ISSN 2229-5518

Simulation of the Framework for Evaluating Academic Performance (FEAP) using WEKA

A. S. Ahmadu[1], S. Boukari[2], E. J. Garba[3], K. J. Danjuma[4]

[1]Department of Computer Science MAUTECH, Yola: ahmaduasabe@mautech.edu.ng

[2]Mathematics programme ATBU Bauchi: bsouley2001@yahoo.com

[3]Department of Computer Science MAUTECH, Yola: e.j.garba@mautech.edu.ng

[4]Department of Computer Science MAUTECH, Yola: jkdanjuma@mautech.edu.ng

Abstract

Decision trees have proven to be efficient in identifying factors responsible for students' success or failure. In this paper, J48 decision tree, a classifier in Waikato Environment for Knowledge Analysis (WEKA) suit was used for the simulation of the model 'A generic Framework for Evaluating Academic Performance (FEAP)'; on the sample data collected from Modibbo Adama University of Technology (MAUTECH) Yola, Nigeria. It has identified the factors responsible for academic performance as: Previous academic performance, Carry-over courses, marital status, parental status, electricity supply, accommodation-type and course-choice-influence. The model's performance is excellent with accuracy of 93.33% this indicates that the results obtained from the training data are optimistic.

1. Introduction

The sole mission of any institution is to produce scholarly graduates because the development of a nation is dependent on its educational efficacy. Educational Data Mining is emerging as a tool for storage and structuring of academic records of students in a form that is adaptive for analysis and forecasting of students' performance using the con8cept learned from the huge accumulated database. Predicting the success or failure of a student in a course helps in warning students to change course of study early enough before being withdrawn. Decision Trees help in discovering factors that could be responsible for good performance, these discoveries could further help school authority take some precautionary measures as well as in making correct placement during enrolment. The aim of this paper is to simulate the generic Framework for Evaluating Academic Performance (FEAP) using WEKA decision tree for the data mining task.

2. Related Work

The following are extracts of some of the works done by some researchers:

Aziz et al.(2014) developed Students' Academic Performance prediction models on first semester Bachelor of Computer Science University Sultan ZainalAbidin (UniSZA) using three selected classification methods; Naïve Bayes, Rule-Based and Decision Tree. The experiment was carried out on five attributes namely: gender, race, hometown, family income and university entry mode to discover the best classification model for prediction. Results show that the models developed using Rule-Based and Decision Tree algorithm gave the best predictions compared to the model developed from the Naïve Bayes algorithm. The result also uncovered influence of the parameters to the students' academic performance (SAP) in the following hierarchy: Race, family income, gender, university entry mode, and hometown location.

Khan (2005) carried out a research on 200 boys and 200 girls of science students at the higher secondary level in the Aligarh Muslim University, Aligarh, India. The clustering and random selection technique was used in the selection with the aim of establishing the predictive value of different measures of cognition, personality and demographic variables for success. The discovery made was that girls with high socio-economic status had relatively higher academic achievement in sciences while boys with low socio-economic status had a relatively higher academic achievement.

Kotsiantis, *et al.* (2004) took sample population of 365 computer science students from distance learning stream of Hellenic Open University, Greece. They made use of their demographic attributes like sex, age, marital status as the independent variable and mark as the dependent variable. They applied five classification algorithms namely Decision Trees, Perception-based Learning, Bayesian Nets, Instance-Based Learning and Rule-learning in order to predict the performance of the students. The variables of importance were selected using filter based variable selection technique. Naïve-Bayes gave the highest predictive accuracy of 74%.

Worley (2007), in her dissertation titled At-Risk Students and Academic Achievement run five regressions on the dependent variable (class) GPA against the attributes Teacher-Student relationship, Parents-Student Relationships, Motivation, Peer-influence and Socioeconomic-status in order to find out if any of the independent variables could predict the dependent variable. The strongest variance found were motivation and peer influence.

3. WEKA

Weka is a flightless bird with an inquisitive nature that is found in the Island of New Zealand. WEKA as an acronym stands for Waikato Environment for Knowledge Analysis. It is a popular suite of machine learning software written in Java, developed at the Waikato University, New Zealand. It is a state-of-the-art collection of machine learning algorithms for data mining tasks. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization (Michael and Gordon, 2004). Research has revealed that there is no single machine learning scheme appropriate to all data mining problems (Witten and Frank 2005). WEKA is a diverse and comprehensive toolkit that has an interface that allows its users compare different methods and identify those that are most appropriate for the given problem.

WEKA is fully implemented in the Java programming language and thus it supports cross platform deployment and usage. By taking advantage of the receptiveness WEKA provides, we have developed an algorithm for the integration of the Human Learning (HL) and the WEKA Machine Learning (ML) platform. Figure 1 shows the working principle of WEKA DM processes.

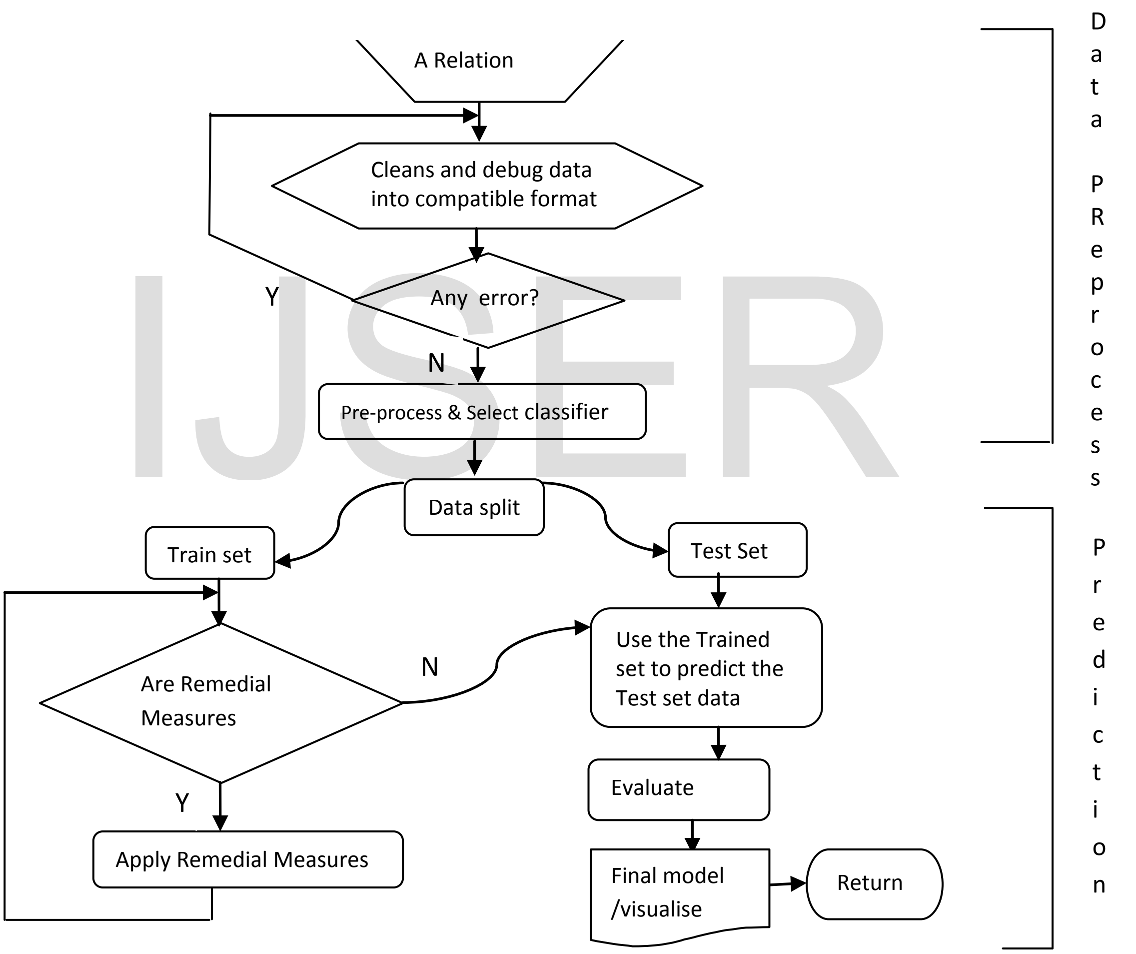

Figure 1: Flowchart of Predictive Data mining

i) Pre-Processing

Pre-processing are the techniques for preparing data for the data mining task. Data can be imported to WEKA in various formats like: ARFF, CSV, C4.5 and binary. Most spreadsheet applications and database programs allow export of data into a file in comma-separated value (CSV) format as a list of records with commas between items. Pre-processing tools in WEKA are called "filters", it contains filters for discretization, normalization, re-sampling, attributes selection, transforming and combining attributes (Aksenova, 2014).

ii) Filters

Filters are tools for pre-processing data in WEKA. They are responsible for data transformation, removing/adding attributes from/to a dataset, discretization of numeric attributes into nominal ones. Some techniques require that data be in numeric type while others require that they be in nominal. For instance when Simple Linear Regression algorithm is to be applied on a data set, the dependent variable is expected to be in numeric type. If it has been captured in nominal form, you do not need to change the data to numeric type manually; you only need to discretize the attribute so as to transform it into numeric form. The same thing applies when it is required to be transformed from numeric to nominal.

There are two basic approaches to the problem of discretization:

a) Unsupervised: Quantize each attribute in the absence of any knowledge of the classes of the instances in the training set, for instance when handling clustering problems.

b) *Supervised*: Takes the class value of the instances into account when creating intervals (discretizing). According to Witten and Frank (2005), WEKA's main unsupervised method for discretizing numeric attributes is: *weka.filters.unsupervised.attribute.Discretize*. It implements these two methods: equal-width (the default) and equal-frequency (when discretizing).

iii) Normalization

Normalization scales all numeric values in the dataset to lie between 0 and 1. It standardizes and transforms them to have zero mean and unit variance (skip the class attribute, if set).

v) Setting Test Options

WEKA has four provisions in its test option settings namely: Use training set, supplied test set, cross-validation (with 10-fold set as its default), and percentage split (66% default split of the data would be used as training while the remaining 34% for the test data). Before classifiers are run, one of the test options radio must be selected otherwise cross-validation being the default would be used as the test option selected. We have tried out the four methods in each of the classifiers we have picked and have used the test option that gives the best performance.

Kirkby (2002), has defined the various test options available in WEKA as thus:

a) **Use training set.** Evaluates the classifier on how well it predicts the class of the instances it was trained on.

b) **Supplied test set.** Evaluates the classifier on how well it predicts the class of a set of instances loaded from a file. Clicking on the 'Set…' button brings up a dialog allowing you to choose the file to test on.

c) **Cross-validation.** Evaluates the classifier by cross-validation – using the number of folds that are entered in the 'Folds' text field.

d) **Percentage split.** Evaluates the classifier on how well it predicts a certain percentage of the data, which is held out for testing. The amount of data held out depends on the value entered in the '%' field.

iii) Attribute selection

There are some attributes that are actually irrelevant in a data set and tend to give misleading or confusing information when using marching learning systems; as such it is common to precede learning with attribute selection. WEKA has a module that handles selection of attributes in order to find which subset works best.

There are two different methods of selecting attribute subset: i) filter method where the attribute set is filtered to produce the most promising subset before learning commences. The selection is based on the general characteristics of the data. ii) The *wrapper* method where the learning algorithm is wrapped into the selection procedure in order to pick the best subset (Witten and Frank 2005). In this experiment we have applied the first method because it gives a satisfactory outcome.

iv) Classifiers

According to Kumar and Chadha (2011), classification is the processing of finding a set of models (or functions) which describe and distinguish data concepts, for the purposes of being able to use the model to predict the class of objects whose class label is unknown. The model is generated based on the analysis of a set of training data and is used to predict the class label of unclassified objects.

Our aim is to explore different algorithms and then pick the best performing model. This is shown in the Figure 2: Classifiers in WEKA make use of SQL statements; and the learning schemes available include Decision trees, Instance-based classifiers, Support Vector Machines, Multi-layer Perceptions, Logistic Regression and Bayes' Nets.

Classifiers obtained from decision tree algorithms have the characteristic of not requiring previous domain knowledge or heavy parameter tuning; making them appropriate not only for prediction but also for exploratory data analysis (Witten and Fran, 2005). Under this process our goal is to generate decision tree with the aim to discover the factors that determine performance of the students. This is shown in Figure 2.

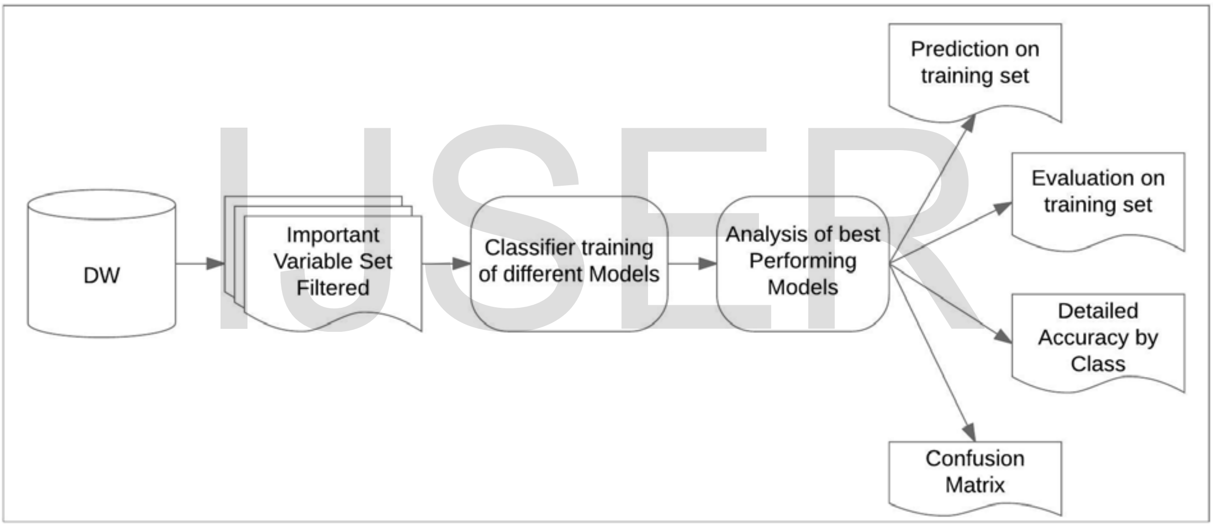

Figure 2: Classifier training and model analysis

v) Extracted Sample Data

To be able to proceed with the processing and analysis of collected data we selected a data set from the DW architecture or the databases. The data extraction model variables are classified as shown in Table 1:

Table 1: Data extraction Variable

VARIABLES	DESCRIPTION	DATA TYPE
State	All states in Nigeria	Nominal
Gender	Male/Female	Nominal
Age range	Less than 20, 20 to 35, 36 to 50, above 50	Nominal
Marital status	Single, married, divorced, separated, widow, widower	Nominal
Course_of_Study	All courses in the University	Nominal
Parental status	Together, separated, mother deceased, father deceased	Nominal

Educational background of fa ther/guardian	Illiterate, attained primary school, attained secondary school, diploma /Nurse, university graduate	Nominal
Educational background mother/ guardian	Illiterate, attained primary school, attained secondary school, diploma /Nurse, university graduate	Nominal
Family size	Less than 5, between 5 and 9, Above 10	Nominal
Parental motivation	Setting a performance target with a reward promised, rewarding whenever I perform well, I have never been motivated with rewards	Nominal
Off campus	Yes, No	Boolean
Accommodation type	A single person in a room, 2 to 4 in a room, 5 to 9 in a room, above 9 in a room, Squatting	Nominal
Electricity supply	Adequate, moderate, inadequate, not at all	Nominal
Level	100, 200, 300, 400, 500	Numeric
Course Choice influence	Self, imposed by university, parents/guardian	Nominal
Student Occupation	Civil servant, self-employed, ordinary, others	Nominal
Health challenge	Frequent ailment, Physical disability, Hearing impairment, visual impairment, Mental retardation	Nominal
Curriculum delivery	Thoroughly, moderately, poorly, not covered	Nominal
Frequency of lectures	Very frequently, frequently, not frequently, rarely	Nominal
Internet usage	Very frequently, frequently, less frequently, not at all	Nominal
Internet connectivity means	Wi-Fi, modem, phone, none	Nominal
Social media account type	Facebook, WhatsApp, Twitter, You Tube, others	Nominal
Mode of Entry	Pre-degree, UTME/JAMB, DE, Inter-University transfer	Nominal
Entry Grade	Distinction, upper credit, lower credit, merit	Nominal
UTME Entry Score	180 to 199, 200 to 220, 221 to 250, above 250	Nominal
Carry over	Yes , No	Boolean
Course	Marks	Numeric
Performance	Yes, No	Boolean
Current CGPA	Below 1.50, 1.50 to 2.39, 2.40 to 3.49, 3.50 to 4.49, above 4.49	Nominal

4. Results

Logistic Regression

=== Run information ===

Scheme: weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4

Relation: WEKA DATA44-weka.filters.unsupervised.attribute.Remove-R1-2,16,20

Instances: 78, Attributes: 24

GENDER
AGERANGE
MARITALSTATUS
COURSECHOICE
LIVELIHOOD
STATE
OFFCAMPUS
ACCOMMODATION
WATER
ELECTRICITY
SOCIALFACTOR
ENTRYMODE
GRADESCORE
CGPA
CARRYOVER
PERFORMANCE1
PARENTALSTATUS
FATHER

MOTHER
FAMILYSIZE
PARENTMOTIVATION
MONTHLYEARNING
CONNECTIVITY
SOCIALMEDIA ACCTS
Time taken to build model: 1.24 seconds

=== Evaluation on training set ===
Time taken to test model on training data: 0 seconds
=== Summary ===

Correctly Classified Instances	76	100 %
Incorrectly Classified Instances	0	0 %
Kappa statistic	1	
Mean absolute error	0	
Root mean squared error	0	
Relative absolute error	0 %	
Root relative squared error	0.0001 %	
Total Number of Instances	76	
Ignored Class Unknown Instances	2	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	2.40 to 3.49
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.50 to 2.39
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	3.50 to 4.49
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Less than 1.50
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Greater than 4.49
Weighted Avg.	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	

=== Confusion Matrix ===
a b c d e <-- classified as
37 0 0 0 0 | a = 2.40 to 3.49
0 12 0 0 0 | b = 1.50 to 2.39
0 0 24 0 0 | c = 3.50 to 4.49
0 0 0 1 0 | d = Less than 1.50
0 0 0 0 2 | e = Greater than 4.49

NaiveBayes

=== Run information ===

Scheme: weka.classifiers.bayes.NaiveBayes

Relation: WEKA DATA44-weka.filters.unsupervised.attribute.Remove-R1-2,16,20
Instances: 78
Attributes: 24
GENDER
AGERANGE
MARITALSTATUS
COURSECHOICE
LIVELIHOOD
STATE
OFFCAMPUS
ACCOMMODATION
WATER
ELECTRICITY
SOCIALFACTOR
ENTRYMODE
GRADESCORE
CGPA
CARRYOVER

PERFORMANCE1
PARENTALSTATUS
FATHER
MOTHER
FAMILYSIZE
PARENTMOTIVATION
MONTHLYEARNING
CONNECTIVITY
SOCIALMEDIA ACCTS

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances	68	89.4737 %
Incorrectly Classified Instances	8	10.5263 %
Kappa statistic	0.8306	
Mean absolute error	0.0662	
Root mean squared error	0.197	
Relative absolute error	25.543 %	
Root relative squared error	55.1521 %	
Total Number of Instances	76	
Ignored Class Unknown Instances	2	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.946	0.077	0.921	0.946	0.933	0.869	0.969	0.962	2.40 to 3.49
	0.750	0.000	1.000	0.750	0.857	0.846	0.927	0.896	1.50 to 2.39
	0.958	0.096	0.821	0.958	0.885	0.831	0.973	0.944	3.50 to 4.49
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Less than 1.50
	0.000	0.000	0.000	0.000	0.000	0.000	0.961	0.393	Greater than 4.49
Weighted Avg.	0.895	0.068	0.879	0.895	0.882	0.832	0.964	0.932	

=== Confusion Matrix ===

```
 a  b  c  d  e   <-- classified as
35  0  2  0  0 |  a = 2.40 to 3.49
 2  9  1  0  0 |  b = 1.50 to 2.39
 1  0 23  0  0 |  c = 3.50 to 4.49
 0  0  0  1  0 |  d = Less than 1.50
 0  0  2  0  0 |  e = Greater than 4.49
```

J48 pruned tree

ISSN 2229-5518

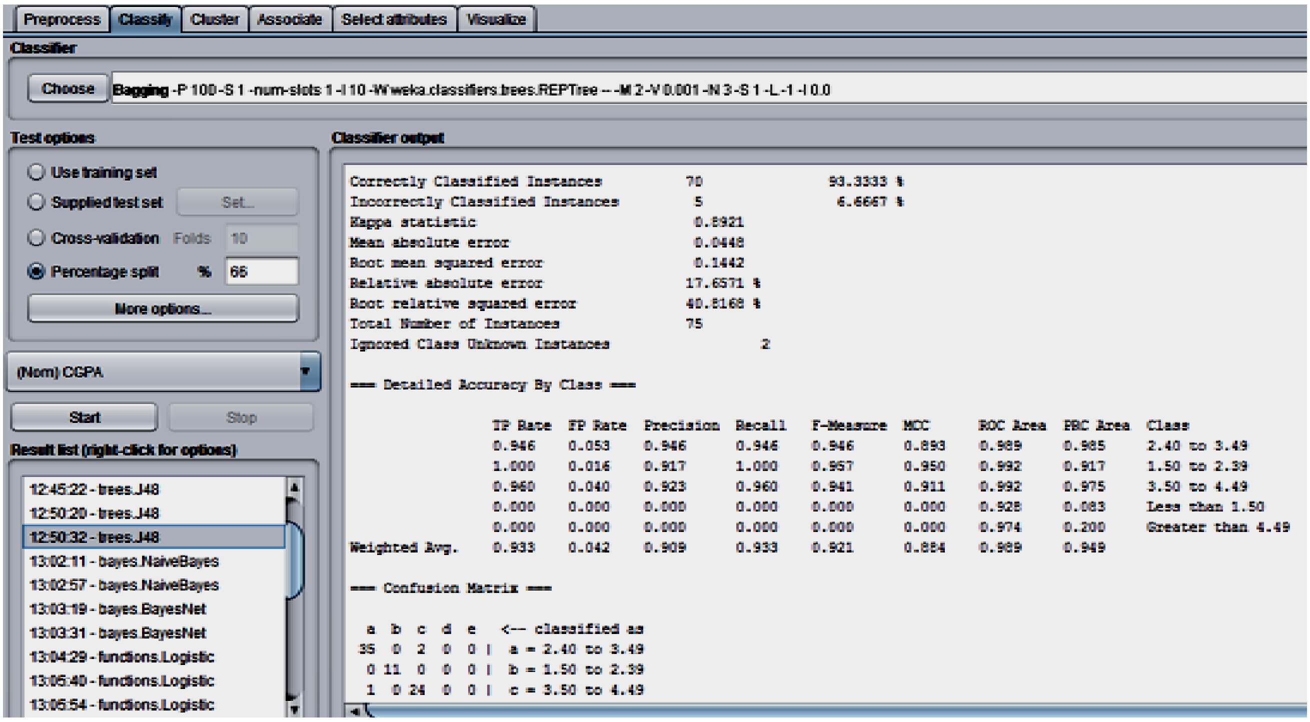

Figure 3: J48 Tree Classifier results

=== Run information ===

Scheme: **weka.classifiers.trees.J48 -C 0.25 -M 2**

Relation: **WEKA DATA4-weka.filters.unsupervised.attribute.Remove-R1-2**

Instances: **77**

Attributes: **24**

GENDER

AGERANGE

MARITALSTATUS

COURSECHOICE

LIVELIHOOD

STATE

OFFCAMPUS

ACCOMMODATION

WATER

ELECTRICITY

SOCIALFACTOR

ENTRYMODE

GRADESCORE

CGPA

CARRYOVER

PERFORMANCE

PARENTALSTATUS

FATHER

MOTHER

FAMILYSIZE

PARENTMOTIVATION

MONTHLYEARNING

CONNECTIVITY

SOCIALMEDIA ACCTS

Test mode: evaluate on training data

```
=== Classifier model (full training set) ===
PERFORMANCE <= 0: 1.50 to 2.39 (12.0/1.0)
PERFORMANCE > 0
|  CARRYOVER = YES: 2.40 to 3.49 (13.0)
|  CARRYOVER = NO
|  |  MARITALSTATUS = Married: 2.40 to 3.49 (8.33/0.17)
|  |  MARITALSTATUS = Single
|  |  |  COURSECHOICE = Self
|  |  |  |  FAMILYSIZE = Greater than 5
|  |  |  |  |  MONTHLYEARNING = 10000 to 30000
|  |  |  |  |  |  ELECTRICITY = Moderate: 2.40 to 3.49 (5.83/2.0)
|  |  |  |  |  |  ELECTRICITY = Adequate: 3.50 to 4.49 (1.0)
|  |  |  |  |  |  ELECTRICITY = Inadequate: 3.50 to 4.49 (7.0/2.0)
|  |  |  |  |  |  ELECTRICITY = Not at all: 3.50 to 4.49 (1.0)
|  |  |  |  |  MONTHLYEARNING = Less than 100000: 3.50 to 4.49 (0.0)
|  |  |  |  |  MONTHLYEARNING = 51000 to 100000: 2.40 to 3.49 (1.0)
|  |  |  |  |  MONTHLYEARNING = Less than 10000: 3.50 to 4.49 (3.75)
|  |  |  |  |  MONTHLYEARNING = 31000 to 50000: 2.40 to 3.49 (2.0)
|  |  |  |  FAMILYSIZE = Greater than 10: 2.40 to 3.49 (4.14/0.14)
|  |  |  |  FAMILYSIZE = 2: 2.40 to 3.49 (0.0)
|  |  |  |  FAMILYSIZE = 4: 2.40 to 3.49 (1.04/0.04)
|  |  |  |  FAMILYSIZE = 3: 3.50 to 4.49 (1.04)
|  |  |  |  FAMILYSIZE = Greater than 4: 2.40 to 3.49 (1.04/0.04)
|  |  |  COURSECHOICE = Religion: 3.50 to 4.49 (0.0)
|  |  |  COURSECHOICE = Glamour of the course: 3.50 to 4.49 (6.0)
|  |  |  COURSECHOICE = Imposed by the University/UTME: 3.50 to 4.49 (5.83)
|  |  |  COURSECHOICE = Influenced by parents/quardian: 2.40 to 3.49 (1.0)
|  |  |  COURSECHOICE = Peer group: 3.50 to 4.49 (0.0)
|  |  |  COURSECHOICE = Financial benefits/Marketability: 3.50 to 4.49 (0.0)
Number of Leaves  :     22
Size of the tree :  29
Time taken to build model: 0.02 seconds
```

ISSN 2229-5518

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances	70	93.3333 %
Incorrectly Classified Instances	5	6.6667 %
Kappa statistic	0.8921	
Mean absolute error	0.0448	
Root mean squared error	0.1442	
Relative absolute error	17.6571 %	
Root relative squared error	40.8168 %	
Total Number of Instances	75	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.946	0.053	0.946	0.946	0.946	0.893	0.989	0.985	2.40 to 3.49
	1.000	0.016	0.917	1.000	0.957	0.950	0.992	0.917	1.50 to 2.39
	0.960	0.040	0.923	0.960	0.941	0.911	0.992	0.975	3.50 to 4.49
	0.000	0.000	0.000	0.000	0.000	0.000	0.928	0.083	Less than 1.50
	0.000	0.000	0.000	0.000	0.000	0.000	0.974	0.200	Greater than 4.49
Weighted Avg.	0.933	0.042	0.909	0.933	0.921	0.884	0.989	0.949	

=== Confusion Matrix ===

a b c d e <-- classified as

35 0 2 0 0 | a = 2.40 to 3.49

0 11 0 0 0 | b = 1.50 to 2.39

1 0 24 0 0 | c = 3.50 to 4.49

0 1 0 0 0 | d = Less than 1.50

1 0 0 0 0 | e = Greater than 4.49

5. Performance Evaluation

In Data Mining, it is necessary to measure performance. Evaluation measures or assesses the predictive performance of the classifier as depicted in table 2 below. The predictions made by a classifier are interpreted from its confusion matrix – the size of which depends on the number of outcomes in the dependent variable (class). Confusion matrix is always a square matrix.

The horizontal and vertical labels represent the same thing i.e. the class label used for the prediction and in our experiment the class labels are: Greater than 4.49, 3.50 to 4.49, 2.40 to 3.49, 1.5 to 2. 39, and less than 1.5. The correctly classified instances are the diagonal values in the matrix which are the intersections of each instance. Values above the diagonally classified instances are incorrectly classified as not meeting the target while values below the diagonally classified instances are incorrectly classified as meeting the target. The illustration made in the two by two matrix in Table 2 below does not in any way restrict confusion matrix to two by two matrix.

i. True positives (*TP*) are the number of students **correctly** classified as having performance within a particular class of degree.
ii. False positives (*FP*) are the number of students **incorrectly** classified as having performance within a particular class of degree.
iii. True negatives (*TN*) are the number of students **correctly classified as not** having performance within a particular class of degree.
iv. False negatives (*FN*) are the number of students **incorrectly classified as not** having performance within a particular class of degree.

Table 2: Confusion matrix

	Predicted positive	Predicted negative
Actual positive	TP	FN
Actual negative	FP	TN

From these entries, there three evaluation measures that can be figure out:

v. Precision (*P*) is the number of students correctly classified as having graduated within the time frame given, divided by the total number of students predicted as having graduated within the time frame (Eq. 1).
vi. Recall (*R*) is the number of students correctly classified as having graduated within the time frame given, divided by the total number of students that graduated within the time frame (Eq. 2);
vii. F-Measure (F1) combines both precision and recall with equal weights into a single measure (Eq. 3).
viii. Accuracy: Compares how close a new test value is to a value predicted by **if ... then** rules (Ciosa and Moore 2002), written as in (Eq.4).

$$P = \text{Precision} = \frac{TP}{TP+FP} \quad ...1$$

$$R = \text{Recall} = \frac{TP}{TP+FN} \quad ...2$$

$$F1 = \text{F-Measure} = \frac{2PR}{P+R} \quad ...3$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}\ 100\% \quad ...4$$

From the evaluation formulas given in section 4 above:

$$\text{Precision} = \frac{TP}{TP+FP} = 70/(70 + 2) = 0.9722222222$$

Recall $= \frac{TP}{TP+FN}$ = 70/(70 + 3) = 0.9589041096

F-Measure = $\frac{2PR}{P+R}$

= 2 * 0.9722222222 * 0.9589041096/(0.9722222222 + 0.9589041096)

= 1.8645357686/1.9311263318= 0.9655172414

Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$ 100% = (70 +0)/(70 + 0 + 2 + 3)* 100 = (70/75)* 100

= 0.9333333333* 100 = 93.33%

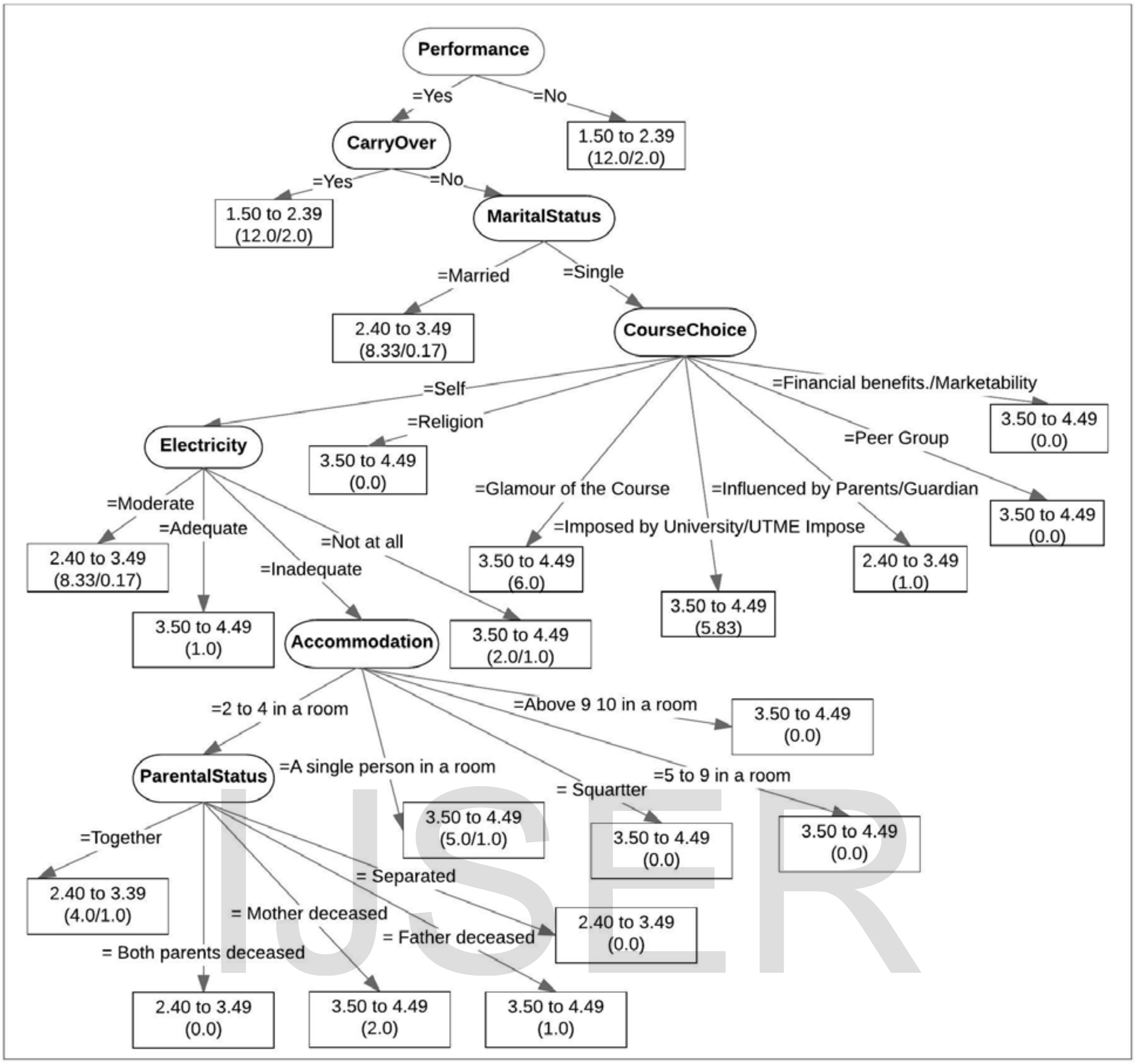

Figure 4: J48 Decision Tree

6. Discussion of Results

The accuracy value obtained above corresponds to the computed value by the machine (93.3333%).

The time taken to evaluate the model was 0.02 sec. The correctly classified instances are 70 with the accuracy performance of 93.3333 %, while the incorrectly classified are 5 making the remaining 6.6667 %.

The Mean absolute error was 0.0448; Relative absolute error is 17.6571%.

Out of the twenty four attributes picked, the classifier identified seven as the most important factors responsible for predicting academic performance. Factors are displayed in a hierarchical order starting at the root with the most important down to the leaves. In the hierarchy above, performance (i.e. previous students' performance measured in terms of cumulative Grade Point

Average (CGPA) >= 2.40 as pass while<2.40 as fail) happens to be the most important attribute for the prediction, followed by carry over (delayed courses), marital status, course choice influence, electricity, accommodation and finally parental status as the factor of least importance.

7. Conclusion

The evaluation of results varied as the classifiers and data sets were varied with some attaining 100% accuracy, but we chose to dwell on J48 tree since it gives evaluation results including attributes that are of importance through the hierarchical tree structure. These factors identified by the model are listed in order of significance as: performance, carry over (delayed courses), marital status, course choice influence, electricity, accommodation and finally parental status which can be vividly seen from the hierarchical structure of J48 tree.

References

Aksenova, S. S. (2004); Machine Learning with WEKA, WEKA Explorer Tutorial Version 3.4.3 California State University, Sacramento California, 95819 pp 10-20 aksenovs@ecs.csus.edu (Access 15/03/2016)

Azwa, A. A., Nor, H. I, and Fadhilah, A. (2014). First Semester Computer Science Students' Academic Performances Analysis by Using Data Mining Classification Algorithms, Proceeding of the International Conference on Artificial Intelligence and Computer Science (AICS 2014), 15 - 16 Bandung, INDONESIA. (e-ISBN978-967-11768-8-7).

Ciosa, K. J. and Moore, G. W. (2002); "Uniqueness of medical data mining," Artificial Intelligence in Medicine, vol. 26, no. 1, pp. 1–24.http://ijacsa.thesai.org, http://maya.cs.depaul.edu/~classes/ect584/WEKA/association_rules.html

Khan, Z. N. (2005), "Scholastic achievement of higher secondary students in science stream", Journal of Social Sciences, Vol. 1, No. 2, pp. 80-87.

Kirkby, R. (2002); "WEKA Explorer User Guide for version 3-3-4, University of Weikato, http://www.cs.waikato.ac.nz/~ml/index.html.

Kotsiantis, S., Pierrakeas, C. and Pintelas, P. (2004). Prediction of Student's Performance in Distance Learning Using Machine Learning Techniques, Applied Artificial Intelligence, 18(5) 411-426.

Kumar, V. And Chadha, A. (2011). An Empirical Study of the Applications of Data Mining Techniques in Higher Education *(IJACSA*) International Journal of Advanced Computer Science and Applications,http://ijacsa.thesai.org

Michael, J. A. B. and Gordon S. L. (2004), Data Mining Techniques, 2nd ed., Wiley Publishing Inc., USA, www.cs.waikato.ac.nz/ml/weka (Access 10/06/2014).

Pedro, S., João, M.M., and Carlos, S. (2014). Educational Data Mining: preliminary results at University of Porto, Pulished by Morgan Kaufmann, New Zealand.www.up.pt (Access 20/01/2015).

Witten, I. H. And Frank, E. (2005); "Data Mining Practical Machine Learning Tools and Techniques", Second Edition, Morgan Kaufmann Publishers is an imprint of Elsevier.500 Sansome Street, Suite 400 San Francisco, CA 94111. pp.267pp320 *www.cs.waikato.ac.nz/ml/weka* (Access 15/03/2016).

Worley, C. L. (2007). At-risk students and academic achievement: the relationship between certain selected factors and academic success. A dissertation submitted by Catherine Lynn Worley to Virginia Polytechnic and State University in partial fulfillment of the requirement for the degree of DOCTOR OF PHILOSOPHY in Educational Leadership and Policy Studies Travis Twiford. https://theses.lib.vt.edu/theses/available/etd-06132007-132141/.../DissPDFone.pdf