DOI:10.26524/jms.2013.48

Preference and performance analysis of community college students using Bayesian classification

S.Balaji PhD
Asst.Professor
D.B.Jain College(Autonomous)
Chennai-97

J.Senthil Kumar PhD
Asst.Professor
D.B.Jain College(Autonomous)
Chennai-97

S.K.Srivatsa PhD
Senior Professor
St.Jospeh's College
Chennai

ABSTACT: *The Community College system in India aims at the empowerment of the disadvantaged persons through appropriate skills development leading to gainful employment..The process of predicting students preference towards the courses and analyzing performance is the biggest challenge for the community college institutions.In addition,community college education is booming in India.There is a need for analyzing the preference of students to provide courses of the need. Data Mining can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students.The investigation considered the bayes theorm based posteriori probabilities of dependent attributes.The objectives of the present investigation were framed so as to assist the enrollment process in vocational education of community colleges and they are:*

(a) Generation of a data source of predictive variables,

(b) Identification of different factors, which effects a student"s preference and performance and

(c) Construction of a prediction model using classification data mining techniques on the basis of identified predictive variables

Keywords: *Data Mining;Bayes theorm;Predictive analysis.*

1.INTRODUCTION

Community college system gives a second chance to those who have dropped out of school.It provides opportunity for those who otherwise would have been excluded[12]. In information systems, classification is one of the key parts for data mining. Such analysis can provide us with a better understanding of the important data classes and predict current and future data trends. The ability to predict/classify a student's performance is very important in community college.The community college having vocational education programmes aimed at three target groups-tenth failed (Primary),tenth passed/Twelth failed(secondary) and twelth passed(higher secondary).The courses for this group can be considered as computer courses,Technician courses and Hand_made_training courses.The computer courses taught are Diploma in DTP Operator (DDTP),Diploma in Animation,Diploma in Computer Applications,Diploma in Multimedia System and Diploma in Computer Hardware Servicing (DCHS).The programmes under Technician category are Diploma in Refrigeration and Air Conditioning Technician (DRAT),Diploma in House Electrician (DHE),Diploma in Plumbing Technician (DPT),Diploma in Four Wheeler Mechanism (DFWM),Diploma in Mobile Phone

Servicing and Diploma in Home Appliances Repair and Servicing.The Hand_made_training courses covers the programmes such as Diploma in Catering Assistant (DCA),Diploma in Fashion Design and Garment Making (DFGM),Diploma in Health Assistant (DHA),Diploma in Early Childhood Care and Education (Kindergarten),
Diploma in Beautician,Diploma in Bakery and Confectionery and Diploma in Food Production.The primary objective of community college system is to provide life,quality education for the needy by removing blockades age bar and minimum qualification.With the changing scenario in the industry and job markets there is a need for identifying the preferences of community college students towards courses and providing the apt course based on demographic and socio-economic variables. In addition,the factors that affect the performance of students in terms of upward and downward trends in results can also be identified to make remidal measures for their upliftment. A very promising arena to attain this objective is the use of Data Mining (DM) [1].

2, RELATED WORK

Data Mining can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students as described by Alaa el-Halees [11]. Mining in educational environment is called Educational Data Mining.
Han and Kamber [10] describes data mining software that allow the users to analyze data from different dimensions, categorize it and summarize the relationships which are identified during the mining process.
In fact, one of the most useful Data mining tasks in education applications is classification. There are different educational objectives for using classification, such as: to discover potential student groups with similar characteristics and reactions to a particular pedagogical strategy [2], to detect students‘ misuse or game-playing [2], to group students who are hint-driven or failure-driven and find common misconceptions that students possess [3], to identify learners with low motivation and find remedial actions to lower drop-out rates [4], to predict/classify students when using intelligent tutoring systems [5], etc. And there are different types of classification methods and artificial intelligent algorithms that have been applied to predict student outcome, marks or scores. predicting student academic success (classes that are successful or not) using discriminant function analysis [6]; classifying students using genetic algorithms to predict their final grade [7]; predicting a student‘s academic success (to classify as low, medium and high risk classes) using different data mining methods [8]; predicting a student‘s marks (pass and fail classes) using regression techniques in Hellenic Open University data [9].
Pandey and Pal conducted study on the student performance based by selecting sixty students from a degree college of Dr. R. M. L. Awadh University, Faizabad, India. By means of association rule they find the interestingness of student in opting class teaching language[13].
Hijazi and Naqvi conducted as study on the student performance by selecting a sample of 300 students (225 males, 75 females) from a group of colleges affiliated to Punjab university of Pakistan. The hypothesis that was stated as "Student's attitude towards attendance in class, hours spent instudy on daily basis after college, students' family income, students' mother's age and mother's education are significantly related with student performance" was framed. By means of simple linear regression analysis, it was found that the factors like mother‘s education and student‘s family income were highly correlated with the student academic

performance[14].Brijesh et al work based on the student database to predict the students division on the basis of previous database using decision tree. Information‟s like Attendance, Class test, Seminar and Assignment marks were collected from the student‟s previous database, to predict the performance at the end of the semester[18].

Galit gave a case study that use students data to analyze their learning behavior to predict the results and to warn students at risk before their final exams[15].

On-line collaborative discussions have significant role in distance education courses. Automatic tools for assessing student activities and promoting collaborative problem solving can offer an improved learning practice for students and also offer useful assistance to teachers. Researchers developed a specific mining tool for making the configuration and execution of data mining techniques easier for instructors and in

order to be of use for decision making, using real data from on-line courses [16][17].

In this paper,we considered the factors that influence the students preference towards the courses of community colleges and their performance in end semester results.The impact of socio-economic,demographic and academic attributes were considered for investigation.

3. NAIVE BAYES METHODOLOGY

Navie Bayes is the basis for many machine learning and data mining methods.In Bayesian classification is a classification method is applicable for huge dataset.Naive Bayesian classifier works with hypothesis H such as that the data tuple X belongs to a specified class C.The determination of P(H/X) that the hypothesis H holds given the evidence or observed data tuple X..P(H/X) is the posterior probability of H conditioned on X.Bayes‟ theorem is useful in that it provides a way of calculating the posterior probability ,P(H/X) ,from P(X/H) and P(X), Bayes theorem is

P(H/X)=P(X/H)P(H)/P(X).

The naive Bayesian classifier, or simple Bayesian classifier, works as follows:

1. Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, X=(x1, x2,…, xn), depicting n measurements made on the tuple from n attributes, respectively, A1, A2,.., An.

2. Suppose that there are m classes, C1, C2,…, Cm. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple x belongs to the class Ci if and only if

P (Ci|X)>P (Cj|X) for $1 \leq j \leq m, j \neq i$

Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes‟ theorem

3. As P(X) is constant for all classes, only P (X|Ci) P (Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, P(C1)=P(C2)

=…=P(Cm), and we would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci)P(Ci). Note that the class prior probabilities may be estimated by P(Ci)=|Ci,D|/|D|**,** where |Ci,D| is the number of training tuples of class Ci in D.

4. Given data sets with many attributes, it would be extremely computationally expensive to compute P(X|Ci).

In order to reduce computation in evaluating P(X|Ci), the naïve assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally

independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes). Thus,

=P(x1|Ci)x P(x2|Ci)x… P(xm|Ci).

We can easily estimate the probabilities P(x1|Ci), P(x2|Ci),… ,P(xm|Ci) from the training tuples. Recall that here xk refers to the value of attribute Ak for tuple X.

For instance, to compute P(X|Ci), we consider the following:

(a) If Ak is categorical, then P(Xk|Ci) is the number of tuples of class Ci in D having the value xk for Ak, divided by |Ci,D|**,** the number of tuples of class Ci in D.

(b) If Ak is continuous valued, then we need to do a bit more work, but the calculation is pretty straightforward. A continuous-valued attribute is typically assumed to have a Gaussian distribution with a mean μ and standard deviation σ, defined by

g(x,,μ,σ)=1/¬2Д σ e-(x- μ)2/2 σ2

So that

P(xk|Ci)=g(xk, μci, σci)

We need to compute μci and σci, which are the mean and standard deviation, of the values of attribute Ak for training tuples of class Ci. We then plug these two quantities into the above equation.

5. In order to predict the class label of X, P(X|Ci)P(Ci) is evaluated for each class Ci. The classifier predicts that the class label of tuple X is the class Ci if and only if

P(X|Ci)P(Ci)>P(X|Cj)P(Cj) for $1 \leq j \leq m, j \neq i$

In other words, the predicted class label is the class Ci for which P(X|Ci)P(Ci) is the maximum.Bayes‘ theorem assumes that all attributes are independent and that the traning sample is a good sample to estimate probabilities.

4. EXPERIMENT AND ITS RESULTS

4.1 Data Preparation

Our dataset considers the students data from Fourteen ommunity colleges in Chennai and Kanchipuram districts of seven each.The dataset having students strength of Three hundred admitted during calendar year admissions of 2011-2012.

4.2 Data selection and Transformation

Data cleaning processes enable us to fill in missing values,reduce noise while recognizing outliers, and accurate inconsistencies in the data (Han J. and Kamber M. 2008).It is known that too many attributes involved will very possibly result in discovered information that is difficult to interpret, or even meaningless. The first task is to remove from the dataset those fields/ attributes/ variables which were irrelevant to the task of analysis.The attributes such as community college code,caste,attendance and date of birth were not considered.Therefore, by in-depth discussion with domain managers, we eliminated some of the attributes and finally came to a conclusion of 9 attributes, namely 1)Place of residence 2)Gender 3) Marital_status 4)Qualification5)Life_generation 6)Employment_status 7)Course_opted 8)Family_financial_status&9)Result_division.The data transformation task is data generalization,where low-level or primitive data are substituted by higher-level concepts during the use of concept hierarchies (Han J. and Kamber M. 2008). In the working set,eighteen

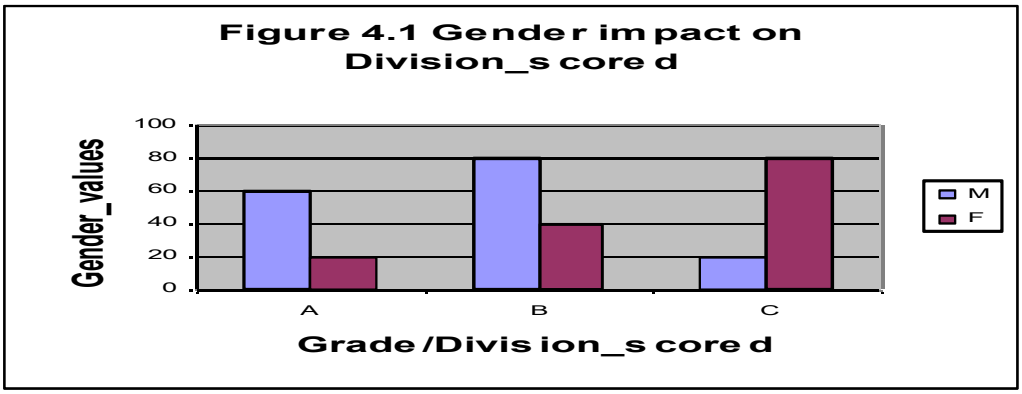
vocational courses of community college stream fall in three categories. The Eighteen different vocational courses taught under community college is being brought under three different classification such as Computer courses,Hands_on_skill and Technician courses.Students academic performance and their preferences towards the courses is analysed with demographic,academic and socio-economic variables.

Given a training set the naïve Bayes algorithm first estimates the prior probability P (Cj) for each class by counting how often each class occurs in the training data. For ach attribute value xi can be counted to determine P (xi). Similarly the probability P (xi | Cj) can be estimated by counting how often each value occurs in the class in the training data. When classifying a target tuple, the conditional and prior probabilities generated from the training set are used to make the prediction. The present investigation used data mining as a tool with naïve Bayes classification algorithm as a technique to design the student performance and preference prediction model. Filtered feature selection technique was used to select the best subset of variables on the basis of the values of probabilities.
Our Bayesian approach based model makes prediction analysis with maximum posteriori hypothesis.Posteriori probabilities of independent variables for the dependent class of result_division attribute is given table 4.1.

Table 4.1 Probabilities of Independent attributes on dependent attribute result_division

Slno	Attribute	Probability
1	Gender	0.37766
2	Marital_status	0.29638
3	Finacial_status	0.25378
4	Course_opted	0.19186
5	Life_generation	0.13702
6	Qualification	0.12216
7.	Employment_status	0.05401
8.	Place_of_residence	0.00369

In order to predict students performance in end_semester is observed with attributes whose probability is greater that 0.30.In the above study,the attributes Gender and Marital_status are the highly influencial variables to predict the dependent attribute result _division score of the students of the community college.

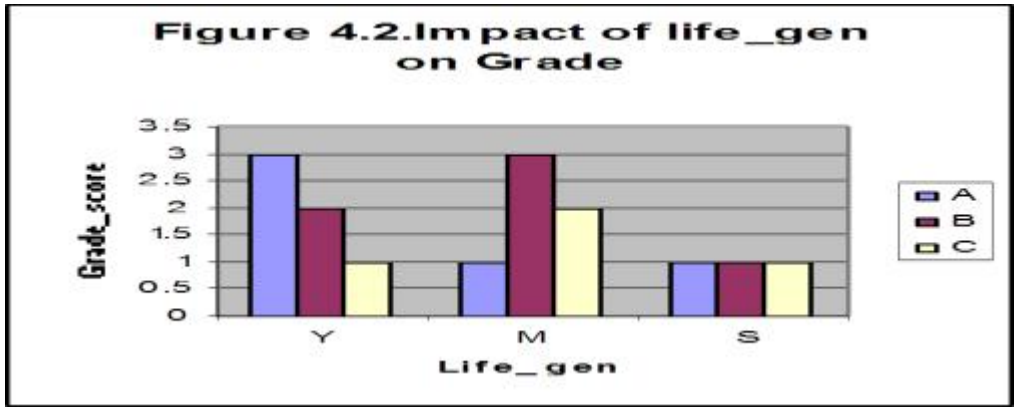


The relationship between the highly influencial attribute gender and dependent attribute divison_obtained is shown in Figure 4.1.From the figure it is evident that female students highly dominated in C-grade where as male student groups equally dominated in A and B-grades.
The preference towards the dependent attribute course is decided by independent attributes and their corresponding posteriori probabilities are given in table 4.2.

Table 4.2 Probabilities of Independent attributes on dependent attribute course_opted

Slno	Attribute	Probability
1	LIFE_GEN	0.2749
2	M_STATUS	0.2506
3	DISTRICT	0.1919
4	DIV	0.1919
5	F-INCOME	0.1183
6	QUALI	0.094
7	EMP	0.0258
8	GENDER	0.0258

From table 4.2,It is evident that the attributes Life_generation and M_status are the highly influencial variables to predict the course_opted by the student.



The figure 4.2 shows the effect of independent attribute Life_generation which has highest probability dependent attribute grade score.Youth ones are the most predominate score in grade A .Grade B is scored by Middle age groups followed by senior age ones,where as senior age

groups normally score symmetric performances in all divisions.From the study of this paper,it is observed that the preference of community college students towards the course are not always depending on their own effort,the impact of life generation has the impact on it.In addition,the present study outlines the dependentability of gender and Marital_status and its impact on academic performances of the students. The life_generation of candidate has an major impact of course preferences of stundents followed by marital status of
the candidate.This proposal will improve the insights over existing methods.

5. CONCLUSION

The interesting possibilities for the education domain can be achieved with the help of highlevel extraction of knowledge from raw data using datamining techniques.This study was based on very specific to vocational courses of community colleges.The students preference towards the vocation courses was observed with Bayesian approach based on demographic and socio-economic variables. The preliminary results presented in this paper give us a first indication on how the information of community college students can be used to enhance better enroll and improve the performance of students as a result of data mining technique.The behaviour of stakeholders can be predicted with a greater degree of accuracy with the models that predict hidden patterns in large databases of community colleges and universities .These predictive models will help educational institutions like community colleges to address issues of students enrollement,performance and better reach of education to the needy.

6. REFERENCES

[1].Romero, C., Ventura, S. Educational Data Mining: a Survey from 1995 to 2005. Expert Systems with Applications, 2007, 33(1), pp.135-146

[2] Baker, R., Corbett, A., Koedinger, K. Detecting Student Misuse of Intelligent Tutoring Systems. Intelligent Tutoring Systems. Alagoas, 2004. pp.531–540.

[3].Yudelson, M.V., Medvedeva, O., Legowski, E., Castine, M., Jukic, D., Rebecca, C. Mining Student Learning Data to Develop High Level Pedagogic Strategy in a Medical ITS. AAAI Workshop on Educational Data Mining, 2006. pp.1-8.

[4]Yudelson, M.V., Medvedeva, O., Legowski, E., Castine, M., Jukic, D., Rebecca, C. Mining Student Learning Data to Develop High Level Pedagogic Strategy in a Medical ITS. AAAI Workshop on Educational Data Mining, 2006. pp.1-8.

[5].Hämäläinen, W., Vinni, M. Comparison of machine learning methods for intelligent tutoring systems. Conference Intelligent Tutoring Systems, Taiwan, 2006. pp. 525–534.

[6].Martínez, D. Predicting Student Outcomes Using Discriminant Function Analysis. Annual Meeting of the Research and Planning Group. California, 2001. pp.163-173.

[7].Minaei-Bidgoli, B., Punch, W. Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System. Genetic and Evolutionary Computation, Part II. 2013. pp.2252–2263.

[8].Superby, J.F., Vandamme, J.P., Meskens, N. Determination of Factors Influencing the Achievement of the First-year University Students using Data Mining Methods.Workshop on Educational Data Mining, 2011. pp.37-44.

[9].Kotsiantis, S.B., Pintelas, P.E. Predicting Students Marks in Hellenic OpenUniversity. Conference on Advanced Learning Technologies. IEEE, 2005. pp.664- 668.

[10].U. Fayadd, Piatesky, G. Shapiro, and P. Smyth, From data mining to knowledge discovery in databases, AAAI Press / The MIT Press, Massachusetts Institute Of Technology. ISBN 0–262 56097–6, 1996.
[11] J. Han and M. Kamber, ―Data Mining: Concepts and Techniques,‖ Morgan Kaufmann, 2000.
[12].Fr.Xavier Alphone.S.J,Journal of Christan Manager,2008,pp.20-25.
[13.] Pandey, U. K. and Pal, S., ―A Data Mining View on Class Room Teaching Language‖, (IJCSI) International Journal of Computer Science Issue, Vol. 8, Issue 2, March -2011, 277-282, ISSN:1694-0814. [14.] Hijazi, S. T., and Naqvi, R.S.M.M., ―Factors Affecting Student‘s Performance: A Case of Private Colleges‖, Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006.
[15]. Galit.et.al, ―Examining online learning processes based on log files analysis: a case study‖. Research, Reflection and Innovations in Integrating ICT in Education 2007.
[16]. Romero C., Ventura S., Espejo P. and Hervas C., Data Mining Algorithms to Classify Students, Proceedings of Educational Data Mining , The 1st International Conference on Educational Data Mining Montreal, Quebec, Canada, June 20-21, 2008 pp. 8-17.
[17]. Vasile Paul Bresfelean, Data Mining Applications in Higher Education and Academic Intelligence Management, Theory and Novel Applications of Machine Learning,2009,pp.209- 228.
[18]. Brijesh Kumar Baradwaj, Saurabh Pal, Mining Educational Data to Analyze Students‟ Performance, International Journal of Advanced Computer Science and Applications,2011,vol.2.No.6,pp.63-69.

*****