BOSTON UNIVERSITY

MET CS 699 Data Mining Project

On Final Report of the Asian American Quality of Life

Philip Chang, Min Cheng

1. Statement of Data Mining Goal

The U.S. Census defines Asian Americans as individuals having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent (U.S. Office of Management and Budget, 1997). Currently, in the U.S, 18.64 million Asian Americans represent 7% of the nation's overall population in 2021. The number is projected to surpass 46 million by 2060. (Pew Research)

Asian American community is a unique community that is diverse in culture, yet shares many similarities within. As Asian American population exponentially grows from minority to majority, we wish to better understand the current social and health state of the community.

Our goal in this project is to predict Asian Americans' satisfaction level with their overall "Quality of life", which is also our class attribute, based on different attributes. There are many attributes such as housing, salary, family ties, religion, and ethnicity which would help us better understand what are the driving forces of Asian Americans' quality of life.

2. Detailed description of the dataset

Our team's dataset was the Final Report of the Asian American Quality of Life (AAQoL), a compiled individual survey which was conducted on Asian American population in the city of Austin, Texas.

Survey questions were divided into answers to 7 different sections: Demographic, Immigration and Acculturation, Health, Special Interest, Social and Community Resources, Life in the City of Austin. The original dataset consists of 231 columns and 2,609 tuples of Asian Americans' survey results, including attributes like Age, Gender, Ethnicity, Marital Status, Education Status, Household Size, Religion, Employment Status, Income, English Level, Family Connection, Transportation Modes, and more. Each 2,609 unique survey represents an individual's information, which can be used for attributes determining the quality of life. Considering the design of the survey, the dataset mainly consisted of nominal and ordinal data.

The full list of attributes and the corresponding descriptions are in the link below.
https://data.austintexas.gov/City-Government/Final-Report-of-the-Asian-American-Quality-of-Life/hc5t-p62z

3. Brief description of data mining tool(s) used

Various tools were used for different purposes throughout the project.
Once the data has been extracted from the source, our team used R for the Data Preprocessing task. We have dropped unnecessary attributes, such as the 'Other'

sections, where surveyors would input character values for further explanation. R was used to detect any missing values, noisy data, and inconsistencies between variables. Outlier detection has been conducted using R as well. Finally, once all cleaning and prepping have been completed, we split the data into training and testing with R.

For attribute selection tasks, we utilized Weka. Weka provides easy access to various attribute selection methods. For model building and testing, performance analysis and visualization, we used Weka.

4. Brief description of classification algorithms you used.

We built 5 different models to predict our class attribute(Quality of life):
Naive Bayes, SimpleLogistic, Bagging, ClassificationViaRegression, RandomForest

- **Naive Bayes** is a Bayesian classification model with the assumption that attributes are related to each other. This model attempts to calculate the probability of class membership. This is a simple but powerful model comparable performance with decision trees and selected neural networks.

$$P(\mathbf{X}|C_i)=\prod_{k=1}^{n} P(x_k|C_i)=P(x_1|C_i)\times P(x_2|C_i)\times\ldots\times P(x_n|C_i)$$

- **SimpleLogistic** is a classifier for building linear logistic regression models. LogitBoost with simple regression functions as base learners is used for fitting the logistic models. The optimal number of LogitBoost iterations to perform is cross-validated, which leads to automatic attribute selection.

- **Bagging** or bootstrap aggregating model averages the prediction over a collection of classifiers. This model is an ensemble method and provides significantly better accuracy than a single classifier derived from the dataset.

- **ClassificationViaRegression** is a class for doing classification using regression methods. Class is binarized and one regression model is built for each class value.

$$f(\mathbf{x};\hat{\mathbf{w}}) = w_0 + \mathbf{x}^T\hat{\mathbf{w}}_1,$$

to classify any new (test) example $\mathbf{x}$ according to

$$\text{label} = 1 \text{ if } f(\mathbf{x};\mathbf{w}) > 0.5, \text{ and label} = 0 \text{ otherwise}$$

- **Random Forest model** is also an ensemble method that builds multiple trees, and each tree classifies a given sample. This model is usually less susceptible to errors and outliers, handles unbalanced data well, and overfitting is not an issue.

Because only a subset of attributes is considered at each node, it is faster than most models.

5. Brief description of attribute selection methods you used.

We used 5 different attribute selection methods:
CfsSubset, Correlation, WrapperSubset, OneR, InfoGain

- **Correlation-based Feature Selection (CFS)** subset is an algorithm that couples this evaluation formula with an appropriate correlation measure and a heuristic search strategy (Waikato).

- **Correlation** evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class. Nominal attributes are considered on a value by value basis by treating each value as an indicator. An overall correlation for a nominal attribute is arrived at via a weighted average.

- **WrapperSubset** method wraps a classifier in a cross-validation loop: it searches through the attribute space and uses the classifier to find a good attribute set. Searching can be forwards, backwards, or bidirectional, starting from any subset. Cross validation is used to estimate the accuracy of the learning scheme for a set of attributes.

- **1R** or One Rule is a simple, robust and efficient classification algorithm that generates one rule for each predictor. The goal is to make rules based on a single attribute. The algorithm chooses the minimum-error attribute as the rule.

- **InfoGain** Evaluates the worth of an attribute by measuring the information gain with respect to the class. In other words, this method measures how each feature contributes in decreasing the overall entropy.

6. The set of attributes selected by each attribute selection method.

Attribute selection method	Set of attributes selected
CFS Subset Evaluator	11,15,23,24,28,31,35,37,38,51,95,106,120 : 13 Achieving.Ends.Meet English.Speaking Present.Mental.Health Present.Oral.Health Regular.Exercise Dentist.Check.up Satisfaction Satisfied.With.Life.1

	Satisfied.With.Life.2 See.Friends Library.Internet.Acess Satisfaction.With.Housing. Public.Meeting
Correlation Ranking Filter	106,38,11,37,13,35,23,72,74,24,15,22,55,54,62,28,45,80,71,14,73,50,56,49,59,92,120,17,16,5,77,53,123,61,33,58,57,78,1,82,20,51,60,85,46,10,31,18,93,101 : 50 0.1536 106 Satisfaction.With.Housing. 0.1416 38 Satisfied.With.Life.2 0.139 11 Achieving.Ends.Meet 0.1358 37 Satisfied.With.Life.1 0.1349 13 Duration.of.Residency 0.1283 35 Satisfaction 0.1275 23 Present.Mental.Health 0.1191 72 Place.to.Live 0.1145 74 Place.to.Work 0.1118 24 Present.Oral.Health 0.1084 15 English.Speaking 0.1058 22 Present.Health 0.1053 55 Similar.Values 0.1049 54 Family.Respect 0.1023 62 Feel.Close 0.102 28 Regular.Exercise 0.1012 45 Advanced.Directives 0.1006 80 Qualtiy.of.Life 0.1 71 Residency 0.0963 14 Primary.Language 0.0954 73 Raising.Children 0.0951 50 Helpful.Family 0.095 56 Successful.Family 0.0946 49 Close.Family 0.0942 59 Family.Pride 0.094 92 EMS.Classes 0.0932 120 Public.Meeting 0.0913 17 Familiarity.with.America 0.0884 16 English.Difficulties 0.0861 5 Education.Completed 0.0848 77 Arts.and.Culture 0.0817 53 Helpful.Friends 0.0816 123 City.Election 0.0814 61 Spend.Time.Together 0.0813 33 Dental.Insurance 0.0813 58 Loyalty 0.0799 57 Trust 0.0799 78 Safety 0.0789 1 Age 0.0777 82 Parks.and.Recs 0.0765 20 Belonging 0.0755 51 See.Friends 0.0743 60 Expression

	0.0735 85 Airport 0.0732 46 Have.an.Advanced.Directive 0.0731 10 Income 0.0695 31 Dentist.Check.up 0.0688 18 Familiarity.with.Ethnic.Origin 0.0683 93 Fire.Alarm 0.0679 101 X3.1.1
Wrapper Subset Evaluator	23,38,59,95,97,122,123 : 7 Present.Mental.Health Satisfied.With.Life.2 Family.Pride Library.Internet.Acess Citizenship.Class Contact.City.Official City.Election
OneR feature evaluator	38,37,23,22,106,24,17,35,80,81,120,78,71,125,121,126,49,50,45,41,86,19,87,65,83,46,13,59,62,61,27,26,42,14,18,8,10,28,57,43,11,44,54,33,48,29,51,53,21,64 : 50 58.8657 38 Satisfied.With.Life.2 58.0183 37 Satisfied.With.Life.1 56.0626 23 Present.Mental.Health 55.2803 22 Present.Health 54.8892 106 Satisfaction.With.Housing. 54.5632 24 Present.Oral.Health 52.9335 17 Familiarity.with.America 52.5424 35 Satisfaction 52.5424 80 Qualtiy.of.Life 52.2164 81 Quality.of.Service 51.6949 120 Public.Meeting 51.369 78 Safety 51.2386 71 Residency 51.1734 125 Informed 50.9778 121 Council.Meeting 50.9778 126 City.Effort.Satisfaction 50.9126 49 Close.Family 50.8475 50 Helpful.Family 50.7171 45 Advanced.Directives 50.6519 41 Prevention 50.6519 86 Austin.Energy 50.6519 19 Identify.Ethnically 50.5867 87 Court 50.5215 65 Religious.Importance 50.5215 83 Libraries 50.5215 46 Have.an.Advanced.Directive 50.4563 13 Duration.of.Residency 50.3911 59 Family.Pride 50.3911 62 Feel.Close 50.3911 61 Spend.Time.Together 50.3911 27 Drinking

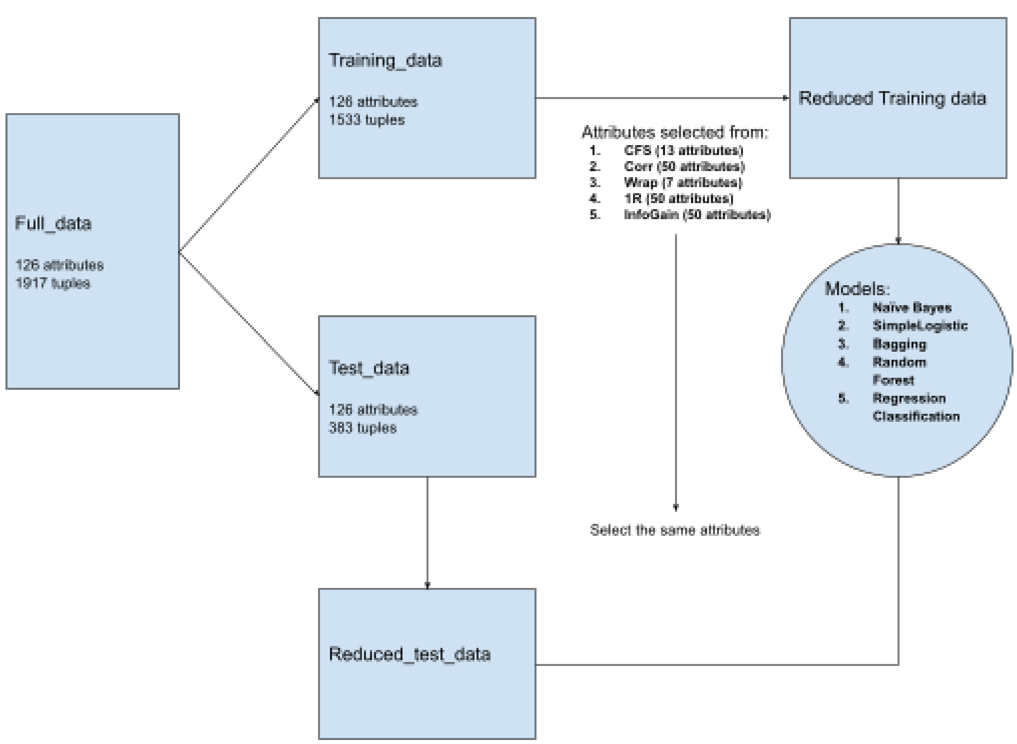
	50.3911 26 Smoking 50.3911 42 Aging..AD. 50.3911 14 Primary.Language 50.3911 18 Familiarity.with.Ethnic.Origin 50.3911 8 Retired 50.3911 10 Income 50.3911 28 Regular.Exercise 50.3911 57 Trust 50.3911 43 Cure..AD. 50.3911 11 Achieving.Ends.Meet 50.3911 44 Nursing.Home..AD. 50.3911 54 Family.Respect 50.3911 33 Dental.Insurance 50.3911 48 See.Family 50.3911 29 Healthy.Diet 50.3911 51 See.Friends 50.3911 53 Helpful.Friends 50.3911 21 Discrimination 50.3911 64 Religious.Attendance
Information Gain Ranking Filter	37,38,23,24,22,15,17,106,10,11,34,16,74,80,59,56,72,35,55,62,75,54,58,60,77,70,33,73,82,69,57,68,61,81,51,78,67,18,28,85,53,31,1,39,101,76,88,14,13,3 : 50 0.2846 37 Satisfied.With.Life.1 0.2714 38 Satisfied.With.Life.2 0.1715 23 Present.Mental.Health 0.1402 24 Present.Oral.Health 0.1305 22 Present.Health 0.1087 15 English.Speaking 0.0877 17 Familiarity.with.America 0.0813 106 Satisfaction.With.Housing. 0.0806 10 Income 0.0708 11 Achieving.Ends.Meet 0.0672 34 Language 0.0601 16 English.Difficulties 0.0583 74 Place.to.Work 0.0524 80 Qualtiy.of.Life 0.0517 59 Family.Pride 0.0516 56 Successful.Family 0.0509 72 Place.to.Live 0.0503 35 Satisfaction 0.0502 55 Similar.Values 0.0494 62 Feel.Close 0.0484 75 Small.Businesses 0.0477 54 Family.Respect 0.0448 58 Loyalty 0.0429 60 Expression 0.0425 77 Arts.and.Culture 0.038 70 Community.Trust 0.0379 33 Dental.Insurance 0.0367 73 Raising.Children 0.0362 82 Parks.and.Recs

	0.0355 69 Get.Along
	0.0354 57 Trust
	0.0353 68 Community.Shares.Values
	0.0352 61 Spend.Time.Together
	0.0342 81 Quality.of.Service
	0.0332 51 See.Friends
	0.0325 78 Safety
	0.032 67 Helpful.Community
	0.0315 18 Familiarity.with.Ethnic.Origin
	0.0314 28 Regular.Exercise
	0.0314 85 Airport
	0.0306 53 Helpful.Friends
	0.0301 31 Dentist.Check.up
	0.0295 1 Age
	0.0282 39 Knowledge
	0.0277 101 X3.1.1
	0.0273 76 Place.to.Retire
	0.0268 88 Social.Services
	0.0259 14 Primary.Language
	0.0257 13 Duration.of.Residency
	0.0256 3 Ethnicity

7. Detailed description of data mining procedure.

Our team followed a full testing process of the selected models, described in the Diagram below.

Full Process Diagram

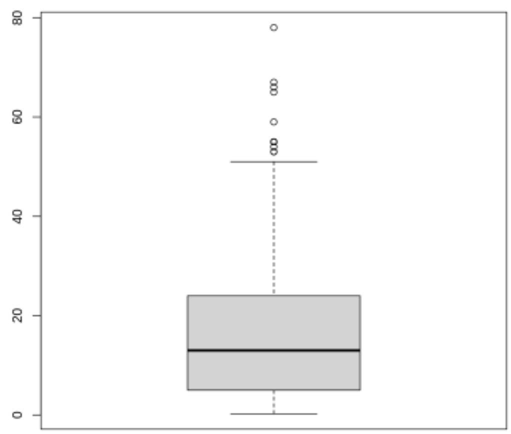


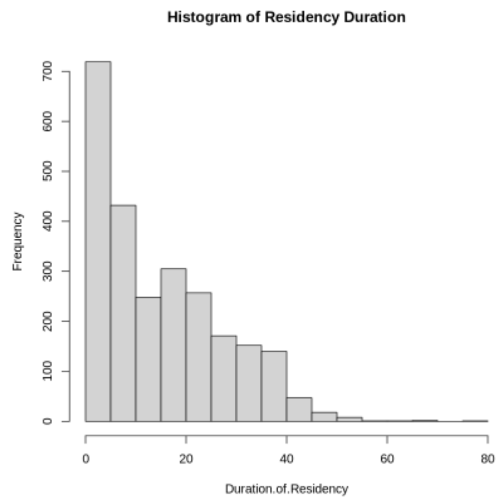
7.1 Data Preprocessing

- **Step1**. Checking for Missing Value
  We removed NA values on the surveyor's demographic information.
- **Step2**. Encoding categorical data
  The "No.One" column is categorical data with 2 levels, "living with no one" and "0", we encoded it into "1", "0".
- **Step3**. Checking for Inconsistent data
  We removed inconsistent data of "household size" and "living with no one".
- **Step4**. Checking for Outliers
  6 people over the age of 80 are detected as outliers. 2 Surveyors with duration of residency over 50 years are detected as outliers. We decided against excluding age as a factor of outlier detection and removed outliers detected with Duration of Residency.
- **Step5**. Data Reduction
  Because this is survey data, many entries are not subject to mining. These surveyor inputs are unnecessary, and therefore dropped.
- **Step6**. Reformatting the class attribute
  We reformatted the class attribute "Quality of life" to factors with wider bin size.
- **Step7**. Splitting the dataset into the training and test set
- **Step8**. Data formatting and export

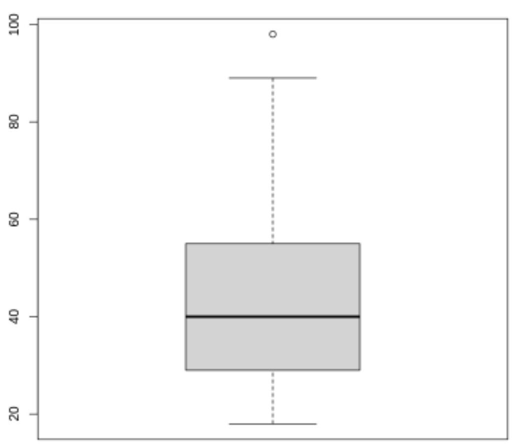
```
csv <-
read.csv('Final_Report_of_the_Asian_American_Quality_of_Life__AAQoL_.
csv')
tib <- as_tibble(csv)
head(tib)

# Removing NA values on surveyor's demographic information
tib1 <- tib[complete.cases(tib[ , 2:7]),]

# Encoding categorical data
tib2$No.One = factor(tib2$No.One,
              levels = c('Living with no one', '0'),
              labels = c(1, 0))

# Checking for Inconsistent data
subset1 <- subset(tib2, tib2$No.One == 1 & tib2$Household.Size > 1)
subset1
anti_join(tib2, subset1) -> tib3
```

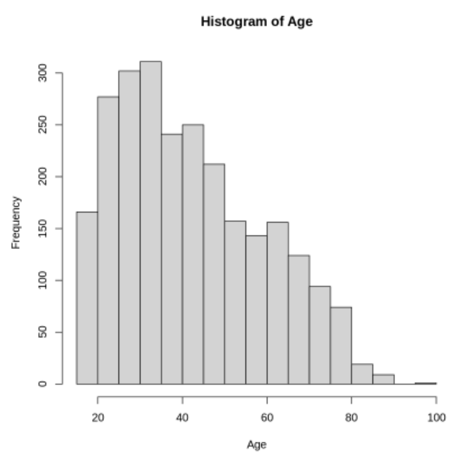
```
tib3

# Checking for Outliers
summary(tib3)
boxplot(tib3$Duration.of.Residency)
hist(tib3$Duration.of.Residency, xlab = "Duration.of.Residency", main
= "Histogram of Residency Duration")
```





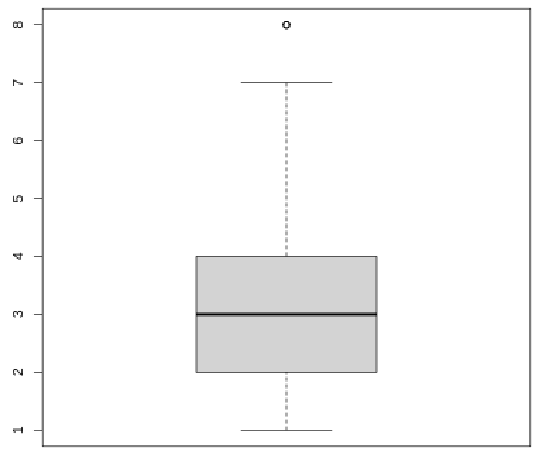

```
boxplot(tib3$Age)
hist(tib3$Age, xlab = "Age", main = "Histogram of Age")
```

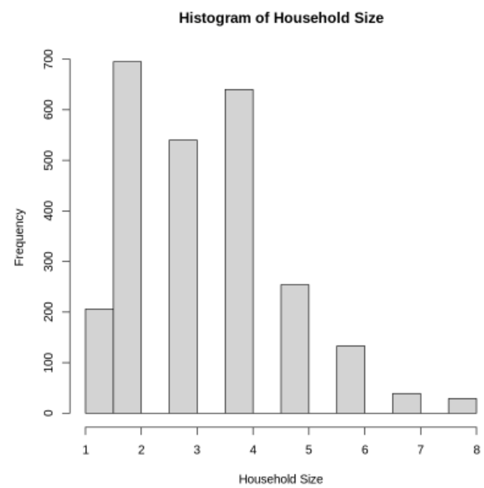





```
boxplot(tib3$Household.Size)
```

```
hist(tib3$Household.Size, xlab = "Household Size", main = "Histogram
of Household Size")
```






```
# Removing Outliers detected with Duration of Residency. Based on the
technique covered in class.
Q1 <- quantile(tib3$Duration.of.Residency, .25, na.rm = T)
Q3 <- quantile(tib3$Duration.of.Residency, .75, na.rm = T)
IQR <- IQR(tib3$Duration.of.Residency, na.rm = T)
tib4 <- subset(tib3, tib3$Duration.of.Residency > (Q1 - 1.5*IQR) &
tib3$Duration.of.Residency < (Q3 + 1.5*IQR))

dim(tib3)
dim(tib4)

##### Data Reduction #####
# Dimensionality Reduction
tib5 <- tib4[-c(8:14, 19:20, 47:56, 69, 71, 73:81, 85:88, 195:209)]
colnames(tib4[c(8:14, 19:20, 47:56, 69, 71, 73:81, 85:88, 195:209)])

# Column Unemployed and Disabled are 0.
tib5 <- tib5[-c(14, 15)]
tib6 <- na.omit(tib5)
summary(tib6)
# reformatting the class attribute to factors with wider bin size
```

```
tib7 <- tib6
QoL_factor <- cut(as.numeric(tib6$Quality.of.Life), breaks = c(0, 3,
4, 6, 8, 10), labels = 1:5)
tib7$Quality.of.Life <- QoL_factor

#Creating the training set and test set separately
library(caTools)
set.seed(123)
split = sample.split(tib6$Quality.of.Life, SplitRatio = 0.8)# returns
true if observation goes to the Training set and false if observation
goes to the test set.

training_set = subset(tib6, split == TRUE)
test_set = subset(tib6, split == FALSE)
training_set
Test_set

# Exporting to csv for further mining process
write.csv(tib7, "full_set.csv")
write.csv(test_set, "test_set.csv")
write.csv(training_set, "training_set.csv")
```

7.2 Attribute Selection

With the training and test dataset generated, our team then moved to the attribute selection process. In this process, as our Process Diagram shows, we ran all 5 attribute selections in Weka on the training set generated. Each training set with selected attributes were separately saved.

7.3 Model Generation

With the reduced training dataset of selected attributes, our team started generating 5 models selected from above 5 classification methods. Once the model was generated with each of the 5 reduced training dataset, we tested out models with a corresponding test dataset with the same attributes.

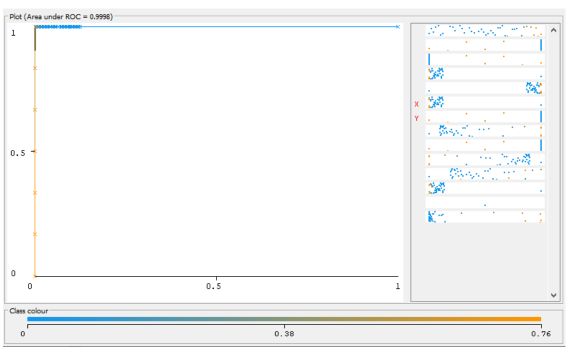
8. Data mining result and evaluation:

Conclusion: According to the performances of 25 classification models below, we concluded that using **WrapperSubset** attribute selection method and **RandomForest** classification algorithm generates the best model. (*Accuracy = 78.0679%, ROC Area= 0.921, PRC Area = 0.874)* PRC Area has been used to evaluate the performance along with ROC Area, given that the dataset is unbalanced.

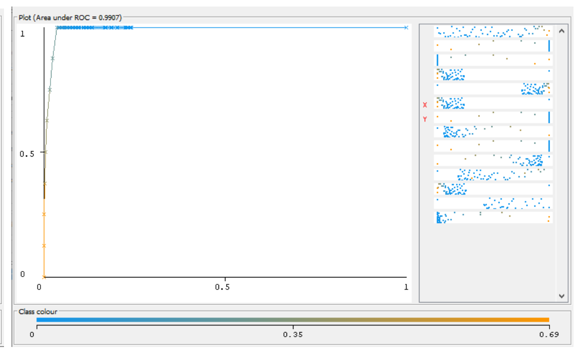
As we can see in the Curves below, for each class, the predicted class has been mostly accurate, especially for the lower ratings for quality of life.

ROC Curves for each class

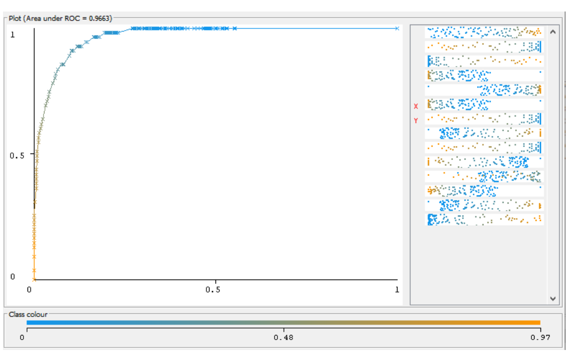
Class1



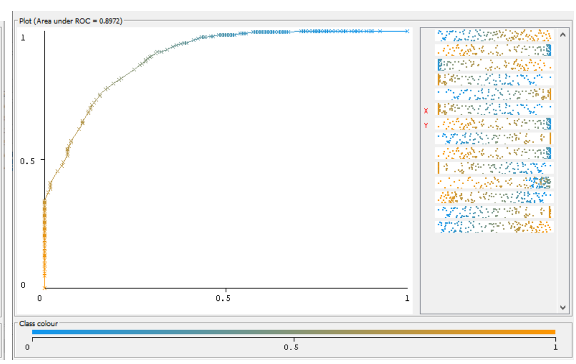
Class2



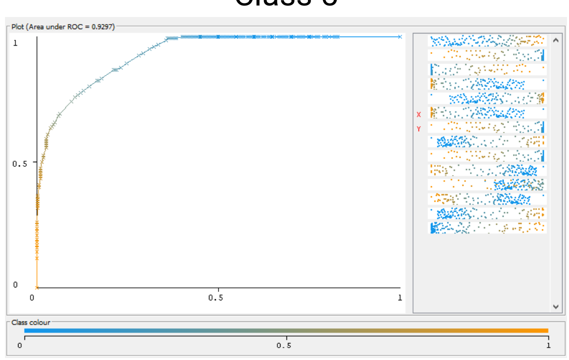
Class 3



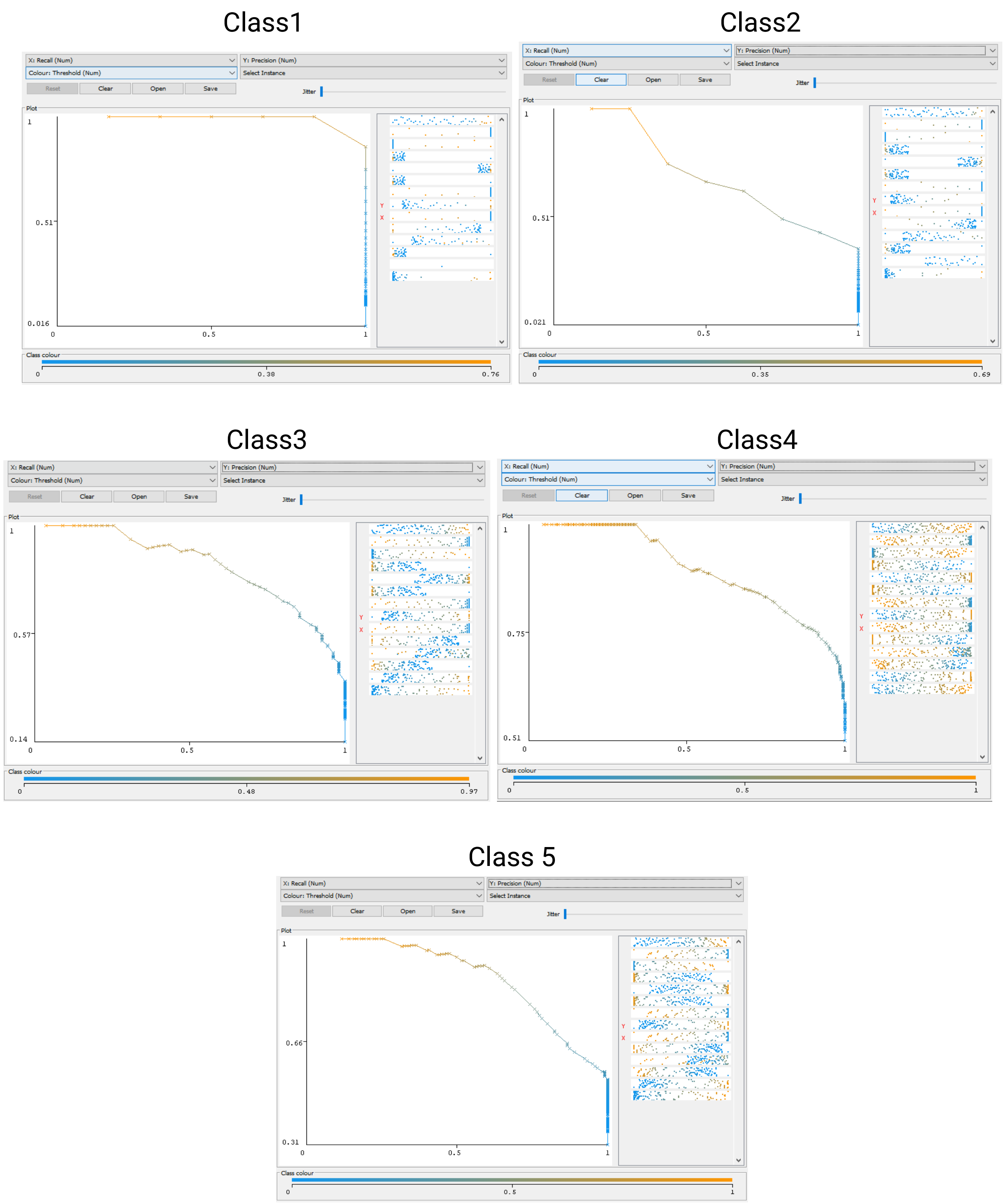
Class4



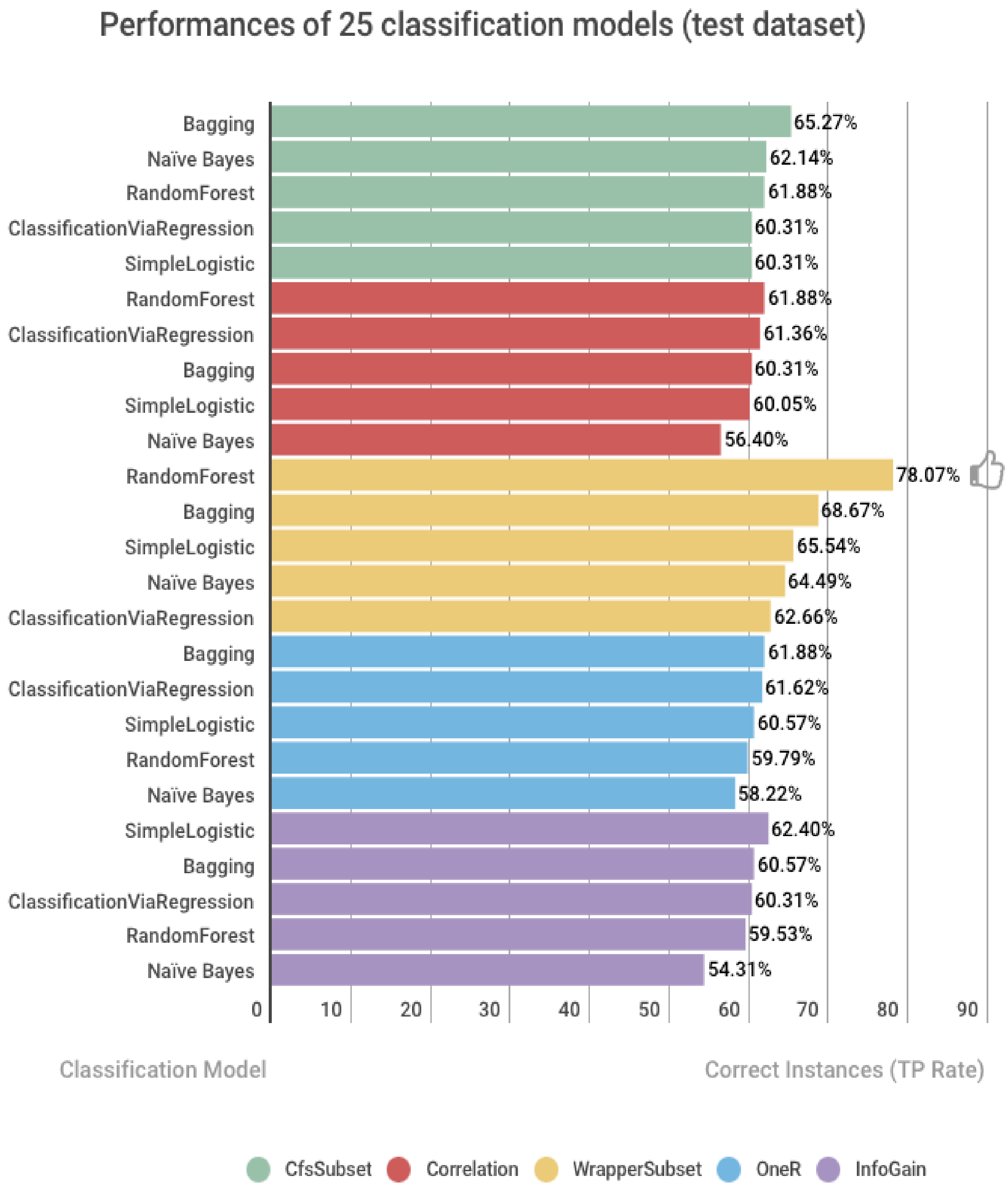
Class 5



PRC for each class

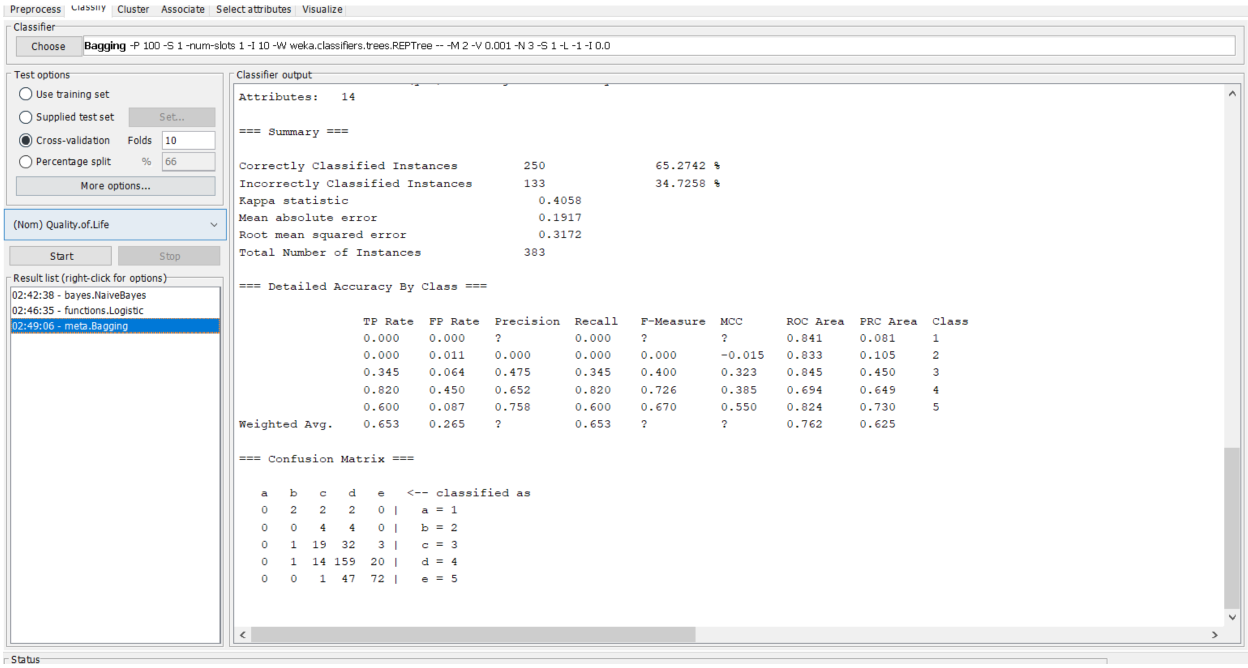


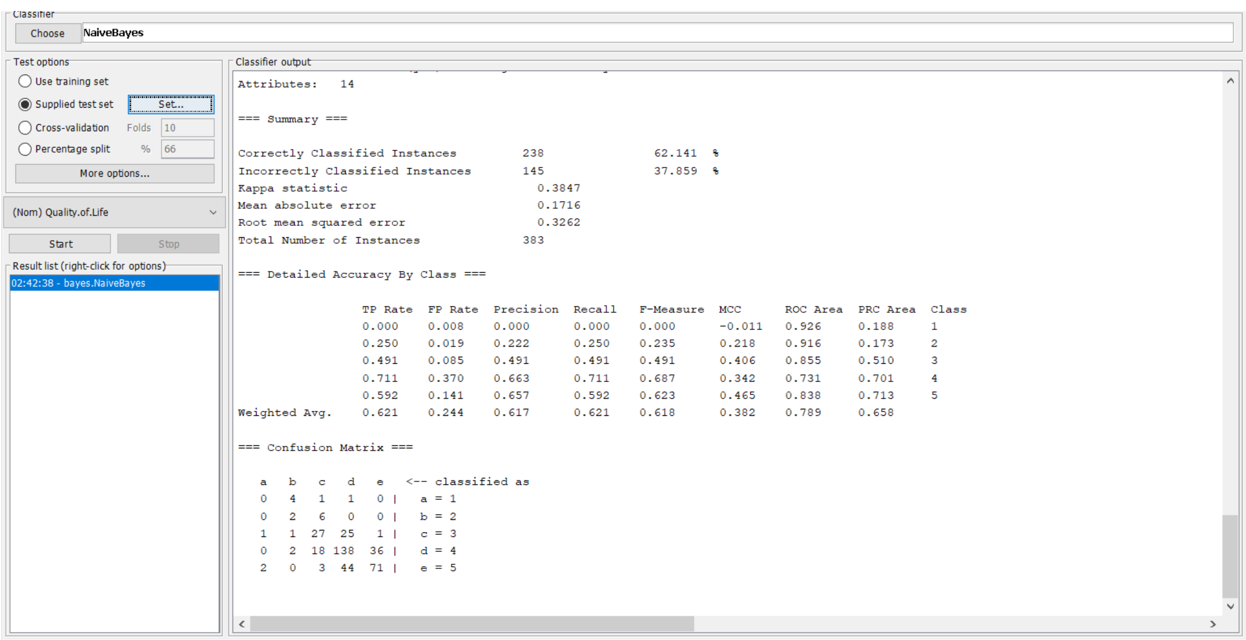
One of the reasons the Random Forest model performs best is because of the unbalanced nature of the dataset. Random forest tries to minimize the overall error rate, so when we have an unbalanced data set, the larger class will get a low error rate while the smaller class will have a larger error rate. Random Forest model builds multiple decision trees and merges them together to get a more accurate and stable prediction. It is also worth noting that the Random Forest model works with subsets of data, and thus works well with high dimensional data.

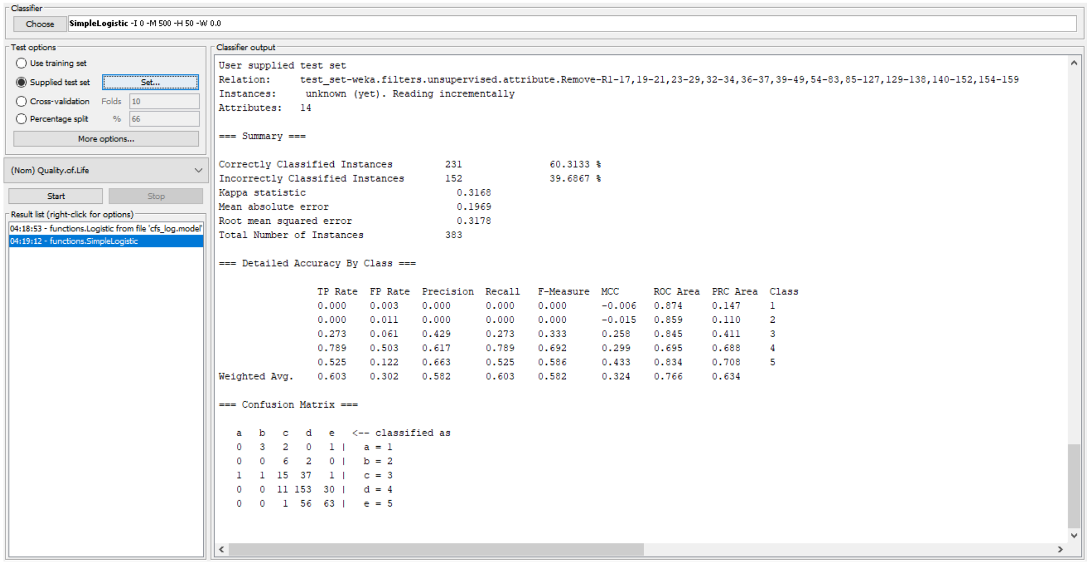


8.1 Results of testing models on reduced test datasets:

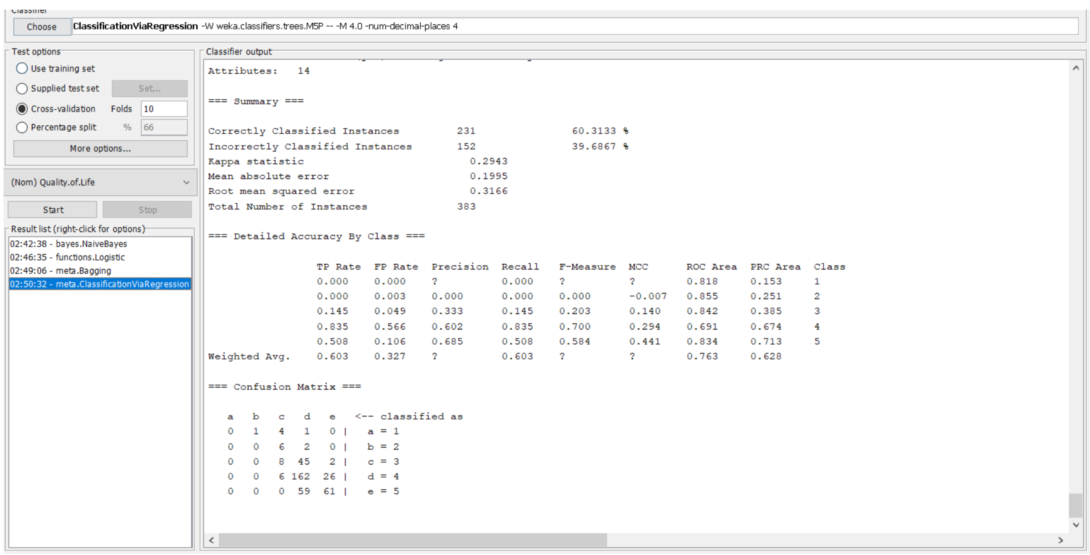
- Reduced test dataset1: CfsSubset



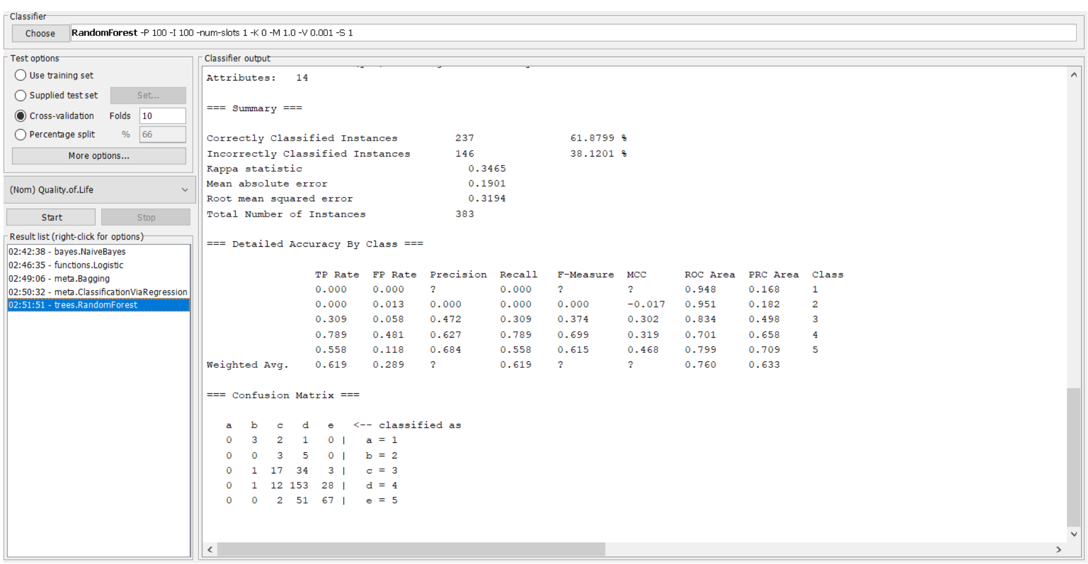




Classifier
Choose SimpleLogistic -I 0 -M 500 -H 50 -W 0.0
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Quality.of.Life
Start
Stop
Result list (right-click for options)
04:18:53 - functions.Logistic from file 'cfs_log.model'
04:19:12 - functions.SimpleLogistic
Classifier output
User supplied test set
Relation: test_set-weka.filters.unsupervised.attribute.Remove-R1-17,19-21,23-29,32-34,36-37,39-49,54-83,85-127,129-138,140-152,154-159
Instances: unknown (yet). Reading incrementally
Attributes: 14
=== Summary ===
Correctly Classified Instances 231 60.3133 %
Incorrectly Classified Instances 152 39.6867 %
Kappa statistic 0.3168
Mean absolute error 0.1969
Root mean squared error 0.3178
Total Number of Instances 383
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.000 0.003 0.000 0.000 0.000 -0.006 0.874 0.147 1
0.000 0.011 0.000 0.000 0.000 -0.015 0.859 0.110 2
0.273 0.061 0.429 0.273 0.333 0.258 0.845 0.411 3
0.789 0.503 0.617 0.789 0.692 0.299 0.695 0.688 4
0.525 0.122 0.663 0.525 0.586 0.433 0.834 0.708 5
Weighted Avg. 0.603 0.302 0.582 0.603 0.582 0.324 0.766 0.634
=== Confusion Matrix ===
a b c d e <-- classified as
0 3 2 0 1 | a = 1
0 0 6 2 0 | b = 2
1 1 15 37 1 | c = 3
0 0 11 153 30 | d = 4
0 0 1 56 63 | e = 5

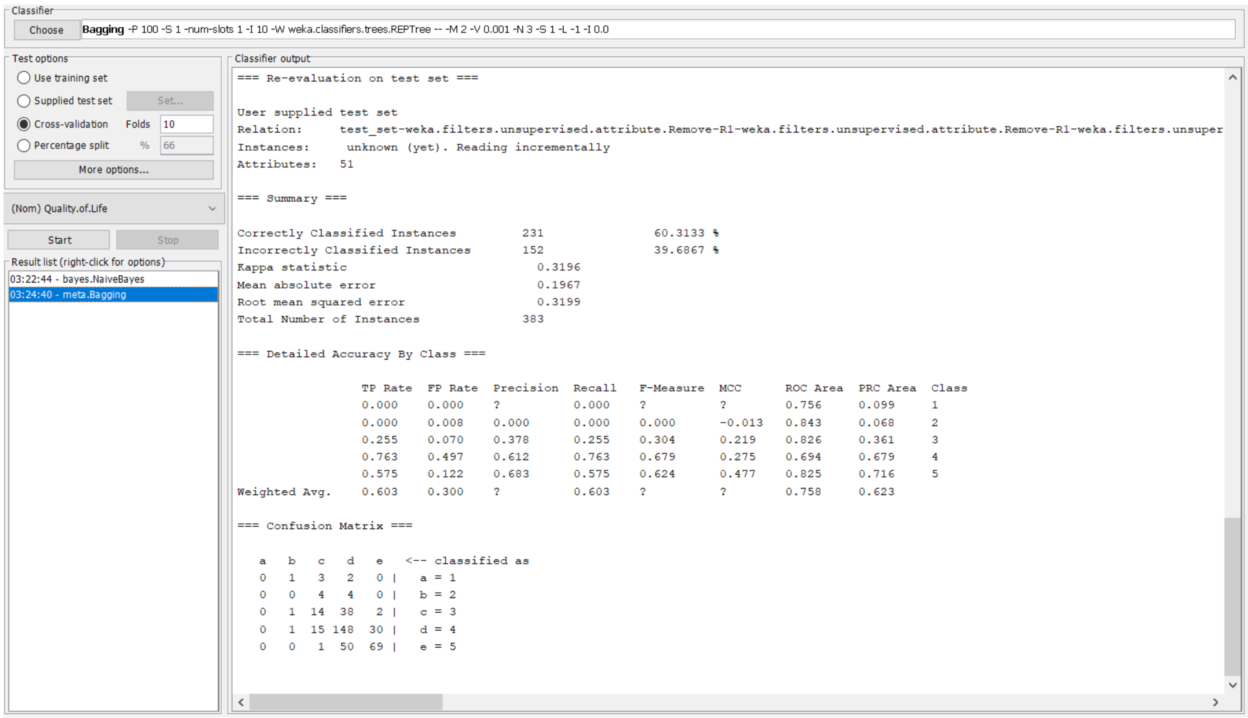


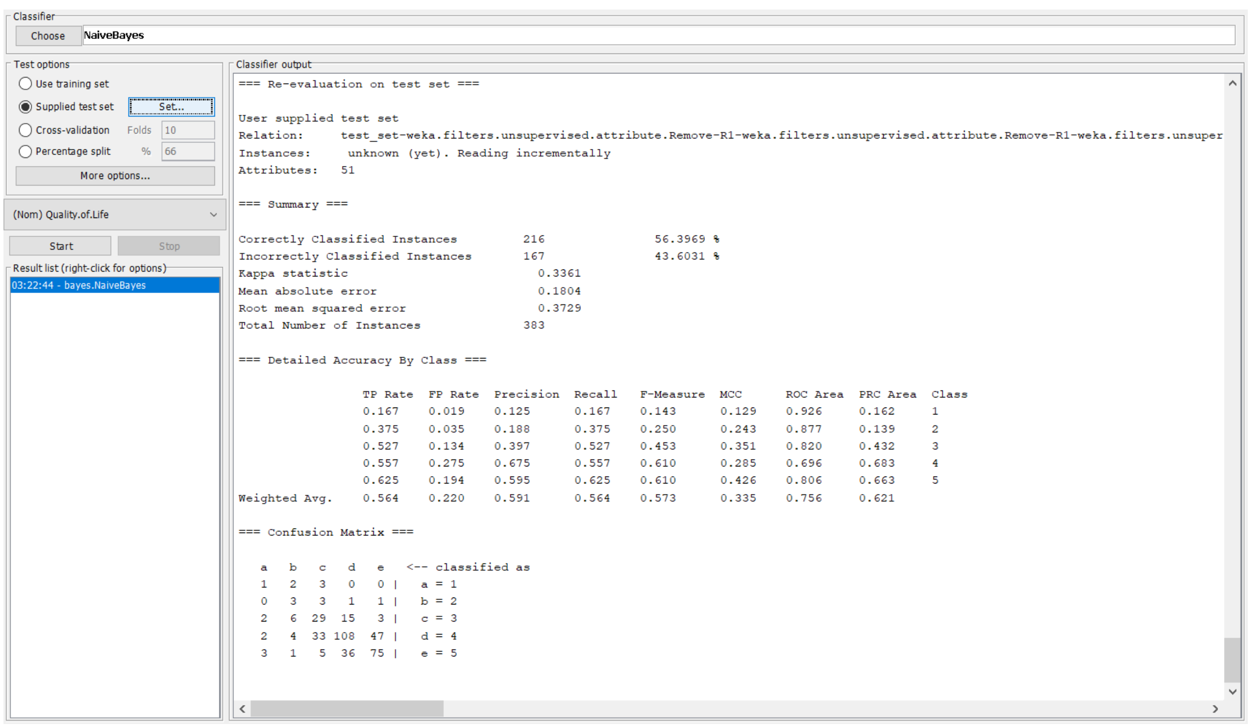
Choose ClassificationViaRegression -W weka.classifiers.trees.M5P -- -M 4.0 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Quality.of.Life
Start
Stop
Result list (right-click for options)
02:42:38 - bayes.NaiveBayes
02:46:35 - functions.Logistic
02:49:06 - meta.Bagging
02:50:32 - meta.ClassificationViaRegression
Classifier output
Attributes: 14
=== Summary ===
Correctly Classified Instances 231 60.3133 %
Incorrectly Classified Instances 152 39.6867 %
Kappa statistic 0.2943
Mean absolute error 0.1995
Root mean squared error 0.3166
Total Number of Instances 383
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.000 0.000 ? 0.000 ? ? 0.818 0.153 1
0.000 0.003 0.000 0.000 0.000 -0.007 0.855 0.251 2
0.145 0.049 0.333 0.145 0.203 0.140 0.842 0.385 3
0.835 0.566 0.602 0.835 0.700 0.294 0.691 0.674 4
0.508 0.106 0.685 0.508 0.584 0.441 0.834 0.713 5
Weighted Avg. 0.603 0.327 ? 0.603 ? ? 0.763 0.628
=== Confusion Matrix ===
a b c d e <-- classified as
0 1 4 1 0 | a = 1
0 0 6 2 0 | b = 2
0 0 8 45 2 | c = 3
0 0 6 162 26 | d = 4
0 0 0 59 61 | e = 5

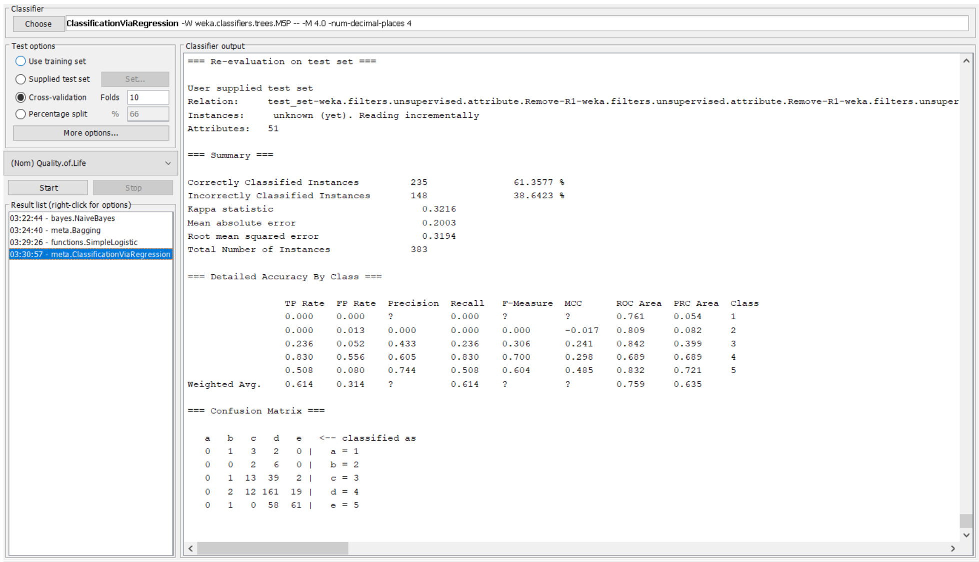


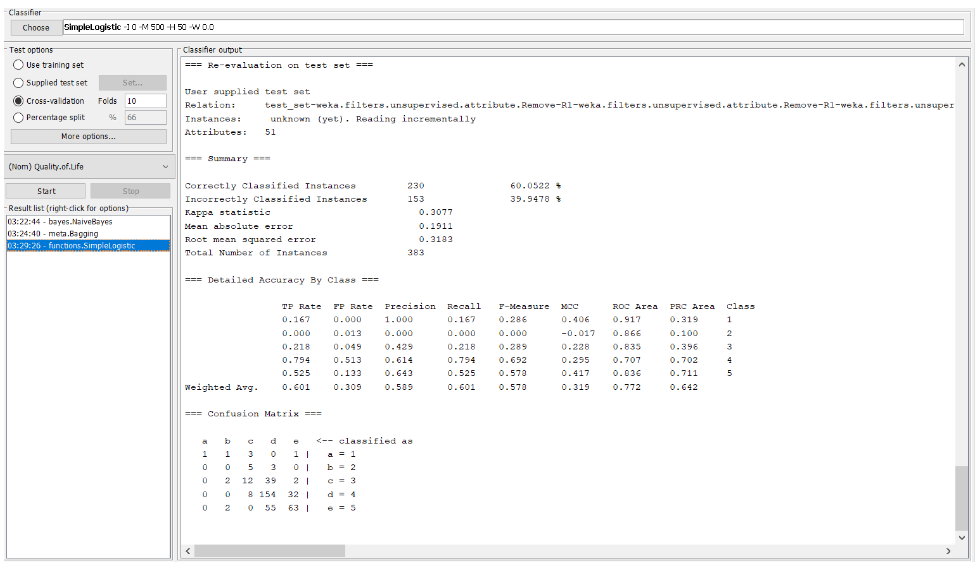
Classifier
Choose RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Quality.of.Life
Start
Stop
Result list (right-click for options)
02:42:38 - bayes.NaiveBayes
02:46:35 - functions.Logistic
02:49:06 - meta.Bagging
02:50:32 - meta.ClassificationViaRegression
02:51:51 - trees.RandomForest
Classifier output
Attributes: 14
=== Summary ===
Correctly Classified Instances 237 61.8799 %
Incorrectly Classified Instances 146 38.1201 %
Kappa statistic 0.3465
Mean absolute error 0.1901
Root mean squared error 0.3194
Total Number of Instances 383
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.000 0.000 ? 0.000 ? ? 0.948 0.168 1
0.000 0.013 0.000 0.000 0.000 -0.017 0.951 0.182 2
0.309 0.058 0.472 0.309 0.374 0.302 0.834 0.498 3
0.789 0.481 0.627 0.789 0.699 0.319 0.701 0.658 4
0.558 0.118 0.684 0.558 0.615 0.468 0.799 0.709 5
Weighted Avg. 0.619 0.289 ? 0.619 ? ? 0.760 0.633
=== Confusion Matrix ===
a b c d e <-- classified as
0 3 2 1 0 | a = 1
0 0 3 5 0 | b = 2
0 1 17 34 3 | c = 3
0 1 12 153 28 | d = 4
0 0 2 51 67 | e = 5

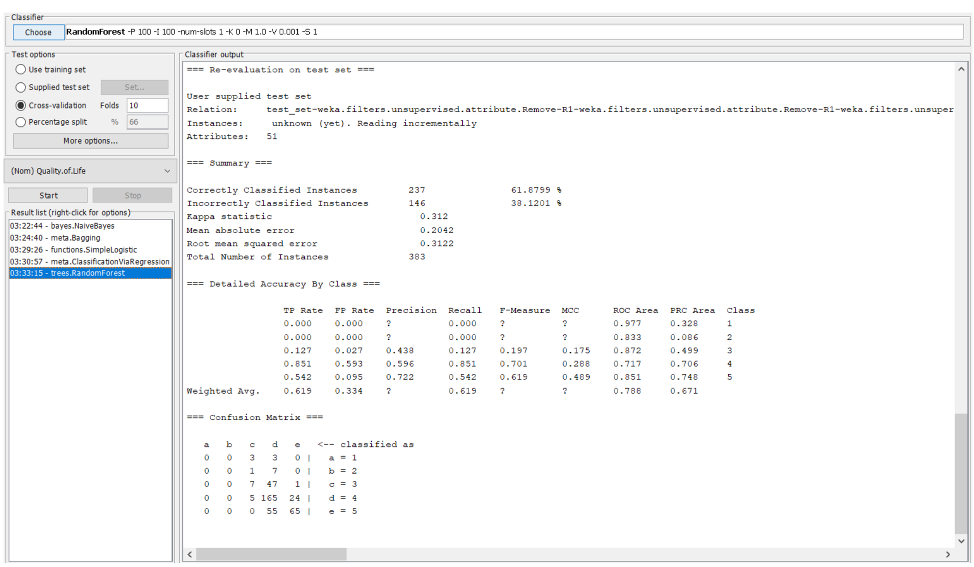
- Reduced test dataset2: Correlation



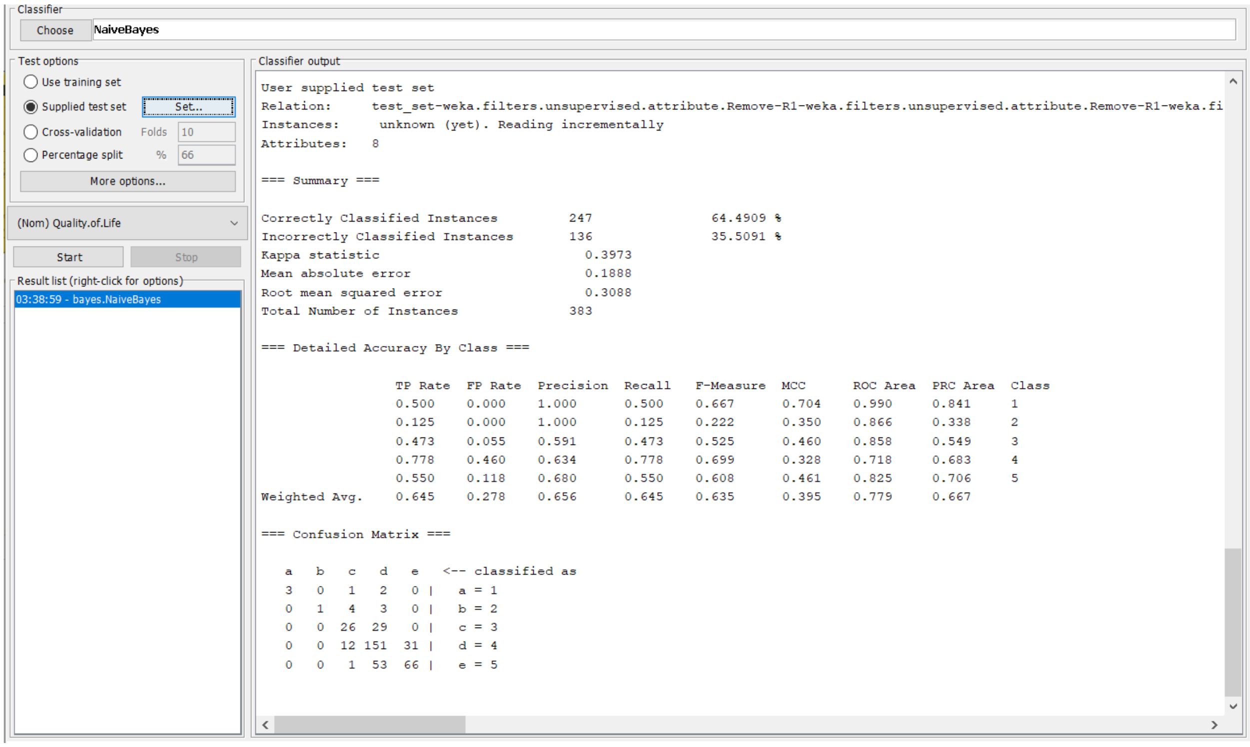


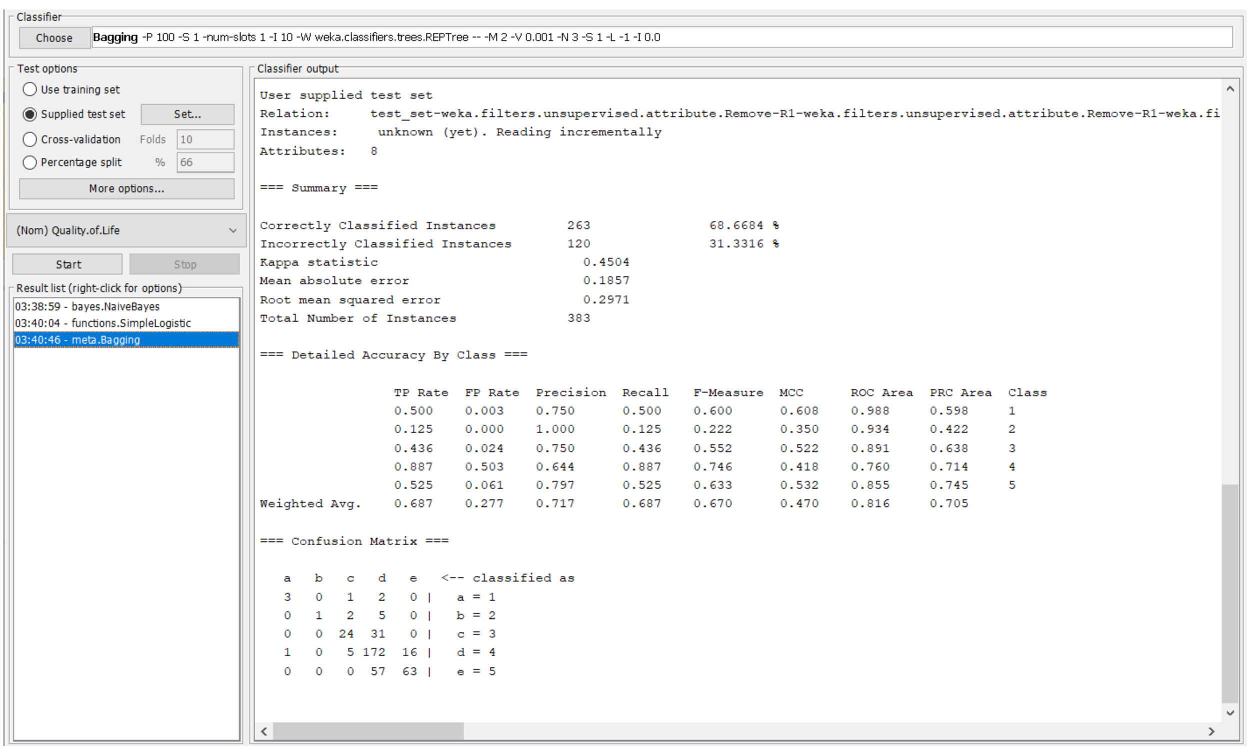

Classifier
Choose
ClassificationViaRegression -W weka.classifiers.trees.M5P -- -M 4.0 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Quality.of.Life
Start
Stop
Result list (right-click for options)
03:22:44 - bayes.NaiveBayes
03:24:40 - meta.Bagging
03:29:26 - functions.SimpleLogistic
03:30:57 - meta.ClassificationViaRegression
Classifier output
=== Re-evaluation on test set ===
User supplied test set
Relation: test_set-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsuper
Instances: unknown (yet). Reading incrementally
Attributes: 51
=== Summary ===
Correctly Classified Instances 235 61.3577 %
Incorrectly Classified Instances 148 38.6423 %
Kappa statistic 0.3216
Mean absolute error 0.2003
Root mean squared error 0.3194
Total Number of Instances 383
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.000 0.000 ? 0.000 ? ? 0.761 0.054 1
0.000 0.013 0.000 0.000 0.000 -0.017 0.809 0.082 2
0.236 0.052 0.433 0.236 0.306 0.241 0.842 0.399 3
0.830 0.556 0.605 0.830 0.700 0.298 0.689 0.689 4
0.508 0.080 0.744 0.508 0.604 0.485 0.832 0.721 5
Weighted Avg. 0.614 0.314 ? 0.614 ? ? 0.759 0.635
=== Confusion Matrix ===
a b c d e <-- classified as
0 1 3 2 0 | a = 1
0 0 2 6 0 | b = 2
0 1 13 39 2 | c = 3
0 2 12 161 19 | d = 4
0 1 0 58 61 | e = 5

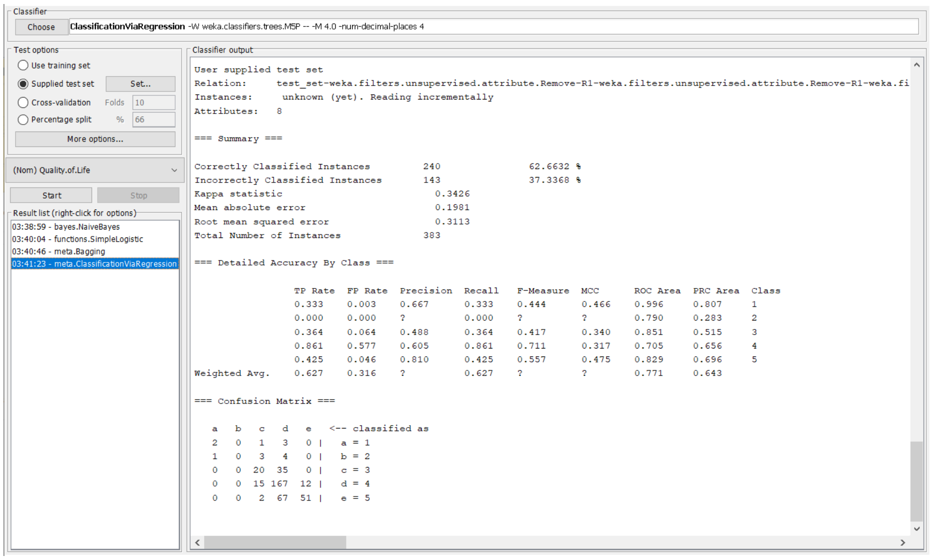

Classifier
Choose
SimpleLogistic -I 0 -M 500 -H 50 -W 0.0
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Quality.of.Life
Start
Stop
Result list (right-click for options)
03:22:44 - bayes.NaiveBayes
03:24:40 - meta.Bagging
03:29:26 - functions.SimpleLogistic
Classifier output
=== Re-evaluation on test set ===
User supplied test set
Relation: test_set-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsuper
Instances: unknown (yet). Reading incrementally
Attributes: 51
=== Summary ===
Correctly Classified Instances 230 60.0522 %
Incorrectly Classified Instances 153 39.9478 %
Kappa statistic 0.3077
Mean absolute error 0.1911
Root mean squared error 0.3183
Total Number of Instances 383
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.167 0.000 1.000 0.167 0.286 0.406 0.917 0.319 1
0.000 0.013 0.000 0.000 0.000 -0.017 0.866 0.100 2
0.218 0.049 0.429 0.218 0.289 0.228 0.835 0.396 3
0.794 0.513 0.614 0.794 0.692 0.295 0.707 0.702 4
0.525 0.133 0.643 0.525 0.578 0.417 0.836 0.711 5
Weighted Avg. 0.601 0.309 0.589 0.601 0.578 0.319 0.772 0.642
=== Confusion Matrix ===
a b c d e <-- classified as
1 1 3 0 1 | a = 1
0 0 5 3 0 | b = 2
0 2 12 39 2 | c = 3
0 0 8 154 32 | d = 4
0 2 0 55 63 | e = 5


Classifier
Choose
RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Quality.of.Life
Start
Stop
Result list (right-click for options)
03:22:44 - bayes.NaiveBayes
03:24:40 - meta.Bagging
03:29:26 - functions.SimpleLogistic
03:30:57 - meta.ClassificationViaRegression
03:33:15 - trees.RandomForest
Classifier output
=== Re-evaluation on test set ===
User supplied test set
Relation: test_set-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsuper
Instances: unknown (yet). Reading incrementally
Attributes: 51
=== Summary ===
Correctly Classified Instances 237 61.8799 %
Incorrectly Classified Instances 146 38.1201 %
Kappa statistic 0.312
Mean absolute error 0.2042
Root mean squared error 0.3122
Total Number of Instances 383
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.000 0.000 ? 0.000 ? ? 0.977 0.328 1
0.000 0.000 ? 0.000 ? ? 0.833 0.086 2
0.127 0.027 0.438 0.127 0.197 0.175 0.872 0.499 3
0.851 0.593 0.596 0.851 0.701 0.288 0.717 0.706 4
0.542 0.095 0.722 0.542 0.619 0.489 0.851 0.748 5
Weighted Avg. 0.619 0.334 ? 0.619 ? ? 0.788 0.671
=== Confusion Matrix ===
a b c d e <-- classified as
0 0 3 3 0 | a = 1
0 0 1 7 0 | b = 2
0 0 7 47 1 | c = 3
0 0 5 165 24 | d = 4
0 0 0 55 65 | e = 5

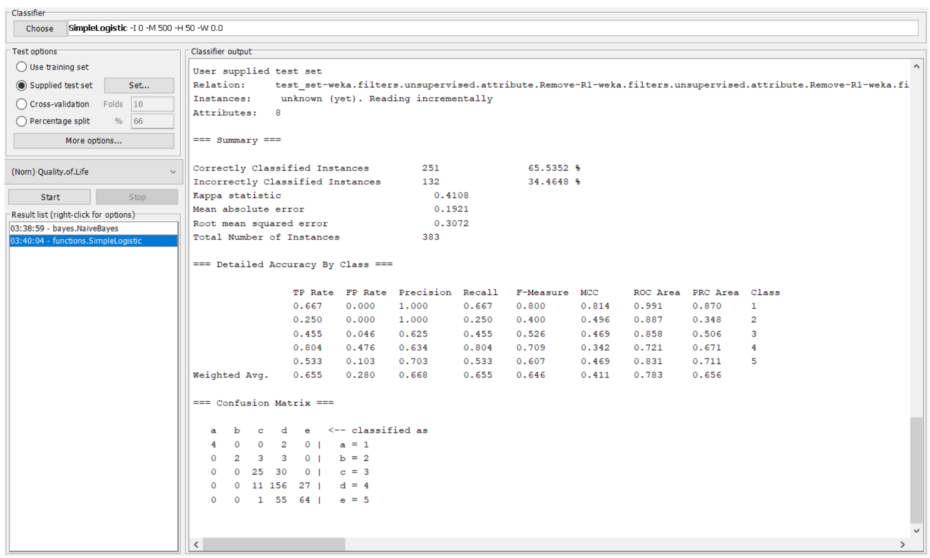
- Reduced test dataset3: WrapperSubset



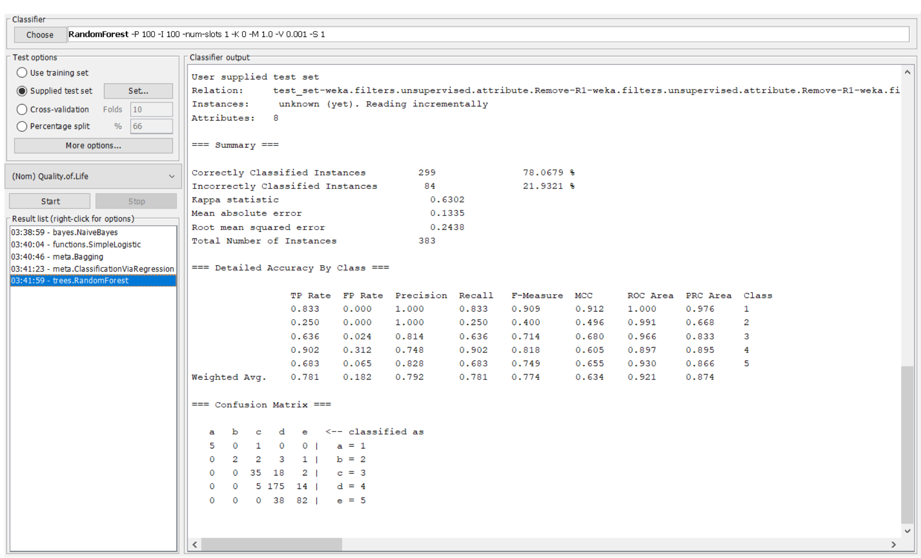





```
User supplied test set
Relation:     test_set-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1-weka.fi
Instances:     unknown (yet). Reading incrementally
Attributes:   8

=== Summary ===

Correctly Classified Instances         240               62.6632 %
Incorrectly Classified Instances       143               37.3368 %
Kappa statistic                          0.3426
Mean absolute error                      0.1981
Root mean squared error                  0.3113
Total Number of Instances              383

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.333    0.003    0.667      0.333    0.444      0.466    0.996     0.807     1
                 0.000    0.000    ?          0.000    ?          ?        0.790     0.283     2
                 0.364    0.064    0.488      0.364    0.417      0.340    0.851     0.515     3
                 0.861    0.577    0.605      0.861    0.711      0.317    0.705     0.656     4
                 0.425    0.046    0.810      0.425    0.557      0.475    0.829     0.696     5
Weighted Avg.    0.627    0.316    ?          0.627    ?          ?        0.771     0.643

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
   2   0   1   3   0 |   a = 1
   1   0   3   4   0 |   b = 2
   0   0  20  35   0 |   c = 3
   0   0  15 167  12 |   d = 4
   0   0   2  67  51 |   e = 5
```

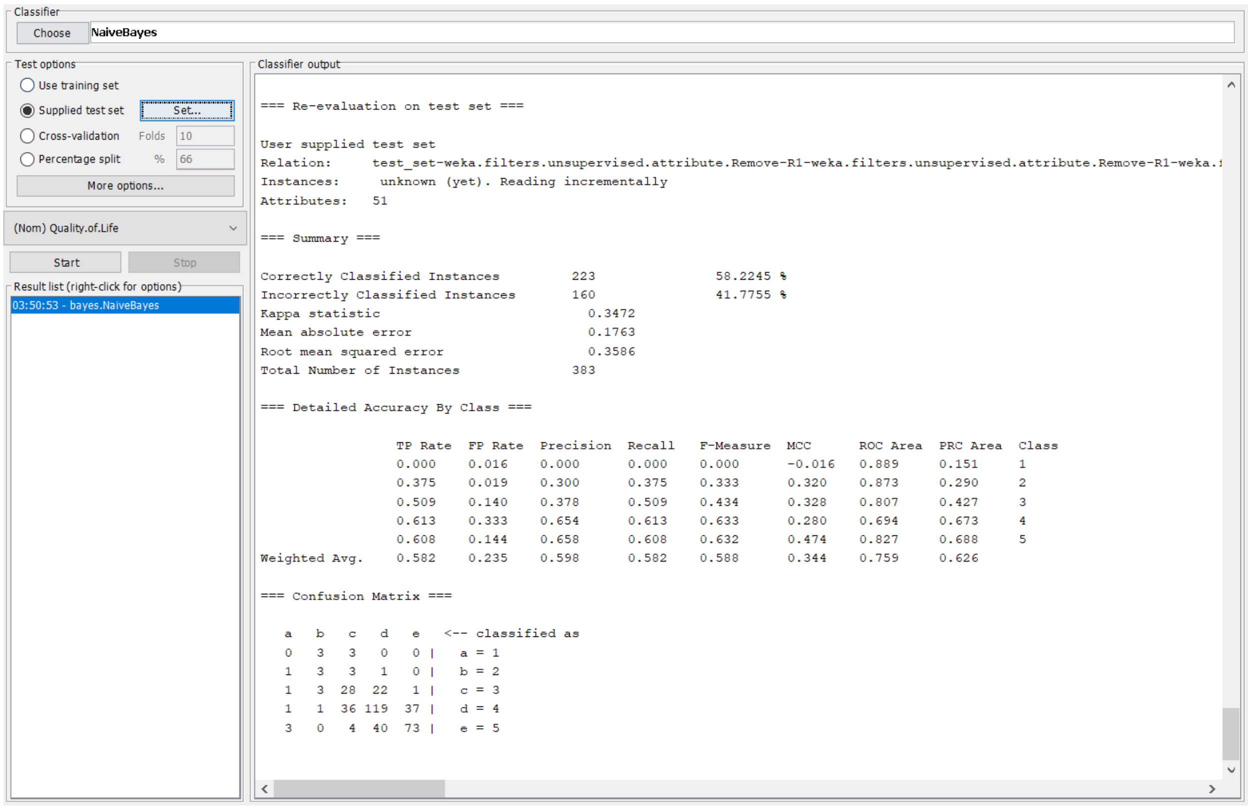


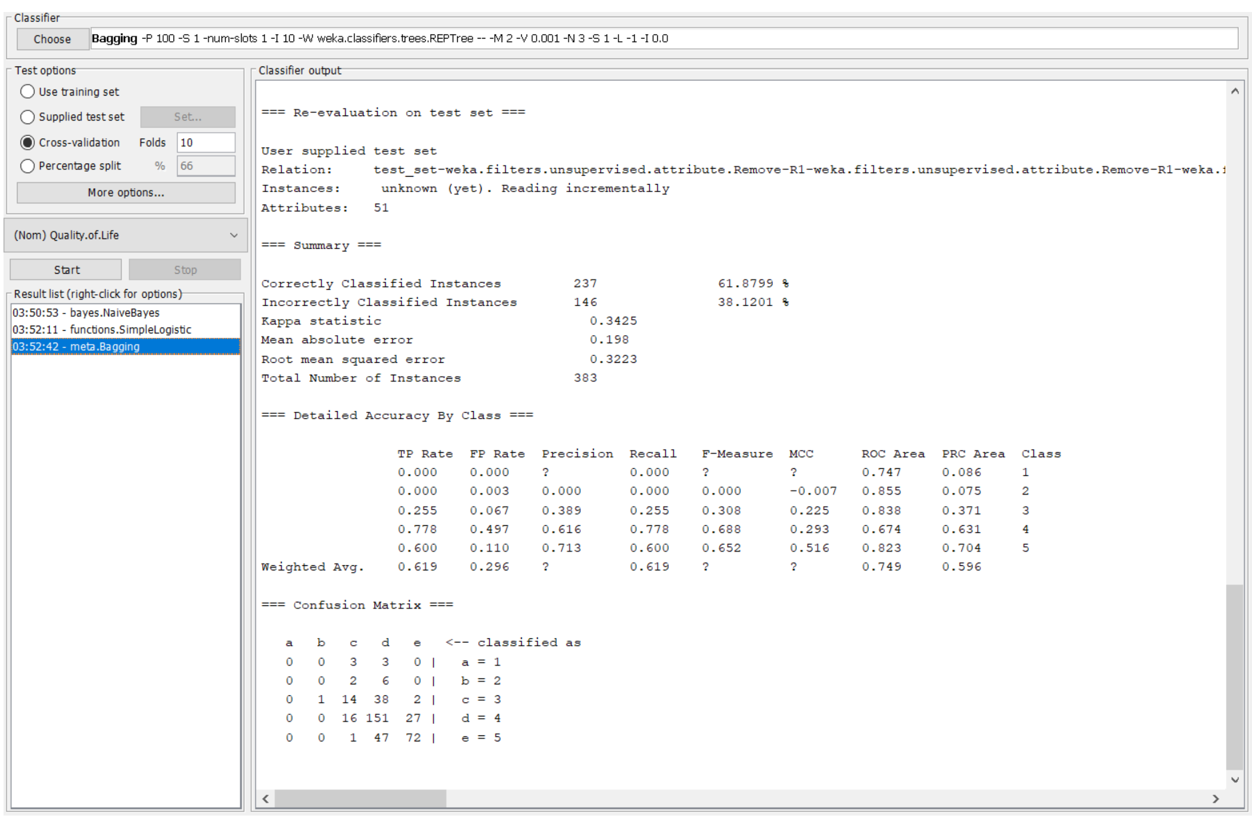

```
User supplied test set
Relation:     test_set-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1-weka.fi
Instances:     unknown (yet). Reading incrementally
Attributes:   8

=== Summary ===

Correctly Classified Instances         251               65.5352 %
Incorrectly Classified Instances       132               34.4648 %
Kappa statistic                          0.4108
Mean absolute error                      0.1921
Root mean squared error                  0.3072
Total Number of Instances              383

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.667    0.000    1.000      0.667    0.800      0.814    0.991     0.870     1
                 0.250    0.000    1.000      0.250    0.400      0.496    0.887     0.348     2
                 0.455    0.046    0.625      0.455    0.526      0.469    0.858     0.506     3
                 0.804    0.476    0.634      0.804    0.709      0.342    0.721     0.671     4
                 0.533    0.103    0.703      0.533    0.607      0.469    0.831     0.711     5
Weighted Avg.    0.655    0.280    0.668      0.655    0.646      0.411    0.783     0.656

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
   4   0   0   2   0 |   a = 1
   0   2   3   3   0 |   b = 2
   0   0  25  30   0 |   c = 3
   0   0  11 156  27 |   d = 4
   0   0   1  55  64 |   e = 5
```

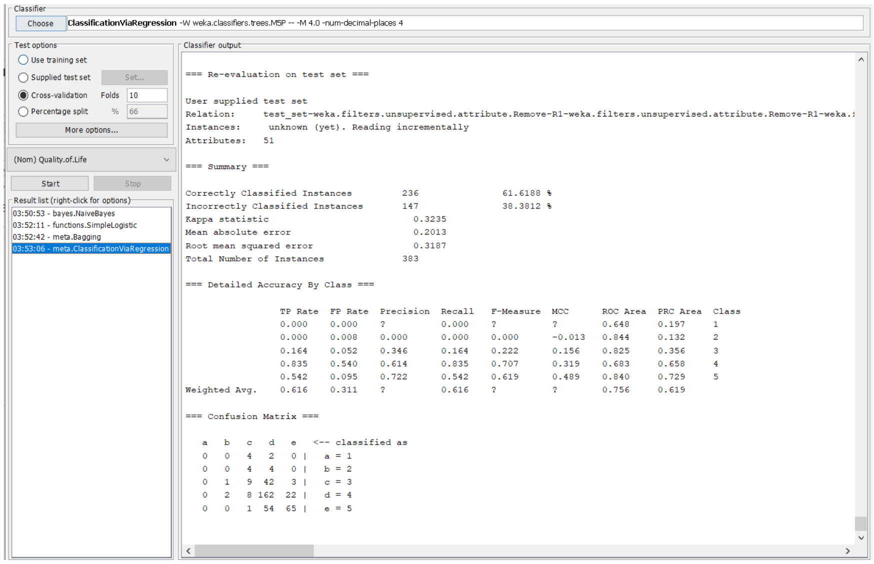



```
User supplied test set
Relation:     test_set-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1-weka.fi
Instances:     unknown (yet). Reading incrementally
Attributes:   8

=== Summary ===

Correctly Classified Instances         299               78.0679 %
Incorrectly Classified Instances        84               21.9321 %
Kappa statistic                          0.6302
Mean absolute error                      0.1335
Root mean squared error                  0.2438
Total Number of Instances              383

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.833    0.000    1.000      0.833    0.909      0.912    1.000     0.976     1
                 0.250    0.000    1.000      0.250    0.400      0.496    0.991     0.668     2
                 0.636    0.024    0.814      0.636    0.714      0.680    0.966     0.833     3
                 0.902    0.312    0.748      0.902    0.818      0.605    0.897     0.895     4
                 0.683    0.065    0.828      0.683    0.749      0.655    0.930     0.866     5
Weighted Avg.    0.781    0.182    0.792      0.781    0.774      0.634    0.921     0.874

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
   5   0   1   0   0 |   a = 1
   0   2   2   3   1 |   b = 2
   0   0  35  18   2 |   c = 3
   0   0   5 175  14 |   d = 4
   0   0   0  38  82 |   e = 5
```

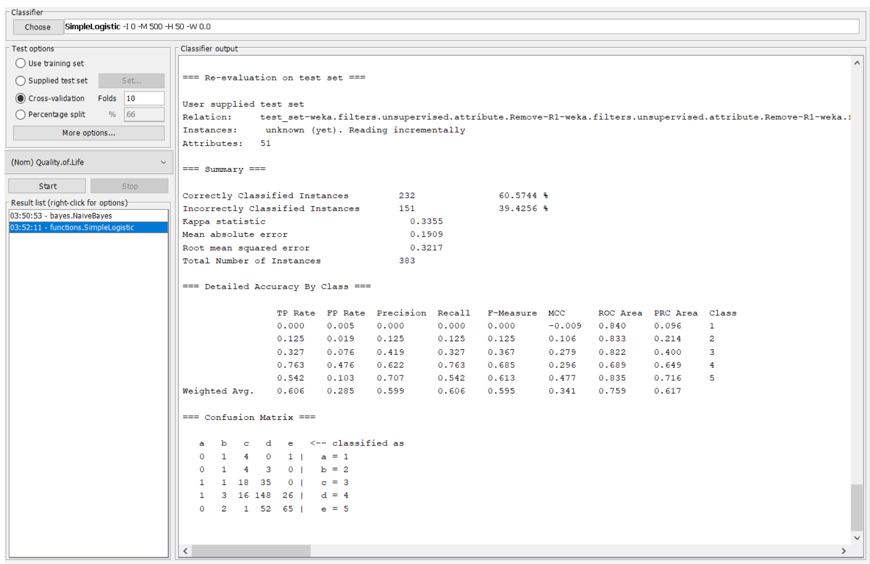
- Reduced test dataset4: OneR



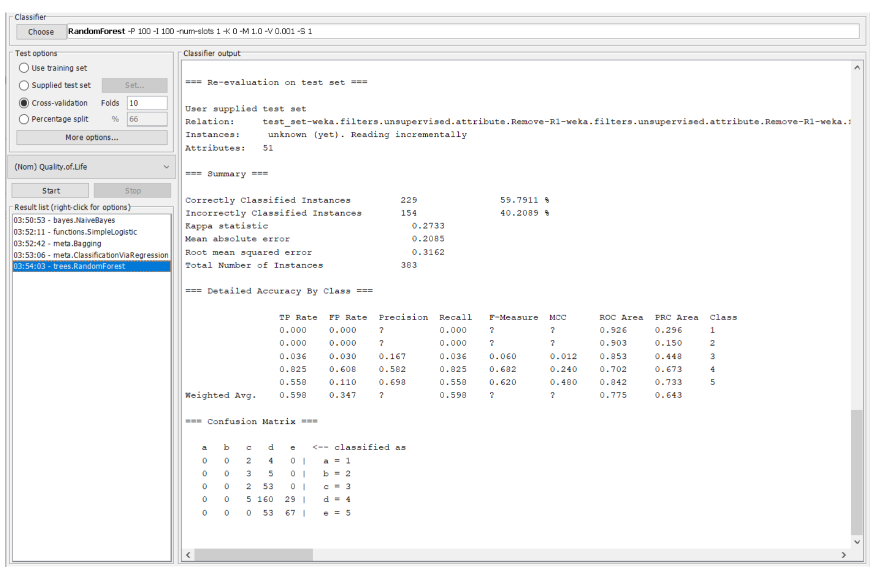




Classifier
Choose
ClassificationViaRegression -W weka.classifiers.trees.M5P -- -M 4.0 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Quality.of.Life
Start
Stop
Result list (right-click for options)
03:50:53 - bayes.NaiveBayes
03:52:11 - functions.SimpleLogistic
03:52:42 - meta.Bagging
03:53:06 - meta.ClassificationViaRegression
Classifier output
=== Re-evaluation on test set ===
User supplied test set
Relation: test_set-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1-weka.
Instances: unknown (yet). Reading incrementally
Attributes: 51
=== Summary ===
Correctly Classified Instances 236 61.6188 %
Incorrectly Classified Instances 147 38.3812 %
Kappa statistic 0.3235
Mean absolute error 0.2013
Root mean squared error 0.3187
Total Number of Instances 383
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.000 0.000 ? 0.000 ? ? 0.648 0.197 1
0.000 0.008 0.000 0.000 0.000 -0.013 0.844 0.132 2
0.164 0.052 0.346 0.164 0.222 0.156 0.825 0.356 3
0.835 0.540 0.614 0.835 0.707 0.319 0.683 0.658 4
0.542 0.095 0.722 0.542 0.619 0.489 0.840 0.729 5
Weighted Avg. 0.616 0.311 ? 0.616 ? ? 0.756 0.619
=== Confusion Matrix ===
a b c d e <-- classified as
0 0 4 2 0 | a = 1
0 0 4 4 0 | b = 2
0 1 9 42 3 | c = 3
0 2 8 162 22 | d = 4
0 0 1 54 65 | e = 5

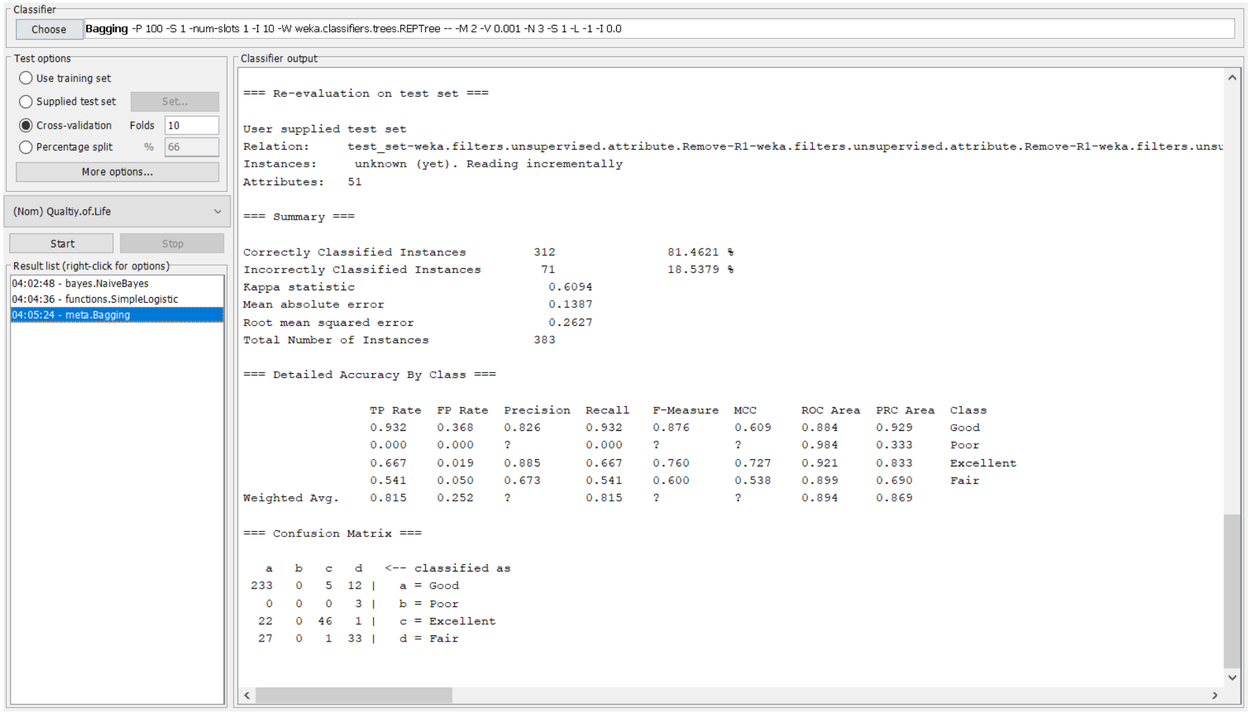


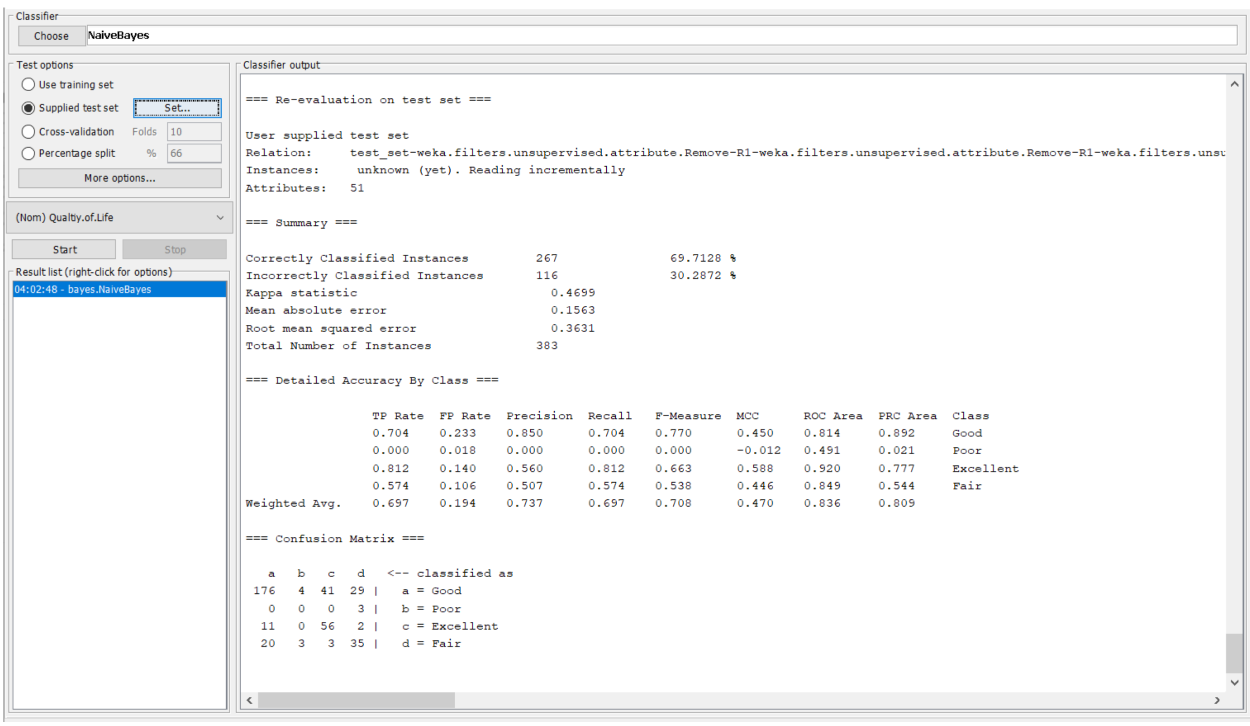
Classifier
Choose
SimpleLogistic -I 0 -M 500 -H 50 -W 0.0
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Quality.of.Life
Start
Stop
Result list (right-click for options)
03:50:53 - bayes.NaiveBayes
03:52:11 - functions.SimpleLogistic
Classifier output
=== Re-evaluation on test set ===
User supplied test set
Relation: test_set-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1-weka.
Instances: unknown (yet). Reading incrementally
Attributes: 51
=== Summary ===
Correctly Classified Instances 232 60.5744 %
Incorrectly Classified Instances 151 39.4256 %
Kappa statistic 0.3355
Mean absolute error 0.1909
Root mean squared error 0.3217
Total Number of Instances 383
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.000 0.005 0.000 0.000 0.000 -0.009 0.840 0.096 1
0.125 0.019 0.125 0.125 0.125 0.106 0.833 0.214 2
0.327 0.076 0.419 0.327 0.367 0.279 0.822 0.400 3
0.763 0.476 0.622 0.763 0.685 0.296 0.689 0.649 4
0.542 0.103 0.707 0.542 0.613 0.477 0.835 0.716 5
Weighted Avg. 0.606 0.285 0.599 0.606 0.595 0.341 0.759 0.617
=== Confusion Matrix ===
a b c d e <-- classified as
0 1 4 0 1 | a = 1
0 1 4 3 0 | b = 2
1 1 18 35 0 | c = 3
1 3 16 148 26 | d = 4
0 2 1 52 65 | e = 5

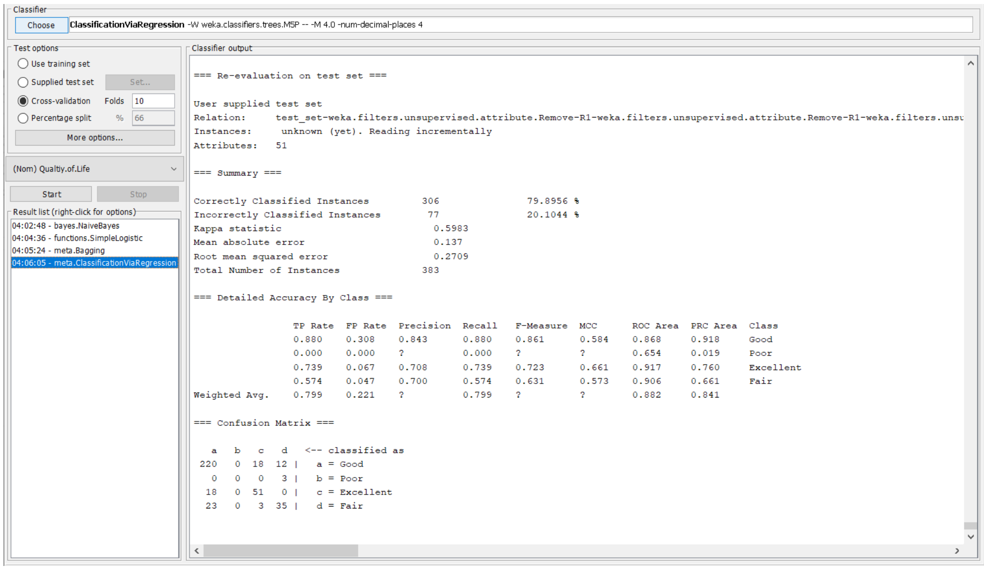


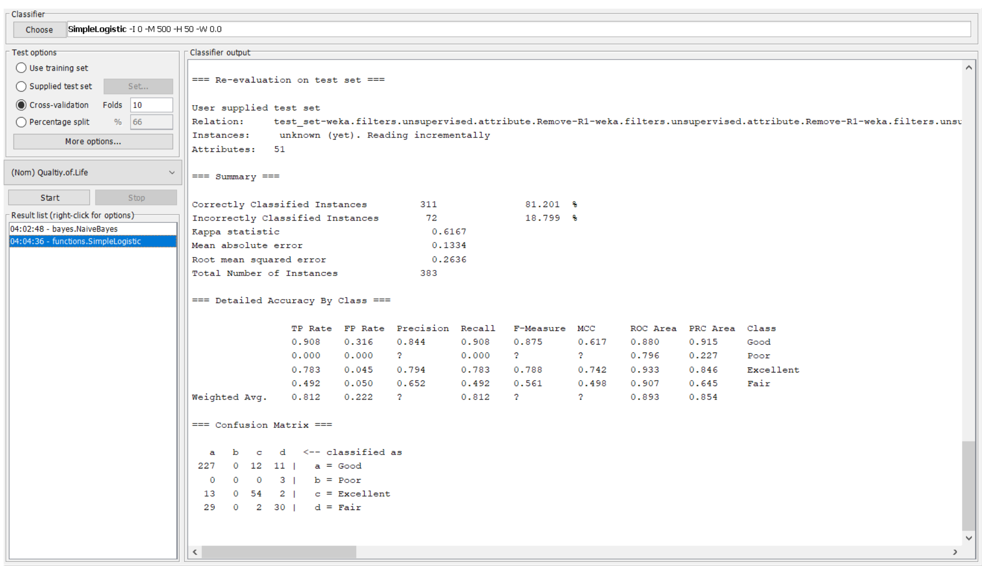
Classifier
Choose
RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Quality.of.Life
Start
Stop
Result list (right-click for options)
03:50:53 - bayes.NaiveBayes
03:52:11 - functions.SimpleLogistic
03:52:42 - meta.Bagging
03:53:06 - meta.ClassificationViaRegression
03:54:03 - trees.RandomForest
Classifier output
=== Re-evaluation on test set ===
User supplied test set
Relation: test_set-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1-weka.
Instances: unknown (yet). Reading incrementally
Attributes: 51
=== Summary ===
Correctly Classified Instances 229 59.7911 %
Incorrectly Classified Instances 154 40.2089 %
Kappa statistic 0.2733
Mean absolute error 0.2085
Root mean squared error 0.3162
Total Number of Instances 383
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.000 0.000 ? 0.000 ? ? 0.926 0.296 1
0.000 0.000 ? 0.000 ? ? 0.903 0.150 2
0.036 0.030 0.167 0.036 0.060 0.012 0.853 0.448 3
0.825 0.608 0.582 0.825 0.682 0.240 0.702 0.673 4
0.558 0.110 0.698 0.558 0.620 0.480 0.842 0.733 5
Weighted Avg. 0.598 0.347 ? 0.598 ? ? 0.775 0.643
=== Confusion Matrix ===
a b c d e <-- classified as
0 0 2 4 0 | a = 1
0 0 3 5 0 | b = 2
0 0 2 53 0 | c = 3
0 0 5 160 29 | d = 4
0 0 0 53 67 | e = 5

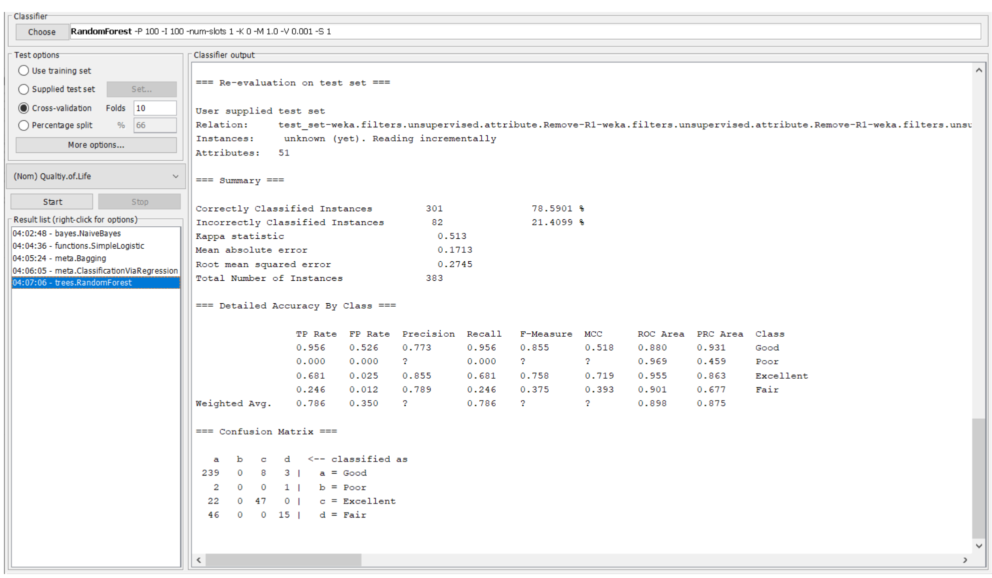
- Reduced test dataset5: InfoGain





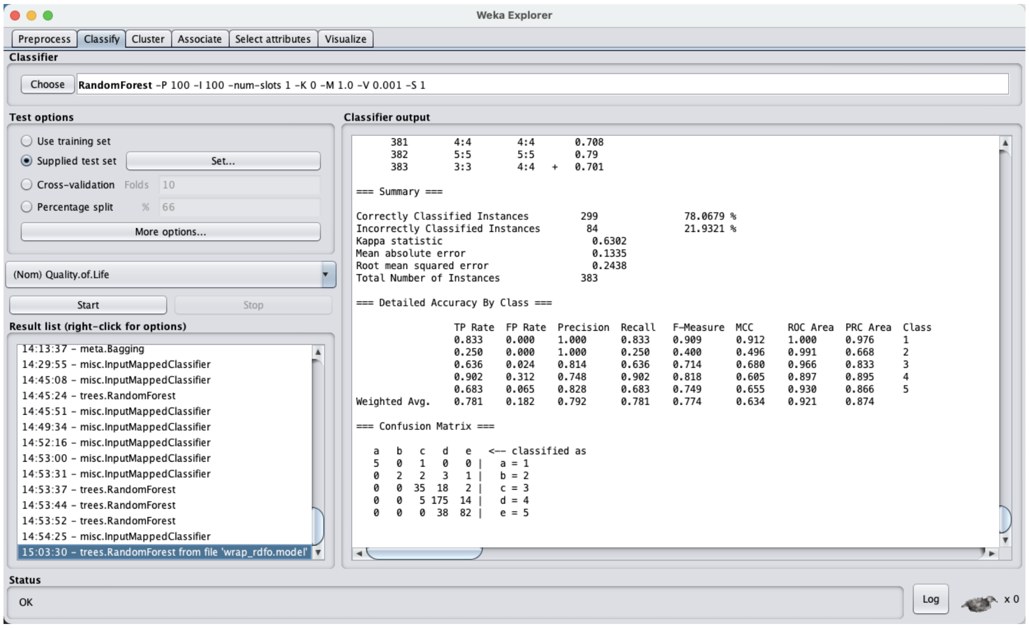

Classifier
Choose
ClassificationViaRegression -W weka.classifiers.trees.M5P -- -M 4.0 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Qualtiy.of.Life
Start
Stop
Result list (right-click for options)
04:02:48 - bayes.NaiveBayes
04:04:36 - functions.SimpleLogistic
04:05:24 - meta.Bagging
04:06:05 - meta.ClassificationViaRegression
Classifier output
=== Re-evaluation on test set ===
User supplied test set
Relation: test_set-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsu
Instances: unknown (yet). Reading incrementally
Attributes: 51
=== Summary ===
Correctly Classified Instances 306 79.8956 %
Incorrectly Classified Instances 77 20.1044 %
Kappa statistic 0.5983
Mean absolute error 0.137
Root mean squared error 0.2709
Total Number of Instances 383
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.880 0.308 0.843 0.880 0.861 0.584 0.868 0.918 Good
0.000 0.000 ? 0.000 ? ? 0.654 0.019 Poor
0.739 0.067 0.708 0.739 0.723 0.661 0.917 0.760 Excellent
0.574 0.047 0.700 0.574 0.631 0.573 0.906 0.661 Fair
Weighted Avg. 0.799 0.221 ? 0.799 ? ? 0.882 0.841
=== Confusion Matrix ===
a b c d <-- classified as
220 0 18 12 | a = Good
0 0 0 3 | b = Poor
18 0 51 0 | c = Excellent
23 0 3 35 | d = Fair


Classifier
Choose
SimpleLogistic -I 0 -M 500 -H 50 -W 0.0
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Qualtiy.of.Life
Start
Stop
Result list (right-click for options)
04:02:48 - bayes.NaiveBayes
04:04:36 - functions.SimpleLogistic
Classifier output
=== Re-evaluation on test set ===
User supplied test set
Relation: test_set-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsu
Instances: unknown (yet). Reading incrementally
Attributes: 51
=== Summary ===
Correctly Classified Instances 311 81.201 %
Incorrectly Classified Instances 72 18.799 %
Kappa statistic 0.6167
Mean absolute error 0.1334
Root mean squared error 0.2636
Total Number of Instances 383
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.908 0.316 0.844 0.908 0.875 0.617 0.880 0.915 Good
0.000 0.000 ? 0.000 ? ? 0.796 0.227 Poor
0.783 0.045 0.794 0.783 0.788 0.742 0.933 0.846 Excellent
0.492 0.050 0.652 0.492 0.561 0.498 0.907 0.645 Fair
Weighted Avg. 0.812 0.222 ? 0.812 ? ? 0.893 0.854
=== Confusion Matrix ===
a b c d <-- classified as
227 0 12 11 | a = Good
0 0 0 3 | b = Poor
13 0 54 2 | c = Excellent
29 0 2 30 | d = Fair

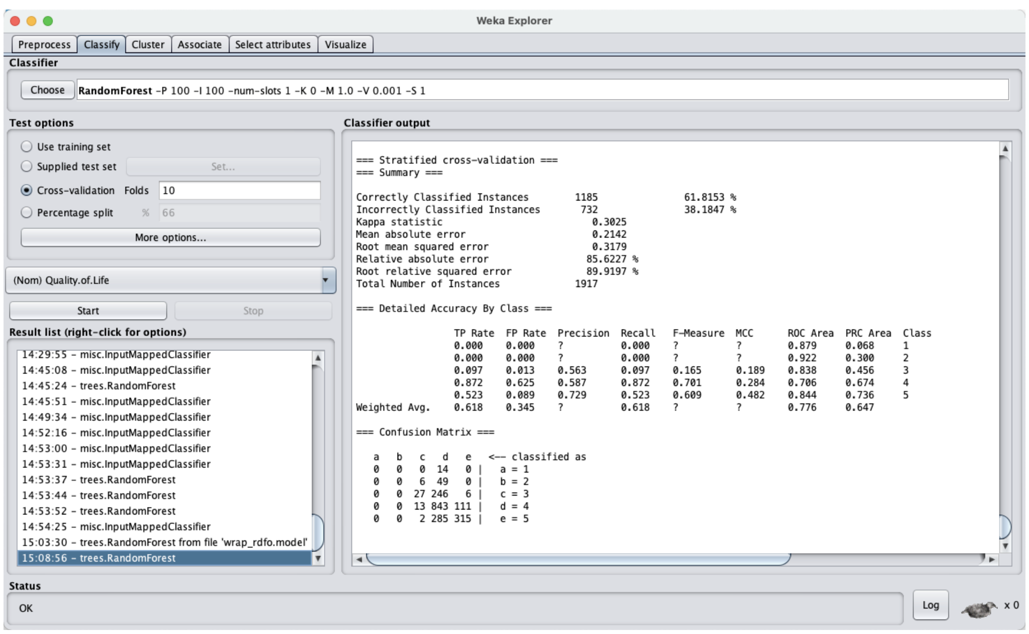

Classifier
Choose
RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Qualtiy.of.Life
Start
Stop
Result list (right-click for options)
04:02:48 - bayes.NaiveBayes
04:04:36 - functions.SimpleLogistic
04:05:24 - meta.Bagging
04:06:05 - meta.ClassificationViaRegression
04:07:06 - trees.RandomForest
Classifier output
=== Re-evaluation on test set ===
User supplied test set
Relation: test_set-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsu
Instances: unknown (yet). Reading incrementally
Attributes: 51
=== Summary ===
Correctly Classified Instances 301 78.5901 %
Incorrectly Classified Instances 82 21.4099 %
Kappa statistic 0.513
Mean absolute error 0.1713
Root mean squared error 0.2745
Total Number of Instances 383
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.956 0.526 0.773 0.956 0.855 0.518 0.880 0.931 Good
0.000 0.000 ? 0.000 ? ? 0.969 0.459 Poor
0.681 0.025 0.855 0.681 0.758 0.719 0.955 0.863 Excellent
0.246 0.012 0.789 0.246 0.375 0.393 0.901 0.677 Fair
Weighted Avg. 0.786 0.350 ? 0.786 ? ? 0.898 0.875
=== Confusion Matrix ===
a b c d <-- classified as
239 0 8 3 | a = Good
2 0 0 1 | b = Poor
22 0 47 0 | c = Excellent
46 0 0 15 | d = Fair

8.2 Comparison of the performances of the best model with the model that was built using the same classification algorithm from the dataset with all attributes.

The best model (using WrapperSubset attribute selection method with 8 attributes selected and Random Forest classification algorithm) performed (*Accuracy = 78.0679%, ROC Area= 0.921, PRC Area = 0.874)* better than the model with all attributes (*Accuracy = 61.8153%, ROC Area= 0.776, PRC Area = 0.647)*. The TP rate of the best model is 16.2526% higher than another one.



Best model with 8 attributes



The model with all attributes (121 attributes)

9. Discussion and conclusion, including what you learned from this project.

9.1 Methods to increase performance

To increase the performance of our classifier models, several methods have been proven to be useful in our project:

- Removing inconsistent data and any outliers
- Reducing dataset
- Encoding the class attribute
  Encoding the class attribute “Quality of life” from “1-10” to “1-5” roughly increased 20% of the model accuracy.

9.2 Findings of the testing results

- Generally, the datasets with attributes selected using WrapperSubset method generated better performance, 5%-20% higher accuracy than other four attribute selection methods (CfsSubset, Correlation, OneR, InfoGain);
- Random Forest algorithm generates the best TP rate for Correlation and WrapperSubset test sets.
- Compared to other attributes selection methods, WrapperSubset selected the least number of attributes (only 8 attributes), so we conclude that less attributes contributes to better performance.
- Apart from the Random Forest classification algorithm, the bagging algorithm performs better than most other classification models. For test sets of CfsSubset and OneR, bagging generates the highest True Positive rates.

9.3 Suggestions for the future steps

- Based on the selected attributes using WrapperSubset, attributes including Present mental health, Satisfaction with life, Family pride, Library internet access, Citizenship class, Contact city official or not, City election, we found that people caring more about the city life and participating in local political activities seems to be important factors in quality of life. Contrary to what we believed, mental health and state were a bigger factor in quality of life than physical health or condition. These findings should be taken into account for future study.

- For the dataset(after preprocessing) using Correlation attribute selection method, we found that attributes like Satisfaction.With.Housing(ρ = 0.1536), Satisfied.With.Life.2(ρ = 0.1416), Achieving.Ends.Meet(ρ = 0.139), Satisfied.With.Life.1(ρ = 0.1358), Duration.of.Residency(ρ = 0.1349) have weakly relationship with the class attribute “Quality of life.” To better understand what factors influence Asian Americans’ quality of life, further research can be taken on exploring its correlation with peoples' satisfaction with housing, achieving ends meet or not, and duration of the residency.

** We do all the work of this project together while Philip contributes a little bit more on R and the first half of this report, and Min contributes a little bit more Weka and the second half of the report.*