**Table of content**

[Introduction 3](#_Toc23867939)

[Description of the dataset 3](#_Toc23867940)

[Detailed description 3](#_Toc23867941)

[Data mining goal 3](#_Toc23867942)

[Names and meanings of all attributes in the dataset 3](#_Toc23867943)

[Data mining tools and algorithms 4](#_Toc23867944)

[Data mining procedures 5](#_Toc23867945)

[Data Preprocessing 5](#_Toc23867946)

[Handling missing values 5](#_Toc23867947)

[Splitting the dataset into training and test sets 7](#_Toc23867948)

[Balancing unbalanced data 7](#_Toc23867949)

[Results of data Preprocessing 8](#_Toc23867950)

[Selected Attributes and Classifiers 9](#_Toc23867951)

[Attribute Selection Algorithm Results: 11](#_Toc23867952)

[Data mining result and evaluation 11](#_Toc23867953)

[Naïve Bayes 11](#_Toc23867954)

[Logistic 14](#_Toc23867955)

[Neural Networks (Multilayer Perceptron) 16](#_Toc23867956)

[J48 + AdaBoostM1 19](#_Toc23867957)

[Justification for our selection of the best model 21](#_Toc23867958)

[What each team member did for this project 22](#_Toc23867959)

[Conclusion 23](#_Toc23867960)

[References 24](#_Toc23867961)

# **Introduction**

## Description of the dataset

Dataset Name: *Mortality Prediction after the First Year of Kidney Transplantation*

The source of the dataset: [*https://plos.figshare.com*](https://plos.figshare.com)

Number of tuples: *3439*

Number of attributes: *31*

The class attribute: *DEATH\_FUNCTGRAFT - Death with a functioning graft: 1 = Yes, 0 = No*

## Detailed description

The dataset contains the history of kidney transplantation, specifically, medical characteristics of recipients and donors, as well as the time duration between the first anniversary of the transplantation and the patient's death with and without a functioning graft. The dataset was taken from the DIVAT (Données Informatisées et VAlidées en Transplantation) cohort containing the data of the patients transplanted between 2000 and 2012 in 6 French centers [1](Lorent M., Giral m.).

## Data mining goal

Our data mining goal is to predict the chances of having successful functioning graft transplantation (e.g. changes of whether or not patients’ body will accept well the functioning graft) based on recipients’ and donors’ characteristics by analyzing both records of patients with a graft who died because of the failure of a graft or other reasons such as car accident after first year of kidney transplantation. We want to build a prediction model for kidney transplantation: 0 is successful kidney transplantation, 1 indicates failure of the kidney functioning graft. The reason is that identification and quantification of the relevant factors for death can improve patients’ individual risk assessment and decision-making.

## Names and meanings of all attributes in the dataset

|  |  |
| --- | --- |
| ID | Patient number |
| TRANSP\_AGE\_R | Age of the recipient at transplantation (years) |
| BMI\_R | Body Mass Index of the recipient at transplantation (kg/m²) |
| TIME\_REG\_TRANSP | Time between registration on waiting list and transplantation (years) |
| TIME\_DIAL\_TRANSP | Time between first dialysis and transplantation (years) |
| CIT\_HOUR | Cold Ischemia Time (hours) |
| 1YR\_CREAT | Serum creatinine at 1-year post-transplantation (µmol/L) |
| SEX\_R | Recipient sex: M = Male, F = Female |
| TRANSP\_RANK | Rank of the kidney transplantation |
| INIT\_DISEASE | Initial disease: 1 = Recurrent disease, 0 = Other) |
| TECH\_DIAL | Technique of dialysis: 1 = Pre-emptive, 2 = Peritoneal dialysis, 3 = Hemodialysis |
| CMV\_R | Cytomegalovirus infection of the recipient at transplantation: 1 = Yes, 0 = No |
| HCV\_R | Hepatitis C Virus infection of the recipient at transplantation: 1 = Yes, 0 = No |
| DIAB | History of diabetes: 1 = Yes, 0 = No |
| HTA | History of high blood pressure: 1 = Yes, 0 = No |
| CARDIOVASC | History of cardiovascular event: 1 = Yes, 0 = No |
| ANGINA | History of cardiac angina: 1 = Yes, 0 = No |
| DYSLIP | History of dyslipidemia: 1 = Yes, 0 = No |
| NEOPLASIA | History of neoplasia: 1 = Yes, 0 = No |
| INCOMP\_ABDR | Number of HLA ABDR incompatibilities |
| NODAT | New Onset Diabetes After Transplantation: 1 = Yes, 0 = No |
| 1YR\_PROTU24h | Daily proteinuria at 1-year post-transplantation (g) |
| AGE\_D | Donor age (years) |
| CREAT\_D | Serum creatinine of the donor (µmol/L) |
| SEX\_D | Donor sex: M = Male, F = Female |
| RELAT\_DR | Relation between donor and recipient: 0 = Living donor, 1 = Deceased donor |
| CAUSE\_DEATH\_D | Cause of the donor death: 1 = Vascular death, 0 = Other |
| CMV\_D | Hepatitis C Virus infection of the donor: 1 = Yes, 0 = No |
| EBV\_D | Epstein-Barr Virus infection of the donor: 1 = Yes, 0 = No |
| TIME\_TRANSP\_DEATH | Time between transplantation and death with a functioning graft (days) |
| DEATH\_FUNCTGRAFT | Death with a functioning graft: 1 = Yes, 0 = No |

**Table 1. Names and meanings of all attributes in the dataset**

## Data mining tools and algorithms

* Weka
* R Studio

# **Data mining procedures**

## Data Preprocessing

### Handling missing values

Step 1:

ID attribute was removed from the dataset.

Step 2:

Since there were 405 missing values in EBV\_D attribute, we just removed those records from the dataset

Step 3:

Removing the NA/NULL values using mean for numeric attributes and mode for categorical attributes. The table below shows the logic of filling the missing values of each attribute the has missing values.

Since medical field is extremely relying on and depend on patient’s health test results and his/her health condition, we should take into consideration all critical values that might affect treatment/medical decision. For this reason, we created the table below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| № | Attribute name | Attribute description | Type | Number of missing values | Replaced by the following logic | Explanation |
| 1 | BMI\_R | Body Mass Index of the recipient at transplantation (kg/m²) | numeric | 24 (1%) | mean (  belonging to the TRANSP\_AGE\_R and SEX\_R) | Body Mass Index of the recipient depends on recipient’s age and gender. Missing values in BMI\_R attribute will be replaced with the attribute mean considering the values of recipient’s age (x <= 30, 30< x<= 60, x> 60) and gender (male/ female) [2] |
| 2 | TIME\_REG\_TRANSP | Time between registration on waiting list and transplantation (years) | numeric | 319 (9%) | mean | Missing values of the TIME\_REG\_TRANSP attribute will be replaced with the attribute mean. |
| 3 | CIT\_HOUR | Cold Ischemia Time (hours) | numeric | 14 (0%) | mean | Missing values of the CIT\_HOUR attribute will be replaced with the attribute mean. |
| 4 | INIT\_DISEASE | Initial disease: Recurrent disease = 1, Other = 0 | binary | 2 (0%) | mode | Missing values of the INIT\_DISEASE attribute will be replaced with the attribute mode. |
| 5 | TECH\_DIAL | Technique of dialysis: Pre-emptive = 1, Peritoneal dialysis = 2, Hemodialysis = 3 | nominal | 1 (0%) | mode | Missing values of the TECH\_DIAL attribute will be replaced with the attribute mode. Since, modality selection process  based on patient motivation, desire, physician and/or nurse bias, patient’s financial situation and overall health condition. [3] |
| 6 | INCOMP\_ABDR | Number of HLA ABDR incompatibilities | numeric | 105 (3%) | mode  (belonging to the attributes AGE\_D and TRANSP\_RANK) | Missing values in attribute INCOMP\_ABDR will be filled considering the values of the donor's age (x <= 30, 30< x<= 60, x> 60) and (TRANSP\_RANK)  Rank of the kidney transplantation (high:1-2, medium: 3-4, low: 5-6)  [4] |
| 7 | AGE\_D | Donor age (years) | numeric | 16 (0%) | mean  (belonging to the SEX\_D attribute) | Missing values of the AGE\_D attribute will be replaced with the attribute mean with considering a donor’s gender SEX\_D. |
| 8 | CREAT\_D | Serum creatinine of the donor (µmol/L) | numeric | 111 (3%) | mean  (belonging to the SEX\_D and AGE\_D attributes) | Missing values of the CREAT\_D attribute will be replaced with the attribute mean with considering a donor’s gender SEX\_D  and age AGE\_D (x <= 30, 30< x<= 60, x> 60)  [5] |
| 9 | RELAT\_DR | Relation between donor and recipient: Living donor = 0, Deceased donor = 1 | binary | 7 (0%) | mode (belonging to the DEATH\_FUNCTGRAFT and TIME\_TRANSP\_DEATH attributes) | Missing values of the RELAT\_DR attribute will be replaced with the attribute mean with Missing values in RELAT\_DR will be filled considering values of attributes DEATH\_FUNCTGRAFT and TIME\_TRANSP\_DEATH, since kidney transplantation with deceased or living donor affects graft survival. |
| 10 | CAUSE\_DEATH\_D | Cause of the donor death: Vascular death =1, Other = 0 | binary | 59 (2%) | mode | Missing values in CAUSE\_DEATH\_D will be filled considering values of attributes AGE\_D (x <= 30, 30< x<= 60, x> 60) and SEX\_D |
| 11 | CMV\_D | Hepatitis C Virus infection of the donor: Yes = 1, No = 0 | binary | 19 (%) | mode | Missing values in CMV\_D will be filled considering values of attributes AGE\_D (x <= 30, 30< x<= 60, x> 60) and SEX\_D |
| 12 | EBV\_D | Epstein-Barr Virus infection of the donor: Yes=1, No=0 | binary | 405 (12%) | mode  (belonging to the SEX\_D) | This attribute is linked with an attribute gender of the donor SEX\_D. Missing values of EBV\_D attribute will be replaced with the attribute mode considering the values of the attribute gender of the donor to fill in the missing values of EBV\_D. Since one of the risk factors associated with acquiring EBV include being female, and also EBV virus can be spread by organ transplants. [6] |

**Table 2. Logic for replacing missing values for the attributes**

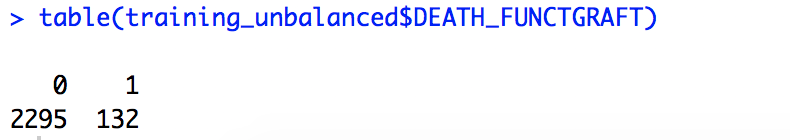
### Splitting the dataset into training and test sets

We split the dataset into two: 80% as of training dataset and 20% as of a test dataset using WEKA.

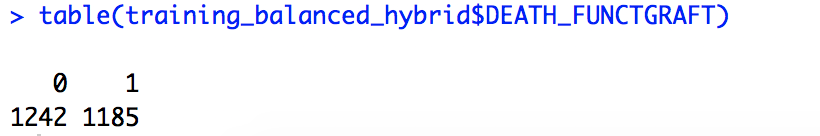
### Balancing unbalanced data

After cleaning the dataset from missing values and splitting the dataset into two sets (training and test), we balanced the training set/ Using hybrid method we balanced an unbalanced training dataset. This method creates possibly balanced samples by random over-sampling minority examples, under-sampling majority examples or combination of over- and under-sampling.

Before balancing:



After balancing:

****

## Results of data Preprocessing

The proportion is shown based on main class attribute: DEATH\_FUNCTGRAFT

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **0** | **1** | **Total** |
| **Initial cleaned data**  ../../../Screen%20Shot%202019-11-04%20at%206.57.54%20PM.png | 2869 | 165 | 3034 |
| **Test set**  ../../../Screen%20Shot%202019-11-04%20at%207.01.18%20PM.png | 574 | 33 | 607 |
| **Training set unbalanced (before balancing)**  ../../../Screen%20Shot%202019-11-04%20at%207.04.50%20PM.png | 2295 | 132 | 2427 |
| **Training set balanced (after balancing)** | 1242 | 1185 | 2427 |

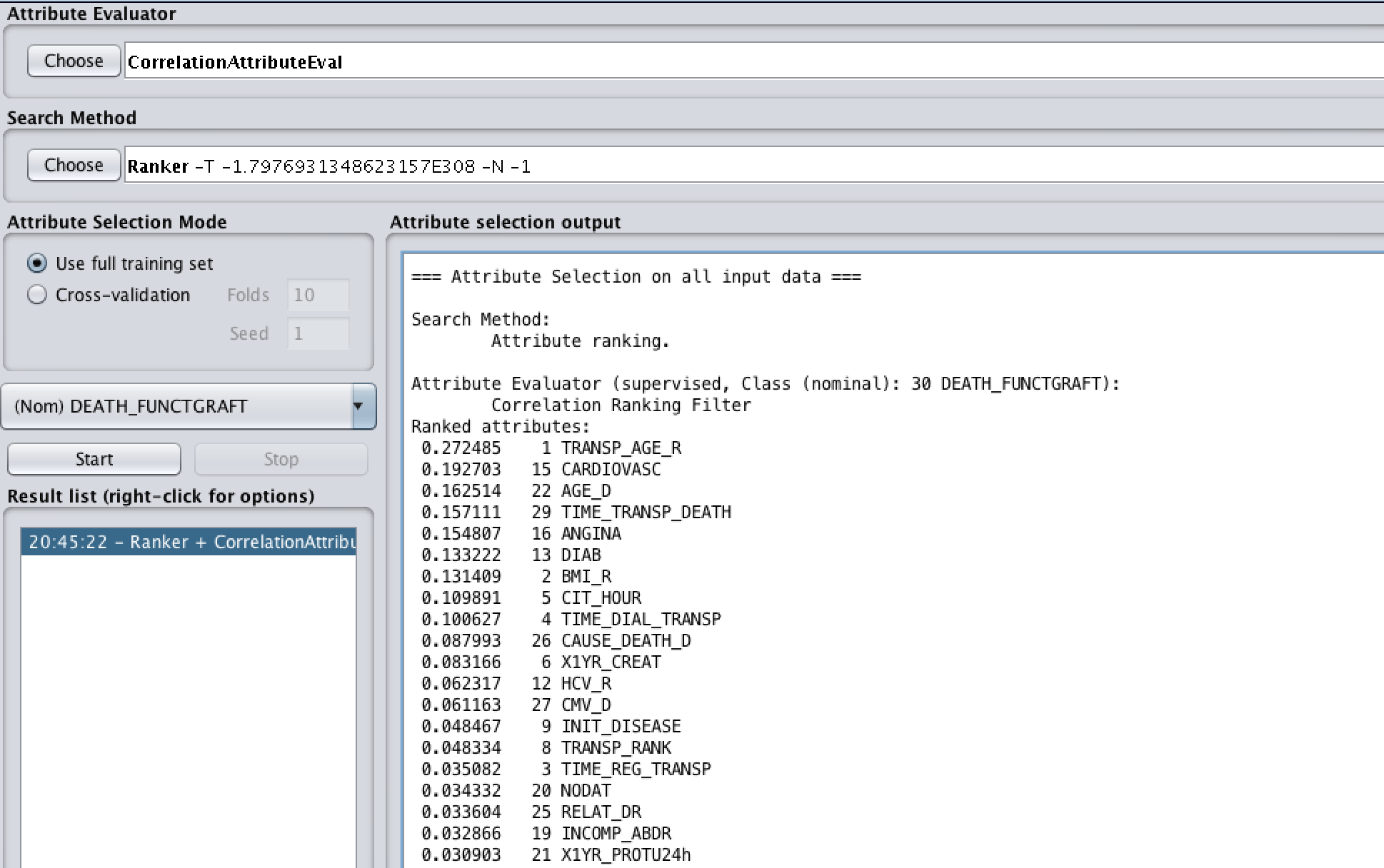
**Table 3. Results of Data Preprocessing based on class attribute values**

# **Selected Attributes and Classifiers**

**CorrelationAttributeEval**

Selected attributes:

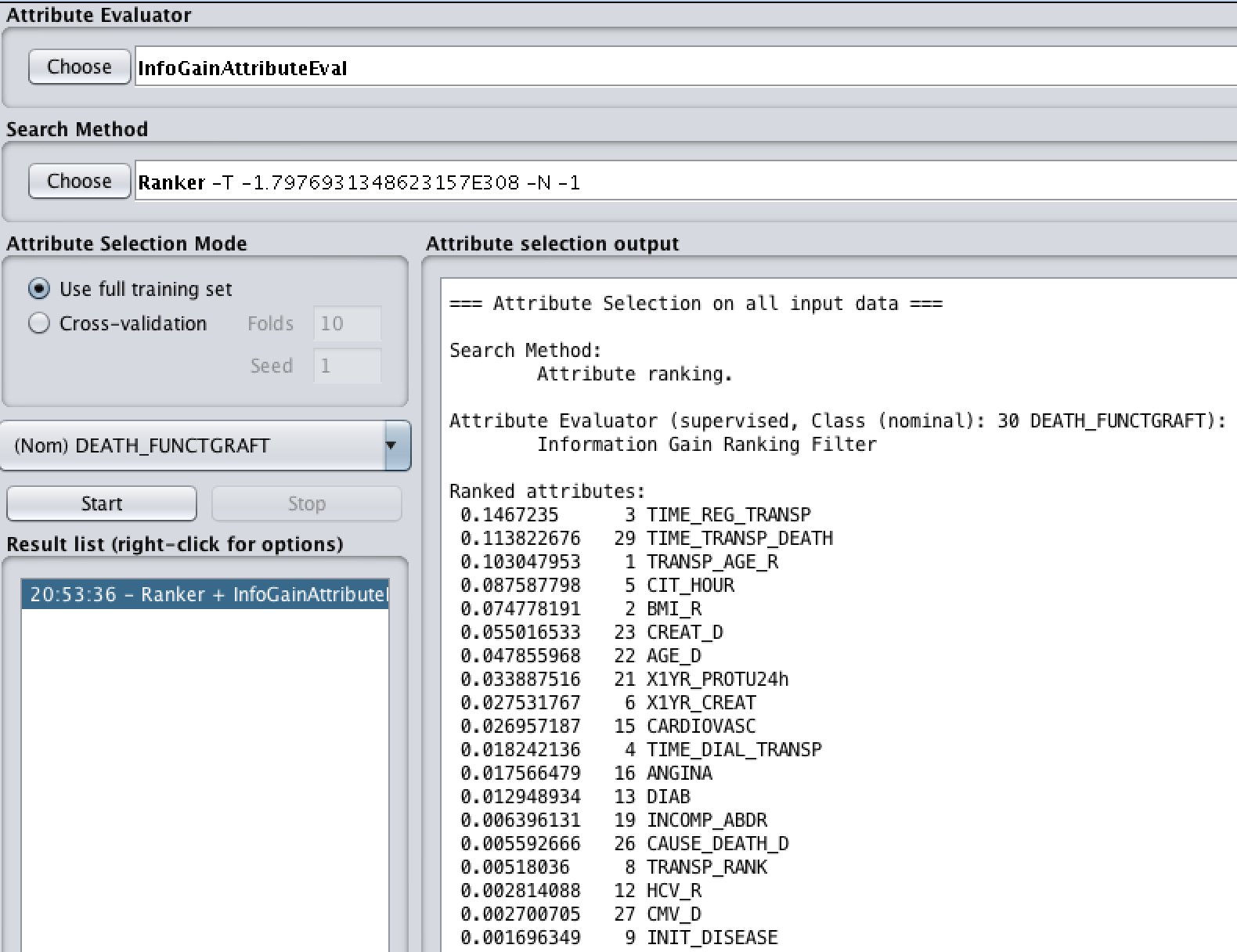
1 2 3 4 5 6 8 9 12 13 15 16 19 20 21 22 25 26 27 29



**InfoGainAttributeEval**

Selected attributes:

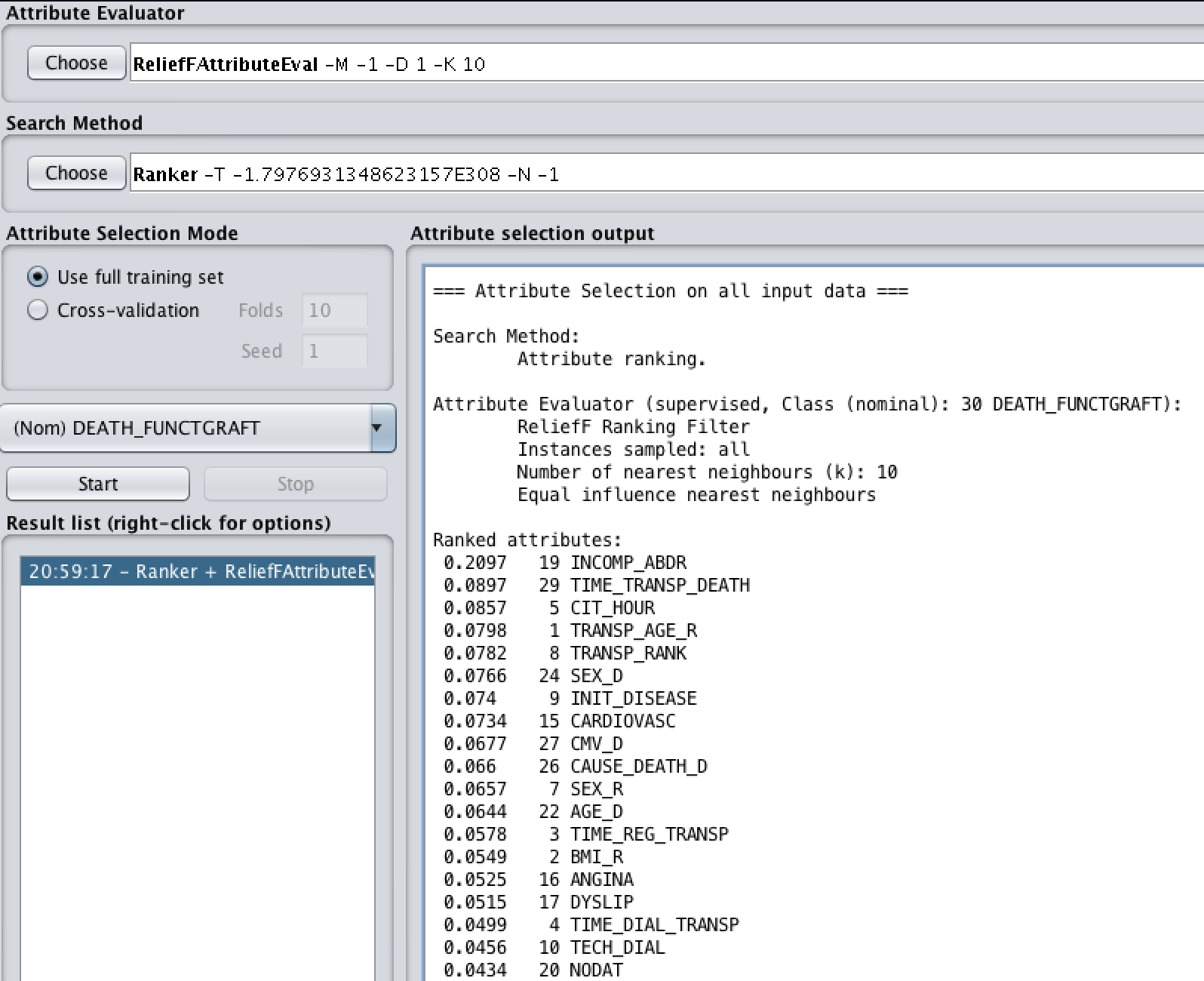
1 2 3 4 5 6 8 9 12 13 15 16 19 20 21 22 23 26 27 29



**ReliefFAttributeEval**

Selected attributes:

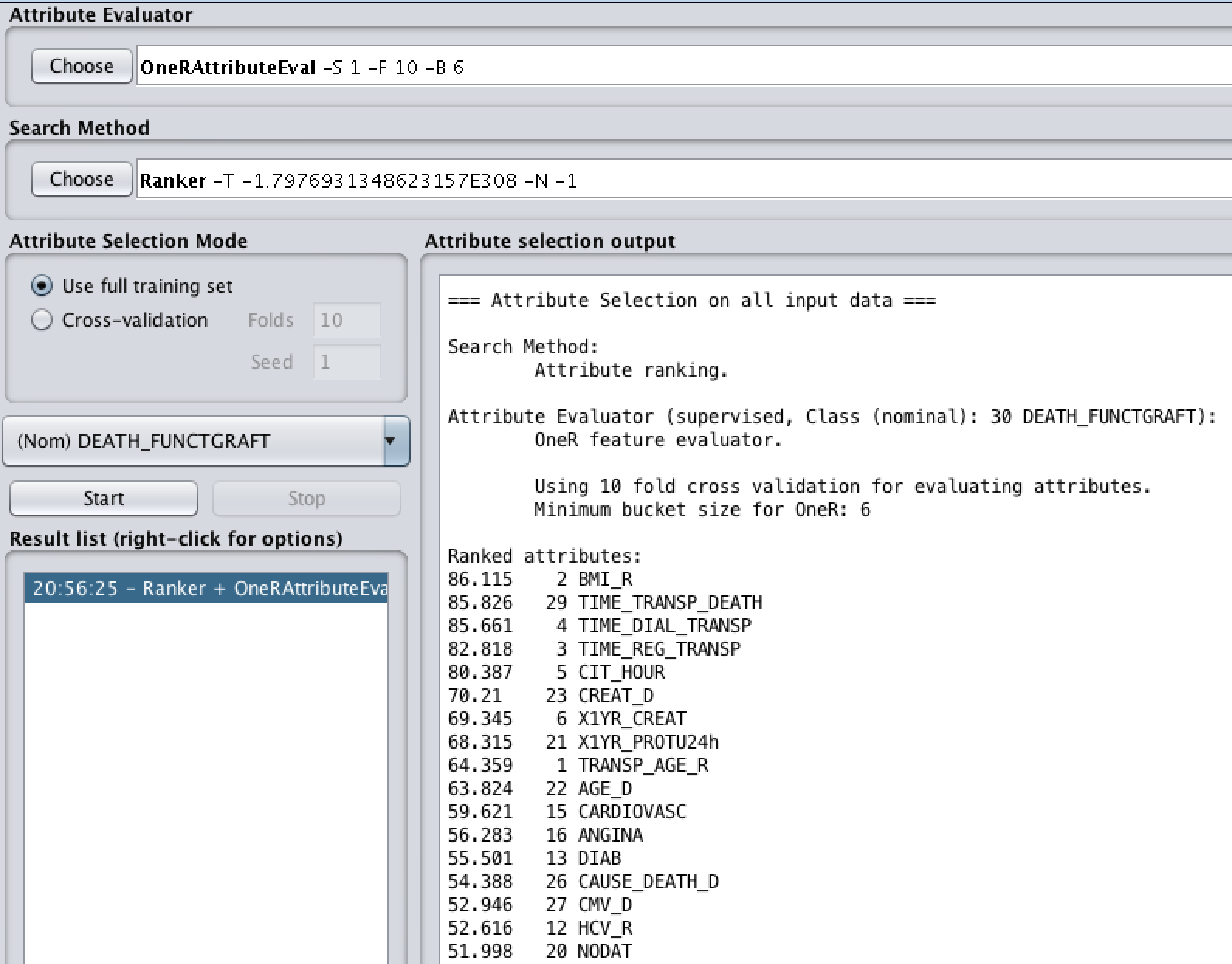
1 2 3 4 5 7 8 9 10 11 15 16 17 19 20 22 24 26 27 29



**OneRAttributeEval**

Selected attributes:

1 2 3 4 5 6 9 12 13 15 16 17 20 21 22 23 26 27 28 29



**ManualAttributeEval**

Selected attributes:

1 2 3 4 5 6 7 8 9 12 13 15 16 22 23 24 25 26 27 29

For the Manual Attribute Selection, we picked up the most chosen ones based on the four attribute selection algorithms above. Some attributes are chosen by three or even four of the previous selectors, we chose them as our self-chosen attributes.

Additionally, we made research on the following values that might affect the kidney functioning graft results.

## Attribute Selection Algorithm Results:

Correlation

F

One R

ReliefF

Manual Attribute Selection

Selected Classifiers:

Naïve Bayes

Logistic

Neural Networks (Multilayer Perceptron)

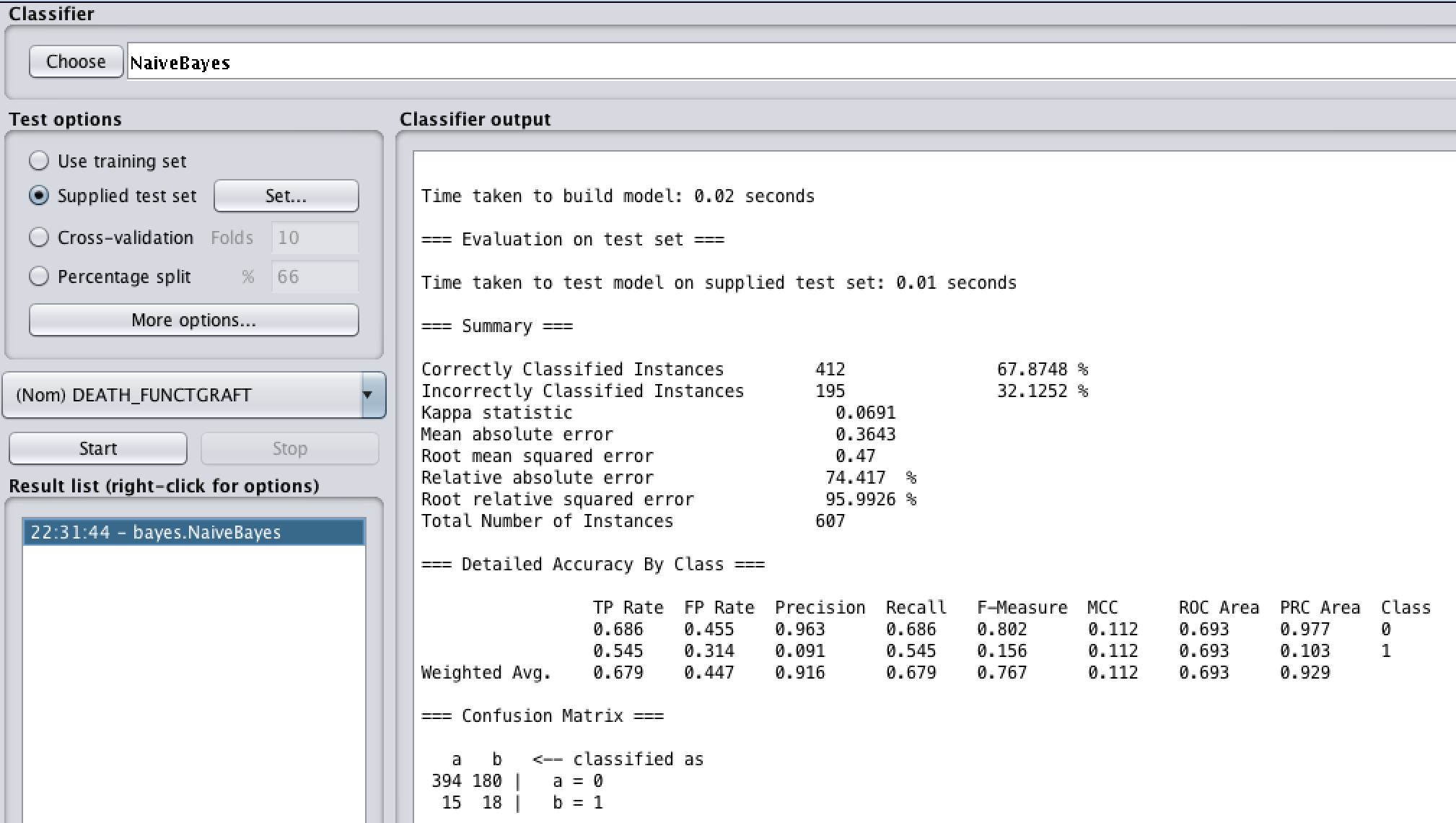
J48 + AdaBoostM1

## Data mining result and evaluation

Building the models:

### Naïve Bayes

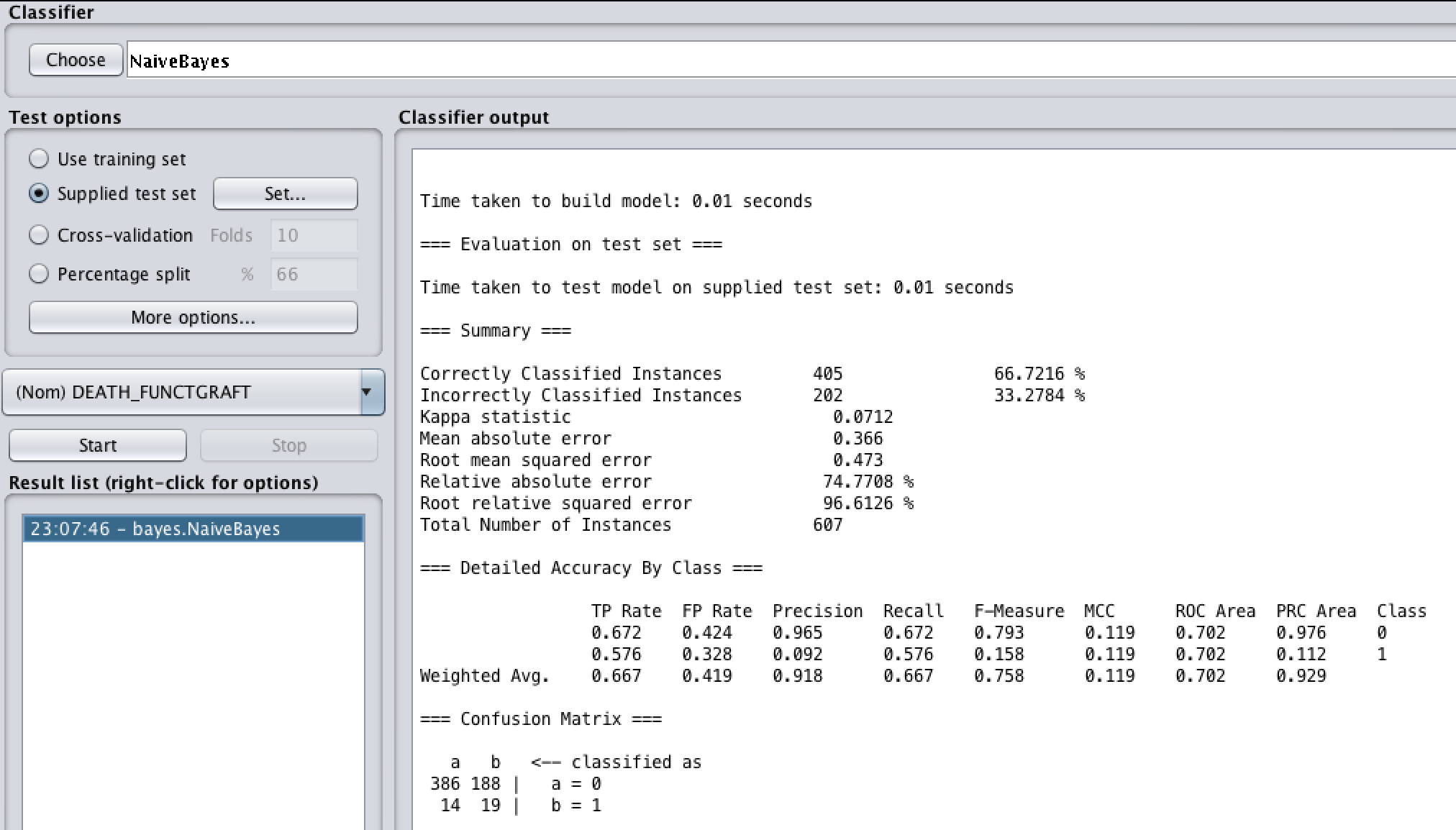
Correlation:



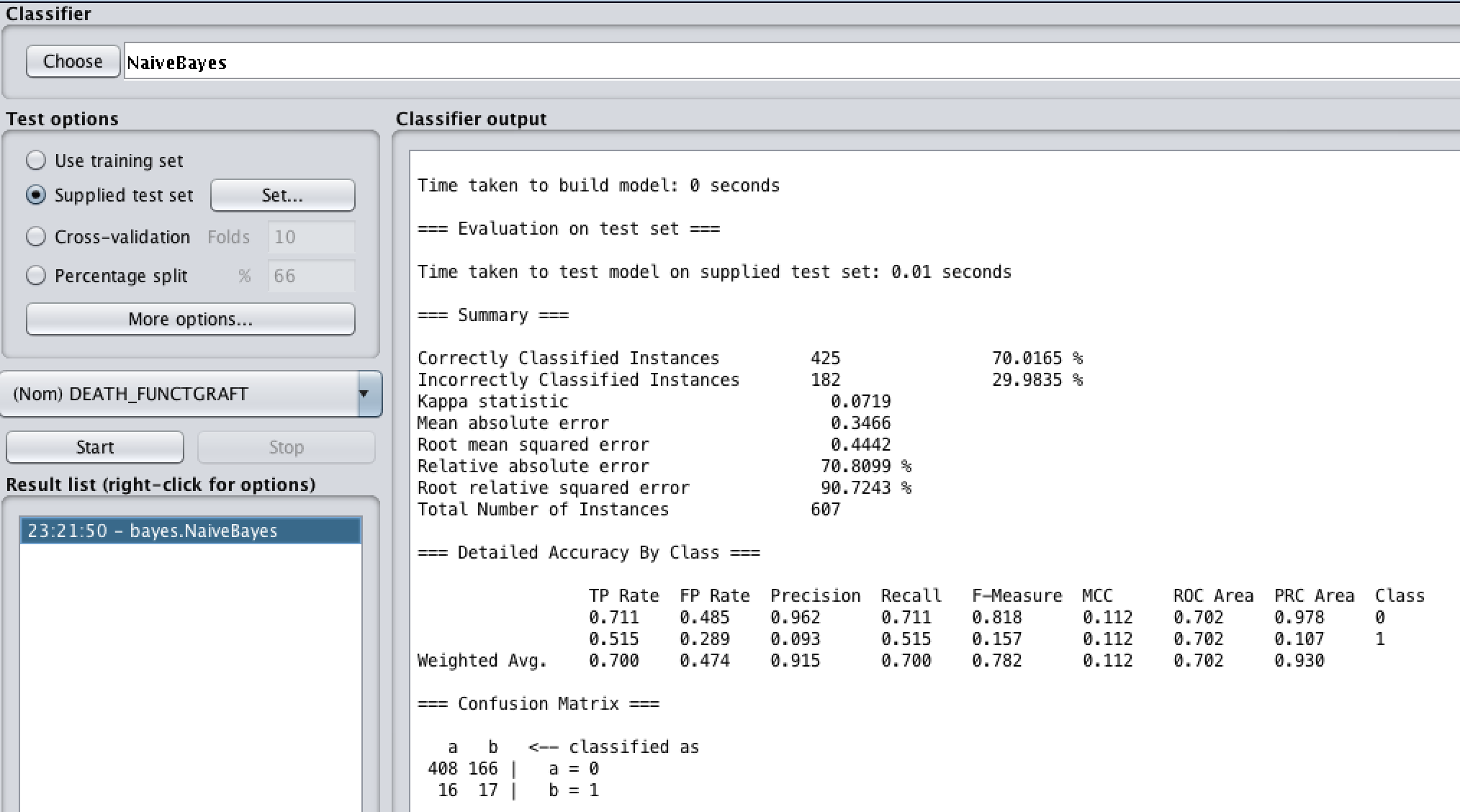
Information Gain:



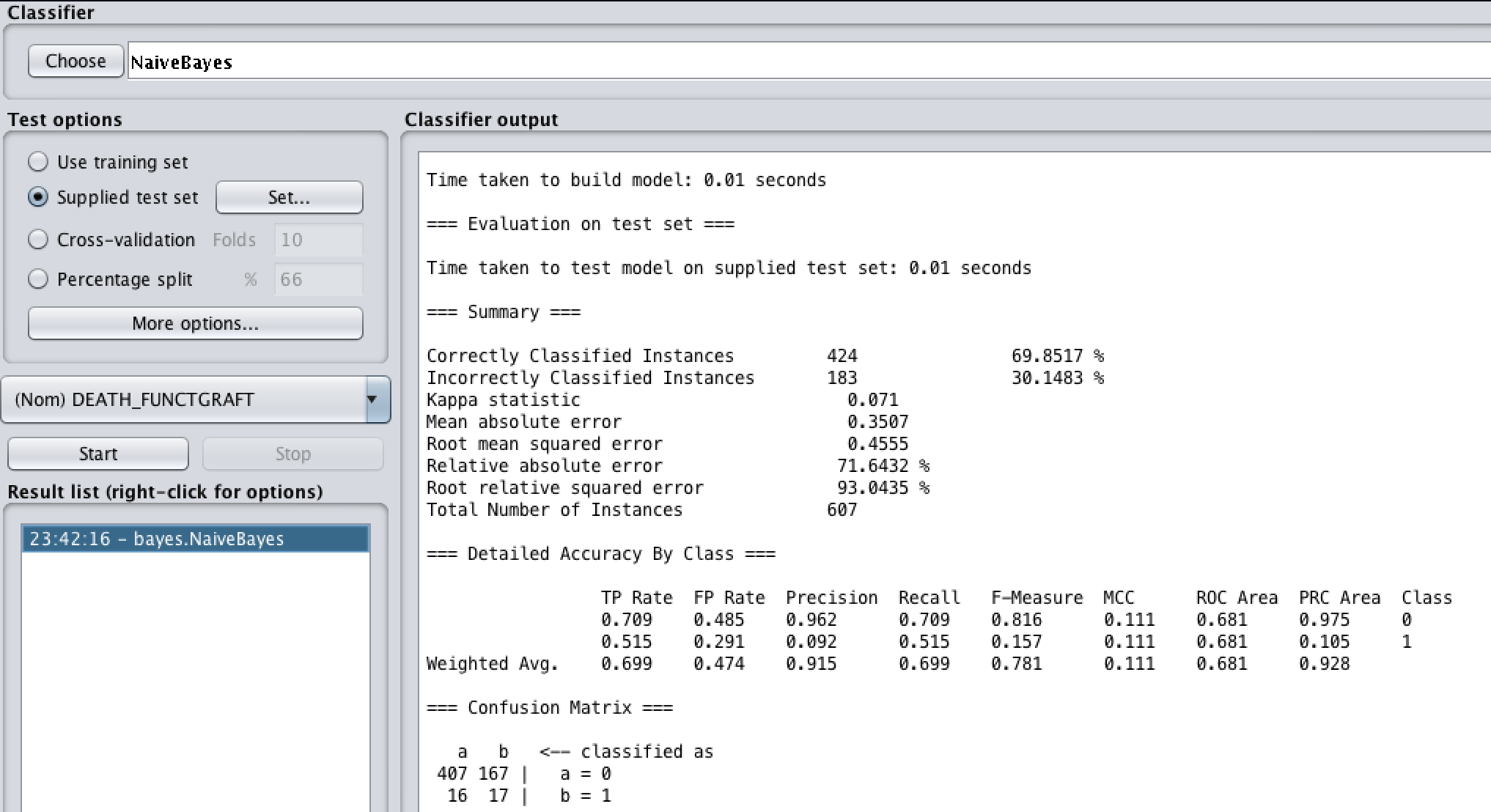
One R:



ReliefF:

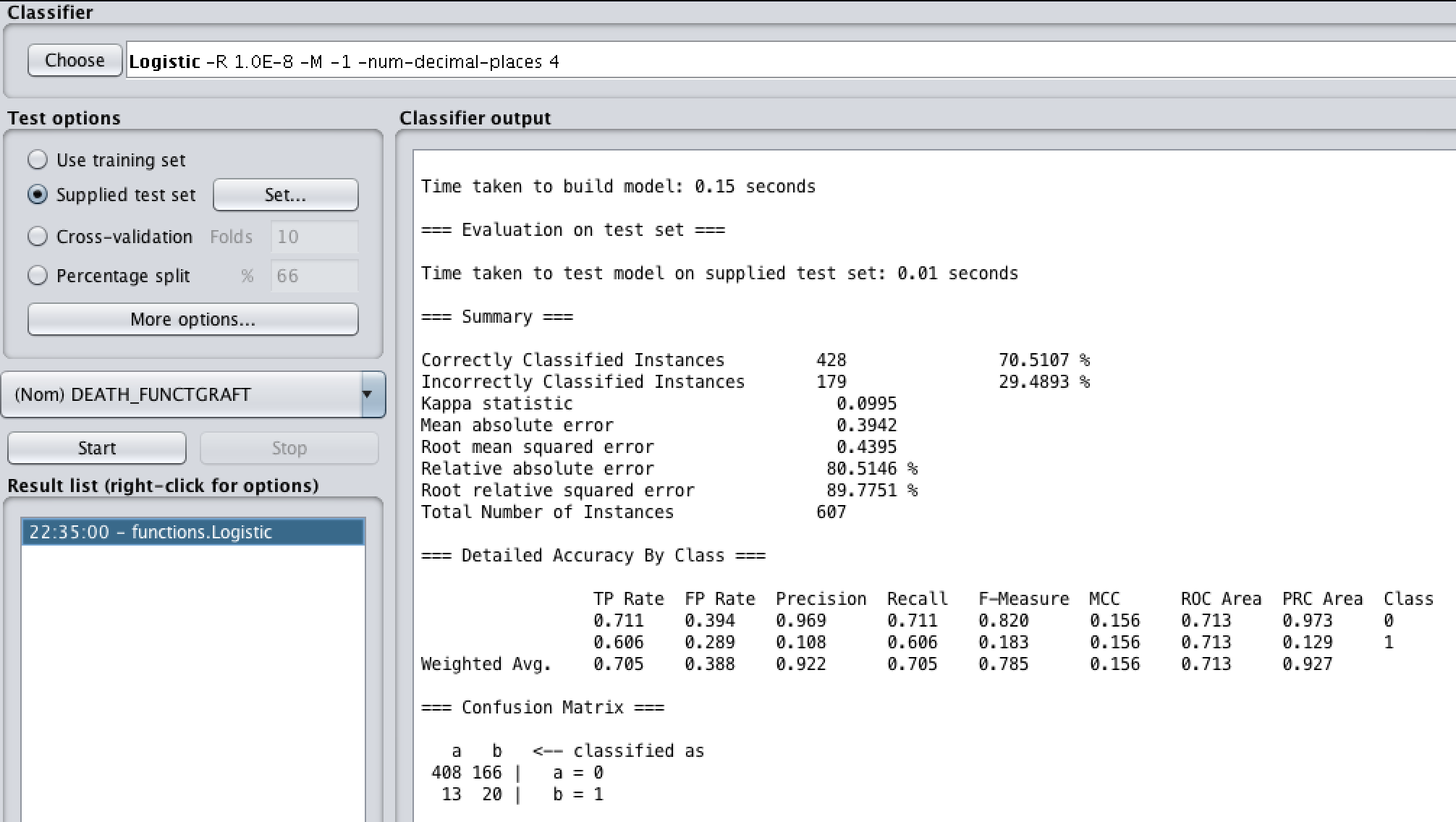


Manual Attribute Selection:



### Logistic

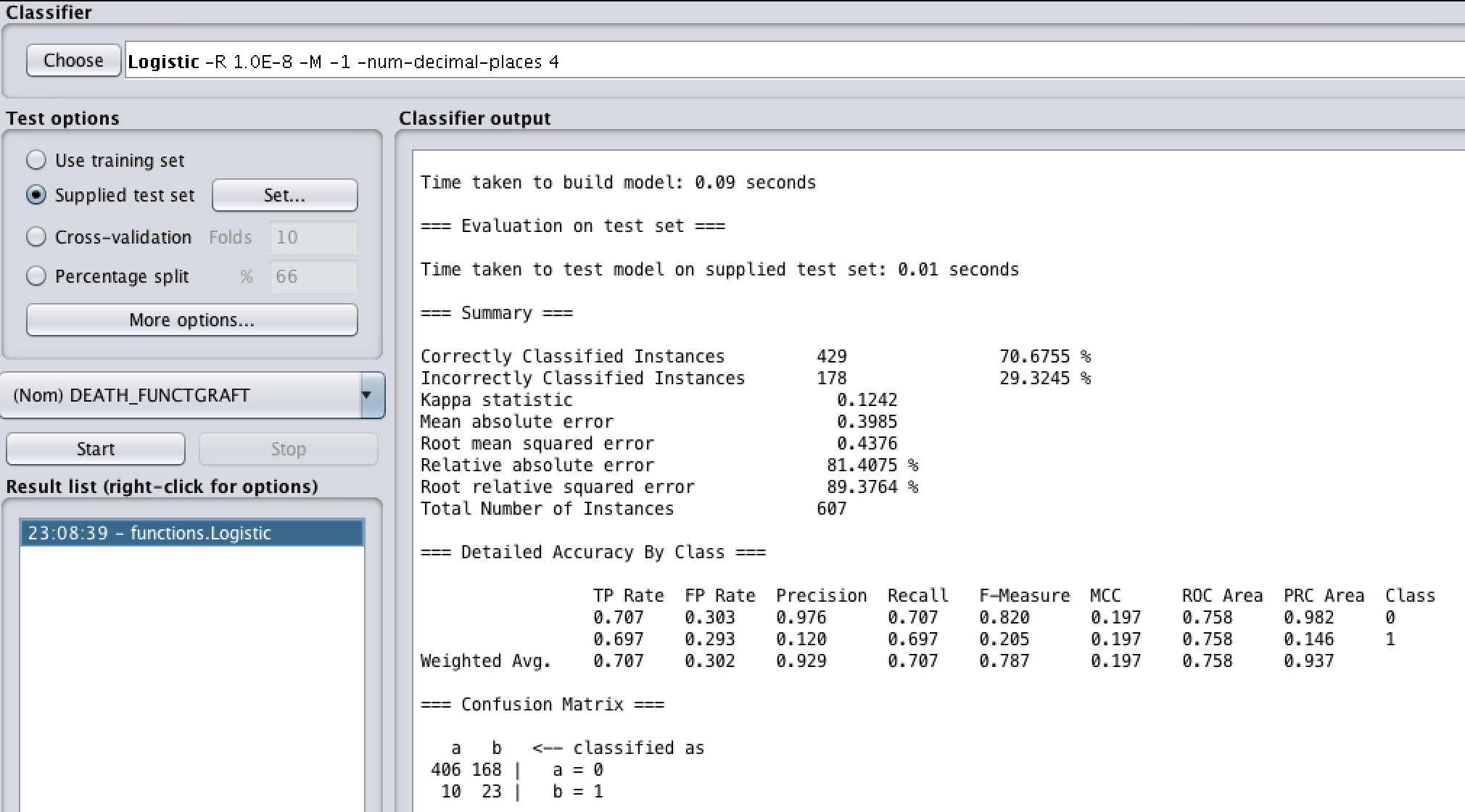
Correlation:



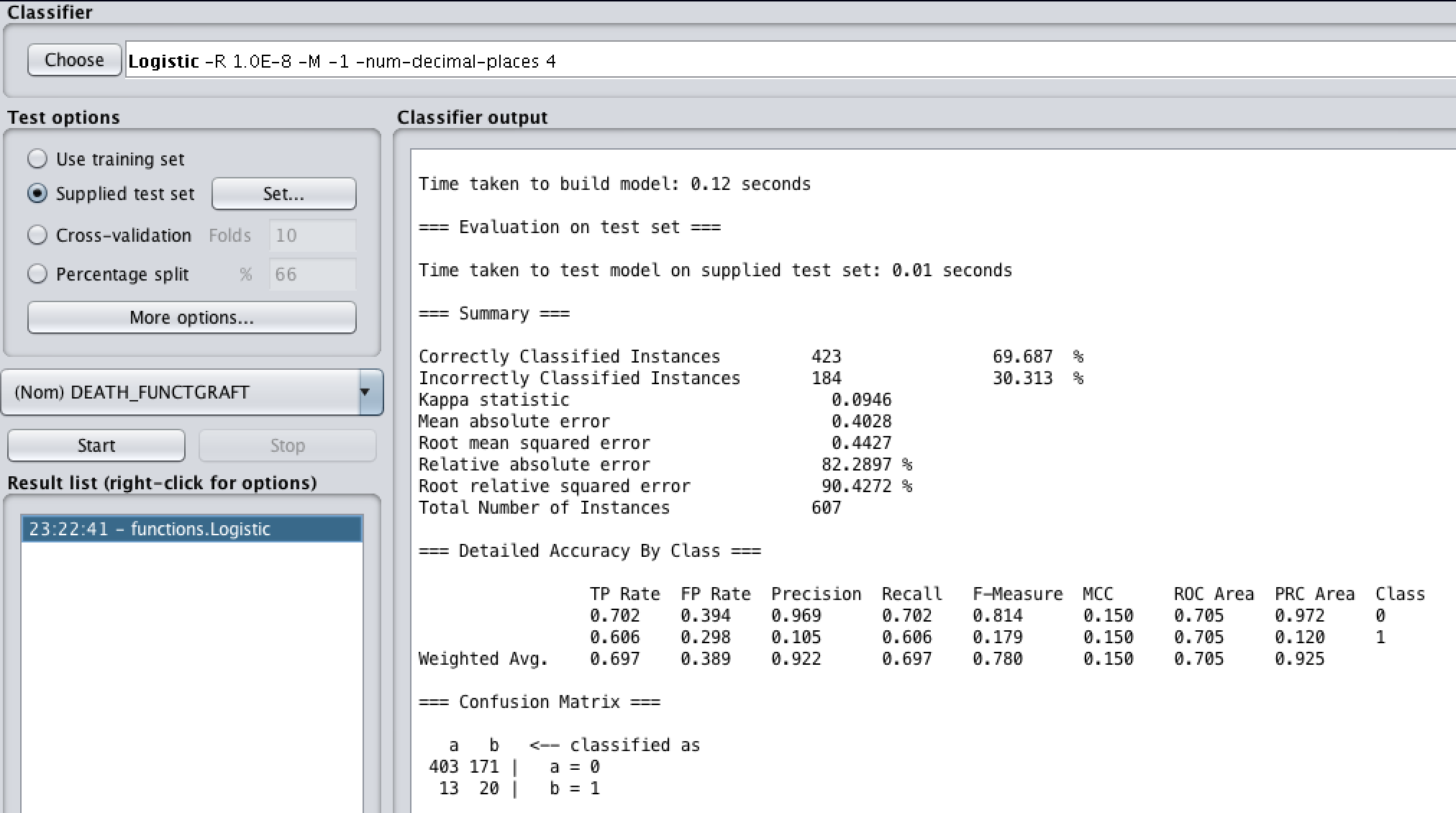
Information Gain:



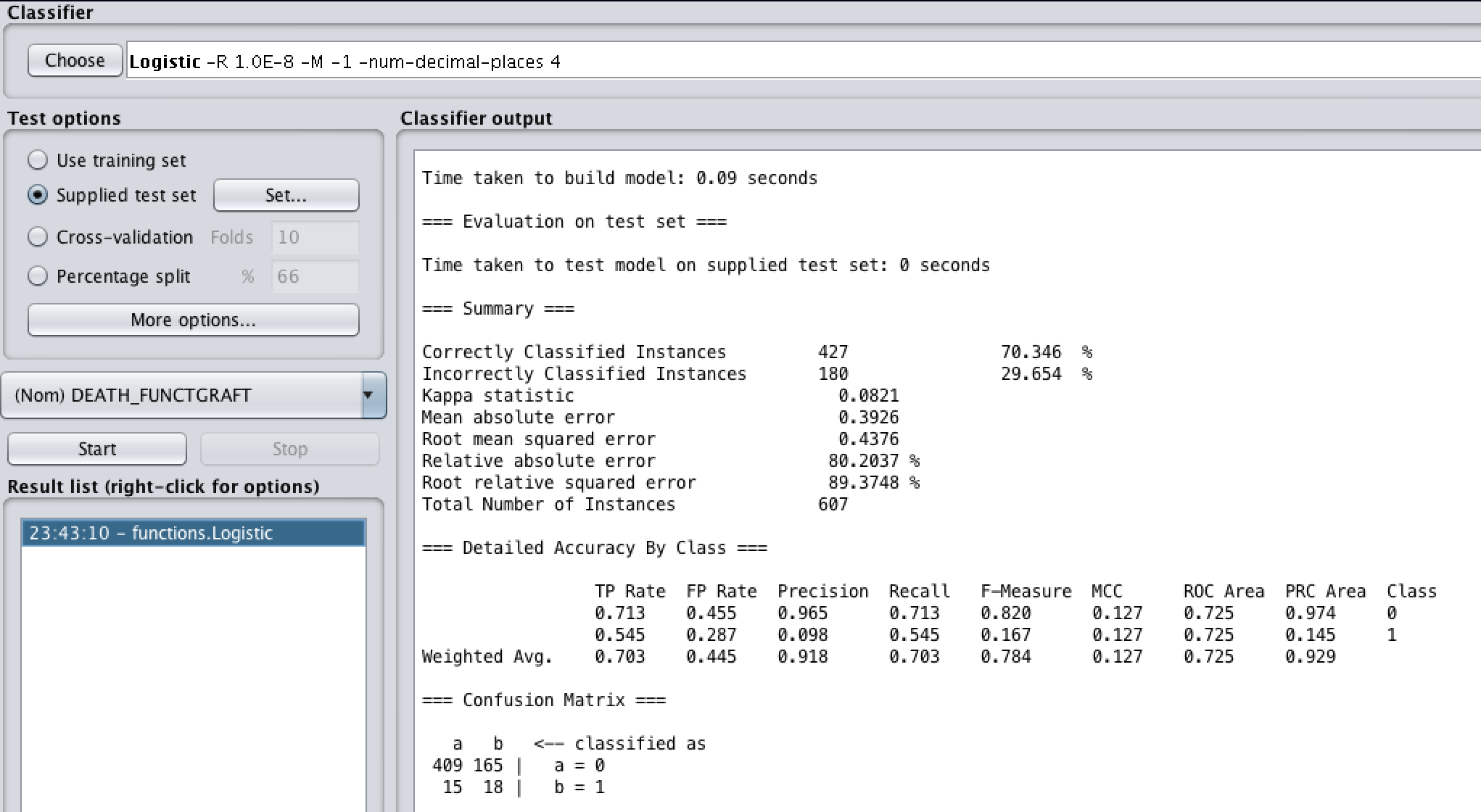
One R:



ReliefF:

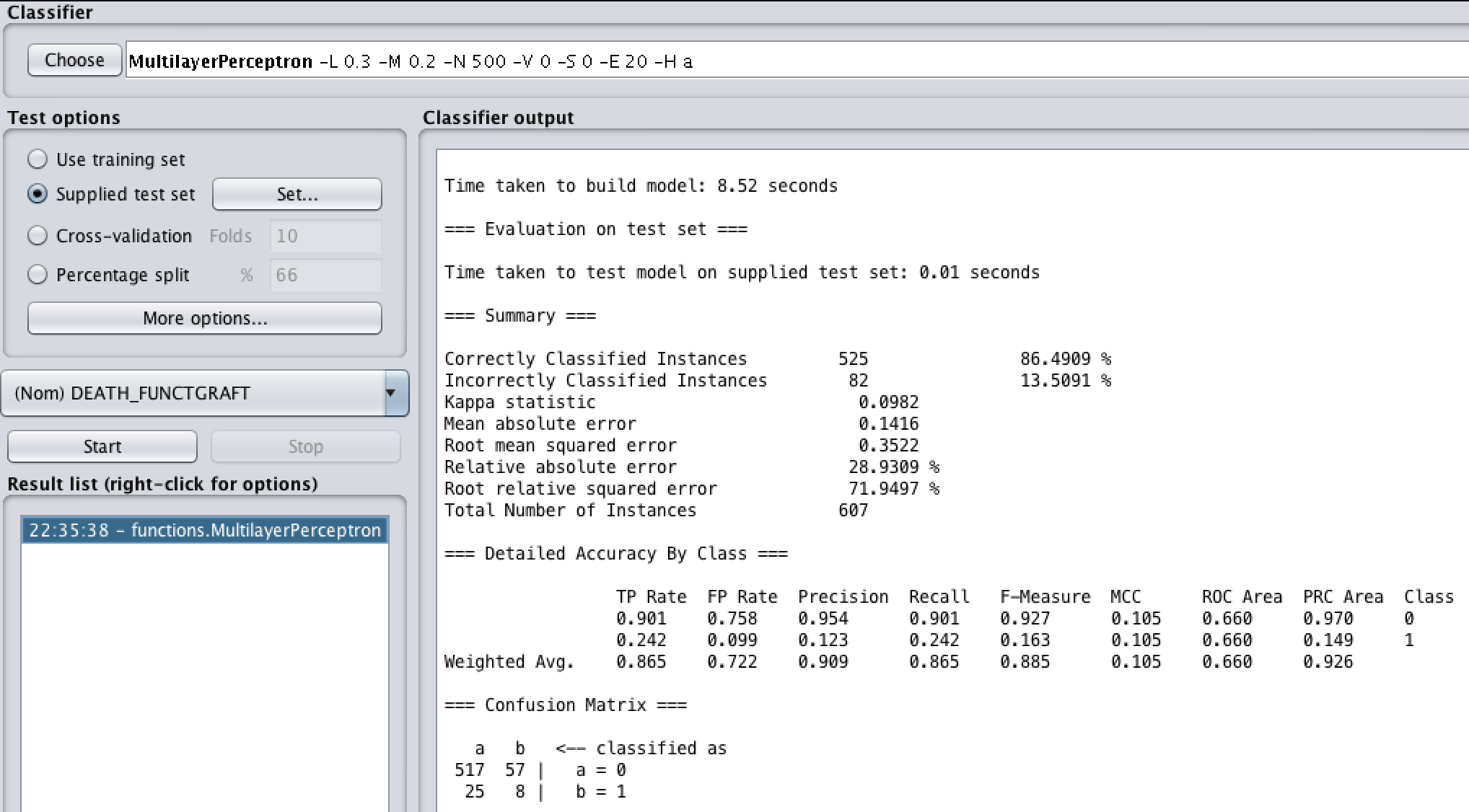


Manual Attribute Selection:

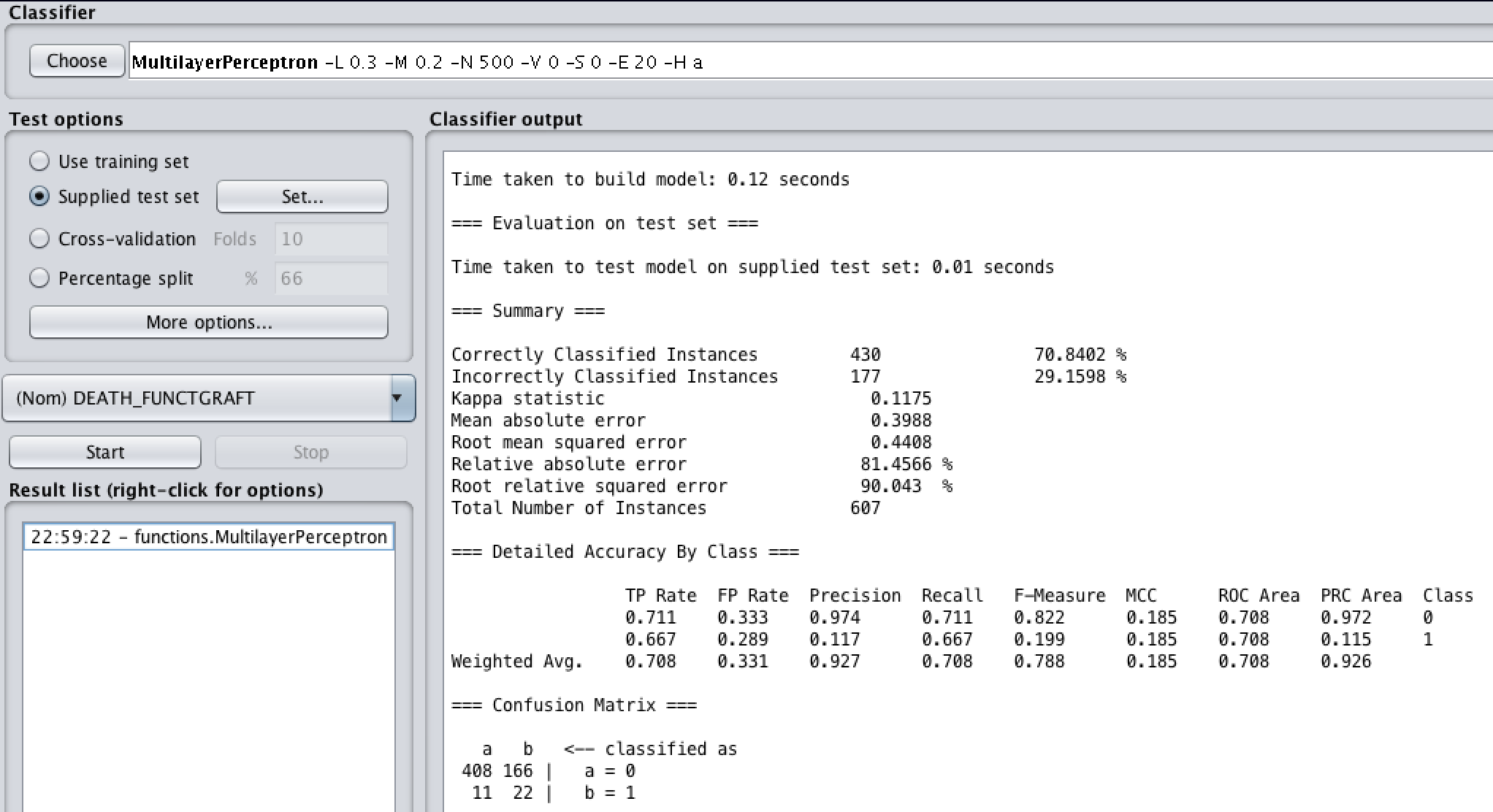


### Neural Networks (Multilayer Perceptron)

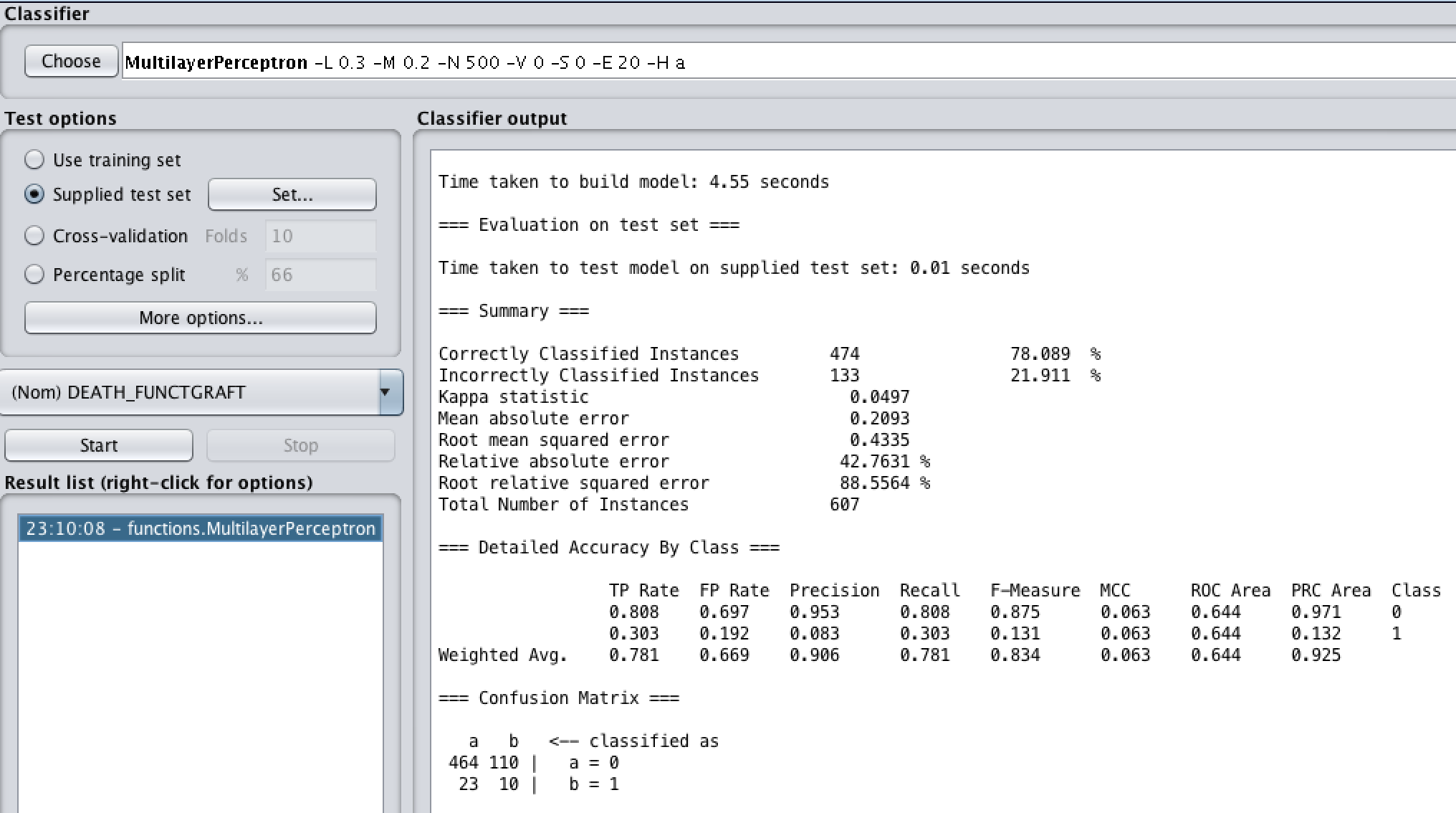
Correlation:



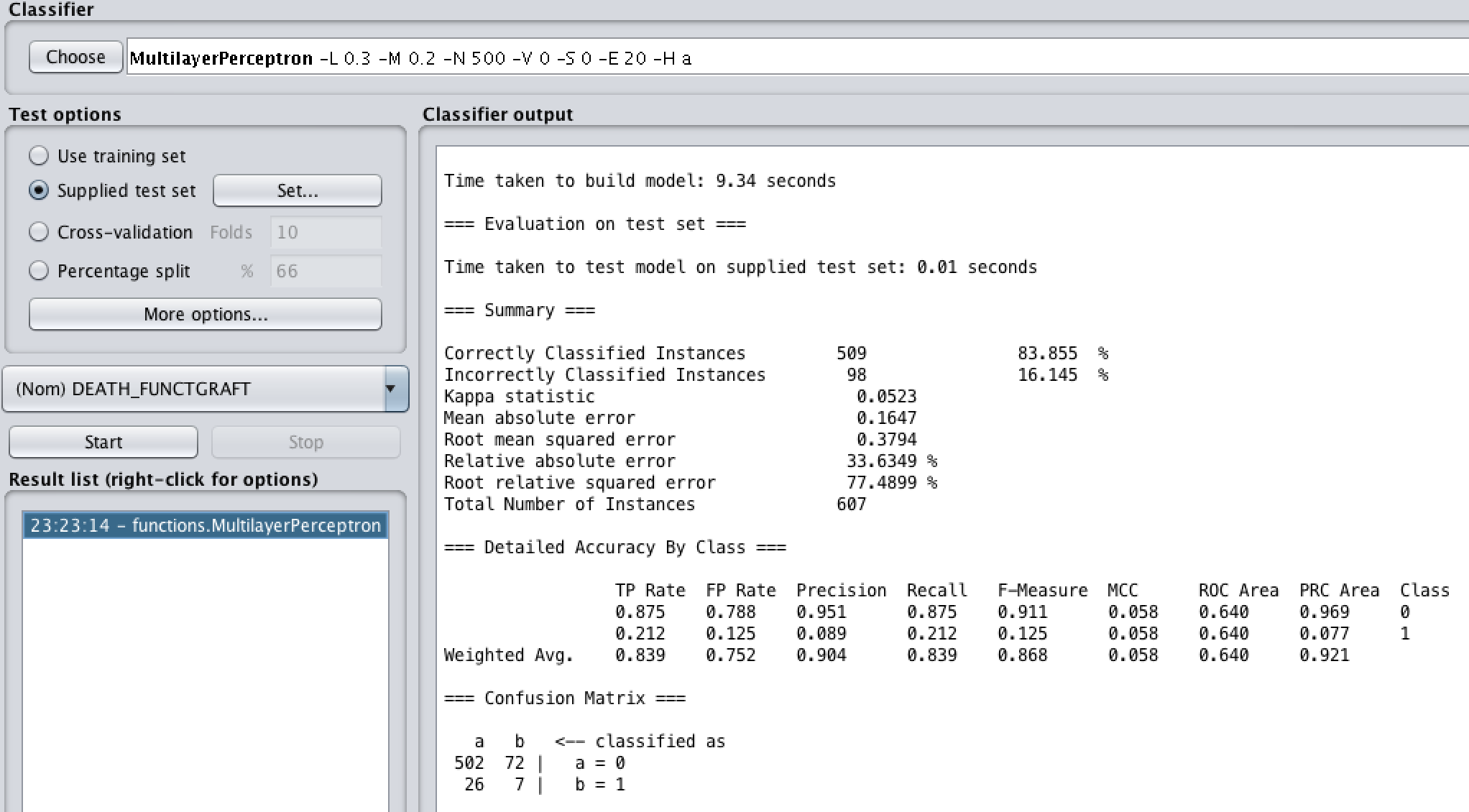
Information Gain:



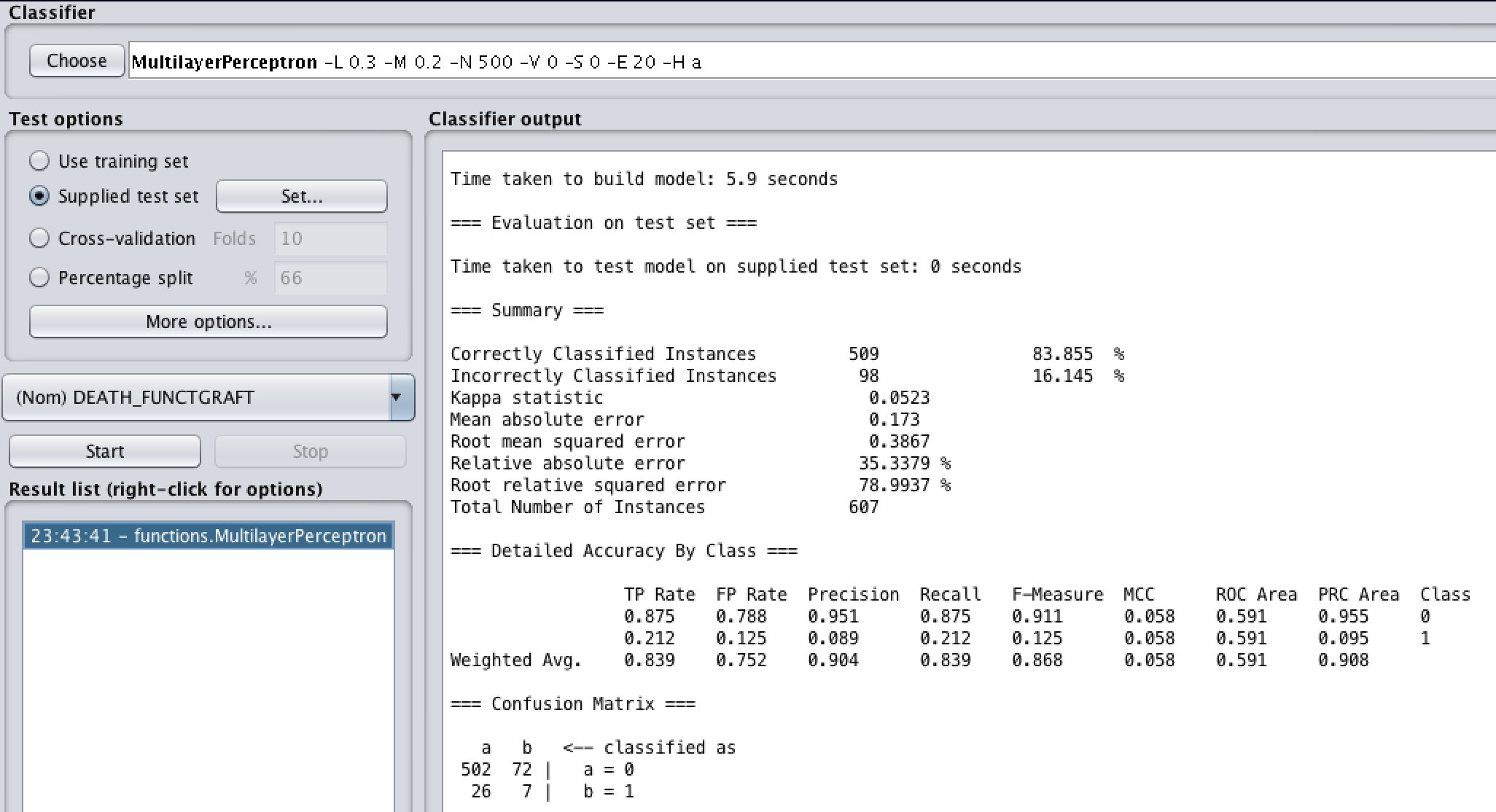
One R:



ReliefF:

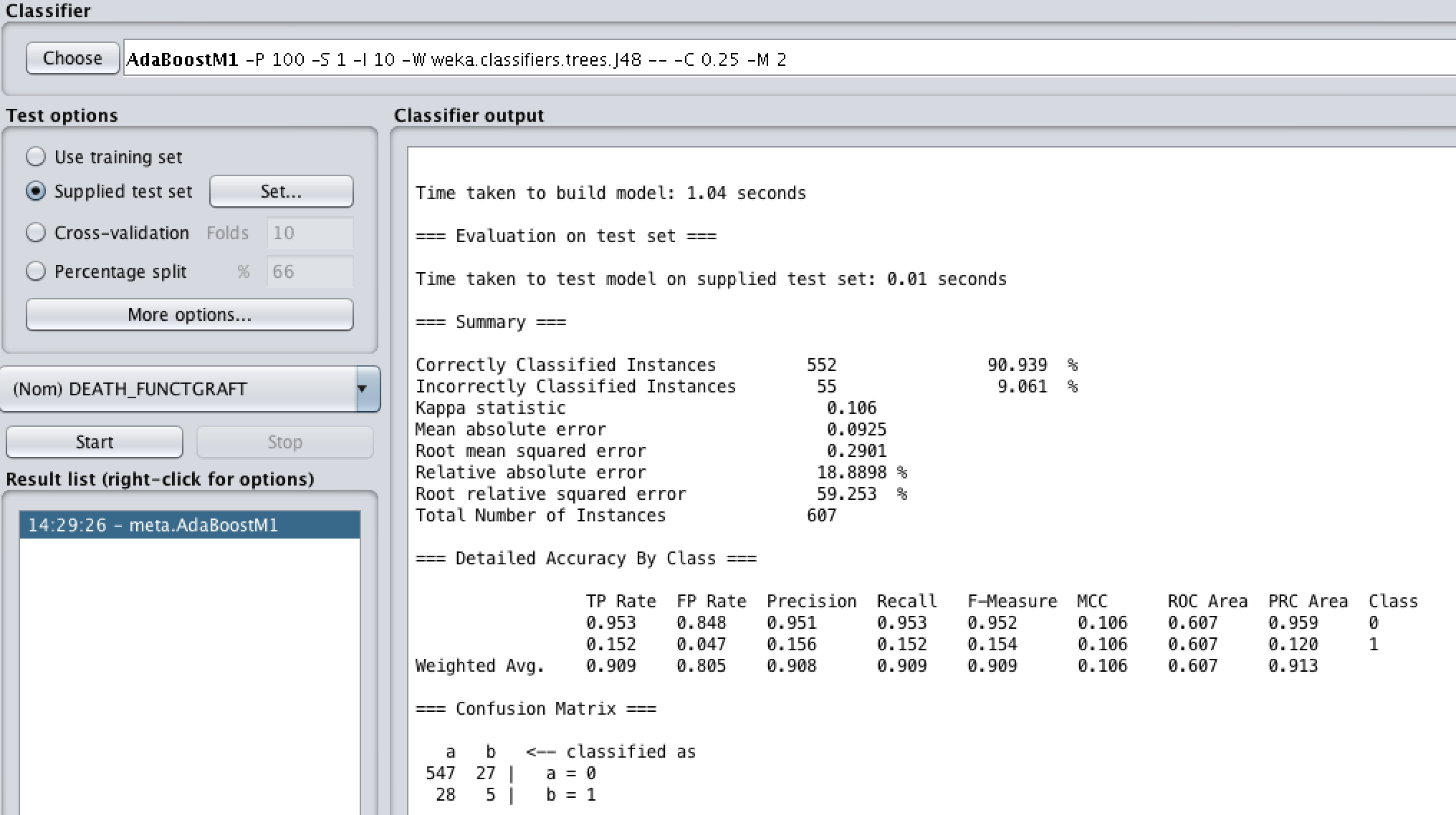


Manual Attribute Selection:

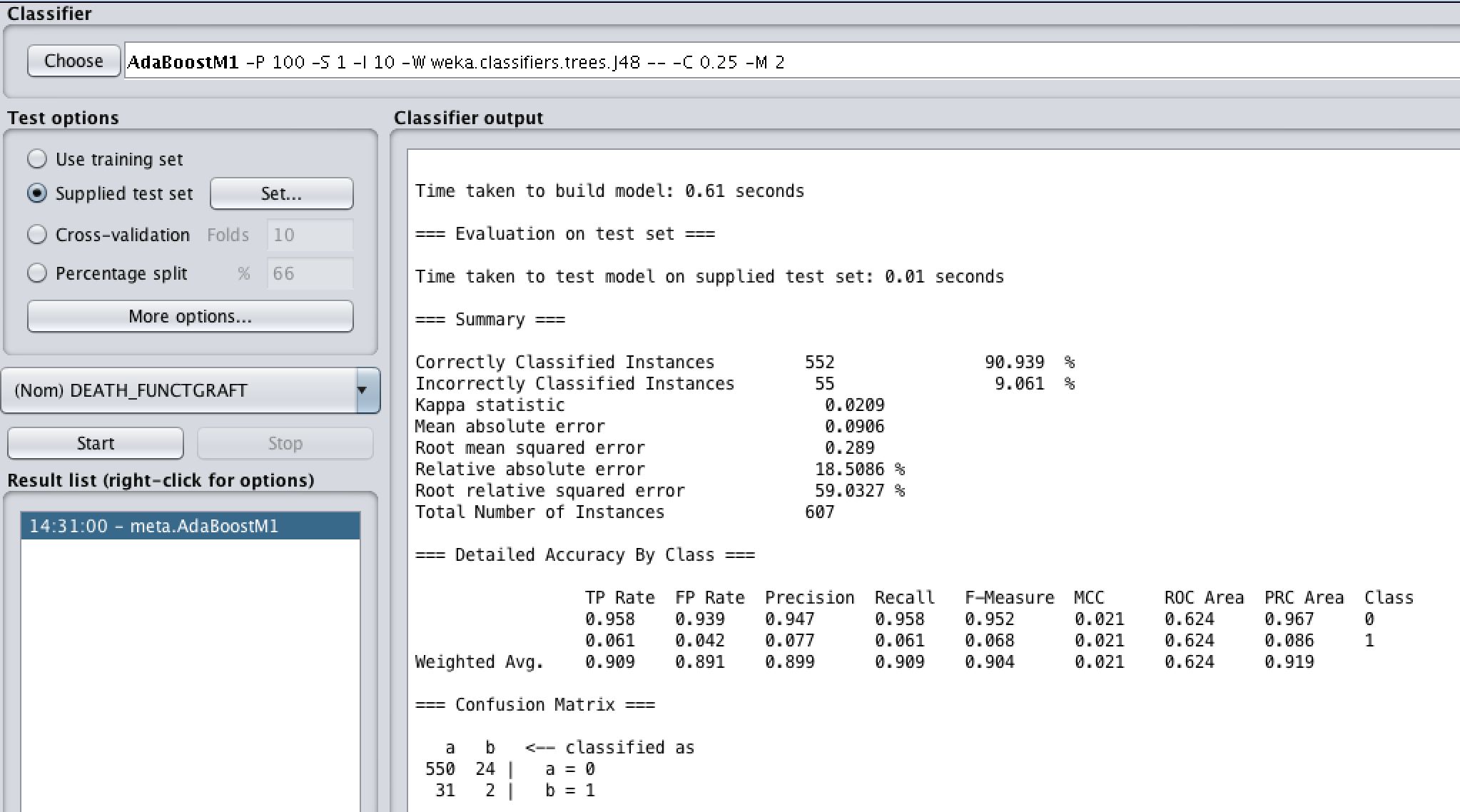


### J48 + AdaBoostM1

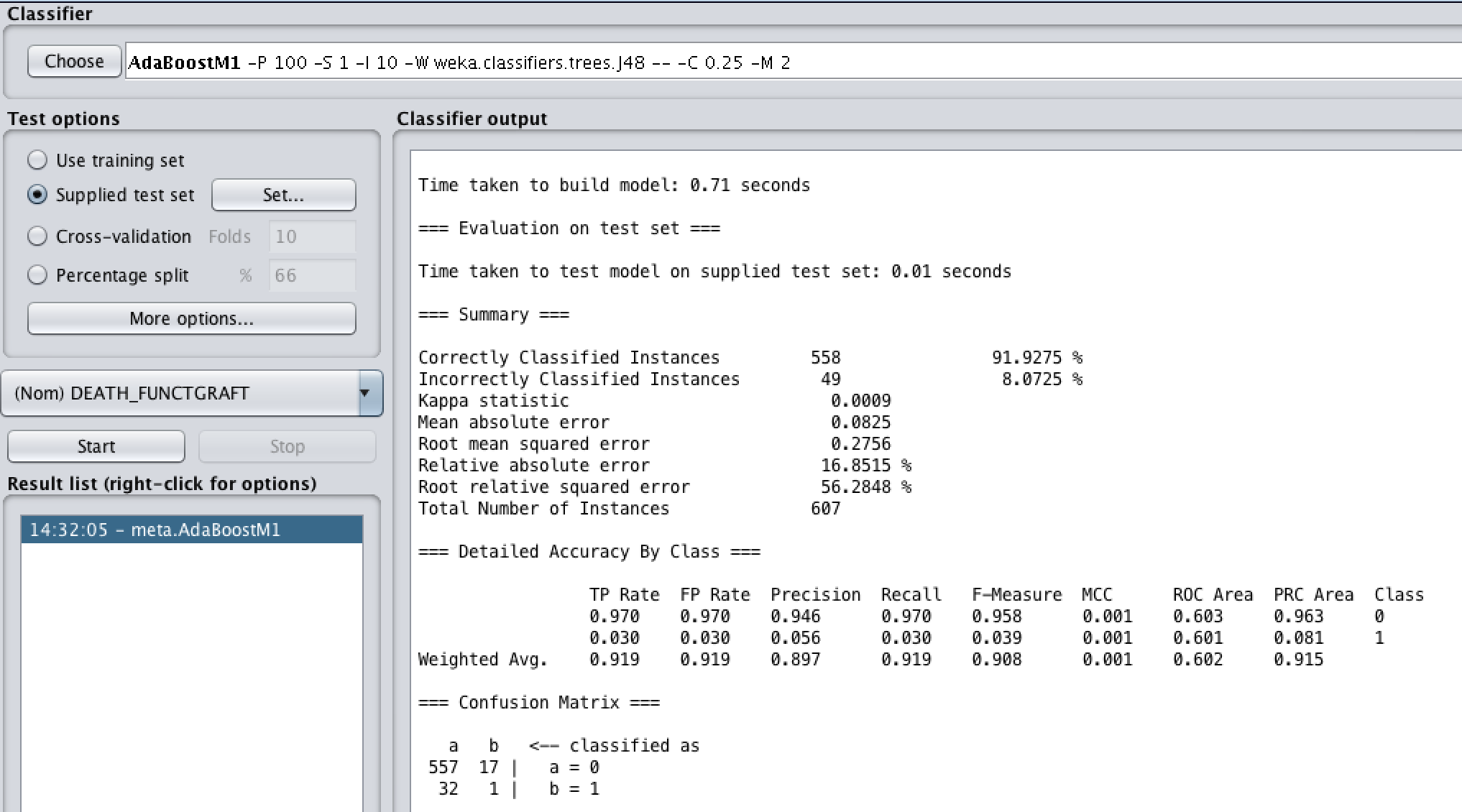
Correlation:



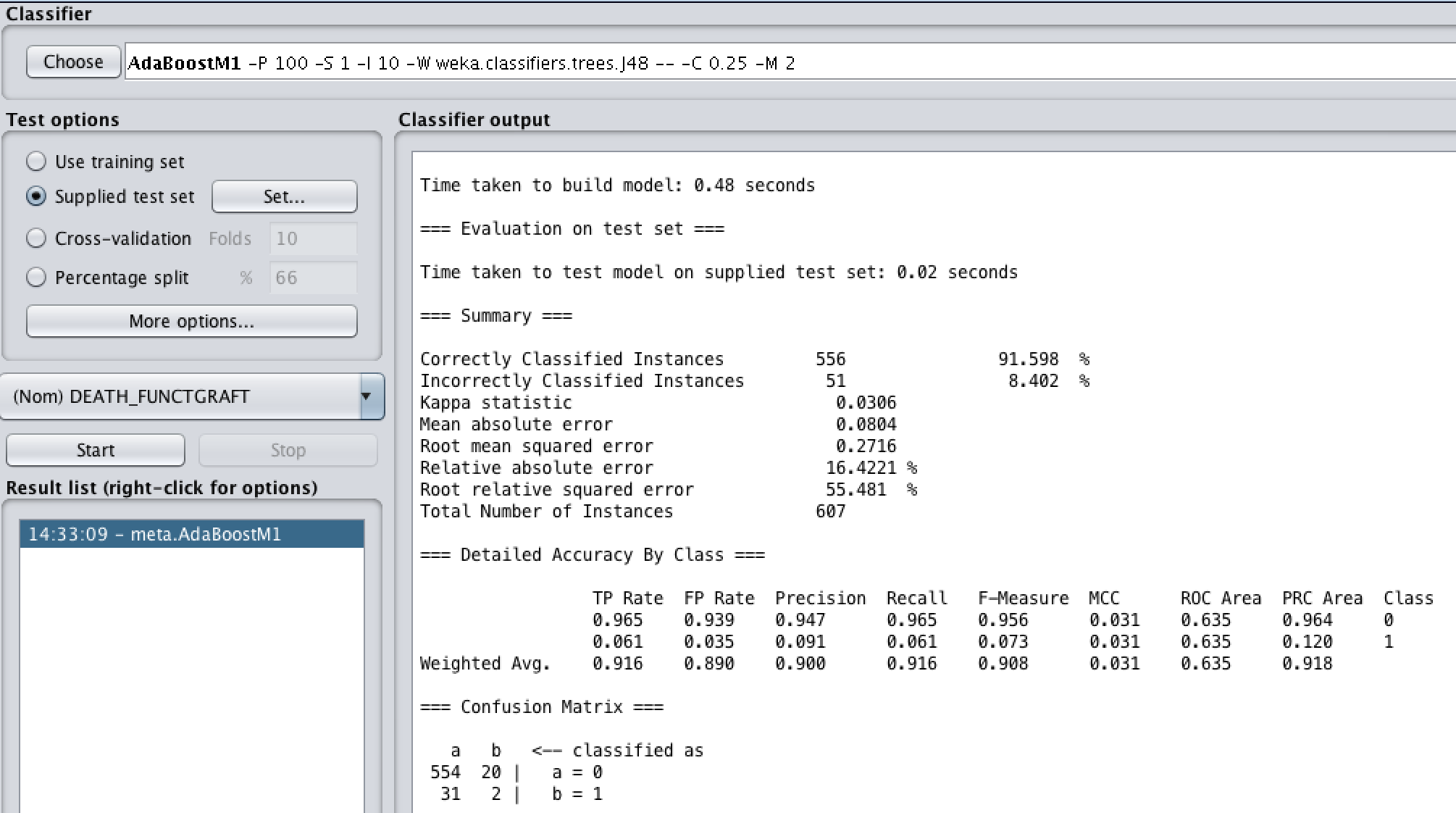
Information Gain:



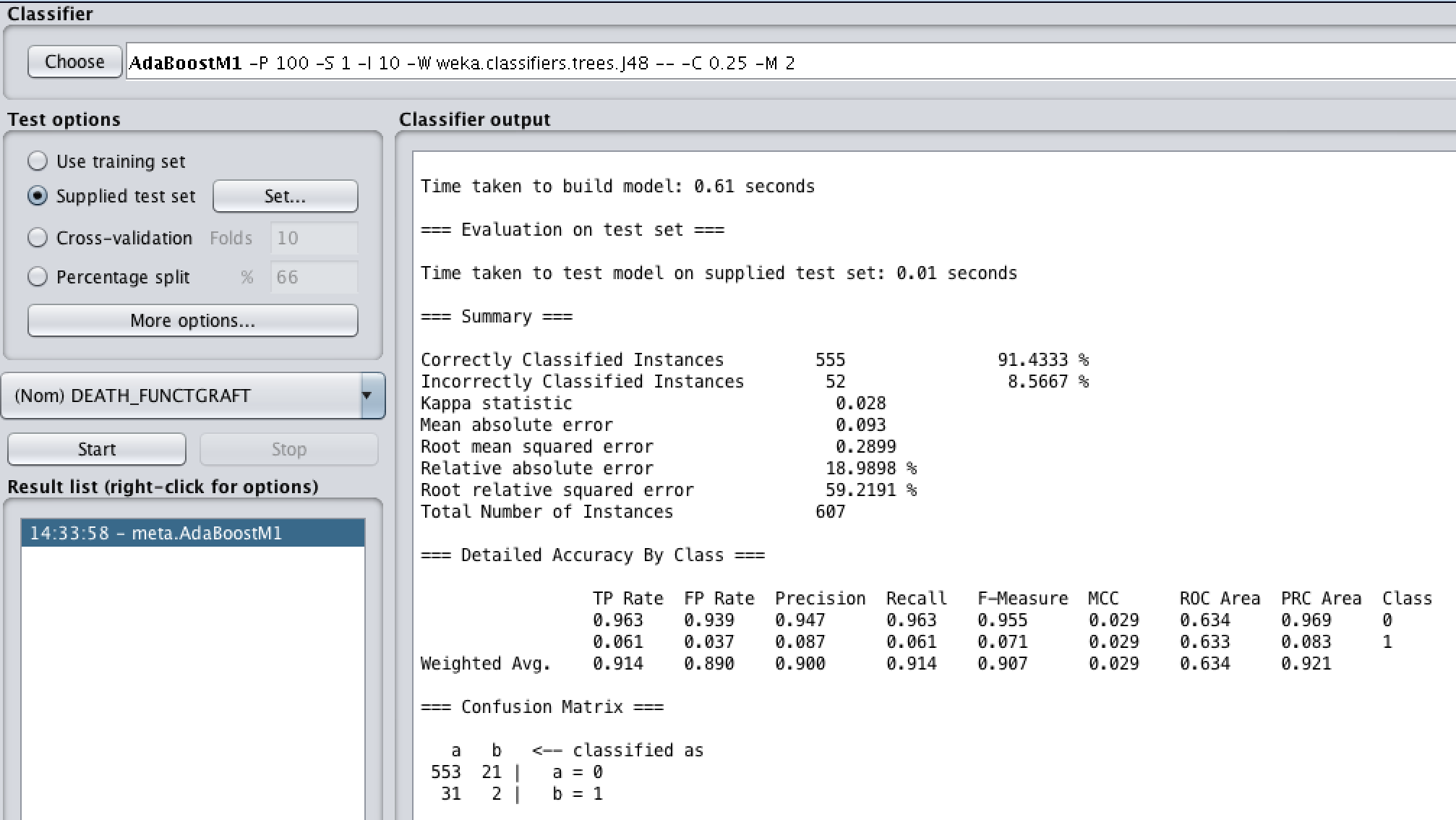
One R:



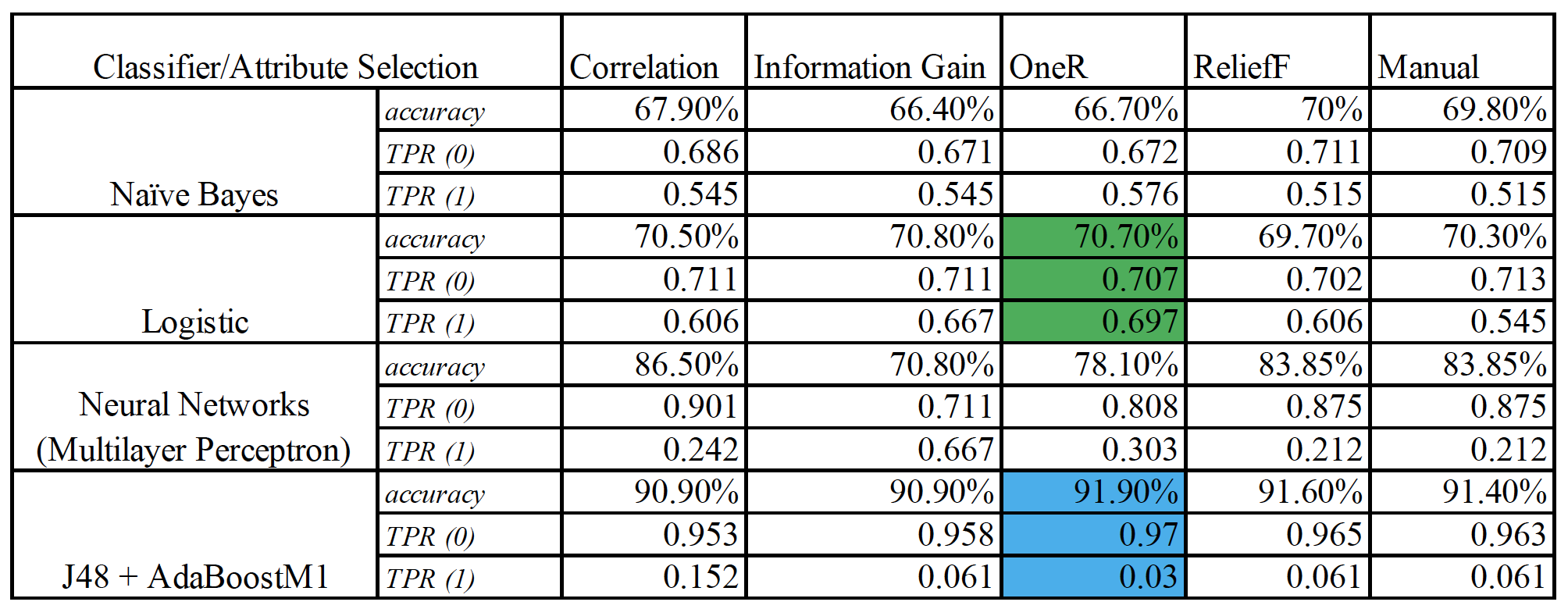
ReliefF:



Manual Attribute Selection:



## Justification for our selection of the best model



**Table 4. Performance measure of all 20 classifier models**

Comparing the performances (i.e. accuracy, TR rates for both classes), we have chosen 2 best classifier models: Logistic classifier with OneR attributes, and J48+AdaBoostM1with OneR attributes. Since, J48+AdaBoostM1 classifier model with OneR attribute selection method works well when the task/assumption is to predict whether or not the functioning graft will be successfully functioning in human’s body, in cases, when 0 (yes) values is more important. On the other hand, Logistic classifier model with OneR would work to predict and assess both values 0 and 1.

## What each team member did for this project

We both put on this project equally amount of effort and contributed to building the model with its best performance.

We analyzed all Attribute Selection Algorithms on Weka such as Information Gain, Gain Ratio, Correlation, One R, ReliefF, Symmetrical Uncertain and then run model building algorithms on them.

Sholpan analyzed all attribute selection algorithms along with model building algorithms such as Simple Logistic, Logistic, Naïve Bayes, J48, Random Forest on Weka

Aizhan analyzed all attribute selection algorithms along with model building algorithms such as Random Tree, Decision Table, One R, Neural Networks (Multilayer Perceptron)) on Weka.

After that we chose best 5 attribute selection algorithms (including manual one) along with 4 model building classifiers.

# **Conclusion**

**According to the results of the models obtained we chose two models, since we had an unbalanced dataset:**

**J48 + AdaBoostM1 with One R attribute selection algorithm.**

In this classifier model, the accuracy level is 91.90% with TP 0.97 for 0 values and 0.03 for values 1. Even if the overall accuracy is high, this model is not useful for functioning graft failure prediction when values 1 matters. This model works well when the task/assumption is to predict whether the functioning graft will be successfully functioning in human’s body, or whether a person's body will accept the functioning graft. In other words, this is a prediction model for the assessment of functioning graft and its compatibility chances with humans’ body based on medical characteristics. So, this model is useful for prediction when 0 (functioning graft success) values matters more.

**Logistic Classifier with One R attribute selection algorithm.**

With this classifier model, the accuracy level is 70% with TP 0.0707 for 0 values and 0.697 for values 1. Even if the overall accuracy is not high, this model is useful when both class attributes are important, and works on both ways, whereas J48 + AdaBoostM1 with One R attribute selection algorithm works only on one way for our dataset.

**What we learned from the project is that there is no classifier model that would work for every scenario equally well. A classifier model should be selected based on different scenarios, where we need to emphasize the attention on which class is more important than the others.**

# **References**

[1] Lorent, M., Giral, M. (2016). Mortality Prediction after the First Year of Kidney Transplantation: An Observational Study on Two European Cohorts.*www. plos.figshare.com.* Retrieved November, 2019 from <https://plos.figshare.com/articles/Mortality_Prediction_after_the_First_Year_of_Kidney_Transplantation_An_Observational_Study_on_Two_European_Cohorts/3366433>

[2] McGee, J., Magnus, J. (2010). Donor-Recipient Gender and Size Mismatch Affects Graft Success after Kidney Transplantation. Retrieved November, 2019 from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2887601/>

[3] Department of Internal Medicine, University of Illinois (2011). Peritoneal dialysis versus hemodialysis: risks, benefits, and access issues. Retrieved November, 2019 from <https://www.ncbi.nlm.nih.gov/pubmed/22098661>

[4] Takemoto, S., Port, F., Claas, F., Duquesnoy, R. (2004). HLA Matching for Kidney Transplantation. Retrieved November, 2019 from <http://www.epitopes.net/pdfs/other/q12Takemoto.pdf>

[5] Davis, C. Creatinine (Low, High, Blood Test Results Explained). Retrieved November, 2019 from <https://www.medicinenet.com/creatinine_blood_test/article.htm>

[6] Cunha, J.,Balentine, J. Epstein-Barr Virus (EPV, Mononucleosis, Mono). Retrieved November, 2019 from <https://www.emedicinehealth.com/epstein-barr_virus_infection/article_em.htm#what_are_risk_factors_and_causes_of_epstein-barr_virus_infections>

[7] Han, J., Kamber, M., Pei. J., (April 13, 2018). Data Mining Concepts and Techniques. *3rd Edition*. Published, 2012 by Elsevier Inc.