Selecting Relevant Features for Classifier Optimization

Mvurya Mgala and Audrey Mbogho

Department of Computer Science, University of Cape Town,
ICT4D Research Centre, 7701 Cape Town, South Africa
{mmvurya,ambogho}@cs.uct.ac.za

Abstract. Feature selection is an important data pre-processing step that comes before applying a machine learning algorithm. It removes irrelevant and redundant attributes from the dataset with an aim of improving the algorithm performance. There exist feature selection methods which focus on discovering features that are most suitable. These methods include wrappers, a subroutine of the learning algorithm itself, and filters, which discover features according to heuristics, based on the data characteristics and not tied to a specific algorithm. This paper improves the filter approach by enabling it to select strongly relevant and weakly relevant features and gives room to the re-searcher to decide which of the weakly relevant features to include. This new ap-proach brings clarity and understandability to the feature selection preprocessing step.

Keywords: feature selection, information gain, wrapper, filter, descriptive statistics

1 Introduction

The trend in education is to achieve universal primary education where children are able to complete a full course of primary schooling. In most developing countries, thousands of children complete primary schools with low grades and are forced to drop out of the school system at an age with no skills for meaningful employment. Education stakeholders; education officers, parents and teachers would like to inter-vene to assist such children, the challenge is to identify this children early enough because of the large numbers of pupil. The teachers in many cases are overwhelmed and cannot offer individual attention to such children whose low performance may need more than just extra lessons. It is necessary to explore methods that can discov-er knowledge from pupil data that allow classification of the children into categories such as those that need high intervention and low intervention. This study seeks to determine the most relevant factors that contribute to academic performance for the purpose of developing an academic prediction model.

Many factors contribute to the challenge of applying machine learning to educational data in rural Africa where education is still based on the traditional classroom teaching because of lack of infrastructure. The quality of data is one such challenge, given that data has to be gathered through surveys and hard copy secondary data. Such data will most likely have irrelevant features, noisy and unreliable entries, making knowledge discovery during training difficult. Feature selection can be seen as the process of eliminating as much of the redundant data as possible so as to remain with an optimum subset of features [1]. Algorithms that select features as preprocessing before learning are categorized as wrappers [5]; they employ a statistical subroutine such as cross validation and are embedded in the learning algorithm. The approach is useful except for the fact that the process is very slow because the learning algorithm has to loop many times.

The other approach is called filters [5]; features are filtered out independent of any learning algorithm, usually before learning commences. Filters have proved to be quicker than wrappers and can therefore be applied to large data sets with many features. One other advantage they have is that they can be used with any algorithm unlike the wrappers which have to be re-run when one is changing algorithms.

This paper presents an enhanced filter approach to feature selection by combining the information gain approach with descriptive statistics. In descriptive statistics, boxplots are used to select the features.

The next section discusses related work. In section 3 we describe a filter approach adopted in this work and the descriptive statistics. Section 4 presents experimental results for both the filter approach and the descriptive statistics. The last section concludes and discusses future work.

2 Related Work

A study conducted by Hall [3] on feature selection for discrete and numeric class machine learning, revealed filters to be more practical than wrappers because they are much faster. Experiments conducted using a correlation-based filter algorithm as a pre-processing step for Naïve Bayes, Instance-based learning decision trees, locally weighted regression and model trees show the approach to be an effective feature selector. It reduces data dimensionality by more than sixty percent in most cases without negatively affecting accuracy. Also decision trees and model trees built from the preprocessed data are often significantly smaller.

Yu and Liu [8] conducted a study on efficient feature selection via analysis of rele-vance and redundancy. They demonstrated that feature relevance alone is insuffi-cient for feature selection of high dimensional data. Based on the previous definition of feature relevance by Kohavi et al. [5], features can be classified into strongly rele-vant, weakly relevant and irrelevant. Strong relevance indicates that the feature is always necessary for the optimal subset; it cannot be removed without affecting the original conditional class distribution. Weak relevance suggests that the feature is not necessary but may become necessary for optimal subset at certain conditions. Irrele-vance indicates that the feature is not necessary at all.

An optimal subset therefore should include all strongly relevant features none of irrelevant and a subset of weakly relevant features. Yu and Liu proposed a new framework of efficient feature selection via relevance and redundancy analysis. They devised a feature selection algorithm that demonstrated efficiency and effectiveness in supervised learning.

Another study that used the filter approach is by Kotsiantis et al. [4]. Their results show an improvement in the accuracy of the algorithms after running the experiments without some of the attributes rated as having no influence.

As a way of comparing the two approaches to features selection, we consider a study conducted using the wrapper approach by Bratu et al. [1]. Their work analyzed the wrapper approach for feature selection with the purpose of boosting classification accuracy. Results show that they were able to reduce the number of attributes considerably by over (50%) which speeded up training and improved classification.

These studies show that there is no universally best feature selection method which produces the highest and most accurate improvement on any dataset. This study proposes a framework of selecting strongly relevant features and some of the marginal (weakly relevant) features, and as a way of saying we agree to the "no-free lunch" theorem of feature selection, we allow the researcher to decide on which weakly relevant features to include.

3 Feature Selection

This section discusses the two techcniques we considered for feature selection, namely, correlation-based feature selection and descriptive statistics.

3.1 Correlation-Based Feature Selection

Correlation [8] is applied widely in machine learning to determine relevance. In this section we describe the correlation based filter approach to feature selection.

There are two types of measure for correlation between random variables: linear and non-linear. In linear correlation, the well-known measure is linear correlation coeffi-cient. However, it is not safe to always assume linear correlation between features in real world. Linear correlation measures may fail to capture correlation measures that are non-linear in nature. Many measures among the non-linear correlation measures are based on the information theoretical concept of entropy. Defined as a measure of the uncertainty of random variables, the entropy of a variable X is defined as:

$$H(X) = -\sum_{i} P(x_i) log_2(P(x_i)) \ . \tag{1}$$

The entropy of variable X after observing another variable Y is defined as:

$$H(X|Y) = -\sum_{j} P(y_i) \sum_{i} P(x_i|y_i) log_2(P(x_i|y_i)) \ . \tag{2}$$

Where $P(x_i)$ are the prior probabilities for all values of X and $P(x_i|y_i)$ is the posterior probabilities of X given the values of Y. The amount by which the entropy of X decreases reflects additional information about X provided by Y and is called information gain [7]. Mitchell [6] defines information gain as a statistical measure that determines how well an attribute separates the training data according to the target classes.

It is expressed as:

$$IG(X|Y) = H(X) - B(X|Y) \ . \tag{3}$$

According to this measure a feature Y is regarded more correlated to X than to another feature Z, if:

$$IG(X|Y) > IG(Z|Y) \ . \tag{4}$$

This study adopts the information gain measure to determine the features that correlate more and rank them according to equation 4. Fig. 1 shows the ranked features.

3.2 Descriptive Statistics

The boxplots [9] give a summary of the descriptive statistics. The box represents the interquartile range bounded by the data values that correspond to the 25^{th} and 75^{th} quartiles. Fifty percent of the data values fall within this box and its length represents the interquartile range. The line within the box is the median. The whiskers are the largest and the smallest data values that are not outliers. Data values that are between 1.5 and 3 interquartile ranges below or above the 25^{th} or the 75^{th} quartiles are considered outliers and are represented with an open circle. Data values that are more than 3 interquartile ranges below and above the 25^{th} and 75^{th} quartiles are called extreme values and are represented with asterisk. Using the boxplots one can see the median clearly. If the median is positioned towards the lower end of the data, it suggests that the data is positively skewed.

4 Methods

The idea of combining two feature selection methods is tested on data collected for the purpose of predicting the academic performance of primary school pupils in a rural county in Kenya. A total of 2546 records are gathered from 55 primary schools. The database contains pupils previous test marks, personal, family and school related information and the national examination marks. A total of 23 features are gathered through semi-structured interviews with education officers and head teachers and from literature as possible causes of low academic performance. These features are: total test marks, sex, religion, age, distance to school, pupil absenteeism, study time, pupil discipline, command in speaking

English, pupil education attitude, pupil motivation, parent encouragement, parents stability, family finance ability, parents education qualification, family size, parents involvement, community involvement, teacher attitude, teacher commitment, teacher absenteeism, school facilities and teacher shortage. The information gain algorithm in the Weka machine learning environment [2] is adopted for part one of the experiments.

The results of the ranked features are illustrated in Fig. 1. Features are ranked according to equation 4, and those that have a high information gain are selected as the optimum subset.

Part two of the experiment involves selecting the features using boxplots. They are created from the same dataset using a statistical application as discussed in section 3.2 above. The results section presents a detailed explanation of how to detect correlations.

Data is digitized using a statistical package. As part of pre-processing, records with missing test marks or final examination marks are deleted. Records only missing some pupil response values are filled, noisy data is removed, spelling mistakes and wrong entries are corrected. Final examination marks columns that had both numbers and letters in the same cell are separated. Table 1 shows all the features gathered, it is the initial stage in preprocessing where pupil responses are coded into digits.

5 Results

This section describes the results obtained after applying the two feature selection approaches.

5.1 Information Gain Feature Selection

Fig. 1 shows a chart of all the features of the dataset that are fed into the feature selection algorithm with their corresponding information gains. The larger the value of information gain, the more strongly relevant the feature is to the training. The fea-tures as given by the information gain algorithm are; test scores (0.21653), pupil sex (0.02252), shortage of teachers (0.01807), pupil motivation (0.01613), family income (0.0133), pupil age (0.01185), study time (0.0108), teacher attitude (0.00972), pupil absenteeism (0.00725), teacher commitment (0.00612), parents encouragement (0.00611), pupil education attitude (0.0045), School facilities (0.00895), Command to speak English (0.008), Distance to school (0.00584), Pupil discipline (0.00584). As seen in Figure 1 test score gives the longest bar because it is overly co-related with the final exam mark.

6 Descriptive Statistics Feature Selection

Figures 2–7 illustrate the various boxplots for each of the selected feature. Different categories in each feature are plotted against a standard scale, the final examination marks (KCPE_TOT). A description of each boxplot is given.

Table 1. Features and their numeric codes

Variable	Description	Domain
AGE	Pupil's age	normal:1, overage: 2
SEX	Pupil's sex	female:1, male: 2
DIST	Distance from home	1, 2, 3, 4, 5 km
ABS	Days absent from school per week	0, 1, 2, 3 times
STUD_T	Time to study at home	0, 1, 2, 3 hours
DISPL	Pupil disciplined how often	1, 2, 3 or more, 4: very often
COM_ENG	Pupil's command of English	speak local language:1, uncertain:2, speak English always:3
PUP_M	Pupil motivated?	motivated:1, neutral:2, not motivated:3
P_ENC	Parent encouragement	encouraging:1, neutral:2, not encouraging:3
P_ATT	Pupil education attitude	positive:1, neutral:2, negative:3
F_HARM	Parents' state of harmony	yes:1, neutral:2, no:3
F_FIN	Parent can pay secondary school fees	yes:1, neutral:2, no:3
PQ	Parent qualification	degree:4, diploma:3, secondary:2, primary:1, none:0
F_SIZ	Family size	3-5:1, 6-10:2, 11 or more:3
P_PART	Parent participates in educ.	yes:1, neutral:2, no:3
C_PART	Community participation	yes:1, neutral:2, no:3
T_ATT	Teacher attitute toward pupils	positive:1, neutral:2, negative:3
T_COMM	Teacher committed to teaching	yes:1, neutral:2, no:3
T_ABS	Teacher absent	never:1, neutral:2, always:3
S_FAC	Lack of school facilities	inadequate:1, neutral:2, sufficient:3
L_TEAC	Lack of teachers	inadequate:1, neutral:2, sufficient:3
T_MARKS	Test scores	400-500:1, 350-399: 2, 300-349: 3, 250-299: 4, 200-249: 5, 0-199:6

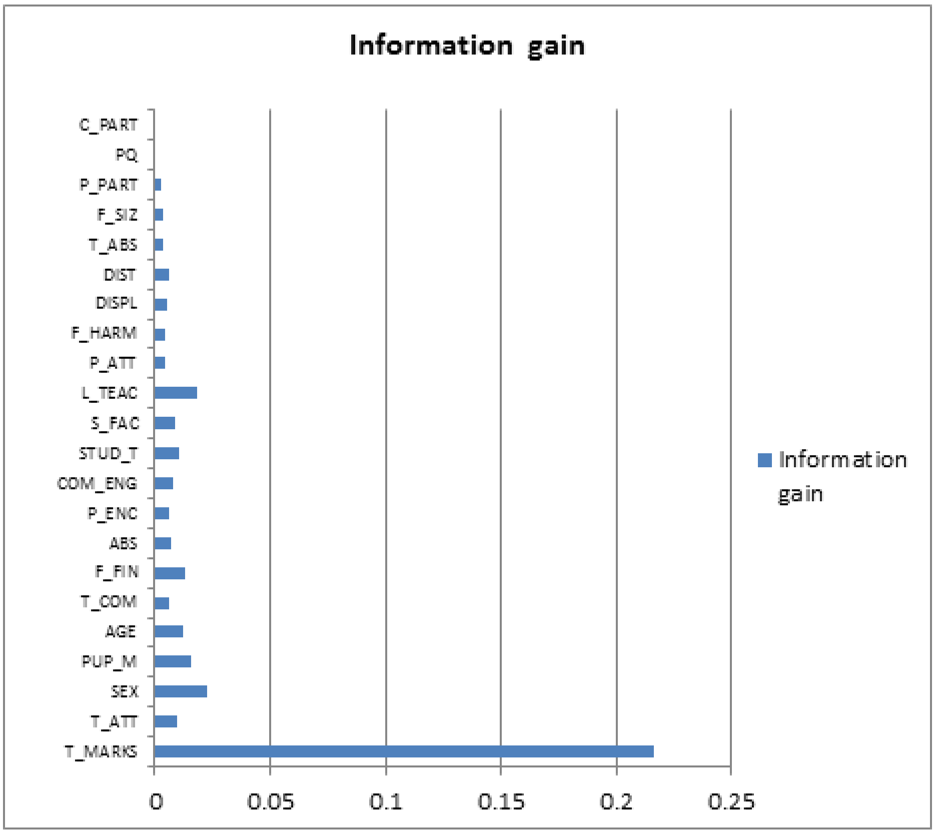

Fig. 1. Information gain for the features

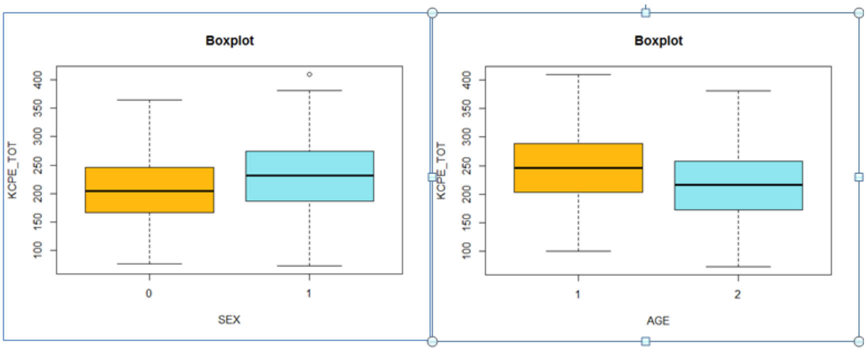

Fig. 2. Final marks against sex and age

The left half of Fig. 2 is a boxplot showing the distribution of the boys (1) and the girls (0). The plot shows that the boys have a higher median of the total score, suggesting sex co-relates with total score.The right half of Fig. 2 is a plot of the pupils ages, normal age (1) and overage (2); those with normal age obtain a higher medium of the total score. This suggests age co-relates with total score.

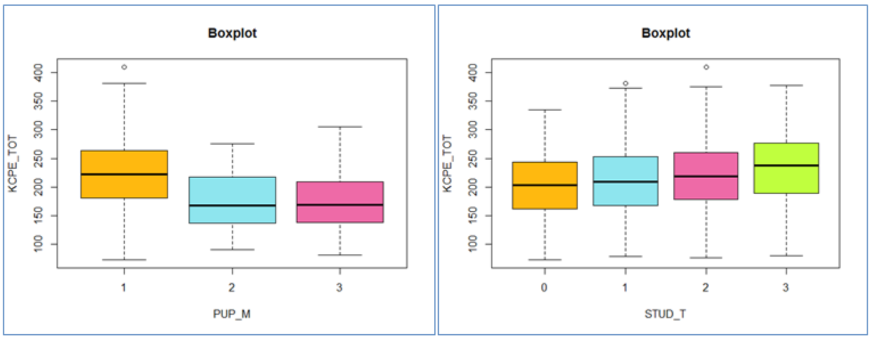

Fig. 3. Final marks against pupil motivation and study time

The left half of Fig. 3 is a boxplot of the pupil motivation on the total score scale. The median score is higher for pupils with high motivation (1). Suggesting pupils motivation co-relates with total score.The right half of Fig. 3 is a boxplot of study time on the total score scale. It is noticed that the median score increases as the amount of study time. This suggests there is a co-relation between study time and total score.

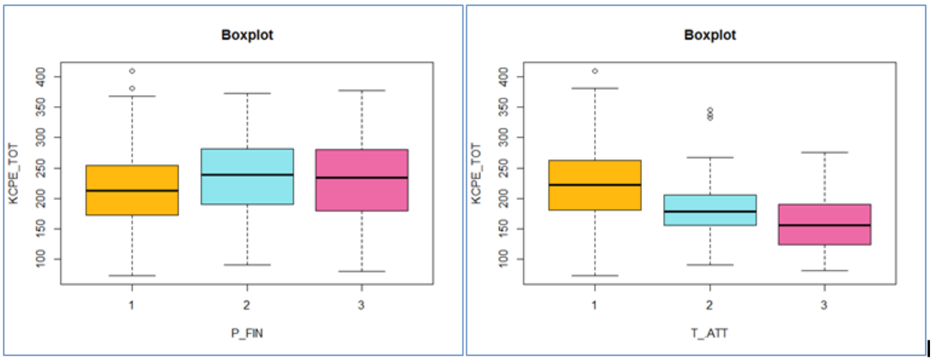

Fig. 4. Final marks against family financial ability and teacher attitude

The left half of Fig. 4 is a boxplot of parents financial ability and the total score. Financial ability (1) does not suggest any co-relation with total score. The poor families (3) however could act as a motivation to work harder. The neutral group (2) could fall either side. On the right is a boxplot of teacher attitude and the total score. Good attitude (1) corresponds to a higher median score, suggesting, there exists a co-relation between teacher attitude and total score.

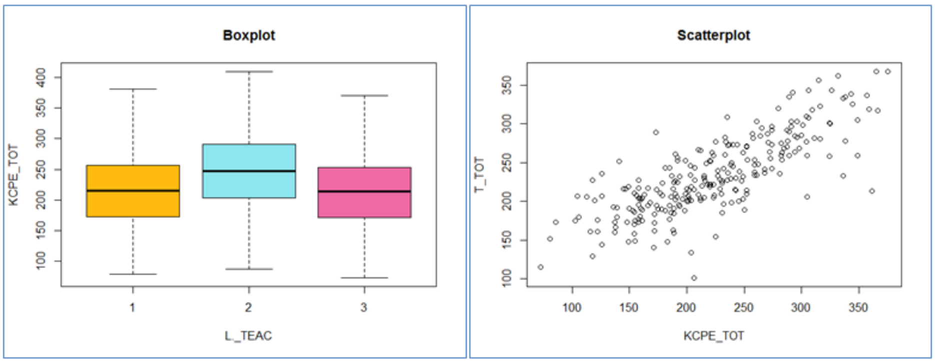

Fig. 5. Final marks against teacher shortage and test marks

Fig. 5 shows a boxplot of teacher shortage and total score. Shortage of teachers (1) and no shortage (3) seem to have the same score median, suggesting there is a no co-relation.The right side is a scatter plot of total score and the test score for three previous years. The plots show a co-relation between the two.

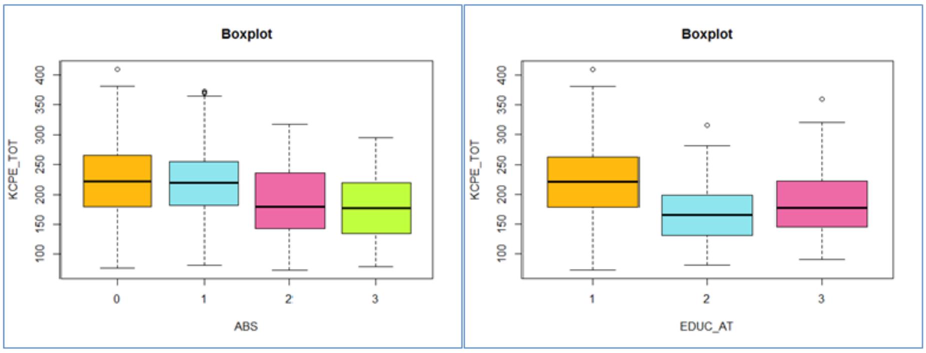

Fig. 6. Final marks against pubil absenteeism and pupil attitude

Fig. 6 is a boxplot of pupil absence from school and total score. Zero days absent (0) and one day absent (1) indicate a higher score median, suggesting a co-relation between absence from school with total score. As seen, the score median decreases as days absent increase. The riht half is a boxplot of pupils attitude towards education and total score. A good attitude (1) corresponds to a higher total score median, showing a correlation between these two variables.

Fig. 7 shows a boxplot of parents encouragement and total score pupil who are encouraged (1) have a higher total score median, suggesting parents encouragement co-relates with total score.Fig. 7also shows a boxplot of teacher commitment and total score. Commitment of teachers (1) indicates a higher median total score, implying a co-relation exists.

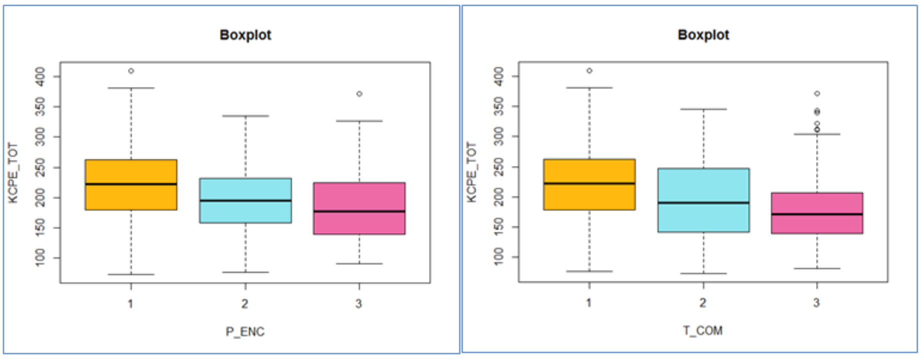

Fig. 7. Final marks against parental encouragement and teacher commitment

6.1 Optimum Subset Verification

Table 2 shows the results of experiments carried out using the two different subsets, 8 features obtained using the information gain approach; total test marks, sex, age, pupil motivation, study time, family finances, teacher attitude and shortage of teachers. Ten features from descriptive statistics; total test marks, sex, age, pupil motivation, study time, teacher attitude, pupil absenteeism, pupil education attitude, parents encouragement and teacher commitment. The subsets are input into the algorithms in turn to obtain the percentage accuracy by each algorithm to determine which of the two an optimum dataset is. The conclusion discusses the finding.

7 Conclustion

Data preprocessing is known to improve the efficiency and effectiveness of learning algorithms. Combining techniques of feature selection has proved to be a better ap-proach to confirm the selected optimum subset. Usually, an optimum subset is a combination of strongly relevant features and some weakly relevant features. The challenge is to identify which weakly relevant features to include given that the irrele-vant features are easily eliminated. This study used both

Table 2. Comparison of subsets

Algorithm	Information Gain Subset Accuracy (%)	Descriptive Statistics Subset Accuracy
LWL	72.8772	73.7016
RepTree	76.2984	75.6389
Logistic	76.7931	76.5045
J48	76.2984	75.7214
Random Forest	74.7321	74.0725
Bayes Net	75.4740	76.5458
SMO	74.9794	74.9794
Average	75.3504	75.3092

information gain and de-scriptive statistics approaches to select the optimum subset features. Information gain approach selected 8 features out of a total of 22 features, namely; total test marks, sex, age, shortage of teachers, pupil motivation, family finances, study time and teacher attitude. The descriptive statistics approach selected 10 out of 22, name-ly; total test marks, sex, age, pupil motivation, study time, teacher attitude, pupil absenteeism, pupil education attitude, parents encouragement, and teacher commitment.

Experiments carried out using the two subsets on 7 different algorithms reveal a marginal difference on the average percentage prediction accuracy. Information gain approach gave 75.3504% while descriptive statistics gave 75.3092%. We conclude that the features that are shared by both subsets are the strongly relevant features, these are; total test marks, sex, age, pupil motivation, study time and teacher atti-tude. The other features appearing in either subset are weakly relevant, these are; shortage of teachers, family finances, pupil absenteeism, pupil education attitude, parents encouragement and teacher commitment. Any of these can be added to the list since they only weakly influence learning. These findings provide a foundation for further work on enhancing the effectiveness of an algorithm for predicting academic performance of primary school pupils in rural Africa.

8 Acknowledgements

Our thanks go to the Hasso Plattner Institute for funding this work, to the university of Cape Town for providing a conducive environment for doing this research, and to the Kenyan NACOSTI for allowing us to collect data.

References

1. Bratu, C. V., Muresan, T., Potolea, R.: Improving Classification Accuracy through Feature Selection. In: Proceedings of the 4th IEEE International Conference on

Intelligent Computer Communication and Processing, pp. 25–32, IEEE Press, New York (2008)

2. Garner, S. R.: Weka, The Waikato Environment for Knowledge Analysis. In: Proceedings of the New Zealand Computer Science Research Students Conference, pp. 57–64, New Zealand (1995)
3. Hall, M. A.: Feature Selection for Discrete and Numeric Class Machine Learning.Technical report, University of Waikato, Department of Computer Science, New Zealand (1999)
4. Kotsiantis, S. Pierrakeas, C. and Pintelas, P.: Efficiency of Machine Learning Techniques in Predicting Students Performance in Distance Learning Systems. Applied Artificial Intelligence 18(5), 411–426 (2004)
5. Kohavi, R., John, G., Long, R., Manley, D., Pfleger, K.: MLC++: A Machine Learning Library in C++. In: Proceedings of the Sixth International Conference on Tools with Artificial Intelligence, pp. 740–743, IEEE Press, New York (1994)
6. Mitchell, T. M.: Machine Learning. Machine Learning. McGraw-Hill, Inc., New York, NY (1997)
7. Quinlan, J. R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)
8. Yu, L., Liu, H.: Efficient Feature Selection via Analysis of Relevance and Redundancy. Journal of Machine Learning Research 5, 1205–1224 (2004)
9. Kessler, F. C.: Understanding Descriptive Statistics, http://nationalatlas.gov/articles/mapping/a_statistics.html