Placement Chance Prediction using Classifiers

Guru Murthy. N
Department of MCA
Global Institute of Management Sciences
Bangalore

***Abstract*— Data Mining in education is an area where in a combination of techniques such as Data mining, Machine Learning and Statistics, is applied on educational data to get valuable information. The purpose of this paper is to help the prospective pharmacist students in providing a right post graduate course viz., Pharmacognosy, Pharmaceutical chemistry, Pharmaceutical Analysis etc., based on the UG course percentage for admission to PG course. Three classification algorithms viz., Decision tree, Neural network and Naïve Bayes algorithms are applied. Algorithms are compared and was found that Naive Bayes algorithm predicts well in terms of precision, accuracy and true positive rate. This paper will help the students in selecting a best course suitable for them which provides best placement chance.**

Keywords: Educational Data mining, Naive Bayes, Neural network, Decision Tree, Prediction and models.

I. INTRODUCTION

Data mining consists of group of techniques to mine the data, such as association rule mining, classification and clustering. In this model, an algorithm is selected from clustering and two from classification models. Pharmacy is the science and technique of preparing and dispensing drugs. It is a health profession that links health sciences with chemical sciences and aims to ensure the safe and effective use of pharmaceutical drugs. Therefore there is a lot of demand for specialization. To excel in the field of pharmacy there is need to select a good specialization in post-graduation. Decision in this regard is arrived by accessing previous year's admission records of pharmaceutical Institute and manually going through the database. The objective of doing this is to predict the future choice of the course. So huge data needs to be processed and patterns need to be compared manually, which is tedious and cumbersome. Data was obtained from pharmaceutical Institute in excel format from 2011 to 2015.Data in the excel format were fed to MYSQL in the form of queries and two databases were constructed, One containing historic data from 2011 to 2014 and another test data i.e., 2015.

II. PROBLEM STATEMENT

Every student dreams to be successful in life. For him to be successful, choosing the right courses while studying is important. Hence a classifier model is proposed which helps the students to choose a course based on type of data or information that he/she furnishes Here student will enter Percentage, Gender, Category and Sector. Among the fields or attributes that he/she enters, the result would be displayed in terms of Excellent [E], Good [G], Average [A] and Poor [P] for the data entered. Each and every course offered is associated with one of the above answers viz., E, G, A, P Such as, Pharmaceutical chemistry with – E, pharmacology with– P and so on..Various mining algorithms from different models are applied on the processed data and tested accordingly. Algorithms are compared based on certain criteria such as accuracy, precision and true positive rate.

III. RELATED WORKS

Many scientists have been working to explore the best mining techniques for solving placement chance prediction problems. Various works have been done in this regard. Few of the similar works are listed below:

Krishna K, Murty M N [1] Propose a novel hybrid genetic algorithm (GA) viz., genetic K- means algorithm that finds worldwide optimal partition of a given data into a specified number of cluster. It is also observed that GKA search quicker than some of the other evolutionary algorithms used for clustering. Zhexue Huang[2] focuses on the practical issues of extending the k-means algorithm to cluster data with categorical value. Outstanding property of k-means algorithm in data mining is its efficiency in clustering large data sets. However, it only works on numeric data limits its use in many data mining applications because of the involvement of categorical data. Leon Bottou, YoshuaBengio[3] Studies the convergence properties of the well-known K- means clustering algorithm. It minimizes the quantization error using the very fast Newton algorithm.Kai mingting, zijianzheng[4] introduce tree structures into naive Bayesian classification to improve the performance of boosting when working with naive Bayesian classification. Yong wang, Hodges.J, Botang[5] focuses upon three aspects of this approach: different event models for the naive Bayes method, disparate chance of smoothing method, and dissimilar feature assortment methods. In the above research paper, we describe the performance of each method in terms of recall, precision, and F-measures.Yongchuan Tang, Yang Xu [6] presents a method to detecting a fuzzy model from data by means of the fuzzy Naive Bayes and a real-valued genetic algorithm. The detection of a fuzzy model is comprised of the mining of "if–then" rules that is followed by the estimation of their parameters. Sreerama K. Murthy [7] Survey existing work on decision tree construction, attempting to recognize the important issues implicated, directions the work has taken and the present state of the art. Elizabeth Murray [8] Studies have been conducted in similar area such as understanding student data. There they apply and evaluate a decision tree algorithm to university records, producing graphs that are useful both for predicting graduation, and verdict factors that lead to graduation. It's always been an active discussion over

which engineering branch is in demand .So this work gives a scientific solution to answer these.Safavian S.R, Landgrebe D [9] presents a survey of current methods for DTC designs and the various existing issues. Past considering potential advantages of DTC's over single stage classifiers, the subjects of tree structure design, characteristic selection at each inner node, decision and search strategy are discussed. Some remarks concerning the relation between decision trees and Neural Networks (NN) are also made. John Mingers [10] the method involve-three main stages—creating a complete tree able to classify all the examples, considering this tree to give statistical reliability, processing the considered tree to develop understandability. This paper is concerned with the initial stage — tree creations which depends on a measure for goodness of split, that is, how well the attributes distinguish between classes. Some problems encountered at this stage are lost data and multi-valued attribute..SudheepElayidom, Suman Mary Idikkula& Joseph Alexander [11] proved that the technology named data mining can be very effectively applied to the domain called employment prediction, which help students to select a good branch that may fetch them placement. A global framework for similar troubles has been proposed. Ajay Kumar Pal, Saurabh Pal [12] presents a proposed model based on classification approach to find an enhanced evaluation method for predicting the placement for students. This replica of a model can determine the associations between academic achievement of students and their placement in campus selection. A K Pal, and S Pal [13] frequently used classifiers are studied and the experiments are conducted to find the best classifier for predicting the student's performance. B K Bharadwaj , S Pal [14] Provides work to identify those students which needed special attention to reduce fail ration and taking appropriate action for the next semester examination. S. K. Yadav, B K Bharadwaj and S Pal [15] Focuses on identifying the dropouts and students who need special attention and allow the teacher to provide appropriate advising/counselling.

IV. DATA DESCRIPTION

Name *– name of the student. It takes only the alphabetical values from A to Z.*

Category *– it is the category of the student that he /she belonging to. It takes string values. The possible values that it can take are 2A, 3A, 2B, 3B, SC/ST and GM.*

Age *– it is the age of the student and it takes only numeric values from 0 to 9.*

Sector *– represents the sector that the student belongs and the possible values that it can take are URBAN and RURAL.*

Percentage *–percentage that a student gets in b-pharma exam and can take values from 0 to 9.*

Address *– it is the address of the student. It can take alphanumeric values from A to Z, 0 to 9.*

Ph.no *– it is the contact number of the student and it takes numerical values from 0 to 9.*

Gender *– it is the gender of the student and the possible values are male, female.*

Specialization *– it is the specialization that the student choses and the possible values are Textile Design, Fashion Design, etc.*

III. METHODOLOGY

A. DATA PREPROCESSING

Number of attributes that were found to be contributing to the result, after applying the chi-square test is as follows.

TABLE I. MAPPING INPUT VALUES TO NUMERIC VALUES.

Category	Input Values	Numeric values
Gender	Female, Male	0 and1
Category	2A,2B,3A,3B OBC,GM, SC,ST	0 and 1
Percentage	1 to 100	0 and 1
Sector	Rural, Urban	0 and 1
Specialization	A to F	0 and 1
Chances	E, G, A, P	0 and 1

a) Percentage: obtained by student in UG entrance examination Range: (0 to 100%)

b) Category: social background Range (2A, 2B, 3A, 3B, GM, SC, ST, OBC).

c) Gender: Range (Male, female).

d) Sector: Range (Urban, Rural).

e) Specialization: Range (A to F).

f) All the input values would be mapped between 0 and 1 as given in the table above.

IV. DATA MINING ALGORITHMS APPLIED

A. NEURAL NETWORKS:

Neural networks or also known as artificial neural networks are computational models.

Category	Input Values	Numeric values
Gender	Female, Male	0 and1
Category	2A,2B,3A,3B OBC,GM, SC,ST	0 and 1
Percentage	1 to 100	0 and 1
Sector	Rural, Urban	0 and 1
Branch	A to N	0 and 1
Chances	E, G, A, P	0 and 1

This paper includes two steps for the process:
Step1: Classification (learning)
Step2: Possibility (Outputs)

Step1: Under the step1 calculation of weight been done. The neural networks accept all the inputs as a weight so that the conversion of characters should be done.
For example if the class of records has gender = male /female conversion of male will be 0 and female will be 1.
Category = Rural/Urban, here the conversion of the rural will 0 and urban will be 1 and the rank mapped between 0 and 1. Based upon table4 the value conversion will be done.

TABLE II. USER INPUT TABLE (AFTER CONVERSION)

Percentage	Category	Sector	Gender
0.90	0	1	0

Step 2: Based up on the input data the flow will get start.

Gender	Sector	Percentage	Branch	Chance
0	0	0.5	0.1	1

First step is loop through data which been tested and stored as a knowledge by this first step (A) the possible neuron and this networks will be created, it gives the output of possibilities for the input.

TABLE III. POSSIBILITIES FOR GIVEN INPUT

Generating the actual output from the table 4.5.10

TABLE IV. OUTPUT BEFORE CONVERSION

GENDER	SECTOR	Percentage	BRANCH	CHANCE
0	0	0.5	0.1	1
0	1	0.5	0.2	1
1	1	0.5	0.2	1

TABLE V. OUTPUT AFTER CONVERSION

Gender	Sector	Percentage	Branch	Chance
MALE	RURAL	2	Civil Law	E

B. NAIVE BAYES:

A Naive Bayes classifier is a probabilistic classifier that works based on the Bayes theorem.
The procedure to be followed while applying this method is as follows

- Data preprocessing
- Finding positive and negative knowledge data
- Application of Bayes theorem

Step 1: Data preprocessing: Filling of the missing values and the dependency check on the attributes listed in the table 8 is performed using chi-square test and Table 4.2.2 is a resultant after preprocessing.

TABLE VI. INPUT FOR NAÏVE BAYES

Name	Age	Gender	sector	Category	Percentage	Specialization
Shiva	21	M	Rural	2a	52	Pharmacology
John	22	M	Urban	3b	90	Pharmaceutical Analysis
Rani	22	F	Rural	SC	95	Pharmaceutical chemistry

Step 2: Finding positive and negative knowledge data: selection constructs are applied on a percentage attribute to get a positive and negative knowledge data.

If (percentage <= 100)//the maximum limit of the possible Percentage
{Positive knowledge data}
Else
{Negative knowledge data}

The above process is repeated for all the attributes listed in table 9 to get the positive knowledge data as given below.

Step 3: Application of Bayes theorem on table 10 gives the resultant output table. At the first instance data in table 10 is converted to the numeric data.

TABLE VII. AFTER PREPROCESSING

Gender	Sector	Category	Percentage	Specialization
M	Rural	2a	52	Pharmaceutical chemistry
M	Urban	3b	90	Pharmaceutical Analysis
F	Rural	SC	95	Pharmaceutical chemistry
M	Rural	3a	80	Pharmaceutical Analysis

TABLE VIII. POSITIVE KNOWLEDGE DATA

Name	Age	Gender	sector	Category	Percentage	Specialization
Shiva	21	M	Rural	2a	78	Pharmaceutical Analysis
John	22	M	Urban	3b	84	Pharmaceutical chemistry
Rani	22	F	Rural	SC	64	Pharmaceutical Analysis

Formulae listed under are used to get the below output table as the resultant.

$$h_{MAP} = \arg\max_{heH} P(h/D)$$

$$= \arg\max_{heH} \frac{P(D|h)P(h)}{P(D)}$$

$$= \arg\max_{he} P(D|h)P(h)$$

Where,
P(h) ≡ Prior Probability of (Correctness of) Hypothesis h
P(h | D) ≡ Probability of h Given Training Data D
P(D) ≡ Probability of D
P(D | h) ≡ Probability of D Given h

TABLE IX. OUTPUT TABLE

Percentage	Gender	Sector	Category	Specialization	Chance
1-60	M	Rural	Any	Pharmaceutical Analysis	E
1-80	M	urban	Any	Pharmaceutical chemistry	E
1-80	F	Rural	Any	Pharmacology	E

C. Decision tree:

Decision tree is the classification method which makes use of top-down tree construction approach, which results in a tree like structure where, each node represents an attribute to be tested and the branch will be the outcome of the test on an attribute. The objective of this algorithm is to generate

a. Knowledge database.
b. Output based on knowledge database for the user input.

TABLE X. INPUT FOR THE ALGORITHM

Name	Category	Age	Sector	Percentage	Address	Ph.No	Gender	Specialization
Ravi	2A	30	Urban	85	Jayanagar	9812346754	Male	Pharmaceutical chemistry
Raj	3B	38	Urban	90	Nagarabhavi	9440213456	Male	Pharmaceutical chemistry
Rani	2A	21	Rural	70	Banikuppe	8050214356	Female	Pharmaceutical Analysis

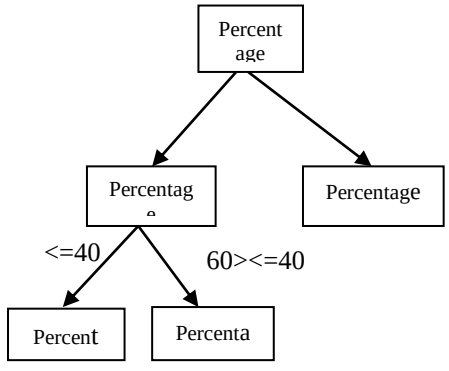
Step1: Priorities are set for the attributes based on the dataset. For our dataset Percentage is taken as the attribute with the top most priority and then sector, category so on.

Step2: based on the conditions set, the prioritized attribute i.e., percentage will be divided basically into two, one with positive values and another with negative values.

Step3: If the Tree contains all the nodes that are positive then create results as yes and then exit the step.

If the tree contains all the nodes that are negative then create result as no and exit the loop.

The partial view of a tree after the application of step 3 for Rank attributes



STEP4:

If step3 fails; expand the tree by selecting the next attribute (F) sector or gender.

STEP5:

Repeat step4 until all the nodes are visited at least once.

Output: The table below represents the knowledge database obtained after the application of the decision tree algorithm on the resultant of the table 4.1.1, after preprocessing.

TABLE XI. REPRESENTATION OF THE KNOWLEDGE DATABASE

Percentage	Sector	Gender	Category	Specialization
100<==>60	Rural	Female	2A	Pharmaceutical chemistry
60<==>40	Urban	Male	2A, 3B	Pharmaceutical Analysis, Pharmacognosy

For the following user input (Percentage, Gender, sector etc…) the table 5 represents the possibilities of choice of specialization as the final output, after processing the knowledge data.

TABLE XII. OUTPUT TABLE FOR THE USER INPUT

Id	Specialization	Chance	Possibilities
1	Pharmaceutical chemistry	E	90%
2	Pharmaceutical Analysis	E	85%
3	Pharmacognosy	E	70%

TESTING

Results obtained after the tests for each algorithm were modeled as confusion matrix. Confusion matrix explains the performance of three algorithms expressed in terms of True Positive rate, Accuracy and Precision.

TABLE XIII. CONFUSION MATRIX TABLE

Algorithms	TPR	Accuracy	Precision
Naïve Bayes	0.83	83%	0.83
Neural network	0.80	77%	0.75
Decision tree	0.81	81%	0.81

From the above table 13 it is clear that the Naïve Bayes algorithm is more accurate with 83% compared to the other algorithms viz., Decision Tree (81%) and Neural Network (77%).Naive Bayes algorithm leads with respect to true positive rate (TPR) with 0.83 correct instances and Precision (0.83).Thus Naïve Bayes predicts the results better than the other algorithms used.

CONCLUSION

Applying data mining techniques on educational data is concerned with developing methods for exploring the unique types of data; In this study, Three classification algorithms viz., Naïve Bayes, Neural Network and decision Tree were applied. Among these algorithms, Naïve Bayes proved to be the best predicting algorithm , for solving placement chance prediction problems. Hence, having the information generated through our study, student would be able to select the appropriate specialization with best chances of getting placed. Furthermore, the work can be extended to solve problems on predictions, using different approaches on data of different disciplines.

BIBLIOGRAPHY

[1] Krishna.k, Murty M.N "Genetic k-means algorithm", volume 29, issue 3, 1999 pages 435-439.

[2] Zhexue Huang "Extensions to the k-means algorithm for clustering large data sets with categorical values", volume 2, issue 3, pages 283-304, 1998.

[3] Leon Bottou, YoshuaBengio "Convergence properties of the k-means algorithms", 1995.

[4] Kai mingting,zijianzheng "Improving the performance of boosting for Naïve Bayesian classification", volume 1574,1999, pages 296-305.

[5] Yong wang, Hodges.J,Botang "Classification of web documents using a naïve Bayes method" ,2003,pages 560-564, Germany, 2005.

[6] Yongchuan Tang, Yang Xu "Application of fuzzy Naïve Bayes and a rel-valued gentic algorithm in identification of fuzz model", volume 169, issue 3-4, 2005, pages 205-226.

[7] Sreerama K. Murthy, Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey, Data Mining and Knowledge Discovery, 345-389 1998.

[8] Elizabeth Murray, Using Decision Trees to Understand Student Data, Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 2005.

[9] Safavian, S.R. , Landgrebe, D "A survey of decision tree classifier methodology", Volume 21, Issue 3,pages 660 – 674.

[10] John Mingers , An empirical comparison of selection measures for decision-tree induction ,volume 3 , issue 4 , pp 319-342, march 1989.

[11] Quinaln, J.R., C4.5: Programs for machine learning, Morgan Kaufmann, San Francisco, 1993.

[12] Wu, X. & Kumar, V., the Top Ten Algorithms in Data Mining, Chapman and Hall, Boca Raton. 2009.

[13] SudheepElayidom, Suman Mary Idikkula& Joseph Alexander "A Generalized Data mining Framework for Placement Chance Prediction Problems" International Journal of Computer Application (0975-8887) Volume 31- No.3, October 2011.

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

[15] A. K. Pal, and S. Pal, "Analysis and Mining of Educational Data for Predicting the Performance of Students", (IJECCE) International Journal of Electronics Communication and Computer Engineering, Vol. 4, Issue 5, pp. 1560-1565, ISSN: 2278-4209, 2013.

[16] B.K. Bharadwaj and S. Pal. "Mining Educational Data to Analyze Students' Performance", International Journal of Advance Computer Science and Applications (IJACSA), Vol. 2, No. 6, pp. 63-69, 2011.

[17] S. K. Yadav, B.K. Bharadwaj and S. Pal, "Data Mining Applications: A comparative study for Predicting Student's Performance", International Journal of Innovative Technology and Creative Engineering (IJITCE), Vol. 1, No. 12, pp. 13-19, 2011.