International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Student Strategizing in Education system using a Machine Learning Model

H S Divyashree[1 *], Avinash N[2], M. Sasi Kumar [3,] S. Dinesh[4]

[1,2,3*]Dept. of CSE, Brindavan College of Engineering, Bangalore, India
[4]Dept. of ISE, Brindavan College of Engineering, Bangalore, India,

[*]*Corresponding Author: divyahosur.shree@gmail.com, Tel.: +91 9972935570*

Available online at: www.ijcseonline.org

Accepted: 18/July/2018, Published: 31/July/2018

Abstract— Strategizing is an important aspect which requires critical analysis to determine performance. The key to solve this issue is by tapping the available student talent within the university. In this paper, we have done research in the domain of education. Strategy considered in the research is assessing the skill set of the students. For this approach, we have constructed our Vision Based Page Segmentation algorithm to extract the data from the university. In Unsupervised Machine Learning and Supervised Machine Learning, We have taken Classification and Regression supervised learning to classify the student's marks. Machine learning models like Neural Network, Random Forest and Logistic Regression have been used to predict the best student team.

Keywords— Strategizing, Neural Network, Random Forest and Logistic Regression

I. INTRODUCTION

There have been several instances in history where the introduction of analytics and statistics completely revolutionized the field of education. The university dataset that has been used for this analysis to provide statistics of about 50,000 students on 30 different attributes, 10 of which are relevant for classification. These attributes are used to find the best student in the university. We have used machine learning classification algorithm to find the best student using marks, attendance and grade.

In the field of computer science Machine learning (ML) is the science of getting computers to act without being explicitly programmed. Machine Learning algorithm such as Decision tree, Naive Bayes Classification, Logistic Regression, Support Vector Machines uses to perform adjective methods to learn information directly from data without predefined equation as a model.

People are using Machine Learning every day to make critical decisions in medical diagnosis, stock trading, energy load forecasting, and more. We have used Machine Learning Algorithm to achieve best student team in university.

There are two types of techniques in Machine Learning one is supervised learning and another one is unsupervised learning. Supervised learning can predict future output on known input. It uses classification technique to classify input data in to three categories such as first class students, second class students and distinction students. Whereas unsupervised learning is used to predict hidden patterns.

To build a strong team, students achieving at every position will give their best performance at that position. This student classification technique can be done using statistical models being used today to make the task faster and more optimized. We trained Machine learning models for this analysis.

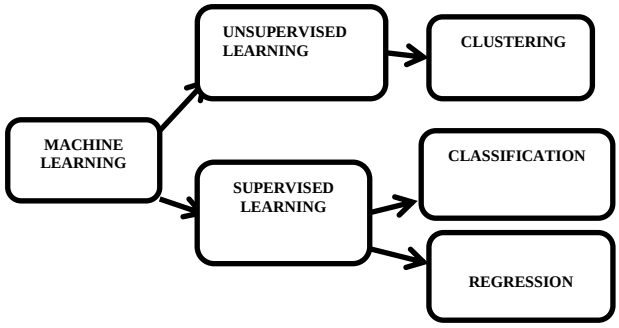

Fig.1: Machine Learning Techniques

Machine learning models are neural network, logistic regression and random forests have been used for extraction of data and making a clear data set. Input the train data with labelled to the classifiers. After the phase of the training, fed to the trained classifiers is the test data whose output class need to be predicted various performance matrices like marks, attendance and grade have been calculated. To

measure the better performance of the model matrix has been plotted.

The construction of the remaining part of the paper is as follows. Section II deals the data set and machine learning models. Overall methodology of research paper presented in section III.Section IV shows the experimental result with the short discussion. Conclusion is given in the section V.

II. BACKGROUND

A. Dataset

Vision Based Page Segmentation algorithm is developed to extract data from university ,the student personnel data along with USN and their name and phone no, address ,email id, gender were collected for further analysis.

The data set contain of around 30 features of which only 10 related features were filtered out Total number of student in the dataset around 50000.Some of the features contain marks in each subject and attendance ,behaviour etc.70% of the data is utilized for the training and the rest 30% is for the testing.

To detect content from university we have used VIPS tag tree approach. VIPS (Vision Based Page Segmentation) algorithm is to transform a student's details into a visual block tree. A visual block tree is actually a segmentation of a web page. The root block represents the whole page, and each block in the tree corresponds to a rectangular region on the web pages.

The leaf blocks are the blocks that cannot be divided further, and they represent the minimum units, such as continuous texts or student images. These block tree is constructed by using DOM (Document Object Model) tree. There is a one main building component in the VIPS algorithm that is DOM tree.

The DOM tree is used to manage XML data or access a complex data structure repeatedly. The DOM is used to Builds the data as a tree structure in memory, Parses an entire XML document at one time, Allows applications to make dynamic updates to the tree structure in memory. Every detail of the students is shown in an XML document. The processor analyses the mark up and passes structured information to an application. Web Pages are designed by using html files and xml files. Now days the web page designers are increasing the complexity of html source code. So we have used VIPS algorithm to extract the data easily.

B. Artificial Neural Network

A collection of connected unit of artificial neurons are called Artificial Neural Network (ANN).Each connection in ALL can transmit a signal from one artificial neurons to another artificial neurons. ANN used Multilayer Perceptron in our approach.

A Multilayer Perceptron (MLP) is a class of feed forward artificial neural network. An MLP consists of at least three layers of nodes. Each node is a neuron that uses a nonlinear function, except for the input nodes.

A Multilayer perceptron in Artificial Neural Network with a single hidden layer can be represented graphically as follows:

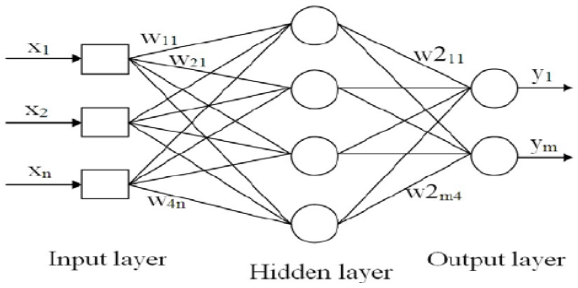

Fig. 2: MLP with single hidden layer

One-hidden-layer in MLP is a function, $f : R^{D} \rightarrow R^{L}$ where D is the input vector size and L is the output vector size of $f(x)$ In matrix notation: F(x) is written as follows: $F(x) = G\left(b^{2} + w^{2}\left(s\left(b^{1} + w^{1}x\right)\right)\right)$ Where b^{1}, b^{2} are bias vectors and w^{1}, w^{2} are weight matrices and activation functions G and s where s include $\tanh(a) = \left(e^{a} - e^{-a}\right) / \left(e^{a} + e^{-a}\right)$ .We will be using tanh in this tutorial because it typically yields to faster training (and sometimes also to better local minima)[1][2].

C. Logistic Regression

Logistic model is a statistical model where the binary dependent variable is taken. It has divided into 2 types based on the binary dependent variable.

1.Binary Logistic Regression: A BLR is a statistical technique is used to predict the relationship between independent variable and dependent variable where the dependent variable can be either continues or categorical.

2. Multinomial Logistic regression: Multinomial Logistic regression is the linear regression analysis use to predict outcome variables. Thus it is an extension of logistic regression, which analyses dichotomous (binary) dependents. In our paper, since the dependent variable is multiclass and thus multinomial logistic regression is used. The input set of the independent variables which might be binary, categorical etc.

D. Random Decision Forests

Random Decision Forests is a supervised learning method for classification, regression and other tasks, which conducted by constructing a multitude of decision trees during training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees [3][4].In case of classification, the final output of the classifier is the mode of the classes predicted by the individual decision trees.

III.METHODOLOGY

The aim of our approach is to get the best student depending on their skill set. In this three output classes are determined: Marks, Attendance and Grade. The complete flow of the process can be shown in below figure.

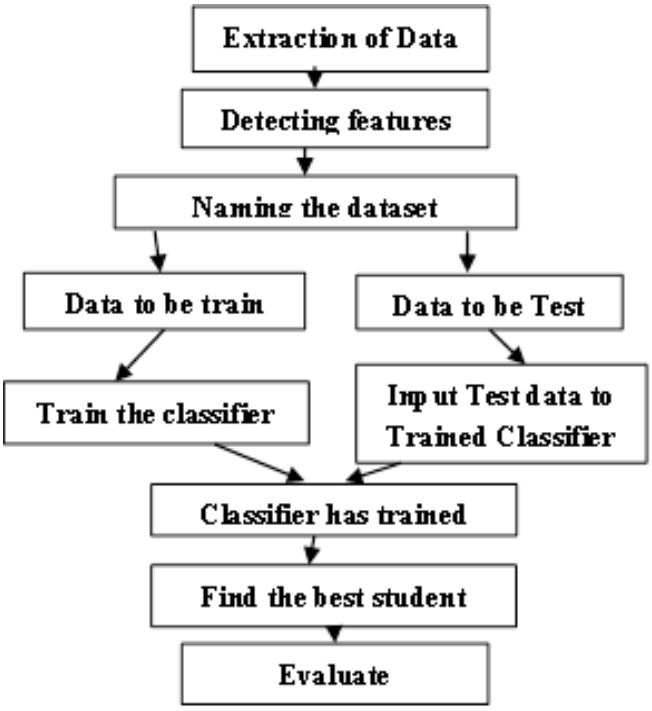

Fig.3: Complete flow of the procedure

1) Data Extracting: The process begins with data extracting from the university which has around 50,000 students with 30 features. The attributes include the student's marks and attendance information. The collected data is stored in a XML file for further processing.
2) Feature Detecting: The next step is feature detecting. Since many features that are not relevant to predict our results, we can drop them. The selection of 10 relevant features is done for improving the accuracy by feeding relevant data to the classifier. For example attributes like marks and attendance are supplying for training the classifier.
3) Naming the Dataset: The dataset has a column where the position of the student started. A total of 10 positions are mapped to the 3 output classes.The value of each feature lies between 0 to 1.After cleaning the dataset,70% of the data is randomly allocated to train the classifier and the rest 30 % is used for testing.
4) Training the Classifier: This approach used machine learning models are Neural network (Multilayer perceptron), Random Forests and Logistic Regression. Performing neural network is selected based on the value of tanh and number of hidden layer. The number of hidden layers and the value of tanh are to be 10 and 0.01 respectively. These arguments are used in training the neural network. In case of Logistic Regression, we can use multinomial Logistic Regression is used since the dependent variable is a multi-class. We can use inbuilt parameters in Random forest. Classifier data stored in XML file in a Chronological order for further process.
5) Data to be tested: After training the classifier, the trained classifier is loaded from the file and test data is fed in the desired output class is determined. The output of the testing phase is then provided for the analysis.
6) Evaluating Performance Metrics: The output from the classifier is determined based on some matrices like marks, attendance and Grade. The primary metric is the Grade. For demonstration, Matrix is plotted as shown in figure 4, 5 and 6.

IV. RESULTS

The performance metrics for a classifier include marks, attendance and Grade with grade being our primary measure and remaining secondary. Grade considered both marks and attendance. These metrics can be represented mathematically as follows.

Marks: It is the ratio of pertinent variable to the extracted variable [5].

$$MARKS = \frac{TRUE\ POSITIVE}{TRUE\ POSITVE + FALSE\ POSITIVE} \quad (1)$$

Attendance: It is the fraction of pertinent variable to the total number of actual pertinent variable [5].

$$ATTENDENCE = \frac{TRUE\ POSITVE}{TRUE\ \ POSITIVE + FALSE\ NEGATIVE} \quad (2)$$

Grade: It is the mean of Marks and Attendance. It is better to use this as a primary metric [6].

$$GRADE = 2 * \frac{MARKS * ATTDENCE}{MARKS + ATTDENCE} \quad (3)$$

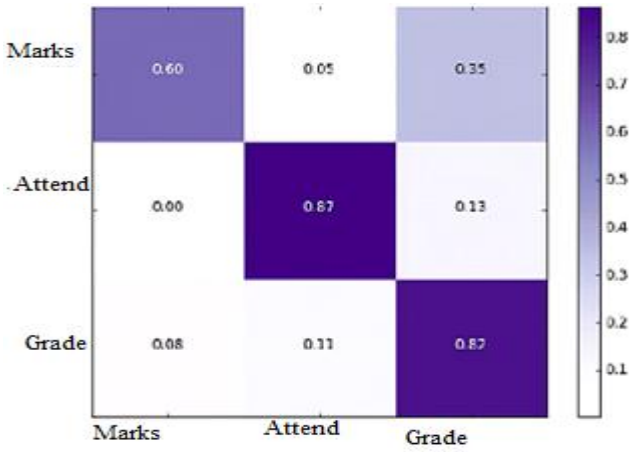

Fig.4: Confusion matrix for Neural Network

By using Neural network analysis we found the grade, attendance and marks are 0.82, 0.87 and 0.60 respectively from the equation 1, 2 and 3.

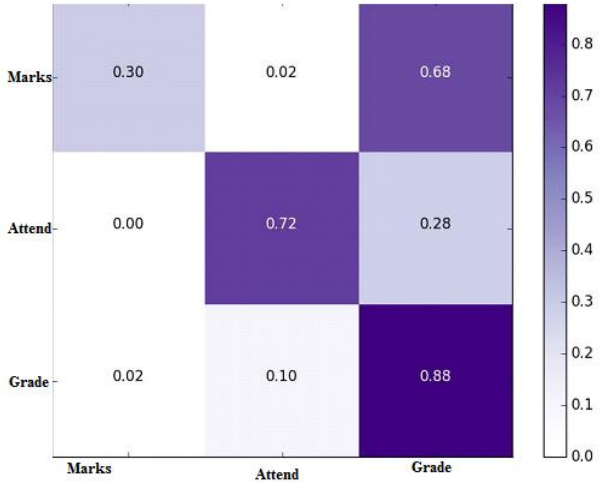

Fig.5: Confusion matrix for Logistic Regression

Fig.5 shows that logistic regression analysis of marks attendance and grade are 0.88, 0.72 and 0.30 respectively. In this approach we have considered grade and attendance because we got less marks than threshold value.

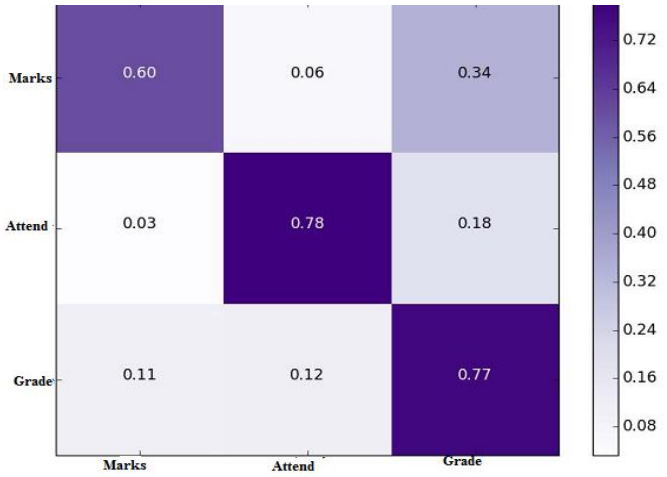

Fig.6: Confusion matrix for Random Forest

According to the equation 1, 2 & 3 random forest approach found the grade and attendance and marks are 0.77, 0.78 and 0.60 respectively.

Calculating the results, it can be pointed out from table 1 that neural network has performed the best with an accuracy of 79.01% and grade of 0.787.Logistis regression with an accuracy and grade as 71.92% and 0.697 respectively. Random forest has performed well too with grade of 0.739 and an accuracy of 74.07%.Confusion matrix can be observed that all three classifiers have correctly predicted the output for the marks attendance and grade. Whereas, the accuracy with which it has predicted the marks metric is quite less compared to other metrics. The less accuracy in the marks is due to fewer students in that position.

TABLE 1: Results

MODEL	MARKS	ATTD	GRADE	ACCURACY
Random Forest	0.739	0.74	0.739	74.07%
Neural Network	0.788	0.790	0.787	79.01%
Logistic Regression	0.747	0.719	0.697	71.92%

V. CONCLUSION

In this paper, Machine learning Techniques have been addressed to achieve to find best student team. Neural Network, Random Forest and Logistics Regression are models used in our research. Matrix and performance metrics determined that neural network has performed best among the other models. We can increase the accuracy of the model in future research.

REFERENCES

[1] Rosenblatt, X.Frank, "*Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms"*, Spartan Books, Washington DC, 1961.

[2] Rumelhart, E. David ,E Geoffrey. Hinton, and R. J. Williams. *"Learning Internal Representations by Error Propagation*", International Journal of Computer Engineering ,Vol. 1 , 1986.

[3] Ho, Tin Kam *," Random Decision Forests* ",In the Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, pp. 278–282..

[4] Ho TK , *"The Random Subspace Method for Constructing Decision Forests"* , IEEE Transactions on Pattern Analysis and Machine Intelligence,vol.8,Issue.20.

[5] T. Landgrebe, P. Paclk, R. Duin, and A. Bradley, *"Precision-recall operating characteristic (P-ROC) curves in imprecise environments"*, In Proceedings of ICPR, 2006.

[6] Y. Baeza and B. R. Neto, *"Modern Information Retrieval",* Boston, 1999

[7] J. Davis and M. Goadrich, "*The relationship between precision recall and ROC curves",* In Proceedings of the 23rd International Conference on Machine Learning, ser. ICML 06. New York, NY, USA: ACM, 2006, pp. 233240

