# A Higher Education Predictive Model Using Data Mining Techniques

Subhalaxmi Panda
Department of Computer Science & Engineering, Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, INDIA.
Ph : +91-9439057546
jinisubha.666@gmail.com

P. A Pattanaik
Department of Computer Science & Engineering, Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, INDIA.
Ph : +91-9692998983
priyadarshiniadyashapattanaik@soauniversity.ac.in

Tripti Swarnkar
Department of Computer Application, Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, INDIA.
Ph : +91- 9437130794
triptiswarnakar@soauniversity.ac.in

## ABSTRACT

The main objective of the higher educational organization is to provide high quality and necessary education to its students. The two goals of data mining in Indian education system is to analyze and enhance the chronicle way of recent educational data mining advances development; the second is to preserve, organize and discuss the content of the result which is produced by a data mining approach. The use of various data mining techniques such as random forest, decision tree, etc in Indian education processes will help to improve students' performance and provide a broad decision management skill in selection of courses as per their retention rate. This paper focuses on the model representation for analyzing the different data mining techniques in an Indian education system. Also the paper reviews a comparative study of ID3, K-Means, Naïve Bayes, Random Forest algorithm. In this paper, we have proposed the approach of Random Forest to predict the career decision for the 12th passing out students. The use of Random Forest has helped the students to take a correct appropriate decision as per their interest and skills and acts a career counselor toolbox.

## Keywords

Indian Education System; Data Mining; Random Forest

## 1. INTRODUCTION

Education is an attempt or effort of the senior people to spread their knowledge to the younger people of society. It is thus an institution, which plays a vital role in maintaining the perpetuation of culture by integrating an individual with his society. But in India, the education system has some serious lacunae[3]. Nowadays the important challenges in the educational organization are, not having more efficient, effective and accurate educational processes. Nowadays the important challenges in the educational organization are, not having more efficient, effective and accurate educational processes.



There exist lack of efficient and enough knowledge in Indian educational system which hampers the system management to get their quality objectives. Thus, data mining is considered as the most suitable technology which provides additional insight into the industrial as well as educational sectors helping in taking better decisions and motivating them to perform effectively. Data mining technology acts as a bridge between the lacunas and Indian educational system. Data mining approach leads to some data mining techniques which will help to improve the effectiveness, efficiency and the accuracy of the processes. As a result, this development helps in improving the Indian educational system by increasing educational system efficiency, minimizing students drop-out rate, gradually increasing students promotion rate, students retention rate, simultaneously educational improvement rate, students success, increase in students learning rate[6]. So, to achieve the overall quality improvement, we need some data mining techniques in the system that helps the decision makers to act smartly. Random Forest is one of the dynamic ensemble learning techniques which helps the students to take correct decision for their appropriate career choices after board exams. This data mining technique instructs the student with a particular pathway to direct his/her brighter career in an effective manner.

## 2. METHODS

Data mining in Indian education system has some extend overcome the lacunae by various techniques. It is gaining popularity because of effective, efficient and accurate towards Indian education system. The dataset used in this study contains records of class 10th and 12th students of career counseling. The data set is used to improve the performances, predict, or focus on skills of students by using different classification techniques.

Figure 1 demonstrates the whole working of the proposed model to give a broad understanding to the students about their career counseling. In the first stage, information about students of class 10th and 12th were collected and is named as data pre-processing stage. In the second stage, remove the unnecessary information and only relevant data will be fed to the database. After addressing the students information, the dataset is tried with different algorithms like ID3, K-Means, Naïve Bayes and Random Forest[1]. K-Means technique is one of the old and most widely

used algorithms used for clustering larger information based databases. Naïve Bayes is one of the statistical classifier techniques which act as a hypothesis for a set of estimating attributes in a database. This technique helps to detect the effectiveness of specific attribute for a given class and its relationship with other classes[2][5]. The other algorithm is the Random Forest, which aims in the first randomization through bagging. This approach of using Random Forest helps on handling missing values and category predictors and problems. The third last stage states the application of Random forest algorithm to the training data set with better output and the performance of each student are evaluated[6].

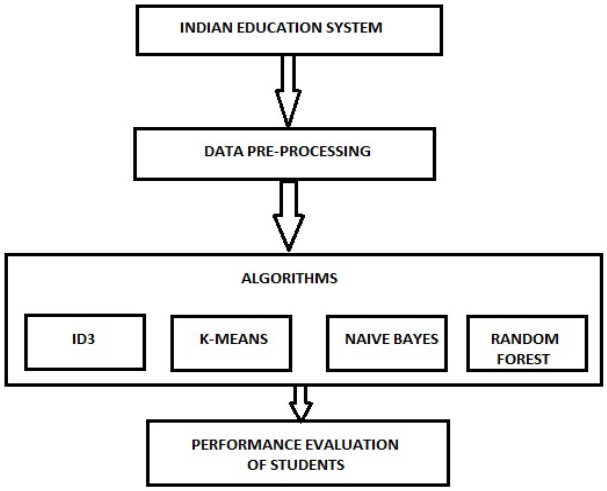


**Figure 1. Paradigm of Proposed model**

The training data set, shown in Table: 1 contains detail information of the student like Student ID, Gender, etc. The whole student information detail is used as the input dataset.

**Table 1. Student related variables**

| ATTRIBUTES | VARIABLES |
|---|---|
| Student ID | Student ID |
| Gender | Male/ Female |
| Students category | Unreserved/ OBC/ SC/ ST |
| Medium of Teaching | Hindi/English/ Local |
| Stream | Science/ Arts/ Commerce |
| 10th Grade | Excellent/ Average/ Poor |
| 12th Grade | Excellent/ Average/ Poor |
| Type of coaching | Online/ offline |
| Scholarship | Yes /No |
| Admission type | Entrance exam/Management |
| Type of coaching | Yes/ No |
| Material | Text book / Online / Both |
| Extra curriculum | NCC /Scout / Guide / Sports & heritage activities |
| | / Both |
| Efficiency | Good/Average/Poor |
| Father's occupation | Service, Business, Agriculture, Retired, NA |
| Mother's occupation | House-wife (HW), Service, Retired, NA |
| Parental income status | High Medium/ Low |

## 2.1 RANDOM FOREST

The Random forest concept was first introduced by Tin Kam Ho. Random forests or random decision forest is a learning technique for classification and regression. It is used in the construction of decision trees at training time and gives output classes that is in the form of the classification classes or mean prediction (regression) of the individual tree[1].

**Basic Random forest Algorithm:-**

Consider $N_{student}$ be the no. of students to create for each of N-students iterations. Where $m_{try}$ is no. of predictors to try at each split.

- Choose a new bootstrap sample from the training set.
- Develop an un-pruned tree on this bootstrap.
- Arbitrarily, choose $M_{try}$ predictors and find the best split using only these predictors at each internal node.
- Each $N_{student}$ leads to the largest extent possible with no pruning.

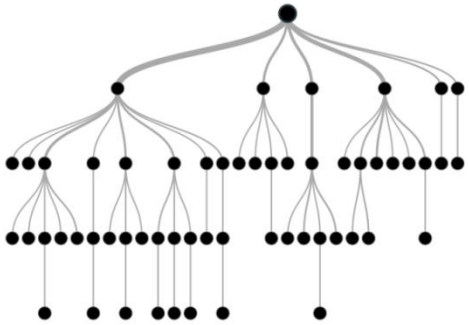

**Figure 2: Description of the working Random forest[1]**

## 3. RESULTS & DISCUSSION

For this experiment, 200 samples were taken into consideration. The table shows the accuracy in terms of percentage for different classifiers with the increasing data set size. To predict the change in behavior, the Random forest technique is used on student database. The technique distinguishes between slow learner and keen learner, recover the failure as soon as possible, takes appropriate action to improve the poor section students in a correct manner. The comparison of students performance using classifier algorithms like decision tree clustering, decision tree, Naïve Bayes, Random forest and outcome concluded that as the size of data set goes on increasing, Random forest gives better result or accuracy.

**Table 2: Prediction accuracy**

| Dataset size | Accuracy (%) | | | |
|---|---|---|---|---|
| | ID3 | K-means | Naïve Byes | Random forest |
| 20 | 62 | 40 | 40 | 60 |
| 80 | 64 | 55 | 62 | 78 |
| 160 | 72 | 43 | 81 | 79 |
| 200 | 75 | 54 | 59 | 80 |

## CONCLUSION

This paper lists a high scope for the students to decide for the brighter future with specific and accurate analysis. As the efficiency, accuracy, and effectiveness play the vital role in the process of Indian education system, use of the Random Forest technique provides us an optimal solution to the real world student's education. In this paper, we have used the approach of Random Forest to predict the career decision for the $12^{th}$ passing out students. The use of Random Forest has helped the students to take a correct appropriate decision as per their interest and skills. The final goal is to give a better insight to design a better Indian Education system for Indian students with the effective outcome. This review may extend to larger features to solve complex decision databases in an efficient manner.

## REFERENCES


[1] Rao, K. Prasada, MVP. Chandra Sekhara, and B. Ramesh "Predicting Learning Behavior of Students using Classification Techniques." International Journal of Computer Applications, Volume 139, Issues 7, pp: 0975 – 8887, April 2016.

[2] P.Veeramuthu "Analysis of Student Result Using Clustering Techniques" International Journal of Computer Science and Information Technologies, Volume 5, Issues 4, pp: 5092-5094, 2014.

[3] Goyal, Monika, and Rajan Vohra "Applications of data mining in higher education." International journal of computer science, Volume 9, Issues 2, pp: 113, March 2012.

[4] Hijazi and Naive, "Factors Affecting Students' Performance" e-Journal of Sociology, Volume 3, Issues 1 , pp: 2, January 2006.

[5] Dutt and Ashish. "Clustering algorithms applied in educational data mining." International Journal of Information and Electronics Engineering, Volume 5, Issues.2, pp:112, March 2015.

[6] Yadav, Surjeet Kumar, Brijesh Bharadwaj, and Saurabh Pal "Data mining applications: A comparative study for predicting student's performance." arXiv preprint arXiv, Volume. 1202, pp.4815, February 2012.