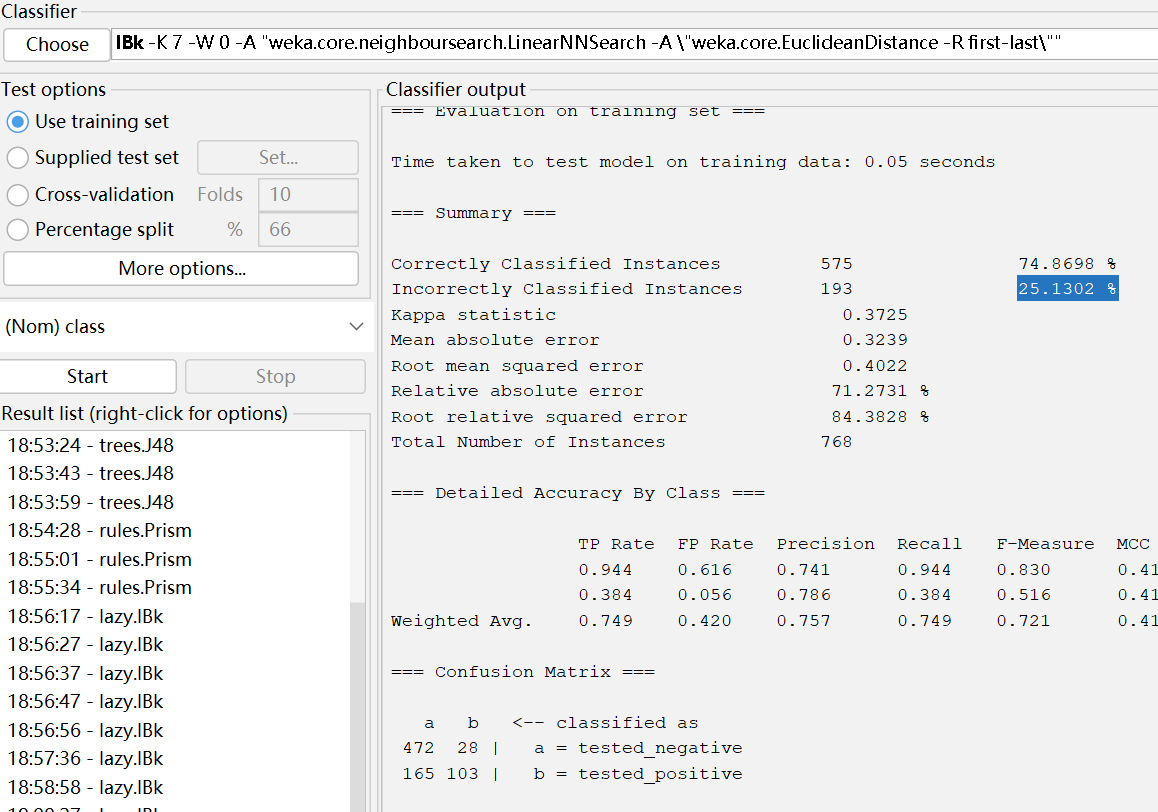
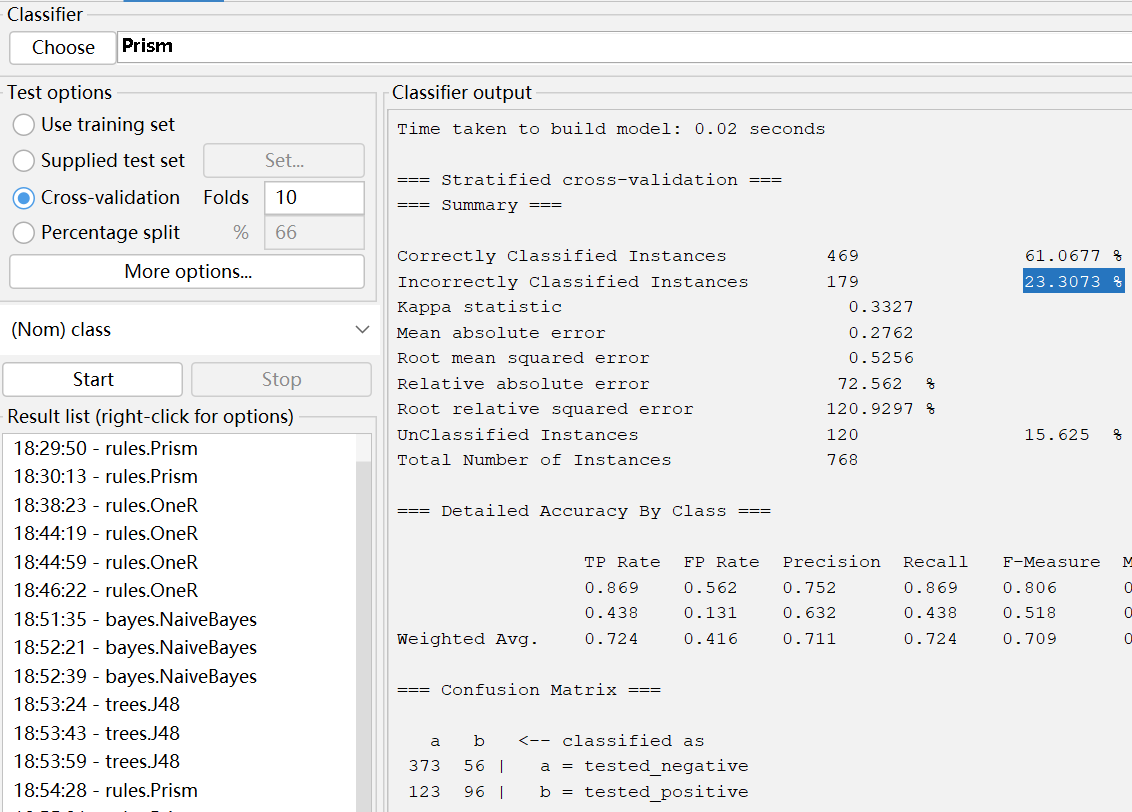
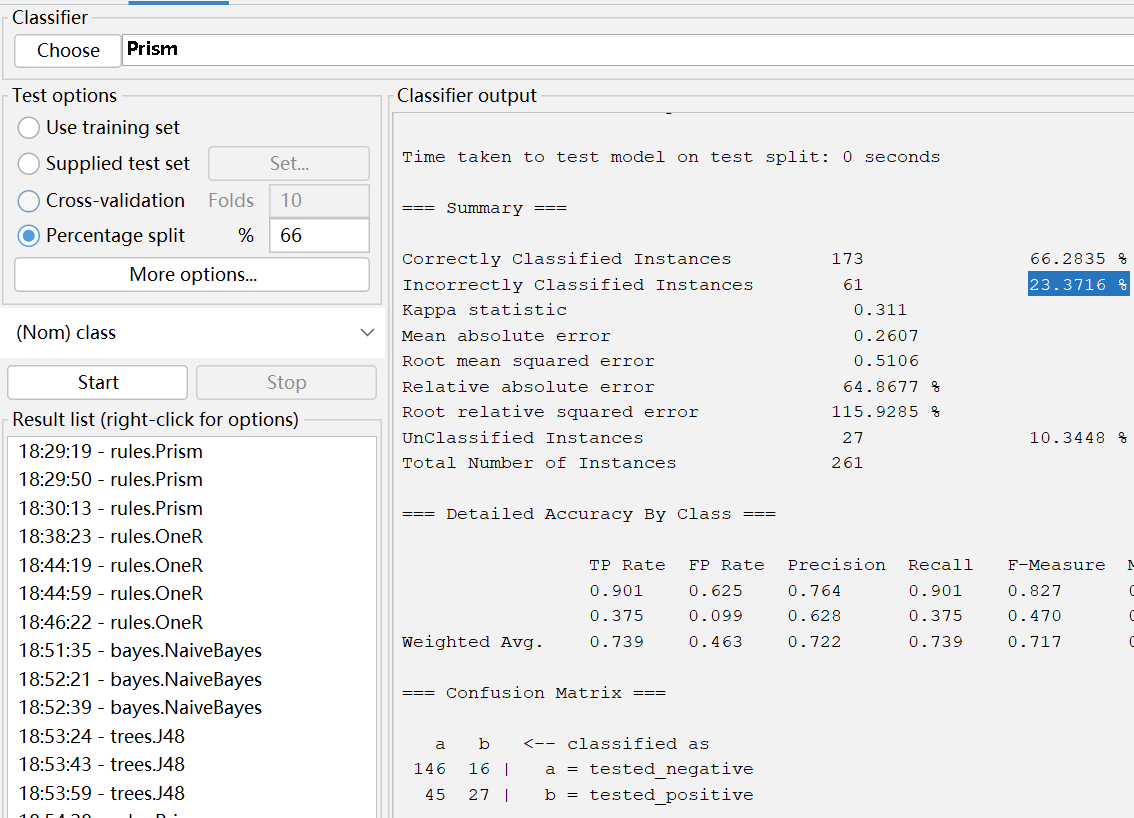
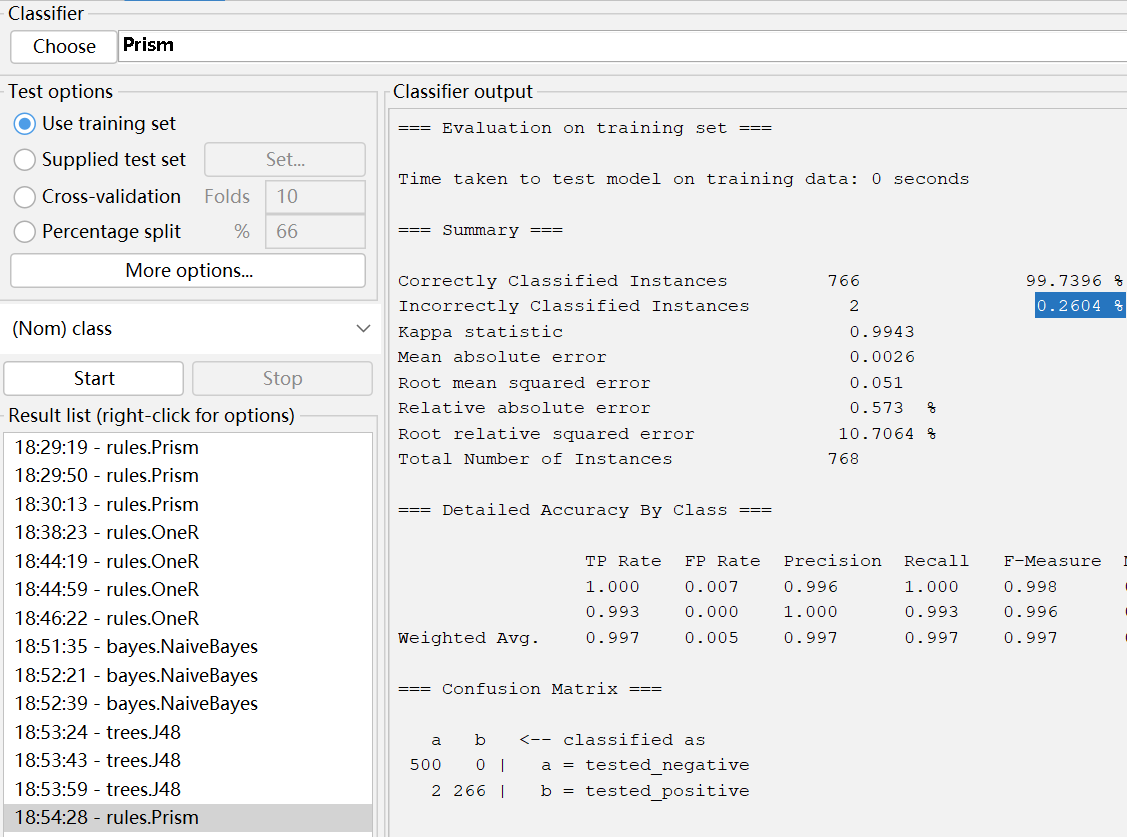
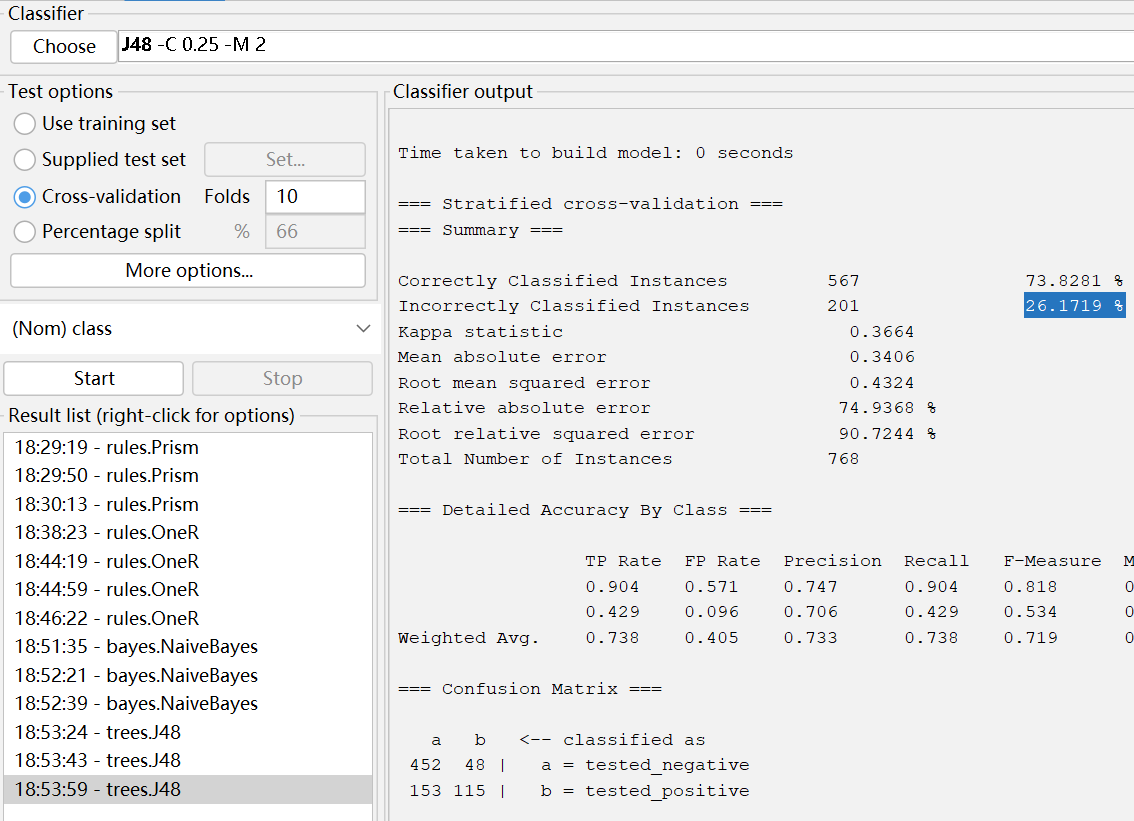
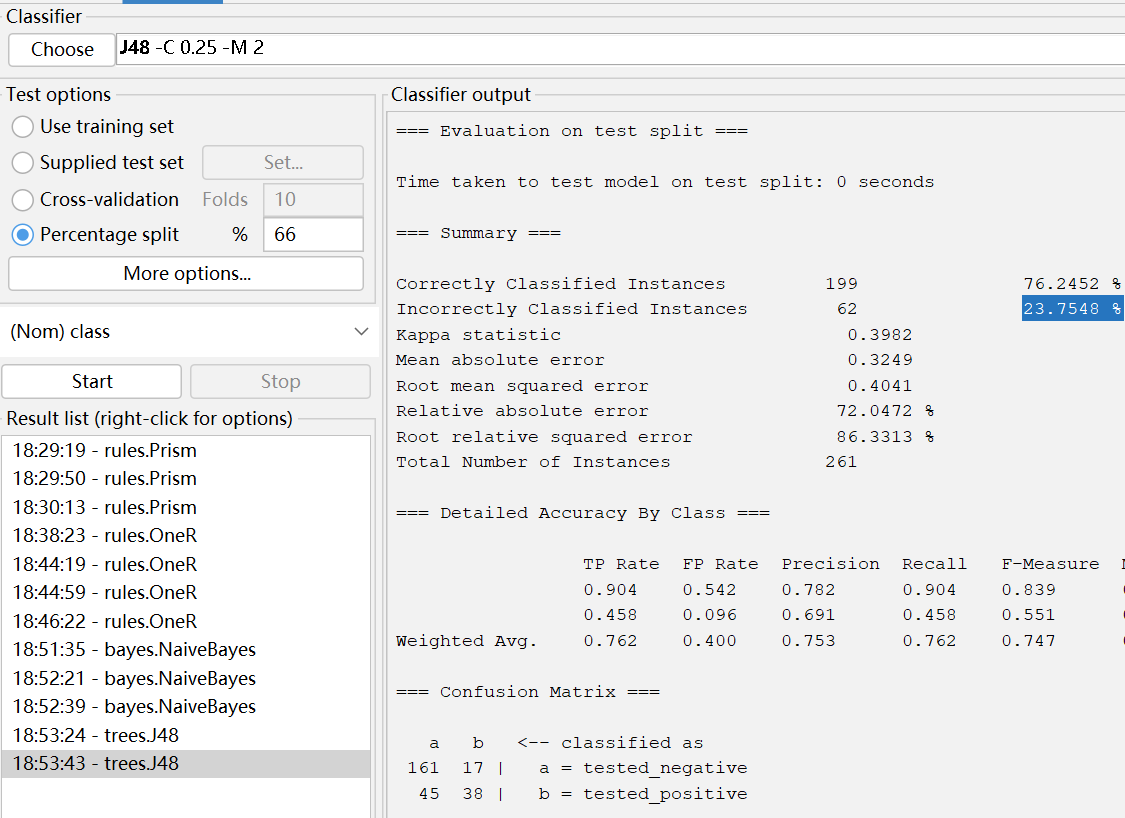
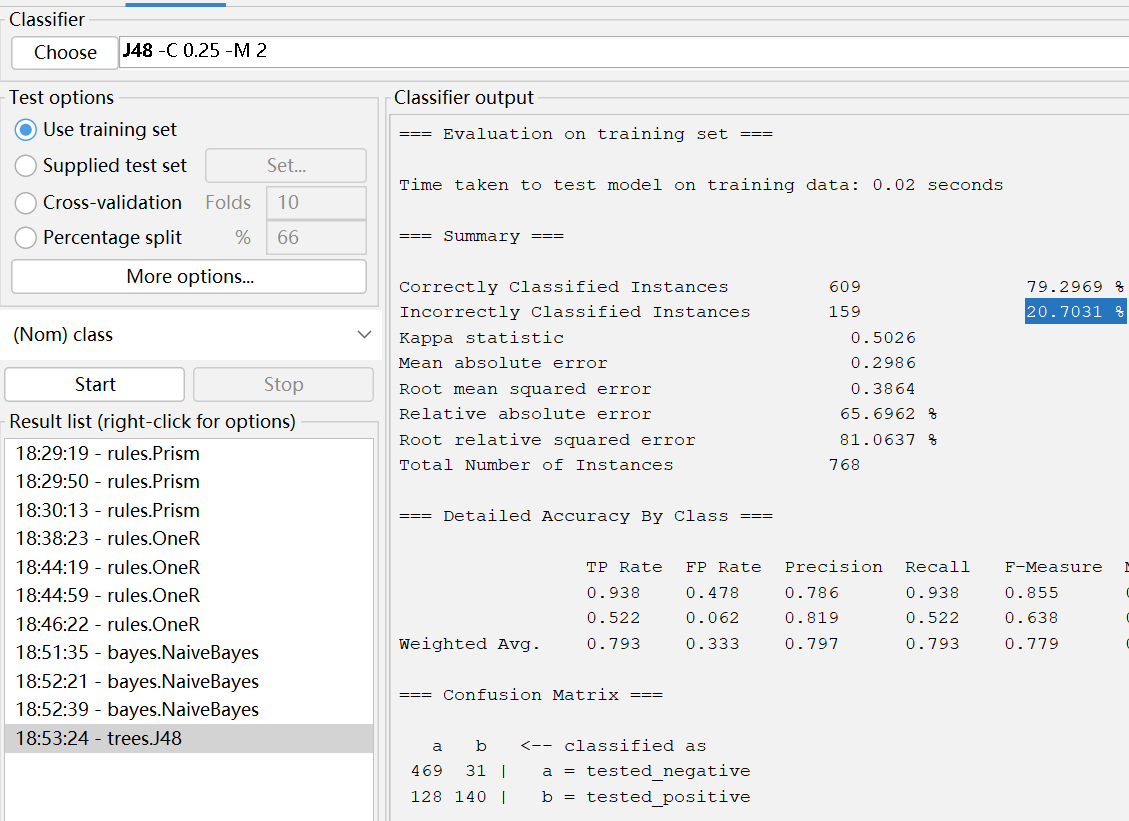
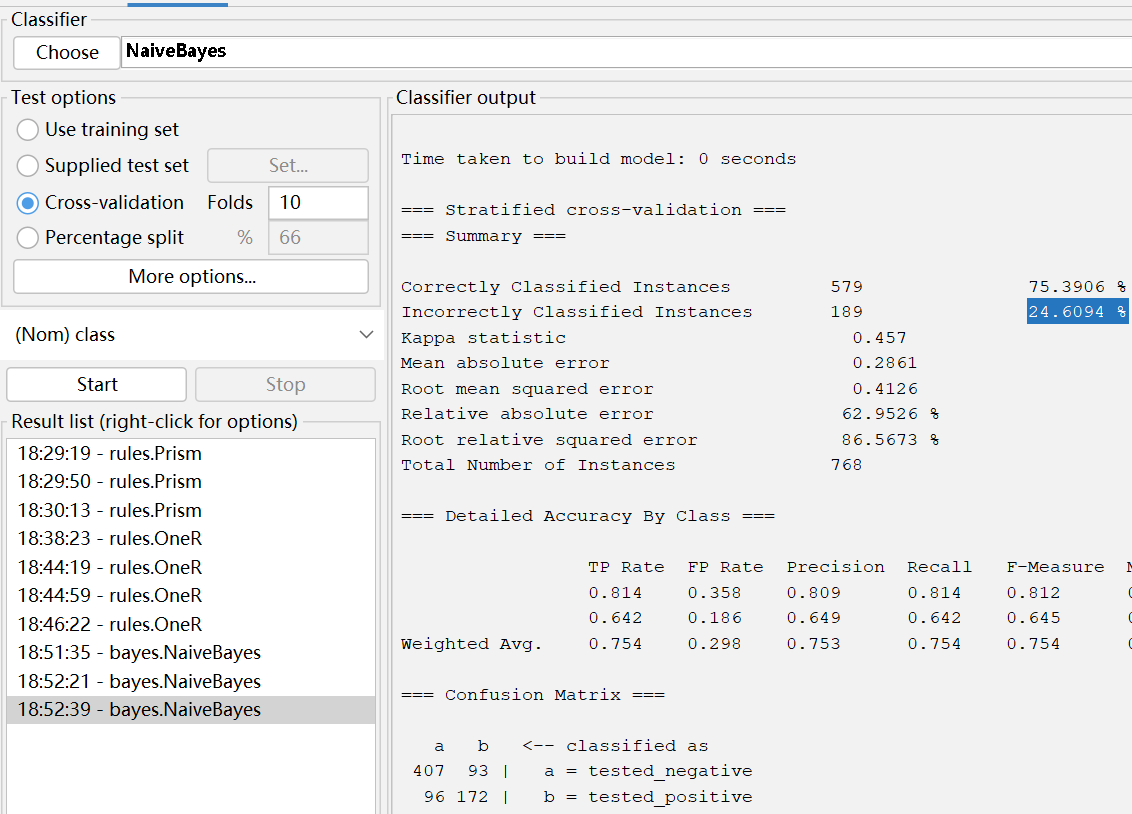
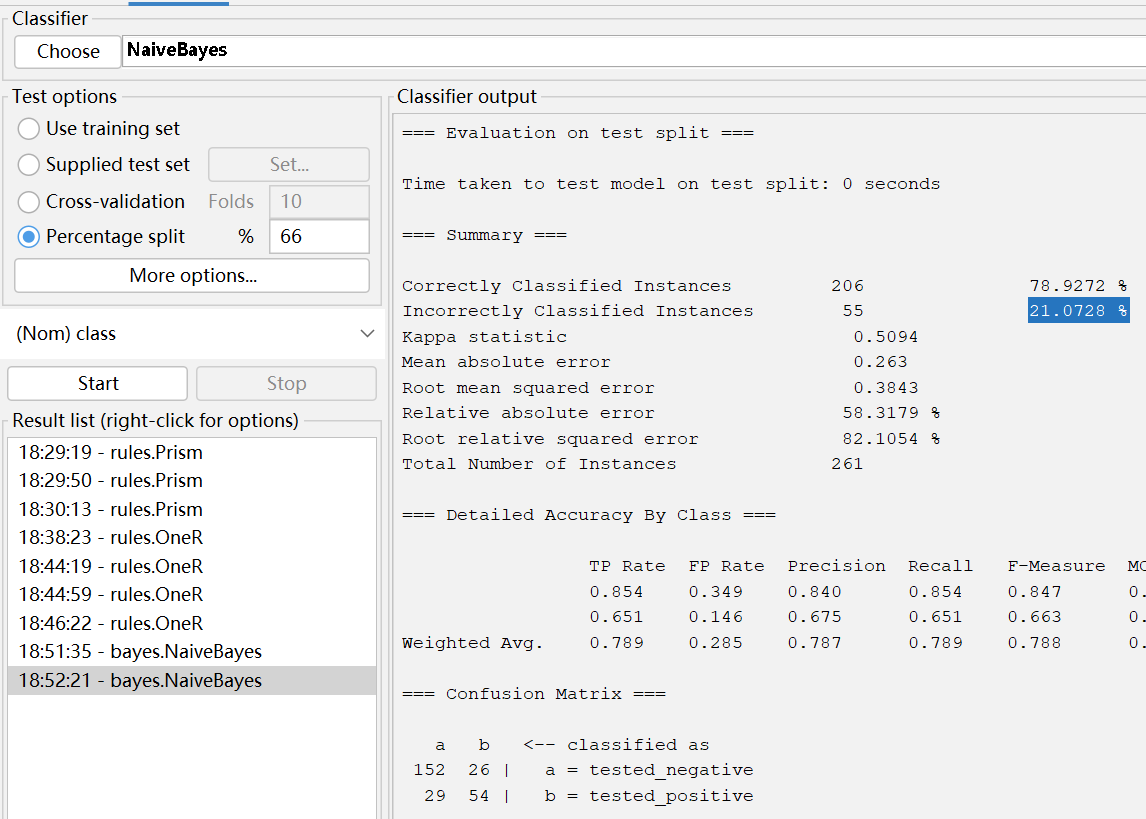
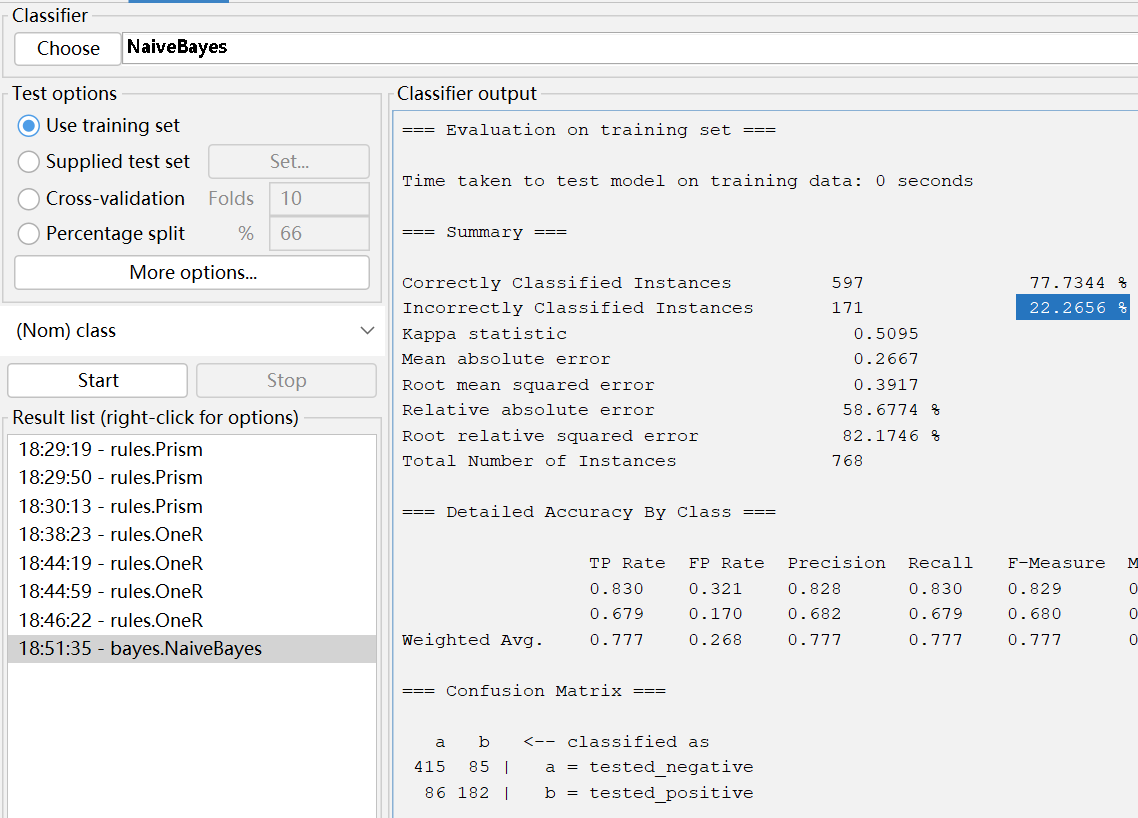
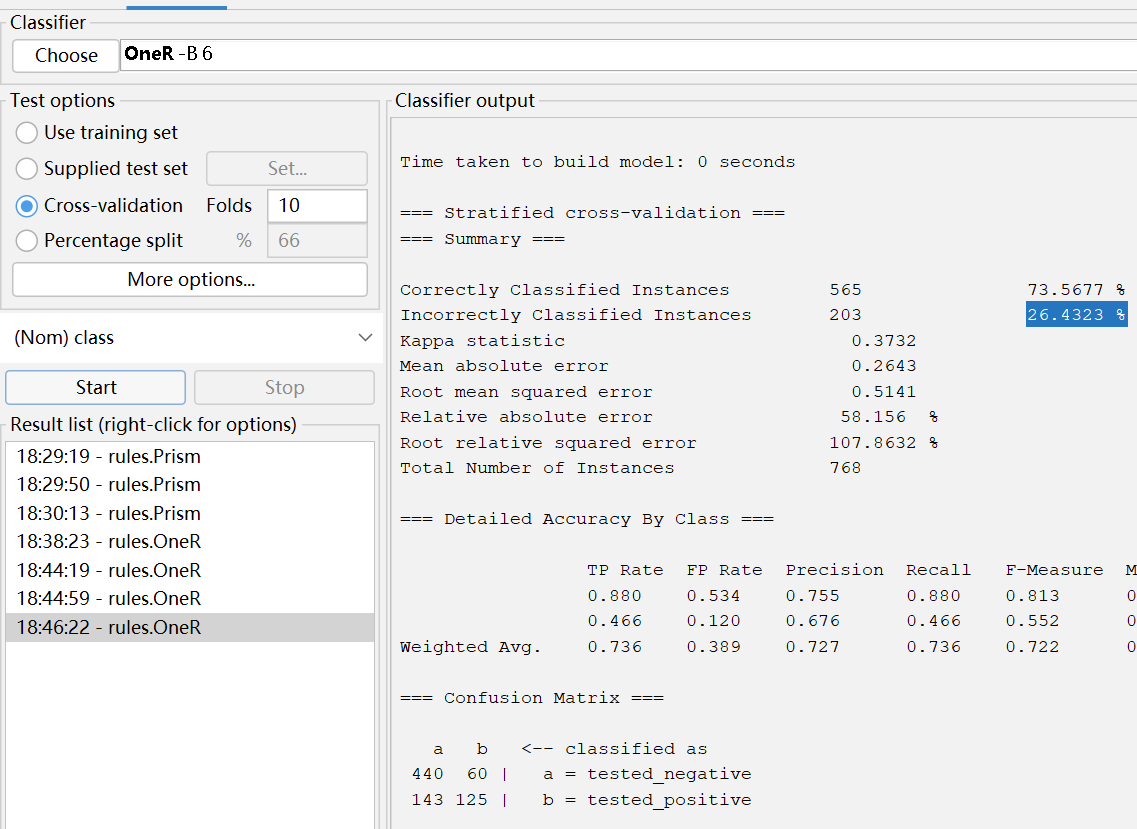
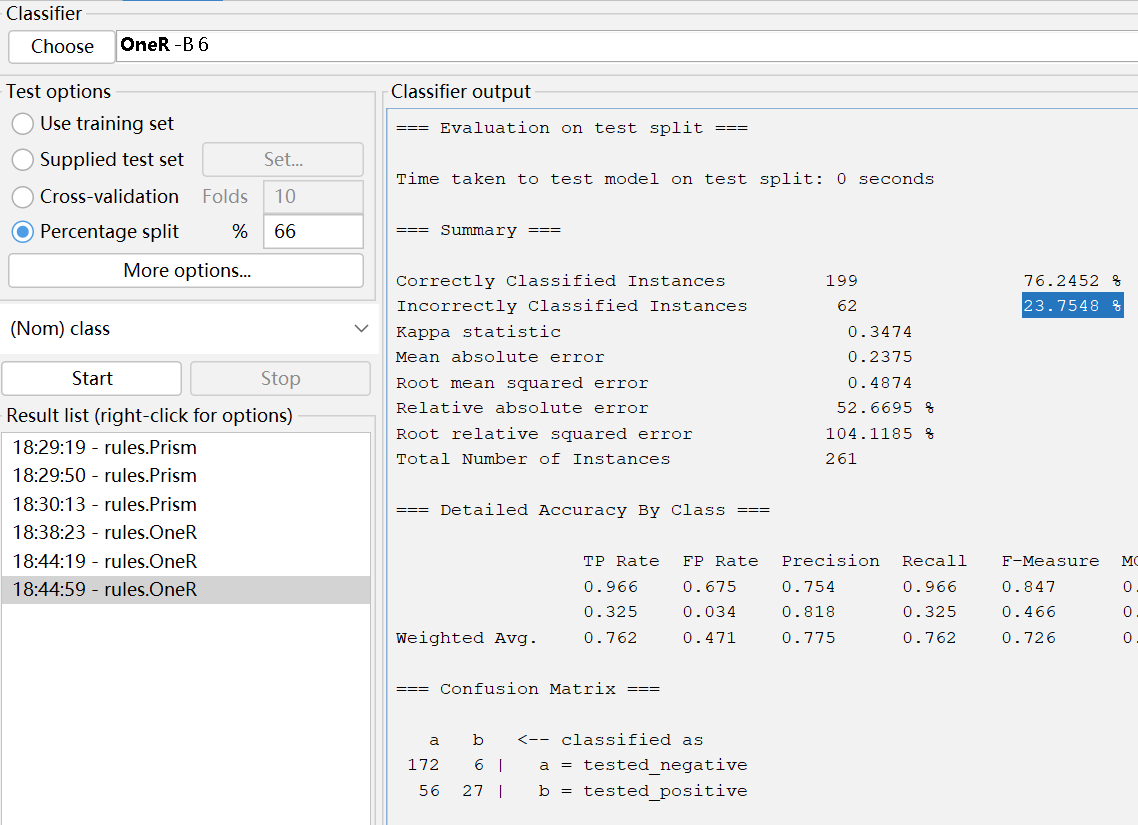
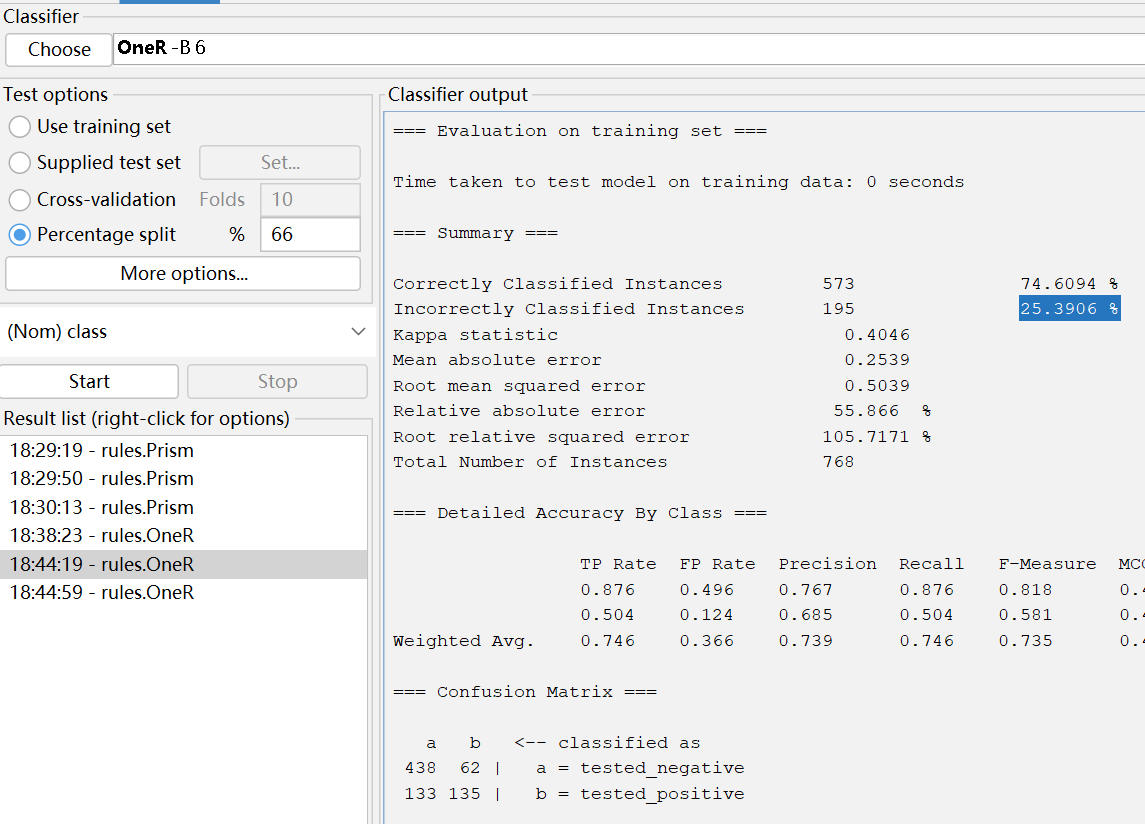


**Appendix:** screenshots for models

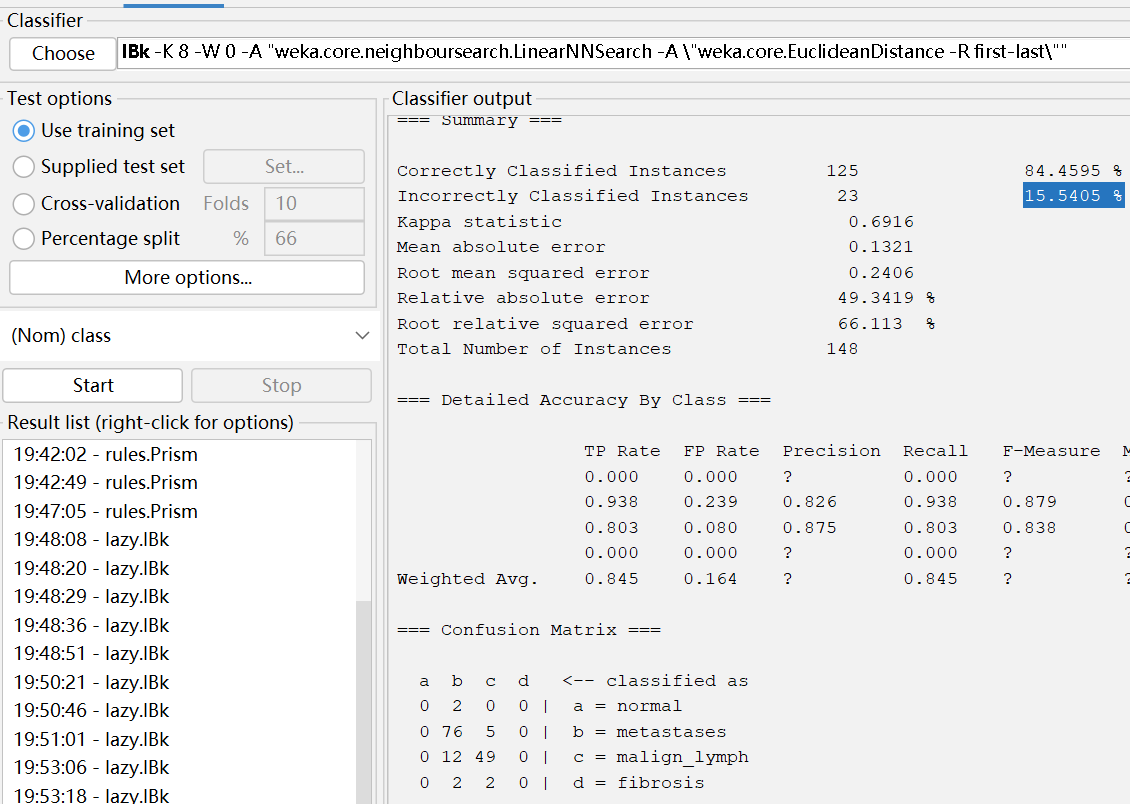
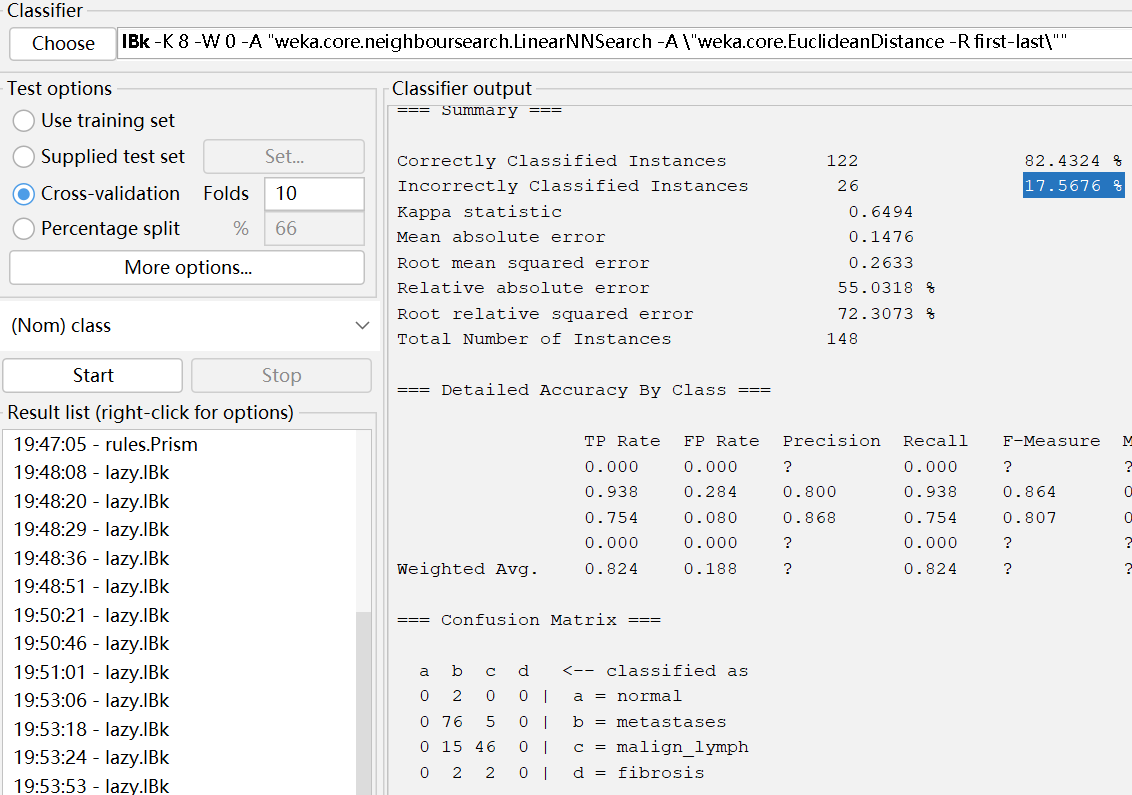
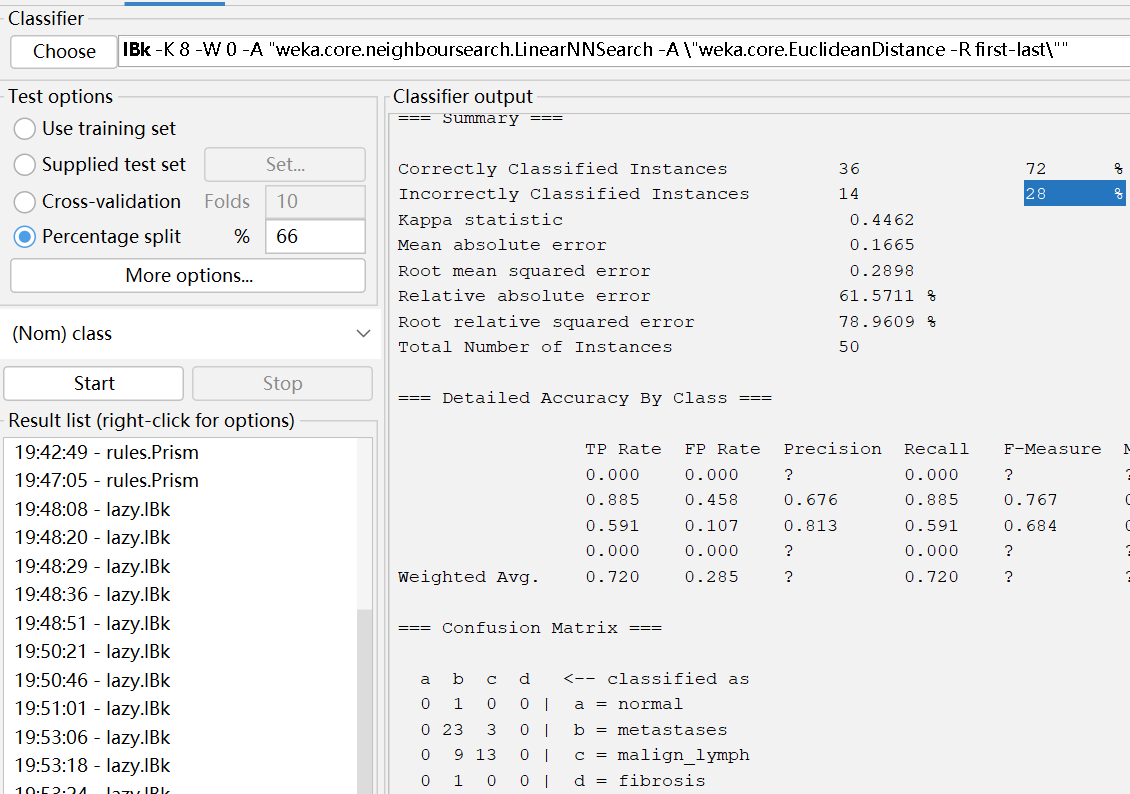
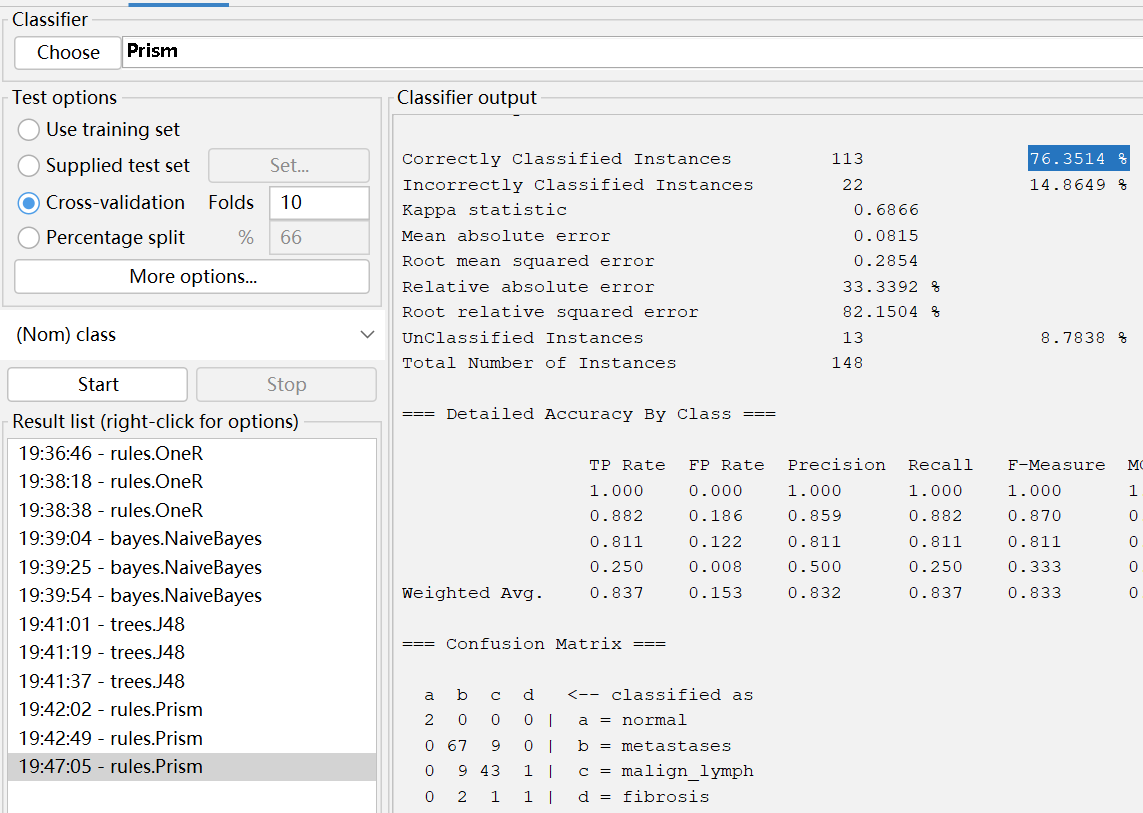
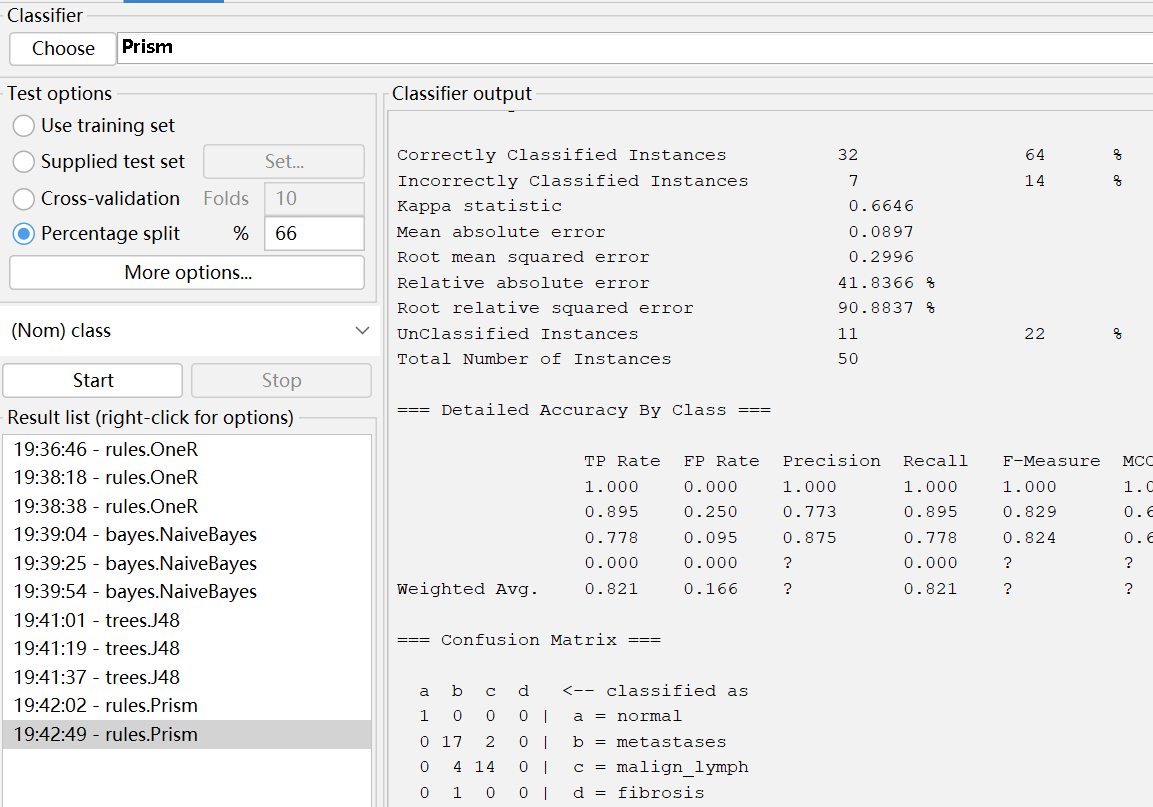
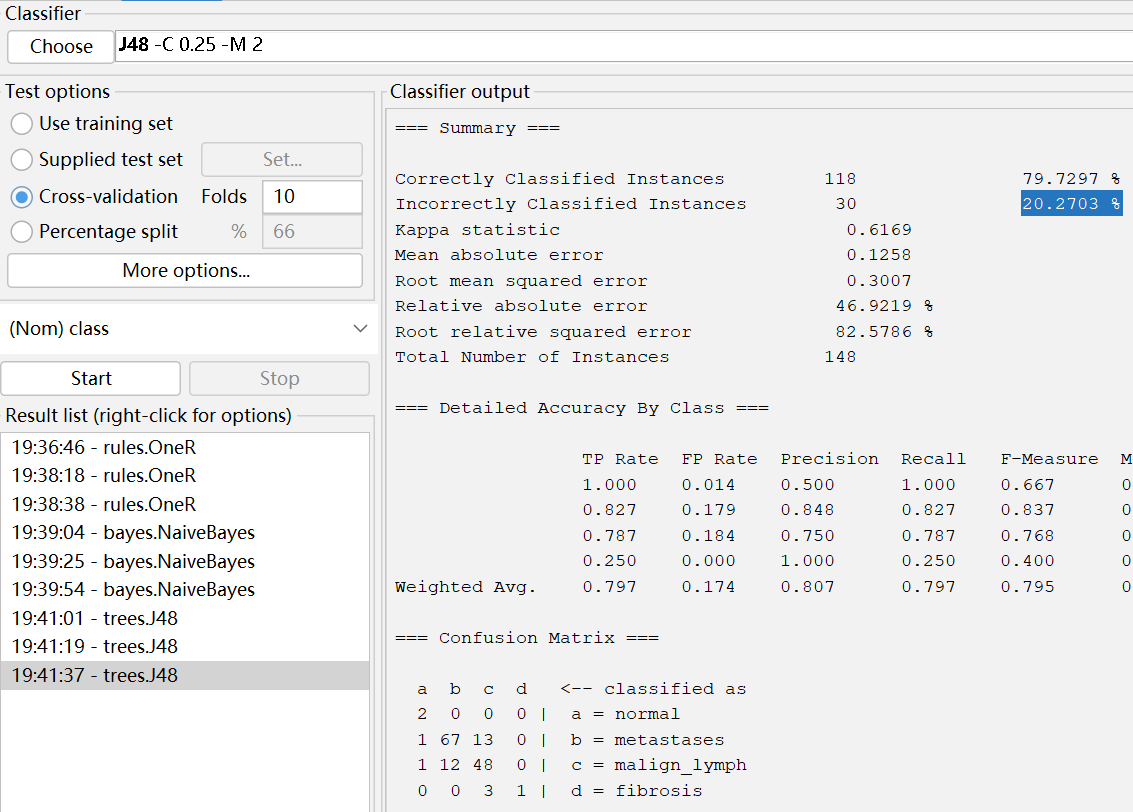
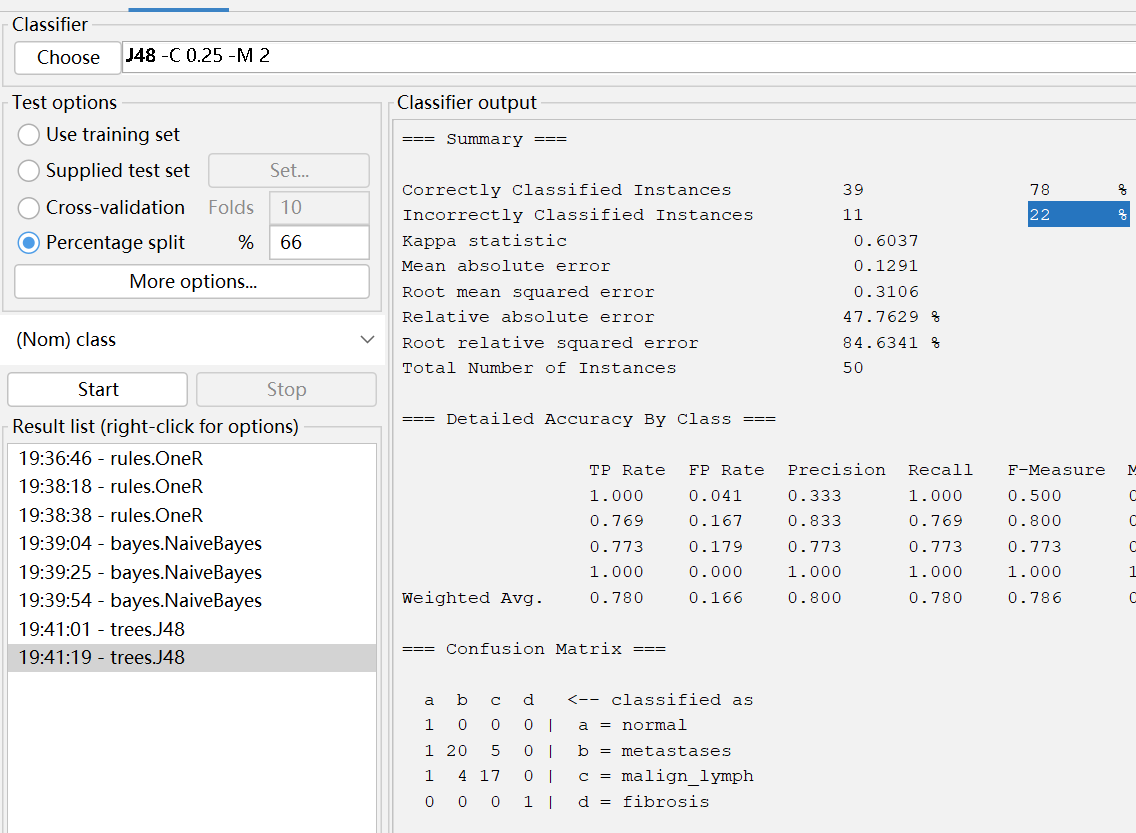
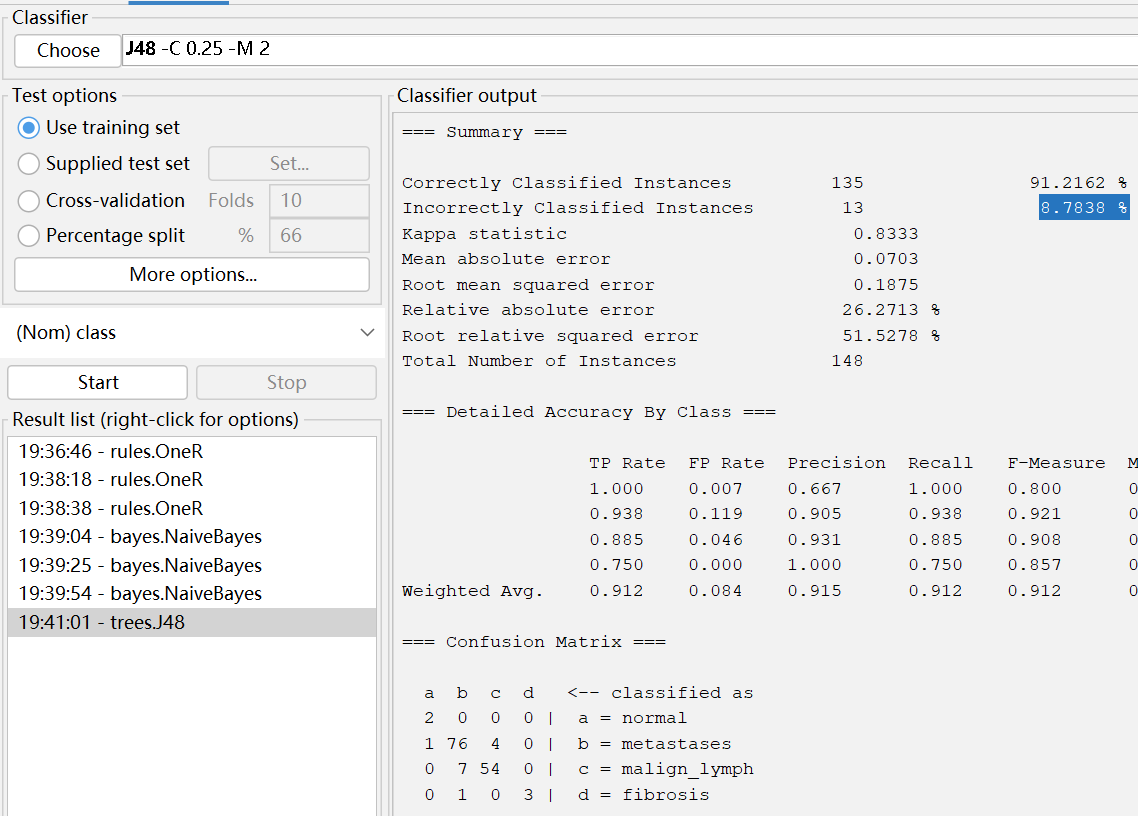
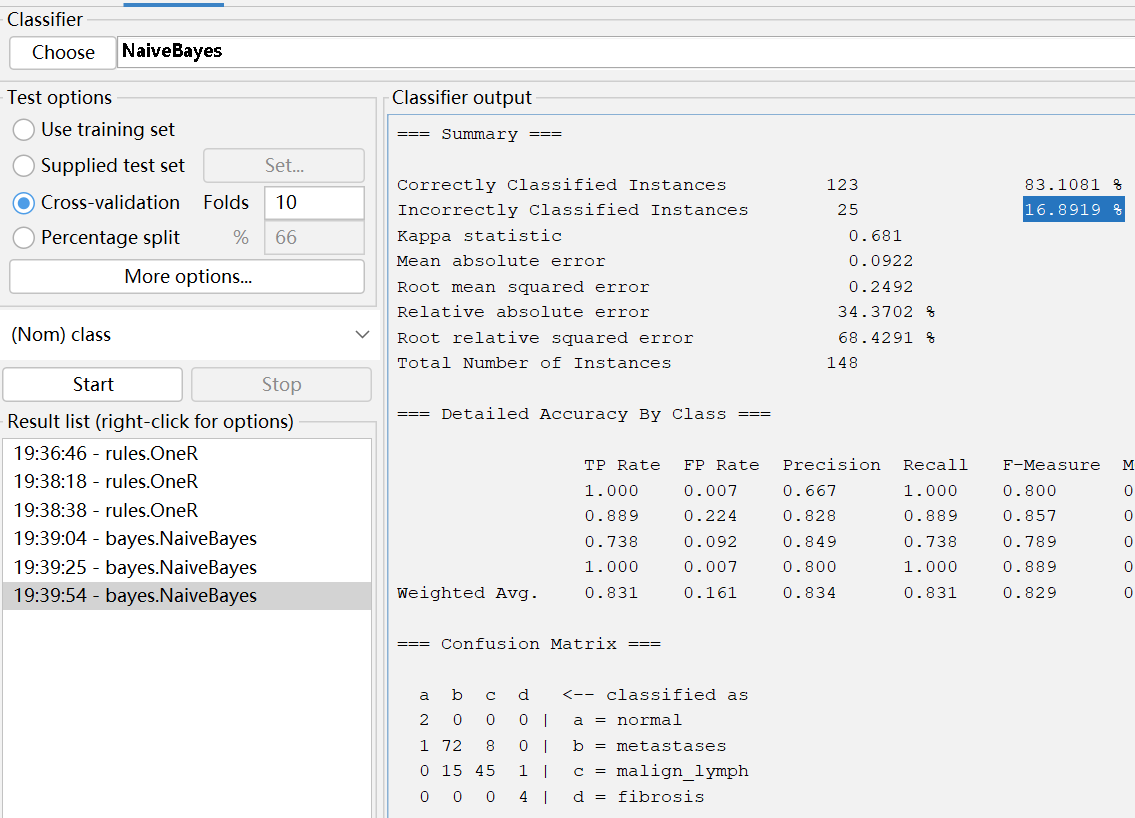
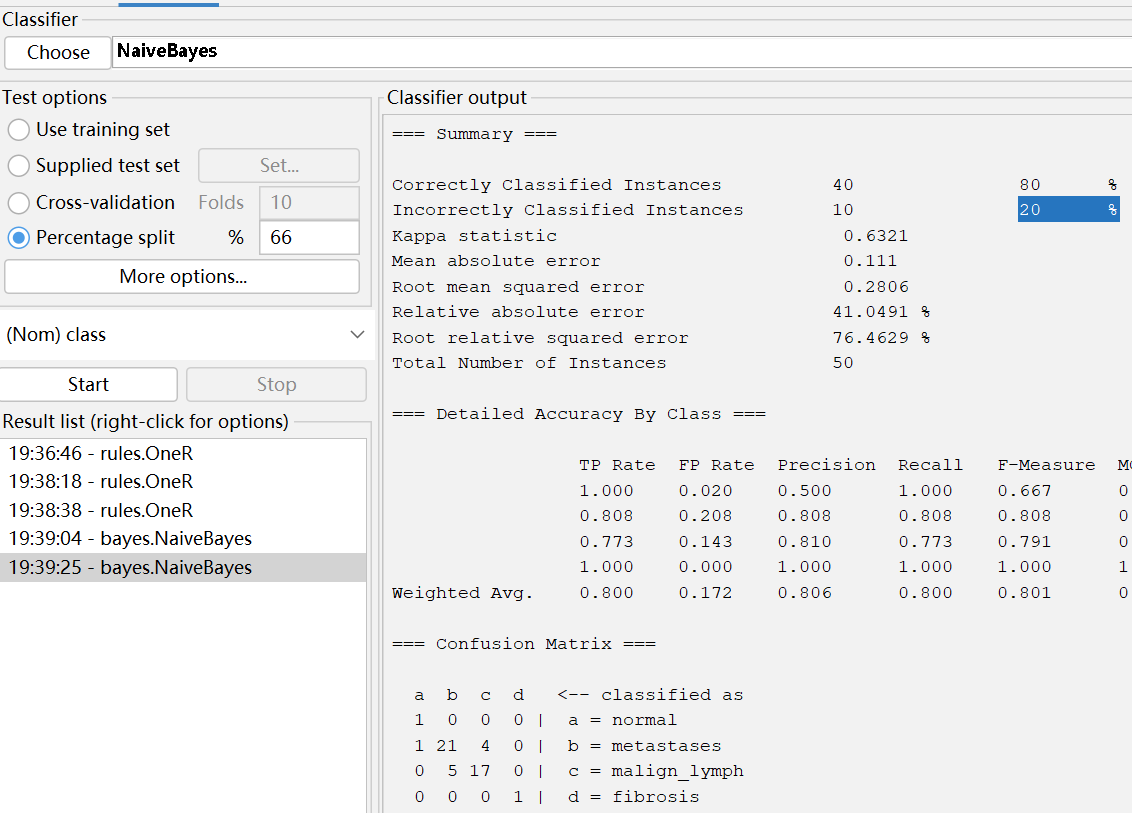
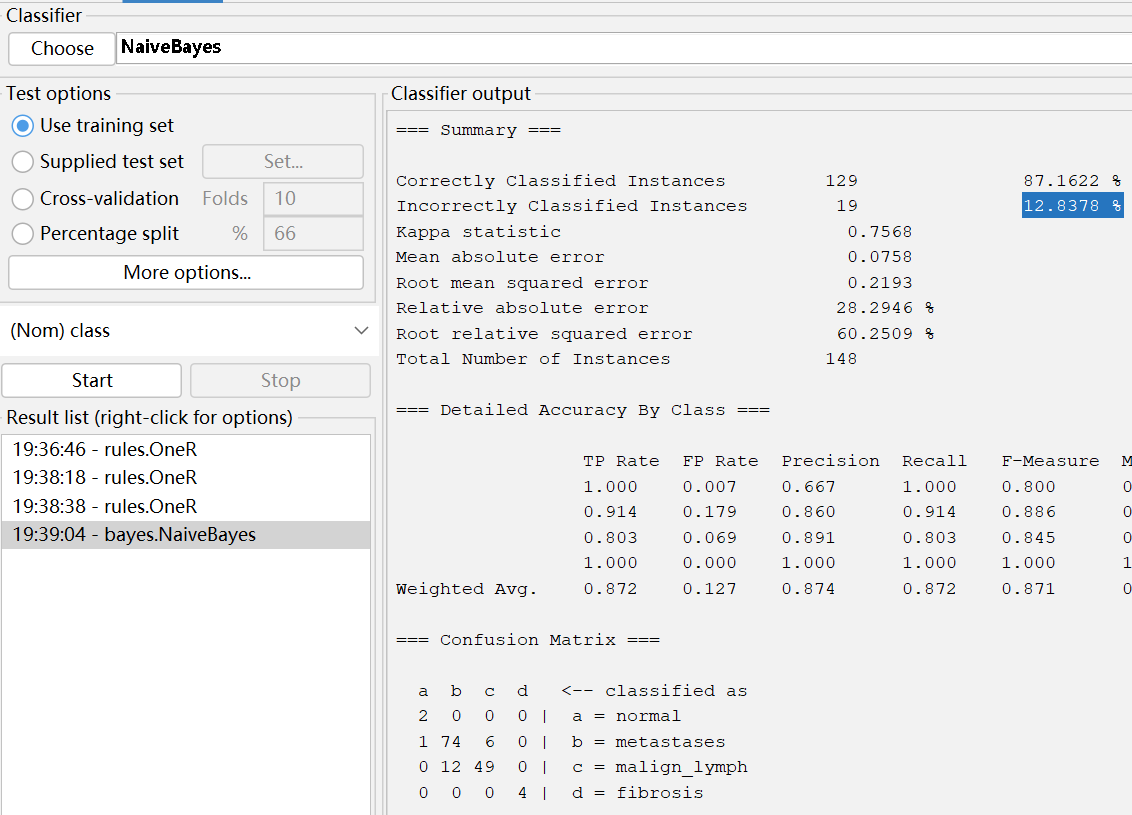
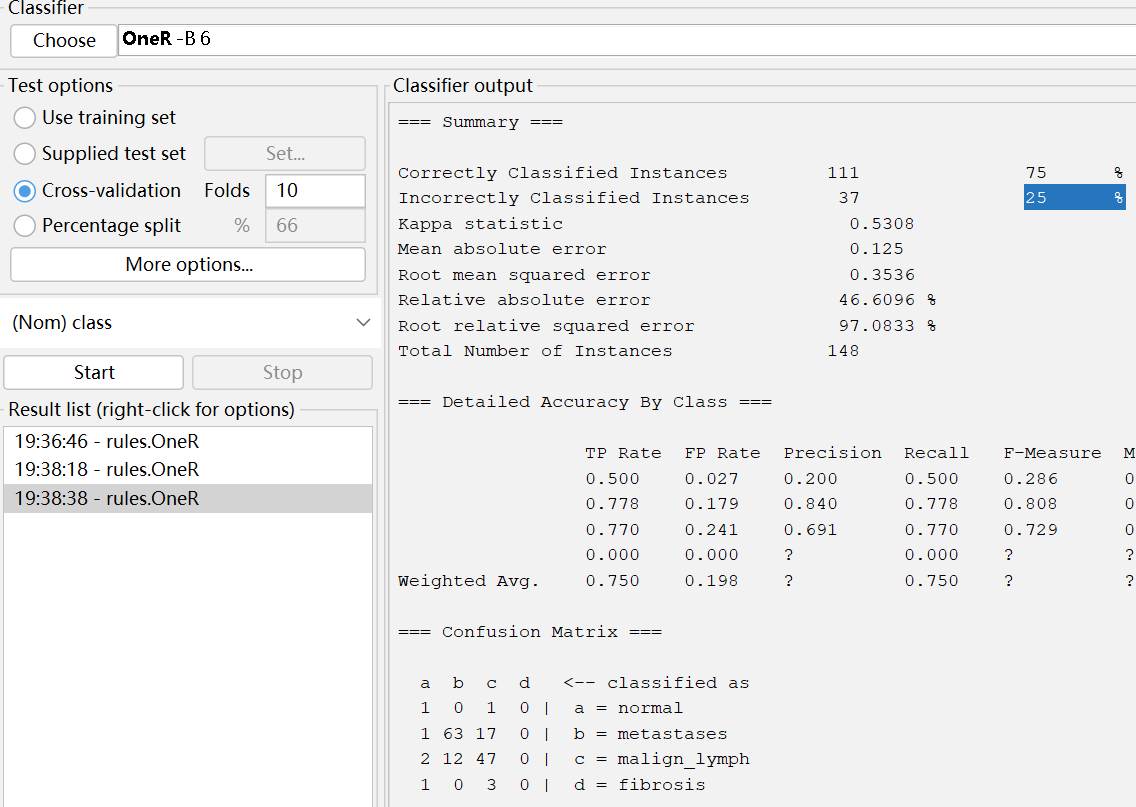
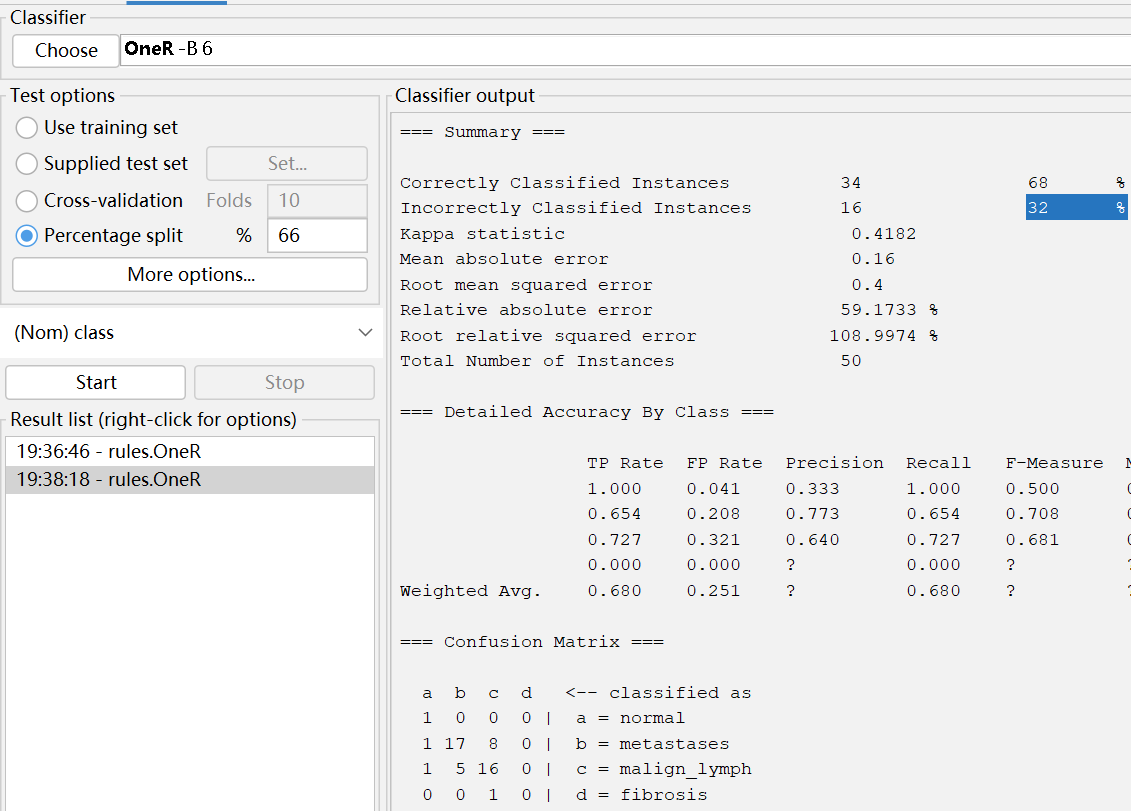
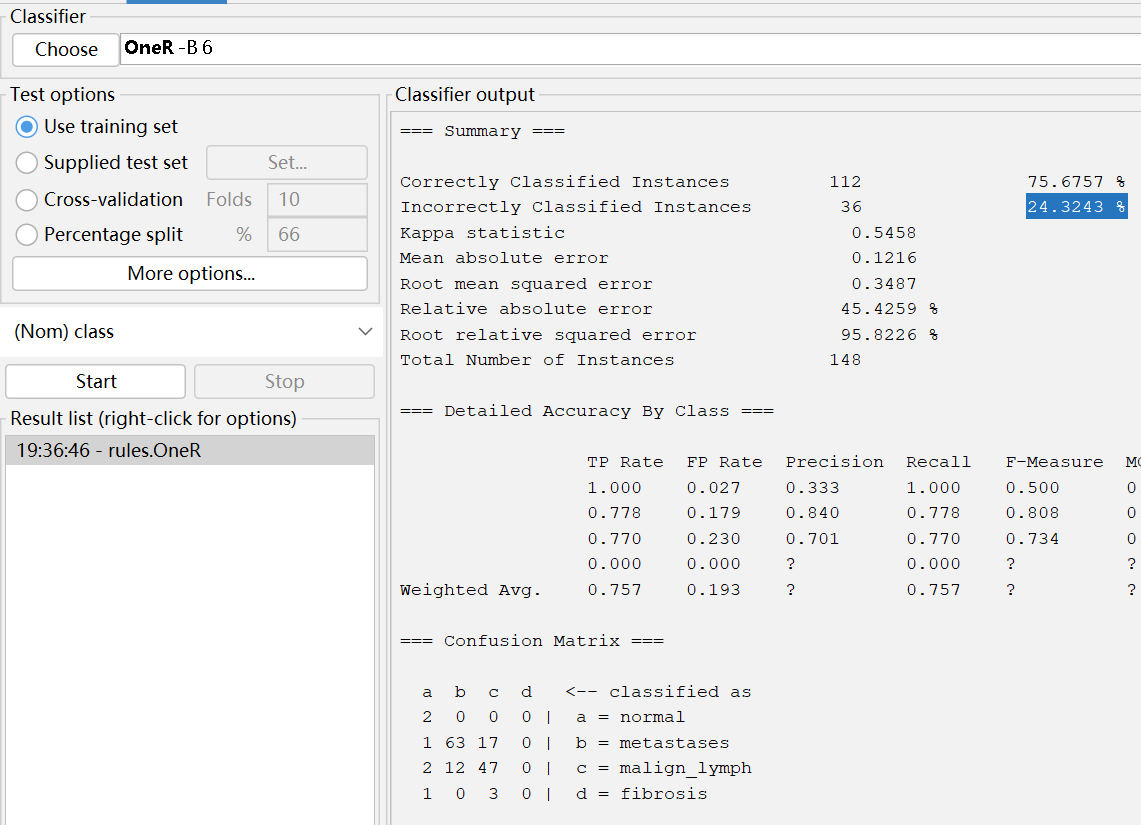
**Breast cancer:**



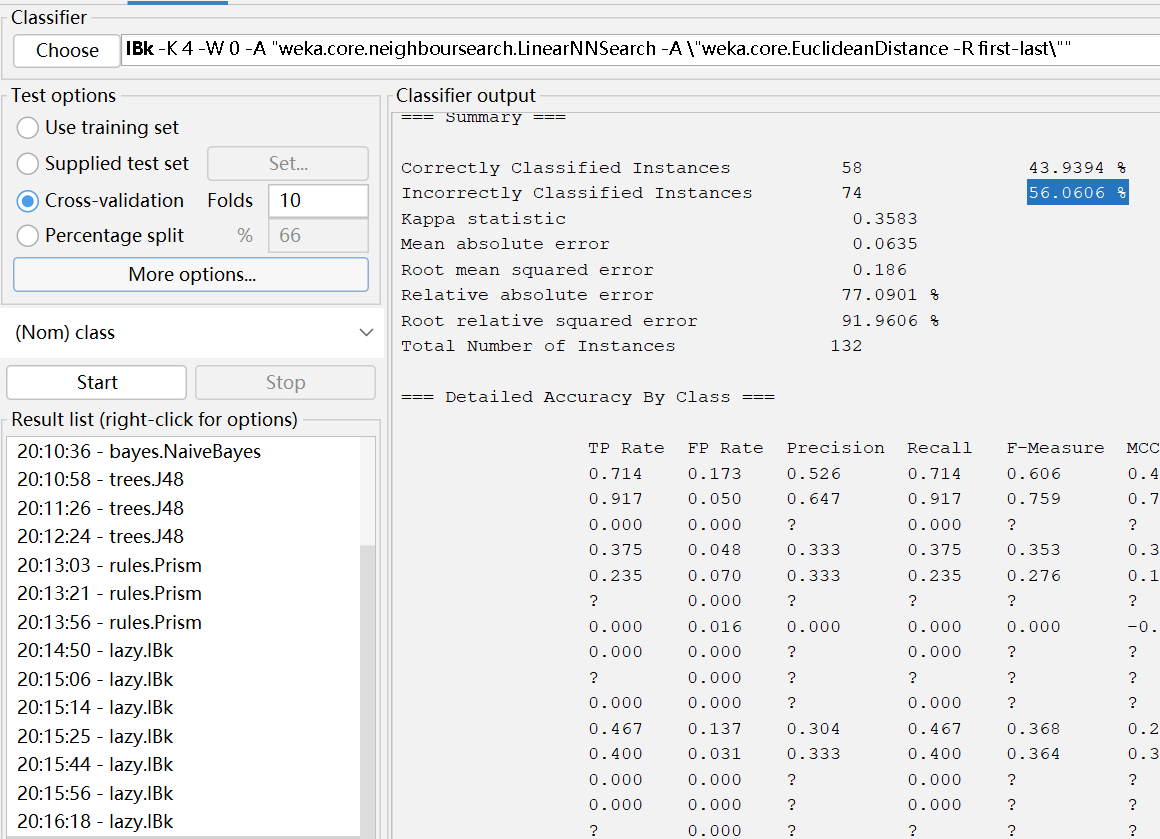
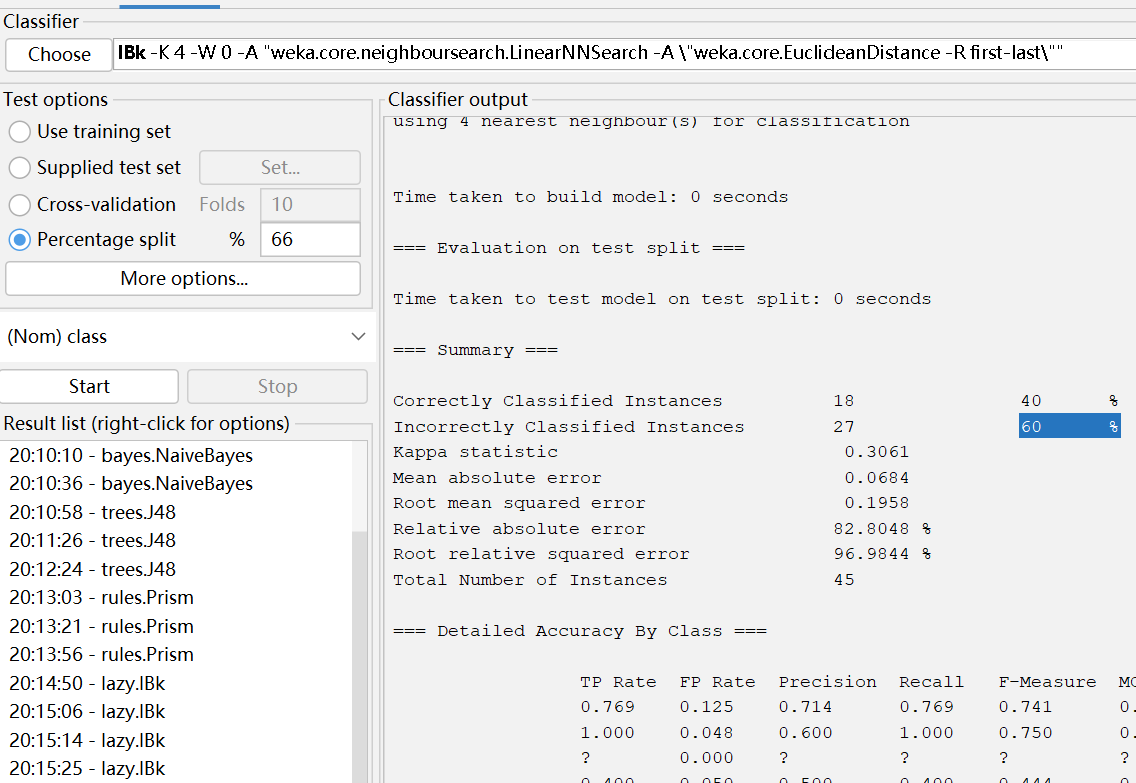
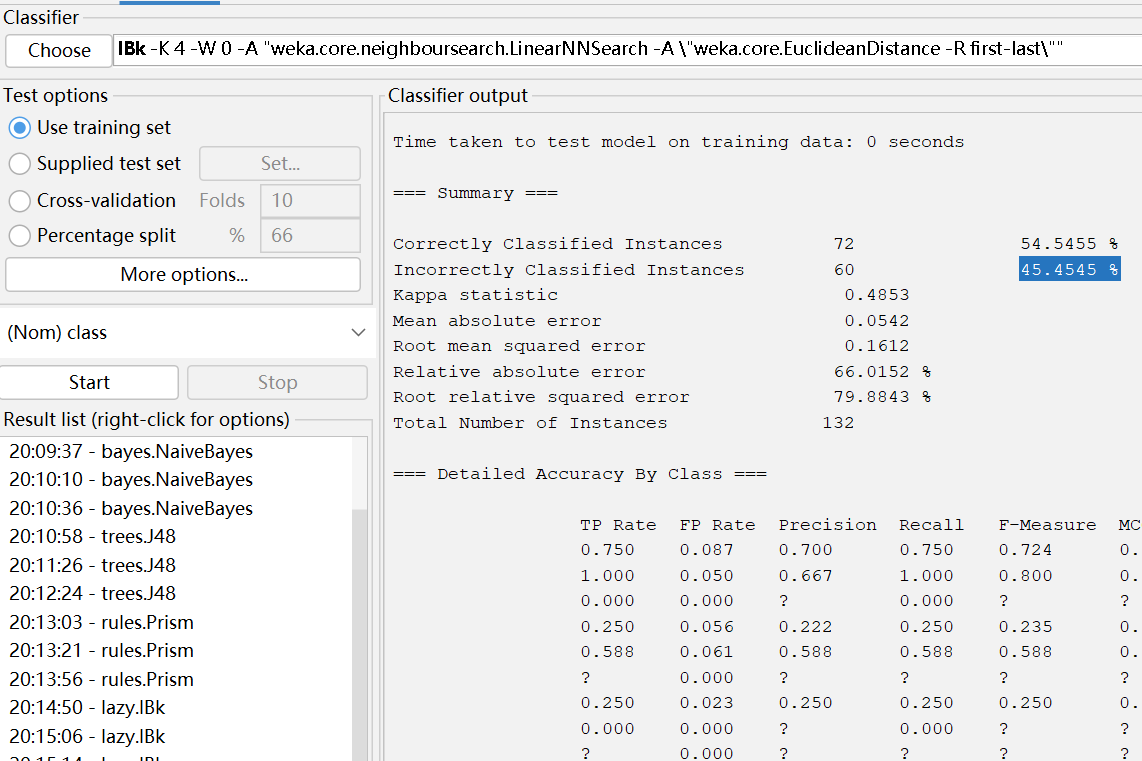
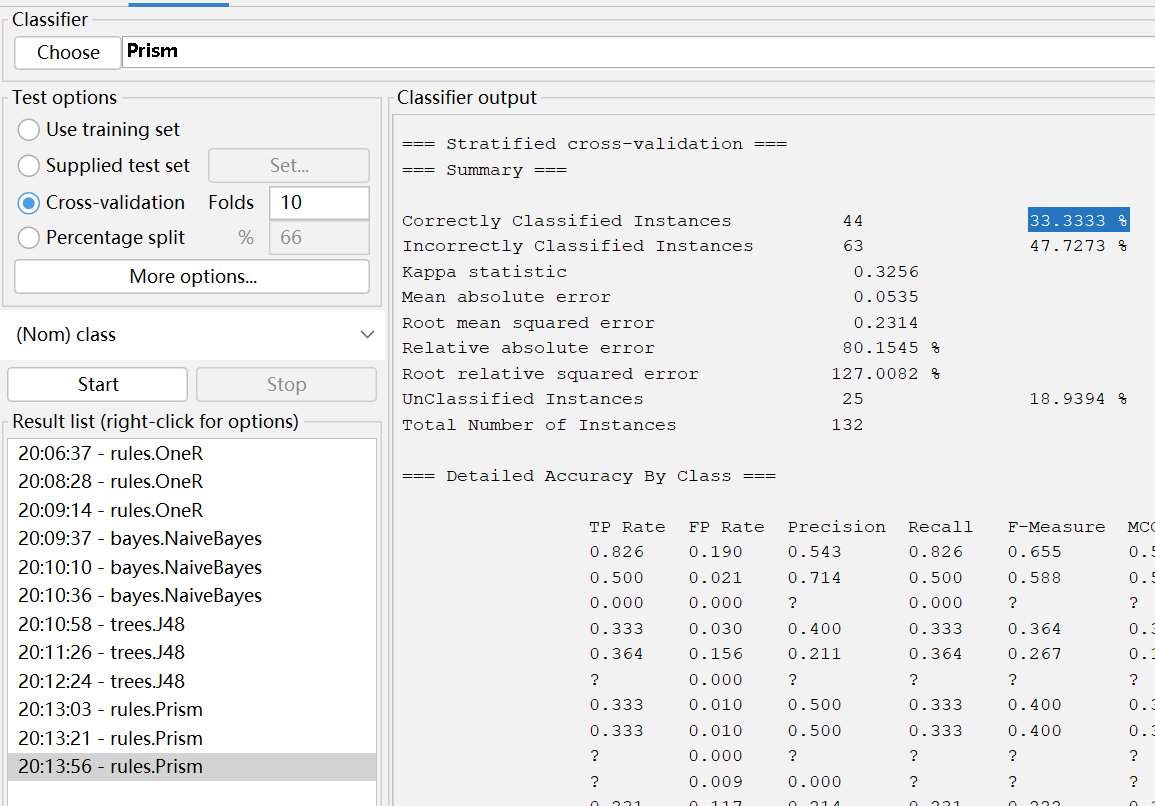
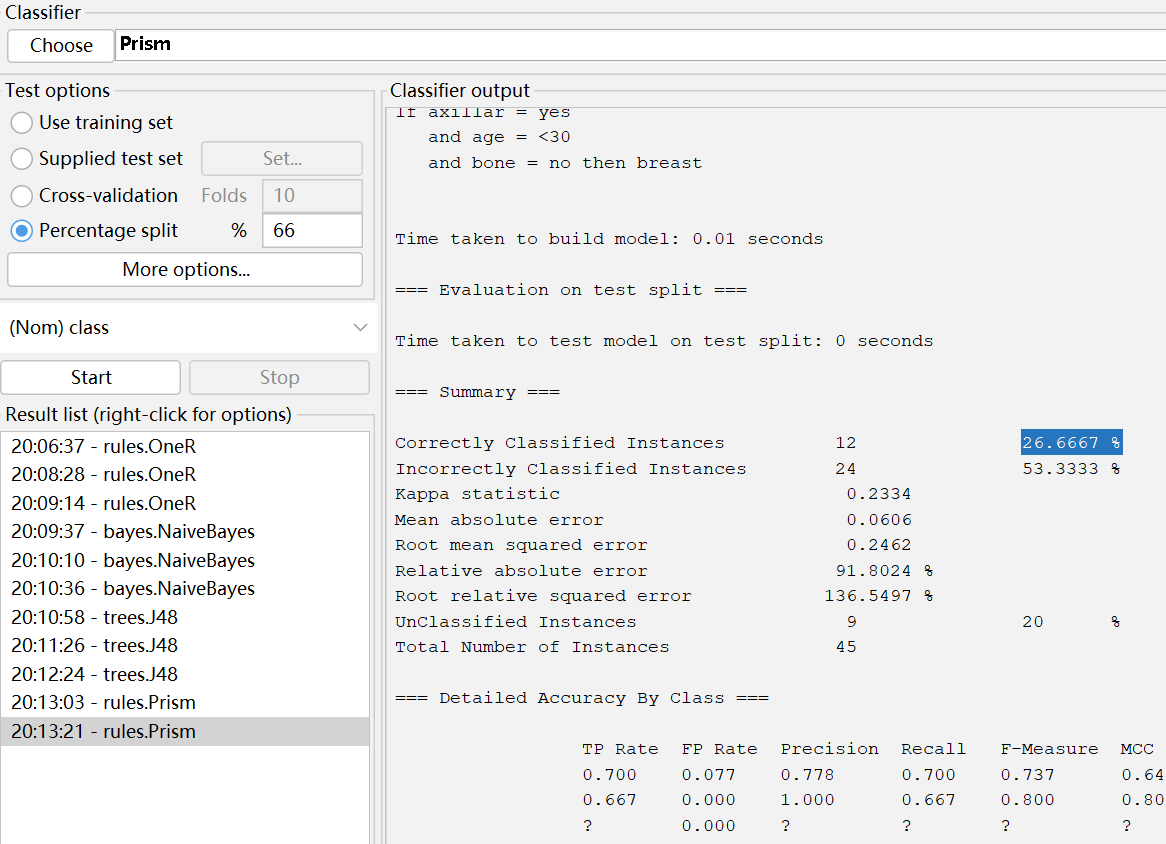
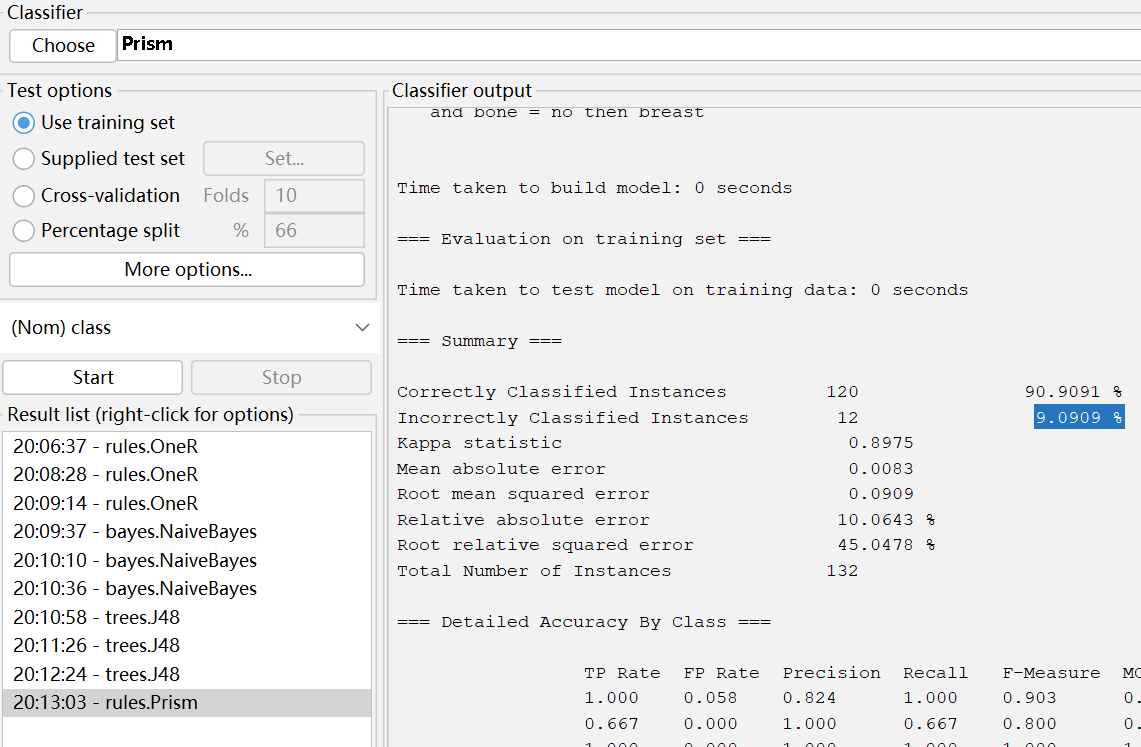
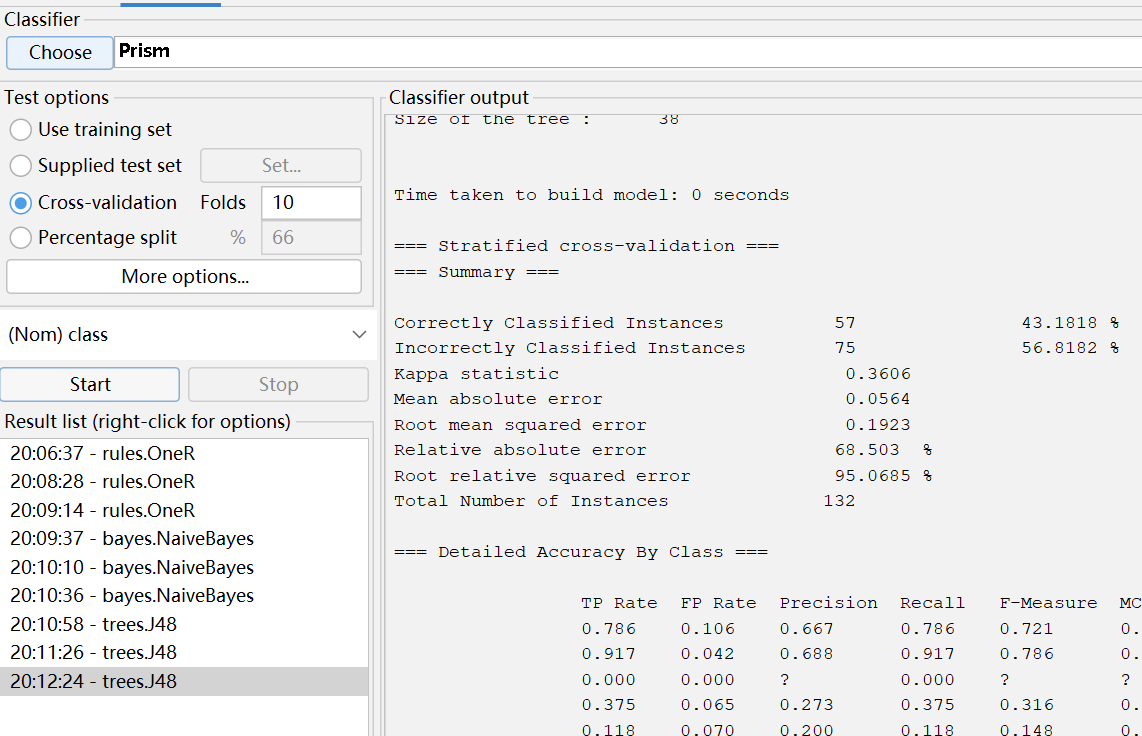
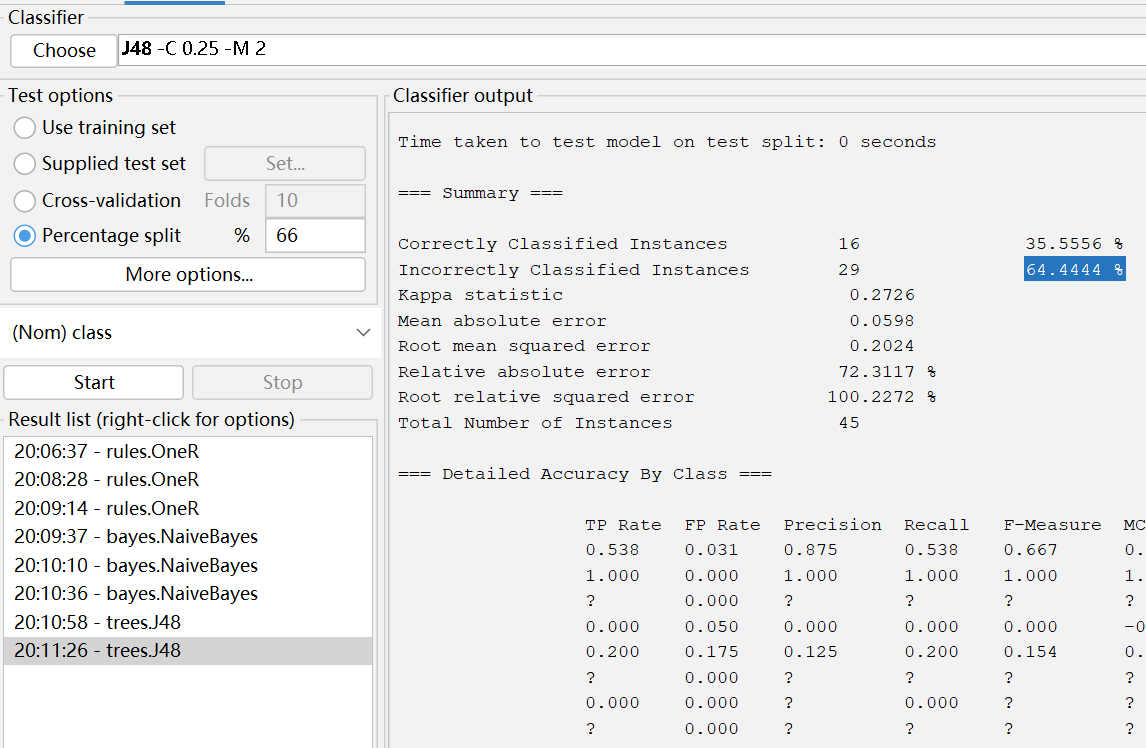
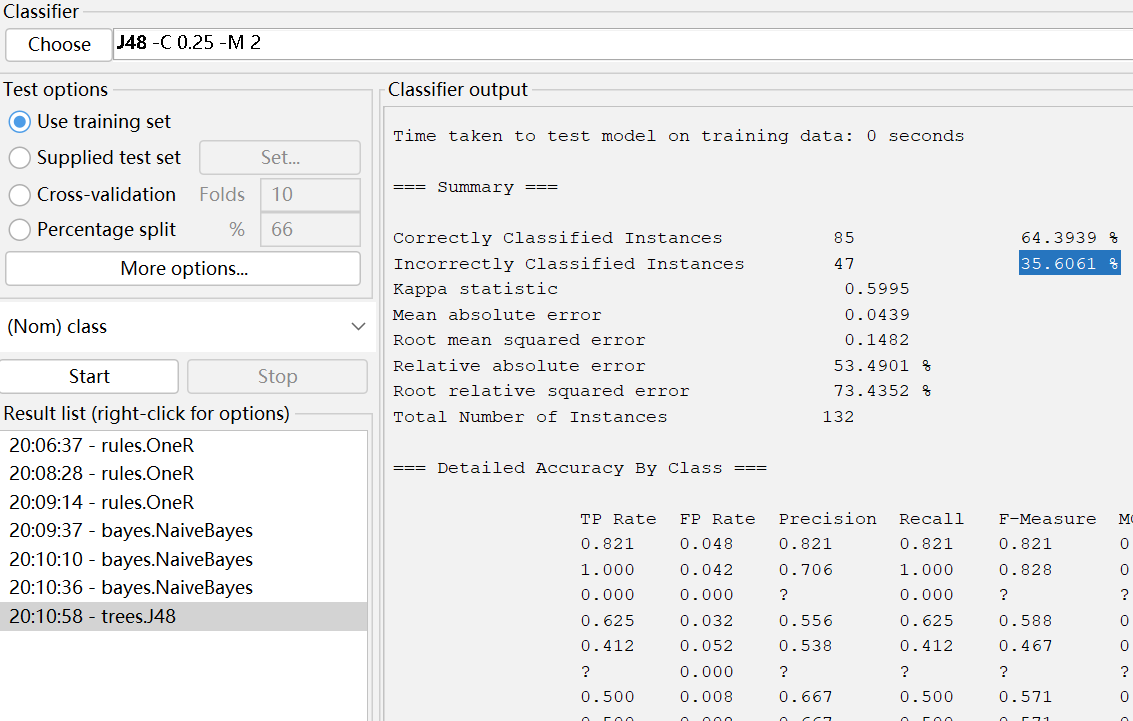
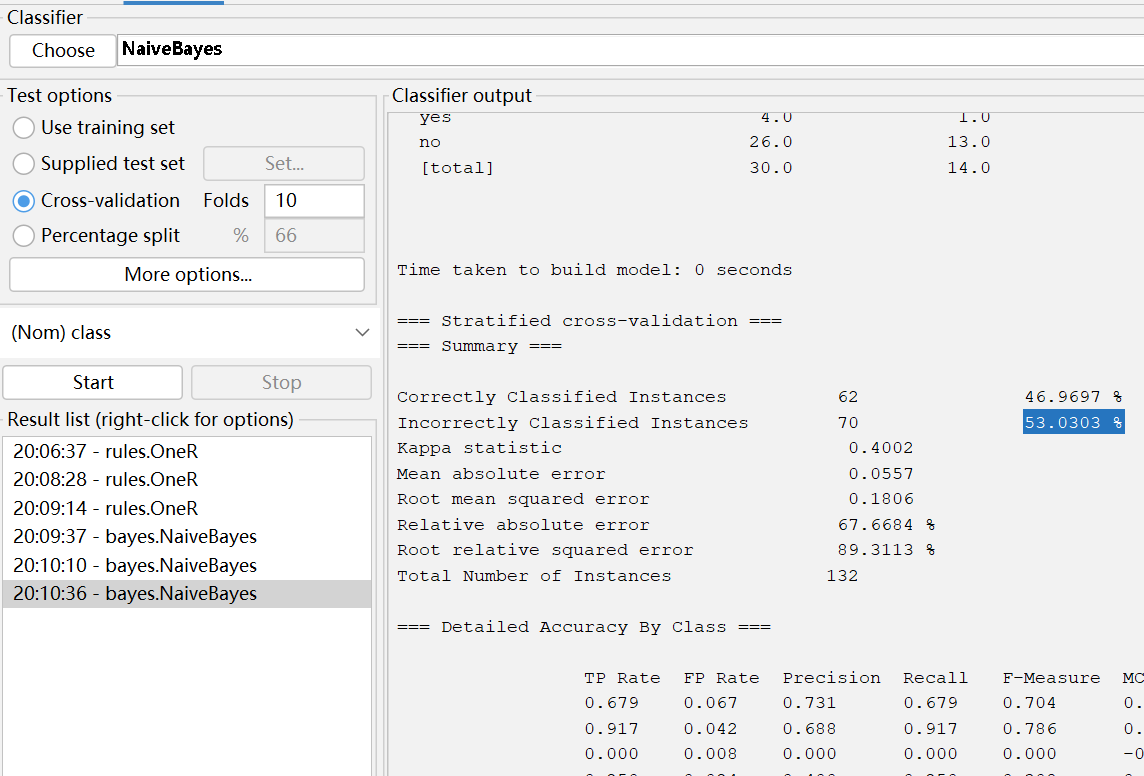
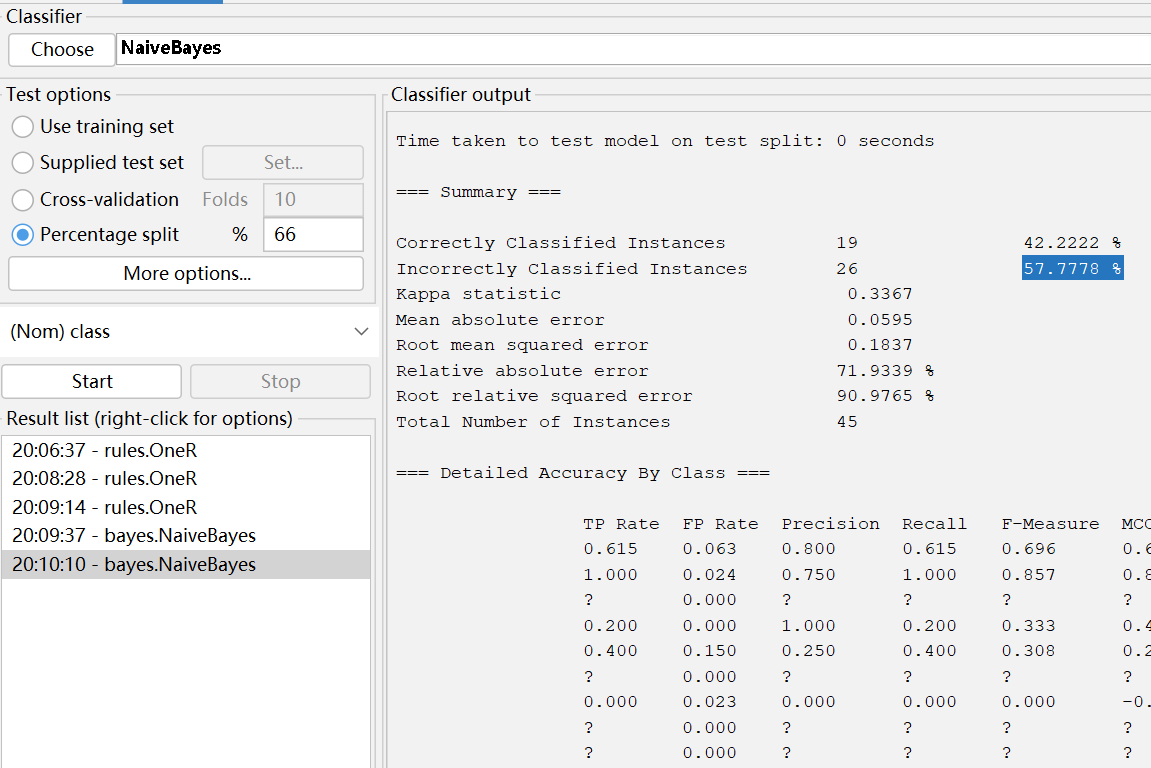
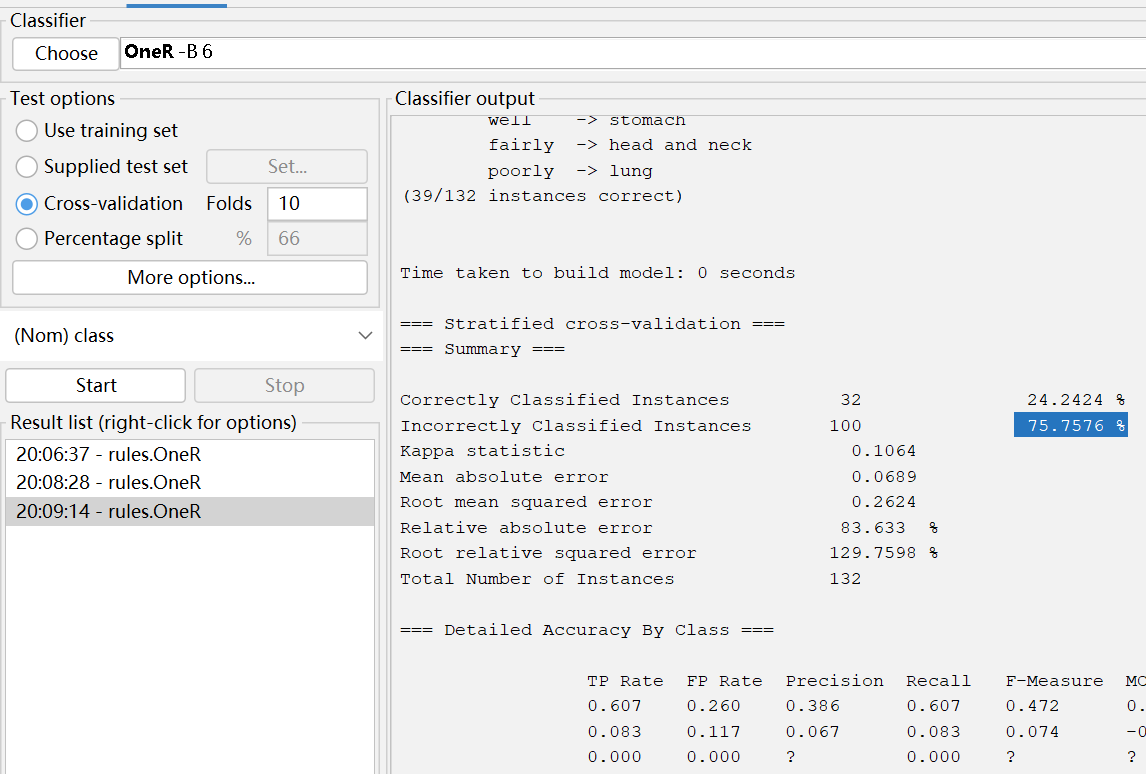
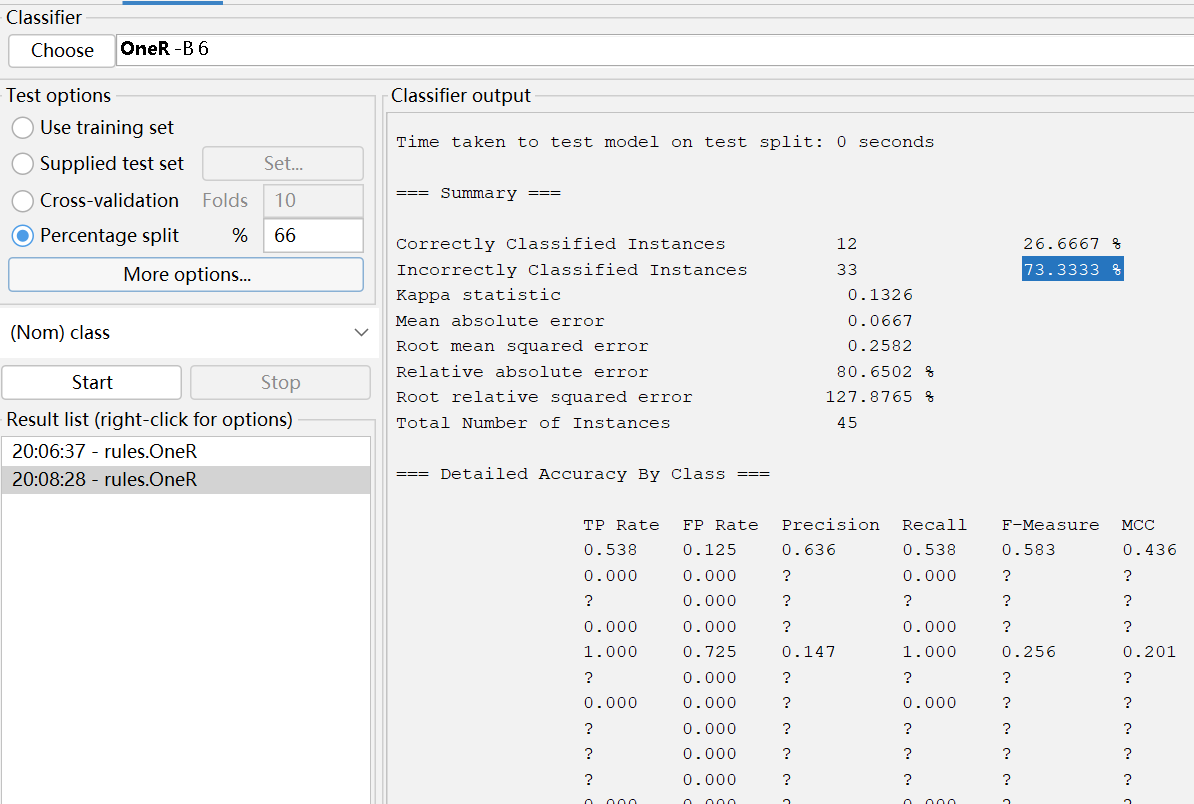
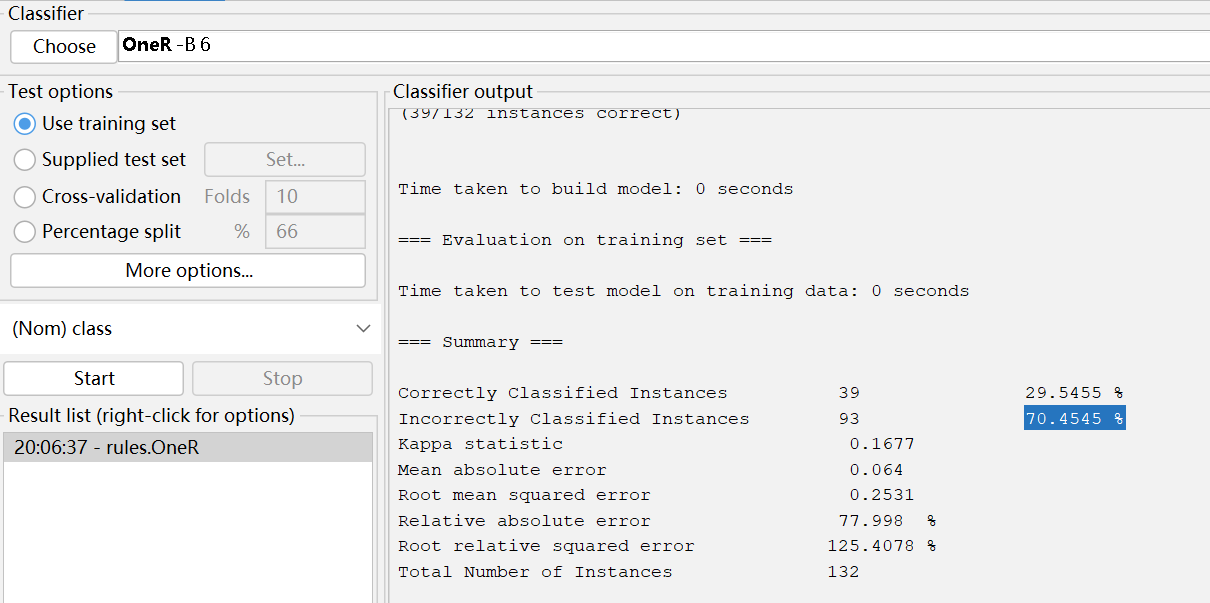
**Diabetes**



**Lymphography**



**Primary-tumor**



**Soybean**

