RESEARCH ARTICLE OPEN ACCESS

Placement Prediction Decision Support System using Data Mining

Joshita Goyal[1], Shilpa Sharma[2]

[1]*M.tech Student, Department of Computer Science and Engineering, Lovely Professional University, Phagwara*
[2]*Assistant Professor, Department of Computer Science and Engineering, Lovely Professional University, Phagwara*

Abstract:

With the growth of voluminous amount of data in educational institutes', the need is to mine the large dataset to produce some useful information out of it. In this research we focused on to form a decision support system for the educational institutes' which can help them to know about the placement possibility of students. Our research is not limited to find out placement possibility but we did multi-level analysis on student performance dataset which will predict that what level of interview process a student is likely to pass. For this we have applied Naïve Bayes and Improved Naïve Bayes which is integrated with relief feature selection technique to obtain the prediction. Data analysis was done using NetBeans and WEKA. For this our proposed technique gave better accuracy than existing naïve Bayes which was 84.7% and naïve Bayes gave 80.96% accuracy.

Keywords —**Naïve Bayes, NetBeans,Relief and WEKA**

I. Introduction

Now a days, data in educational institutes' is growing rapidly and in voluminous amount. To process large volume of data available in databases of educational institutes' we are provided with data mining to figure out the relation between various attributes and to predict the result.Data mining when applied on educational data is called education data mining[1]. Education data mining uses techniques which are analytical tools for extracting and processingdata available for data analysis[2].The various techniques are classification, clustering, feature selection and association rule mining.

Education data mining is a powerful technique to make the data available in the data repositories of institutes' useful. By using previous data for future prediction, data mining can help a lot in raising the institutes' value. By using data mining we focused on to form a decision support system which can assist placement cell of the institute to know the student performance and corresponding to that,particular inputs can be given to the students which can enhance placements. Education data mining is used in this research is to find out various placements related issues which are as follow:

- To find out the correlation between various attributes available in mined dataset[3].
- To find out how particular student is likely to pass a certain level of the interview process.
- To find out what input or subject a student needs more attention in to get placed.

There are a lot of data mining techniques available for knowledge discovery. Some of them are well illustrated below for better understanding.

Classification: Classification is a technique which predicts a particular target class for a particular case in dataset based on the knowledge it gained from previous dataset[4]. Classification techniques are applied on detection of fraud, disease detection and also for placement prediction. Various classification techniques are random forest, j48, ID3 decision tree, naïve Bayes, neural networks, support vector machine and k- nearest neighbor classifier.

Clustering: Clustering is a technique which is used to group similar and dissimilar data based on Euclidian distance.Clustering is used in various fields. We can also group placed and non-placed students using clustering.

Association: Association is used to find out the frequent patterns from the available dataset. It is used to find out that who is likely to be the best customer for increasing profits. Apriori and feature selection is used in association rule mining.

II. LITERATURE REVIEW

To find out the importance of data mining in education, here is an insight of related work which has been done by many researchers.

Dijana, Mario and Milićperformed cluster analysis on set of students and found out cluster which was having higher of students who were having good score in subject like mathematics and English in high school. They were found to be female and not the males so they broke the hypothesis stating that guys can perform better. Hence they proved that such systems can be formed with clustering[5].Maryam Zaffar, Hashmani and Savita did performance analysis of various feature selection algorithms and classifiers and compared their results on basis of recall, precision and f-measure. It was found out that random forest when embedded with principal component analysis gives better results than other techniques[6].Ashok MV and Apoorva worked on to form placement prediction system which is based on students' overall percentage and their skill set. They proposed an algorithm for this which was compared with decision tree, naïve bayes and neural network. The proposed algorithm gave better accuracy than the existing ones[7]. Animesh, Vignesh, BysaniPruthvi and Naini formed a system to assist placement office and students to know where they stand. They took dataset consisting of students' details. They applied KNN, logistic regression and support vector machine. KNN classifier gave better accuracy than logistic regression and support vector machine[8]. MangasuliSheetal and Savitaformed a placement prediction system using fuzzy logic and KNN. The two techniques were compared and it was found that KNN gives more accuracy than fuzzy inference system[9]. Keno C, Dumlao, Melvin and Shaneth formed a system to find out the unemployability rate of countries under Association of south east Asian nations and found out that Philippines has more unemployed people. They applied naïve bayes, J48, SimpleCart, Logistic Regression and Chaid but logistic regression gave high accuracy and low error rate[10].Getaneh and Dr. Sreenivasaraocollaborated to form a system in which students are placed in different university departments based on their scores in entrance exam. For this applied naïve bayes, j48 and random forest and predicted that such systems can be formed using j48 as it gave higher accuracy[11]. Anupam and soumya k. collaborated to evaluate the relation between poor student results and teaching quality. In this authors applied apriori to find it out. Rules formed from association rule mining broke the hypothesis that the reason for poor results is poor student quality and it was determined that there are other factors which affect the student performance[12]. Liang, Huang, Qing, Yunheng and Lang worked on to find out which student isextraversion or introversion. For this used usedsci-kit and applied naïve bayes,classification and regression trees, linear SVM. And it was predicted that students who pay more attention online are likely to be introversion and others are extroversion. For this linear SVM performed better than the other two by giving high accuracy[13].Larian and Muesser found out whether the students access online material available to them on last minute or they procrastinate online submission. They used weka for data analysis and association rule mining to find out[14].

III. BACKGROUND

A. **Data Gathering:**Since we worked on to form placement prediction system which will not just assist the placement office of our own institute but can also become a decision support system for placements of other institutes. So data taken for the placement analysis of students include the student details which are their registration number, names, program, batch start, batch end ,10th score, 12th score, CGPA, standing arrears and scores of pre assessment taken. Scores of pre assessment included scores of Quantitative Ability, Logical Ability, English Competency, Written English, Computer fundamentals, Domain Knowledge andCoding ability. Evaluation of general aptitude of students willing for placements is considered for level 1 prediction of recruitment process and evaluation of general aptitude with coding ability is considered for prediction of level 2 of recruitment process. For predicting that whether a student is able to pass the group discussion level considered as level 3 of recruitment process and technical and HR interview level considered as level 4 of the recruitment process, scores of Group discussion Technical and HR interview were also considered.

B. **Pre-processing:** In this step, some attributes which were registration no, name, program, batch start and batch end were eliminated for training the system to have better predictions. Other attributes were taken in training dataset. This dataset contained in excel was converted to arff (Attribute relation file format) accepted by Weka for data analysis. Training dataset contained all the parameters including eligibility, level 1, level 2, level 3, level 4 and placement possibility.

C. **Naïve Bayes classification:** Naïve Bayes algorithm which works on the principle of Bayes theorem was used. Naïve Bayes classification was used to estimate the value of target variables. As our dataset contains categorical data as well so naïve Bayes was chosen as better option for classification as this algorithm works better for this kind of data. Another reason for using naïve Bayes classification was that there is data independence in our dataset such that one attribute doesn't depend on another attribute for prediction of target variable. This can be understood by considering the well-known example of a fruit say apple. An apple can be predicted as apple by naïve Bayes if the attributes taken for its prediction are color as red, diameter as 3inchesand shape as round. This implies that all features contribute independently for prediction of apple as apple. That's why naïve Bayes is called naïve. Bayes theorem which forms the foundation for naïve Bayes algorithm can be expressed as the following equation for two events A and B:

$$P = (A|B) = \frac{P(B|A)P(A)}{P(B)}$$

In this above equation P is probability of occurrence of event A if we are given event B is true.

P(B|A) is probability of occurrence of event B if event A occurs.

P(A) and P(B) are individual probabilities of both the events.

D. **Improved Naïve Bayes algorithm:** As simple Naïve Bayes gives less accurate prediction so we embedded relief feature selection technique in it. Feature selection whenever used in classifier improves its accuracy to greater extent. Relief filter chooses the best attribute which contributes more to prediction and assigns weights and ranks to all attributes. More the value of weight and higher the rank then that attribute will contribute more to the prediction. Weights assigned to attributes lies in the range of -1 to 1[15]. To understand more about the Relief algorithm, say an instance is taken from dataset. Then the filter finds its nearest neighbor if the nearest neighbor belong to same class then it is called nearest hit else we call it nearest miss. Also a change in attribute value accompanied by a change inclassleads up to weighting of the attribute based on the intuition that the attribute change could be responsible for the class change. On theother hand, a change in attribute value accompanied by no change in class leads to down weighting of the attribute based on theobservation that the attribute change had no effect on the class. This procedure of updating the weight of the attribute is performed fora random set of samples in the data or for every sample in the data. The weight updates are then averaged so that the final weight is inthe range [−1, 1].This is how naïve Bayes was improved for better analysis. Comparison was made between naïve Bayes and Improved Naïve Bayes based on certain parameters which were correctly classified instances, incorrectly classified instances, accuracy, precision, recall, F-measure. Results of all these parameters were obtained from confusion matrix.

Confusion matrix tells us about the performance of the classification algorithm. From this, true positive rate which is termed as sensitivity, false positive rate or type 1 error, true negative rate also called as specificity and false negative rate or type 2 error is evaluated.Below is confusion matrix table followed by equations to obtain above mentioned values :

	Predicted yes	Predicted No
Actual Yes	True positive	False negative
Actual No	False positive	True negative

CONFUSION MATRIX

$$True\ Positive\ rate = \frac{TP}{TP+FN}$$ Equation (1).

$$False\ Positive\ rate = \frac{FP}{FP+TN}$$ Equation (2).

$$False\ Negative\ rate = \frac{FN}{TP+TN}$$ Equation (3).

$$True\ Negative\ rate = \frac{TN}{TN+FP}$$ Equation (4).

Pseudo code for improved naïve Bayes is as follow:

Input: Dataset
Output: Ranking of the attributes

```
set W[a] = 0
for each attribute a for his = 1 to n do
select sample $S_{his}$ from data at random
find nearest hit $S_h$ and nearest miss $S_m$
for each attribute a do
$\Delta W_{his}$ [a] = diff(a, $S_{his}$ , $S_m$ ) – diff(a, $S_j$ , $S_h$ )
W[a] = W[a] + $\Delta W_{his}$ [a]
```

```
end for
end for
for each attribute a do
W[a] = W[a] / n end for
where diff (a, $S_{his}$ , $S_j$ ) = 0, if $s_{his}$ [a] = $S_j$ [a]
                                   = 1, if $S_{his}$ [a] ≠ $S_j$ [a]
```

These ranked attributes are given as an input to naïve Bayes classification.

E. **Post Processing:** For cross-validation, data was divided into chunks so as get predictions of eligibility, Level 1, Level 2, Level 3, Level 4 and placement possibility in different interfaces created for prediction of each variable.

F. **Tools used:** For data analysis Weka was used. Weka stands for Waikato environment for knowledge analysis. Weka contains all the data mining algorithms used for analysis. It is open to use and freely available. All machine learning approaches are written in java language. Data accepted by Weka should be in csv or arff format. It contains graphical user interface and visualization tools so it's easy to use. In this research, we included weka.jar file in NetBeans to have access to all libraries available in weka and for data analysis. NetBeans is integrated development environment which is used for writing business logic and creating softwares. All interfaces for data prediction were created in it.

IV. TESTING AND RESULTS

For placement analysis, first naïve Bayes was applied on student performance dataset having 560 instances which gave accuracy of 80.96% and then improved naïve Bayes was applied on the same dataset which gave accuracy of 84.7%.

TABLE 1.
COMPARISON BASED ON CORRECTLY AND INCORRECTLY CLASSIFIED INSTANCES

Parameters	**Correctly classified instances**	**Incorrectly classified instances**
Naïve Bayes	455	107
Improved Naïve Bayes	476	86

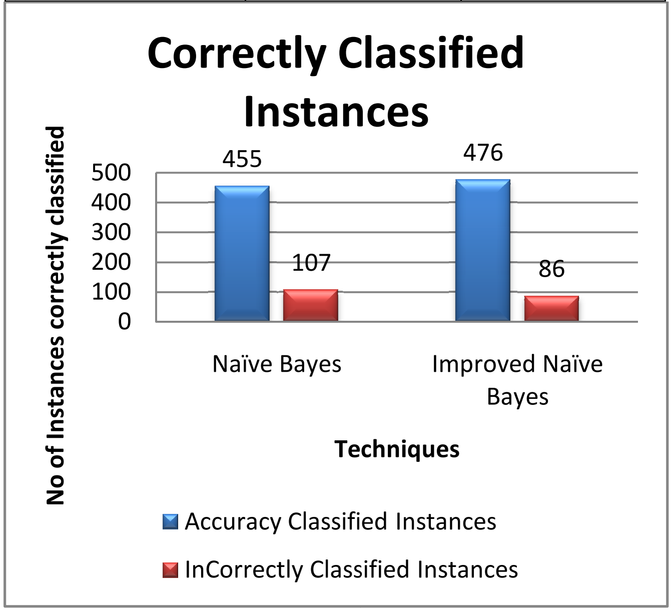

Fig 1. Pictorial representation of correctly and incorrectly classified instances.

Naïve Bayes and improved Naïve Bayes are compared based on accuracy, TP rate, Precision, recall and F-measure.

Accuracy obtained from confusion matrix is the percentage of true positive rate such that correctly predicted instances in dataset. Accuracy obtained after implementation of naïve Bayes and improved naïve Bayes is 80.96% and 84.7% respectively.

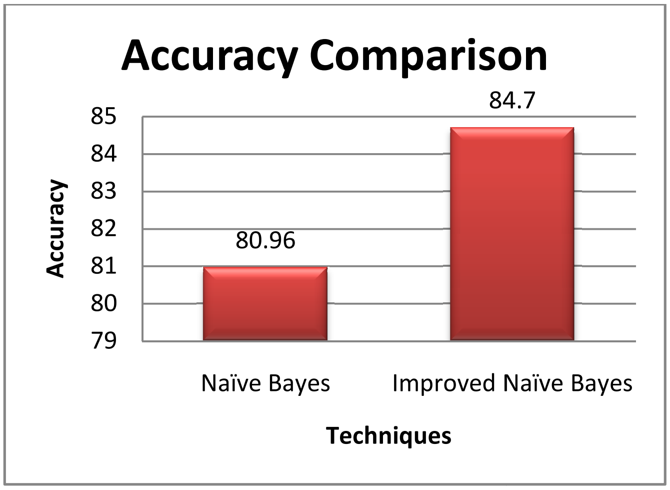

Fig 2. Pictorial representation of accuracy comparison

Precision can be better understood by considering an example, say our system predicts that out of a dataset of twenty students, twelve can get placed. But out of those twelve, ten can actually get placed and two can't. So precision is the ratio of actual true positive to the number of predicted true positive. So it describes that how useful the prediction is.

Recall is another comparison parameter taken. Recall tells the completeness. If we consider the example taken to understand precision then recall will the fraction of actual true positives and total instances in dataset.

F-measure also called F_1 score signifies test's accuracy. It can be expressed as the ratio of correct positive results to the positive results returned by the classifier.

TABLE 2.
RESULTS BASED ON COMPARISON PARAMETERS

Parameters	TP rate	Precision	Recall	F-measure
Naïve Bayes	0.81	0.87	0.81	0.823
Improved Naïve Bayes	0.847	0.881	0.847	0.855

.

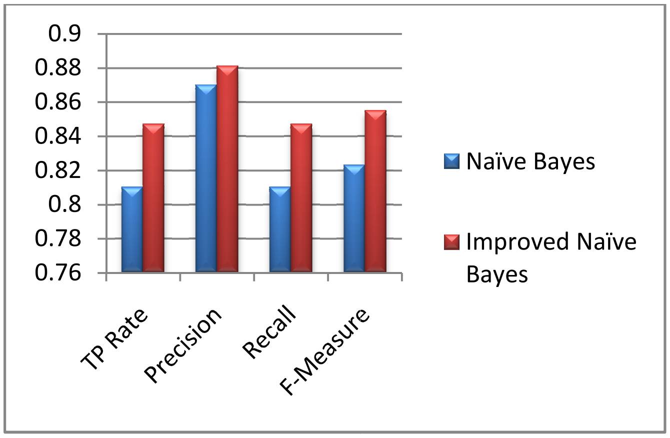

Fig 3.Pictorial representation of comparison between two techniques based on class parameters.

V. CONCLUSION AND FUTURE SCOPE

After the comparison, improved naïve Bayes was applied on testing dataset for cross validation consisting of hundred instances. It was predicted that out 100 students, 88 were eligible, 87 were able to pass level 1, 64 were able to pass level 2, 62 were able to pass level 3, 51were able to pass level 4 and finally placement possibility was predicted as yes for 61 students and no for 39students. This can serve as better placement prediction system for any institute as they will come to know that which student is likely to clear which level of the recruitment process so that corresponding inputs can be given to students who are unable to pass a certain level. Representation of every level individually in their corresponding interfaces can lead to better understanding of student performance. So thissystem can enhance future placements by providing the adequate knowledge to the Institute. We would like to extend our work in future by using other machine learning algorithms for better predictions.

REFERENCES

1. *Febrianti and VianyUtami ,"Predicting Students Performance in Final Examination using Linear Regression and Multilayer Perceptron" 2017 10th International Conference on Human System Interactions (HIS),2017, Pages: 188 – 192*
2. *Sentkil Kumar, P. DivyaBkaratki and AbijitkSankar,"A data mining techniques for campus placements*

prediction in higher education"2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS),2017,Pages: 1 – 5

3. *AngelosCharitopoulos; Maria Rangoussi; DimitriosKoulouriotis,"Educational data mining and data analysis for optimal learning content management: Applied in Moodle for undergraduate engineering studies, 2017 IEEE Global Engineering Education Conference (EDUCON),2017, Pages: 990 – 998*
4. *Siddhi Parekh, Ankit Parekh, AmeyaNadkarni and RiyaMehta,"Results and Placement Analysis and Prediction using Data Mining and Dashboard"International Journal of Computer Applications (0975 – 8887) Volume 137 – No.13, March 2016*
5. *DijanaOreški, Mario Konecki and Luka Milić,"Estimating profile of successful IT student: data mining approach,"2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),2017, Pages: 723 – 727*
6. *Maryam Zaffar, Manzoor Ahmed Hashmani and K. S.Savita," Performance analysis of feature selection algorithm for educational data mining",2017 IEEE Conference on Big Data and Analytics (ICBDA),2017, Pages:7-12*
7. *Ashok MV and Apoorva A," Data Mining Approach For Predicting Student and Institution's Placement Percentage", 2016 International Conference on Computational Systems and Information Systems for Sustainable Solutions,2016,Pages: 336 – 340*
8. *AnimeshGiri, M Vignesh V Bhagavath, BysaniPruthvi, NainiDubey,"A Placement Prediction System Using K-Nearest Neighbors Classifier" 2016 Second International Conference on Cognitive Computing and Information Processing (CCIP),2016,Pages: 1 – 4*
9. *MangasuliSheetal and Prof. SavitaBakare," Prediction of Campus Placement Using Data Mining Algorithm-Fuzzy logic and K nearest neighbor,"International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 6, June 2016*
10. *Keno C. Piad, MenchitaDumlao, Melvin A. Ballera and Shaneth C. Ambat,"Predicting IT Employability Using Data Mining Techniques,"2016 Third International Conference on Digital Information Processing, DataMining, and Wireless Communications (DIPDMWC),2016,Pages: 26 – 30*
11. *GetanehBerieTarekegn and Dr.VudaSreenivasarao,"Application of Data Mining Techniques to Predict Students Placement in to Departments, "International Journal of Research Studies in Computer Science and Engineering (IJRSCSE) Volume 3, Issue 2, 2016, PP 10-14*
12. *Anupam Khan and Soumya K. Ghosh," Analysing the Impact of Poor Teaching on Student Performance", 2016 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE),2016,Pages: 169 – 175*
13. *Liang Ge, Huang Tang, Qing Zhou, Yunheng Tang and Jiangtao Lang," Classification Algorithms to Predict Students' Extroversion – Introversion Traits",2016 International Conference on Cyberworlds (CW),2016,Pages:135-138.*
14. *Larian M Nkomo and Muesser Nat," Discovering students use of learning resources with education data mining",2016 HONET-ICT, 2016,pages:98-102*
15. *S. Francisca Rosario and Dr. K. Thangadurai,"RELIEF: Feature Selection Technique", 2015, International journal of innovative research and development, October 2015, Vol 4, Issue 11, Pages: 218- 224.*