AI Programming Assignment 2,

Chetan There, 1001248929.

**Introduction**:

As part of this assignment I am using machine learning algorithms namely J48 (decision trees), NaiveBayes, and JRip to classify given data sets and comparing their results with and without preprocessing using Weka. I am performing preprocessing techniques like feature selection (to use only relevant attributes) and replacing missing values to check whether accuracy is getting increase with these techniques.

**Datasets:**

Below are three datasets that has been given and used for experiments.

1. Hepatitis Data : Mostly Boolean or numeric attributes.

Number of Instances: 155

Number of features: 19

Class is binary attribute

Attribute Details:

1.Class: DIE, LIVE   
2. AGE: 10, 20, 30, 40, 50, 60, 70, 80   
3. SEX: male, female   
4. STEROID: no, yes   
5. ANTIVIRALS: no, yes   
6. FATIGUE: no, yes   
7. MALAISE: no, yes   
8. ANOREXIA: no, yes   
9. LIVER BIG: no, yes   
10. LIVER FIRM: no, yes   
11. SPLEEN PALPABLE: no, yes   
12. SPIDERS: no, yes   
13. ASCITES: no, yes   
14. VARICES: no, yes   
15. BILIRUBIN: 0.39, 0.80, 1.20, 2.00, 3.00, 4.00   
16. ALK PHOSPHATE: 33, 80, 120, 160, 200, 250   
17. SGOT: 13, 100, 200, 300, 400, 500,   
18. ALBUMIN: 2.1, 3.0, 3.8, 4.5, 5.0, 6.0   
19. PROTIME: 10, 20, 30, 40, 50, 60, 70, 80, 90   
20. HISTOLOGY: no, yes

1. Wine Quality Data Sets : I have merged red and white wine datasets for my experiment

Number of Instances: 4898

Number of features: 12

Class attribute: Quality is multi class attribute (takes value from 0 to 10)

Attribute Details:

1 - fixed acidity   
2 - volatile acidity   
3 - citric acid   
4 - residual sugar   
5 - chlorides   
6 - free sulfur dioxide   
7 - total sulfur dioxide   
8 - density   
9 - pH   
10 - sulphates   
11 - alcohol   
12 - quality (score between 0 and 10)

1. Bank Marketing Data Set

Number of Instances: 45211

Number of features: 17

Class: “y”, it’s binary class attribute

Attribute Details:

1 - age (numeric)

2 - job : type of job(categorical: admin,bluellar,entrepreneur,housemaid,..etc)

3 - marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown')

4 - education (categorical: basic.4y,basic.6y,basic.9y,high.school,illiterate,..etc)  
5 - default: has credit in default? (categorical: 'no', 'yes', 'unknown')  
6 - housing: has housing loan? (categorical: 'no', 'yes', 'unknown')  
7 - loan: has personal loan? (categorical: 'no', 'yes', 'unknown')  
8 - contact: contact communication type (categorical: 'cellular', 'telephone')   
9 - month: last contact month of year (categorical: 'jan', 'feb', ..., 'nov', 'dec')  
10 – day of week: last contact day of the week (categorical: 'mon', 'tue','wed', 'thu', 'fri')  
11 - duration: last contact duration, in seconds (numeric).   
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)  
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)  
14 - previous: number of contacts performed before this campaign and for this client (numeric)  
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success')

16- balance   
17 - y - has the client subscribed a term deposit? (binary: 'yes', 'no')

**Preprocessing:**

I am using two preprocessing techniques

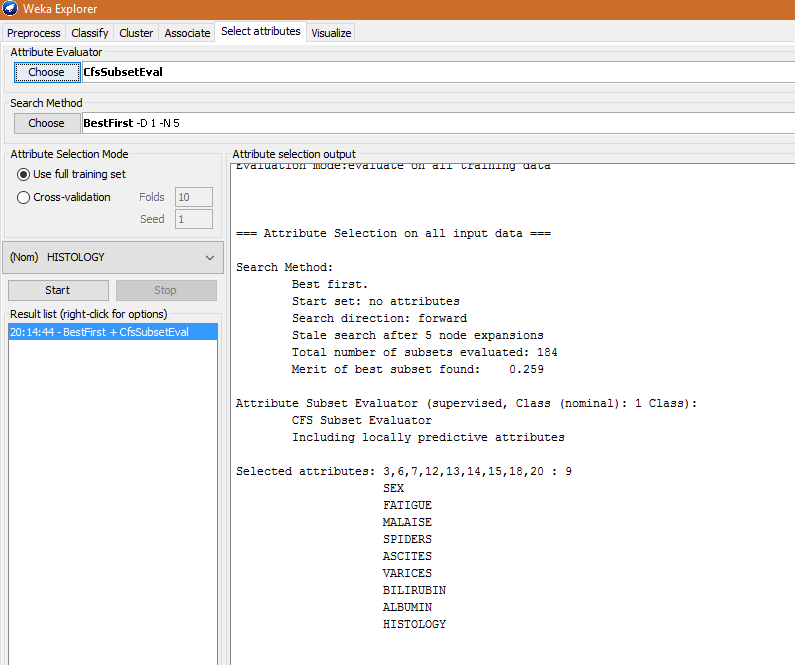
1. Feature selection(CfSsubsetEval option in Weka)
2. Replace missing values technique(Relplace Missing values under preprocess option in Weka)

I am comparing two results for each data set

1. With given data set
2. Using preprocessing methods of Feature Selection and replace missing values

**Feature Selection Process:**

I found out most important attributes using “Select Attributes” in Weka. With “CfsSubsetEval” attribute evaluator .Please find screenshot given below.



Once I get attributes I am using only these attributes followed by running replace missing values and using this new data set for next experiment.

**Results:**

All Are below experiments done by using 10 fold cross validation.

1. **Hepatitis Data Set**
2. Naïve Byes



With Preprocessing

1. J48



With Preprocessing



1. JRip



With Preprocessing



1. **Wine Quality : I am merging red wine and white wine data sets together**
2. Naïve Byes



With Preprocessing



1. J48



With Preprocessing



1. JRip



With Preprocessing



1. **Bank-full dataset**
2. Naïve Bayes



With Preprocessing



1. J48



With Preprocessing



1. JRip



With Preprocessing



**Conclusions:**

Conclusion with respect to these data sets as follows.

Accuracy of all three algorithms is nearly same with respect to given data sets.

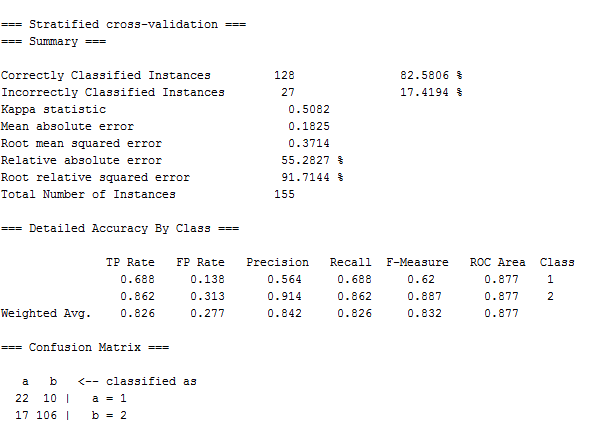
Preprocessing haven’t affect much to accuracy. Slight increase/decrease in accuracy with preprocessed datasets.

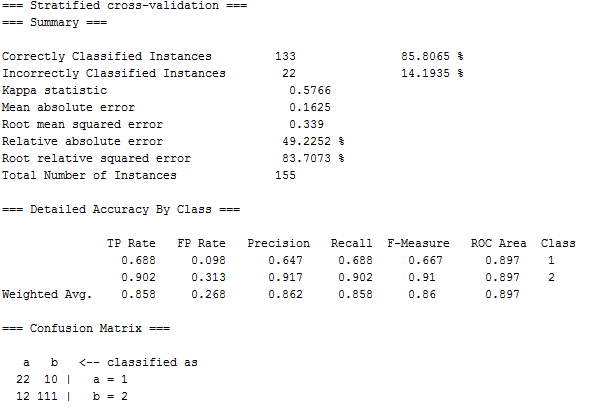
Naïve Byes was fastest followed by J48.Jrip was slowest while building model.

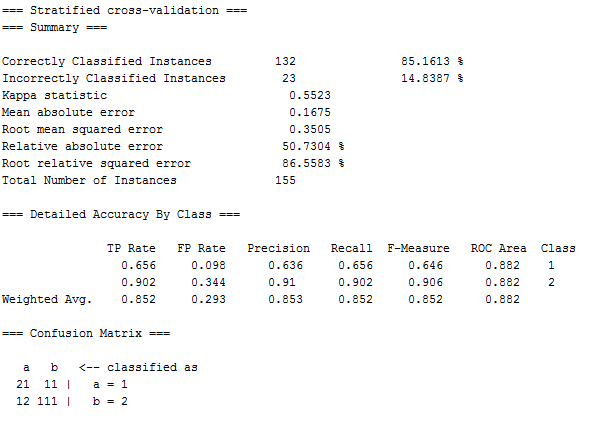
Preprocessing helped a lot for computation (building model) especially with large data set of bank details.

**References:**

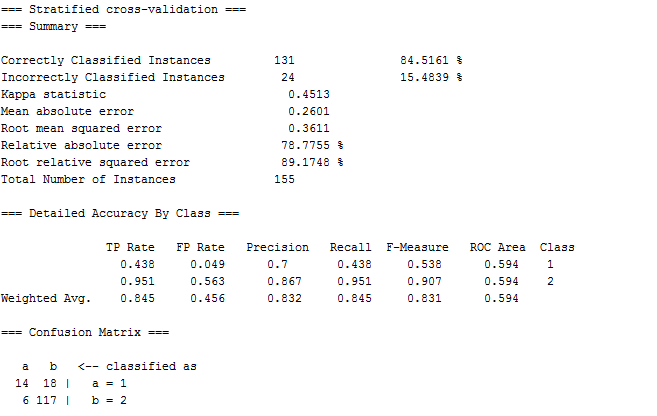
1. Weka : <http://www.cs.waikato.ac.nz/ml/weka/>
2. Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
3. P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.   
   Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.
4. [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

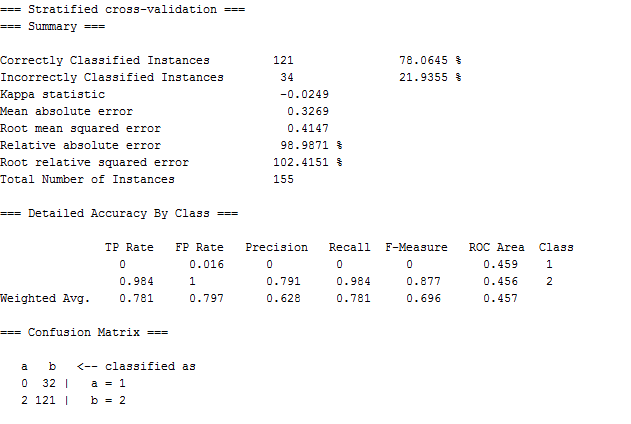


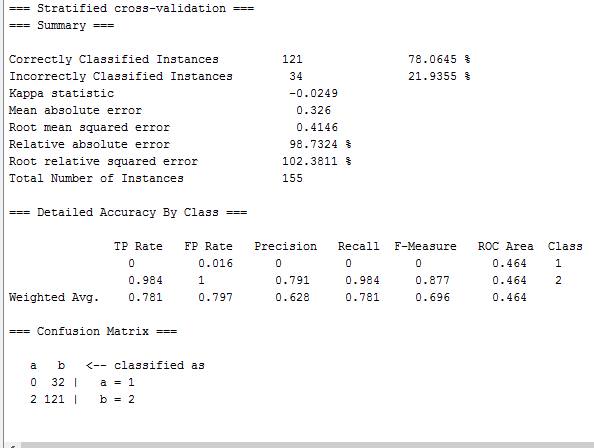




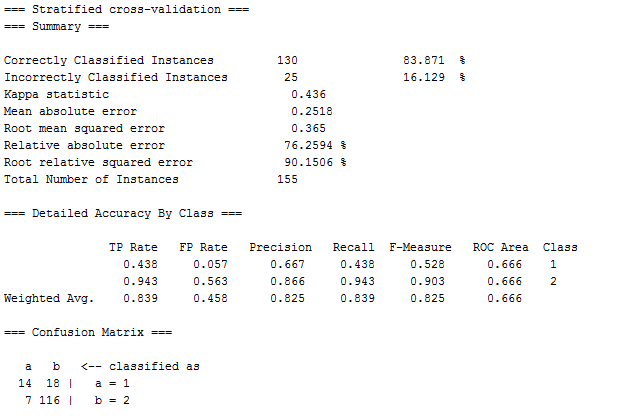
1. J48 (decision trees):

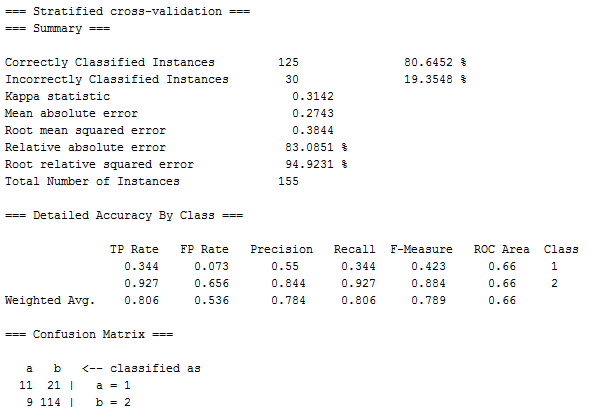


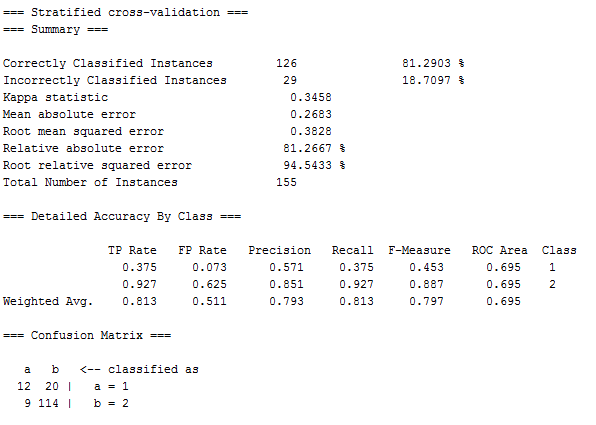




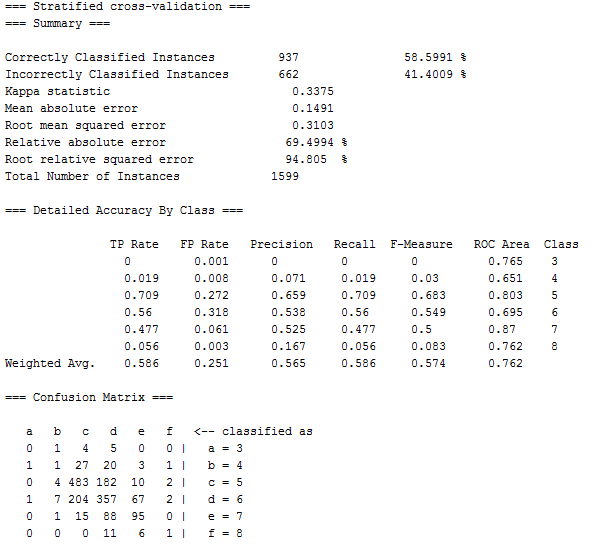
1. Jrip

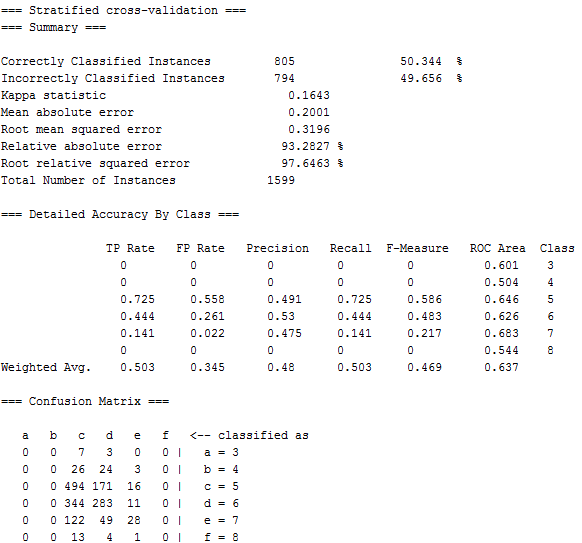




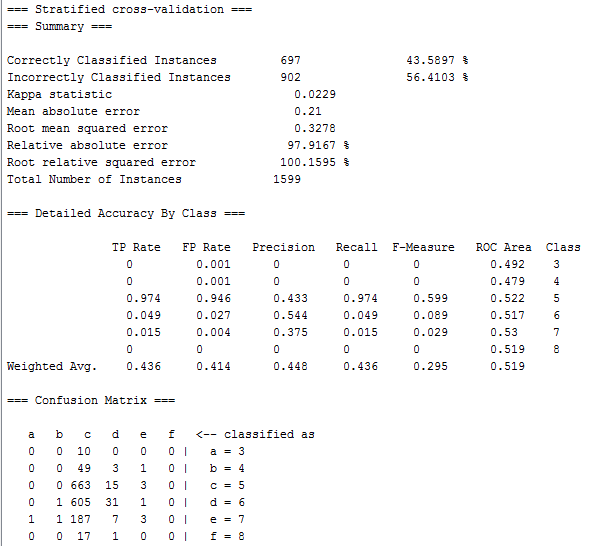


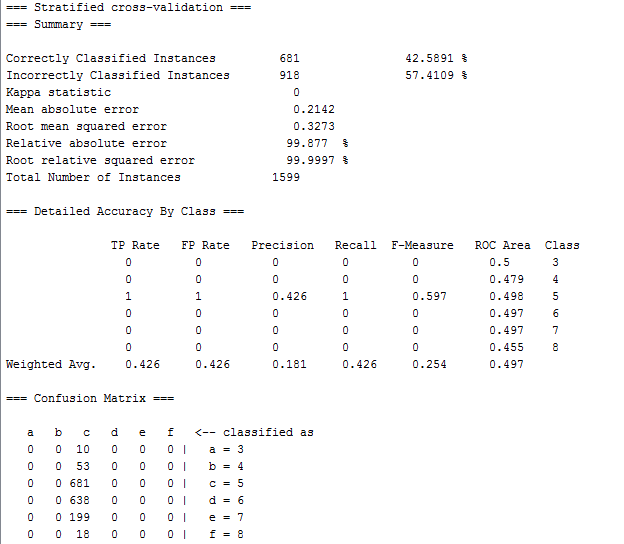
1. Red Wine Dataset
2. Naïve Bayes





1. J48





c.jrip

