Placement Predictor and Course Recommender System

Priyanka P. Wadekar[1], Yedhukrishnan P. Pillai[2], Manodeep U. Roy[3], Prof. Neelam Phadnis[4]

[1,2,3,4] *Department of Computer Engineering, Shree L.R. Tiwari College of Engineering, Maharashtra, India*

Abstract - *The performance in education sector in India is a turning point in the lives of all students. As this academic performance is influenced by many factors, it is essential to develop predictive data mining model for students' performance so as to identify the slow learners and study the influence of the dominant factors on their academic performance. The educational sector in IT includes the student records namely aptitude skills, certification courses, technical abilities in various languages or web development and academic performance. It may be an important consideration to analyze various trends since all the systems are now computer based information system so data availability, modification and updating are a common process now. Student achievement is highly influenced by past evaluations which involves various relevant features (e.g. attendance in lectures/practical, participation in various intercollegiate college events, test scores, etc.). As a direct outcome of this project, more efficient student prediction tools can be developed, improving the quality of education and enhancing school resource management.*

***Key Words*: Placement, Course, Academics, Data Mining, C 4.5, Naïve Bayesian**

1. INTRODUCTION

The current education system does not involve any prediction about fail or pass percentage based on the performance. The system doesn't deal with dropouts. It doesn't identify the weak student and inform the teacher. Students are often found struggling to find help with coursework, or having difficulty choosing (or getting into) the courses they need, many students are daunted by the task of working through the collegiate bureaucracy. One of the biggest challenges that higher education faces today is predicting the paths of students. Institutions would like to know, which students will enroll in which course, and which students will need more assistance in particular subject. The performance, success of student's in the examination as well as their overall personality development could be exponentially accelerated by thoroughly using data mining technique to evaluate their admission academic performance and finally the placement. Currently there is no system to predict the chances of a student being placed in a company. The Placement Predictor and Course Recommender (PPCR) is a new system that replaces the current methodology of hiring fresher and eliminating unnecessary rounds and incapable applicants. The system applies data mining techniques using decision tree and Naïve Bayes classifier. It will recommend courses to students to improve their chances of getting placed in a company. It will also help companies to analyze the quality of students in a particular institute. The system will also help the educational organization to analyze weak students and provide them with academic help.

The paper is organized into five sections. The first section gives a brief introduction about the system. The second section is about data mining and the study of related existing systems. The third section details out the implementation of the system. The fourth section provides the results obtained using mining algorithms. Finally the conclusion gives the summary and future scope about the system.

2. LITERATURE REVIEW

In Literature review, we discuss about the various aspects of the project by taking reference of the existing projects that are similar to the makers of this current project.

Bhullar, Manpreet Singh and Amritpal Kaur [1] have stated different levels of Data Mining, their various phases, and advantages and also classification of data is done using WEKA data mining tool. The mining algorithm used here is J48. Since there was no clustering, the domain and the interests of students couldn't be evaluated in a particular field. The accuracy of this system is 77.4%.

Agarwal, Sonali, G. N. Pandey, and M. D. Tiwari [2] have performed and have done a comparative analysis using the best classifier with maximum accuracy and minimum root mean square error (RMSE) using Support Vector Machines (SVM). Comparison of 8 classification algorithms including Logistic, Multilayer Perceptron, LIBSVM, RBFNetwork, Simple Logistic, SMO, Voted Perceptron, Winnow are done in which LIBSVM with Radial Base Kernel gives the best and accurate result of 97.03% accuracy. Other algorithms like RBF network and Multilayer Perception gives 96.05% and 95.85% respectively.

Ramaswami, M., and R. Bhaskaran [4] have done a survey cum experimental methodology to generate a database using a primary and a secondary source. The primary data was collected from the regular students and the secondary data was gathered from the school and office of the Chief Educational Officer (CEO). A total of 1000 datasets of the year 2006 from five different schools in three different districts of Tamil Nadu were collected out of which 772 student records were collected after data transformation, which were used for CHAID prediction model construction. A set of prediction rules were extracted from CHIAD prediction model and the efficiency of the generated CHIAD prediction model was found. The accuracy of the present model was compared with other model and it has been found to be

satisfactory. The prediction obtained from the proposed model gave a result with accuracy of 44.69%.

Zaíane, Osmar R. [6] have used Web Mining Techniques to build such an agent that recommends on-line learning activities or shortcuts in a course web site based on learners' access history to improve course material navigation as well as assist the online learning process. Visualisation techniques are being used here. This process conveys information which a user quickly understands and digests completely.

Han and Kamber [7] describes about data mining and its various algorithms which helps the users to analyse data from different dimensions, classify it and the relationships which are identified during the mining process.

Ajay Kumar Pal and Saurabh Pal [8] have conducted tests using three set of algorithms viz. Naïve Bayes, Multilayer Perceptron and J48 using WEKA tool. Since all these algorithms provide different results, they have taken the average value of all the algorithms and final result is displayed in the form of variables ranking, instead of selecting one algorithm and trusting it. The highest accuracy belongs to the Naïve Bayes Classifier with 86.15% followed by Multilayer Perceptron function with a percentage of 80.00% and subsequently J48 tree with 75.38%.

3. DATA MINING

Data mining basically means "mining' or 'extraction' of knowledge from a large amount of data. Here we examine a large pre-existing databases to generate new information. It is a process of discovering various patterns in huge data-sets that involves methods of machine learning, statistics and database systems.

In Data Mining a function that assigns different types of items in a collection to target classes or categories is called "Classification". The main goal is to predict the target class for each case in data. The Bayesian classifiers are statistical classifiers that can predict class membership probabilities. On the other hand, Decision Tree is in the form of a tree structure which builds classification and regression models. The goal is to break down a data-set into smaller and smaller subsets and create a model that predicts the value of a target variable based on different input variables.

3.1 C 4.5 Algorithm

The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set of already classified samples. Each sample consists of a p-dimensional vector, where they represent attribute values or features of the sample, as well as the class in which falls.
At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sub lists.

3.2 Naïve Bayesian Classification

Naïve Bayes Classifiers are basic classifiers that work on probabilistic values based on the Bayes' theorem. This algorithm requires a number of parameters linear in number of features/variables. Bayesian classification gives a practical knowledge and prior information on learning algorithms. It calculates probabilities for hypothetical values and is robust to noise in input.

Naïve Bayesian algorithm gives a maximum or near to maximum accuracy in lesser number of data. Since it is a generative model hence it returns a probabilistic value. The greatest advantage of this system is that a small change in the training set ill not make a big change in the model. [10]

3.3 Multilayer Perceptron

Multilayer Perceptron (MLP) algorithm is one of the most widely used and popular neural networks. The network consists of a set of sensory elements that make up the input layer, one or more hidden layers of processing elements, and the output layer of the processing elements. [11]

MLP is especially suitable for approximating a classification function which sets the example determined by the vector attribute values into one or more classes.

4. SYSTEM ARCHITECTURE

The Placement Predictor and Course Recommender system predicts the chances of a student being placed in a company. It will recommend courses to the students before predicting the result in case they do not pass the tests.

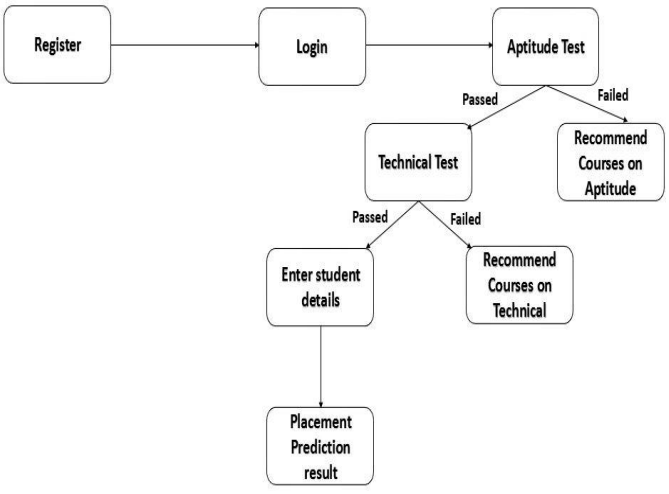

Fig -1: System Architecture

5. IMPLEMENTATION OF THE SYSTEM

Here we will discuss about how we implemented our system and is represented in a flowchart manner in Figure 3.

5.1 Register And Login

In Fig. 2 we can see that first the student registers himself first and then logs in to the system using his/her registered username and password.

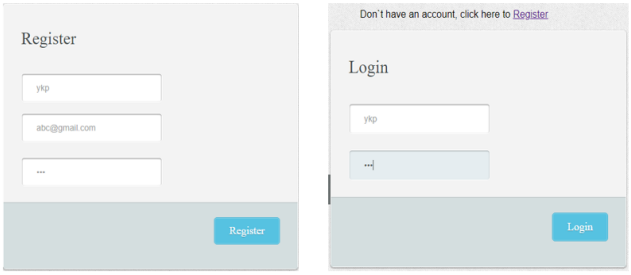

Fig -2: GUI of Register and Login

5.2 Eligibility Criteria

In this part we have provided two simple tests to identify the students who are eligible for placement prediction.

Step 1: Aptitude Test

It will test the basic aptitude skills of the student who are giving the test. Students who fail in the test are recommended to improve their aptitude skills. Students who have cleared are eligible for technical test.

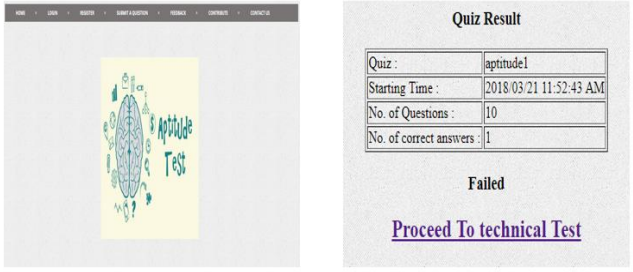

Fig -3: GUI of Aptitude test

Step 2: Technical Test

It will test the technical knowledge of the student in languages such as Java, PHP, SQL, Mongo DB, Python, Linux, JavaScript and CSS. Students who fail in the test are recommended to improve that particular programming language. Students who have cleared are eligible for placement prediction.

Fig -4: GUI of Technical test

5.3 Prediction Process

Step 1: Data Acquisition

The data set used in this study was obtained from the computer engineering batch of 2016-17 Shree LR Tiwari College of Engineering. Initially the size of the training data is 200 records.

Step 2: Data selection and Transformation

In this step only those fields were selected which were required for data mining.

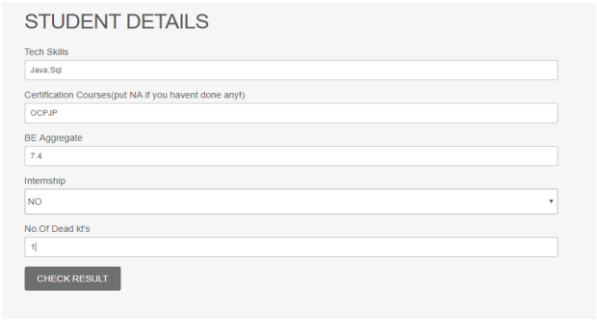

Fig -5: Placement related variables

Step 3: Data Mining Algorithms

For the prediction process classification algorithms C4.5 and Naïve Bayes Classifier were used on the training data.

- **Steps of C 4.5 algorithm**

1. Check for the base case.
2. Construct a DT using random training data.
3. Find the attribute with the highest info gain (A_Best).
4. A_Best is assigned with entropy minimization.
5. Partition S into S1, S2, S3...
6. According to the value of A_Best.
7. Repeat the steps for S1, S2, and S3.
8. For each ti€D, apply the DT. [12]

- **Steps of Naïve Bayesian algorithm**

1. Assume D to be training set of tuple. Every record can be represented by n-dimensional attribute vector i.e. X=(x1, x2…, xn), predicting n measurements on tuple from n attributes, i.e. A1 to An.

2. Let m number of class for prediction (C1, C2……. Cm). As for record X, the classifier predict that X will belong to the class with maximum posterior probability that is conditioned on X. Naïve Bayes predict that the tuple x will belong to class Ci only if P (Ci|X)>P (Cj|X). Therefore we have to maximize P (Ci|X).

$$P\left(\frac{C_i}{X}\right) = \frac{P\left(\frac{X}{C_i}\right) * P(C_i)}{P(X)}$$

3. Because P(X) is constant in all classes, therefore P (X|Ci)* P (Ci) need be maximized.

4. As then assumption of class conditional independence is done. Therefore it is pre assumed that value of attributes are conditionally independent of each other.

Thus,

$$P\left(\frac{X}{Ci}\right) = \prod_{k=1}^{m} P\left(\frac{X_k}{C_i}\right) = P\left(\frac{X_1}{C_i}\right) * P\left(\frac{X_2}{C_i}\right) \ldots\ldots P\left(\frac{X_m}{C_i}\right)$$

5. To predict class of X, P (X|Ci) P (Ci) is calculated for each class Ci. Naive Bayes predict that class label of X is Ci class if [12]

$$P\left(\frac{X}{C_i}\right) P(C_i) > P\left(\frac{X}{Cj}\right) P(Cj) \; for \; 1 \le j \le m, \; j \ne i$$

Step 4: Testing

Weka is an open source software that implements a large collection of machine leaning algorithms and is widely used in data mining applications. Under the "Test options", the 10-fold cross-validation is selected as our evaluation approach. The data set used for testing was obtained from the computer engineering batch of 2017-18.

Step 5: Placement Prediction

After running the model on the test set C4.5 algorithm was found to be more accurate. The final placement prediction is based on the more accurate classifier.

Hi !!!! You Have Passed !!!You are most likely to get placed !!Good Job :)

Fig -6: Prediction result

6. RESULTS

The Placement Predictor and Course Recommender system applies data mining techniques using decision tree and Naïve Bayes classifier. Decision trees are considered easily understood models because a reasoning process can be given for each conclusion.

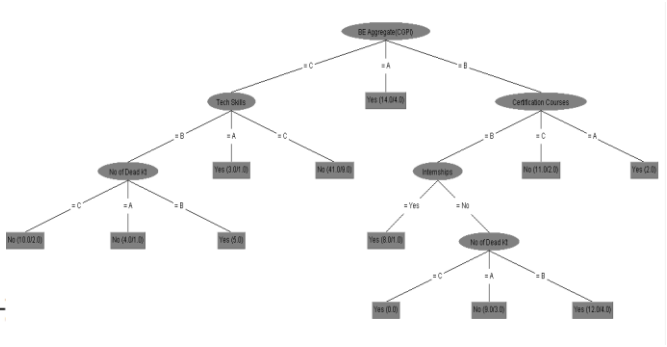

Fig -7: Decision Tree of PPCR

```
BE Aggregate(CGPI) = C
|   Tech Skills = B
|   |   No of Dead Kt = C: No (10.0/2.0)
|   |   No of Dead Kt = A: No (4.0/1.0)
|   |   No of Dead Kt = B: Yes (5.0)
|   Tech Skills = A: Yes (3.0/1.0)
|   Tech Skills = C: No (41.0/9.0)
BE Aggregate(CGPI) = A: Yes (14.0/4.0)
BE Aggregate(CGPI) = B
|   Certification Courses = B
|   |   Internships = Yes: Yes (8.0/1.0)
|   |   Internships = No
|   |   |   No of Dead Kt = C: Yes (0.0)
|   |   |   No of Dead Kt = A: No (9.0/3.0)
|   |   |   No of Dead Kt = B: Yes (12.0/4.0)
|   Certification Courses = C: No (11.0/2.0)
|   Certification Courses = A: Yes (2.0)

Number of Leaves  :     12

Size of the tree :      18
```

Fig -8: IF-THEN rules of PPCR

Experiments were carried out in order to evaluate the performance and usefulness of different classification algorithms for predicting students' placement. The results of the experiments are shown below:

```
=== Summary ===

Correctly Classified Instances          92               77.3109 %
Incorrectly Classified Instances        27               22.6891 %
Kappa statistic                          0.5287
Mean absolute error                      0.3374
Root mean squared error                  0.4107
Relative absolute error                 68.8653 %
Root relative squared error             82.9992 %
Total Number of Instances              119
```

Fig -9: Results using C 4.5

```
=== Summary ===

Correctly Classified Instances          80               66.6667 %
Incorrectly Classified Instances        40               33.3333 %
Kappa statistic                          0.2925
Mean absolute error                      0.3901
Root mean squared error                  0.4512
Relative absolute error                 79.4011 %
Root relative squared error             91.0608 %
Total Number of Instances              120
```

Fig -10: Results using Naïve Bayes Classifier

The percentage of correctly classified instances is often called accuracy or sample accuracy of a model. So Naïve Bayes classifier has more accuracy than other two classifiers. The Accuracy of the predictive model is calculated based on the precision, recall values of classification matrix.
PRECISION is the fraction of retrieved instances that are relevant.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

RECALL is fraction of relevant instances that are retrieved.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Classifier	TP	FP	Precision	Recall	Class
C 4.5	0.853	0.333	0.773	0.853	No
	0.667	0.147	0.773	0.667	Yes
Naïve Bayesian	0.838	0.558	0.663	0.838	No
	0.442	0.162	0.676	0.442	Yes

Fig -11: Comparison of evaluation measures

The performance of the learning techniques is highly dependent on the nature of the training data. Confusion matrices are very useful for evaluating classifiers. The columns represent the predictions, and the rows represent the actual class where a= No and b= Yes.

C4.5		Naïve Bayes	
A	B	A	B
58	10	57	11
17	34	29	23

Fig -12: Confusion Matrix

7. CONCLUSION

To conclude, we have predicted the placement results using C4.5 algorithm and Naïve Bayes Classifier. The best algorithm based on the training data is C4.5 with an accuracy of 77.31%. Naïve Bayes classifier has the accuracy of 66.67%. In educational field, C4.5 gives much better prediction than any other classification algorithms. We have observed that as the number of data increases the prediction accuracy using C4.5 also increases. This system can be further used in various college's placement cells to help enhance the placement procedure.

REFERENCES

[1] Bhullar, Manpreet Singh, and Amritpal Kaur. "Use of data mining in education sector." Proceedings of the World Congress on Engineering and Computer Science. Vol. 1. 2012.

[2] Agarwal, Sonali, G. N. Pandey, and M. D. Tiwari. "Data mining in education: data classification and decision tree approach." International Journal of e-Education, e-Business, e-Management and e-Learning 2.2 (2012): 140.

[3] Salazar, A., et al. "A case study of knowledge discovery on academic achievement, student desertion and student retention." Information Technology: Research and Education, 2004. ITRE 2004. 2nd International Conference on. IEEE, 2004.

[4] Ramaswami, M., and R. Bhaskaran. "A CHAID based performance prediction model in educational data mining." arXiv preprint arXiv: 1002.1144 (2010).

[5] Cortez, Paulo, and Alice Maria Gonçalves Silva. "Using data mining to predict secondary school student performance." (2008)

[6] Zaíane, Osmar R. "Building a recommender agent for e-learning systems." Computers in education, 2002. Proceedings International conference on. IEEE, 2002.

[7] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000.

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

[9] Sujith Jayaprakash, Balamurugan E., Vibin Chandar "Predicting Students Academic Performance using Naive Bayes Algorithm".

[10] http://gerardnico.com/wiki/data_mining/naive_bayes

[11] Witten, I.H. & Frank E., Data Mining– Practical Machine Learning Tools and Techniques, Second edition, Morgan Kaufmann, San Francisco, 2000.

[12] S. Nagaparameshwara Chary, Dr. B.Rama "Analysis of Classification Technique Algorithms in Data mining-A Review"