IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Students' Placement Prediction Using Classification Techniques

Sumaira Asif[1*], Tabrez Nafis[2]

[1,2]Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India

[*]*Corresponding Author: asifsumaira25@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i4.289293 | Available online at: www.ijcseonline.org

Accepted: 17/Apr/2019, Published: 30/Apr/2019

Abstract— Data mining techniques are being extensively used in almost every field to gain insights into the large data that is being generated worldwide. These techniques are also being used in analysing the student's performance in several educational institutions. Some of the student data of a particular college was collected with information about their academic performances and individuals' skills such as their aptitude level and communication skills. This data was gathered in order to analyse and predict the placement of the students by applying classification techniques such as decision tree algorithm (J48), K-Nearest Neighbour (IBk) and Naïve Bayes using one of the Data Mining tools known as WEKA. The rules were generated using the decision tree algorithm which helped us in visualizing the decision tree using the above dataset. The accuracy of all the three classifier models was compared and it was shown that the highest accuracy was shown by KNN classifier (IBk in Weka) followed by J48 classifier and Naïve Bayes classifier. This system is thus much reliable and can be used to predict if the student can be placed or not.

Keywords— Data Mining, Classification, Decision Tree, Naive Bayes, Prediction.

I. INTRODUCTION

Nowadays, data is being generated at such a tremendous rate that the mining of data i.e. finding useful information from that data has become the need of the hour. Data mining classification can be on the basis of several factors such as the database, technique used, knowledge type and applications adapted. It has various disciplines such as machine learning, logic programming, Bayesian learning, neural networks etc. [1]. Data mining is being used massively in almost each and every field of study, be it Medicine or Engineering. Data mining finds its application in fraud detection, disease prediction, bio-informatics, insurance industry, as well as in analysing the customer preferences etc.

Over the years, establishment of Universities and colleges in India has increased astoundingly which had a huge impact on the level of education imparted to the students admitted to those institutions. The most significant aspect of any private of Government Academic Institute is its level of Placement which is the foremost criteria that the students take in consideration before deciding their future college. Getting a good placement opportunity right from their own college campus is what almost every student expects when they are on the verge of completing their course. The academic institutes need to understand the crucial role of a placement in a life of a student and then aim to adopt better placement strategies to improve the placement of students. Also, better analysis of the educational details of students can help in predicting beforehand the chances of placement of students. This can be done by building a predictive model using different techniques of data mining and machine learning such as Classification and Regression techniques. Classification techniques include Decision Trees, SVM, Naïve Bayes, and K -Nearest Neighbour etc. Some of the Regression techniques are Linear regression and Logistic regression. Some of the factors on which placement mostly depends are the overall percentage attained by the students and their potential skills. This model can thus be used to predict the student's placement status which can help the students in figuring out where they stand and hence plan out further improvements for themselves accordingly. It can also in a way help the Placement Committee members to take well-informed decisions which can escalate the results of placement which shall ultimately lead to the amelioration of the students and the institution as a whole.

II. RELATED WORK

Data mining is being widely used across the world in different problem domains. The techniques of data mining have also been used in predicting the chance of placement of students efficaciously. Some of the related work has been listed down so as to get an insight into the work that has been

carried out so far which can then help us in realizing further opportunities for growth.

A system (PPS) model was developed using a regression technique called Logistic regression to predict the placement of students in the recruitment session considering the scores of students, subjects and demographics having a good accuracy of 83.3% as shown in [2].

Mansi and Shivani [3] have used decision tree algorithm on the past year's data of companies to build an effective model that can predict if the students are eligible for a particular category of a company or not, so that they prepare only for the companies they have a chance in.

Data mining techniques like decision trees and clustering were applied using WEKA on some students' academic performance data to predict if the students are placed in a core IT company or a consultancy and also to predict the company's name they have been placed in such as the Microsoft, Deloitte or Samsung etc. giving an accuracy of 95% and 62% respectively [4].

Pratiyush and Manu [5] have applied the Bayesian theorem on an educational dataset and Naïve Bayes classifier evaluated the results obtained using tools such as WEKA and Rapid Miner. This was done for placement prediction which can, not only help the students but also the educators in improving the strategies of placement.

A model and an algorithm had been proposed to predict the student's placement chance and also the placement percentage of the academic institute in [6]. Naïve Bayes, Neural Network and Decision tree classification algorithms were then compared to the proposed model showing that the proposed algorithm is better in placement prediction that the others algorithms.

A prediction system was also proposed using K-nearest neighbours classification technique to predict if an under-grad student would get placed in an IT company taking in account the student's academic information and skills[7]. Also the obtained results were compared with the results of other algorithms such as the logistic regression and SVM.

Senthil , Divya and Abijith [8] proposed a recommendation system which could predict one of the five placement statuses for students such as the Core company, Dream company, Mass recruiters, Not interested in placements and Not eligible. This system can help in identifying the potential students and work on improving the others.

III. METHODOLOGY

A. Data Preparation & Description

The dataset used in this study is the Placement data of all the engineering students of a particular batch of a college. The total number of student records in this data set is 70. Out of which, 70% has been used for training the model and the remaining, for testing.

There are large numbers of factors that can affect the placement of a student, personal as well as professional. We have considered the factors that we find to be the most influential, as the attributes of our dataset. The goal of the system is to use these factors and predict the placement of students. The attributes and their description is given as follows-

Table 1: Student related attributes

ATTRIBUTES	DESCRIPTION	POSSIBLE VALUES
B.Tech %	The aggregate percentage attained by the student in B.Tech course	{58,59,60,…100}
Coding	The level of proficiency in Coding	{Good, Average, Poor}
Aptitude	Aptitude skills (Quantitative Aptitude & Logical Reasoning)	{High, Medium, Low}
Comm.	Communication skills	{Good, Average, Poor}
Backlogs	Indicates if the student has any backlogs	{0- Yes, 1- No}
PLACED	Placement status of the Student	{YES, NO}

B. Techniques Used

The classification techniques that we have used for Classification are the Decision Trees, K-Nearest Neighbour and Naïve Bayesian. These algorithms have been used to predict if the student is going to be placed or not, depending upon their academic performance and other measures. We have considered two classes for our Classification i.e. Yes or No.

Decision Tree (J48)

The C4.5 algorithm is implemented in WEKA tool as a classifier called J48 for building decision trees. C4.5 is basically an extension to the ID3 algorithm which was

developed by Quinlan Ross. It has the ability to not only manage the categorical attributes (as in ID3) but also continuous attributes. Gain Ratio, which is one of the attribute selection measures, is used by C4.5 to generate the decision tree. As per the algorithm, the gain ratio for each attribute is calculated and the attribute which gives the highest gain ratio is taken as the root node. Some unnecessary branches in the decision tree are also removed by C4.5, which is known as pruning to increase the accuracy of classification [9].

K-Nearest Neighbor

K-nearest neighbor which is one of the supervised learning algorithms is used for classification. Classification is performed on the basis of various distances such as the Euclidean distance that is between the training data and the testing data. Based on the similarity between the training and the testing data, the nearest k neighbors are selected [7].

Naïve Bayesian

The Naïve Bayesian is another technique of Classification which is mostly suited when the dimensionality of the inputs is high. Though this technique may be very simple in use, it has often proved to perform better than other advanced techniques. Naïve Bayes classifier gives the probability of all of its input attributes for the predictable state. It is basically a very simple probabilistic classifier particularly based on applying Bayesian theorem (from Bayesian statistics) with strong (naive) independent assumptions [10].

C. Implementation

We have used WEKA tool for our implementation that has been developed at the University of Waikato in New Zealand. It comprises of an extensive collection of state-of-the-art machine learning and data mining algorithms which are written in Java. WEKA consists of all the tools that may be required to perform regression, classification, clustering, association rules etc.

We collected the placement data of students having 6 attributes of each of the 70 instances and made it into two Excel files (csv format), one for training the classifier and the other for testing. 49 instances have been kept for training and remaining 21 for our testing. The dataset then had to be converted to an (arff) format in order to make it suitable for the WEKA tool. Three of the classification techniques that we have used and implemented on our dataset using Weka to build the classification models to predict the placement of the students are: J48 decision tree algorithm, the K-Nearest Neighbour (known as Instance-Based learning, IBk in Weka) and Naïve Bayes algorithm. The results that were obtained after building the models are further shown.

IV. Results and Discussion

With the help of the above mentioned attributes, a decision tree is built in Weka tool. The J48 classification algorithm is used to generate the decision tree. The root node that is chosen for the decision tree is BTECH % based on the gain ratios. If the percentage attained by the student is less than or equal to 68%, then the student will not be placed. If the percentage is greater than 68%, then the communication skills of the student are evaluated which can have one of the three possible values i.e. good, average or poor. If the communication skills are average or poor, the student will be placed. And if the communication skills are good, the aptitude level of the student is considered, based on which the placement is predicted. If the aptitude level is high or medium, the student will be placed and if the aptitude level is low, he will not be placed.

The decision tree obtained having six leaf nodes is shown below. The leaf nodes indicate the value if the student has been placed or not.

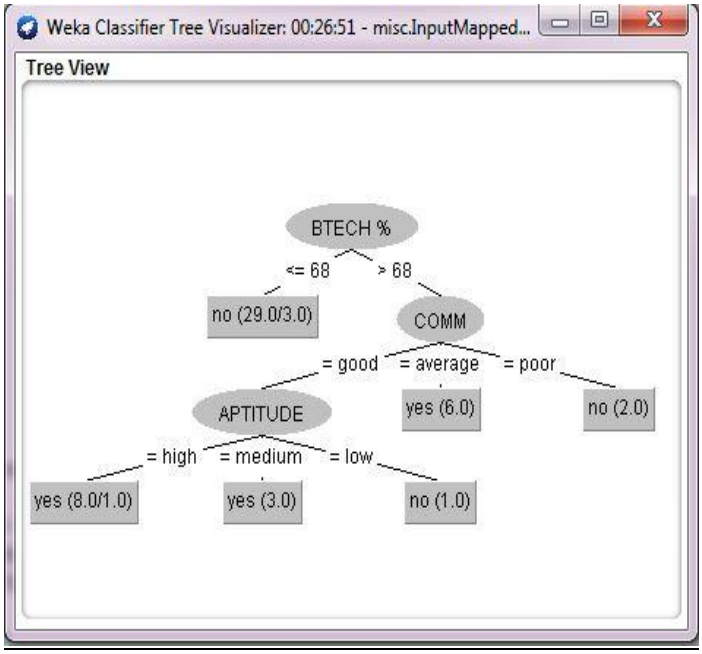

Fig.1 Weka Classifier Tree Visualizer

After applying our dataset using Naive Bayes classifier in WEKA Tool, we obtained a classifier model with each attribute and its class as shown.

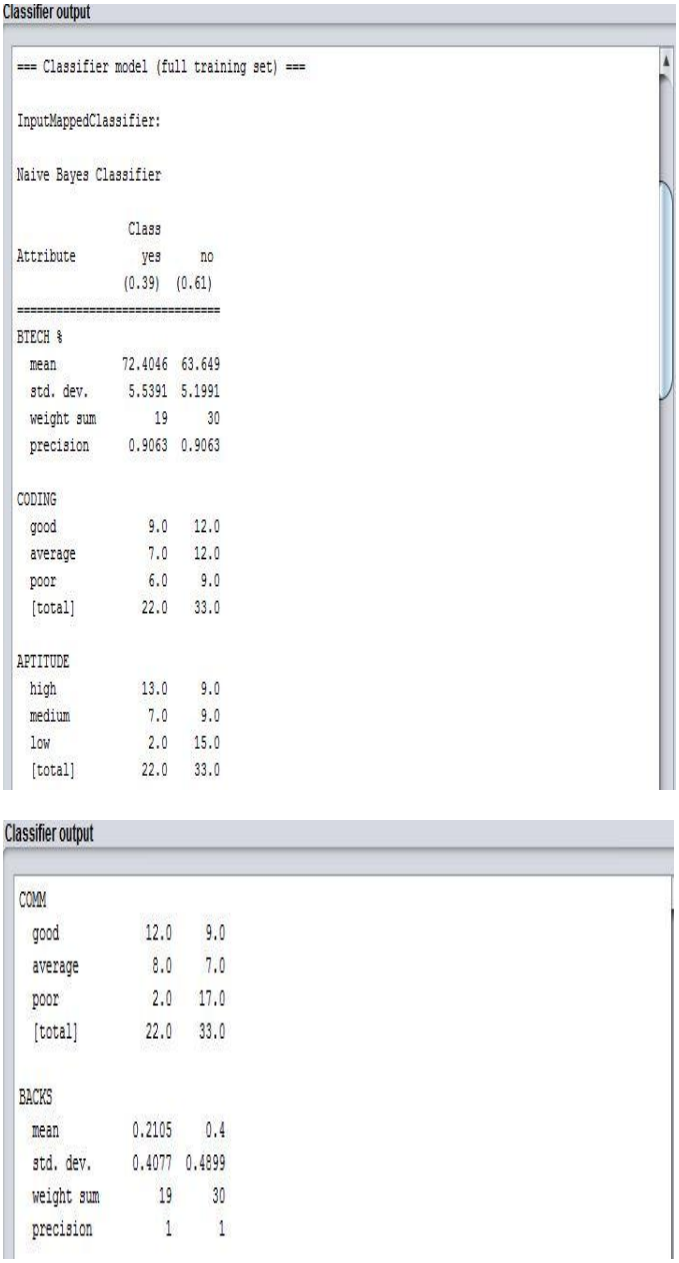

Fig.2 Naïve Bayes Classifier

After building different classifiers using different classification algorithms and some thorough experimentation, we were determined to evaluate the performance and also the usefulness of each of these models in predicting the opportunity of placement of students in any reputable organization in order to boost their career. The experimental results are given in table 2.

Table 2: Performance of Classifiers

Evaluation criteria	**Classifiers**		
	J48	**KNN (IBk)**	**Naïve Bayes**
Total no. of instances	21	21	21
Correctly Classified Instances	15	18	17
Incorrectly Classified Instances	6	3	4
Accuracy (%)	71.43%	85.71%	80.95%

It can thus be seen from the above table, that from the total number of instances in our testing data, i.e. 21, the number of correctly classified instances is found to be 15, 18 and 17 in the case of J48, KNN, and Naïve Bayes classifiers respectively. Accuracy of any model is basically the percentage of correctly classified instances by it. Therefore we have observed K-Nearest Neighbour (IBk) classifier to have the highest accuracy of 85.71% followed by the other two classifiers Naïve Bayes and J48, having the accuracy of 80.95% and 71.43% respectively.

Confusion matrices can also prove to be very effective in evaluating the various classifiers. The columns of the matrices represent the predictions, and the rows represent their class. The confusion matrices for the above mentioned classification models are shown in the table 3.

Table 3: Confusion Matrices

Classifier	Yes	No	Class
J48	7	2	Yes
	4	8	No
KNN	8	1	Yes
	2	10	No
Naïve Bayes	8	1	Yes
	3	9	No

Some of the error measure statistics have been shown below for all the three classification models. These error measures basically tell us how different our estimated values are from the true value in different ways.

Table 4: Training & Simulation Errors

Evaluation measures	Classifier		
	J48	KNN (IBk)	Naïve Bayes
Kappa statistic	0.4324	0.7123	0.6216
Mean absolute error	0.3438	0.1555	0.2175
Root mean squared error	0.4646	0.3709	0.3947
Relative absolute error	70.9548%	32.0867%	44.8798%
Root relative squared error	93.6329%	74.7553%	79.5527%

V. CONCLUSION AND FUTURE SCOPE

The placement session that takes place in the final year is the most crucial point for almost every graduate or post-graduate student in their academic paths. It not only gives them a decent job opportunity but also uplifts their overall morale in an exceptional manner. It is the prime responsibility of any educational institution to ensure that their institute guarantees maximum placement of all their students in well-known companies where students can showcase their potential skills effectively. This kind of placement prediction system can help these institutions in analysing their student's performance in placements and also analyse in what area of the student's individual placement skills they can contribute, for their improvement. They can hence make effectual strategies and decisions for increasing the institution's overall placement status. The classification methods that we have used for classifying students if they are placed or not, are Decision trees(J48), K-Nearest Neighbour(IBk) and Naïve Bayes. According to the results obtained by our system, accuracy shown by KNN is around 85% which makes it the most accurate method for classification compared to J48 and Naïve Bayes.

Another set of machine learning algorithms can also be implemented to analyse prediction chances or even ensemble of classifier models can be built to predict placement of students effectively as a future scope.

REFERENCES

[1] H. A. Madni, Z. Anwar and M. A. Shah, "Data mining techniques and applications — A decade review," *2017 23rd International Conference on Automation and Computing (ICAC)*, Huddersfield, pp.1-7, 2017.

[2] A. S. Sharma, S. Prince, S. Kapoor and K. Kumar, "PPS — Placement prediction system using logistic regression," *2014 IEEE International Conference on MOOC, Innovation and Technology in Education (MITE)*,Patiala,pp.337-341, 2014.

[3] M. Gera and S. Goel, "A model for predicting the eligibility for placement of students using data mining technique," *International Conference on Computing, Communication & Automation*, Noida, pp.114-117, 2015.

[4] K. Pruthi and P. Bhatia, "Application of Data Mining in predicting placement of students," *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*,Noida,pp.528-533,2015.

[5] P. Guleria and M. Sood, "Predicting student placements using Bayesian classification," *2015 Third International Conference on Image Information Processing (ICIIP)*, Waknaghat,pp.109-112,2015.

[6] Ashok M V and Apoorva A, "Data mining approach for predicting student and institution's placement percentage," *2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, Bangalore, pp. 336-340, 2016.

[7] A. Giri, M. V. V. Bhagavath, B. Pruthvi and N. Dubey, "A Placement Prediction System using k-nearest neighbors classifier," *2016 Second International Conference on Cognitive Computing and Information Processing (CCIP)*, Mysore, pp.1-4, 2016.

[8] S. K. Thangavel, P. D. Bkaratki and A. Sankar, "Student placement analyzer: A recommendation system using machine learning," *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*,Coimbatore,pp.1-5,2017.

[9] Pal, Saurabh, "Analysis and Mining of Educational Data for Predicting the Performance of Students", International Journal of Electronics Communication and Computer Engineering, 4, pp.1560-1565, 2013.

[10] Pal, Saurabh, "Classification Model of Prediction for Placement of Students", I.J.Modern Education and Computer Science, 2, pp. 49-56, 2013.

