A Review On Predicting Student Performance Using Deep Learning Technique

Dr. S. MAHESWARI

Associate Professor, National Engineering College, Kovilpatti, Tamil Nadu, India
Email: maheswaricse@nec.edu.in

PREETHI J

PG Student, National Engineering College, Kovilpatti, Tamil Nadu, India
Email: 182805@nec.edu.in

Abstract

Predicting student's success based on this academic performance becomes more complicated for the educational organization. Determine the deep learning technique to enhance the performance of the student's placement. The deep learning method could guide welfare and effect on educators, students, and educational institutions. The dataset is collected from the student's academic records. The dataset comprises variant information regarding their previous and present academic records and then apply a Logistic Regression algorithm using the Anaconda platform for examining the student's success in their campus placement. The proposed methodology denotes the classification algorithms for analysis from the student dataset to the predicted outcome. This project can determine the relations of student's academic success and their investment in placement selection.

Keywords: *Classification, deep learning, Logistic Regression, placement, prediction, student performance.*

1.INTRODUCTION

Placement is a process of allocating a secure job to each of the selected candidates. Placement is not an easy procedure and very hard to adapt to a new employee who is entirely unknown to the job environment. For this purpose, the employee is generally put on a probation time ranging from one year to two years. Concluding that the trial period, if the employee shows excellent performance, then the employee is confirmed as a regular employee of the organization. Thus, the trial period is the transition period at the end of which management takes a decision whether to make the employee consistent or discharge him from the job.The term placement means assigningthe specific job rank and responsibility to a newly selected and appointed employee. Basically, it is matching the employee to the job requirement. A candidate should be placed on the right job is a real motive.A student should be placed in the campus selection according to the necessity of the work, such as sensual and conceptual ability, sight, hear, etc. The job shouldn't be regulated according to the qualification and abilities of the employees.

Providing placement opportunities increases an institution's reputation for the graduate employee as standard placements are detected to upgrade student employability. Therefore, the institution's well-being from more satisfied students and graduates having a betterquality experience. This could seriously enhance the attractiveness of the education organization to prospective students who are progressively mindful of the worth of their studies to employers and improving their job opportunities. The profile chart displays an evaluation of both job requirements and student abilities for critical features of the job so that management can quickly determine how well a candidate fits a job. Students who are completing their higher education will join the universities to get a better job. Placement is essential for both the organization and the staff that each student should be placed on a better job.

This paper can denote the association ofthe campus placement for the students of particular institutions based on their past and present academic reports. Apply a Logistic Regression algorithm to predict the student's performance in placement and examine which algorithm gives better accuracy for the given dataset.

1.1. Logistic Regression:

Logistic Regression is one of the famous deep learning algorithms for binary classification. This is a supervised classification algorithm. In a classification problem, the target variable Y can take only discrete values for a given set of features X. Logistic Regression frame a regression model to project the possibility that a given input entry belongs to the category numbered as "1". Logistic Regression is the go-to method for binary classification. It provides a discrete binary outcome between 0 and 1.

Logistic Regression is used in disparate fields, including machine learning, deep learning, medical fields, and social sciences. Many other medical scales used to assess the harshness of a patient have been refined using logistic Regression. Logistic Regression is allowed to predict the exposure of establishing a given disease (e.g., diabetes, and heart disease), based on observed characteristics of the patient (age, sex, body mass index, results of various blood tests, etc.). The approach can also be used in engineering, specifically for concluding the feasibility of loss in a given development, structure, or device. Logistic Regression is also used in marketing applications such as the prediction of a customer's tendency to investment a product or stop a contribution, etc. In economics, Logistic Regression can be used to predict the probability of a person's choosing to be in the labor force. Conditional random fields, the addition of logistic regression to sequential data, are used in natural language processing.

2. MATERIALS AND METHODS

2.1. Anaconda Navigator

Anaconda is a free and open-source dispensation of the Python and R programming languages for scientific computing (data science, deep learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. Anaconda is well-liked because it brings many of the tools used in data science and deep learning with one installation, so it's significant for having a small and comfortable structure. Anaconda also uses the theory of creating environments so as to isolate different libraries and versions. In the prediction system, Anaconda gives better performance to determine the accuracy of the trained dataset. The large volume of the dataset is used to predict the outcome of the results in this platform. Using Anaconda Navigator, especially for

prediction in python language, is a trouble-free platform. In Anaconda, using the jupyter tool for executing the codes is unchallenging.

2.2. Logistic Regression Algorithm

Logistic Regression is one of the most elementary and widely used deep learning algorithms. Logistic Regression is a supervised learning classification algorithm used to predict the possibility of a target variable. Logistic Regression does not use a regression algorithm but a possibilities classification model. It is a predictive scanning algorithm and based on the concept of probability. Logistic Regression uses a more compound cost function, and this cost purpose can be explained as the 'Sigmoid function' or also familiar as the 'logistic function.' The hypothesis of logistic regression tends to maximize the cost responsibility between zero and one. Therefore, linear tasks fail to represent it as a value greater than one or less than zero, which is not possible as per the hypothesis of logistic regression. The nature of the target or dependent variable is dichotomous, which means there would be only two possible classes. The dependent variable is binary in identity, having data coded as either 1 or 0.

2.3. A multiclass logistic regression model

In a multi-classification problem, using the training dataset with a classification algorithm, the trained classifier is to predict the target value out of more than two possible outcomes. In this project, the multi-class problem is used to predict the placement results with a "multinomial" algorithm. Multinomial Logistic Regression, also known as "softmax regression," is used to find the predicted probability of each class. It is a supervised learning algorithm that can be used in several problems, including "text classification." According to this project, Predicting the campus placement results using the given student features, and the target outcomes are CTS, TCS, Zoho, and Not Placed.

3 PROPOSED SYSTEM

Predict the student placement from the dataset. The proposed system denotes the Logistic Regression algorithm for analyzing the outcome. From the dataset collection, eliminate the redundant data and fill the lost values. Then, fitting the multi-class logistic regression algorithmto the trained dataset. Predict the test results of the trained dataset. Evaluate the probability for true and false class and find the accuracy, recall, precision by using a confusion matrix. Logistic Regression algorithm is predicting the best outcome of positive and negative class in the confusion matrix.

3.1. Dataset Collection

Collect the datasetfrom the students. The dataset contains academic records of each student, personal data, and their information. On the other hand, another dataset was collected from the company that provides input from their demand.

3.2. Data preprocessing

Inpreprocessing, eliminate the redundant feature from the dataset and lost data filling. Then, apply the unified databasemethod to the dataset. This module can remove replicated data. Eventually, use the regression model to the dataset, which can predict the class. Preprocessing can produce a training dataset.

3.3. Prediction

Prediction refers to the outcome of an algorithm after it has been trained on a historical dataset and applied to the given dataset. It predicts a particular issue. In this module, apply the logistic regression to the trained dataset, and the logistic regression algorithm is used to predict the results for each student.

3.4. Classification

Classification is a procedure of categorizing a given set of data into classes, and it can be achieved on both shaped or unshaped data. The system starts by predicting the type of information provided. The models are frequently referred to as target, label, or categories. The classification predictive modeling is the duty of approaching the mapping function from input variables to discrete output variables. The main goal is to identify the class or category of the dataset.

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k \qquad (1)$$

Where, Y is the predicted or expected value of the dependent variable, X_1 through X_k are k distinct independent or predictor variables, b_0 is the value of Y when all of the independent variables (X_1 through X_k) are equal to zero, and b_1 through b_p is the estimated regression coefficients. Logistic Regression classifiers can be effectively used to train small data sets and a classifier that can manage high dimensional data samples.

3.5. Performance Metric

- The predicted data is then calculated from the confusion matrix.
- For that, calculate True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN).
- Evaluate the accuracy of the dataset.
- Simulate the recall and precision values.

Considering the genuinely positive data points plotted as true positive, and genuinely negative data points as true negative. (TP+TN)/(TP+FP+FN+TN) determine the accuracy of the given dataset. The precision is provided by the true positive cases to the sum of true positive and false positive class (TP/(TP+FP)). The capacity of the estimator recognizes relevant instances using (TP/(TP+FN)), and harmonic mean is used to examine the F1 score, which is a union of recall and precision values.

3.6. Accuracy:

Accuracy (ACC) is denoted the number of all corrected predictions divided by the total number of the dataset. The best skill is 1.0, whereas the worst accuracy is 0.0. Efficiency can also be calculated as 1 – ERR(Error)

$$\text{Accuracy}=(TP+TN)/(TP+FP+FN+TN) \qquad (2)$$

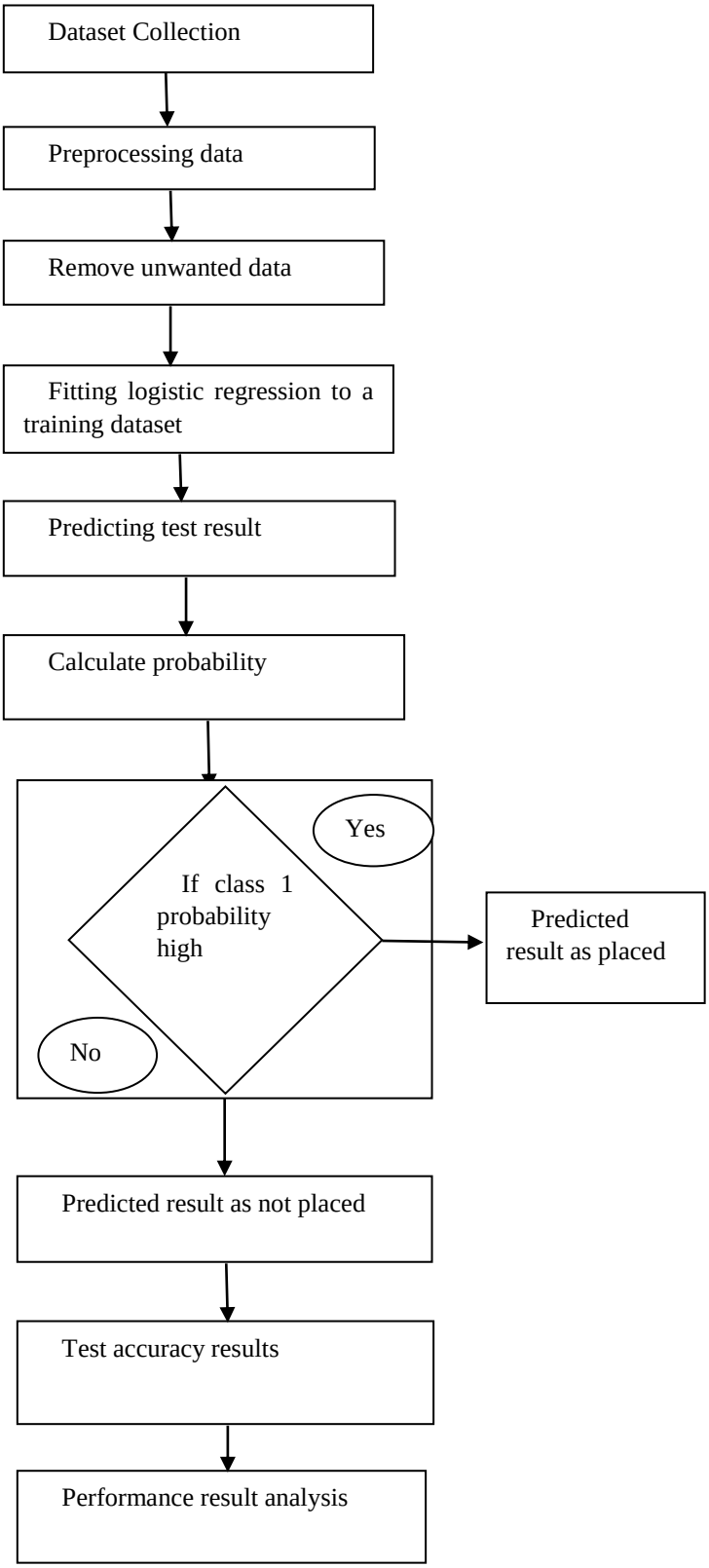

Figure 1: Work Flow Diagram

3.7. Precision:

Precision is the balance of the correctly positive labeled by our program to all positive tagged. Precision is calculated by the number of correct positive predictions is divided by a total number of positive predictions. Precision is also called a Positive Predictive Value (PPV). The best precision value is 1.0, whereas the worst is 0.0.

Precision = TP/(TP+FP) (3)

3.8. Recall:

Sensitivity (SN) is calculated by the number of accurate positive predictions divided by the total number of positives. Sensitivity is also called recall (REC) or true positive rate (TPR). The best sensitivity value is 1.0, whereas the worst is 0.0.

Recall = TP/ (TP+ FN) (4)

4 IMPLEMENTATION

4.1 Dataset Collection

Collecting dataset from the student, which contains student personal, academic records, extra activities, and their information. The student's data ((i.e.), Email, Name, Register Number, Department, Year, 10^{th} %, 12^{th} cut off, Medium, Mode, No. of outside participation, No. of online course completed, carrier interest, Skill Set, No. of arrears, CGPA) are gathered.

Table 1: Sample Collected Data

Name	10th	12th	CGPA	Year	Arrear	Department
Sankar	92	167	7.7	2	0	CSE
Balaji M	93	136.75	6.4	2	4	CSE
Naveen	89	138.5	7.0	2	1	CSE
Krishna	90	153	7.9	2	0	CSE
Ramya	94	144	7.9	4	0	CSE
Swetha	98	158	8.5	4	0	CSE
Kokila	80	138	7.6	4	0	CSE
Arun S	98	178	9.4	4	0	CSE
Surya	82	144	7.7	3	0	CSE
Rahul	97	134	8.5	4	0	CSE
Vignesh	94	149	8.3	3	0	CSE
Gokul	92	150	8.7	4	0	CSE
Rajesh	94	141	7.4	4	2	CSE
Aparna S	95	163	8.5	3	0	CSE
Dhanya M	95	160	8.7	4	0	CSE
Sujith	91	148	8.2	4	0	CSE
Sri mathi	90	136	7.9	4	0	CSE
Merlin	97	158	8.0	2	0	CSE
Bharathi	81	117	7.1	2	2	CSE

The dataset is collected from the company. It contains input from the company on their demand.

Table 2: Company Dataset

Company	10th	12th	CGPA	Communication Skill	Person Skill	Arrears
Zoho	90	190	9.0	1	1	0
CTS	80	180	8.0	1	1	2
TCS	80	160	7.0	0	0	1

4.2 Preprocessing

Eliminateredundant features from the dataset and lost values of data filling. It can remove the replicated data. Finally, apply the logistic regression to the trained dataset, which can predict the class. In preprocessing, generate the training data from the collected student's dataset.

Table 3: Pre-processed data

Name	10th	12th	CGPA	Year	Arrear	Department
Sankar	92	167	7.7	2	0	CSE
Balaji M	93	136.75	6.4	2	4	CSE
Naveen	89	138.5	7.0	2	1	CSE
Krishna	90	153	7.9	2	0	CSE
Ramya	94	144	7.9	4	0	CSE
Swetha	98	158	8.5	4	0	CSE
Kokila	80	138	7.6	4	0	CSE
Arun S	98	178	9.4	4	0	CSE
Surya	82	144	7.7	3	0	CSE
Rahul	97	134	8.5	4	0	CSE
Vignesh	94	149	8.3	3	0	CSE
Gokul	92	150	8.7	4	0	CSE
Rajesh	94	141	7.4	4	2	CSE
Aparna S	95	163	8.5	3	0	CSE
Dhanya M	95	160	8.7	4	0	CSE
Sujith	91	148	8.2	4	0	CSE
Sri mathi	90	136	7.9	4	0	CSE
Merlin	97	158	8.0	2	0	CSE
Bharathi	81	117	7.1	2	2	CSE

4.3 Remove Redundant Data

Applying the unified database module to the dataset. Finally, use the logistic regression to the trained database, which can predict the class. The preprocessed data contains predicted attributes, i.e., 10^{th} %, 12^{th} cut off, CGPA, No. of outside participation, and Arrear.

Table4: Trained dataset

Name	10th	12th	CGPA	Person Skill	Arrears
Sankar	92	167	7.7	0	0
Balaji M	93	136.75	6.4	2	4
Naveen	89	138.5	7.0	0	1
Krishna	90	153	7.9	3	0
Ramya	94	144	7.9	5	0
Swetha	98	158	8.5	1	0
Kokila	80	138	7.6	5	0
Arun S	98	178	9.4	2	0
Surya	82	144	7.83	3	0
Rahul	97	134	6.52	2	0
Vignesh	94	149	8.35	4	0
Gokul	92	150	8.93	2	0
Rajesh	94	141	7.89	3	2
Aparna S	95	163	8.1	1	0
Dhanya M	95	160	9.76	2	0
Sujith	91	148	9.1	2	0
Sri mathi	90	136	8.7	2	0
Merlin	97	158	8.72	3	0
Bharathi	81	117	8.8	2	2

4.4 Predicted Result

The predicted results for each studentcontain two types of classes they are placed and not placed. The predicted attributes of 10^{th}, 12^{th}, CGPA, Communication skill, Person skill, and Arrears. Apply the logistic regression to the trained dataset, and the logistic regression algorithm is used to predict the results for each student.

Table 5: Predicted Result

Name	10th	12th	CGPA	Person Skill	Arrears	Predicted result	Companies
Balaji M	93	136.75	6.4	0	4	Not Placed	-
Sankar	92	167	7.7	2	0	Placed	Zoho
Bharathi	81	117	7.1	1	2	Not Placed	-
Kokila	80	138	7.6	0	0	Not Placed	-
Gokul	92	150	8.7	3	0	Placed	CTS
Sujith	91	148	8.2	2	0	Placed	Zoho
Merlin	97	158	8.0	5	0	Placed	CTS
Sri mathi	90	136	6.4	5	0	Not Placed	-
Arun S	98	178	9.4	3	0	Placed	CTS

4.5 Calculate Probability

The calculation of probability denoted the predicted results and determined the accuracy, precision, and recall.

The calculation for two students:

$Y=b_0+b_1x_1+b_2x_2$

Input: The dependent variable is denoted Y. It contains the value 1 and 0 for placed and not placed respectively, and the independent variable denoted X1 and X2. It contains the value of the 10th percentage and current CGPA for two students. The input for one student is 10th percentage = 52 and CGPA = 6.2 and the input for other student is 1oth percentage = 94 and CGPA = 8.9.

$$b0 = \bar{y} - b1\overline{x1} - b2\overline{x2}$$
$$x1 = 146\ and\ x2 = 15.1$$
$$b1 = 0\ and b2 = 0$$
$$b0 = \frac{1}{2} - 0 * \frac{146}{2} - 0 * \frac{15.1}{2} = 0.5$$
$$y = 0.5 + 0x1 + 0x2$$

Among the two students, one student gets placed, and one student is going to get not placed. Similarly, the remaining values are calculated.

5 RESULT AND DISCUSSION

According to the classifier algorithm, based on analysis from the student dataset, the algorithm gives efficient accuracy results. Training data determines to remove the replicated data, lost values filling, and eliminated the redundant attributes. Apply the logistic regression to the training dataset, and the algorithm is used to predict the results for each student. The classification part stimulated the predicted attributes and finding the predicted results using Logistic Regression to find accuracy, precision, and recall.

Table 6: Confusion matrix results

Algorithm	TP	TN	FP	FN	Accuracy
Logistic Regression	54	33	7	6	0.87

Accuracy=54+33/54+33+7+6=0.87
Precision=54/54+7=0.88
Recall=54/54+6=0.90

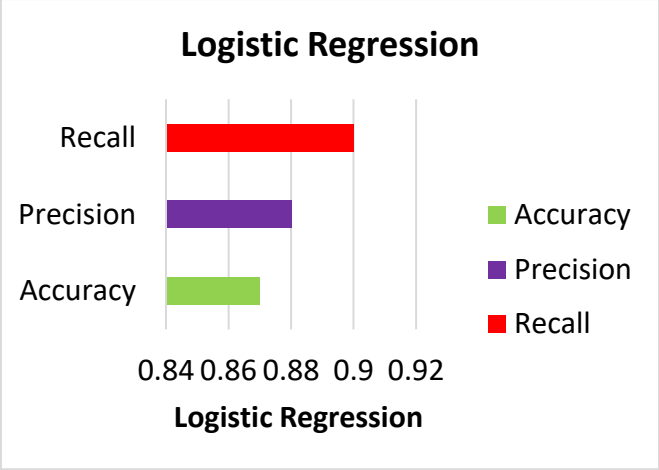

Figure2: Prediction Accuracy for Logistic Regression

The imaginary form representsthe confusion matrix using the classifier algorithm. The accuracy of the logistic regression algorithm is 0.88. The precision and recall values for the classifier are 0.90 and 0.88, respectively.

6 CONCLUSION

The reason for choosing the logistic regression algorithm is that it is easy to explain and well predict the output compared to the other classification methods. The multi-class logistic regression algorithm gives better performance in the predicted outcome. In this proposed logistic regression model, the final result authorizes the educational institutions and analyzes the students who are getting the least prospect in their current academic year.

References

[1] Bashir Khan, et al.," Final Grade Prediction of Secondary School Students using Decision Tree," International Journal of Computer Applications (0975 – 8887) Volume 115 – No. 21 April 2015. b

[2] HeenaSabnani, Mayur More, Prashant Kudale, Prof. Surekha Janrao, "Prediction of Student Enrolment Using Data Mining Techniques," International Research Journal of Engineering and Technology (IRJET) Volume: 05 Issue: 04 | Apr-2018.

[3] Mrs.Varsha. P. Desai, "Classification Technique for Predicting Learning Behaviour of Student in Higher Education" International Journal of Trend in Scientific Research and Development (IJTSRD) Oct-2018.

[4] Maria Koutina and Katia LidaKermanidis, "Predicting Postgraduate Students' Performance Using Machine Learning Techniques" IFIP International Federation for Information Processing 2011.

[5] Amirah Mohamed Shahira, et al., "A Review on Predicting Student's Performance using Data Mining Techniques" ELSEVIER/ (PCS)Procedia Computer Science 72 (2015) 414 – 422

[6] Manisha Sahane, et. Al, "Prediction of Primary Pupil Enrollment in Government School Using Data Mining Forecasting Technique" International Journal of Advanced Research in Computer Science and Software Engineering 4(9), September - 2014, pp. 656-661

[7] Amandeep Kaur, et al., "Machine Learning Approach to Predict Student Academic Performance" International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 6 Issue IV, April 2018

[8] Havan Agrawal and HarshilMavani, "Student Performance Prediction using Machine Learning" International Journal of Engineering Research & Technology (IJERT) Vol. 4 Issue 03, March-2015

[9] Ajay Kumar Pal and Saurabh Pal, "Classification Model of Prediction for Placement of Students" I.J.Modern Education and Computer Science, 2013, 11, 49-56.

[10] Anal Acharya et al.," Early Prediction of Students Performance using Machine Learning Techniques" International Journal of Computer Applications (0975 – 8887) Volume 107 – No. 1, December 2014.

[11] U. K. Pandey, et al., "Data Mining: A prediction of performer or underperformer using classification" International Journal of Computer Science and Information Technology(IJCSIT), Vol. 2(2), pp.686-690, ISSN: 0975-9646, 2011.

[12] Z. J. Kovacic, "Early prediction of student success: Mining student enrollment data," Proceedings of Informing Science.

[13] Ioannis E. Liveris, Tassos A. Mikropoulos, et al., "A decision support system for predicting students' performance" Themes in Science & Technology Education, 9(1), 43-57, 2016

[14] Alaa Khalaf Hamoud, "Selection of Best Decision Tree Algorithm for Prediction and Classification of Students' Action" (AIJRSTE)American International Journal of Research in Science, Technology, Engineering &MathematicSSN (Print): 2328-3491, ISSN (Online): 2328-3580, ISSN (CD-ROM): 2328-3629.

[15] Ravi Kumar Rathore, J. Jayanthi, Gurpreet Kaur, "Student Prediction System Using Data Mining Algorithm- Survey" International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 5 Issue VI, June 2017 ISSN: 2321-9653

[16] Hashmi Hamsa, Simi Indiradevi, Jubilant J, et al., "Student academic performance prediction model using decision tree and fuzzy genetic algorithm" Global Colloquium in Recent Advancement and Effectual Researches in Engineering, Science, and Technology (RAEREST 2016)

[17] Dr.Anjali B Raut, Ms. Ankita A Nichat, "Students Performance Prediction Using Decision Tree Technique" International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 13, Number 7 (2017), pp. 1735-1741

[18] Manmohan Singh, Harish Nagar, Anjali Sant, "Using Data Mining to Predict Primary School Student Performance" IJARIIE-ISSN(O)-2395-4396 Vol-2 Issue-1 2016

[19] S. Kotsiantis et al., "Predicting Students' Performance in Distance Learning Using Machine Learning Techniques" Applied Artificial Intelligence, 18:411--426, 2004 Copyright # Taylor & Francis Inc. ISSN: 0883-9514 print/1087-6545 online DOI: 10.1080=08839510490442058

[20] Raheela Asif et al., "Predicting Student Academic Performance using Data Mining Methods" IJCSNS International Journal of Computer Science and Network Security, VOL.17 No.5, May 2017