PREDICTIVE ANALYTICS USING DEEP NEURAL NETWORKS

Ranjana Durai[1], Krithika Selvam[2], Dr. S. Saranya[3]

[1,2,3] UG Scholars and Assistant Professor, Department of Electronics and Communication Engineering, Easwari Engineering College, Bharathi Salai, Ramapuram, Chennai, Tamil Nadu, India

Email: [1]iitianranj2017@gmail.com, [2]krithikaselvam1004@gmail.com, [3]nimmycharan@gmail.com

Abstract: Several studies indicate that teaching methods have a significant impact on the academic performance of the students. Similarly, they also connote a substantial dependance of a student's performance on the methods and techniques they employ in their study schedule. In this project, a model based on Deep Neural Network (DNN) has been deployed for the Predictive Analysis of a student's performance was tested for results. Furthermore, this model also aims at recommending study methodologies to students based upon the psychometric features derived from a self-analysis questionnaire that was prepared for the same.

Various factors that influence the performance of a student were identified and corresponding data was collected. This includes internal assessment scores, scores in the 10th grade and 12th grade board exams, attendance percentage for the particular semester and travel duration, availability internet connection facility, number of arrears among the others. This model was implemented on data of the student batches graduating in years in the duration of 2016-22. Test data evaluation presents an accuracy of 96.3% in the prediction of the student performance.

***Keywords*: Deep Neural Networks, Artificial Intellingence, Deep Learning, Machine Learning, Education, Technology, Computer Science, Artifical Neural Networks, Performance Prediction, Predictive Analytics, Learning Methodology Recommendation, Psychometric Evaluation and considerations in performance prediction**

1. Introduction

Numerous competitive exams in India are conducted for entrance into undergraduate and post graduate professional courses as well as for securing services in the government. The students undergo a lot of pressure from their family to score well in these exams in order to attain a admission slot in their dream institution or to secure the best possible designation as freshers. The pressure undergone by the students has manifested in multiple ways. Some students take it up as a challenge and prepare conscientiously for the exams whereas for the rest it has a negative influence and indulges them into the fear of not meeting the minimum cutoff set by the examination board. As a result, the subcontinental media observes a surge in the student suicide rates.

In India, students are required to enter higher secondary school after spending a minimum of five years of Primary Education. A student then spends a minimum period of five years in Secondary School at the end of which they are required to take the Secondary School Examination (SSE) after which they are required to uptake the All Indian Senior School Examination (AISSE).To get admission in colleges, students must have a minimum score of 40-50% in their AISSE examinations corresponding to their caste categories for an eligible candidature for NEET and a position in the top 20 percentile for their candidature in JEE. The eligibility criteria for GATE cites a minimum percentage of 50% in their undergraduate course. Considering the fact that almost all students possess the candidature eligibility criteria, the cutoff to achieve admission eligibility in a renowned institution poses the actual threat. The current standard states a requirement of 97%-30%, 58-33% and 94-55% in NEET, JEE mains and GATE respectively corresponding to their caste categories. For careers, Gate requires a score ranging between 84-47% corresponding to one's caste identity under the renewed list of castes.

This interactive approach seeks to predict the performance of the students and to recommend a befitting study methodology that when incorporated in their study session, gives the best performance in accordance with the effort they put in. The recommendation is personalized to each and every student based upon their psychometric characteristics derived from self-analyzing psychometric test provided to them. There are basically four categories of learners: Visual, Auditory, Kinesthetic and Read/Writing learner. Every learner falls into these categories and exhibits unique traits that are utilized to assort them based on the characteristic features inferred from self-analysis questionnaires provided in the model.

Here, the strategy implemented to deduce fruitful results is the DNN (Deep Neural Network). The DNN is an extension of ANN (Artificial Neural Network) and includes more than two hidden layers in the architecture. The architectural skeleton of the ANN consists of variable number of Dense class layers which imitate the functioning of the human brain. It has been deployed fruitfully to Speech

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Recognition, Analysis and adaptive control of Images, Predictive Analysis, Construction of software agents (3D games and computer) or robots.

2. Objectives:

The objectives of this project are:

1) To identify the factors influencing the performance of a student
2) To develop a predictive model for efficient academic prediction
3) To deduce vital features that help in categorizing the students based on their psychometric evaluation
4) To recommend competent learning methodology.

3. Literature Review:

This section consists of various papers on a variety of topics that were reviewed for coming up with an efficient approach. These include: Approaches to student performance prediction, Comparisons of efficient classification and prediction of such techniques, Dependance of the performance on various factors listed in the dataset collected and so on.

The study by Amirah Mohamed Shahiri and his colleagues in 2015 depicts the comparison between various approaches taken up in order to predict student performance so far. They summarize various techniques such as Neural Networks, Decision Trees, Support Vector Machine algorithm, K-Nearest Neighbors and the Naïve Byes Classifier method along with their accuracies in the descending order. He identifies and tabulates various factors that prominently influence and shape the efficiencies of each algorithm.

Adedeji (2001) noted the correlation between the scores of the student enrolment examination (UME) and their academic performance in Nigerian universities, using the school of Technology, Nigeria, as the case for the dataset. He investigated the connection of the scores in university level by utilizing the correlation and multivariate analysis . He concluded that a positive relationship exists between student admission scores and their UG performance. Recent trends following Adedeji's study, however, indicate the unreliability of the UME scores.

Geraldine Gray reports on the classification and regression models to identify college students at risk of failing in first year of study. Students were sampled from fourteen academic courses in five disciplines during the years 2010-12, and with diversity in their academic backgrounds. Metrics used included noncognitive psychometric indicators that can be assessed in the early stages after enrolment, specifically factors of personality, motivation, self-regulation and approaches to learning. Models were trained and yielded that classification models identifying students at risk of failing had a predictive accuracy greater than 79% on courses that had a significant proportion of high-risk students (over 30%).

4. Paradigms Applied: An Introduction

A.DNN

A computational model or statistical model that is influenced by the structure and/or functional features of biological neural networks is an artificial neural network (ANN), which is often typically referred to as a neural network (NN). The presence of more than one hidden layers is present in the structure, ANNs transform into what is known as Deep Neural Networks (DNNs)In contrast to traditional model-based methods, deep neural networks (DNNs) are a non-linear data driven self-adaptive approach. They are strong modelling tools, particularly when the underlying data relationship is unknown. NNs can identify correlated patterns between input data sets and corresponding target values and learn them. DNNs may be used to forecast the outcome of new independent input data after preparation and testing. It is said to replicate the human brain's learning process and can tackle situations concerning difficult and non-linear data, particularly though the data is imprecise and noisy.

B. Predictive Analytics

In order to make a fair assumption about the future, predictive analytics is a mathematical tool used to evaluate present and historical evidence. The model allows companies to determine patterns and trends and to support decision-makers in making effective decisions. The application cuts through numerous sectors and also plays a leading role in the flow of education. In order to provide accurate forecasts based on massive data sets fed by policy makers, it blends efficient analyzing technology with automated exploration. The datasets provided may be structured or un-structured. Together, these data are analyzed to classify hazards and trends that enable the meaningful forecast of potential events.

5. Framework:

Through an extensive literature search, a number of socio-economic, environmental, academic, and other factors have been identified that have an influence on a university student's performance. The influencing factors were categorized as input variables. In terms of the present school grading system, the output labels represent the levels of a candidate's performance. In addition to performance prediction, our model also seeks to suggest individual learners the most suitable learning approach based on their psychometric trait analysis.

6. The Dataset

The dataset utilized to test the above proposed model was generated from the student database of the batches graduating in the years 2015-2021 from the Electronics and Communication Engineering Department of SRM Easwari Engineering College, Chennai, India.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

A.PERFORMANCE PREDICTION

1) Input Variables: The factors that were determined to influence the scores of a student are listed below:

- Score in Continuous Assessment Test 1 (Internal assessment scores graded upon hundred)
- Scores obtained in CAT 2
- Scores obtained in CAT 3
- Secondary School Examination score (Class 10 board examination score)
- All India Senior School Examination score (Class 12 board examination score)
- Total hours of lecture per unit in the subject syllabus
- Total study hours spent by the student for the number of lecture hours for each unit.
- Attendance percentage in the current semester
- Average number of arrears per semester/ Failure percentage with respect to all the examinations undertaken by the student in the current semester.
- Number of hours of commute between the location of university and place of residence
- Availability of internet connection facility and
- Whether the student revised the concepts before taking the semester examinations.

Table I : Input Data Transformation

S/N	Input	Transformation	Formula
1.	CAT 1 score, CAT 2 score, CAT 3 score	Nil	Nil
2.	SSC score	CGPA to percentage wherever necessary.	Percentage= CGPA x 10
3.	Lecture hours per unit	Computed using data that requires the curriculum to follow the predefined hours of lectures for each subject.	Data given: 45-60 hours per subject. Average is calculated for the current semester= 45(x)+60(6-x) /6 =12
4.	Travel hours	Transformed as binary values depending upon the given criteria	Above 45 mins => 1 Below 45 mins => 0

5.	Internet connection availability	Binary values representing different categories	Yes => 1
			No => 0

6.	Revision	Binary values representing different categories	Yes => 1
			No => 0

The input features are transformed according to the specifications stated in Table 6.1.1, to achieve a better correlation between the factor.

2. Output Variables:

The output of the proposed model is a label that indicates the category into which the score obtained by the student falls into. The transformation is depicted in Table 6.1.2

Table II : Output Data Transformation

Output	Transformation	
Semester GPA	Percentage= GPA x 10 Sub classification into categories	
	Percent	Cadre
	0-4.99 5-	0
	9.99 10-	1
	14.99 15-	2
	19.99 20-	3
	24.99 25-	4
	29.99 30-	5
	34.99 35-	6
	39.99 40-	7
	44.99 45-	8
	49.99 50-	9
	54.99 55-	10
	59.99 60-	11
	64.99 65-	12
	69.99 70-	13
	74.99 75-	14
	79.99 80-	15
	84.99 85-	16
	89.99 90-	17
	94.99 95-100	18
		19

B.PSYCHOMETRIC EVALUATION

1) The input Values: Psychometrical evaluation of the student seeks to deduce insights on the various traits that are involuntarily exhibited from time to time. The evaluation is based upon the rating that the student gives on the self-analyzing questionnaire. A set of questions that gave the best insights to the evaluation were prepared to be put up on the were prepared to be put up in the questionnaire. The inputs were used to calculate the category scores of each classification of learning methodologies

2) The Classification: The output for the psychometric evaluation is the classification of the students which entirely depends upon their category score. The categories are decided as per the variety of predefined characteristics manifested in generations of learners under each category.

The four categories into which the students are classified into are:

- **VISUAL LEARNER**
- **AUDITORY LEARNER**
- **KINESTHETIC LEARNER**
- **READING/WRITING LEARNER**

7. Model

A. Topology

The topology of the model refers to the structural arrangement of the neural network. This feature illustrates the number of layers, the class in which the layers belong to, the activation functions deployed and so on.

The proposed model is based on the Multi-Layer Perceptron principle that has been developed for small-scale, static and simple analysis problems. The further specifications to be incorporated in the topology of the model is enlisted below

- Number of hidden layers
- Number of neurons per layer
- Activation functions to be deployed in each layer
- Loss functions to be applied

B. Splitting of Dataset

The records of 2000 students were used for the deployment of this model. The test and training set were split in the ratio of 0.29 and the validation data was partitioned from the training set in the ratio of 0.4. This yielded the dataset split of about 850 candidates as training data, 568 candidate's data as the validation set and 780 records for the test set.

C. Network Topology

The proposed model was deployed using 6 hidden layers with numerically decreasing value of neurons per layer. The output layer consisted of 20 neurons corresponding to the classes (cadre) under which the students are classified.

8. Training And Validation

The model was trained in batch sizes of 32 for 200 epochs with a validation split of 40%. The training efficiency was evaluated based on the given below loss function.

The Sparse Categorical Cross Entropy Loss

$$J(w) = (-1/N) \text{ x }$$
$$[\ \textstyle\sum_{i=1}^{N} Y_i \log(\tilde{Y}_i) + (1-Y_i)\log(1-\tilde{Y}_i)]$$

Where,

- W refers to the Weights of the Neural Network
- Y_i is the TRUE label
- $\tilde{Y}_i$ is the PREDICTED label.

This formula applies for both categorical cross entropy as well as for the case of sparse categorical cross entropy. The value of Yi is the only fact that differentiates the use of the formula for each case.

The Yi format determines the role of loss to be used. Using one hot encoded format of Yi along with categorical cross entropy is highly recommended. In addition, in models that consist of samples that may belong to multiple other classes, the categorical cross entropy loss function is implemented.

To deploy sparse categorical cross entropy, models with samples belonging to mutually exclusive classes are suggested. The nature of Yi remains discrete because of this trait. Such a format performs well hen paired with the sparse categorical cross entropy loss function.

9. Results

After the training and validation process was implemented as per the specifications mentioned above, the performance measures used to evaluate the model (DNN) for multi-class classification methods are:

- Accuracy
- Precision
- Recall
- F1 score.

The precision is defined as the proportion of relevant results in the list of all returned search results and the recall is defined as the ratio of the relevant results returned by the search engine to the total number of the relevant results that could have been returned.

As per the definition of precision, it gives us the ratio of True positive cases (TP) to all the expected positive set of cases (TP+FP) and the formula of recall is given as:

$$\text{Recall} = \frac{TP}{TP+FN}$$

High precision and recall value suggest that, relative to misclassified cases, the model returns a larger percentage of correctly classified cases.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

The accuracy metric provides an efficiency measure based upon the number of correctly predicted data points out of all the data points.

The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall.

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})}$$

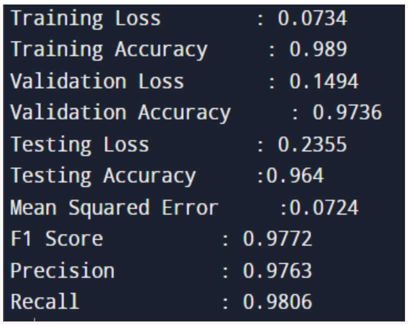
Where,

- TP -> number of TRUE positive cases
- FP -> number of FALSE positive cases
- FN -> number of FALSE negative cases

Our model yields a loss value of 0.0734 along with an accuracy of 98.9% in the training set and the values 0.23 and 96.4% as the testing loss and test set accuracy respectively. The validation loss is calculated over the 200 epochs and settles at a minima at the value of 0.1494.

Results of the metrics passed at the model.fit() function are given in the image enclosed below.

The personalized learning methodology recommendation gives and output as a printed statement which displays the best methodology suitable to the individual based solely upon his/her choices that were marked in the self-analysis questionnaire.



```
Training Loss        : 0.0734
Training Accuracy     : 0.989
Validation Loss       : 0.1494
Validation Accuracy     : 0.9736
Testing Loss         : 0.2355
Testing Accuracy     :0.964
Mean Squared Error     :0.0724
F1 Score           : 0.9772
Precision          : 0.9763
Recall             : 0.9806
```

10. Conclusions And Further Enhancement

This research has shown the potential of Deep Neural network for student performance prediction to help them gain insights into the study method they are using to prepare for competitive or semester examinations. The model has achieved an accuracy of over 97 percent, thus demonstrating the potential effectiveness of Deep Neural Networks as a prediction tool and a selection criterion for applicants seeking career opportunities or admission to a university. It is possible to further improve this project by including emotional intelligence concepts that help to analyse the emotional stability of a student throughout the entire duration of the course. By using it as an attribute to whether or not a counselling session needs to be allocated, this information can be put into a fair use. This enhancement is solely proposed for the benefit of students by bringing in changes in the methods of teaching and so on.

References

[1] E. Osman Begovic, M. Sulji ´ c "Data mining approach for predicting student performance" in *Economic Review 10*

[2] S. T. Jishan, R. I. Rashu, N. Haque, R. M. Rahman, "Improving accuracy of student's final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique", *Decision Analytics 2* (1) (2015)

[3] D. M. S. Anupama Kumar, "Appraising the significance of self-regulated learning in higher education using neural networks", *International Journal of Engineering Research and Development Volume 1* (Issue 1) (2012)

[4] V. Oladokun, A. Adebanjo, O. Charles-Owaba, "Predicting students' academic performance using artificial neural network: A case study of an engineering course", *The Pacific Journal of Science and Technology* 9 (1) (2008)

[5] P. M. Arsad, N. Buniyamin, J.-l. A. Manan, "A neural network students' performance prediction model (nnsppm)", in: *Smart Instrumentation, Measurement and Applications (ICSIMA)*, 2013 *IEEE International Conference* on, IEEE, 2013, pp.

[6] G. Gray, C. McGuinness, P. Owende, "An application of classification models to predict learner progression in tertiary education", in: *Advance Computing Conference (IACC)*, 2014 *IEEE International*, *IEEE*, 2014, pp. 549–554.

[7] Wang, A. Mitrovic, "Using neural networks to predict student's performance", in: *Computers in Education*, 2002. *Proceedings. International Conference* on, *IEEE*, 2002.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731