A SURVEY ON STUDENT'S ABSENTEEISM AT UNDER GRADUATE LEVEL USING NAÏVE BAYES ALGORITHM FOR CATEGORICAL DATASET

[1]Dr. N. Venkatesan, [2]S. Muthukumaran & [3]K. Arunmozhi Arasan

[1]Associate Professor, [2]Assistant Professor &[3]HOD and Assistant Professor,
[1]Department of IT, [2]Department of Computer Science and Applications,
[3]Department of Computer Science and Applications,
[1]Bharathiyar College of Eng&Tech, Karaikal.
[2,3]Siga College of Management and Computer Science, Villupuram.

ABSTRACT-*Educational data mining is used to study the data available in the educational field and bring out the hidden knowledge from it. Classification methods like can be applied on the educational data for predicting the students behavior. This paper focus on the reason for the leave taken by the student in an academic year. The first step of the study is to gather student's data by using a questionnaire. We collect data from 123 students who were under graduate from a private college which is situated in a semi-rural area. The second step is to clean the data which is appropriate for mining purpose and choose the relevant attributes the classification is done using the gender attribute. This paper presents the use of Naïve Bayes Algorithm in predicting the reason for student's absenteeism. The efficiency of Naïve Bayes algorithm on classifying a given dataset was measured using the confusion matrix obtained in tanagra. This knowledge is used to identify the reason for the leave taken by the student and help to improve the quality of the environment and also to improve the performance of the student.*

Keywords:**Educational Data Mining, Naïve Bayes Algorithm**

I. INTRODUCTION

Currently many educational institutions especially small-medium education institutions are facing problems with the lack of attendance among the students. The universities will allow the students who have attendance above than 80% to the semester exam, if a student who have attendance percentage below 80% will lack attendance and are not permitted to write the semester exam[1]. All educational institutions are facing this problem so this research aims to find the reason for a student to put leave to the college and take immediate actions to overcome this problem.

II. LITERATURE SURVEY

V. Muralidharan, V. Sugumaran[2] presented a paper and in it a vibration based condition monitoring system is presented for monoblock centrifugal pumps as it plays relatively critical role in most of the industries. This paper presents the use of Naïve Bayes algorithm and Bayes net algorithm for fault diagnosis through discrete wavelet features extracted from vibration signals of good and faulty conditions of the components of centrifugal pump. LeventKoc, Thomas A. Mazzuchi, ShahramSarkani[3] presented a paper and in it the Hidden Naïve Bayes (HNB) model can be applied to intrusion detection problems that suffer from dimensionality, highly correlated features and high network data stream volumes. HNB is a data mining model that relaxes the Naïve Bayes method's conditional independence assumption. Their experimental results show that the HNB model exhibits a superior overall performance in terms of accuracy, error rate and misclassification cost compared with

the traditional Naïve Bayes model. Alfonso Ibáñez n, ConchaBielza,PedroLarrañaga[4]weproposetwogreedywrapper forwardcost-sensitiveselective naïve Bayesapproaches.Bothapproachesreadjusttheprobability thresholds ofeachclasstoselecttheclasswiththeminimum-expectedcost.The first algorithm(CS-SNB-Accuracy) considersaddingeachvariabletothemodelandmeasurestheperformanceoftheresulting modelonthetrainingdata. Incontrast,thesecondalgorithm(CS-SNB-Cost)considersaddingvariablesthat reduce themisclassification cost,thatis,thedistancebetweenthereadjustedclassandactualclass. Chung-Chian Hsu a, Yan-Ping Huang a,b,*, Keng-Wei Chang[5] we propose a classification method, Extended Naive Bayes (ENB), which is capable for handling mixed data. The experimental results have demonstrated the efficiency of our algorithm in comparison with other classification algorithms ex.

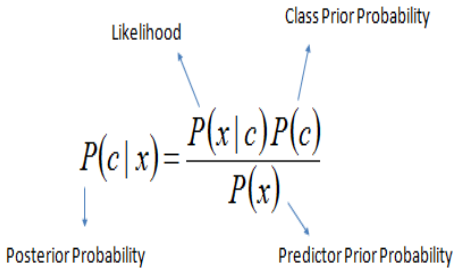
CART, DT and MLP's.Dewan Md. Farid a, Li Zhang ,ChowdhuryMofizurRahman[6] presented a paper, we introduce two independent hybrid mining algorithms to improve the classificationaccuracy rates of decision tree (DT) and naive Bayes (NB) classifiers for the classification of multi-classproblems. in our firstproposed hybrid DT algorithm, we employ a naive Bayes (NB) classifier to remove the noisy troublesomeinstances from the training set before the DT induction. Thus, in the second proposed hybrid NB classifier, we employ a DT induction to select acomparatively more important subset of attributes for the production of naive assumption of class conditionalindependence.Pablo Bermejo, Jose A. Gámez, Jose M. Puerta[6] propose a new method based on learning and sampling probability distributions for e-mail classifications. Their experiments over a standard corpus (ENRON) with seven datasets (e-mail users) show that the results obtained by Naive Bayes Multinomial significantly improve when applying the balancing algorithm first. For the sake of completeness in our experimental study we also compare this with another standard balancing method (SMOTE) and classifiers.

III. BACKGROUND KNOWLEDGE

Databases are rich with hidden information that can be used for intelligent decision making. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large. Classification and prediction have numerous applications, including fraud detection, target marketing, performance prediction, manufacturing and medical diagnosis[7]. Classification is used to find the class label for the data and prediction is used find the value in the class label.

A. CLASSIFICATION BY NAÏVE BAYESIAN CLASSIFICATION

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors[8]. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets[9]. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence[10,11].



$$P(c|X) = P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$$

- P(c|x) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

Algorithm: for discrete attributes
Learining Phase: Given a training set S,
For each target value of c_i(c_i=c1,c2,,,,,cn)
P(C=c_i)←estimate P(C=c_i) with examples in S;
For every feature value x_β of each feature H_j(j=1,,,n;k=1,…N_j)
P(Xj= x_β |C=c_j)←estimate P(Xj= x_β |C=c_j) with examples in S;
Output:conditional probability tables; for X_j,N_j×L elements.
Test Phase: Given an unknown instance X'=(a',…..a'$_n$)
Look up tables to assign the label c* to X' if
[P(a'$_1$|c*)…. P(a'$_n$|c*)]P(c*)>[P(a'$_1$|c)… P(a'$_n$|c)]P(c), c≠c*,c=c_1,…,c_L.

B. DATA COLLECTION

The data are collected from a private college at Ulundurpet in Villupuramdistrict[12]. There were 123 records collected from the students who are doing under graduate course who belongs to the age group 18 to 23. Among the 123 students 85 were male and 38 were female candidates. The data used for data mining contains 123 records and have 30 dimensional attribute namely name, gender, age, department, year, mode of transport, college location, home location, test, cinema, festival, sick, miss bus, friend leave, subject boring, staff question, exam study, result, occasionally, institution work, part time job, assignment, pay fees, native, accident, dress code, commitment friends, college care, impress, problem in college. For our study name is not necessary so we omit the attribute and take the 29 attribute for classification[13].

C. SYSTEM FRAMEWORK

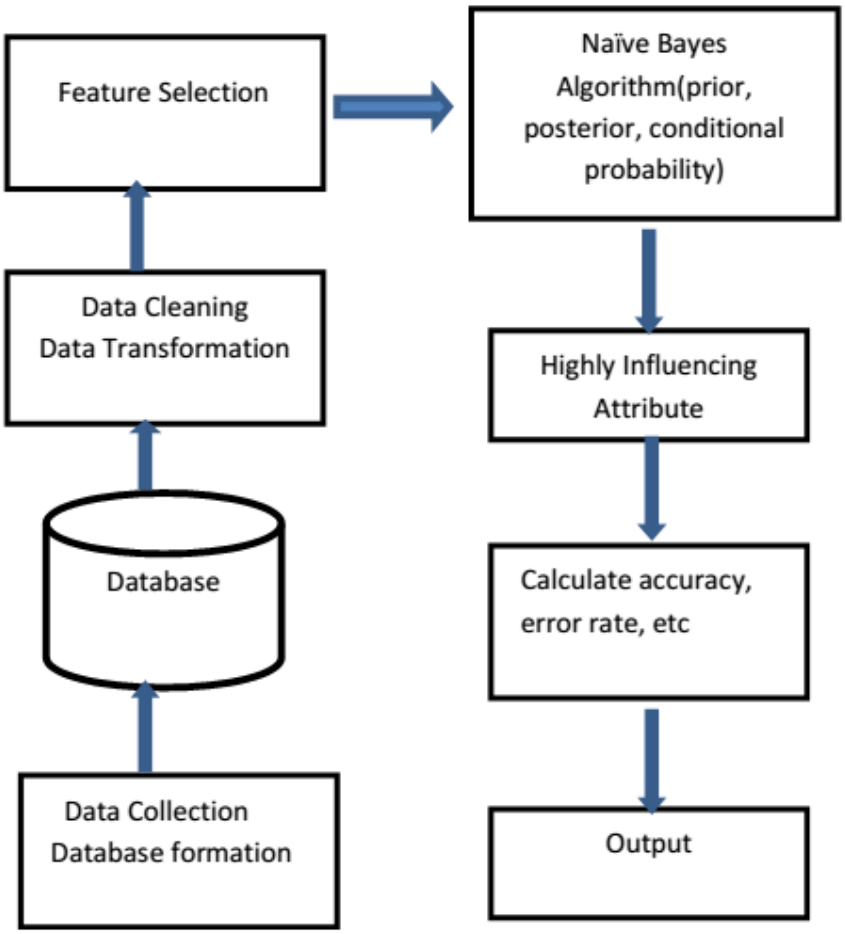


Figure 1: System Frame Work

IV. EXPERIMENTAL SETUP

Tanagra is a open source software used for data mining. It supports all the basic type of data formats like *.xls, *.txt etc. and it is very user friendly. The data set is implemented in Tanagra by the following method[14].

- Open the Tanagra software and go to the file menu and click open then insert the data you want to evaluate
- Select the view dataset and right click and click execute and then click view then the data is displayed on the right side screen
- Drag the Define status from the icon and give the target attribute as gender and in input select all the attributes.
- Go to spv learning and select Naïve Bayes and drag into the Define status and right click it and click execute and then click view then the

V. EXPERIMENTAL RESULTS

The dataset was implemented in tanagra with the following parameters lapacian=1, Lambda for Lablacian=1.0000, Show Conditional Pobabilities=1, Show Model Description= yes. The results obtained from Tanagra is shown in the figure below.

Figure 2: Result Obtained In Tanagra

Prior probability for gender

Probability	Value
P(gender=male)	$\left(\frac{85}{123}\right)$=0.691
P(gender =female)	$\left(\frac{38}{123}\right)$=0.308

Table 1: Prior Probability for gender

Conditional probabilities for each attribute value calculated using the leave dataset and the value whose threshold exceeds one are shown below.

Descriptors	Male	Female	Decision Function
constant	-60.3916	-64.9751	4.583515
Job = Strongly Agree	0.550046	-1.94591	2.495956
Job = Agree	0.659246	-1.54045	2.199691
cinema = Strongly Agree	0.459532	-1.70475	2.16428
Test = Strongly Agree	-0.13976	-2.30259	2.162823
festival = Strongly Agree	2.61496	0.847298	1.767662
Friendleave = Strongly Agree	0.753772	-1.0116	1.765373
festival =	2.397895	0.847298	1.550597

Agree			
cinema = Neutral	0.510826	-1.0116	1.522427
Friendleave = Neutral	0.405465	-1.0116	1.417066
festival = Strongly Disagree	1.098612	-0.18232	1.280934
Staff problem = Neutral	0.262364	-0.91629	1.178655
Friend leave = Disagree	1.139434	0	1.139434
College environment = Strongly Agree	0.847298	-0.28768	1.13498
Friendleave = Strongly Disagree	1.252763	0.167054	1.085709

Table 2: Conditional Probability for all attribute

The results obtained clearly shows that student put leave to college for going to job for yearning money to pay their college fees. If there is any test in the college student put leave to college. If a new cinema release then student put leave to college and go to the cinema. If there is a festival in the students village then they put leave to college. If the student's best friend put leave to college then the student also put leave to college.

A. PERFORMANCE MEASURES

The confusion matrix is a tool for the analysis of a classifier [15]. Given m classes, a confusion matrix is a table of at least size m by m. An entry $CM_{i,j}$ in the first m rows and m columns indicates the number of tuples of class i that were labelled by the classifier as class j. For a classifier to have good accuracy, ideally most of the tuples would be represented along the diagonal of the confusion matrix, from entry $CM_{1,1}$ to entry $CM_{m,m}$ with the rest of the entries being to close to zero[16].

	C1	C2
C1	True Positives	False Negatives
C2	False Positives	True Negatives

Table 3: Confusion Matrix

The performance of the naive bayes classification algorithms were evaluated by the accuracy % on the student leave data set[19]. The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier[17].

	Male	**Female**	**sum**
Male	70	15	85
Female	4	34	38
	74	49	123

Table 4: Confusion Matrix of Naive Bayes Algorithm for our dataset

$$\text{Accuracy (M)} = \frac{TP+TN}{TP+TN+FP+FN}$$

For our data set using C4.5 algorithm we have,

$$\text{Accuracy (M)} = \frac{70+34}{70+34+4+15} = 0.846$$

A true positive (TP) occurs when a classifier correctly classified class1. A true negative (TN) occurs when a classifier correctly classified class2[18]. A false positive (FP) occurs when a classifier incorrectly classified class1. A false negative (FN) occurs when a classifier incorrectly classified class2.

$$\text{Error Rate} = 1 - \text{Accuracy (M)}$$

For our data set using C4.5 algorithm we have,

$$\text{Error Rate} = 1 - 0.846 = 0.154$$

The sensitivity measures the proportion of the actual positives which are correctly identified.

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

For our data set using C4.5 algorithm we have,

$$\text{Sensitivity} = \frac{70}{70+15} = 0.823$$

The specificity measures the proportion of negatives which are correctly identified.

$$\text{Specificity} = \frac{TN}{TN+FP}$$

For our data set using C4.5 algorithm we have,

$$\text{Specificity} = \frac{34}{34+4} = 0.895$$

Precision is the percentage of true positives (TP) compared to the total number of cases classified as positive events[13].

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{1-Precision} = 1 - \frac{TP}{TP+FP}$$

For our data set using C4.5 algorithm we have,

$$\text{Precision} = \frac{70}{70+4} = 0.945$$

$$\text{1-Precision} = 1 - 0.945 = 0.054$$

Recall is the proportion of the total number of predictions that were correct.

$$\text{Recall} = \left(\frac{TP}{TP+FN}\right)$$

$$\text{Recall} = \left(\frac{70}{70+15}\right) = 0.823$$

The following table values are calculated by using the confusion matrix obtained from Tanagra tool.

Measures	Naive Bayes
Accuracy	0.846
Error Rate	0.154
Recall	0.823
Specificity	0.895
Precision	0.945
1- Precision	0.054

Table 5: Performance values of the Naive Bayes Algorithms

B. RECOMMENDATIONS

From the above experimental results we know that the job attribute plays a key role on job going students to earn money. To improve good learning environment and the quality of education in the rural and semi-rural areas, our suggestion is that change the college timings such as morning and evening sessions to avoid the students absenteeism for the classes. It was found that student put leave

due to the purpose of studying for the examinations so if we enough study holiday we can avoid students putting leave to college.

VI. CONCLUSION

This research aims to study the pattern of students who put leave to the college frequently and the reason behind the students to put leave. In this research, the naïve bayes algorithm technique has been used because it is easy to interpret and understand. This algorithm helps to understand the most influencing independent attributes in the dataset. The college management along with the educational experts should take necessary steps to prevent student's absenteeism and help to improve the educational level of the students.

REFERENCES

[1] *Muralidharan, V., &Sugumaran, V." A comparative study of Naïve Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis", Applied Soft Computing, Vol.12 Issue 8, 2012 pp.2023-2029.*

[2] *Koc, L., Mazzuchi, T. A., &Sarkani, S. (2012). A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier. Expert Systems with Applications, Vol.39 Issue 18, 2012, pp.13492-13500.*

[3] *Soria, D., Garibaldi, J. M., Ambrogi, F., Biganzoli, E. M., & Ellis, I. O, "A 'non-parametric'version of the naive Bayes classifier", Knowledge-Based Systems, Vol.24 Issue 6, 2011, pp.775-784.*

[4] *Ibáñez, A., Bielza, C., &Larrañaga, P. " Cost-sensitive selective naive Bayes classifiers for predicting the increase of the h-index for scientific journals", Neurocomputing, Vol.135, 2014 pp. 42-52.*

[5] *Hsu, C. C., Huang, Y. P., & Chang, K. W,"Extended Naive Bayes classifier for mixed data", Expert Systems with Applications, Vol.35 Issue 3, 2008,pp.1080-1083.*

[6] *Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M. A., & Strachan, R. (2014). Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. Expert Systems with Applications, Vol.41 Issue 4, 2014,pp.1937-1946.*

[7] *Bermejo, P., Gámez, J. A., &Puerta, J. M,"Improving the performance of Naive Bayes multinomial in e-mail foldering by introducing distribution-based balance of datasets", Expert Systems with Applications, Vol.38 Issue3, 2011, pp.2072-2080.*

[8] *Jiang, L., Cai, Z., Wang, D., & Zhang, H,"Improving Tree augmented Naive Bayes for class probability estimation", Knowledge-Based Systems,Vol.26, 2012, pp.239-245.*

[9] *Ali, W., Shamsuddin, S. M., & Ismail, A. S, "Intelligent Naïve Bayes-based approaches for Web proxy caching", Knowledge-Based Systems, Vol.31, 2012,pp.162-175.*

[10] *Mukherjee, S., & Sharma, N."Intrusion detection using naive Bayes classifier with feature reduction", Procedia Technology, Vol.4, 2012,pp.119-128.*

[11] *Peng, F., Schuurmans, D., & Wang, S. "Augmenting naive bayes classifiers with statistical language models", Information Retrieval, Vol.7, Issue 3-4, 2004, pp.317-345.*

[12] *Jiang, L., Cai, Z., Zhang, H., &Wang, D,"Not so greedy: Randomly selected naive bayes", Expert Systems with Applications, Vol.39, Issue 12, 2012, pp. 11022-11028.*

[13] *Lakoumentas, J., Drakos, J., Karakantza, M., Sakellaropoulos, G., Megalooikonomou, V., &Nikiforidis, G. "Optimizations of the naïve-Bayes classifier for the prognosis of B-Chronic Lymphocytic Leukemia incorporating flow cytometry data", Computer methods and programs in biomedicine, Vol.108, Issue 1, 2012, pp.158-167.*

[14] *Han, J., Kamber, M., & Pei, J., "Data mining, southeast asia edition: Concepts and techniques", Morgan kaufmann, 2006.*

[15] *Kumar, S. A., &Vijayalakshmi, M. N.,." Efficiency of decision trees in predicting student's academic performance", In First International Conference on Computer Science, Engineering and Applications, CS and IT , Vol. 2, 2011, pp. 335-343.*

[16] *e Sun, H., " Research on Student Learning Result System based on Data Mining", IJCSNS, Vol. 10, Issue 4, 2010, pp.203.*

[17] *Lu, S. H., Chiang, D. A., Keh, H. C., & Huang, H. H., "Chinese text classification by the Naïve Bayes Classifier and the associative classifier with multiple confidence threshold values", Knowledge-based systems,Vol . 23, Issue 6, 2010, pp.598-604.*

[18]Lin, K. C., Liao, I. E., Chang, T. P., & Lin, S. F. "A frequent itemset mining algorithm based on the Principle of Inclusion–Exclusion and transaction mapping". Information Sciences, Vol.276, 2014, pp.278-289.

[19]Liu, Y. H., & Wang, C. S. Constrained frequent pattern mining on univariate uncertain data. Journal of Systems and Software, Vol.86, Issue 3, 2013, pp.759-778.

[20]Yu, K. M., & Zhou, J." Parallel TID-based frequent pattern mining algorithm on a PC Cluster and grid computing system". Expert Systems with Applications, Vol 37, Issue 3,2010, pp.2486-2494.