PERFORMANCE EVALUATION OF CPU FOR REGRESSION TECHNIQUE USING DATA MINING TOOL-WEKA

Arvind Kumar Sharma[*]

P.C. Gupta[**]

Abstract:

The objective of this paper is to predict the CPU performance using regression technique. A regression is a statistic technique that helps in finding out how the dependent variable i.e. CPU performance is related to the independent variables. In this paper we have been proposed an approach which predicts how the CPU performance is correlated with the attributes like Machine Cycle Time, Minimum Main Memory and Memory.

Keywords- Data Mining, Linear Regression, Weka

[*] Ph.D Scholar, Dept. of Comp.Sc. & Engineering, Jaipur National University, Jaipur, Rajasthan-India

[**] Associate Professor, Dept. of Comp.Sc. & Informatics, University of Kota, Kota, Rajasthan-India

I. INTRODUCTION

In 21st century the human beings are used in the different technologies to adequate in the society. Each and every day the human beings have been using the huge data and these data are in the different areas. It may be in the form of documents, may be graphical formats, may be videos, etc. As the data are available in the different formats so that the proper action to be taken. Not only to analyze these data but also take a good decision and maintain the data. As and when the user or customer will require the data should be extracted or retrieved from the database and make the better decision. This technique is actually known as data mining or simply Knowledge Discovery in Databases or Knowledge Hub. Several algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbour, etc. have been used in data mining. The ultimate aim of data mining is prediction. Therefore, predictive data mining is the most common type of data mining and is one of the areas that has the most application to businesses or life concerns.

Paper is organized in different sections: Section-II explains data mining and predictive techniques. Related works are shown in Section-III. Section-IV contains experiments and result of the complete work. Conclusion is shown in Section-V while references are mentioned in the last section.

II. DATA MINING

Data mining is usually defined as the automatic or semi automatic process of finding meaningful information in large quantities of data. The process needs to be automatic due to the huge amount of available data and the patterns found needs to be meaningful. Data Mining is the extraction of interesting and potentially useful patterns and implicit information from activity related to the World Wide Web [1]. It is a detailed process of analyzing large amounts of data and picking out the relevant information. It also refers to mining knowledge from large amount of data [12]. It is the fundamental stage inside the process of extraction of useful and comprehensible knowledge, previously unknown, from large quantities of data stored in different formats, with the objective of improving the decision of companies, organizations where the data can be collected. Data mining techniques have been applied in a great number of fields including retail sales, finance, marketing, manufacturing, health care, customer relationship, medical sciences, web education, e-commerce, bioinformatics, engineering applications, and counterterrorism. In [9] various models of data mining are classified into two groups: predictive and descriptive. Following fig.1 shows the most commonly used data mining models and their respective tasks.

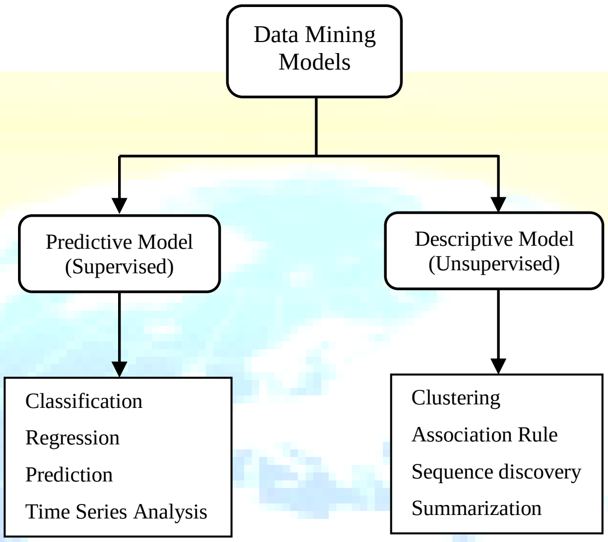


Figure 1: Classification of Data Mining Models and Tasks

The predictive model makes prediction about unknown data values by using the known values. The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined [2]. There are several predictive data-mining techniques such as regression, neural networks, decision tree, etc. but in this paper, only the regression models are discussed.

A. Regression

The term regression is defined as an analyzing or measuring the relation between a dependent variable and one or more independent variable. Regression techniques can be categorized in two categories such as: Linear regression and Logistic regression which are shown in figure2 below.

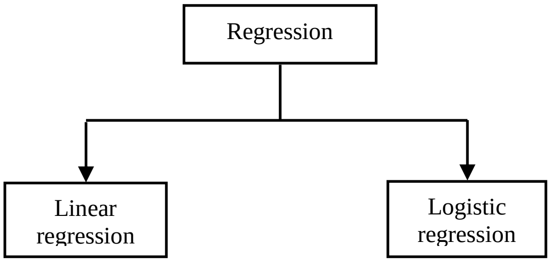


Figure 2: Regression Categories

A.1 Linear Regression

Linear Regression was historically the earliest predictive method and is based on the relationship between input variables and the output variable. Linear regression is a simple technique suitable for numeric prediction that is frequently used in statistical application. The idea is to find the amount of how much each of the attributes $\mathbf{a_1, a_2 ,..., a_k}$ in a data set contributes to the target value x . Each attribute is assigned a factor $\mathbf{w_i}$ and one extra factor is used to constitute the base level of the predicted attribute.

$$\mathbf{x = w + w_1a_1 + w_2a_2 + ... + w_ka_k}$$

The aim of linear regression is to find optimal weights for the training instances by minimizing the error between the real and the predicted values. As long as the data set contains more instances than attributes this is easily done using the least square method [11]. Linear regression is quite intuitive and easily understood but the downside is that it handles non-numerical attributes poorly and that it can't handle more complex nonlinear problems [3].

A.2 Logistic Regression

Logistic regression is a generalization of linear regression [4]. Basically it is used for estimating binary or multi-class dependent variables and the response variable is discrete, it cannot be modeled directly by linear regression i.e. discrete variable changed into continuous value. It also provides the difference in the percentage of dependent variable and provides the rank of individual variable according to its importance. Thus the main aim of Logistic regression is to determine the result of each variable correctly. Logistic regression also known as nominal regression [18]. It is a statistical technique for classifying records based on values of input fields. It is analogous to linear regression but takes a categorical target field instead of a numeric one.

Both binomial models (for targets with two discrete categories) and multinomial models (for targets with more than two categories) are supported. It works by building a set of equations that relate the input field values to the probabilities associated with each of the output field categories. Once the model is generated, it can be used to estimate probabilities for new data. For each record, a probability of membership is computed for each possible output category. The target category with the highest probability is assigned as the predicted output value for that record. Linear regression models are often quite accurate. They can handle symbolic and numeric input fields. They can give predicted probabilities for all target categories. Logistic models are most effective when group membership is a truly categorical field. Logistic regression is related to some other statistical analysis techniques but it offers more flexibility and robustness [15,16]. It does not assume linear relationship between the input and output variables, nor normal distribution and equal variance within input variables.

B. Requirements of Regression

There are various reasons for using regression technique in data mining [17]. Some of these are listed below:

- Regression models are tested by computing various statistics which measure the difference between the predicted values and the expected values.
- A regression task begins with a data set in which the target values are known [13]. For example, a regression model that predicts children's height could be developed based on observed data for many children over a period of time. The data may be: track age, height, weight, developmental milestones, family history, and so on. Height would be the target, the other attributes will be the predictors, and the data for each child would constitute a case.
- In the model build (training) process, a regression algorithm estimates the value of the target as a function of the predictors for each case in the build data. These relationships between predictors and target are summarized in a model, which can be applied to a different dataset in which the target values are unknown.

III. RELATED WORKS

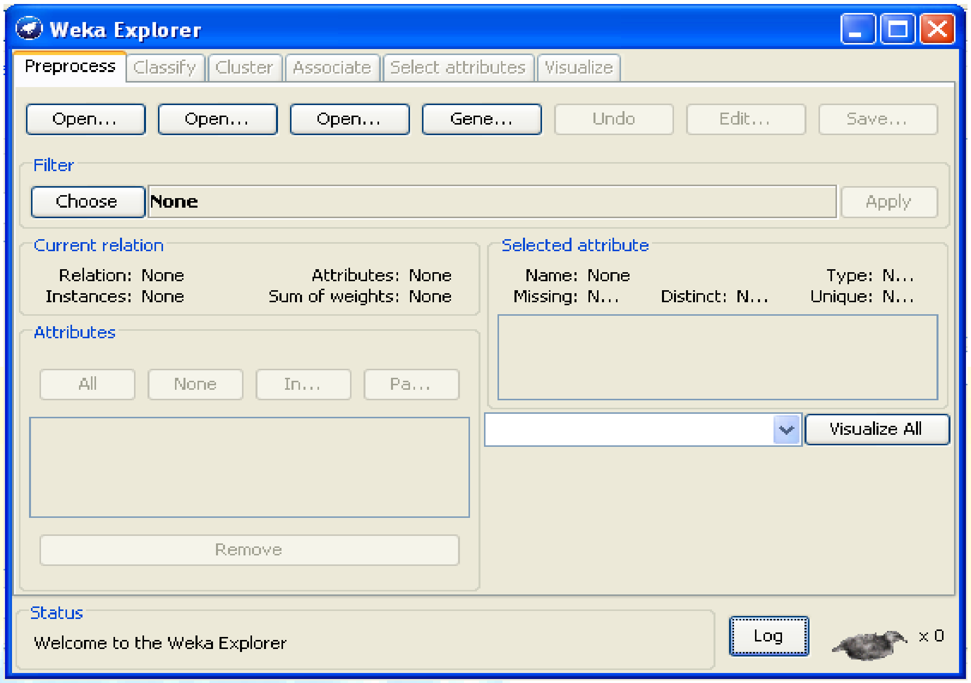
Some of the works have been done by different researchers. In one work the methodology has been proposed to solve software related problems which had been used the data mining approach [1]. In another work, the quest for patterns in data has been studied for a long time in many fields, including statistics, patterns recognition and exploratory data analysis [4,5]. Analyzing data can provide further knowledge about a business by going beyond the data explicitly stored to derive knowledge about the business. This is where data mining has obvious benefits for any enterprise. Data mining is an approach currently receiving great attention and is being recognized as a newly emerging analysis tool [6]. Recently, data mining has given a great deal of concern and attention in the information industry and in society as a whole. This is due to the wide accessibility of huge amount of data and the important need for turning such data into useful information and knowledge [7]. Logistic regression (LR), which is helpful for predicting the presence or absence of a characteristic or outcome based on values of a set of predictor variables, is a multivariate analysis model [10]. Over the years, data mining has involved several techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbour. Besides that, Data mining has been applied in many fields such as Data mining has been applied in a great number of fields including retail sales, finance, marketing, customer relationship, manufacturing, medical science, web education, e-commerce, bioinformatics, counterterrorism, and engineering applications. Nevertheless, an application of Data mining in predicting CPU performance is rare [8].

IV. EXPERIMENTS AND RESULTS

In this paper we conduct experiments using WEKA tool to predict the CPU performance using Logistic regression. We propose an approach that is applied to find how the CPU performance is correlated with the attributes like Machine Cycle Time, Minimum Main Memory and Memory.

A. Experimental Setup

Our experiments were conducted using a PC with Core i3 CPU 2.13 GHz with 2 GB of RAM memory, and they involve the data set CPU.arff. We have performed all experiments using WEKA. Following figure-3 shows the WEKA Interface, on top are panels to select that give access to the others components.



Filter: None

Relation: None Attributes: None
Instances: None Sum of weights: None

Name: None Type: N...
Missing: N... Distinct: N... Unique: N...

Welcome to the Weka Explorer

Figure 3: Weka Explorer

B. Data set Used

The dataset has acquired from UCI Machine Learning Repository [15]. It can be available at following link: www.ics.uci.edu/~mlearn/MLSummary.html.

The experimental setup is as follows:

Database used	:	CPU.arff
Attributes of Database	:	8 Attributes
Algorithms Used	:	Logistic Regression

The used attributes of the dataset CPU.arff are shown in following table-1.

Table-1: Attributes of the dataset CPU.arff

Vendor	Name of the Vendor
MYCT	Machine cycle time in nanoseconds (integer)
MMIN	Minimum main memory in kilobytes (integer)
MMAX	Maximum main memory in kilobytes (integer)
CACH	Cache memory in kilobytes (integer)

CHMIN	Minimum channels in units (integer)
CHMAX	Maximum channels in units (integer)
Class	Name of the Dependent variable

A part of the data file in .arff format is as follows:

```
@relation cpu
@attribute MYCT numeric
@attribute MMIN numeric
@attribute MMAX numeric
@attribute CACH numeric
@attribute CHMIN numeric
@attribute CHMAX numeric
@attribute class numeric
@data
125,256,6000,256,16,128,198
29,8000,32000,32,8,32,269
29,8000,32000,32,8,32,220
29,8000,32000,32,8,32,172
29,8000,16000,32,8,16,132
26,8000,32000,64,8,32,318
23,16000,32000,64,16,32,367
23,16000,32000,64,16,32,489
23,16000,64000,64,16,32,636
```

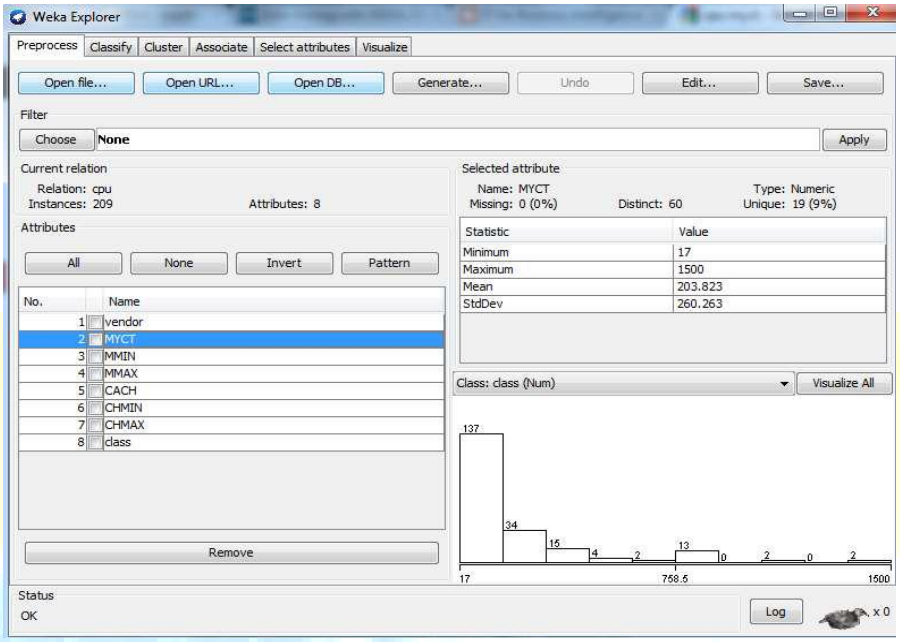
C. Implementation of Regression in Weka

In this work we have been used the regression technique through WEKA [14]. The algorithm we are going to implement to classify is Weka’s Linear Regression-S0-R 1.0E-8.

Following steps have been performed during the complete work.

Step-1. Create data file and open the Weka.

Step-2. Load the regression data file CPU.arff into Weka. Click on open file and browse for the file that is shown in figure 4 below.



Filter: None

Current relation
Relation: cpu
Instances: 209
Attributes: 8

Selected attribute
Name: MYCT Type: Numeric
Missing: 0 (0%) Distinct: 60 Unique: 19 (9%)

Statistic	Value
Minimum	17
Maximum	1500
Mean	203.823
StdDev	260.263

Attributes

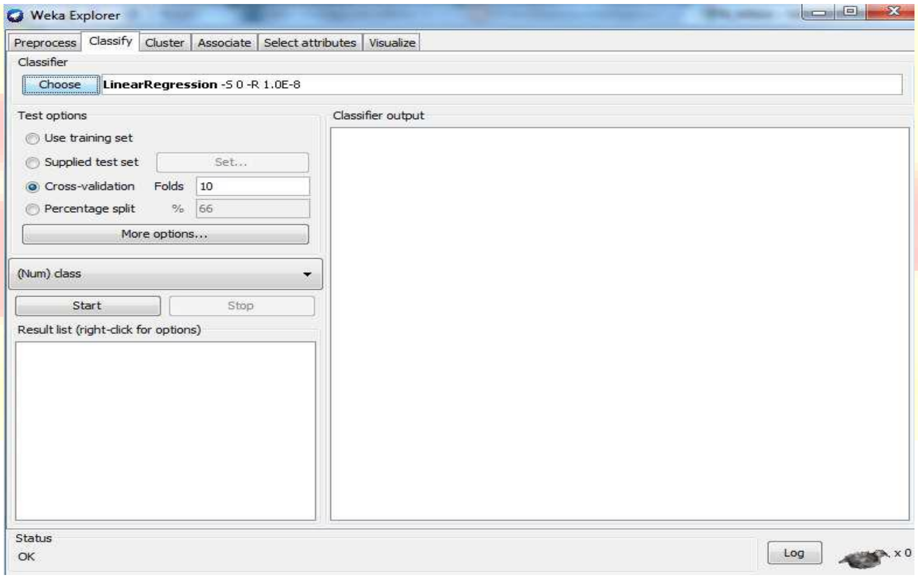
No.	Name
1	vendor
2	MYCT
3	MMIN
4	MMAX
5	CACH
6	CHMIN
7	CHMAX
8	class

Class: class (Num)

Status
OK

Figure 4: Loading Dataset CPU.arff

Step-3. We click on the **Classify** tab to create the model and choose **Linear Regression** from the node under function. This is shown in figure 5.



Classifier
LinearRegression -S 0 -R 1.0E-8

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Num) class

Result list (right-click for options)

Status
OK

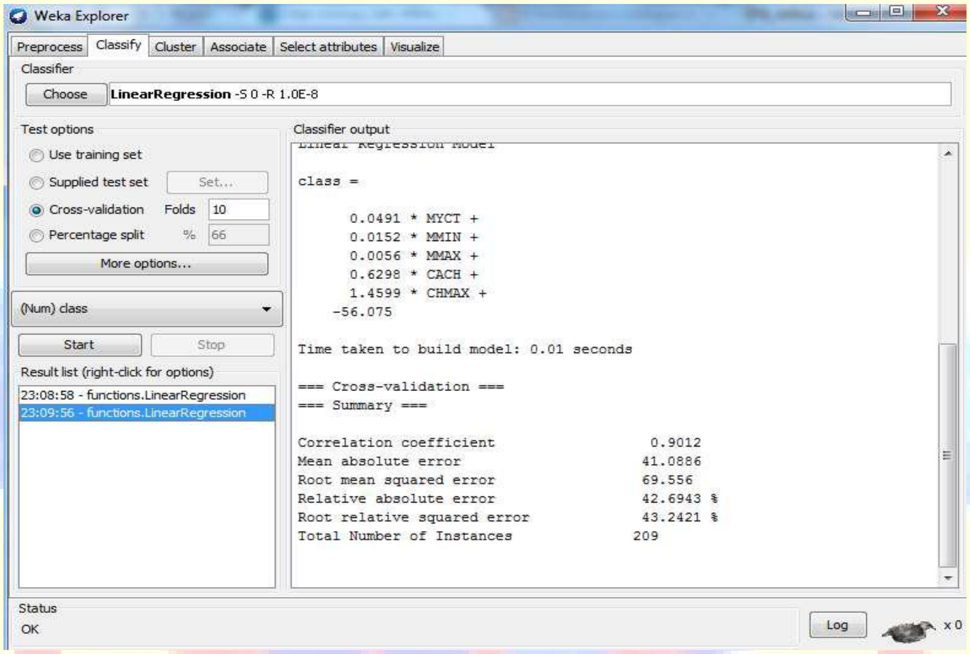
Figure 5: Linear Regression in Weka

This asks Weka that we want to build a regression model. As one can see from the other choices, though, there are lots of possible models to build. There is another choice called **Simple Linear**

Regression in the same node under function. We do not use this here, because simple regression only works at one variable, and we have more than one.

Step-4. Run the regression.

Finally, the last step to creating our model is to choose the dependent variable (the attribute we are looking to predict). In this work the dependent variable is Class. Click on start that displays results in the classifier output window that gives regression equation, shown in figure 6 below.



Classifier: **LinearRegression** -S 0 -R 1.0E-8

Test options:
- ○ Use training set
- ○ Supplied test set
- ● Cross-validation Folds 10
- ○ Percentage split % 66

(Num) class

Result list (right-click for options)
- 23:08:58 - functions.LinearRegression
- 23:09:56 - functions.LinearRegression

Classifier output

```
Linear Regression Model

class =

      0.0491 * MYCT +
      0.0152 * MMIN +
      0.0056 * MMAX +
      0.6298 * CACH +
      1.4599 * CHMAX +
    -56.075

Time taken to build model: 0.01 seconds

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

Correlation coefficient                  0.9012
Mean absolute error                     41.0886
Root mean squared error                 69.556
Relative absolute error                 42.6943 %
Root relative squared error             43.2421 %
Total Number of Instances              209
```

Status: OK

Figure 6: Results of Regression

D. Classifiers

The Classifiers **class** generates a decision tree classifier for the dataset given as input. Also, a 10-Fold Cross-validation estimation of its performance is calculated. The Classifiers package implements the most common techniques separately for categorical and numerical values which are described. (see Appendix-A)

E. Interpretation of the Results

From the above results shown in figure 6, we can observe that the Correlation Coefficient is obtained **0.912**, which is very high. So that the result suggests that the dependent variable is strongly associated with the independent variables. Thus the CPU performance is more dependent on CHMAX (Maximum Channel) and then CACH (Cache Memory). Similarly we can also investigate the new CPU performance by using the regression technique, if we have the values of the attributes.

V. CONCLUSION

In this paper we have used the Linear regression to determine the factors which significantly affect the performance of the CPU in a Computer System. This technique helps the computer professionals and users to form an opinion about the CPU performance to be determined. It may be observed that the CPU performance is more dependent on CHMAX (Maximum Channels) and then CACHE (Cache Memory). The correlation coefficient **0.912** has obtained after the experiments which is very high. Therefore the result suggests that the dependent variable is strongly associated with the independent variables. We can also determine the new CPU performance by using the regression if we have the values of the attributes.

REFERENCES

[1] Berners-Lee T.J., Cailliau R., Groff J.F., Pollermann B. (1992) Electronic Networking: Research, Applications and Policy, 2(1).

[2] Arvind Sharma, P.C. Gupta, “Predicting the Number of Blood Donors through their Age and Blood Group by using Data Mining Tool”, International Journal of Communication and Computer Technologies, ISSN: 2278-9723, Vol.1-No.6, Issue: 02, September 2012.

[3] M. H. Dunham. Data mining : introductory and advanced topics, Prentice Hall/Pearson Education, Upper Saddle River, NJ, 2003.

[4] De Mantaras & Armengol E. (1998), “Machine learning from example: Inductive and Lazy methods”, Data & Knowledge Engineering 25: 99-123.

[5] http://en.wikipedia.org/wiki/Regression_analysis.

[6] Tso, G.K.F. and K.K.W. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks". Energy, 2007. 32: p. 1761 - 1768.

[7] Han, J. and M. Kamber, "Data Mining: Concepts and Techniques", San Francisco: Morgan Kaufmann Publisher, 2006.

[8] Chien, C.F. and L.F. Chen, "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry", Expert Systems and Applications, 2008. 34(1): p. 380-290.

[9] Margret H Dunham, “Data Mining Introductory and Advanced Topics”, Pearson education ISBN 978-81-7758-785-2.

[10] Application of likelihood ratio and logistic regression models to landslide susceptibility mapping using GIS. Environmental Management 34(2), 223-232.

[11] I. H. Witten and E. Frank. Data mining: practical machine learning tools and techniques, Morgan Kaufmann, Amsterdam, 2005.

[12] Jaiwei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Second Edition, Morgan Kaufmann Publishers.

[13] Nanhay Singh et al., "Data mining with regression technique", Journal of Information Systems and Communication, Volume 3, Issue-1, 2012, pp.-199-202.

[14] http://www.cs.waikato.ac.nz/ml/weka/

[15] B. Hamadicharef, et al., "Performance evaluation and fusion of methods for early detection of alzheimer disease", In Proc. Int. Conf. BioMedical Engineering and Informatics BMEI 2008, vol. 1, pp. 347–351, May 2008.

[16] B. G. Tabachnick and L. S. Fidell, "Using Multivariate Statistics", 5th ed., Allyn & Bacon, Inc., Needham Heights, MA, USA, pp. 437-505, 2007.

[17] www.bioinfo.in/uploadfiles/13476038793_1_41_JISC.pdf

[18] SPSS Clementine help file. http//www.spss.com

Appendix–A

(Sample Executions for Classifiers)

Classifiers for Numerical Prediction

1.	weka.classifiers.LinearRegression	Linear regression
2.	weka.classifiers.m5.M5Prime	Model trees
3.	weka.classifiers.Ibk	K-nearest neighbor learner
4.	weka.classifiers.LWR	Locally weighted regression
5.	weka.classifiers.RegressionByDiscretization	Uses categorical classifiers

Sample Executions of the Linear Regression Algorithm

Linear Regression Model:

```
> java weka.classifiers.LinearRegression -t data/cpu.arff

Linear Regression Model

class =
  -152.7641 * vendor=microdata,formation,prime,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=formation,prime,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,am
dahl +
    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 +
    21.4119 * vendor=ipl,ibm,cdc,ncr,basf,gould,siemens,nas,adviser,sperry,amdahl +
   -41.2396 * vendor=gould,siemens,nas,adviser,sperry,amdahl +
    32.0545 * vendor=siemens,nas,adviser,sperry,amdahl +
  -113.6927 * vendor=adviser,sperry,amdahl +
   176.5204 * 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

=== Error on training data ===

Correlation coefficient                  0.963
Mean absolute error                     28.4042
Root mean squared error                 41.6084
Relative absolute error                 32.5055 %
Root relative squared error             26.9508 %
Total Number of Instances              209

=== Cross-validation ===

Correlation coefficient                  0.9328
Mean absolute error                     35.014
Root mean squared error                 55.6291
Relative absolute error                 39.9885 %
Root relative squared error             35.9513 %
Total Number of Instances              209

=== Run information ===
Scheme:       weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8
Relation:     machine_cpu
Instances:    209
Attributes:   8
              MYCT
              MMIN
              MMAX
              CACH
              CHMIN
              CHMAX
              class
Test mode:    evaluate on training data
=== Classifier model (full training set) ===
Linear Regression Model
class =
      0.0491 * MYCT +
      0.0152 * MMIN +
      0.0056 * MMAX +
      0.6298 * CACH +
      1.4599 * CHMAX +
    -56.075
Time taken to build model: 0.1 seconds
=== Evaluation on training set ===
=== Summary ===
Correlation coefficient              0.9012
Mean absolute error                 41.0886
Root mean squared error             69.556
Relative absolute error             42.6943 %
Root relative squared error         43.2421 %
Total Number of Instances           209
```

Here **class** represents (PRP) Published Relative Performance, which is dependent variable.