**International Journal of Information Retrieval Research**
Volume 12 • Issue 1

# Artificial Intelligence and Deep Learning-Based Information Retrieval Framework for Assessing Student Performance

S. L. Gupta, Birla Institute of Technology, International Centre, Oman
Niraj Mishra, Waljat College of Applied Sciences, Oman

## ABSTRACT

Improving the quality of education is a challenging activity in every educational institution. Through this research paper, a model has been proposed representing the challenges in order to manage the trade-off to maintain the philosophy of continuous quality improvement and strict control based on higher education institutions (HEIs). Several standards criteria, performance parameters, and key performance indicators are studied and suggested for a quality self-assessment approach. After the data is collected, the significant features are selected for analysis of data using dedicated gain, which are designed by integrating the information gain and the dedicated weight constants. After that, deep learning methodologies like regression analysis, the artificial neural network, and the Matlab model are used for evaluating the academic quality of institutions. Finally, areas of development have been recommended using the probabilistic model to the administrators of the institutions based on the prediction made using a deep neural network.

## KEYWORDS

Academic Quality, Artificial Intelligence, Deep Learning, Machine Learning, Performance Evaluation, Student Academic Performance Quality Improvement in Higher Education

## 1. INTRODUCTION

The knowledge sector has observed a fast increase where the universities/colleges strive for global presence and inserts the supplementary feature for globalization process. Online educational repositories are increasing thereby enhancing the learning platforms due to the advancements in the technology (Treasure-Jones et al., 2019) which are significantly demonstrating the impact on Higher Education Institutes. Lot of education mining techniques are being used using deep learning and artificial intelligence to analyse the education data and predict the performance of students in order to improve the Institution achievement and ranking and thereby also enhancing student academic achievements (Agrawal & Pandya, 2015). In depth knowledge is needed for the higher education institutions to evaluate, assess, plan and make decisions in order to remain competitive with other educational institutions. Due to the accumulation of lot of educational data, it had led

DOI: 10.4018/IJIRR.2022010101

various researchers and research communities to work on data analytics for predicting the learner's behaviour and formulate performance indicators for optimizing the formulations of policies for higher educational institutions (Azcona and Smeaton, 2017;Viberg et al., 2018) ). Thus, a new term developed as 'Educational Data Science' is the field which explores the educational data in order to perform academic analytics, predictive analytics and learning analytics (Piety et al., 2014).

The Artificial Neural Network (ANN) is the most used methodology in the Educational Data Mining (Coelho & Silveira, 2017). Few of the drawbacks of this methodology were removed with the emergence of the Deep Learning methodology (Lecun et al., 2015). Deep learning is a branch of machine learning where several computational layers enable the model to learn from patterns (Wang et al., 2017) or events learning (Nawaz et al., 2012). Not much literature is available for Deep Learning ANN's but research has proved that the Deep Learning is being used in learning analytics and is used to evaluate performance of students (Okubo et al., 2017) and assessment of the students (Li et al., 2017) in academics.

Through this research paper, we propose a model for predicting the performance of student and helping the higher educational institutions to have an edge over other competitors and improve their learning process and performance of the institute. Section 1 gives and introduction about deep learning and learning analytics. Section 2 elaborates our objective for our research work and Section 3 discusses the contribution of the various researchers which has helped us to carry forward our research. Section 4 describes the methodology and Section 5 represents the implementation of our proposed model through regression analysis, deep learning and Matlab. Section 6 represents the results and discussion and Section 7 describes the conclusion for our research work.

## 2. PROBLEM STATEMENT AND OBJECTIVE

The factors which act as a predictor for performance of students in higher education institutes using ANN has not been studied much by researchers and this research gap has been the main focus area of the present study.

Hence the study was done with following objectives.

The purpose of the research is to study the various aspects of quality, namely commitment of Board of Trustees towards quality management, improvement in teaching and learning, mapping of stakeholders expectations, and professional development assistance by affiliating university. The basic purpose of the research is to predict the academic quality of institutions using deep neural network, and to improve the areas using probabilistic based recommendation model.

This research has contributed to the knowledge in area of ANN, but there are several limitations of this study which should be kept in mind when interpreting the findings, The study has been conducted on students of 10 universities and colleges in Oman and thus generalization of the findings should be done with caution. It is also suggested to have further empirical investigation to establish whether the constructs in the proposed model vary across countries and types of higher education institutes.

## 3. RELATED WORK

Learning analytics is the area related to prediction in academics which can thereby focus on students or higher education institutes. The data is gathered, assembled, examined and analyzed for information of students and higher education institutes in learning analytics in order to understand the overall learning environment and optimizing the performance of students and institute (Siemens & Long, 2011). It helps the higher education institutions in assessing their academic performance, framing and formulating strategies and policies and helps in effective decision making (Leitner et al., 2017). Learning analytics in higher education focuses on predicting academic growth of institutions, predicting student's performance, reducing attrition rate and formulating policies which help in increasing the stability of the institution. Thus, synonyms to Learning analytics would be Educational analytics

and Academic Analytics. Deep learning and machine learning along with Artificial Intelligence help in performing prediction analysis of students learning, teaching methodologies and thereby use the educational repositories to extract all the meaningful information by generation of patterns and graphs which can help the management of higher educational institutions to make effective decisions.

Deep learning methodologies along with data mining techniques are being used to predict the performance of students and help in identifying the weak students and also help the students in their future growth (Hussain et al., 2019). Bendangnuksung & Prabu (2018) proposed a model using deep neural network to identify the weak students who may fail in examination in order to help them provide extra classes and improve their performance. Similarly, earlier educational data mining was done on historical and operational data of higher education institutions to help them in assessing, evaluating the policies of the institute and lend a hand in decision making process (Beikzadeh & Delavari, 2004; Aher & Lobo, 2011) also suggested the use of data mining for prediction analysis in education institutions. Zohair (2019) has studied the possibility of training and modeling a small dataset size and the feasibility of creating a prediction model with credible accuracy rate. Faculty consultation is a key influencer predictor for performance of students in higher education institutes. In recent studies investigating predictors for student performance (Rivas et al., 2021; Kaur & Vadhera, 2021) it has been seen that faculty consultation is a key factor affecting the performance.

The method and models proposed by various researchers takes information about different aspects which are responsible towards the building of a progressive institute and performs regressive analysis and goes through deep learning to bring out what is necessary and how it should be executed. Regression analysis being one of the tools to depict relationship between variables is considered the primary factor for implementation of deep learning methodologies (Sykes, 1993). Regression analysis is basically used for statistical analysis (Freund et al., 2006). Deep learning utilizes a large data set and represents the data in different levels of abstraction (Lecun et al., 2015). (Deng & Yu, (2013) and Deeley (2014) have used the deep learning techniques to calculate the employability skills of students enrolled in an institution. The influential selected variables in each of the Regression-model are further processed in Artificial Neural Network (Yegnanarayana, 2012); (Hassoun, 1995); (Naser, 2012) modelling through R-program (Matloff, 2011). After the analysis each of these variables form a node in the neural network that is processed in the Neural Network model through R-program. The first analysis is through the training model and a part of data set has been considered for the analysis during the training and testing of the model. The weightages given by the neural network model, to the influential input variables that predict the outcomes of factor analysis, impacts the academic quality of higher institutes (Aggarwal, 2018; Schmidhuber, 2014). This brings out the role of deep learning and we proceed to the MATLAB functions.

Matlab tool for carrying out numerical computations which is beneficial for universities and education institutions (Higham & Higham, 2016). R- Tool is another tool which is used for facilitating the used of datga mining algorithms such as neural networks n classification and regression analysis (Cortez, 2010). The MATLAB functionality plays a vital role in the proposed system (Hingham & Hingham, 2016; Kim, 2017). The Linear Regression for multiple variables has been used to train the models and test each model with the test dataset that has not been passed to the application earlier. This helped in getting the full statistics of a new dataset after an optimised training model has been developed. The trained models were optimised with the PCA, "Principal Component Analysis" and also the Feature Selection options. The best result model was selected for the testing procedure and statistics reported in the analysis. The high coefficient predictors of the regression model were processed as the input nodes in Artificial Neural Networks for building the model with training and then improving the performance, to reduce the error and increase the acceptability factor, similar to R-Squared (Craven & Shavlik, 1997). These methods are also applicable in various fields such as online tutorials, enhancement techniques for different organizations (Clare, 2007; Campagni et al., 2015; Tam et al., 2012; Gaudioso & Méndez, 2005).

## 4. METHODOLOGY

The research is based on empirical research. The survey was conducted among students of 10 universities and colleges in Oman. Convenience sampling was used as sampling tool for collecting the data. The data was collected from structured questionnaire having Likert scale. 420 responses were received out of which 350 responses were found suitable for analysis after scrutiny.

The primary intention of this research is to design and develop an approach for accessing quality of academic performance of students in technical education based on deep learning model. Thus, an integrative approach was designed and developed using deep neural networks for evaluating academic quality. In order to develop an integrative framework, the service quality parameters related to academic activities are collected from the students, alumni, parents and recruiters of various technical institutions. Accordingly, the questionnaire shared with all the stakeholders comprises of various items, like maximum learning time, extra academic activities, practical orientation in education, prompt service of the supporting staff, as like the detailed parameters given below are prepared and data with respect to these parameters are collected.

The important parameters considered in the questionnaire include:

- Maximum learning time
- Extra academic activities
- Practical orientation in education
- Prompt service of the supporting staff
- Effective classroom management
- Faculty available regularly for students' consultation

Initially, the data is gathered, and the significant feature components are chosen by the dedicative gain, which shall be newly designed by integrating the information gain with the dedicative weight constants. Then, deep neural network is modelled to evaluate the academic quality of institutions. At last, areas of development can be recommended through the probabilistic model to the administrators of the institutions based on the prediction made.

### 4.1 Design of Questionnaire

The questionnaire designed has been classified into two important factors – Internal Factors and External Factors. A sample size of 10 academic institutes was considered for our research. In internal factors, location, maximum learning time, extra academic activities, practical orientation in education, fee structure, recommendations of teachers / counsellors, course offered, course load/credits, college infrastructure, international collaborations, hostel facilities, advertisement , prompt service of the supporting staff, effective classroom management, faculty available regularly for students' consultation, college website, career growth prospects, college placements (including training-related opportunities) are considered.

In the external factors, international certification, word-of-mouth, scholarship and/or sponsorship, alumni feedback, employer feedback and Job placement are considered.

The variables selected belong to one of the two factors, Internal or External. All of these are directly or indirectly influencing the academic quality of the Higher Institutions. A very close relationship between the factors and nature of the variables has been observed with regards to the type of variables. Any prediction model needs to have an "Outcome" as a result of model based on "Inputs".

The inputs are the "Independent" variables that are directly in control and are the "Predictors" or the "Influencers". The independent variables are called as such because independent variables predict or forecast the values of the dependent variable in the model.

The output variable of the model, is the ones that is not directly in the control, rather is the result. The "dependent" variables refer to that type of variable that measures the response of the independent variable(s) on the test units. The dependent variables are named as such because they are the values that are predicted or assumed by the predictor / independent variables.

Considering the nature of questionnaire and an initial analysis gave insights as to which factors are in control (Internal) and which are not in control (External). Taking this forward, a correlation and co-variance was done for each of the variables to confirm their categories of either being an "Independent" or "Dependent".

- 15 variables from Internal factors gave a high to medium correlation & covariance to confirm as their selection as "Independent" variables.
- 4 variables from External factors were more output and result oriented indicating their selection as "Dependent" variables.

The details of the number of variables from each category are mentioned in Table 1. The questions of the students, parents and faculty questionnaire were matched to the above listed variables.

Multiple regression has been used to predicted value of dependent variable based on the 15 selected independent variables in R-tool. ANN and Matlab has been used to train the data so as to build a good training model for better results. R and Matlab has been used simultaneously to validate the results of the training model.

**Table 1. Variable for Internal and External Factors**

| Factors | Type of Variables | R_Label Name | Variables |
|---|---|---|---|
| Internal | Independent | LT | Location_Transport |
| Internal | Independent | LRT | Learning_Time |
| Internal | Independent | EC | Extra_curricular |
| Internal | Independent | PO | Practical Orientation |
| Internal | Independent | CO | Courses_offered |
| Internal | Independent | CF | Credit_feedback |
| Internal | Independent | II | Institute_Infrastructure |
| Internal | Independent | HF | Hostel_facilities |
| Internal | Independent | IA | Institute_Advertisements |
| Internal | Independent | CM | Classroom_management |
| Internal | Independent | FC | Faculty_consultation |
| Internal | Independent | EP | Exhibition_participation |
| Internal | Independent | IW | Institute_Website |
| Internal | Independent | CG | Career_Growth |
| Internal | Independent | FS | Fee_Structure |
| External | Dependent | PLM | Job_Placement |
| External | Dependent | IS | International_Status |
| External | Dependent | SS | Scholarship & Sponsorship |
| External | Dependent | RF | Reference & Feedback |

## 5. IMPLEMENTATION

The objective of deriving a prediction model for the "Academic Quality", has been tried using different methodologies. The 4 dependent variables - International Status, Scholarship & Sponsorship, Reference & Feedback and Job Placement are considered for prediction analysis:

- The first variable International Status is related to all parameters of "International External factor". This is one of the Key output variables that strongly relates to the "Quality of Academics of Higher Education Institutes". This variable reflects the International Positioning of the institutes.
- Scholarship & Sponsorship output variable strongly relates to the "Quality of Academics of Higher Education Institutes". This variable considers the Scholarships & Sponsorships to students of the institutes. It considers the recommendations and feedbacks of students, alumni, parents & teacher.
- Reference and Feedback is one of the Key output variable that strongly relates to the "Quality of Academics of Higher Education Institutes". This variable considers the recommendations and feedbacks of students, alumni, parents & teachers.
- Job Placement "PLM" is an output variable that strongly relates to the "Quality of Academics of Higher Education Institutes". This variable considers the job and training opportunities for students.

Firstly, prediction has been done based on "Multiple Regression". Multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables. Thus, we have identified the "Independent" and "Dependent" variables. The value of dependent variable will be predicted based on the 15 selected independent variables in R-tool. The result will give us a simple model to check the relationship of independent and dependent variables through the outputs. Multiple regression will also help determine the overall fit of the model and the relative contribution of each of the predictors selected through their coefficient outputs.

Based on the multiple regression output, we would be able to analyse the influence of the variables and the ones which do not contribute in an acceptable weight, would be rejected. As there are four dependent variables, there are four predicted models with each one having the same set of 15 variables. Every model has selected the key variables with their relevant weightages give as "Coefficients" by the R-programming technique. The Key outputs analysed through the regression model are explained. These statistics help to figure out how well a regression model fits the data. Based on these output statistics, a model could be accepted or rejected.

Secondly, we have used deep learning which is an artificial intelligence function that imitates the workings of the human brain in processing data and creating patterns for use in decision making. Deep learning, a subset of machine learning, utilizes a hierarchical level of artificial neural networks to carry out the process of machine learning. The artificial neural networks are built like the human brain, with neuron nodes connected together like a web. While traditional programs build analysis with data in a linear way, the hierarchical function of deep learning systems enables machines to process data with a nonlinear approach. The power of neural networks lies in their ability to handle non-linear data relationships. They are able to create relationships and patterns between variables that would prove impossible or too time-consuming for human analysts. Deep learning learns from vast amounts of unstructured data that would normally take humans decades to understand and process.

Lastly, The data set with same set of Independent and Dependent variables, has been partitioned for training and test to be modelled in Matlab, using the Multiple Linear Regression algorithm. The training and test data have been partitioned in 70:30 ratio so as to build a good training model for better results. The 70% dataset was trained and optimised to build four different models for the four selected output or dependent variables. All the models had same set of 15 independent variables or the "predictors".

## 5.1 Regression Analysis

Multiple regression analysis is done on the 4 dependent variables - International Status, Scholarship & Sponsorship, Reference & Feedback and Job Placement. R-Tool is used for prediction analysis. The statistical summary for the 4 dependent variables is explained in the next sections.

### *5.1.1 International Status*

The model design is based on IS (International Status) as a dependent variable "Y", that is predicted based on the 15 independent variables in R-program for multiple regression. All 15 independent variables were put in the regression model and based on the significance, "0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1". Green marked variables with asterisks, in the R-output for IS as dependent, have been selected by the model as influencing input parameters as shown in Figure 1. The lower the significant code value, higher is the percentage of the influence. The six selected independent variables as below:

- CO: Courses Offered
- CF: Credit Feedback
- HF: Hostel Facility
- IA: Institute Advertisement
- FC: Faculty Consultations
- CG: Career Growth

Three of the variables, "Courses Offered", "Hostel Facility" and "Institute Advertisement" from Internal factors are showing approximately 20% each weightage on International Status.

### *5.1.2 Scholarship and Sponsorship*

The model design is based on SS (Scholarship & Sponsorship) as a dependent variable "Y", that is predicted based on the 15 independent variables in R-program for multiple regression. All 15 independent variables were put in the regression model and based on the significance, "0 '***'

**Figure 1. Significant code for International Status**

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.157241   0.340103   0.462  0.64431
LT          -0.051858   0.059859  -0.866  0.38727
LRT          0.029729   0.066059   0.450  0.65313
EC          -0.039454   0.065336  -0.604  0.54657
PO           0.100127   0.094262   1.062  0.28932
CO           0.201864   0.088807   2.273  0.02401 *
CF           0.062573   0.026473   2.364  0.01898 *
II          -0.006990   0.098451  -0.071  0.94346
HF           0.222866   0.074230   3.002  0.00299 **
IA           0.174393   0.067107   2.599  0.01000 *
CM          -0.014879   0.067144  -0.222  0.82483
FC           0.139094   0.062642   2.220  0.02743 *
EP          -0.008392   0.061908  -0.136  0.89230
IW           0.029565   0.047298   0.625  0.53258
CG           0.114019   0.038175   2.987  0.00315 **
FS           0.020019   0.041405   0.483  0.62924
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1". Green marked variables with asterisks, in the R-output for SS as dependent, have been selected by the model as influencing input parameters as shown in Figure 2. The lower the significant code value, higher is the percentage of the influence. The seven selected independent variables as below:

- LT: Location & Transport
- LRT: Learning Time
- PO: Practical Orientation
- CO: Courses Offered
- CF: Credit Feedback
- FC: Faculty Consultations
- FS: Fee Structure

The most significant variable, Faculty Consultation, "FC" has almost 100% significance as an influencing variable. The weightage of FC variable is the highest at 43%. The others from 6 selected are Courses Offered "CO" at 36% and Learning Time "LRT" at 23%. The variable Practical Orientation "PO" is giving a negative impact of 39%, which is being ignored, as an outlier due to possible ambiguity around the survey questionnaire.

### *5.1.3 Reference and Feedback*

The model design is based on RF (Reference & Feedback) as a dependent variable "Y", that is predicted based on the 15 independent variables in R-program for multiple regression. All 15 independent variables were put in the regression model and based on the significance, "0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1". Green marked variables with asterisks, in the R-output for RF as dependent, have been selected by the model as influencing input parameters as shown in Figure 3. The lower the significant code value, higher is the percentage of the influence. The five selected independent variables as below:

- EC: Extra Curricular Activities
- CO: Courses Offered

**Figure 2. Significant code for Scholarship and Sponsorship**

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.52694    0.46865  -1.124  0.26210
LT           0.14496    0.08248   1.757  0.08026 .
LRT          0.23540    0.09103   2.586  0.01037 *
EC          -0.11930    0.09003  -1.325  0.18654
PO          -0.39681    0.12989  -3.055  0.00253 **
CO           0.36174    0.12237   2.956  0.00346 **
CF           0.07613    0.03648   2.087  0.03806 *
II           0.17928    0.13566   1.322  0.18773
HF          -0.05173    0.10229  -0.506  0.61356
IA           0.04619    0.09247   0.500  0.61793
CM           0.10642    0.09252   1.150  0.25135
FC           0.43649    0.08632   5.057  9.1e-07 ***
EP           0.06596    0.08531   0.773  0.44024
IW          -0.08296    0.06518  -1.273  0.20444
CG           0.02172    0.05260   0.413  0.68009
FS           0.14054    0.05705   2.463  0.01455 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Figure 3. Significant code for Reference and Feedback**

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.140580   0.308087   3.702 0.000271 ***
LT          -0.019553   0.054224  -0.361 0.718754
LRT         -0.034632   0.059840  -0.579 0.563362
EC          -0.110051   0.059185  -1.859 0.064327 .
PO           0.094996   0.085388   1.113 0.267153
CO           0.174556   0.080447   2.170 0.031112 *
CF          -0.012268   0.023981  -0.512 0.609459
II          -0.030476   0.089183  -0.342 0.732890
HF          -0.026436   0.067242  -0.393 0.694593
IA          -0.009991   0.060790  -0.164 0.869613
CM           0.009828   0.060823   0.162 0.871785
FC           0.156150   0.056745   2.752 0.006431 **
EP           0.090713   0.056080   1.618 0.107216
IW           0.059751   0.042846   1.395 0.164579
CG           0.213529   0.034581   6.175 3.24e-09 ***
FS           0.107527   0.037507   2.867 0.004556 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- FC: Faculty Consultations
- CG: Career Growth
- FS: Fee Structure

The most significant variable, Career Growth, "CG" has almost 100% significance as an influencing variable. The weightage of CG variable is the highest at 21%. The others from 5 selected are Faculty Consultation "FC" at 17% and Courses Offered "CO" at 15%. The variable Extra Curricular Activities "EC" is giving a negative impact of 11%, which is being ignored, as an outlier due to possible ambiguity around the survey questionnaire.

### *5.1.4 Job Placement*

The model design is based on PLM (Job Placement) as a dependent variable "Y", that is predicted based on the 15 independent variables in R-program for multiple regression. All 15 independent variables were put in the regression model and based on the significance, "0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1". Green marked variables with asterisks, in the R-output for PLM as dependent, have been selected by the model as influencing input parameters as shown in Figure 4. The lower the significant code value, higher is the percentage of the influence. The three selected independent variables as below:

- LRT: Learning Time
- FC: Faculty Consultations
- CG: Career Growth

The most significant variable, Faculty Consultation "FC" has almost 99.9% significance as an influencing variable. The weightage of FC variable is the highest at 30%. The others from 3 selected are Learning Time "LRT" at 27% and Career Growth "CG" at 14%.

**Figure 4. Job Placement**

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.413749   0.515743  -0.802  0.42330
LT           0.051111   0.090772   0.563  0.57397
LRT          0.276504   0.100174   2.760  0.00627 **
EC          -0.031088   0.099078  -0.314  0.75400
PO           0.011999   0.142942   0.084  0.93318
CO          -0.006731   0.134670  -0.050  0.96019
CF          -0.018603   0.040144  -0.463  0.64353
II           0.158422   0.149294   1.061  0.28981
HF          -0.095136   0.112564  -0.845  0.39895
IA           0.126102   0.101763   1.239  0.21663
CM           0.087049   0.101819   0.855  0.39353
FC           0.299551   0.094993   3.153  0.00184 **
EP           0.098056   0.093879   1.044  0.29743
IW          -0.006079   0.071724  -0.085  0.93253
CG           0.141457   0.057890   2.444  0.01535 *
FS           0.054593   0.062788   0.869  0.38555
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## 5.2 Deep Learning

The influential selected variables in each of the Regression-model are further processed in Neural Network modelling through R-program. The Feedforward technique of data flow direction is from input nodes to the output nodes through the hidden layers, inserted to enhance the model performance.

Model construction method:

1. **Input layers:** Layers that take inputs based on existing data.
2. **Hidden layers:** Layers that use backpropagation to optimise the weights of the input variables in order to improve the predictive power of the model.
3. **Output layers:** Output of predictions based on the data from the input and hidden layers.

### *5.2.1 International Status*

The input nodes here are the factors that have a high significance and influence on the dependent variable, "International Status" which impacts the academic quality of the Higher Institutions. These six selected factors through multiple regression are as below:

- CO: Courses Offered
- CF: Credit Feature
- HF: Hostel Facility
- IA: Institute Advertisement
- FC: Faculty Consultations
- CG: Career Growth

Each of these variables form a node in the neural network that is processed in the Neural Network model through R-program. The first analysis is through the training model and a part of data set has been considered for the analysis during the training and testing of the model.

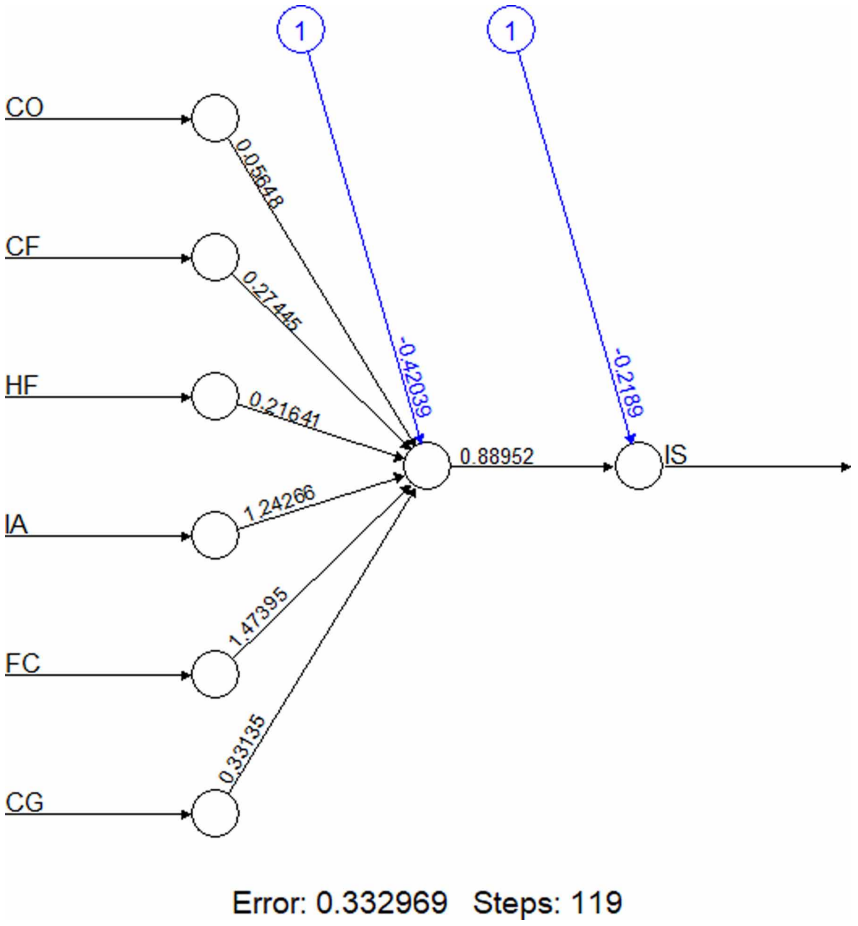
The weightages given by the neural network model, to the influential input variables that predict the outcome of IS, impacting the academic quality of higher institutes, is shown in Table 2. The top three variables are “IA”, “FC” and “CG”.

The Figure 5 represents the number of Training steps and an error measure called the SUM OF SQUARED ERRORS (SSE), which is the sum of the squared differences between the predicted and

**Table 2. Weightage for International Status**

| Input Variable | Feature Name | Weightage |
|---|---|---|
| CO | Courses Offered | 0.05648 |
| CF | Credit Feature | 0.27445 |
| HF | Hostel Facility | 0.21641 |
| **IA** | **Institute Advertisement** | **1.24266** |
| **FC** | **Faculty Consultations** | **1.47395** |
| **CG** | **Career Growth** | **0.33135** |

**Figure 5. Sum of Squared Errors for International Status**



actual values. The lower the SSE, more closely the model conforms to the training data, which tells about performance on the training data. The iteration process in this model involved 119 steps to give the error of 0.332969, approximately 33%.

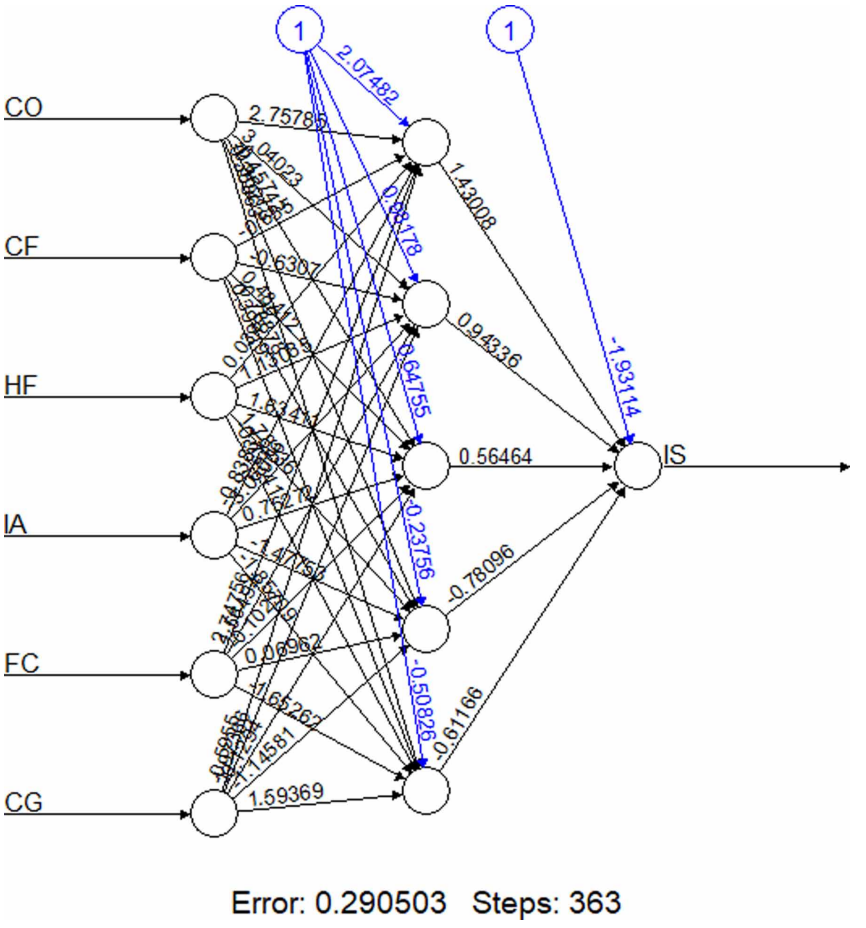
After improving the performance of the model as shown in Figure 6, the SSE has been reduced from 0.33 to 0.29. Additionally, the number of training steps rose from 119 to 363. Applying the same steps to compare the predicted values to the true values, we now obtain a correlation around 0.5428797. The correlation more or less remained similar but the error value reduced.

### *5.2.2 Scholarship and Sponsorship*

There are effectively seven outputs in Multiple Regression model and we have ignored the one with negative impact due to possible ambiguity in the survey question. In addition we have included two more variables that have over 10% impact, Institute Infrastructure "II" and Classroom Management "CM". These eight factors, ignoring the negative impact factor, were selected for the inputs of Neural Network:

- LT: Location & Transport
- LRT: Learning Time

**Figure 6. Sum of Squared Errors after improved performance for International Status**



- CO: Courses Offered
- CF: Credit Feature
- FC: Faculty Consultations
- FS: Fee Structure
- II: Institute Infrastructure
- CM: Classroom Management

Each of these variables form a node in the neural network that is processed in the Neural Network model through R-program. The first analysis is through the training model and a part of data set has been considered for the analysis during the training and testing of the model.

The weightages given by the neural network model, to the influential input variables that predict the outcome of SS, impacting the academic quality of higher institutes, is shown in Table 3. We see that there are many predictors that have a negative relationship with the output variable through this weight analysis of neural network. Learning Time "LRT" and Courses Offered "CO" have a high absolute value but with a negative relationship. The dataset of the sample is small for doing any further deep analysis into this. We could either ignore the relationship and consider the absolute value of these parameters or just discard them in further consideration. Depending on these features commonality with other dependent variable, we should take the decision. The bold highlighted features have significant positive impact.

As shown in Figure 7, the SSE, the iteration process in this model involved 251 steps to give the error of 0.320839, approximately 32% and on improving the performance of the model, SSE has been reduced from 0.32 to 0.23 as shown in Figure 8. Additionally, the number of training steps rose from 251 to 369. Applying the same steps to compare the predicted values to the true values, we now obtain a correlation around 0.5572707.

### *5.2.3 Reference and Feedback*

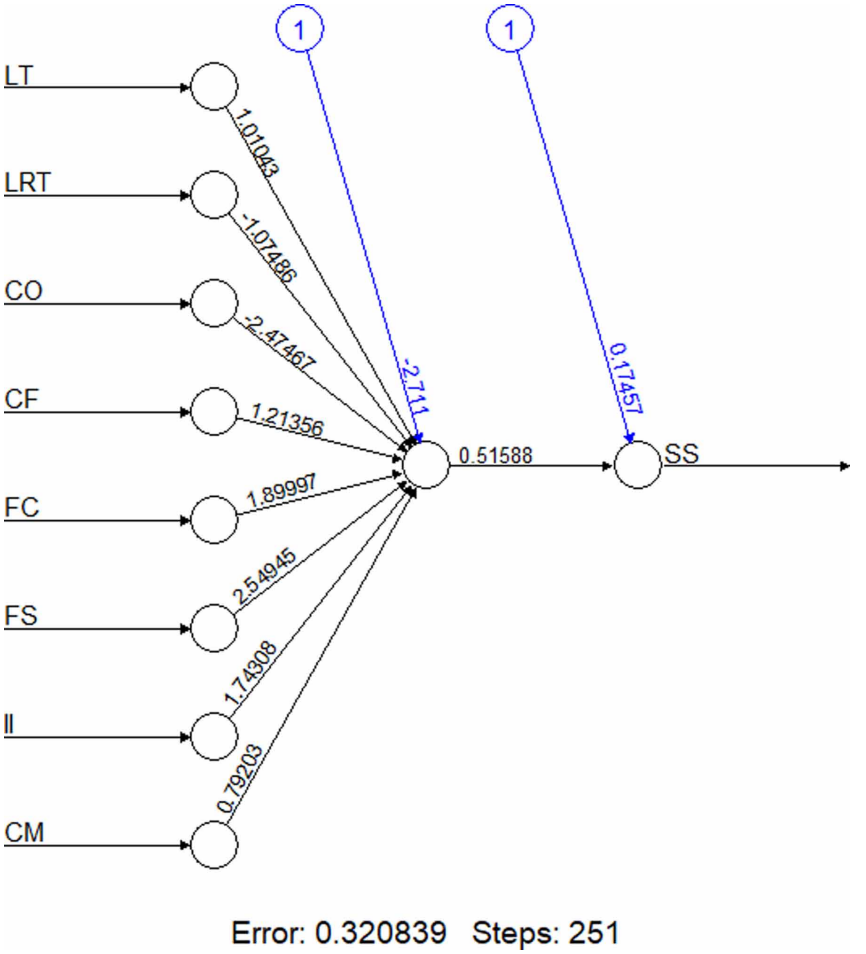
The five selected factors through multiple regression are given below. Extra-Curricular Activities have got a negative relationship as per the output of Regression model. Hence eliminating this variable from the input list of this dependent:

- EC: Extra Curricular Activities
- CO: Courses Offered
- FC: Faculty Consultations
- CG: Career Growth
- FS: Fee Structure

**Table 3. Weightage for Scholarship and Sponsorship**

| Input Variable | Feature Name | Weightage |
|---|---|---|
| **LT** | **Location & Transport** | **1.01043** |
| LRT | Learning Time | -1.07486 |
| CO | Courses Offered | -2.47467 |
| **CF** | **Credit Feature** | **1.21356** |
| **FC** | **Faculty Consultations** | **1.89997** |
| **FS** | **Fee Structure** | **2.54945** |
| **II** | **Institute Infrastructure** | **1.74308** |
| CM | Classroom Management | 0.79203 |

**Figure 7. Sum of Squared Errors for Scholarship and Sponsorship**




Each of these variables form a node in the neural network that is processed in the Neural Network model through R-program. The first analysis is through the training model and the part of data set has been considered for the analysis during the training and testing of the model.

The weightages given by the neural network model, to the influential input variables that predict the outcome of RF, impacting the academic quality of higher institutes, are shown in Table 4. The top two from these, we have "FC" and "CG" with very high weightage when the Reference & Feedback output is being considered for the academic quality.

As shown in Figure 9, the SSE, the iteration process in this model involves 497 steps to give the error of 0.412407, approximately 41%. and on improving the performance of the model, SSE has been reduced from 0.412407 to 0.338132 23 as shown in Figure 10. In this case, the number of training steps reduced from 497 to 212. In this model, correlation value is not giving the desired acceptable result. This certainly could be further analysed with some changes in the input variables and doing few more iterations.

### *5.2.4 Job Placement*

The three selected factors through multiple regression and one more "II" with coefficient of 0.158422 are inputs for neural network:

**Figure 8. Sum of Squared Errors after improved performance for Scholarship and Sponsorship**

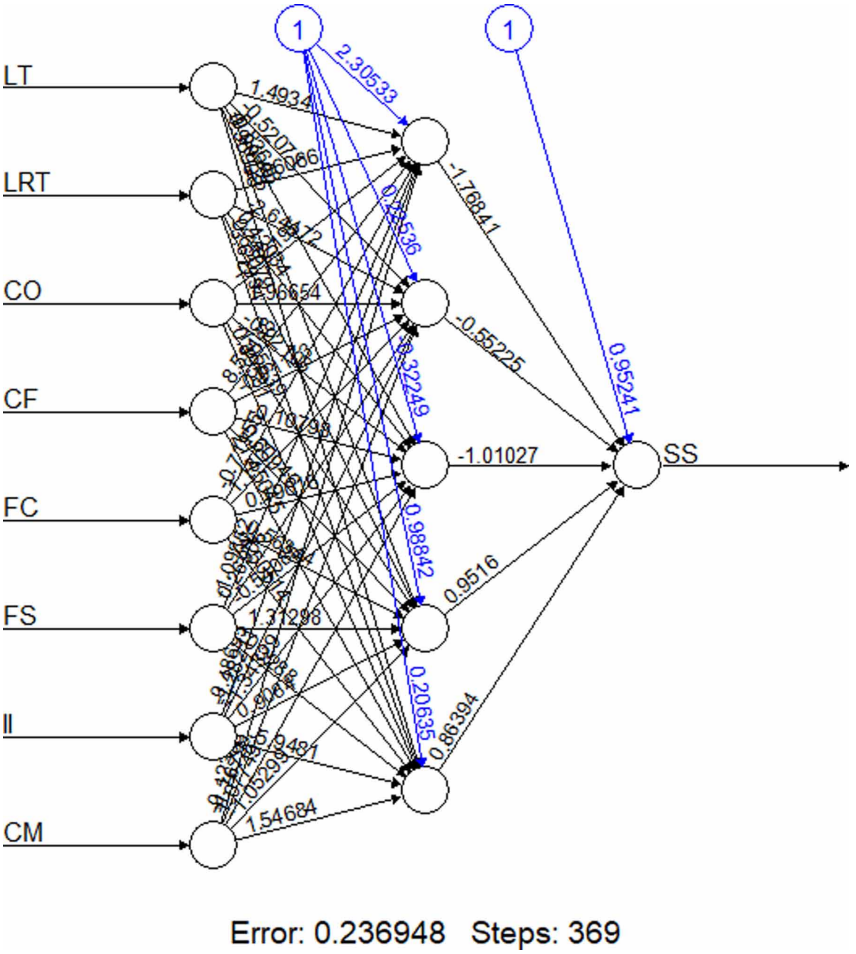
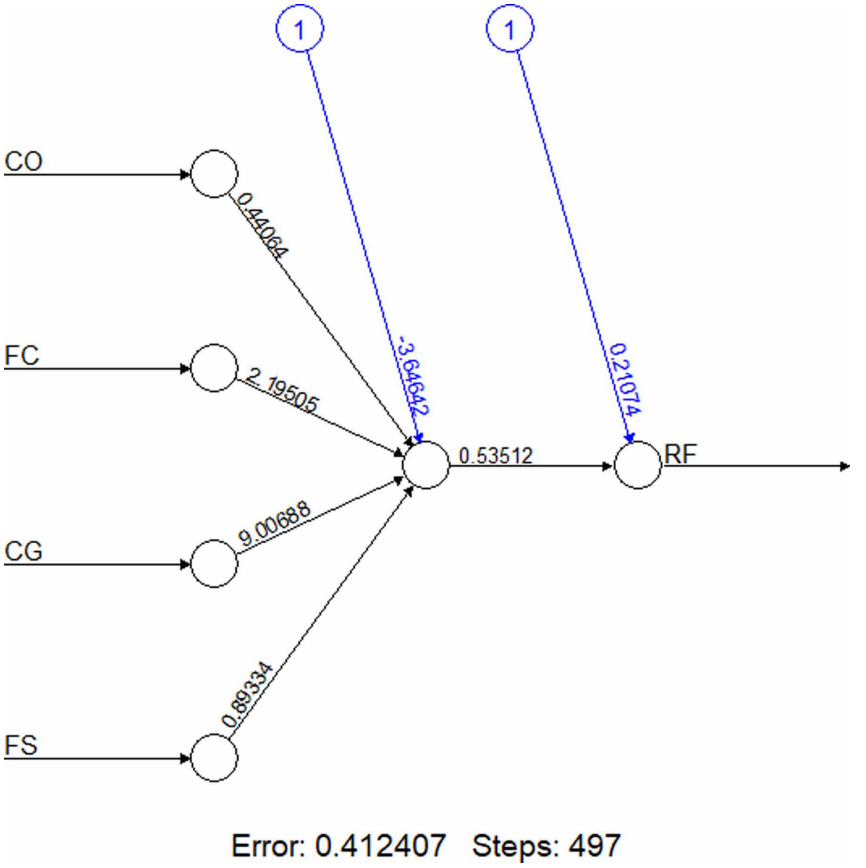



**Table 4. Weightage for Reference and Feedback**

| Input Variable | Feature Name | Weightage |
|---|---|---|
| CO | Courses Offered | 0.44064 |

| | | |
|---|---|---|
| **FC** | **Faculty Consultations** | **2.19505** |
| **CG** | **Career Growth** | **9.00688** |
| FS | Fee Structure | 0.89334 |

**Figure 9. Sum of Squared Errors for Reference a Feedback**




- LRT: Learning Time
- FC: Faculty Consultations
- CG: Career Growth
- II: Institute Infrastructure

Each of these variables form a node in the neural network that is processed in the Neural Network model through R-program. The first analysis is through the training model and a part of data set has been considered for the analysis during the training and testing of the model.

The weightages given by the neural network model, to the influential input variables that predict the outcome of PLM, impacting the academic quality of higher institutes, as shown in Table 5. From the table, we see that three out of four inputs have weightage in a negative relation. The only positive relation input is Career Growth "CG".

As shown in Figure 11, the SSE, the iteration process in this model involved 183 steps to give the error of 0.524, approximately 52%and on improving the performance of the model, SSE has been reduced from 0.524 to 0.509908 as shown in Figure 12. The number of training steps decreased from 183 to 100. Applying the same steps to compare the predicted values to the true values, we now obtain a correlation around 0.4488912. The correlation slightly increased.

**Figure 10. Sum of Squared Errors after improved performance for Reference and Feedback**

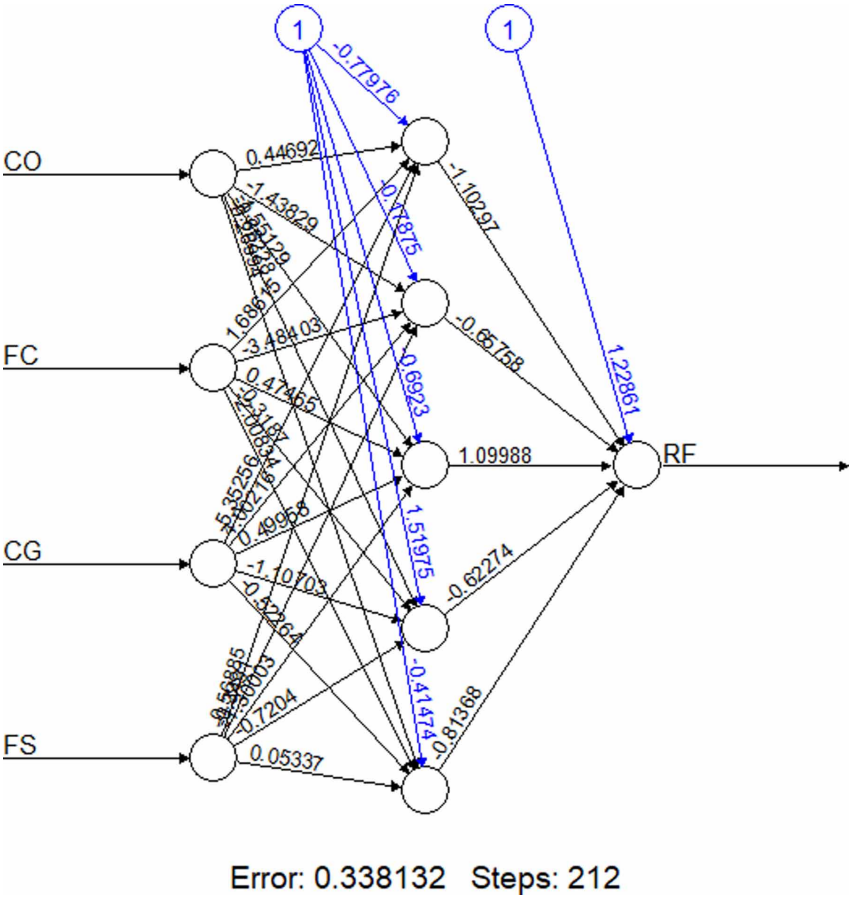



**Table 5. Weightage for Job Placement**

| Input Variable | Feature Name | Weightage |
|---|---|---|
| LRT | Courses Offered | -0.97795 |
| FC | Faculty Consultations | -0.863 |
| CG | Career Growth | 0.30491 |
| II | Institute Infrastructure | -0.06357 |

## 5.3 Matlab Model

The Linear Regression for multiple variables has been used to train the models and test each model with the test dataset that has not been passed to the application earlier. This helped in getting the full statistics of a new dataset after an optimised training model has been developed. The trained models were optimised with the PCA, "Principal Component Analysis" and also the Feature Selection options. The best result model was selected for the testing procedure and statistics reported in the analysis.

**Figure 11. Sum of Squared Errors for Job Placement**

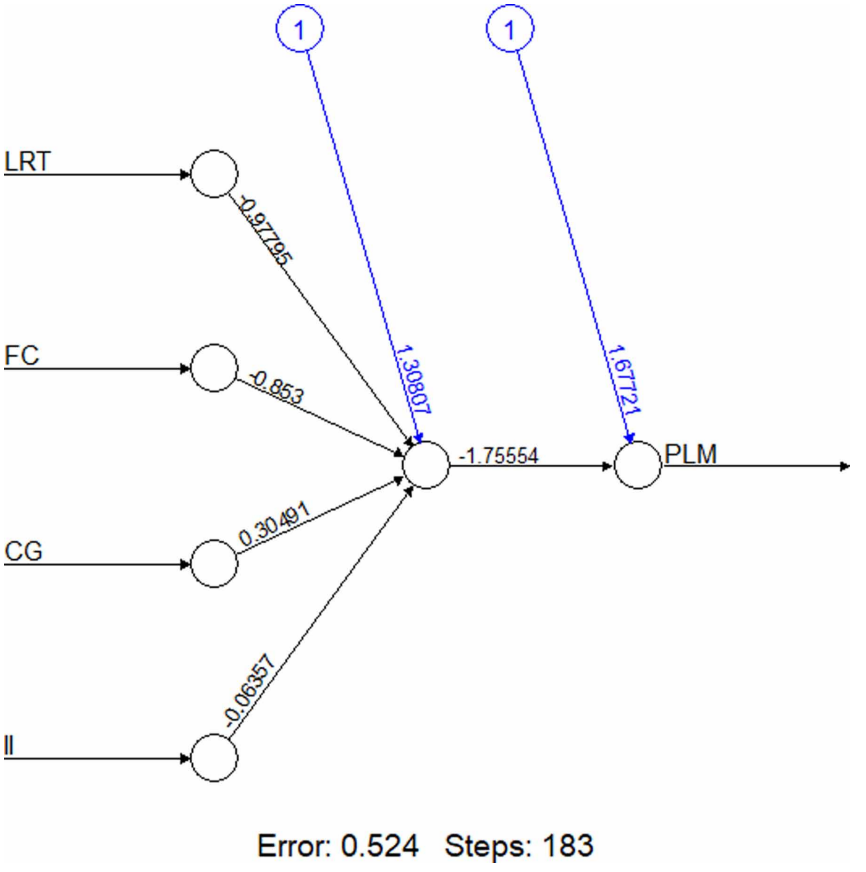
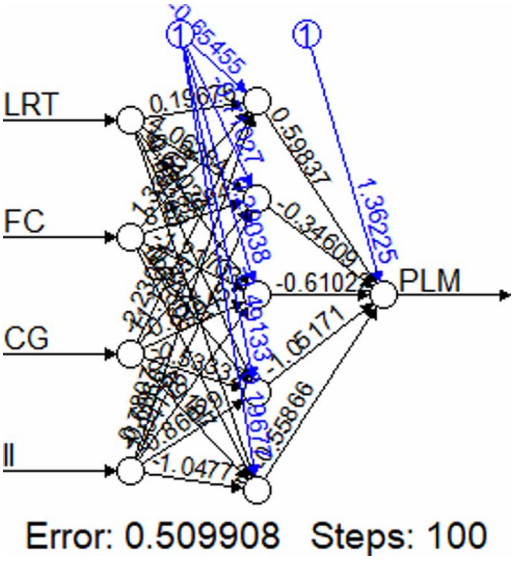



**Figure 12. Sum of Squared Errors after improved performance for Job Placement**



### 5.3.1 International Status

In training the model, the dataset was run for three scenarios, the normal Linear training, the training with PCA and training with Feature selection. Three different set of statistics were reported with improved performance as shown in Figure 13.

As we observe from the statistics that the iterations of training to improve the model performance has increased the R-Squared value from 0.55 to 0.58. First the normal dataset was trained, then the PCA was enabled and it marginally improved by 1% and then finally, few combinations of Feature selection was done to get the best R-Squared probability rate at 58% and this model was selected for testing the data for prediction.

Once the model has been trained and optimised, a new data set needs to be passed through the model for prediction. This test data set has been created by 30% partitioned data. The highest value of R-Squared model optimised during training process has been selected for the test data for prediction. The Prediction model was tested on the optimised trained model with 0.58 R-Squared value, loaded with the test dataset and the "fitlm" command displayed the below statistics for prediction model. The Predicted model gave a similar response with R-Squared value of 0.577 and a good acceptable p-value of 0.0000464 as shown in Figure 14. This shows that the model has 57.7% probability of successful prediction with data in future.

From the prediction statistics above, three key influencer predictors with high coefficients and acceptable p-values are:

- Learning Time – Coefficient 0.6484 and p-value 0.0086463
- Courses Offered - Coefficient 0.5327 and p-value 0.046497
- Faculty Consultation - Coefficient 0.39982 and p-value 0.010676

Another variable, Classroom Management with coefficient 0.20496 may be influencing but with a p-value of 0.15483 can be ignored and is not considered as a significant predictor.

### 5.3.2 Scholarship and Sponsorship

In training the model for "Scholarship & Sponsorship" output variable, the dataset was run for multiple scenarios for normal Linear training, training with PCA and training with Feature selection. The two outputs selected for further analysis are Normal Linear trained model and Model with PCA as shown in Figure 15.

The R-Squared Value increased from 0.35 to 0.39, an increase of model acceptability by 4%. Other training iterations for improving performance did not show significant improvement in the R-Squared value. The model for testing for prediction has been selected with R-Squared value of 0.39.

Once the model has been trained and optimised, a new data set needs to be passed through the model for prediction. This test data set has been created by 30% partitioned data. The highest value of R-Squared model optimised during training process has been selected for the test data for prediction.

**Figure 13. Statistics of Training Model for International Status**

**Model 1**: Trained

**Results**

| | |
|---|---|
| RMSE | 0.48186 |
| R-Squared | 0.55 |
| MSE | 0.23219 |
| MAE | 0.38145 |
| Prediction speed | ~2700 obs/sec |
| Training time | 1.0353 sec |

**Model 2**: Trained

**Results**

| | |
|---|---|
| RMSE | 0.47592 |
| R-Squared | 0.56 |
| MSE | 0.2265 |
| MAE | 0.37535 |
| Prediction speed | ~1000 obs/sec |
| Training time | 1.4114 sec |

**Model 6**: Trained

**Results**

| | |
|---|---|
| RMSE | 0.46648 |
| R-Squared | 0.58 |
| MSE | 0.21761 |
| MAE | 0.36968 |
| Prediction speed | ~2200 obs/sec |
| Training time | 1.0747 sec |

**Figure 14. Regression Statistics for International Status**

```
31/7/20 5:25 PM     MATLAB Command Window                    1 of 1

>> load('trainedModel_IS.mat')
>> load('datatest_IS.mat')
>> fitlm(datatest)

ans =


Linear regression model:
    International_Status ~ [Linear formula with 16 terms in 15 predictors]

Estimated Coefficients:
                              Estimate        SE        tStat       pValue
                              ________    ________    ________    _________

    (Intercept)                -0.95908     0.95898     -1.0001      0.32228
    Location_Transport         0.085381     0.12402     0.68846      0.49448
    Learning_Time                0.6484     0.23682       2.738    0.0086463
    Extra_curricular          -0.070154     0.11894     -0.5898      0.55809
    PracticalOrientation       -0.26877     0.17123     -1.5697      0.12306
    Courses_offered              0.5327     0.26066      2.0436     0.046497
    Credit_feedback            0.060467    0.036481      1.6575      0.10395
    Institute_Infrastructure   -0.15194     0.20434    -0.74357      0.46076
    Hostel_facilities          -0.16752     0.21223    -0.78936      0.43378
    Institute_Advertisements   0.097724     0.11502     0.84963      0.39975
    Classroom_management        0.20496      0.1418      1.4454      0.15483
    Faculty_consultation        0.39982     0.15048      2.6569     0.010676
    Exhibition_participation    0.16222     0.12415      1.3067      0.19753
    Institute_Website           -0.1651     0.10681     -1.5457      0.12875
    Career_Growth               0.01246    0.045858      0.2717      0.78702
    Fee_Structure              0.070938    0.097325     0.72888      0.46962


Number of observations: 64, Error degrees of freedom: 48
Root Mean Squared Error: 0.378
R-squared: 0.577,  Adjusted R-Squared: 0.445
F-statistic vs. constant model: 4.37, p-value = 4.64e-05
>>
```

The Prediction model was tested on the optimised trained model with 0.39 R-Squared value, loaded with the test dataset and the “fitlm” command displayed the below statistics for prediction model. The Predicted model gave a very high response with R-Squared value of 0.594 as compared to 0.39 in trained model as shown in Figure 16. A good acceptable p-value of 0.0000207. This shows that the model has 59.4% probability of successful prediction with data in future.

From the prediction statistics above, four key influencer predictors with high coefficients and average acceptable p-values are:

- Location Transport – Coefficient 0.26368 and p-value 0.078498
- Practical Orientation - Coefficient -0.39564 and p-value 0.056565
- Faculty Consultation - Coefficient 0.39982 and p-value 0.010676
- Fee Structure - Coefficient 0.26435 and p-value 0.02604

Variable, Institute Infrastructure with coefficient 0.28264 may be influencing but with a p-value of 0.24794 can be ignored and is not considered as a significant predictor. Another variable, Practical

**Figure 15. Regression Statistics for Scholarship and Sponsorship**

```
MATLAB Command Window                                                Page 1

>> load('datatest_SS.mat')
>> load('trainedModel_SS.mat')
>> fitlm(datatest)

ans =


Linear regression model:
    Scholarship ~ [Linear formula with 16 terms in 15 predictors]

Estimated Coefficients:
                                Estimate        SE         tStat       pValue
                               _________    ________    _________    ________

    (Intercept)                  0.69711      1.1341      0.61467     0.54168
    Location_Transport           0.26368     0.14667       1.7978    0.078498
    Learning_Time               -0.14424     0.28007     -0.51501     0.60891
    Extra_curricular          -0.0023788     0.14067    -0.016911     0.98658
    PracticalOrientation         -0.39564      0.2025      -1.9538    0.056565
    Courses_offered           0.00045146     0.30827    0.0014645     0.99884
    Credit_feedback              0.069191    0.043144       1.6037     0.11534
    Institute_Infrastructure      0.28264     0.24165       1.1696     0.24794
    Hostel_facilities            -0.28201     0.25098      -1.1236     0.26677
    Institute_Advertisements       0.1279     0.13602      0.94023     0.35181
    Classroom_management          0.13986     0.16769      0.83403      0.4084
    Faculty_consultation          0.37493     0.17797       2.1067    0.040392
    Exhibition_participation       0.0706     0.14682      0.48087      0.6328
    Institute_Website           -0.090857     0.12632     -0.71927     0.47546
    Career_Growth               -0.045292    0.054233     -0.83513     0.40778
    Fee_Structure                 0.26435      0.1151       2.2967     0.02604


Number of observations: 64, Error degrees of freedom: 48
Root Mean Squared Error: 0.447
R-squared: 0.594,  Adjusted R-Squared: 0.467
F-statistic vs. constant model: 4.68, p-value = 2.07e-05
```

Orientation has a high negative coefficient and acceptable p-value. This can be taken into consideration if it plays an important role for the Organisation.

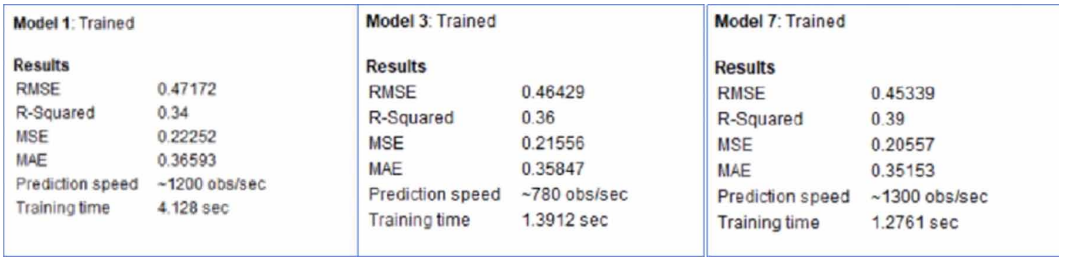
### *5.3.3 Reference and Feedback*

In training the model for "Reference & Feedback" output variable, the dataset was run for multiple scenarios for normal Linear training, training with PCA and training with Feature selection. The three outputs selected for further analysis and the model with highest R-Squared value was used for test data and prediction as shown in Figure 16.

The R-Squared Value increased from 0.34 in Normal Linear to 0.36 in Linear with PCA and finally to 0.39 in Linear with Feature Selection, an increase of model acceptability by 5%. Other training iterations for improving performance did not show significant improvement in the R-Squared value. The model for testing for prediction has been selected with R-Squared value of 0.39.

Once the model has been trained and optimised, a new data set needs to be passed through the model for prediction. This test data set has been created by 30% partitioned data. The highest value of R-Squared model optimised during training process has been selected for the test data for prediction. The Prediction model was tested on the optimised trained model with 0.39 R-Squared value, loaded

**Figure 16. Statistics of Training Model for Reference and Feedback**

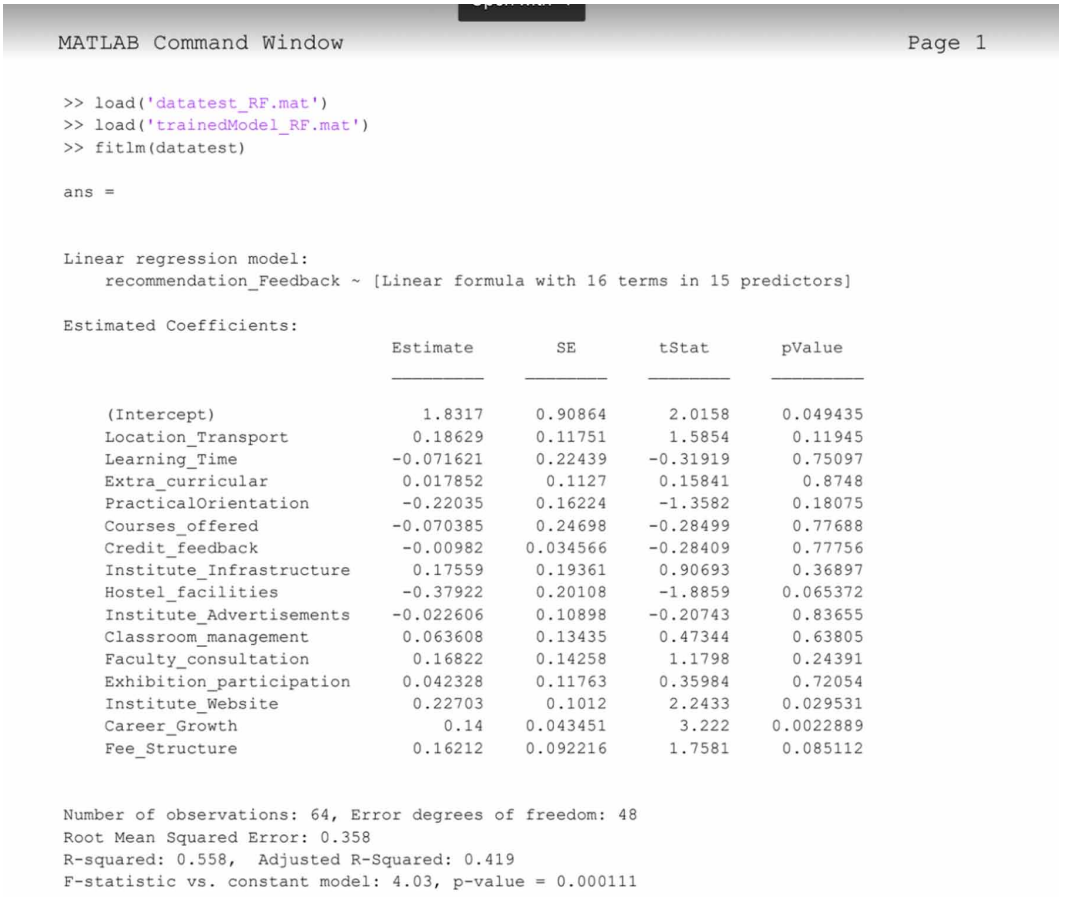


| **Model 1**: Trained | |
|---|---|
| **Results** | |
| RMSE | 0.47172 |
| R-Squared | 0.34 |
| MSE | 0.22252 |
| MAE | 0.36593 |
| Prediction speed | ~1200 obs/sec |
| Training time | 4.128 sec |

| **Model 3**: Trained | |
|---|---|
| **Results** | |
| RMSE | 0.46429 |
| R-Squared | 0.36 |
| MSE | 0.21556 |
| MAE | 0.35847 |
| Prediction speed | ~780 obs/sec |
| Training time | 1.3912 sec |

| **Model 7**: Trained | |
|---|---|
| **Results** | |
| RMSE | 0.45339 |
| R-Squared | 0.39 |
| MSE | 0.20557 |
| MAE | 0.35153 |
| Prediction speed | ~1300 obs/sec |
| Training time | 1.2761 sec |

**Figure 17. Regression Statistics for Reference and Feedback**

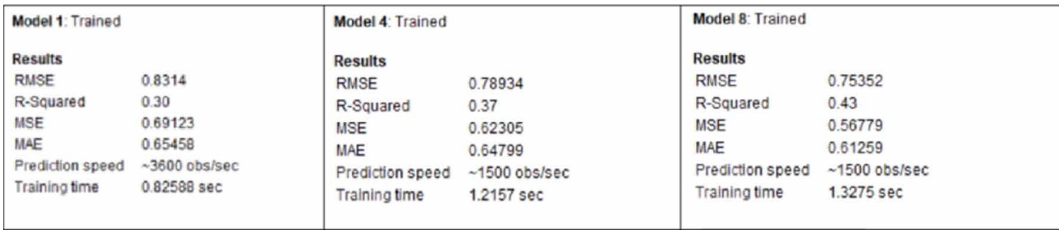


```
MATLAB Command Window                                          Page 1

>> load('datatest_RF.mat')
>> load('trainedModel_RF.mat')
>> fitlm(datatest)

ans =


Linear regression model:
    recommendation_Feedback ~ [Linear formula with 16 terms in 15 predictors]

Estimated Coefficients:
                                Estimate        SE        tStat       pValue
                                ________    ________    ________    _________

    (Intercept)                   1.8317     0.90864      2.0158     0.049435
    Location_Transport           0.18629     0.11751      1.5854      0.11945
    Learning_Time              -0.071621     0.22439    -0.31919      0.75097
    Extra_curricular            0.017852      0.1127     0.15841       0.8748
    PracticalOrientation        -0.22035     0.16224     -1.3582      0.18075
    Courses_offered             -0.070385     0.24698    -0.28499      0.77688
    Credit_feedback              -0.00982    0.034566    -0.28409      0.77756
    Institute_Infrastructure      0.17559     0.19361     0.90693      0.36897
    Hostel_facilities            -0.37922     0.20108     -1.8859     0.065372
    Institute_Advertisements    -0.022606     0.10898    -0.20743      0.83655
    Classroom_management         0.063608     0.13435     0.47344      0.63805
    Faculty_consultation          0.16822     0.14258      1.1798      0.24391
    Exhibition_participation     0.042328     0.11763     0.35984      0.72054
    Institute_Website             0.22703      0.1012      2.2433     0.029531
    Career_Growth                    0.14    0.043451       3.222    0.0022889
    Fee_Structure                 0.16212    0.092216      1.7581     0.085112


Number of observations: 64, Error degrees of freedom: 48
Root Mean Squared Error: 0.358
R-squared: 0.558,  Adjusted R-Squared: 0.419
F-statistic vs. constant model: 4.03, p-value = 0.000111
```

with the test dataset and the “fitlm” command displayed the below statistics for prediction model. The Predicted model gave a very high response with R-Squared value of 0.558 as compared to 0.39 in trained model. A good acceptable p-value of 0.000111 as shown in Figure 18. This shows that the model has 55.8% probability of successful prediction with data in future.

From the prediction statistics above, four key influencer predictors with high coefficients and average acceptable p-values are:

**Figure 18. Statistics of Training Model for Job Placement**




- Institute Website – Coefficient 0.22703 and p-value 0.029531
- Career Growth - Coefficient 0.14 and p-value 0.0022889
- Fee Structure - Coefficient 0.16212 and p-value 0.085112
- Hostel Facilities - Coefficient -0.37922 and p-value 0.065372

Variable, Hostel Facilities has a high negative coefficient and average acceptable pvalue. This can be taken into consideration if it plays an important role for the Organisation.

### *5.3.4 Job Placement*

training the model for "Job Placement" output variable, the dataset was run for multiple scenarios for normal Linear training, training with PCA and training with Feature selection. The three outputs selected for further analysis and the model with highest R-Squared value was used for test data and prediction as shown in Figure 18.

The R-Squared Value increased from 0.30 in Normal Linear to 0.37 in Linear with PCA and finally to 0.43 in Linear with Feature Selection, an increase of model acceptability by 13%. Multiple training iterations kept increasing the R-Squared value till max of 0.43 was reached. The model for testing for prediction has been selected with RSquared value of 0.43.

Once the model has been trained and optimised, a new data set needs to be passed through the model for prediction. This test data set has been created by 30% partitioned data. The highest value of R-Squared model optimised during training process has been selected for the test data for prediction. The Prediction model was tested on the optimised trained model with 0.43 R-Squared.

value, loaded with the test dataset and the "fitlm" command displayed the below statistics for prediction model. The Predicted model gave a similar response with R-Squared value of 0.441 as compared to 0.43 in trained model as shown in Figure 19. An acceptable p-value of 0.00768. This shows that the model has 44.1% probability of successful prediction with data in future.

From the prediction statistics above, only one key influencer predictors with high coefficients and acceptable p-values has been reported on this output variable.

- Faculty consultation – Coefficient 0.4287 and p-value 0.057839

There are other predictor variables with high coefficients but the p-values are not in the acceptable range. However, based on Organisation's preferences, these variables can also be considered for improvements:

- Location Transport - Coefficient 0.22056 and p-value 0.23097
- Classroom management - Coefficient 0.20745 and p-value 0.32327
- Exhibition Participation - Coefficient 0.22463 and p-value 0.22307

**Figure 19. Regression Statistics for Job Placement**

```
MATLAB Command Window                                                Page 1

>> load('datatest_PL.mat')
>> load('trainedModel_PL.mat')
>> fitlm(datatest)

ans =


Linear regression model:
    Placement ~ [Linear formula with 16 terms in 15 predictors]

Estimated Coefficients:
                                Estimate        SE         tStat       pValue
                                ________    ________    _________    ________

    (Intercept)                   1.7597      1.4057       1.2518      0.2167
    Location_Transport           0.22056     0.18179       1.2132     0.23097
    Learning_Time                0.45131     0.34715       1.3001     0.19979
    Extra_curricular            -0.016644    0.17436    -0.095457     0.92435
    PracticalOrientation          -0.1128       0.251      -0.4494     0.65517
    Courses_offered              -0.35288      0.3821      -0.92355     0.36034
    Credit_feedback              -0.039872    0.053477     -0.74559     0.45955
    Institute_Infrastructure      -0.27286     0.29953      -0.91096     0.36687
    Hostel_facilities             -0.23284      0.3111      -0.74847     0.45783
    Institute_Advertisements      0.13702      0.1686       0.81265     0.42043
    Classroom_management          0.20745     0.20786       0.99804     0.32327
    Faculty_consultation           0.4287     0.22059        1.9434    0.057839
    Exhibition_participation      0.22463     0.18198       1.2344     0.22307
    Institute_Website             -0.1023     0.15657      -0.6534     0.51661
    Career_Growth                -0.058027    0.067222     -0.8632     0.39232
    Fee_Structure                -0.036162     0.14267     -0.25347     0.80099


Number of observations: 64, Error degrees of freedom: 48
Root Mean Squared Error: 0.553
R-squared: 0.441,  Adjusted R-Squared: 0.267
F-statistic vs. constant model: 2.53, p-value = 0.00768
```

## 6. RESULT AND DISCUSSION

The predictive final analysis was done comparing the Machine Learning Regression (MLR) model and the Hybrid model of Regression and Artificial Neural Network (ANN). The high coefficient predictors of the regression model were processed as the input nodes in ANN for building the model with training and then improving the performance, to reduce the error and increase the acceptability factor, similar to R-Squared. All the four models were evaluated to measure the accuracy, comparing the calculated value of coefficient of determinant also known as R-Squared. Higher the value of R-squared, the more acceptable forecasts are related to the actual data. The comparison has been done with R-Squared value of Regression and the Correlation value of ANN, which gives a similar explanation. R-Squared value indicates the probability of success of future predictions and Correlation value indicated how close the predictions are to the true values. Correlation values are also analysed together with the error rate generation after the performance evaluation has been done by introducing more hidden layers.

Multilinear Regression test was done on same set of variables on the same sample dataset in Matlab Tool to compare the results. A comparative output is as shown in the Table 6.

**Table 6. Final Prediction Statistics using Deep Learning Algorithms**

| | R-Squared (R-Program) | ANN Correlation | R-Squared (Matlab) |
|---|---|---|---|
| International Status | 0.5359 | 0.5428 | 0.577 |
| Scholarship & Sponsorship | 0.4745 | 0.5572 | 0.594 |
| Reference & Feedback | 0.4403 | 0.2891 | 0.558 |
| Job Placement | 0.4279 | 0.4488 | 0.441 |

## 7. FINDINGS AND CONCLUSION

The findings as evident from the data suggests that faculty consultation is a key influencer predictor for performance of students in higher education institutes. This finding is similar to the findings reported by studies investigating predictors for student performance (Rivas et al., 2021; Kaur & Vadhera, 2021). The study has also found that class room management and exhibition participation also act as predictor for academic performance of students

For predictive analytics models to be successful at predicting outcomes, there needs to be a huge sample size representative of the population. If datasets are smaller than the predictive analytics models will be unduly influenced by anomalies in the data, which will distort findings. For this research, the data set is relatively small, hence the model prediction accuracy of 50 percent average is a fair value to consider the influencing factors and focus on the improvement areas. The Analysis has been done on sample data following the procedure of data cleaning, compilation, comparison, transformation, and running through different Machine Learning algorithms. The statistics of the algorithms have been further analysed and reported on the output values.

The future will see predictive analytics models play an integral role in every business. These models may not be perfect but they do offer immense value to organisations. With predictive analytics, organisations have the opportunity to take action proactively in a variety of functions.

## REFERENCES

Abu Zohair, L. M. (2019). Prediction of Student's performance by modelling small dataset size. *International Journal of Educational Technology in Higher Education*, *16*(1), 27. doi:10.1186/s41239-019-0160-3

Aggarwal, C. C. (2018). Neural Networks and Deep Learning. In *Neural Networks and Deep Learning*. Springer International Publishing. doi:10.1007/978-3-319-94463-0

Agrawal, R. S., & Pandya, M. H. (2015). Survey of papers for Data Mining with Neural Networks to Predict the Student's Academic Achievements. *International Journal of Computer Science Trends and Technology*, *3*(15).

Aher, S. B., & Lobo, L. M. R. J. (2011). Data mining in educational system using weka. *International Conference on Emerging Technology Trends (ICETT)*, 20–25.

Azcona, D., & Smeaton, A. F. (2017). Targeting at-risk students using engagement and effort predictors in an introductory computer programming course. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10474 LNCS, 361–366. doi:10.1007/978-3-319-66610-5_27

Beikzadeh, M. R., & Delavari, N. (2004). A New Analysis Model for Data Mining Processes in Higher Educational Systems. *Proceedings of the 6th Information Technology Based Higher Education and Training*, 7–9.

Bendangnuksung, & Prabu, P. (2018). Students' Performance Prediction Using Deep Neural Network. *International Journal of Applied Engineering Research, 13*(2), 1171–1176. http://www.ripublication.com

Campagni, R., Merlini, D., Sprugnoli, R., & Verri, M. C. (2015). Data mining models for student careers. *Expert Systems with Applications*, *42*(13), 5508–5521. doi:10.1016/j.eswa.2015.02.052

Clare, B. (2007). Promoting deep learning: A teaching, learning and assessment endeavour. *Social Work Education*, *26*(5), 433–446. doi:10.1080/02615470601118571

Coelho, O. B., & Silveira, I. (2017). Deep Learning applied to Learning Analytics and Educational Data Mining: A Systematic Literature Review. *Anais Do XXVIII Simpósio Brasileiro de Informática Na Educação (SBIE 2017), 1*(1), 143. 10.5753/cbie.sbie.2017.143

Cortez, P. (2010). Data mining with neural networks and support vector machines using the R/rminer tool. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6171 LNAI, 572–583. doi:10.1007/978-3-642-14400-4_44

Craven, M. W., & Shavlik, J. W. (1997). Using neural networks for data mining. *Future Generation Computer Systems*, *13*(2–3), 211–229. doi:10.1016/S0167-739X(97)00022-8

Deeley, S. J. (2014). Summative co-assessment: A deep learning approach to enhancing employability skills and attributes. *Active Learning in Higher Education*, *15*(1), 39–51. doi:10.1177/1469787413514649

Deng, L., & Yu, D. (2013). Deep learning: Methods and applications. *Foundations and Trends in Signal Processing*, *7*(3–4), 197–387. doi:10.1561/2000000039

Freund, R., Wilson, W., & Sa, P. (2006). *Regression Analysis* (2nd ed.). Academic Press. https://www.elsevier.com/books/regression-analysis/freund/978-0-12-088597-8

Gaudioso, E., & Méndez, L. J. T. (2005). Data mining to support tutoring in virtual learning communities: Experiences and challenges. *WIT Transactions on State-of-the-Art in Science and Engineering*, *2*, 207–225.

Hassoun, M. (1995). *Fundamentals of Artificial Neural Networks*. MIT Press.

Higham, D. J., & Higham, N. J. (2016). *MATLAB Guide*. Society for Industrial and Applied Mathematics.

Hussain, S., Muhsin, Z. F., Salal, Y. K., Theodorou, P., Kurtoğlu, F., & Hazarika, G. C. (2019). Prediction Model on Student Performance based on Internal Assessment using Deep Learning. *International Journal of Emerging Technologies in Learning*, *14*(8), 4–22. doi:10.3991/ijet.v14i08.10001

Kaur, N., & Vadhera, R. P. (2021). Predicting students' achievement in science from selected affective factors. *SN Social Sciences*, *1*(1), 33. doi:10.1007/s43545-020-00040-2

Kim, P. (2017). MATLAB Deep Learning With Machine Learning, Neural Networks and Artificial Intelligence. In *MATLAB Deep Learning*. Apress. doi:10.1007/978-1-4842-2845-6

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. doi:10.1038/nature14539 PMID:26017442

Leitner, P., Khalil, M., & Ebner, M. (2017). Learning analytics in higher education—a literature review. In Studies in Systems, Decision and Control (Vol. 94, pp. 1–23). Springer International Publishing. doi:10.1007/978-3-319-52977-6_1

Li, J., Wong, Y., & Kankanhalli, M. S. (2017). Multi-stream Deep Learning Framework for Automated Presentation Assessment. *2016 IEEE International Symposium on Multimedia (ISM)*, 222–225. doi:10.1109/ISM.2016.0051

Matloff, N. (2011). *The Art of R Programming: A Tour of Statistical Software Design*. No Starch Press.

Naser, S. S. A. (2012). Predicting learners performance using artificial neural networks in linear programming intelligent tutoring system. *International Journal of Artificial Intelligence & Applications*, *3*(2), 65–73. doi:10.5121/ijaia.2012.3206

Nawaz, R., Thompson, P., & Ananiadou, S. (2012). Identification of Manner in Bio-Events. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 3505–3510. http://www.lrec-conf.org/proceedings/lrec2012/pdf/818_Paper.pdf

Okubo, F., Shimada, A., Yamashita, T., & Ogata, H. (2017). A neural network approach for students' performance prediction. *ACM International Conference Proceeding Series*, 598–599. doi:10.1145/3027385.3029479

Piety, P. J., Hickey, D. T., & Bishop, M. J. (2014). Educational data sciences - Framing emergent practices for analytics of learning, organizations, and systems. *ACM International Conference Proceeding Series*, 193–202. doi:10.1145/2567574.2567582

Rivas, A., González-Briones, A., Hernández, G., Prieto, J., & Chamoso, P. (2021). Artificial neural network analysis of the academic performance of students in virtual learning environments. *Neurocomputing*, *423*, 713–720. doi:10.1016/j.neucom.2020.02.125

Schmidhuber, J. (2014). Deep Learning in Neural Networks: An Overview. *Neural Networks*, *61*, 85–117. doi:10.1016/j.neunet.2014.09.003 PMID:25462637

Siemens, G., & Long, P. (2011). Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE Review*, *46*(5), 30–40.

Sykes, A. O. (1993). *An Introduction to Regression Analysis*. Law School, University of Chicago. https://books.google.com.om/books?id=hxyaHAAACAAJ

Tam, V., Lam, E. Y., & Fung, S. T. (2012). Toward a complete e-learning system framework for semantic analysis, concept clustering and learning path optimization. *Proceedings of the 12th IEEE International Conference on Advanced Learning Technologies, ICALT 2012*, 592–596. doi:10.1109/ICALT.2012.66

Treasure-Jones, T., Sarigianni, C., Maier, R., Santos, P., & Rosemary, D. (2019). Scaffolded contributions, active meetings and scaled engagement: How technology shapes informal learning practices in healthcare SME networks. *Computers in Human Behavior*, *95*, 1–13. doi:10.1016/j.chb.2018.12.039

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. In *Computers in Human Behavior* (Vol. 89, pp. 98–110). Elsevier Ltd. doi:10.1016/j.chb.2018.07.027

Wang, L., Sy, A., Liu, L., & Piech, C. (2017). Deep knowledge tracing on programming exercises. *L@S 2017 - Proceedings of the 4th (2017) ACM Conference on Learning at Scale*, 201–204. doi:10.1145/3051457.3053985

Yegnanarayana, B. (2012). *Artificial Neural Networks*. Prentice Hall.