A Survey On Investigation Of Students Placement Log Using Machine Learning Algorithms

S.Rajeshkumar, Dr.V.Seenivasagam,D.Vijayakumar,S.Saravanaselvi

Abstract: Placement is an initial expectation regarding both the institute and the student's perspective. Achieving placement at the right time depends on various factors. This paper makes an investigation into the placement dataset of professional students from diverse domains. Twin fold aspects of assessment have taken into consideration. First, examining student's different performance attributes during all four years of the course. Second, studying the ratio of growth in the placement with the influence of ICT. The analysis returns two classes of students: The Marginal and Virtuous attaining students. This paper also focuses on various classifier's accuracy using WEKA tool. The discoveries in this study indicate that the list of techniques must account to increase the probability of placement as well as an optimal classifier to predict the appropriate class of students.

Index Terms: Data mining, Placement data, Classifier model, WEKA,Placement prediction, ICT, Classifier performance

1. INTRODUCTION

DATA mining is commonly used in the academic environment to explore new ideas and issues that arise in this sector. Student placement is an essential concern in the academic institutes where numerous factors may disturb the expected outcome. For prediction, the essential gears are Constraints that distress the student Excellence, Data mining methodologies, and tools to mining the data. These gears may be psychological, personal, and societal. For a clear understanding, let's consider the following figure 1 represents the taxonomy of machine learning algorithms.

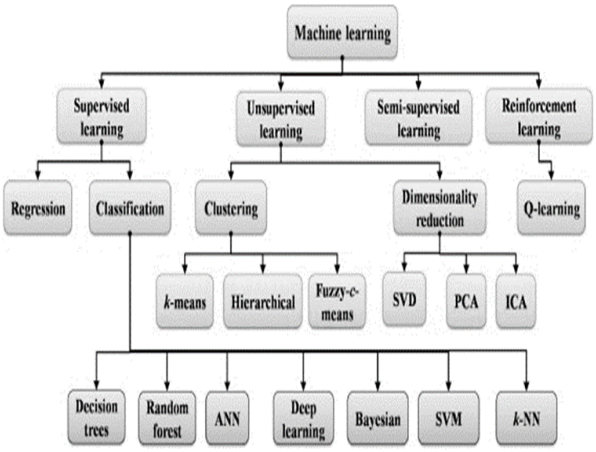

Fig 1: *The machine learning paradigm*

Research findings expected to deliver precise result based on the given samples for that the researcher must observe the techniques which are suitable for their work. In concern with this, the finding is as follows. Supervised Learning Uses the labeled data to build the predictive model. The input and output are known here. Finally, the model used to represent the learned relationship between the given input and known output. The outcome of the findings has directed the analysis using the classification techniques. This research paper allows us to study and improves the placement strategy by diminishing the various erroneous factors on student's performance. In this Paper, Prediction of student Placement analyzed by applying a set of four prominent classification techniques using the WEKA tool.

- *Rajeshkumar.S is working as Assistant Professor, Department of Computer Science and Engineering, National Engineering College, K.R.Nagar, Kovilpatti E-mail: suyamburajesh@gmail.com*
- *Dr.V.Seenivasagam is working as Professor, Department of Information Technology, National Engineering College, K.R.Nagar, Kovilpatti. E-mail: yespee1094@gmail.com*

2 BACKGROUND

Recently Indian education domain has taken more advanced action to become self-balanced with new generation students. Advancement through ICT is more and more essential against the rapid growth of the Internet era. Datamining is the promising technology focusing on evaluating the useful pattern by the automatic or semi-automatic process. The newly derived pattern expected to produce meaningful predictions [13]. Educational data mining creates endless opportunities for research in higher education institutions data. It supports various tools and methodologies to review student's data for assessing performance in various aspects [11]. In this way, a study has taken into consideration to identify individual student characteristics using the MDT classification technique. Deviation in student performance recorded using two primary classification models from higher educational data are students with different modes of course registration and different levels of family income [12].

A critical aspect of educational data mining is to predict the performance of slow learners as early as possible. The challenging task here is to find the best-suited algorithm from a variety of classification algorithms are available. Algorithms play a vital role in classification accuracy against multi-dimensional data. The experiment on high school data using prominent classification algorithms shows multilayer perception coming back with 75% accuracy on early prediction [16]. As the world moves faster, an essential new course should introduce in a timely fashion. Hidden risk must be taken into consideration while new courses introduced. Reducing the failure rate in the early stages needed an efficient data mining approach. Support vector machine classifiers preciously identified students likely to fail in introductory courses offered by Brazilian universities through online and offline mode [7]. The primary focus should be given in the area of algorithm selection to review the student data to obtain the precious

classification. Classifier accuracy can be varying under any circumstances based on data heterogeneity. The Possible well-known issues missing data, irrelevant data are adding more and more in comforts to the research. The review says Researcher should have a concise focus on characteristics of data to be used [10]. The classification algorithms primarily focusing on attaining the maximum accuracy concerning the target; several classification algorithms have included in data analysis open-source tools provide an immense opportunity for analyzing the data from a different perspective. Based on the nature of the analysis, different algorithms possess different behavior on data, so the comparison is becoming essential to predict the correlation between accuracy level achieved by algorithm against the nature of data. For instance, in the research paper [15], liver disorder Medical diagnosis data has been analyzed using the decision tree, Naive Bayes, K-Nearest neighbor algorithms. Using various tools has yields useful information that the open-source tool KNIME can provide the highest accuracy with all these algorithms. In WEKA tool provides the lowest accuracy with the Naive Bayes algorithm. Keeping the result in mind, the researcher can choose the appropriate algorithm and tool based on their dataset utilization.

2.1 Scope of Mining in Education Data

An IJSTR, the review [4] says: Research in education data mining has witnessed rapid growth in recent days due to the unavoidable influence of ICT. Enormous availability of technologies enforces e-learning behavior among the student community. Education data mining also deals with e-governance through the massive data has produced from an educational organization, online forum, e-learning sites. Recent research initiatives in education data mining have to produce soulful predictions about the early identification of dropout, failure, behavioral change, and psychological factors of young minds for their attainment, poverty, region, and confidence. In [3], the application of Logistic regression modeled the research that map the correlation between ability and self-confidence of the students from the poverty line. It could be unbelievable that the outcome says the quality and experience of teachers also negatively affect the student's self-confidence. Similarly, the research article [1] to prevent academic dropout points out the application of new academic action plans during the crucial week, of course, is marked as a useful action plan for reducing the academic dropout. The research has used Various data classifiers such as the J48 decision tree, Multilayer perception, Naïve Bayes. Performs the sequential mining analyses on Turkey student evaluation record returns the useful information that the instructor's performance being a critical factor in improving the quality of education.[8] Survey [6] reports, during the year 2010 to 2013, there is 240 EDM research work published and is made up of 222 education data mining approaches and 18 Education data mining tools. Research article [2] briefs about the application of machine learning algorithms to predict the most clinical feature of depression due to the smoking habits of youths.

Though student behavior in a higher educational institution is an essential criterion, the paper aims to investigate tools to find the best fitted one. Analysis has done with Apriori and FP growth algorithms on rapid minder and Taranga tools [9] The study carried out an efficient teaching-learning process on educational institutions to understand the most significant features to implement computer-assisted learning system [14].

3 METHODOLOGY

The study aims to investigate the usage pattern in student's placement logs using widely used powerful machine learning algorithms. Also, the study has expressed the interestingness performance of algorithms on the dataset. The study carried out in the following comportments

3.1 Data Collection

The targeted audience for data is that the students are aiming the placement from various academic backgrounds. The Questionnaire with 18 well-defined attributes has been dispatched to selected government, aided, autonomous, self-financed institutions over the southern part of TamilNadu for data collection. By integrating datasets, a single view database is modeled in a unified way to accomplish the scope of this study by performing data analysis using classification techniques. The questionnaire includes the student's gender, nationality, place of birth, CGPA, course id, course topic, semester, absence days. ICT activities such as seminar presentations, web resource utilization, project activities, and communication. Regarding family aspects, parent education, parent satisfaction about the institution, parents answering the survey, and three categories of the class label has defined as low risk, medium risk, and high risk. Since each attribute assigned with weights, during data analysis, the cumulated weight acts as the best indicator to predict the class label. Similarly, each attribute evaluated to describe the student's success ratio on placements.

3.2 Experimental Setup

The study emphasis supervised learning classification techniques, which highly connected to the efficient analysis of modeled dataset. Decision tree, Naïve Bayesian, neural network, support vector machine classifiers used for the finding of convenient Patten. To improving classification accuracy, extensive K-fold cross-validation used. Since 70 percent of data assigned to test data, the remaining 30 percent utilized for the test. WEKA toolkit is performing the classification analysis task using the most rated classification techniques mentioned above.

3.3 Evaluation metrics

The study decided the following evaluation metrics to assess classification accuracy. The cumulative dataset includes 480 samples in which the correctly classified and wrongly classified counts measured to evaluate the accuracy of the classifier. In addition to that, the following metrics also used

- Precision
  The confusion matrix yields true-positive, true-negative, false-positive, false-negative rates. From the outcome, the precision measured by the ratio between correctly predicted positive instances and total predicted positive instances.
- Recall
  The yield of a confusion matrix is deriving the new measure recall by calculating the ratio between correctly predicted positive instances and all the instance in the class
- F-measure
  F-Measure represented as the harmonic mean between recall and precision. F measure reports how many instances classified correctly by

indicating with a value between o to 1.

- Root mean Squared error
  Root Mean Squared error evaluated as the average Square of difference between the exact original value and predicted values.

$$Meansquarederror = \frac{1}{N}\sum_{j=1}^{N}\left(y_j - \hat{y}_j\right)^2 \quad (1)$$

- Mean absolute error
  Mean absolute error evaluated as the average of the difference between the exact original value and predicted values.

$$Meanabsoluteerror = \frac{1}{N}\sum_{j=1}^{N}|y_j - \hat{y}_j \quad | \quad (2)$$

4 EXPERIMENTAL ANALYSIS

The primary focus of this paper is to ensure building an efficient classification model. Several criteria have to guaranteed from invoking the dataset to applying the classifier. The instance categorized into two-level as 70 percent of total count for the training process and the remaining 30 percent for the testing process. While dealing with the attributes, few derived attributes removed from the dataset. Regarding achieve enriched classification accuracy, the model was validated with a small set of samples by implementing k-fold cross-validation. As concerned with the optimized bias, and variance value of k has assigned to 10. The best split can be guaranteed by the level of impurity of child nodes(t). Entropy, Gini is enforced to reduce classification error. The correlation is as follows in Equations (3), (4), respectively.

$$\text{Entrophy(t)} = -\sum_{i=0}^{c-1} \text{p(i / t)} \log_2 \text{p(i/t)} \quad (3)$$

$$Gini(t) = -\sum_{i=0}^{c-1}\left((p/t)\right)^2 \quad (4)$$

When c is considered as a number of classes the classification error has derived as

$$\text{Classification error(t)} = 1 - \max_i\left(p(i/t)\right) \quad (5)$$

The best-suited classification model has brought into the notice from the literature. Decision tree classifier, naïve Bayesian classifier, neural network-Multi layer perceptron classifier, support vector machine classifier is found the sound correlation with the data to be analyzed. Since the data is categorical, classification trees fit with the model design instead of regression trees. As decision variable a_i is a discrete binary recursive partitioning process splits the data in an iterative manner. An ideal heuristic recursive partitioning process followed to split and creating subsets on the way to develop a decision tree. The case overfitting has avoided by fixed the limit on depth. The model has accomplished on achieving the highest information gain. Decision tree classifier pessimistic error can consider in the way to improves accuracy. For all the records if node t has correctly classified the training records n(t) and misclassified the records e(t), the error produced by decision tree classifier in Equation (5)

$$e_g(t) = \frac{\sum_{i=1}^{k}(e(t_i)+\Omega(t_i))}{\sum_{i=1}^{k}(n(ti))} = \frac{e(T)+\Omega(T)}{N_t} \quad (5)$$

Simple and widely used classification model naïve Bayesian classifier has built with an assumption that shows the best accuracy. NaïveBayes process with the logic that all the attributes are independent. The model calculates the posterior probability of any attribute based on class c. The classifier prediction is the ratio between predicted prior probability and product of likelihood and class prior probability which is denoted in Equation (6)

$$P(c / x) = \frac{P(x / c) \cdot P(c)}{p(x)} \quad (6)$$

In addition to that, the classifier aims to find the maximum probability represented in terms of conditional and class probabilities are shown in Equation (7)

$$Y = \text{argmax}_x P(y)\prod_{i=1}^{n} P(x_i / y) \quad (7)$$

The error rate of the Bayes classifier model for the multiclass dataset can be calculated using the function shown in Equation (8)

$$p = 1 - \sum c_i \neq \quad \text{Cmax,x} \int_{x \in H_i} P(c_i / x)P(x)\,\mathrm{d}x \quad (8)$$

The neural network multilayer perceptron provides precise results on complex data sets. N number of hidden layers h_n can have the place between input and output layers. Each instance from data set product with weight for hidden layer and added with bias combined feed into an activation function, which produces an output for the first hidden neuron layer. The outcome of the hidden layer's dot product added with a bias to produce the class output. Activation function f passed with the hidden layer output value z as follows in Equation (9)

$$z = \sum_{i=1}^{n} W_i^{h_1} h_i^1 + bias \quad (9)$$

Where the classification accuracy loss function as squared mean error denoted in Equation (10)

$$\text{Loss function: Square mean error} = \sum_{1=1}^{n}(f_i(x) - y_i)^2 \quad (10)$$

In concerned with optimal computation power support vector machine model have been built. SVM classifier manages the process by enabling hyperplane in N-Dimensional space. Choosing the plane with the highest distance between datapoints separates the data into classes. Maximizing the margin among data and hyperplane to achieve higher accuracy. Cost function concentrate on accuracy measure is derived as follows in Equation (11)

$$Cost\ function = c\left(x, y, f(x)\right) = \left((1-y) * f(x)\right) \quad (11)$$

Loss function with regularization denoted in Equation (12) is utilized to maximize the distance between nodes and plane.

$$\text{Min}_w \lambda\|w\|^2 \sum_{i=1}\left(1 - y_i(x_i, w)\right) \quad (12)$$

5 RESULT AND DISCUSSION

An investigation from the study has deep insights into student's placement relevant factors. Primarily students involved in ICT activities have a high correlation with placement. Enrollment in web-based courses, project activities, event participation isgrouped into ICT activities. As the CGPA has taken as a key indicator to analyze, the interesting pattern reveals an unfortunate fact that major students from rural regions fall in low possibilities for placement category even having higher CGPA. The key finding from the study is the region of student's birth have a considerable impact on failure to attain placement. Rural and municipality residents have witnessed under strong struggle to reach the placement goals.

TABLE 1

KEY ATTRIBUTES AND DESCRIPTIONS

Web resource utilization	Online courses, contest-based learning
Project activities	Internship, Consultancy, funded from agencies
Presentation	Events participation, communication
Place of birth	Corporation, Municipal, Unions, villages
Class Indicator Medium	Medium possibilities for placement
Class Indicator High	High possibilities for placement

The multilayer perceptron has shown the highest accuracy rate of 80% with cross-validation k value k=10. The classifier's error rate measured in two aspects by mean absolute error referred to in equation (2), and root means squared error referred to in Equation (1).SVM classifier and decision tree classifiers ranked second and third position, respectively, based on accuracy measures denoted in table 2.

TABLE 2

PERFORMANCE OF CLASSIFIER

Model / Result	Total Instances	Correctly classified	Wrongly classified
Multilayer perceptron	480	382	98
Support vector machine	480	378	102
Decision Tree	480	374	106
Naïve Bayesian	480	349	141

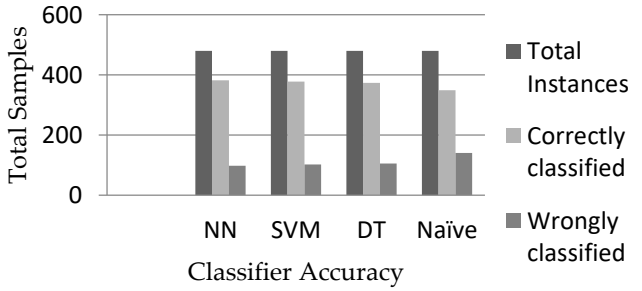

***Fig2: Performance** of the classifier*

Prediction error plays a vital role in deviation of classification accuracy. This paper also concentrates on the analysis of prediction error produced by the classifier using an efficient common metrics Mean absolute error, Squared mean error, receiver operating characteristics curve.

TABLE 3

CLASSIFICATION ACCURACY EVALUATION METRICS

Model / Result	Precision	Recall	F-measure	AUROC
Multilayer perceptron	0.78	0.77	0.78	0.90
Supportvector machine	0.83	0.91	0.87	0.87
Decision Tree	0.79	0.75	0.68	0.81
Naïve Bayesian	0.69	0.80	0.74	0.83

***Fig 3:** ROC curves of Classifiers*

Variations in sensitivity and specificity have analyzed without changing the threshold using the receiver operating characteristic curve (ROC). The ROC curve for the classifiers is shown in figure 3.

Plotting the graph between true positive rates against false positive rate allows setting an optimal threshold value for achieving higher sensitivity and specificity.

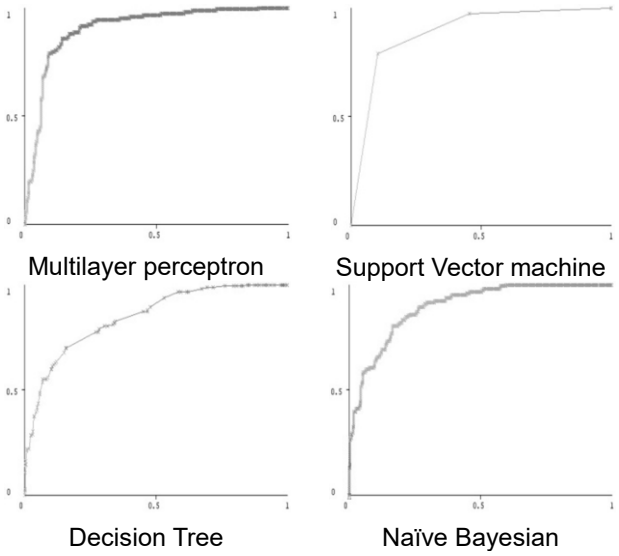

***Fig 3:** ROC Curves of Models*

In the case study, the discussion student's placement log analysis has disclosed the interesting pattern that students involved in various ICT activities have the highest probability to achieve placement; also, application of various classifies exploits the deep knowledge in performance metrics of classifiers. This study results from a Multilayer perceptron classifier that has found as the more promising one, which assured the highest true positive and false positive rate as 77%, 89%, respectively. The error rates produced by the multilayer perceptron classifier have shown in figure 4.

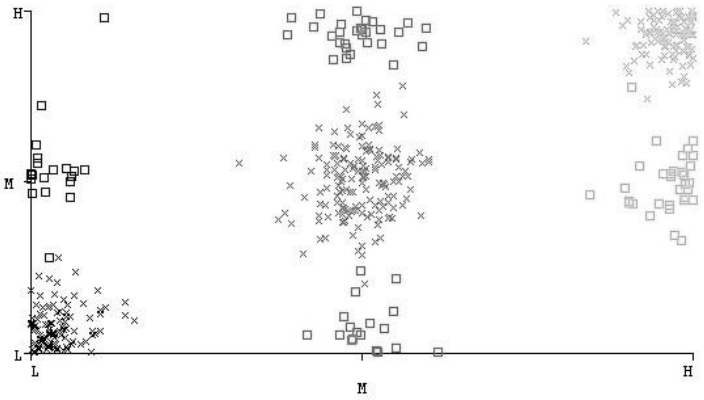

Fig 4: *Error rates of MLP*

Error rates of MLP expressed the precise information that the distance between misclassified data points is very low. So, the optimization techniques are implemented on the classifier than the model would become the best-fitted one for the dataset.
Total 480 number of samples provided to classification model out of that 70 percent attributes are supplied to train the model remaining 30 percent attributes are supplied to evaluation. Since multilayer perceptron took more time to train model, in the aspect of accuracy, MLP predicts the correct classes higher than the remaining three classifiers. In the meantime, the support vector machine also expresses nearby accuracy percentage on less execution time.

7 CONCLUSION AND FUTURE WORK

In this study, the investigation has done on the student's placement log collected from heterogeneous data sources. Two major works, such as predicting the interesting pattern for student placement opportunities and analyzing the data set using four prominent classifier models. Analysis carried out in the aspects of analyzing the performance metrics of different classifiers. Upon completion of the analysis, deep insights in classifier accuracy, error rate, validating the accuracy are assessed and presented in the discussion. The study concludes with the classifier multilayer perceptron is an efficient one on the student's placement log dataset. In the future, the study can be moved towards finding the best optimization technique to improve the classification accuracy by Influencing the parallel processing on MLP.

REFERENCES

[1] Ahmed Mohamed Ahmed, Ahmet Rizaner, Ali Hakan Ulusoy," Using data mining to predict instructor performance", Procedia Computer Science 102 (2016) 137 – 142.
[2] Alejandro Peña-Ayala,"Educational data mining: A survey and a data mining-based analysis of recent works", Expert Systems with Applications 41 (2014) 1432–1462.
[3] Amirah Mohamed Shahiri,Wahidah Husain, Nur'aini Abdul Rashid,"A Review on Predicting Student's Performance using Data Mining Techniques", Procedia Computer Science 72 (2015) 414 – 422.
[4] Amrita Naik, Lilavati Samant, "Correlation review of classification algorithm using data mining tool: WEKA, Rapidminer , Tanagra ,Orange and Knime", Elsevier-Procedia Computer Science 85 (2016) 662 – 668
[5] K.Arunmozhi Arasan, Dr.E.Ramaraj, S.Muthukumaran,"Generating Association Rules To Identify Adolesence Behavior Of Students In Higher Educational Institutions", international journal of scientific & technology research volume 8, issue 09, september 2019
[6] Concepción Burgos , MaríaL. Campanario , Daviddela Peña, JuanA. Lara , David Lizcano, María A. Martínez,"Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout" Computers and Electrical Engineering 66 (2018) 541–556.
[7] Dalynés Reyes-Colón, Gilberto Crespo-Pérez,"Blended Learning: An Alternative For Undergraduate Anatomy Teaching In Developing Countries", international journal of scientific & technology research volume 7, issue 3 , march 2018.
[8] Evandro B. Costa, Baldoino Fonseca, Marcelo Almeida Santana, Fabrísia Ferreira de Araújo, Joilson Rego ," Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses", Computers in Human Behavior 73 (2017) 247e256.
[9] Huseyin Guruler , Ayhan Istanbullu , Mehmet Karahasan , "A new student performance analysing system using knowledge discovery in higher educational databases" Computers & Education 55 (2010) 247–254.
[10] Ian H. Witten, Eibe Frank Martinez, "Data Mining: Practical Machine Learning Tools and Techniques", -2d-Ed - Morgan Kaufmann series in data management systems ISBN: 0-12-088407-0[Book Type]
[11] Neesha Jothi, Nur'Aini Abdul Rashid, Wahidah Husain, " Data Mining in Healthcare – A Review" Elsevier- Procedia Computer Science 72 (2015) 306 – 313
[12] Parneet Kaur,Manpreet Singh,Gurpreet Singh Josan, "Classification and prediction based data mining algorithms to predict slow learners in education sector" Elsevier-Procedia Computer Science 57 (2015) 500 – 50
[13] Siti Khadijah Mohamad, Zaidatun Tasir," Educational data mining: A review", Elsevier- Procedia - Social and Behavioral Sciences 97 (2013) 320 – 324.
[14] Srecko Natek, Moti Zwilling, "Student data mining solution–knowledge management system related to higher education institutions", Expert Systems with Applications 41 (2014) 6400–6407.
[15] Steve Humble, Pauline Dixon,"The effects of schooling, family and poverty on children'sattainment, potential and confidence—Evidence from Kinondoni, Dar es Salaam, Tanzania Steve Humble, Pauline Dixon", International Journal of Educational Research 83 (2017) 94–106.
[16] Sukaina Alzyoud, Mohammad Kharabsheh, Rola Mudallal," Predicting Depression Level of Youth Smokers Using Machine Learning", international journal of scientific & technology research volume 8, issue 11, November 2019.