GROUP MEMBERS
ANNAN LESLIE - 10671152
LAUREEN ZORMELO - 10678718
GYAMERA COLLINS BREFO -10671724
ARYEE JOSHUA - 10677215
BAFFOE CLEMENT-10680839

UNIVERSITY OF GHANA
SCHOOL OF ENGINEERING SCIENCES
DEPARTMENT OF COMPUTER ENGINEERING
CPEN 403:ARTIFICIAL INTELLIGENCE
LAB REPORT ON WEKA CLASSIFICATION

1 DATASET

The machine.data contains 209 instances and 8 attributes. With one relation of 'CPU'. Predictive attributes vendor, MYCT, MMIN, MMAX, CACH, CHMIN, and CHMAX in order to predict the attribute CLASS in the dataset.

Pictures

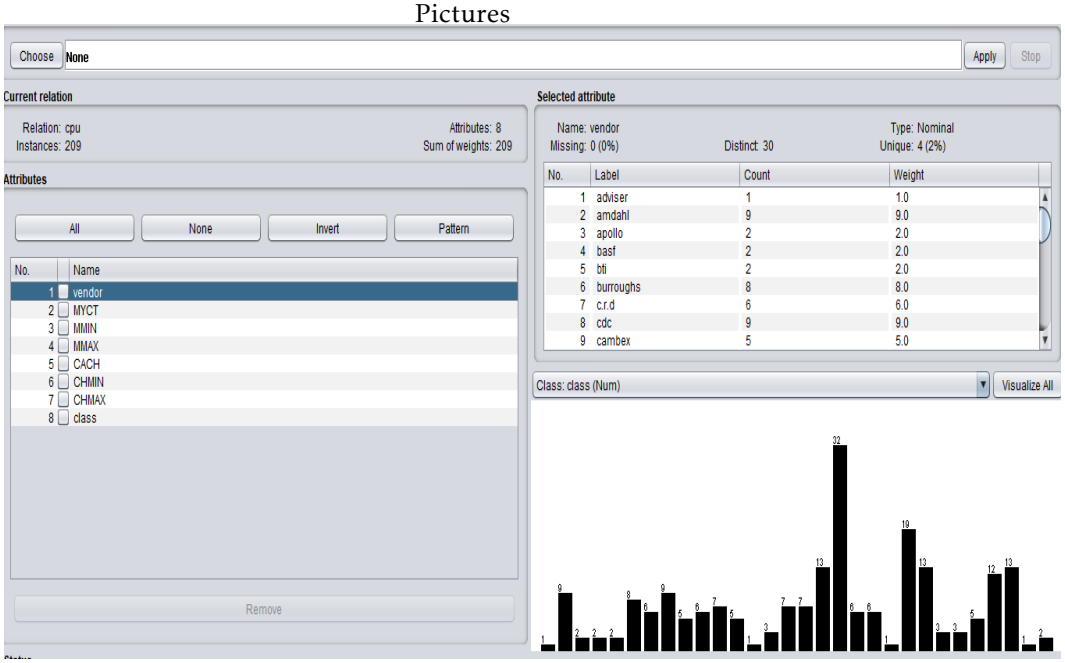


Figure 1: From figure above the CPU performance dataset has been loaded into Weka. the CPU performance dataset from Table 1.5 (page 16) has been loaded into Weka.

2 MODEL CLASSIFICATION

I. M5

The model tree inducer M5 has been chosen as the classifier.This is a supervised model

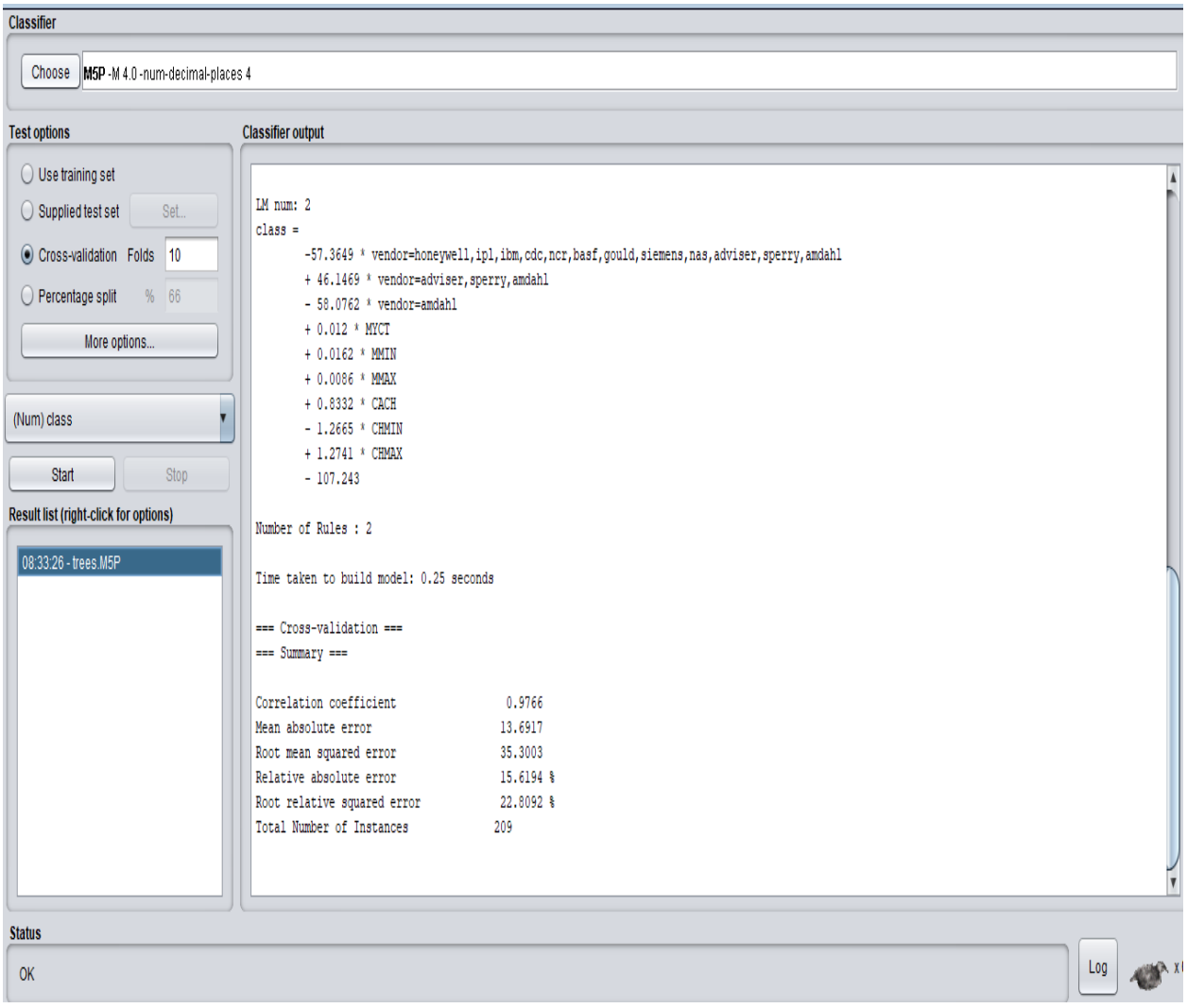


Figure 2: Cross validation with 10 fold.

Divide dataset into 10 parts(folds),hold out each part in turn, averages the results and each data point used once for testing 9 times for training. Also weka invokes the learning algorithm 11 times

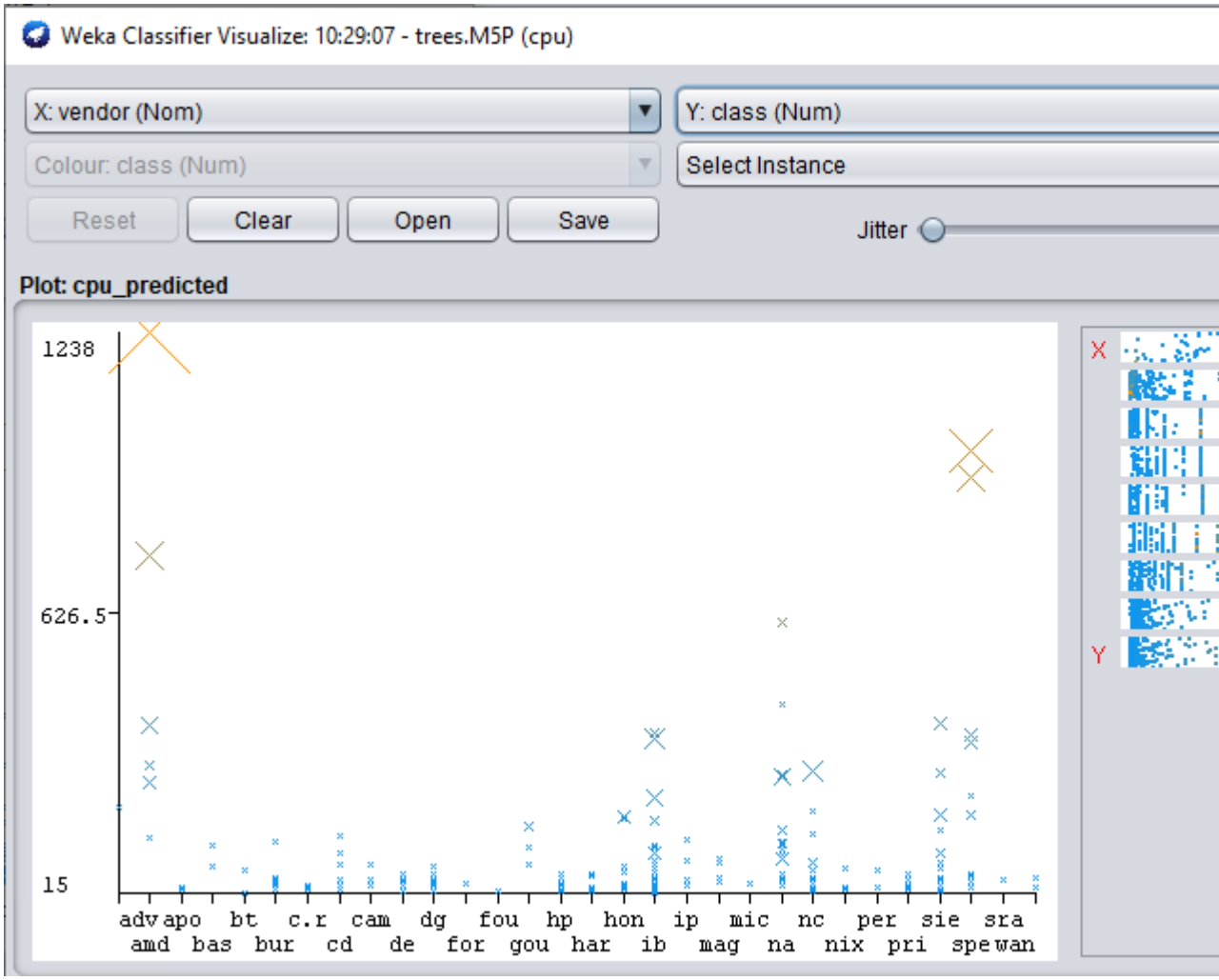
Figure 2 shows the output. t. The pruned model tree is simply a decision stump with a split on the MMAX attribute and two linear models, one for each leaf. Both models involve a nominal attribute, vendor, as well as some numeric ones. The expression vendor = adviser,sperry,amdahl is interpreted as follows: if vendor is either adviser, sperry, or amdahl, then substitute 1; otherwise, substitute 0. The description of the model tree is followed by several figures that measure its performance.

3 RESULTS AND ANALYSIS

```
=== Run information ===

Scheme:       weka.classifiers.trees.M5P -M 4.0
Relation:     cpu
Instances:    209
Attributes:   8
              vendor
              MYCT
              MMIN
              MMAX
              CACH
              CHMIN
              CHMAX
              class
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

M5 pruned model tree:
(using smoothed linear models)

MMAX <= 14000 : LM1 (141/4.178%)
MMAX >  14000 : LM2 (68/50.073%)

LM num: 1
class =
     -2.0542 *
     vendor=honeywell,ipl,ibm,cdc,ncr,basf,gould,siemens,nas,adviser,sperry,amdahl
     + 5.4303 * vendor=adviser,sperry,amdahl
     - 5.7791 * vendor=amdahl
     + 0.0064 * MYCT
     + 0.0016 * MMIN
     + 0.0034 * MMAX
     + 0.5524 * CACH
     + 1.1411 * CHMIN
     + 0.0945 * CHMAX
     + 4.1463

LM num: 2
class =
     -57.3649 *
     vendor=honeywell,ipl,ibm,cdc,ncr,basf,gould,siemens,nas,adviser,sperry,amdahl
     + 46.1469 * vendor=adviser,sperry,amdahl
     - 58.0762 * vendor=amdahl
    + 0.012 * MYCT
    + 0.0162 * MMIN
    + 0.0086 * MMAX
    + 0.8332 * CACH
    - 1.2665 * CHMIN
    + 1.2741 * CHMAX
    - 107.243

Number of Rules : 2

Time taken to build model: 1.37 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9766
Mean absolute error                     13.6917
Root mean squared error                 35.3003
Relative absolute error                 15.6194 %
Root relative squared error             22.8092 %
Total Number of Instances              209
```

Result using the M5 tree classifier.

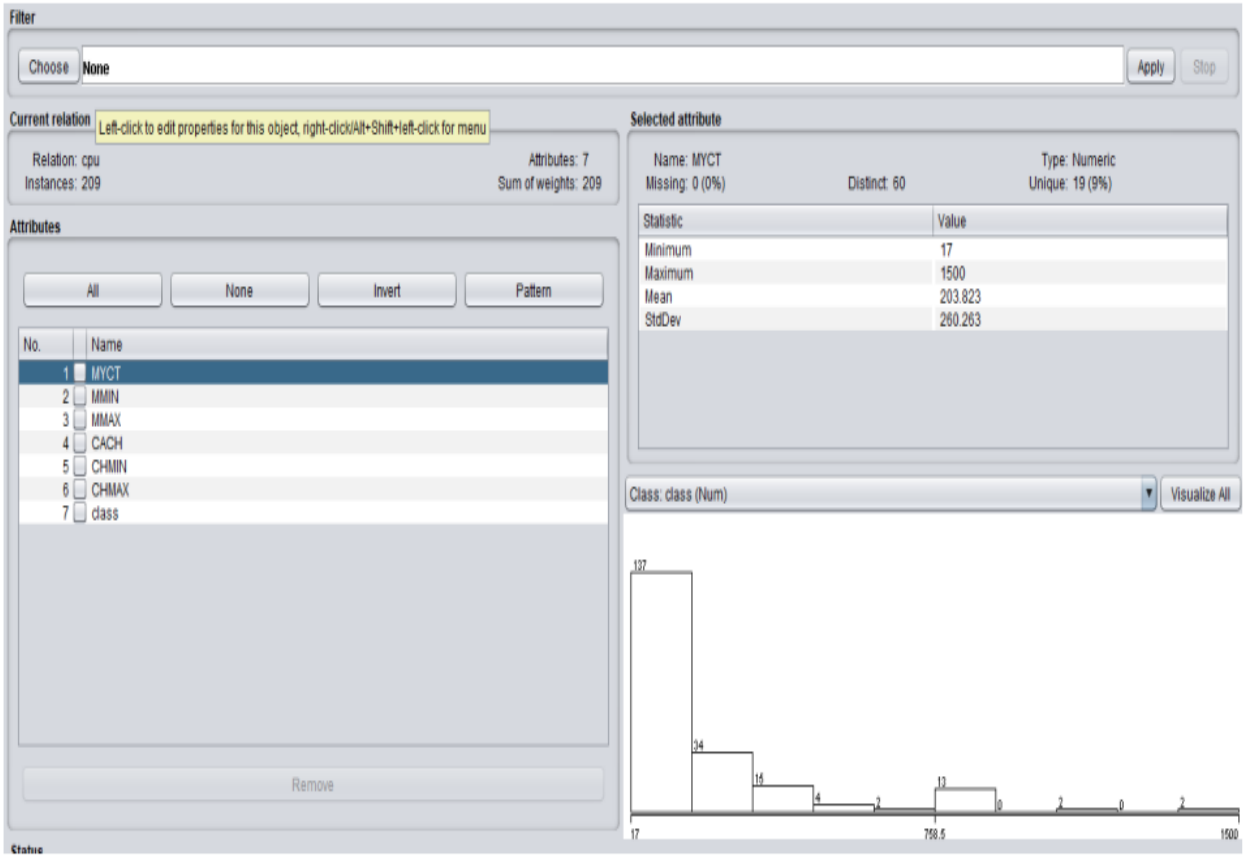
The correlation coefficient ranges from 1 for perfectly correlated results, through 0 when there is no correlation to -1.
To get a feel for their relative performance, let's visualize the errors the scheme using the visualization classifier error.



Vendor attribute has been selected for the X-axis and the predictedERP has been chosen for the Y-axis because this gives a good spread of points

II. Linear Regression

For the supervised linear regression, the attribute 'vendor ' is deleted to make data consistent. This makes the total number of attributes to 7 with the same number of instances as 209.



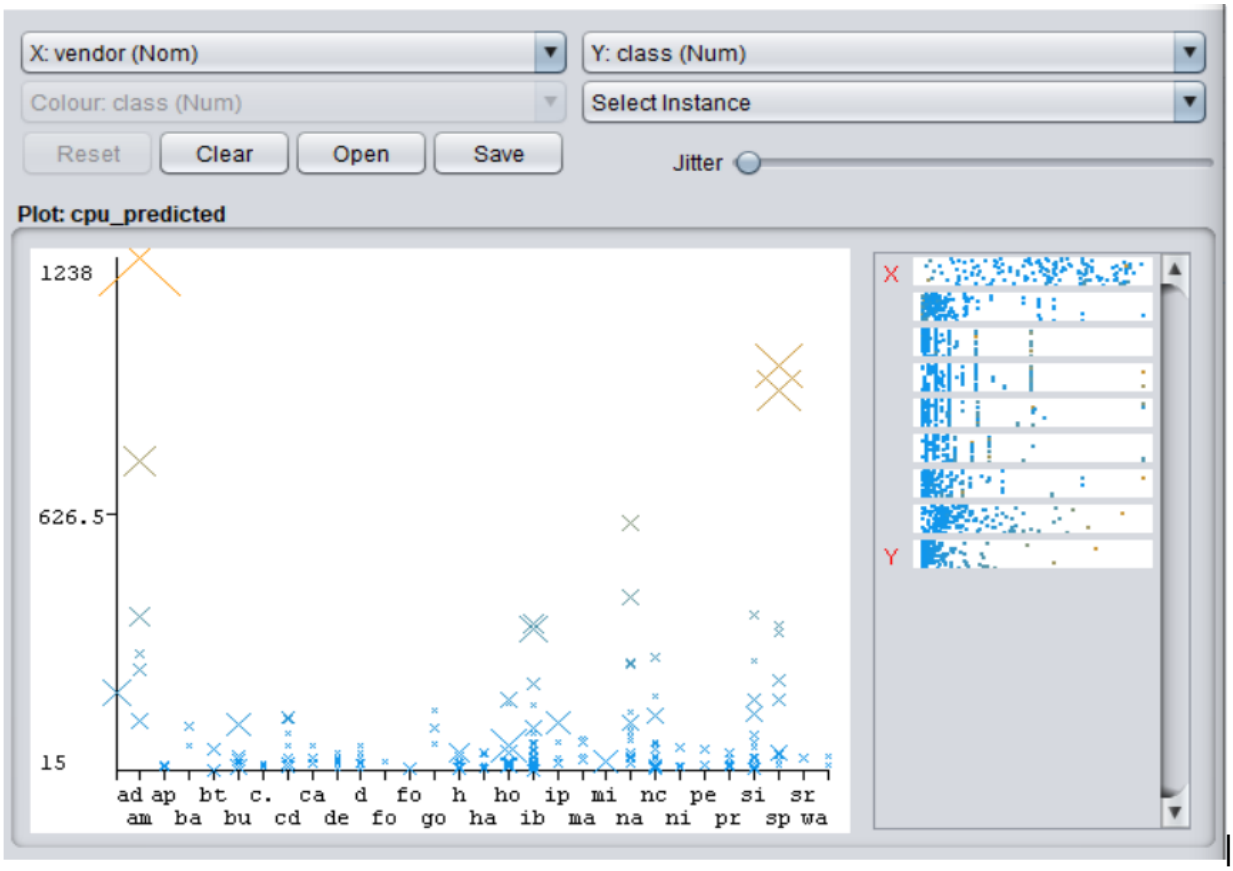
Vendor attribute has been selected for the X-axis and the predictedERP has been chosen for the Y-axis because this gives a good spread of points

4 RESULT

```
=== Run information ===

Scheme:       weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4
Relation:     cpu
Instances:    209
Attributes:   8
              vendor
              MYCT
              MMIN
              MMAX
              CACH
              CHMIN
              CHMAX
              class
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


Linear Regression Model

class =

   -152.7641 * vendor=microdata,prime,formation,harris,dec,wang,perkin-elmer,nixdorf,bti,sratus,dg,burroughs,cambex,magnuson,honeywell,ipl,ibm,cdc,ncr,basf,gould,siemens,nas,adviser,sperry,amdahl +
    141.8644 * vendor=prime,formation,harris,dec,wang,perkin-elmer,nixdorf,bti,sratus,dg,burroughs,cambex,magnuson,honeywell,ipl,ibm,cdc,ncr,basf,gould,siemens,nas,adviser,sperry,amdahl +
    -38.2268 * vendor=burroughs,cambex,magnuson,honeywell,ipl,ibm,cdc,ncr,basf,gould,siemens,nas,adviser,sperry,amdahl +
     39.4748 * vendor=cambex,magnuson,honeywell,ipl,ibm,cdc,ncr,basf,gould,siemens,nas,adviser,sperry,amdahl +
    -39.5986 * vendor=honeywell,ipl,ibm,cdc,ncr,basf,gould,siemens,nas,adviser,sperry,amdahl +
     21.412  * vendor=ipl,ibm,cdc,ncr,basf,gould,siemens,nas,adviser,sperry,amdahl +
    -41.2397 * vendor=gould,siemens,nas,adviser,sperry,amdahl +
     32.0545 * vendor=siemens,nas,adviser,sperry,amdahl +
   -113.6927 * vendor=adviser,sperry,amdahl +
    176.5205 * vendor=sperry,amdahl +
    -51.2583 * vendor=amdahl +
      0.0616 * MYCT +
      0.0171 * MMIN +
      0.0054 * MMAX +
      0.6654 * CACH +
     -1.4159 * CHMIN +
      1.5538 * CHMAX +
    -41.4854

Time taken to build model: 0.01 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9262
Mean absolute error                     36.7763
Root mean squared error                 58.2233
```

Figure3: The result using Linear Regression

Evaluation: Correlation coefficient R=0.9012 or R2=0.8123 5 variables can predict PRP about 81



OBSERVATION OF TWO MODELS

Each data point is marked by a cross whose size indicates the absolute value of the error for that instance.The smaller crosses in Figure 4 for M5 , when compared with those in Figure 5 for linear regression , show that M5 is superior. Hence M5 performance is slightly worse than the linear regression