International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Educational Data Classification and prediction using Data Mining Algorithms

J. Jayapradha, Kishore Jagan Jothi Kumar, Binti Deka

Abstract: *Data Mining is the process of extraction interesting patterns from huge data sets and converts the patterns into logical structure for further Analysis. Predictive Modeling processes that make use of data mining, Machine learning and probability methods to forecast. Engineering is the most widely accepted stream of education in India. Students are uncertain about which department to join in engineering. It is important to improve the individual performance and help the students make the perfect choice regarding the department. In this paper, the hidden information from the previously recorded enrollment details during admission process is used to solve the students' uncertainty in their choice of department. In addition to this, the performance of alumnae also needs to be analyzed by the teachers to have a clear idea about the future of existing students. Our main goal is to unravel these problems using predictive Modeling. Here, we are focusing on three classification algorithms namely, support vector machine, Random Forest and Naïve Bayes. Data has been collected, normalized and applied to the three different classification algorithms, from which the best model is formulated using various parameters of evaluation. In this paper, we present our approach towards implementing the best model which is built based on the profession of parents, demographic features, type of location of the student and correlation between high school and higher secondary examinations. The Result of this research work shows that Random forest is efficient for the data set used when compared to the other two Classification algorithms.*

Keywords: *Predictive modeling, Classification algorithm, Support Vector Machine, Random Forest, Naïve Bayes, Data mining.*

I. INTRODUCTION

This is an International reputed journal that published research articles globally. In this global era, engineering is the most popular career option for the students in our country. Since, even an undergraduate degree in engineering guarantees a decent pay scale, the popularity of this field in India is highly comprehensible. It is well to note that Engineering consists of various streams like computer science, IT, mechanical and so on. One should choose their department prudently based on their interest and how good their performance is. It is the responsibility of every parent to orient their child towards finding their natural interest and the aptitude for the subject or branch of study. India is one among the top countries where many engineers are passing out every year. It is important for the universities to deliver quality education as well as make the students choose the best stream of engineering to shine brighter in future [1]. In general, engineering is a professional degree with a course duration of 4 years. Hence it is significant for each student to choose the stream according to their interest. Many students come to engineering counselling without a complete idea on their choice of department. Therefore, it is the responsibility of the university to give the students proper consultancy for their career and future. Also after the commencement of classes, if the student struggles to pass in examinations during the initial stage, it psychologically affects the student and creates a huge obstacle in pursuing the subjects in further semesters [2]. Based on their model test results we cannot come to a clear picture of the semester results as the conditions are entirely different for semesters compared to the model exams. It is more important to predict the poorly performing students and guide them solely for attaining success [1].

In this paper, we propose a betterment model to help the students by predicting which department will be good for them to outshine in that field. It has been done based on the implementation of classification algorithms considering various attributes like demographic features, the profession of the parents and his previous high school and higher secondary records. Moreover, to predict the performance of the students, attendance and extracurricular activities are added along with the other attributes. We have chosen the best algorithm for the creation of a model by comparing the error rate of three classification algorithms namely Support Vector Machine, Random Forest, and Naïve Bayes algorithm. We have created the resultant model after multi-fold cross-validation of data set. The platform used here is R [14].

II. LITERATURE SURVEY

Similar work was carried out by creating two models for the enrolment and performance tracking [1]. We collected 10,000 data considering 20 parameters including demographic features and academic statistics. The WEKA tool served as the working platform, where they implemented two rule learners, decision tree, and two Naïve Bayes algorithms

Manuscript published on 30 September 2019
* Correspondence Author

Jayapradha J*, Assistant Professor of SRMIST, Chennai.
Kishore Jagan Jothi Kumar, Master of Computer Science at Arizona State University, Tempe, AZ, USA .
Binti Deka, Case Manager in Onco.com, Banglore, KA, India.

Retrieval Number: C6457098319/2019©BEIESP
DOI:10.35940/ijrte.C6457.098319
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The current trend is the educational data mining. The contemporary system of educational data mining is done using the class X results of the Central Board of Secondary Education (CBSE) for the prediction of the 1st-semester results. [2]. The system involves pre-processing and normalization of data. It helps the petabytes of information to get processed with traditional algorithms translated as MapReduce algorithms on the Hadoop clusters where storage and computing take place.

Various data mining algorithms have been compared using Special data mining tool, MOODLE, which was created to study the performance of the students based on their web-based courses and final marks [3].

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

Previously for implementing the learning analytics for teachers, a toolkit named ELAT: (Exploratory Learning Analytics Toolkit) was created using all the implementation algorithms executed in WEKA tool. ELAT was made exclusively for EDM [4]. The data fed with multiple parameters were made easy with ELAT, and even the visualization UI made greater understanding for the users.

It is to note that by comparing various papers relating to educational data mining, neural network algorithm has the highest accuracy of prediction followed by decision trees and support vector machines, k-Nearest methods and lastly the Naïve Bayes theorem [5].

The result of analysis made on only a small dataset of 250 students, k-Nearest, and Native Bayes performed with an accuracy of 80% [6]. But it is also stated that SVM also performed better than the other algorithms when we included more parameters.

III. PROPOSED SYSTEM

Prediction of proper department and University performance record are two systems to be built and maintained. Data mining involves various processes like data collection, data pre-processing, feature extraction, model creation and cross- validation, algorithm selection, and implementation. We have used three classification algorithms namely, Random Forest, Support Vector Machines and Naïve Bayes algorithm for model creation. We fed the data for analysis with some attributes depending on the requirement. So, based on needs, these algorithms are executed, and results are validated using multi-fold cross-validation . Finally, the best algorithm with efficient accuracy and kappa statistic have been selected and implemented.

Random Forest algorithm relies on decision trees concept. It takes the input data and creates more subsets of data for predicting the classification. Each subset of data has different rules for classification to form a tree [7]. We spread the diversified data and subset across for all the trees. A collection of all the trees combines to develop the random forest algorithm. We will finally acquire the generalized solution for the classification and prediction [8].

Support vector machine algorithm confides on the hyperplane separation of classes [9]. Multidimensional planes are mapped as per the requirement and based on the kernel chosen. The optimal hyperplane with more margins will be picked out as the classifier. Numerous vectors which can classify between two classes are spotted out and construct the hyperplane. The distance between the hyperplane and the closest point to the plane is called the margin [10]. Faraway the margin, more optimal the hyperplane is. Cost is the number of vectors restricted to define the hyperplane. Gamma factor is to outline the length of the vector can be drawn to define the hyperplane [11].

Naïve Bayes algorithm saves the run time of execution comparing other algorithms. This algorithm depends on the Bayes theorem [5].

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

Here the comparison of conditional probability is made for given features with each classifier of the output and the class with more probability is chosen as the result [12].

Pros: Easy to implement, Fast in execution, Not sensitive to irrelevant features

The term accuracy relates to the correctness of prediction made by the model.

Kappa statistics acquaintances with the exactness associated with the distribution of the prediction made comparing the actual outcome [6].

Cross Validation is the method to train the model by taking the statistics set as different folds of data and split them into training and testing table in a loop so that the result will be a model which could train and test all the data using the data classification [13] .

IV. IMPLEMENTATION

We made the Data collection for the study by using questionnaire of 13 questions. The survey included factors like gender, age during admissions, department, secondary and higher secondary exam percentage, parent's occupation, history of arrear, year of pass out, extracurricular activities, branch during high school, etc. The Alumnae gave 265 responses via crowdsourcing of data. After collection of data, pre-processing of data was done to normalize the data. The data fed into R language and feature extraction are done based on the correlation coefficient exhibited by the attributes to the outcome variable [14].

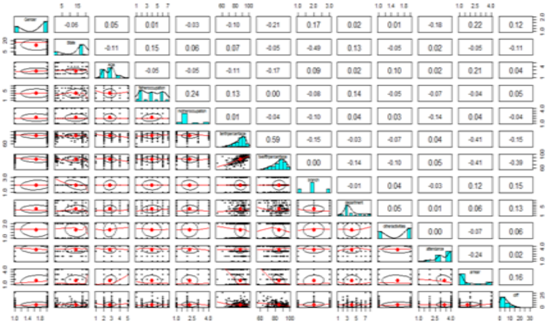
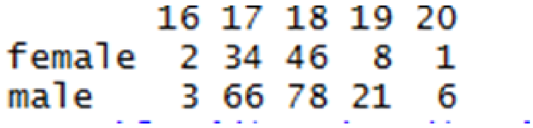


Fig. 1.Correlation coefficient relationship of various factors

We should also see how to distribute each feature among the dataset. 91 female alumnae and 174 male alumnae responded to the questionnaire. Most of the students joined the course at the age of 17 or 18.

We have taken the factor: history of arrear as the key output variable. So, we need to study how the other factors are related to it. History of arrear is under four levels 0 for nil arrear 1,2 for a respective number of arrears and 3 for 3 or more arrears.

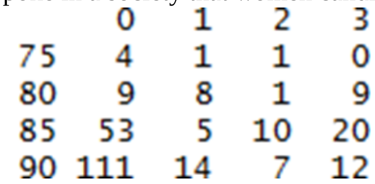


```
        16 17 18 19 20
female   2 34 46  8  1
male     3 66 78 21  6
```

Fig. 2.Relationship between age and gender

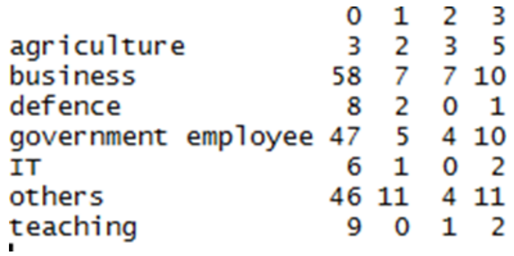
In Fig. 2, it is observable that female students start their graduation courses by maximum 19 years of age. But men do extend their entry to engineering until the age of 20.

In Fig. 3, most female alumnae have less history of arrears compared to the proportion of male alumnae. The general conscience spoke in a society that women candidates always perform better than male candidates became factual again. The correlation factor between father occupation and the history of arrear is very less as in Fig. 1.



```
     0   1   2   3
75   4   1   1   0
80   9   8   1   9
85  53   5  10  20
90 111  14   7  12
```

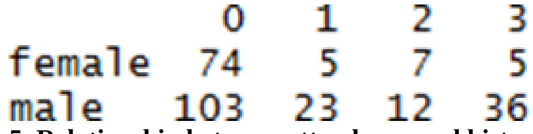
Fig. 3.Relationship between gender and no. of arrear



```
                     0  1  2  3
agriculture          3  2  3  5
business            58  7  7 10
defence              8  2  0  1
government employee 47  5  4 10
IT                   6  1  0  2
others              46 11  4 11
teaching             9  0  1  2
```

Fig. 4.Relationship between father occupation and the history of arrear

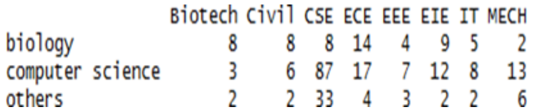
From the correlation coefficient relationship in Fig. 1, we could find that apart from other demographic factors, the academic scores play leading role in determining the performance of the student. The values of coefficient relate to around -0.412 for twelfth percentage and tenth percentage. The negative value indicates that it is inversely related to the history of arrears, i.e., as the proportion of school academics increases for a person, the history of arrear is found to be low. But the variation between tenth and twelfth percentage doesn't relate much to the history of arrears. In parallel to these observations, we should also note that we need to consider attendance and extra activities column for predicting the performance of the students.



```
         0   1   2   3
female  74   5   7   5
male   103  23  12  36
```

Fig. 5 .Relationship between attendance and history of arrear.

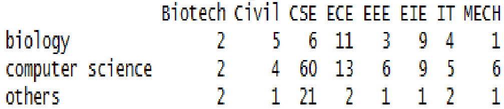
In Fig. 5, we can interpret that students having a higher calculation of presence had performed better. But students with consistent attendance did relatively lesser compared to the students having very low turnout percentage. The correlation coefficient is turned out to be -0.24.



```
                 Biotech Civil CSE ECE EEE EIE IT MECH
biology                8     8   8  14   4   9  5    2
computer science       3     6  87  17   7  12  8   13
others                 2     2  33   4   3   2  2    6
```

Fig. 6.Relationship between branch and department

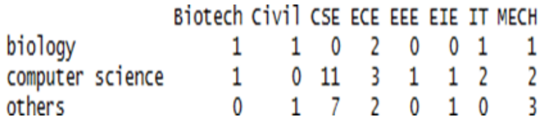
In Fig. 6, we could see the overall view of how students have transformed from their branch of study in high school to numerous streams of engineering. Considering the facts of biology division, very few students retain their field of study. Most of them shift to ECE or EIE Department. But most of the students who took computer science branch, recollect their area of study in engineering too. To get more insight on this information we need to look upon the no of arrears also.



```
                 Biotech Civil CSE ECE EEE EIE IT MECH
biology                2     5   6  11   3   9  4    1
computer science       2     4  60  13   6   9  5    6
others                 2     1  21   2   1   1  2    1
```

Fig. 7.Relationship between branch and department on null arrear category.

In Fig. 7, the students who studied under computer science branch in their high school have performed well in computer science engineering too. Similarly, students who shifted from biology branch to ECE department also performed better.



```
                 Biotech Civil CSE ECE EEE EIE IT MECH
biology                1     1   0   2   0   0  1    1
computer science       1     0  11   3   1   1  2    2
others                 0     1   7   2   0   1  0    3
```

Fig. 8. Relationship between branch and department on or more arrears category.

In Fig. 8, we could observe that 1/8th of computer science students whose department is also computer science fails miserably with more arrears. Interestingly, out of 33 students of another branch, only seven have failed very badly in semester examinations.

We have implemented all the three algorithms for the history of arrears with the enrolment data [2]. Confusion matrix was used to show the nature of the data with their accuracy and kappa statistic value. Moreover, the result was analyzed even for tracking the academics of the student.

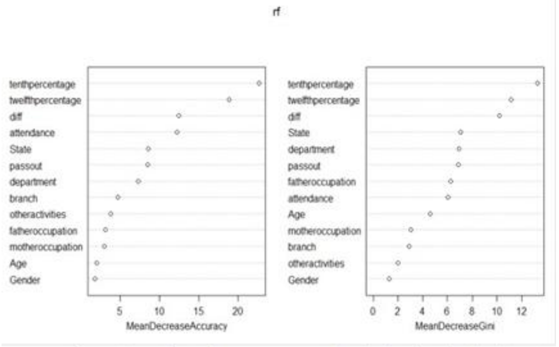


Fig. 9. Decreasing importance of features in Random Forest Algorithm.

In Figure 9, it is evident that for Random Forest classifier, academic performance has more importance than the demographic features.

Retrieval Number: C6457098319/2019©BEIESP
DOI:10.35940/ijrte.C6457.098319
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Decreasing importance of features enables us a better understanding of the impact of each feature considered for data understanding.

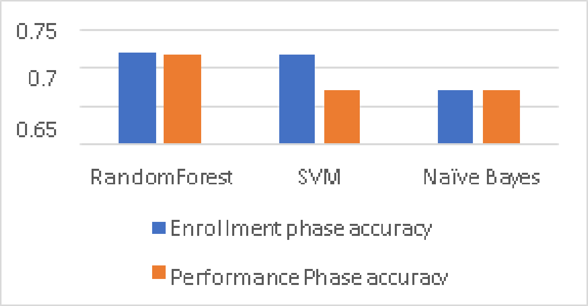


Fig. 10.Accuracy rate of different classification algorithms during enrolment phase and performance phase

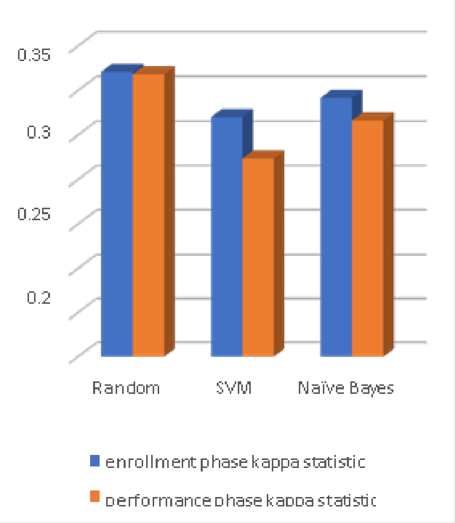


Fig. 11.Kappa Statistics of dissimilar classification algorithms during admission phase and presentation phase

We could see from Fig. 10 and 11 that the accurateness rate for Random Forest algorithm is higher among the classification algorithms. The value of accuracy didn't change when we added more features for classification of Random Forest and Naïve Bayes, but when the kappa statistics varies; it decreases for all the algorithms. More the value of kappa statistics more is the efficiency to build the algorithm. The accuracy decreases in SVM when we add more features.

The result is the same when tested with a larger size of the database using synthetic data of 8888 rows from which we can conclude that Random Forest algorithm is efficient classification algorithm regarding this data set.

V. CONCLUSION AND FUTURE WORKS

In this paper, prediction of proper department and University performance record are the systems built and maintained. We have used three classification algorithms namely, support vector machine, Random Forest and Naïve Bayes. Relationship between various factors of our dataset has been analyzed to enable us to understand the impact of different feature. A questionnaire was filled in by 273 members out of which, 265 remained after data cleaning and pre-processing. This data later was used for studying the students' performance, which helps to frame models for creating the system to predict the perfect choice of department. This could also contribute for tracking the performance of the students. For model creation using R language, the comparison of nature of three different classifier algorithms are done. The variation of accuracy and kappa statistic differs for a different dataset and by for changing the size of the attributes. Using the model created, a student can have an idea about choosing their department. Further after admissions, a teacher could also track the performance of the student and guide him before examinations with more concern. The Experimental evaluation in this paper has shown that the Random Forest is the efficient algorithm when compared to the other two algorithms. For future works, we can combine the hybrid of two supervised algorithms for model creation for more efficient results.

REFERENCES

1. Predicting Student Performance by using Data Mining Methods for Classification DorinaKabakchieva Sofia University "St. Kl. Ohridski", Sofia 1000 CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume 13, No 1Sofia • 2013 Print ISSN: 1311-9702; Online ISSN: 1314-4081DOI: 10.2478/cait- 2013-0006
2. A Big Data Approach for Classification and Prediction of Student Result Using Map Reduce by Midhun Mohan M G &Siju K Augustin and Dr.KumariRoshni V S 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) | 10-12 December 2015 | TrivandrumH. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985, ch. 4.
3. C. Romero, S. Ventura, Educational data mining: A review of the state of the art, Trans. Sys. Man, Cyber Part C 40 (6) (2010) 601–18.doi:10.1109/TSMCC.2010.2053532.
4. Dyckhoff, A. L., Zielke, D., Bültmann, M., Chatti, M. A., & Schroeder, U. (2012). Design and Implementation of a Learning Analytics Toolkit for Teachers.Educational Technology & Society, 15 (3), 58–76.
5. The Third Information Systems International Conference-'A Review on Predicting Student's Performance using Data Mining Techniques' (2015) by Amirah Mohamed Shahiria , Wahidah Husaina, Nur'aini Abdul Rashida,Procedia Computer Science 72 (2015) 414 – 422
6. Comparison of Machine Learning Methods for Intelligent Tutoring Systems Wilhelmiina H¨am¨al¨ainen and Mikko Vinni Finland M. Ikeda, K. Ashley, and T.-W. Chan (Eds.): ITS 2006, LNCS 4053, pp. 525–534, 2006._c Springer-Verlag Berlin Heidelberg 2006
7. http://blog.echen.me/2011/03/14/laymans-introduction-to-random-forests/
8. https://www.slideshare.net/m80m07/random- forest
9. http://www.svm-tutorial.com/2014/11/svm-understanding-math-part-1/
10. https://www.quora.com/What-are-Kernels-in-Machine-Learning-and-SVM
11. http://scikit- learn.org/stable/auto_examples/svm/plot_rbf_parameters.html
12. Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction Mr.Chintan Shah Dr. Anjali G. Jivani 4th ICCCNT 2013 July 4-6, 2013, Tiruchengode, India
13.]https://www.analyticsvidhya.com/blog/015/11/improve-model-performance-cross-validation-in- python-r/
14. https://data-flair.training/blogs/classification-in-r/

AUTHORS PROFILE

Jayapradha J, pursuing PhD at SRMIST, Chennai, TN, India. Currently working as a Assistant Professor of SRMIST, Chennai. Three years of Industry ecperience and six years of teaching experience in SMIST. Published eight papers in the Scopus journals and two paper in the conference. More than 20 workshops has been conducted and participated. Second rank Holder in M.tech, SRMIST 2011. Nine certification courses have been completed in the field of CSE.Holding Membership in ISCA**.**

Kishore Jagan Jothi Kumar, Completed Bachelor's degree at SRM University, Chennai, TN, India. Former Programmer Analyst Trainee of Cognizant Technology Solutions and currently pursuing Master of Computer Science at Arizona State University, Tempe, AZ, USA .

Binti Deka, Completed Bachelor's degree at SRM University, Chennai, TN, India. Former Business Analyst of GoFrugal Technologies, Chennai, TN, India and currently working as Case Manager in Onco.com, Banglore, KA, India.