Journal of Electrical Engineering 6 (2018) 229-237
doi: 10.17265/2328-2223/2018.04.006

The Design of Predictive Model for the Academic Performance of Students at University Based on Machine Learning

Barnabas Ndlovu Gatsheni and Olga Ngala Katambwa

Department of Applied information Systems, School of Consumer Intelligence and Information Systems, University of Johannesburg, Auckland Park 2006, South Africa

Abstract: Students in South African Universities come from different socio-cultural backgrounds, countries and high schools. This suggests that these students have different experiences which impact on their levels of grasping information in class as they potentially use different lenses on tuition. The current practice in Universities in contributing to the academic performance of students includes the use of tutors, the use of mobile devices for first year students, use of student assistants and the use of different feedback measures. What is problematic about the current practice is that students are quitting university in high numbers. In this study, knowledge has been drawn from data through the use of machine learning algorithms. Bayesian networks, support vector machines (SVMs) and decision trees algorithms were used individually in this work to construct predictive models for the academic performance of students. The best model was constructed using SVM and it gave a prediction of 72.87% and a prediction cost of 139. The model does predict the performance of students in advance of the year-end examinations outcome. The results suggest that South African Universities must recognize the diversity in student population and thus provide students with better support and equip them with the necessary knowledge that will enable them to tap into their full potential and thus enhance their skills.

Key words: Machine learning, Bayesian networks, support vector machines, decision trees and predictive model.

1. Introduction

Universities generally make use of different ways in order to enhance the academic performance of students. Some of these practices include the use of tutors, use of mobile devices for first year students, use of student assistants and the feedback measures. These practices are not effective enough for the intended purpose as evidenced by the number of students who are quitting tertiary education. The 2014 annual report on one particular university faculty states low attendance of tutorial sessions, insufficient preparation for lectures and tutorials; and minimal support from lecturers tended to lead to the poor performance of students.

South African Universities accommodate students from different socio-cultural backgrounds, countries and high schools, thereby suggesting that all these students have different learning experiences and levels of education. These universities must recognize this diversity in order to provide students with better support and equip them with the appropriate knowledge that will help students to realize their full potential with regard to academic performance. A university is mandated to allow students to thrive and benefit from cultural differences. However, if students feel intimidated it could negatively affect their behavior, their confidence, their participation in class and academic performance. For instance, a number of students could be passive in class and not interrogate knowledge or express themselves owing to the fear of being mocked by their peers or having the question ignored by the lecturer [12]. This study therefore developed a predictive model for the academic performance of students, in light of the stated diversity.

Corresponding author: Barnabas Ndlovu Gatsheni, Dr., research fields: computational intelligence.

The factors identified include commuting and non-commuting students, lecturer's competency, not owning desktop computers, tablets and laptops, language and type of high school of attended (private or public), etc.

A student who is unable to afford the taxi commuter fare to university resorts to walking, which may result in them being exhausted, late for lectures and hence a lack of focus. Such students do skip some lectures. Nowadays, most learning activities can be done electronically. The language of instruction at the University of Johannesburg is English. There is a large contingent of students coming from francophone and also Portuguese speaking countries and hence language can be a problem. In addition, to most South Africans English is a second language. It is against these diverse university conditions that this study conceptualized its focus. In this study a predictive model was developed from machine learning algorithms using attributes that include lecture attendance, self-study, lecturer competency, average matriculation results (high school leaving results) and first semester university results. This predictive model does predict the performance of students well in advance of the year end examinations outcome.

2. Related Work

The results of Hansen [4] show that performance of students has been affected by learning abilities, race and gender among others. Hijazi and Naqvi [6] focused on the factors affecting the performance of pupils from 3rd and 4th grade in a private college in Pakistan and have demonstrated that empowering mothers on education can lead to a better educated society. Martinez and Gomez [11] using clustering, association generators and decision trees showed that the profile of the students considered low academic performers corresponds to 37.73% of the student population and the one considered average performers corresponded to 36.44%. The profile of the students considered high academic performers, corresponding to 25.78%. What is problematic with their result is that their total statistic does not add to 1.

Ruby and David [16] predicted the academic students' performance using J48, ID3, REP Tree, NB Tree, BF Tree, Decision Table, MLP, Bayes net and simplecart. The MLP gave the best result of 74.8% accurate prediction while the ID3 was 73%. The remaining algorithms gave a poorer performance.

Aziz, Ismail and Ahmad [1] presented a framework based on Naïve Bayes for predicting first year students' performance.

Sembiring et al. [17] focused on the prediction of academic performance through the use of smooth support vector machine (SVM) for classification and kernel k-means for clustering.

Hashim et al. [5] used C4.5 to determine the academic performance of students in mathematics.

Kurniawan and Halim [8] used a star schema to model the data warehouse in order to support the data mining analysis. The results presented a model that identified the performance of students before the end of the semester.

3. Methodology

This research used a quantitative approach. A survey design was used and a questionnaire to collect data on attributes that potentially affect the academic performance of first year students at the University of Johannesburg. Data and machine learning techniques were used to construct a prediction model.

3.1 Candidate Techniques for Academic Performance Prediction

The candidate techniques for this paper are the different machine learning algorithms classified under supervised Guerra et al. [3]. The supervised prediction includes algorithms that use a priori-defined class.

3.1.1 Artificial Neural Network

An artificial neural network (ANN) algorithm provides a general technique for learning real, discrete and vector value for a given problem and for

interpretation of complex real-world sensor data [14]. The multi-layer perceptron (MLP) is an ANN that consists of perceptrons and sigmoid units [15] to produce solution for a specific tasks [9]. Its input is transformed through the use hidden layer units shown in Fig. 1. The perceptron takes a vector of real-valued inputs, calculates a linear combination of these inputs using Eq. (1). If the outcome is greater than some threshold, the output is 1 otherwise it is -1.

$$0(x1 \dots \dots \dots \dots xn) = \{\frac{1 \; if \; \omega o + \omega 1x1 + \omega 2x2 + \cdots . + \omega nxn > 0}{-1 \; otherwise}\}$$

$$E = \frac{1}{2} \sum_{examples} (t-o)^2 \quad (1)$$

The weights w_i associated with input x_i are updated as in Eqs. (2) and (3).

$$\Delta w_i = \eta (t-o) x_i \quad (2)$$

t is the target output for the current training instance, o is the generated output, η is the learning rate and it determines the step size in the gradient descent search and influences the extent to which weights are changed at each step. During training one of the parameter's values were changed whilst other parameters were held constant. The process was repeated for each parameter until all of them had been used.

When $(t-o)=0$, the training example is correctly classified, a situation that must be avoided as it corresponds to a rot learner. The objective is to attain the direction of steepest descent along the error surface by computing the derivative of E with respecting to each component of vector $\vec{w}$ in Eq. (3). A negative E of the $\vec{w}$ moves the weight vector in the direction that decreases E.

$$w_i \leftarrow w_i + \Delta w_i, \text{ where } \Delta w_i = -\eta \frac{\partial E}{\partial w_i} \quad (3)$$

Thus steepest descent is achieved by altering each component w_i of $\vec{w}$ in proportion to $\frac{\partial E}{\partial w_i}$. This process is repeated, iterating through training instances until the weight that minimises E is found and thus all instances are classified correctly. Although the error-weight surface can have multiple local minima, when using the gradient descent the MLP does converge to acceptable solutions (hypotheses) [4] when used with the correct η and the correct momentum α. Thus α tends to keep the search out of the small

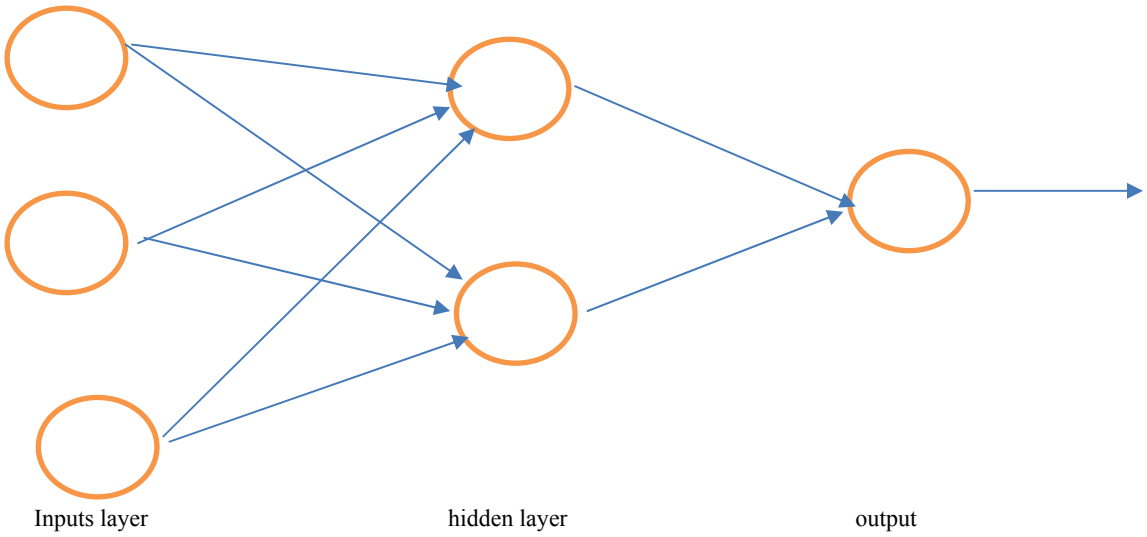


Fig. 1 Shows a multi-layer perceptron.

local minima or flat regions. The α term also has the effect of increasing the step size of the search in regions where the gradient is unchanging, resulting in faster convergence. A large η can result in the search overshooting the global minimum. A very small η will land the search on the global minimum (the desired solution), at a cost of an extremely long time. Despite all this, backpropagation has been found to be an effective function approximation in many real world applications.

3.1.2 Bayesian Networks (BN)

BN is a statistical model based on graphical probability and represents a set of events and their causal relations through the use of a directed acyclic graph (DAG). What makes BN unique is the way that it models a complex problem that is categorized by direct or indirect effects and uncertainty [10]. In BN, a node characterizes a random variable while an arc designates the relation between nodes. Nodes at the extremities of the arcs are called children and nodes at the heads of the arcs are called parents. A BN shown in Fig. 2 starts first with a selection of random variables.

A network is formed by joining a pair of nodes utilizing directed acyclic arcs based on relationships. The probability distribution of every single node is specified in its own probability table [7]. The table for a node without a parent (s) is called the prior probability table whilst for those with a parent (s) is a conditional probability table.

3.1.3 Naïve Bayes (NB)

NB algorithm based on Bayes' theorem works well [19] with a small training data set to estimate the essential parameters [2]. NB makes a decision by choosing the class which has the highest likelihood. To estimate each of the prior probabilities $P(v_j)$ count the frequency with which each target value v_j occurs in the training data. Because of the assumption that the attribute values of an instance ($a_1, a_2 \ldots\ldots a_n$) are conditionally independent given the target value (v_j) of the new instance, the probability of observing $a_1, a_2 \ldots\ldots a_n$ is the product of the probabilities for the individual attributes as shown in Eq. (4).

$$\text{NB}: v_{NB} = \frac{\arg\max}{v_j \in V} P(v_j) \prod_i P(a_i \mid v_i) \quad (4)$$

where, The $P(a_i \mid v_i)$ is the probability of a_j given that a_j belongs to class v_j. The $P(a_i \mid v_i)$ terms estimated from the training data is the product of the number of distinct attribute values and target values. With v_{NB} there is no explicit search through the space of possible values that can be assigned to $P(v_j)$ and $P(a_i \mid v_i)$ terms when classifying the novel instance.

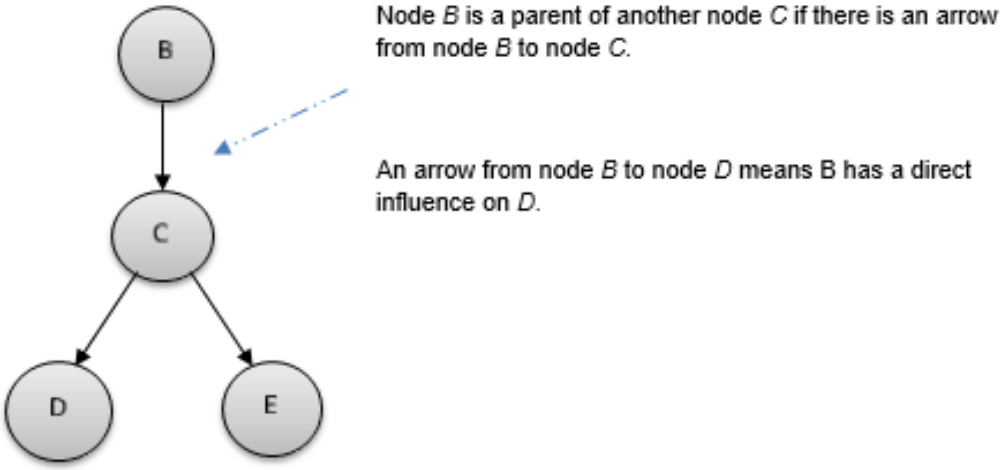


Fig. 2 A Bayesian network.

3.1.4 Decision Trees

The structure of a decision tree model is similar to a natural tree in terms of branches, leaves and roots but it is an inverted tree with the root at the top. Classifications are represented by leaves whereas branches characterize unions of features that lead to classifications. Thus a series of nodes and branches are terminated by a leaf. The class of an instance is defined by tracing the path of nodes and branches to the terminating leaf [18].

$$G(S,A) = entrpy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} entrpy(S_v)$$

A consists of $+p$ and $-p$ instances. S is the sample.

The information gain measure is used to determine the attribute that must be the root of the tree and succeeding attributes.

The advantage is that the tree can easily be changed into a set of production rules which a human being can understand. It can classify categorical and numerical data and there is no need of having a-priori assumption on the nature of data. The ID3 algorithm developed by J. Ross Quinlan in 1993 is based on the concept learning [13] and applies a top-down search via the space of possible decision trees. It starts by selecting an attribute that will be tested at the root of the tree. J48 and C4.5 decision tree algorithms are variations of ID3. In this work J48 is used.

3.1.5 SVMs

SVMs' SMO was used for the experiment as follows: the kernel transformed data that are linearly non-separable in a lower-dimensional feature space into data that are linearly separable in a higher dimensional feature space and this process is called the Kernel Trick. Kernel functions are based on calculating the inner products of 2 vectors. The hypothesis space for SVM is a subset of all hyperplanes of the form $f(x) = w \cdot x + b$ defined in some space. The kernel defines a dot product in that space. An optimal hyperplane was defined to maximize the margin that separates the two classes as shown in Eq. (5).

$$\omega, \varepsilon \{\tfrac{1}{2} \| \omega \|^2 + C \sum_i \varepsilon_i\} \quad (5)$$

ε is the slack is variable, C is the regularization parameter and ω is the weight. In the experiment for this paper model parameters for SVM were determined by varying one parameter at a time and results noted. The process was repeated until all the parameters were varied. To make predictions for a new input x we compute

$$g(\mathbf{x}) = \text{sgn}(\mathbf{w}.\mathbf{x} + w_0)$$
$$= \text{sgn}\left(\sum_{i=1}^{n} \alpha_i t_i (\mathbf{x}_i.\mathbf{x}) + w_0\right)$$

x enters this expression only on terms of the inner product $\mathbf{x}.\mathbf{x}_i$. Variables used were average matric, self-study, competent lecturers, lecture's attendance, first semester average percentage and the class called performance. SVM is a sparse kernel and sparsity is in the solution which is based on data points (called support vectors) that lie on the margins and thus SVM can provide an optimal solution from a small sample size and it is a convex optimization problem meaning that any local solution is also a global optimum.

SVM for the non-separable case

Data points for which $0 < \xi \leq 1$, lie inside the margin but on the correct side of the decision boundary. Data points for which $\xi > 1$, lie on the wrong side and are misclassified. For $\xi = 0$ data points lie on the margin or on the correct side. The penalty for misclassifications increases linearly with ξ. ξ is said to relax the hard margin constraint to give a soft margin by allowing some of the data points to be on the wrong side of the margin boundary and are thus misclassified but with a penalty C that increases with the distance from that boundary.

$C > 0$ controls the trade-off between the ξ penalty and the margin. SVM for k > 2 classes, C was found by using cross-validation. Predictions are linear combinations of kernel functions that are centred on training data points. The complexity of SVM was

controlled by using the radial basis kernel function. A high *C* means a higher penalty to errors. If *C* is made small, the outlier points are de-emphasized. Minimizing *W* for the linear case maximizes the margin. SVM has direct methods of limiting over-fitting. ANN has good performance in many applications, however, it cannot control generalization ability as it cannot control empirical risk and confidence interval.

3.1.6 Feature Selection

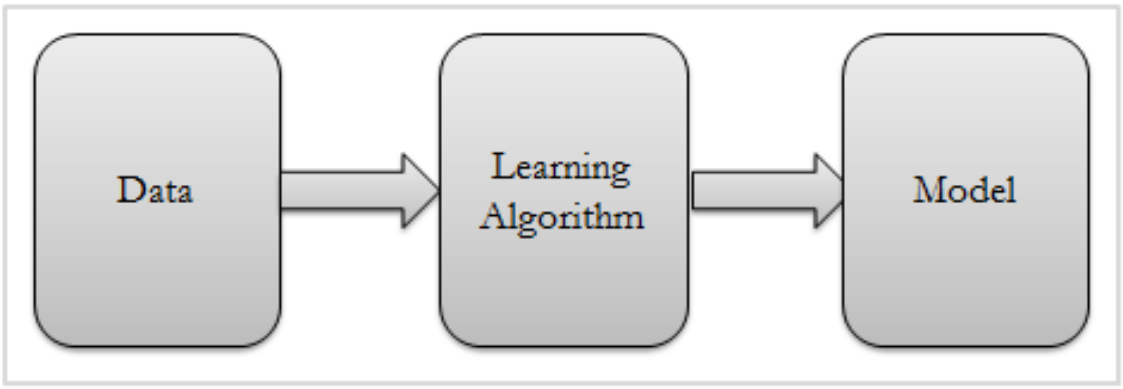
Feature selection was used for choosing a subset of highly discriminative features for creating robust learning models. The architecture of feature selection involves the use of a set of features as a training set representative of positive and negative examples of the classes for which classification is required. A search procedure was used to search the space of all subsets of the specified feature group. The performance of each one of the designated feature subset is measured using an evaluation function. The subset of features that achieve the highest classification accuracy are chosen.

4. Experimentation

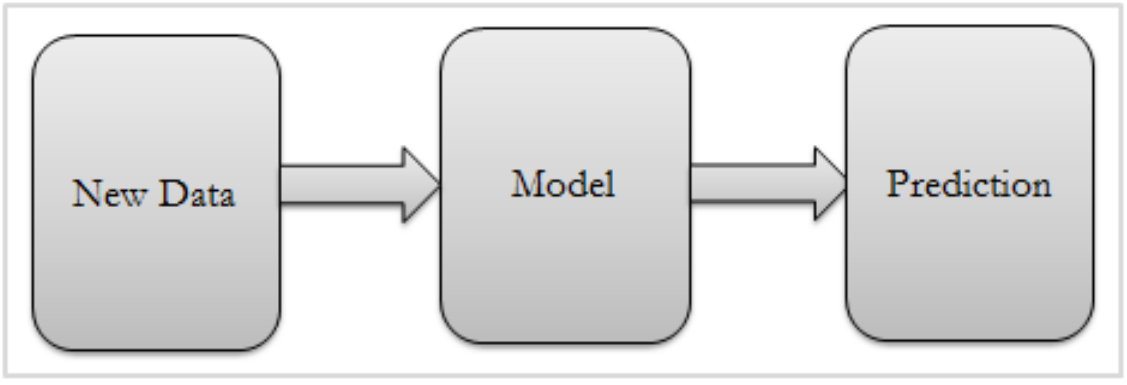
Data Preprocessing

The experiments were conducted using a data set of 247 instances consisting of 5 attributes namely average matriculation results, self-study, competent lecturer, lecture's attendance and first semester's average results. SPSS was used to determine which attributes were significant and could be used to build the model shown in Fig. 3. Machine learning algorithms housed in WEKA were used to construct the student academic performance predictive model. Different models were built but for all of them the accuracy for prediction (percentage of correct prediction) and the root mean square error (RMSE defined by Eq. 6) were used to measure the performance of the model and also to evaluate the model respectively.

$$\sqrt{\sum \frac{(Xi-yi)^2}{M}} \quad (6)$$



(a)



(b)

Fig. 3 (a) constructing a model. (b) using a model for prediction.

4.1 Experiment Using SPSS

Regression analysis was used to provide the relationship between the independent and dependent variables. Table 1 shows that the model is significant because the *p*-value is less than 0.05 and was considered in the selection of attributes to use for the model.

Results in Table 2 mean that the average matric attribute and competent lecturer are important for our model. If a *p*-value is inferior to 0.05, this means that there is a 95% chance that the relationship between variables has a good predictive analysis.

Table 3 means that the model is significant but the independent variables used for the model are not significant enough except for the average matric attribute. The significant ones were used to construct the student performance predictive model.

4.2 Experiment Using WEKA

Student performance prediction models were constructed using algorithms that include SVM, BN, NB, Decision trees' J48 and MLP. A dataset of 247 instances with 5 different attributes was used.

Table 3 is a collection of highest percentage of correctly classified instances for each algorithm. Results Table 4 mean that NB academic prediction performance model with an accuracy of 75.30% and a RMSE of 0.3555 is considered to be adequate for this study therefore recommended for prediction of academic performance.

Table 1 ANOVA.

Model		Sum of squares	df	Mean square	F	Sig.
1	Regression	23.412	2	11.706	6.282	.002
	Residual	454.637	244	1.863		
	Total	478.049	246			

Table 2 Coeffs (multi-regression analysis).

Model		Sig. Beta	
1	(Constant)		.000
	Competent lecturers	.144	.023
	Average matric	-.189	.003

Table 3 Coefficients (from multi-regression analysis).

Model		Sig.
1	(Constant)	.000
	Average matric	.001
	Library resources	.234
	Laptop	.444
	Competent lecturers	.023
	Self-study	.030

Table 4 Algorithms comparison.

Algorithms	Correctly classified	RMSE
Bayesian networks	72.87%	0.3698
Naïve Bayes	75.30%	0.3555
MLP	71.66%	0.3877
J48	72.47%	0.3813
SVM	72.87%	0.3822

Table 5 Confusion matrix for SVM.

		Predicted class		
		Good	Average	Poor
Actual class	Good	45	40	0
	Average	8	121	3
	Poor	2	14	14

Table 6 Loss matrix.

	Good	Average	Poor
Good	0	3	1
Average	3	0	1
Poor	2	2	0

The constructed model comes with a confusion matrix (shown in Table 5) that shows the number of instances that were correctly and incorrectly predicted respectively. The selection of the best model was based on the model's performance in prediction and also on the cost associated with prediction which is computed in the next section.

Post-processing

The models were evaluated using a loss matrix in Table 6 in combination with a confusion matrix in Table 5.

The same loss matrix was used for all the models. The cost of a model is calculated as follows:

Cost = Confusion matrix *x* Loss matrix

The cost of the model built with Bayesian networks is 181 and the cost of NB is 163. The cost of MLP is 165, the cost of J48 is 180 and finally the cost of SVM is 139. The SVM turned to be the best model because it has the lowest cost of prediction.

5. Discussion

The results mean using attributes that include lecture attendance, self-study, lecturers' competency, average matriculation (high school) results, type of high school the student attended and first semester university results for first year students can potentially determine the academic performance of a first year University student in advance (well before the yearend examinations). We had expected the NB to perform well as it has done so in related work. Previous work mentioned in this paper has not used features such as high school leaving results, commuting and non-commuting students attributes. The implications of the results are the university can put measures in place to mitigate the problem of student performance. Most comprehensive universities can benefit from the model.

6. Conclusions

This study identified variables that influence the academic performance of students which were used for constructing a student academic performance prediction model. This model can tell in a case of an intelligent student, coming from a poor performing high school at admission at university and if this student is given a good lecturer, attends classes regularly, spends more time doing self-study, that he/she will perform to the standard if not come out at the top of the class by the end of academic year. The model does predict the performance of students well in advance of the year end examinations outcome. The SVM model showed an accuracy of 72.87% with a cost of 139. This model was chosen because its prediction cost is lower than that of other models. The model will help the university to increase the graduation rate through the identification of the specific factor (s) affecting the performance of a specific student before the end of the academic year.

References

[1] Aziz, A. A., Ismail, N. H., and Ahmad, F. 2013. "Mining Students' Academic Performance." *Journal of Theoretical*

and Applied Information Technology 53 (3).

[2] Bramer, M. 2007. *Principles of Data Mining.* Springer London, 24-30.

[3] Guerra, L., McGarry, L. M., Robles, V., Bielza, C., Larranaga, P., and Yuste, R. 2011. "Comparison between Supervised and Unsupervised Classifications of Neuronal Cell Types: A Case Study." *Dev Neurobiol.* 71 (1):71-82.

[4] Hansen, J. B. 2000. "Student Performance and Student Growth as Measure of Success: A Evaluator'S Perspective." Paper Presented at Annual Meeting of the American Educational Research Association New Orleans, Louisiana, April 25, 2000.

[5] Hashim, H., Talab, A. A., Satty, A., and Talab, S. A. 2015. "Data Mining Methodologies to Study Student's Academic Performance Using the C4.5 Algorithm." *International Journal on Computational Sciences & Applications (IJCSA)* 5 (2).

[6] Hijazi, S. T., and Navqi, S. M. M. R. 2006. "Factors Affecting Students' Performance." *Bangladesh E-Journal of Sociology* 3 (1).

[7] Holland, A., Fathi, M., Abramovici, M., and Neubach, M. 2008. "Competing Fusion for Bayesian Applications." In *Prcceedings of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems,* 378-85.

[8] Kurniawan, Y., and Halim, E. 2013. "Use Data Warehouse and Data Mining to Predict Student Academic Performance in Schools: A Case Study (Perspective Application and Benefits)." In *proceedings of the IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE).*

[9] Lee, D., and Yeo, H. 2015. "A Study on the Rear-end Collision Warning System by Considering Different Perception-Reaction Time Using Multi-layer Perceptron Neural Network." *IEE Intelligent Vehicles Symposium (IV).*

[10] Lee, C., Song, B., Cho, Y., and Park, Y. 2010. "A Bayesian Belief Network Approach to Operationalization of Multi-scenario Technology Roadmap." *IEEE publication*.

[11] Martinez, D. L., and Gomez, C. E. 2014. "Contributions from Data Mining to Study Academic Performance of Students of a Tertiary Institute." *American Journal of Educational Research* 2 (9): 713-26.

[12] McKaiser, E. 2015. Are you Comfortable at Your University? Discussion on Book Being Home: Race, Culture and Transformation at SA Higher Education.

[13] Mitchell, T. 1997. *Machine Learning.* McGraw Hill.

[14] Mitchell, T. 2006. "The Discipline of Machine Learning." Accessed 12 May 2015. Available at: http://www-cgi.cs.cmu.edu/~tom/pubs/MachineLearningTR.pdf.

[15] Pirooznia, M., Gong, P., Guan, X., Inouye, L. S., Yang, K., Perkins, E. J. and Deng, Y. 2007. "Cloning, Analysis and Functional Annotation of Expressed Sequence Tags from the Earthworm." *BMC Bioinformatics.* 8 (Suppl 7): S7.

[16] Ruby, J., and David, K. 2015. "Analysis of Influencing Factors in Predicting Students Performance Using MLP—A Comparative Study." *International Journal of Innovative Research in Computer and Communication Engineering* 3 (2).

[17] Sembiring, S., Zarlis, M., Hartama, D., Wani, E., and Ramliana, S. 2011. "Prediction of Student Academic Performance by an Application of Data Mining Techniques." *International Conference on Management and Artificial Intelligence IPEDR* (6).

[18] Williams, L. M., Brown, K. J., Palmer, D., Liddell, B. J., Kemp, A. H., Olivieri, G., Peduto, A., and Gordon, E. 2006. "The Mellow Years?: Neural Basis of Improving Emotional Stability over Age." *The Journal of Neuroscience* 26 (24): 6422-30.

[19] Zhang, H. 2004. "The Optimality of Naïve Bayes." Accessed 23 April 2015. Available at: http://www.aaai.org/Papers/FLAIRS/2004/Flairs04-097.pdf.