Psychology assisted Prediction of Academic Performance using Machine Learning

Radhika R Halde, Arti Deshpande, Anjali Mahajan

Abstract—**The psychological state of the student has deep influence on their academic performance is being proved by various studies. The paper demonstrates the impact of student's psychology and their learning and study skills in the predicting their academic performance. The experiment was performed on the real time data collected from final year students. The matriculate and pre-university examination scores, five semester scores along with data on the motivation level, information processing ability and other learning and study skills were taken as input to the model to predict the Cumulative Grade Point Average(CGPA) of the sixth semester. Two machine learning algorithms were used to test the impact of students' psychology on prediction which includes Neural Network for numeric prediction of sixth semester CGPA and Decision Tree for classification of failures in sixth semester. The performance of the models were evaluated using the coefficient of correlation R and the Mean Squared Error. The accuracy of the prediction increases about 4 to 6%. The study reveals that level of motivation in student's life and the way they perceive the information and use the available study materials for the examination, all counts in prediction of their examination performance.**

Keywords—Predictive, Analytics, Psychology, Neural network,Decision Tree,Learning and Study Skills;

I. INTRODUCTION

Predictive analytics is amalgamation of number of statistical modeling techniques, data mining and machine learning which tries to find trends in past data and predicts the future.. It has application in various fields like banks, business, healthcare, marketing, manufacturing industries and academics. Administrators of educational institutes are making use of predictive analytics for taking vital decisions regarding efficient management of student's enrollment, student's retention, maintaining long term relationship with alumni and the recruiters and for knowing the possible percentage of placements in advance. Machine Learning techniques are most popular ones for predictive analytics for they work robustly on large and noisy datasets. Various Machine Learning algorithms are used to predict student's performance. The paper tries to describe the experiment were psychological factors were considered in prediction process to increase accuracy of prediction of student's academic performance.

Radhika R Halde Department of Computer Engineering, Thadomal Shahani Engineering College, Mumbai, IndiaNagpur, India, (radhikah91@gmail.com)

Arti Deshpande, Department of Computer Engineering, G.H. Raisoni College of Engineering, Nagpur, India, (artideshpande75@gmail.com)

Anjali Mahajan, Department of IT, Government Polytechnic, (armahajan@rediffmail.com)

II. LITERATURE REVIEW

The study done in the paper [1] showed the application of Artificial Neural Network for prediction of final year score using the scores of the fundamental subjects in the first and third semesters of Matriculate and Diploma students. Further the relationship between the fundamental subjects and final year performance was investigated using Neural Network and Linear Regression. The accuracy of the Neural Network was found to be more compared to the Linear Regression [2].For depicting the fact that demographic features play important role in prediction of student's academic performance, the experiment was performed using Naive Bayes and Support Vector Machine (SVM) algorithms [3].Along with the academic performance the prediction of placements was done using Placement Prediction System. The Logistic Regression was fed up with all semester marks and demographic information to get the probability of the student's placement [4]. In paper [5],a study was conducted on 200 students to identify the correlation between mental health and academic performance of student. Using the Pearson correlation coefficient, the relation between variables of mental health and academic performance was studied [5].The Support Vector Regression was implemented on the data collected through questionnaires filled by120 undergraduates in Taiwan to show that there is close relation between student's performance and their personality characteristics [6].

III. PROPOSED MODELING APPROACH

The Neural Networks performs with higher accuracy in capturing the nonlinear trends in data and giving numeric prediction [2].It used in current experiment to find the CGPA of individual student in the sixth semester. For classification of students into Pass or Fail class the Decision Tree algorithm is being applied[3].The main aim of the experiment is to demonstrate how student's psychological factors plays vital role in the prediction of pre-final year score.

A. Regression using Neural Network

Neural Network in terms of regression analysis is to approximate functional relationships between dependent and independent covariates. The neural network calculates an output o(x)for given inputs x and current weights. It calculates the function

$$o(\mathbf{x}) = f\left(w_0 + \sum_{i=1}^{n} w_i x_i\right) = f\left(w_0 + \mathbf{w}^T \mathbf{x}\right) \qquad (1)$$

where w0 denotes the intercept, w = (w1, . . . ,wn) the vector consisting of all synaptic weights without the intercept, and x = (x1, . . . , xn) the vector of all covariates. As the main focus is on supervised learning, the error E is calculated and weights are adjusted using the learning algorithm. The error is given as

978-1-5090-0774-5/16/$31.00 © 2016 IEEE

$$E = \frac{1}{2}\sum_{l=1}^{L}\sum_{h=1}^{H}(o_{lh} - y_{lh})^2 \qquad (2)$$

Which measures the difference between predicted and observed output, where l = 1, . . . , L indexes the observations, i.e. given input-output pairs, and h = 1,.. , H the output nodes. The weights are adjusted by the following rule

$$w_k^{(t+1)} = w_k^{(t)} - \eta \cdot \frac{\partial E^{(t)}}{\partial w_k^{(t)}} \qquad (3)$$

in traditional back propagation, where t indexes theiteration steps and k the weights [13] [14].

B. Classifiction using Decision tree

Decision Tree is a classification algorithm that decides whether a specific value should be accepted or rejected, and it provides with the set of the IF-Then rules for transforming present state to future state. The tree structure is used to represent decision tree in which variant types of the nodes are connected by the branches where the topmost node is called as root node and the leaves are called decision node[7][8][9].

C. Learning and Study Skills Inventory(LASSI)

The LASSI [10] is a diagnostic measure that studies the student thought processes and behaviors which influence the learning process. The main aim is on the covert and the overt thoughts that are needed for success in the learning process and can be changed using the educational interventions. It has ten scales through which the thought process can be captured which include Attitude, Motivation, Time Management, Anxiety, Concentration, Information Processing, Selecting Main Ideas, Study Aids, Self Testing, and Test Strategies.
Online questionnaire forms are given to individual students and based on the responses they provide the score for each scale is found. Questionnaire based on LASSI scales plays leading role in this experiment to know what is level of motivation in student's life, how do they make use of the study materials, in what way do they understand the things taught in the class, how do they make use of the available study material for the examination. All these things are considered in PHASE 2for prediction of the CGPA.

IV. Data Gathering and Compilation

Real time data is collected by surveying 150 students of Thadomal Shahani Engineering College [11] from Computer Department by making them fill the online questionnaire. The questionnaire consists of the 98 questions regarding the previous year scores, Scales of LASSI and personality traits. The data was compiled in the CSV file separately for two phases of implementation. The compiled data consisted of first and last name of student, registration number, matriculate score, pre-university examination score, CGPA of six semesters, score of LASSI scales, personality type. Data is then normalized using Min Max Algorithm [12]. The splitting of the data in training set and the test set is done in the ratio of 70:30.

V. Methodology

The experiment was implemented step by step in three phases using two techniques of regression and classification for numeric value prediction and classifying students in to two types of classes i.e. success and failure respectively.

A. Implementation of PHASE 1

In the PHASE 1, the numeric prediction of SEM6Score of individual student is done using Neural Network and Decision Tree is used to classify students in two classes of PASS or FAIL. Both the models are trained by using input data of the matriculate and pre-university board exam scores and CGPA of five semesters. The models were evaluated using the coefficient of correlation, MSE, accuracy and confusion matrix. Fig 1 describes the implementation plan of PHASE 1.

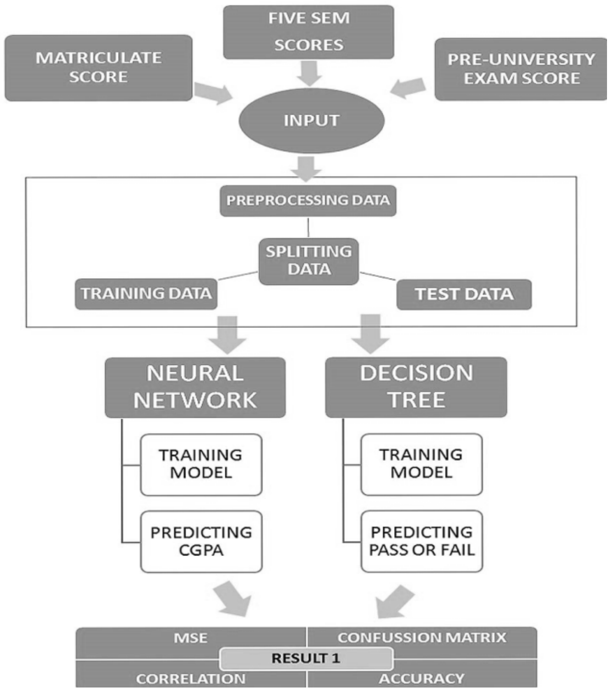


Fig. 1. Proposed implementation of PHASE 1

TABLE I. INPUT AND OUTPUT VARIABLES OF PHASE 1

Input variables	Output variable
MATRICULATE PERCENTAGE PRE-UNIVERSITY EXAM PERCENTAGE SEM1 CGPA SEM2 CGPA SEM3 CGPA SEM4 CGPA SEM5 CGPA	SEM6 CGPA

B. Modelling the Neural network for PHASE 1

The Neural Network for PHASE 1was designed using the Feed forward network topology and activation function of log sigmoid. It is implemented in R language using neural net package with resilient back propagation algorithm as learning algorithm. The model consisted of only one layer of hidden nodes. The accuracy of the predicted CGPA of Sixth semester was evaluated using Mean Squared Error. The model will be evaluated based on coefficient of correlation R, MSE. Fig2 describes the neural network for

PHASE1

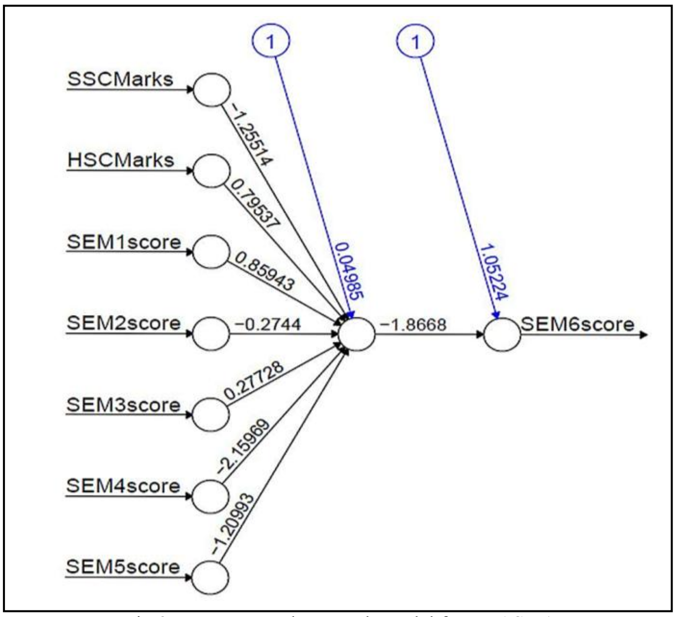


Fig.2. Neural Network model for PHASE 1

C. Modelling the Decision tree forPHASE 1

The Decision Tree in PHASE 1 described in Fig 3gave the binary prediction of whether the student will PASS (1) or FAIL (0).In present phase the Decision tree was built using the Recursive partitioning algorithm in R programming .This model is evaluated on the basis of confusion matrix and sensitivity, specificity, accuracy.

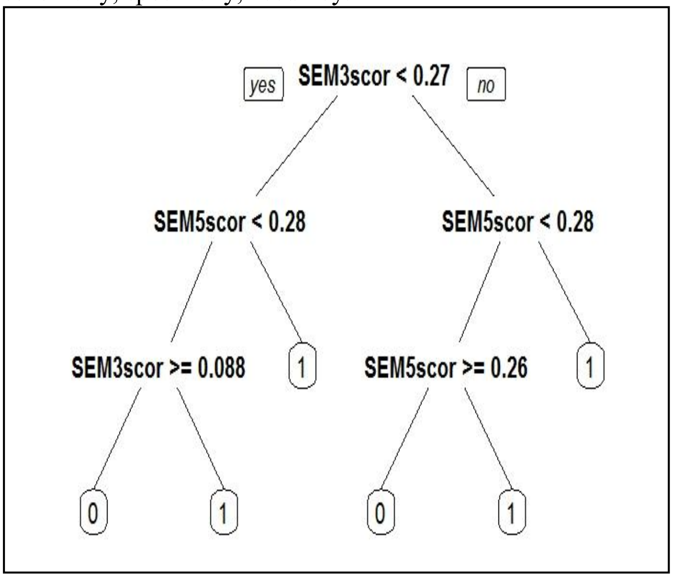


Fig. 3. Decision Tree for PHASE 1

D. Implementation of PHASE 2

In the PHASE 2 the input for the Neural network and Decision tree included psychological factors like Motivation, Concentration, Information processing, Time management, Self-testing, Study Aids ,Study Main Ideas along with the previous phase inputs of matriculate and pre-university board exam scores and CGPA of five semesters. The models were again evaluated using the coefficient of correlation, MSE, accuracy and confusion matrix.

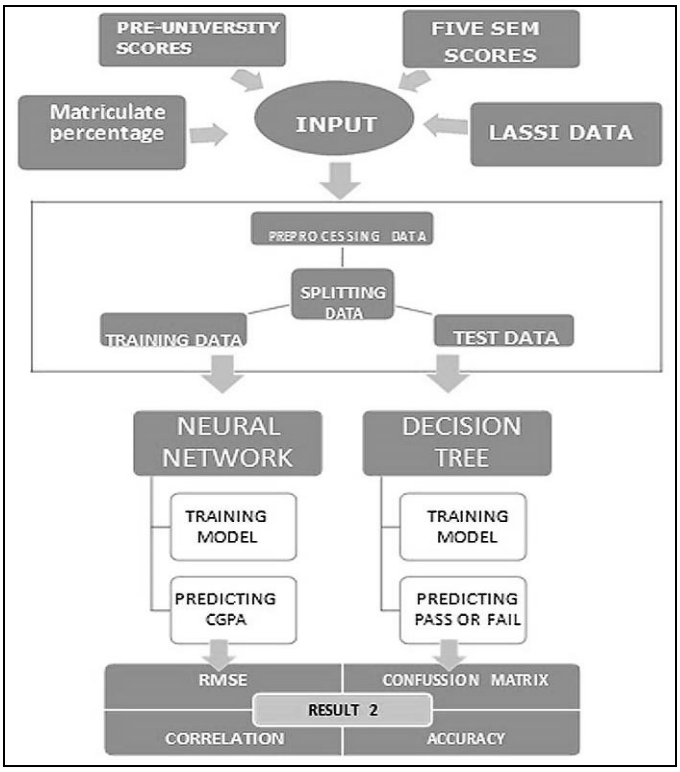


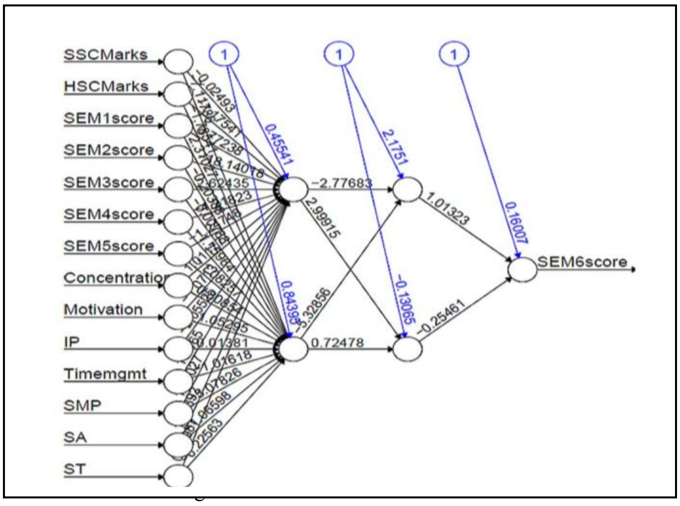
Fig.4.Proposed Implementation of PHASE 2

TABLE II. INPUT AND OUTPUT VARIABLES OF PHASE 2

Input variables	Output variable
MATRICULATE PERCENTAGE PRE-UNIVERSITY EXAM PERCENTAGE SEM1 CGPA SEM2 CGPA SEM3 CGPA SEM4 CGPA SEM5 CGPA MOTIVATION CONCENTRATION INFORMATION PROCESSING TIME MANAGEMENT SELF TESTING STUDY AIDS STUDY MAIN IDEAS.	SEM6 CGPA

E. Modelling the Neural Network for PHASE 2

Considering the new data set new Neural Network model given in Fig 5 was designed and implemented with addition of two new hidden layer each consisting of two neurons. The predicted SEM6 Score was then compared to actual score to find its accuracy. The model was evaluated using coefficient of correlation and MSE.



F. *Modelling the Decision tree for PHASE 2*

The Decision Tree in this phase gave the same binary prediction of Pass(1) or Fail(0) class but using new input variables of PHASE 2 and tree structure is of longer depth and is shown in Fig 6.

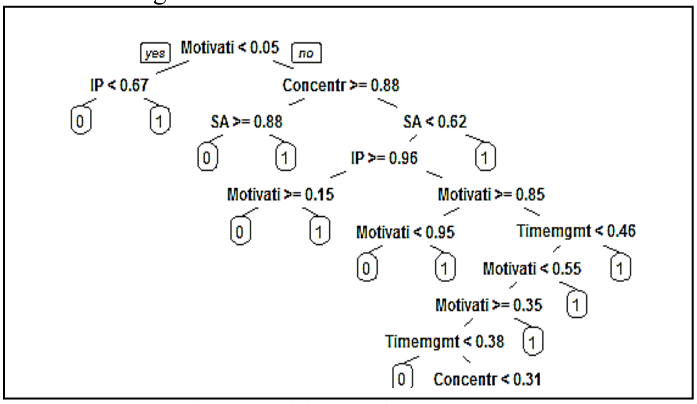


Fig6. Decision Tree for PHASE 2

G. *Implementation of PHASE 3*

In PHASE 3 the results of PHASE 1 and PHASE 2 were compared using performance parameters. The models trained in PHASE 1&2 were tested using test data set and the results were validated.

VI. Experimental Results

The Coefficient of correlation was used to find the amount of similarity between predicted and actual values which helps to know the accuracy of prediction. The Mean Square Error (MSE) signifies the error in the prediction. The models are first trained using the training dataset and the results are validated using the test dataset. Table III and IV represent performance results of PHASE 1 models and Table V and VI show the performance of PHASE 2 models. Table III and V clearly shows that the coefficient of Correlation in PHASE 1 increased from R=0.8860 to R=0.9335 in PHASE 2 and there was drastic reduction in MSE from 0.0107 to 0.0073 resulting in increase in accuracy.

TABLE III. Evaluation of Neural network without psychological factors

Performance Parameters	Results
Coefficient of Correlation	**0.8860**
Mean square Error	**0.0107**

TABLE IV. Evaluation of Decision tree without pyschological factors

Performance Parameters	Results
Specificity	**0.1379**
Sensitivity	**0.7931**
Accuracy	**0.9310**

TABLE V. Neural network Model with pyschological factors

Performance Parameters	Results
Coefficient of Correlation	**0.9335**
Mean square Error	**0.0073**

TABLE VI.Evaluation of Decision tree with pyschological factors

Performance Parameters	Results
Specificity	**0.1724**
Sensitivity	**0.7931**
Accuracy	**0.9999**

In evaluation of Decision Trees, sensitivity denoted the proportion of correctly classified passed students and specificity described the proportion of correctly classified failed students and accuracy defines the proportion of correct classification of success and failure. Table IV & Table VI presents the performance of Decision trees of two phases. The rate of accuracy was found to be more in the PHASE 2 compared to the PHASE 1. After adding the psychological factors and learning and study skills data the number of true positives and false negatives increased. Correlation analysis was made between the output variable of SEM6 Score and the LASSI scales. In figure 7, the correlation matrix showed that SEM6 Score had positive correlation with motivation, information processing, study aids. This demonstrated that motivation of the student and the way they perceived the information given to them and their usage of the study aids influenced their predicted score. The graphs presented in the figure 8&9 compares the predicted and actual values. It can be seen that at the peaks the models were either underperforming or over performing which leads to large difference between the

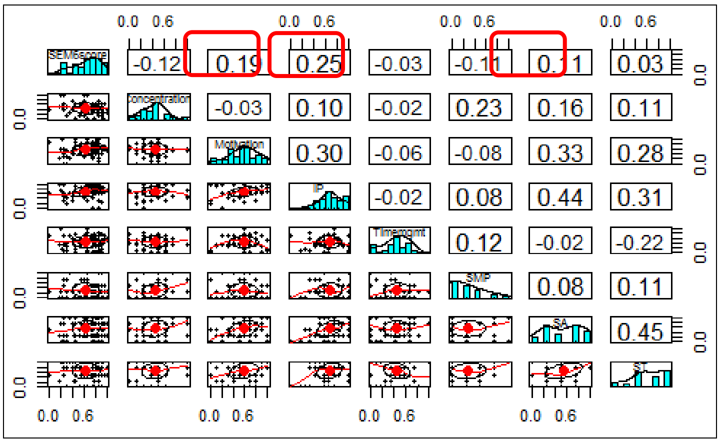


Fig.7. Correlation Matrix

individual student's predicted and their actual score. This was observed clearly for prediction of the toppers of the class.

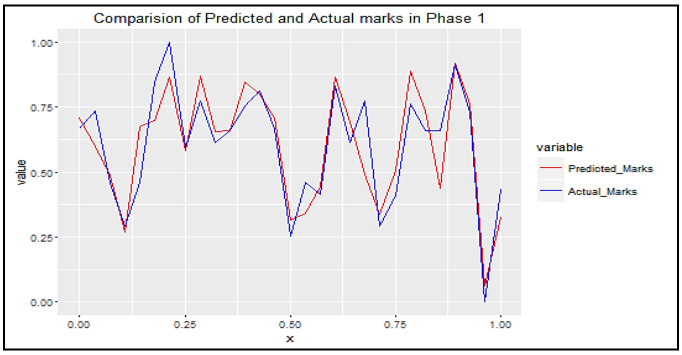


Fig.8. Comparison of Actual and Predicted Score in Phase 1

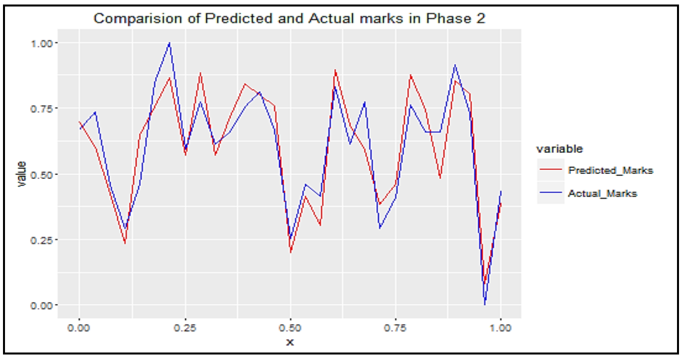


Fig.9. Comparison of Actual and Predicted Score in Phase 2

VII. CONCLUSION

The results shown in the Table V and Table VI showed that after adding the psychological factors and learning and study skills data to the previous semesters score the accuracy of the prediction in PHASE 2 increased. Both the models in PHASE 2 showed6-7% increment in the accuracy over PHASE1 models. The improvement in the overall prediction accuracy was ensured by considering thought process and learning skills of the student. The study thus proves the impact thought process in the predicting the academic performance of the students. The accurate prediction will help in planning of remedial coaching to the students who are week performers and help them to improve their scores in upcoming semesters. A faculty can get the overview of his class performance well before the examination and can plan counselling of students and extra teaching hours for the needy student.

The future work includes the cloud base application which will implement the Models of Phase 2 for getting prediction of marks. This application will get faculty members prediction of whole class by just uploading the data on previous year marks and LASSI data of the students. The TPO of the college can also make use of the application to know the number of students eligible for the placement process in the pre-final year and arrange for special classes for training the student for placement process.

ACKNOWLEDGMENT

The authors sincerely acknowledge to Thadomal Shahani Engineering College for granting permission to conduct survey with the students of final year and collect valuable data.

REFERENCES

[1] Pauziah Mohd , Norlida Buniyamin, Jamalul-lail Ab Manan, “A Neural Network Students’ Performance Prediction Model (NNSPPM*)”, IEEE International Conference on Smart Instrumentation, Measurement and Applications*,pp.1-5,November 2013

[2] Pauziah Mohd , Norlida Buniyamin, Jamalul-lail Ab Manan, “Neural Network and Linear Regression Methods for Prediction of Students’ Academic Achievement”,*IEEE Global Engineering Education Conference (EDUCON)*,pp. 916-921, 3-5 April 2014

[3] Bruno Trstenjak, Dženana ,“Determining the impact of demographic features in predicting student success in Croatia” , *MIPRO 2014*,pp.1223-1227, 26-30 May 2014

[4] Ajay Shiv Sharma , Swaraj Prince , Shubham Kapoor , Keshav Kumar “PPS — Placement prediction system using logistic regression”,*MOOC, IEEE International Conference on Innovation and Technology in Education (MITE)*,pp. 337-341, 19-20 Dec. 2014

[5] Mehdi Bostani, Amir Nadri , Azami Rezaee Nasab,“A Study of the Relation between Mental health and Academic Performance of Students of the Islamic Azad University Ahvaz Branch” presented in *5th World Conference Educational Sciences -WCES 2013, Procedia - Social and Behavioral Sciences 116 (2014)* ,pp. 163 – 165,2013

[6] Jui-Hsi Fu, Jui-Hung Chang, Yueh-Min Huang , Han-Chieh Chao, “A Support Vector Regression-Based Prediction of Students' School Performance”, presented in *International Symposium on Computer, Consumer and Control 2012*,pp.84 – 87, 4-6 June 2012

[7] I. Nincevic, M. Cukusic, Z Garaca. "Mining Demographic Data With Decision tree," *MIPRO*, May, 2010.

[8] P.M Arsad, N. Buniyamin, J. A. Manan, N. Hamzah. "Proposed Academic Students' Performance Prediction Model: A Malaysian Case Study," *3rd International Congress on Engineering Education (ICEED)*, December, 2011.

[9] R. Chuchra. "Use Data Mining Tecknique for The Evaluation of Student Performance: A Case Study," *International Journal of Computer Science and Management Researchs*, Vol 1,October, 2012.

[10] Claire E. Weinstein, Ph.D ,David R. Palmer**, “**LASSI-HS USER’S MANUAL”,1990

[11] Thadomal Shahni Engineering College[Online],Available:www.**tsec**.edu

[12] Normalization scaling between 1&0[Online],Available:https://docs.tibco.com/pub/spotfire/6.0.0-november-2013/userguide-webhelp/norm/norm_scale_between_0_and_1.htm

[13] Frauke Günther and Stefan Fritsch,“Neuralnet: Training of Neural Networks” ,June 2010

[14] R. I. I. I. Ismail M J, "Development of Neural Network Prediction Model of Energy Consumption,"2011