ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 10, October 2017

Students Grades Predictor using Naïve Bayes Classifier – A Case Study of University of Education, Winneba

Delali Kwasi Dake[1], Esther Gyimah[2]

Lecturer, Department of Information and Communication Technology, University of Education, Winneba, Ghana[1]

Lecturer, Department of Information and Communication Technology, University of Education, Winneba, Ghana[2]

ABSTRACT: Educational Data Mining is an emerging research area with vast algorithms to analyse hidden patterns in student's data and discover new knowledge for academic counselling and improvement. Recent deployment of Classification and Clustering algorithms on educational data has proven to be more successful in classifying data instances correctly. One key focus of higher education is to provide quality education and guidance throughout the academic year. This can only be achieved if student'sfailure rate can be averted in final exams by determining under-performing students earlier on in the semester and sampled for academic counselling and recommendation by school authorities. This study presents a Naïve Bayes Classification approach in predicting student's final grade. The Classifier model was built on the training dataset of previous students who offered the same course with a predictor accuracy rate of 88%. This deployed algorithm will help under-performing students to improve their final exam score.

KEYWORDS: Educational Data Mining, Naïve Bayes Classification, Performance, Prediction

I. INTRODUCTION

Educational Data Mining (EDM) is a fast growing research area with vast application in tertiary institutions. Data mining is a research area that deals in knowledge discovery in large data sets using data mining algorithms. The importance of data mining algorithms is to find interesting and hidden patterns in volumes of data. Academic authorities and students could benefit from the application of data mining techniques on large data sets from tertiary institutions.

One area of concern in tertiary institutions is to provide students with the necessary guidance in the learning process. A successfully implemented Naïve Bayes Classifier's discovered knowledge can help students and academic instructors in carrying out better and enhanced educational quality, by identifying possible under-performer's mid-semester and counsel them for educational growth and results. It is quite difficult to use manual mechanisms to track students' performance over past academic years and make any meaningful decision based on that.

Educational Data Mining techniques such as Decision Trees, Neural Networks, Naïve Bayes, K-Nearest neighbour, are useful techniques for discovering knowledge in large data sets that can be used for prediction regarding enrolment of students in a particular course, alienation of traditional classroom teaching model, detection of unfair means used in online examination, detection of abnormal values in the result sheets of the students, prediction about students' performance and so on.

This study will focus on using one data mining algorithm, Naïve Bayes, to predict a student as either failing or passing after tracking the academic variables of the student. Naïve Bayes are predictor classification algorithms and used to study historical data to generate comprehensive and precise analysis [1].

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 10, October 2017

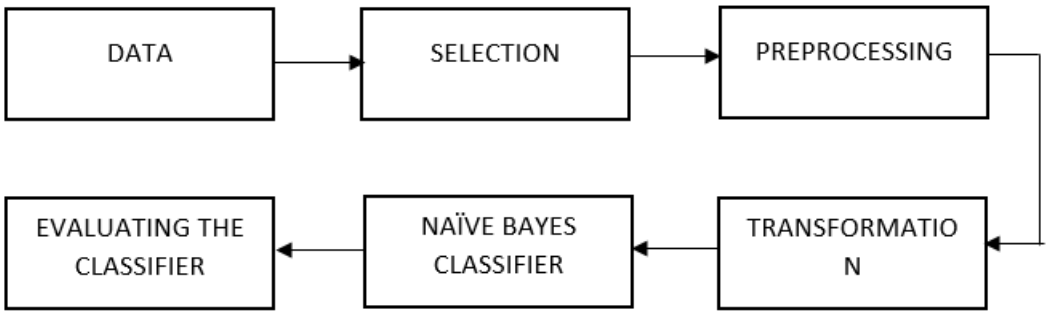

Fig 1: Steps of extracting knowledge using Naive Bayes Classifier

Problem Statement

Grading students using class score, assignment, project score and final exam is one of the main academic focus of tertiary institutions. Especially for students, terminal grading forms a core part of their final Grade Point Average (GPA) calculations and significantly affects their future opportunities. It becomes difficult for academic counsellors to advise failing students' mid-semester due to large classroom size and the manual approach of detecting such under-performing students. It even becomes more difficult for academic authorities to prevent any future failure occurrence of different set of students due to a non-deployment of machine learning algorithms to study data patterns.

Supervised learning algorithm will be deployed in this research to automate large data set and detect failing students for counselling. This algorithm will also inform academic authorities on the variables that normally lead to high failure rates and a possible course re-design.

II. RELATED WORK

Recent research on data mining techniques for predicting students' academic performance in higher education institutions has focused on classification algorithms such as decision trees among others.

[2] analysed the accuracy of the decision tree algorithm using the C4.5 decision tree algorithm on student's internal assessment data to predict their performance in the final exam. The results of this research proved that decisions tree is an effective classification technique for predicting and improving students' academic performance and hence, the efficiency of various decision tree algorithms can be analyzed based on their accuracy and time taken to derive the tree.

Another paper investigated the accuracy of decision tree techniques for predicting students' academic performance by identifying the dropouts and students who need special attention and allowing the teacher to provide appropriate counselling earlier in time [3]. A similar study conducted by [4] used the decision tree technique to predict the drop out feature of students to enable academic counselors take the needed interventions to retain students.

[5] compared the accuracy of decision tree and bayesian network algorithms for predicting the academic performance of undergraduate and postgraduate students at two very different academic institutions. The results showed the decision tree was consistently 3-12% more accurate than the bayesian network.

A paper by [6] also discussed the use of decision trees in educational data mining. The algorithm was applied on engineering students' past performance data to generate a model that could be used to predict students' performance to enable the teacher identify failing students in advance and give appropriate help to such students.

[7][8] used Weka's J48 algorithm to build decision trees with the purpose of differentiating and predicting students' choice in continuing their education with post university studies.

This paper will investigate the accuracy of the decision tree algorithm as a predictor model for determining students' grades for courses before end of semester examinations.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 10, October 2017

III. METHODOLOGY

In this research, probabilistic Naïve Bayes classification algorithm based on statistical inference is adopted [9]. This classification algorithm will output a corresponding posterior probability of the test instance been a member of each of the possible classes. The posterior probability relates to observing the specific characteristics of the test instance. On the other hand the prior probability is simply the fraction of training records belonging to each particular class, with no knowledge of the test instance. This training records in normal database terms is referred to as the attributes or the fields. With this model, decision theory to determine class membership for each new instance is deployed. Bayes classifier relies on independency of training set data or features. Consider a test instance with d different features, which have values $X = \langle x_1 \dots x_d \rangle$ respectively. It is desirable to determine the posterior probability that the class $Y(T)$ of the test instance T is i. To determine the posterior probability $P(Y(T)= i|x_1...x_d)$. Then, the Bayes rule can be used in order to derive the following: [10][11]

$$P(Y(T) = i|x_1 \dots x_d) = P(Y(T) = i) \cdot \frac{P(x_1 \dots x_d|Y(T) = i)}{P(x_1 \dots x_d)} \quad (1.1)$$

$P(Y(T)= i|x_1...x_d)$ = *Posterior probability of instance* x_d *being in class* $Y(T)$

$P(x_1 \dots x_d|Y(T) = i)$ = *Probabiltiy of generating instance* x_d *given class* $Y(T)$

$P(x_1 \dots x_d)$ = *Probability of instance of* x_d *occurring*

Since the denominator is constant across all classes, and one only needs to determine the class with maximum posterior probability, this expression can be approximated as follows:

$P(Y(T)= i|x_1...x_d) \propto P(Y(T)= i) \cdot P(x_1...x_d|Y(T)= i)$ (1.2)

IV. DATA DESCRIPTION

No	Attribute	Description	Values
1	Attendance	Is the student regular in class?	Poor <6/12; Average 6<&<10; Good >10
2	Assignment	Average of students assignment score	Weak < 10/20; Average 10<&<15; Excellent 15>&<=20
3	Test Score	Average of Students Test Score	Weak < 15/30; Average 15<&<20; Excellent 20>&<=30
4	Class Participation	How often does student answer or take part in class work	Normal; High
5	Hostel Proximity	Hostels proximity to lecture hall and campus	Near; Far
6	Gender		Male, Female
	PSF	Pass Final Exam?	Yes >60%; No <60 {Class Predictor}

Table 1: Attribute description of training dataset

The attribute and their values were selected based on the discretion that they can have an effect on a student's ability to Pass Final Exam.

Attendance: Students attendance can have an effect on the final exam. The attendance is for the twelve (12) weeks of active class. Poor 6/12, Average is >6 and <10, Good is >10.

Assignment: The average performance of two assignments was taken out of 20 marks. Weak <10, Average >10 and <15, Excellent >15. Students were expected to perform better in assignments.

Test Score: The average of two Test Score was taken out of 30 marks. Weak <15, Average >15 and <20, Excellent >20

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 10, October 2017

Class Participation: Relevant index tag was given to students who answered questions in class and were active during lessons. Active students were given High tag against Normal.
Hostel Proximity: Hostel proximity can affect student's attendance, group discussions and class concentration. Students were recorded and tag as Far, Near from campus
Pass Final Exam: This is the Class Predictor attribute. A score of >60% is a pass and <60% a fail. The high pass value of 60% is to help include more failing students in the counseling section.

V. RESULTS ANALYSIS

To run the Classifier, a total of 50 instances were taken for analysis. The Comma-Separated Values (CSV) was converted to the Weka Attribute-Relation File Format (ARFF) using the ARFF-Viewer[12].

ARFF-Viewer - E:\2017_2018_first_sem\research\research1_student_grades\result_file\grade_predictor.arff

File Edit View

grade_predictor.arff

Relation: grade_predictor

No.	1: Attendance Nominal	2: Class Participation Nominal	3: Gender Nominal	4: Hostel Proximity Nominal	5: Assignment Nominal	6: Test Score Nominal	7: **Pass Final Exam** Nominal
1	Good	Normal	Male	Near	Average	Weak	No
2	Good	Normal	Male	Near	Excellent	Average	Yes
3	Good	High	Male	Far	Average	Average	Yes
4	Good	Normal	Male	Near	Excellent	Average	Yes
5	Average	High	Male	Near	Excellent	Weak	No
6	Good	Normal	Male	Near	Weak	Weak	No
7	Average	Normal	Male	Near	Excellent	Excellent	Yes
8	Poor	Normal	Male	Far	Weak	Average	No
9	Good	High	Female	Near	Excellent	Average	Yes
10	Good	Normal	Male	Far	Average	Average	No
11	Average	Normal	Male	Near	Average	Average	Yes
12	Good	Normal	Male	Near	Excellent	Average	Yes
13	Good	Normal	Male	Far	Average	Weak	No
14	Good	High	Male	Near	Excellent	Excellent	Yes
15	Average	Normal	Male	Near	Average	Excellent	Yes
16	Good	High	Male	Far	Excellent	Weak	No
17	Good	High	Female	Near	Weak	Weak	No
18	Good	Normal	Male	Far	Average	Average	Yes
19	Good	High	Male	Near	Average	Weak	No
20	Good	Normal	Male	Near	Weak	Average	No
21	Poor	Normal	Male	Far	Excellent	Average	Yes
22	Good	High	Male	Near	Average	Excellent	Yes
23	Good	Normal	Female	Far	Excellent	Average	Yes
24	Good	Normal	Male	Near	Average	Excellent	Yes
25	Good	Normal	Male	Far	Excellent	Average	Yes
26	Good	Normal	Male	Near	Excellent	Excellent	Yes
27	Good	Normal	Male	Near	Average	Average	Yes
28	Average	High	Male	Near	Average	Average	Yes
29	Good	Normal	Male	Near	Excellent	Average	Yes
30	Good	Normal	Female	Far	Average	Weak	No
31	Average	Normal	Male	Near	Excellent	Average	Yes
32	Good	Normal	Male	Near	Excellent	Average	Yes
33	Good	Normal	Male	Near	Average	Average	Yes
34	Good	Normal	Male	Near	Average	Weak	No
35	Average	High	Male	Near	Excellent	Excellent	Yes
36	Good	Normal	Female	Far	Excellent	Average	Yes
37	Good	Normal	Female	Near	Average	Excellent	Yes
38	Good	Normal	Male	Near	Average	Excellent	Yes
39	Good	High	Male	Near	Excellent	Excellent	Yes
40	Good	High	Male	Far	Average	Weak	No
41	Average	Normal	Male	Near	Excellent	Average	No
42	Good	Normal	Male	Near	Average	Average	No
43	Good	Normal	Male	Far	Excellent	Average	Yes
44	Good	Normal	Male	Near	Excellent	Excellent	Yes
45	Good	Normal	Female	Far	Average	Excellent	No
46	Good	High	Male	Far	Excellent	Excellent	Yes
47	Good	Normal	Male	Near	Average	Excellent	No
48	Good	Normal	Male	Near	Average	Average	Yes
49	Good	Normal	Male	Near	Average	Weak	No
50	Good	Normal	Male	Far	Excellent	Average	Yes

Table 2: Training Dataset viewed in ARFF format of Weka

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 10, October 2017

The Training Set Data was subjected to Naïve Bayesian Classifier. The results obtained from the Classifier gave 88% accuracy rate for correctly classified instances.

06:25:28 - bayes.NaiveBayes

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     grade_predictor
Instances:    50
Attributes:   7
              Attendance
              Class Participation
              Gender
              Hostel Proximity
              Assignment
              Test Score
              Pass Final Exam
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

                        Class
Attribute                 No    Yes
                       (0.37) (0.63)
===================================
Attendance
  Good                  16.0   26.0
  Average                3.0    7.0
  Poor                   2.0    2.0
  [total]               21.0   35.0

Class Participation
  Normal                14.0   25.0
  High                   6.0    9.0
  [total]               20.0   34.0

Gender
  Male                  16.0   29.0
  Female                 4.0    5.0
  [total]               20.0   34.0

Hostel Proximity
  Near                  12.0   24.0
  Far                    8.0   10.0
  [total]               20.0   34.0

Assignment
  Average               12.0   13.0
  Excellent              4.0   21.0
  Weak                   5.0    1.0
  [total]               21.0   35.0

Test Score
  Weak                  12.0    1.0
  Average                6.0   21.0
  Excellent              3.0   13.0
  [total]               21.0   35.0
```

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 10, October 2017

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          44               88      %
Incorrectly Classified Instances         6               12      %
Kappa statistic                          0.7191
Mean absolute error                      0.2413
Root mean squared error                  0.3403
Relative absolute error                 52.027  %
Root relative squared error             70.6809 %
Total Number of Instances               50

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.667    0.000    1.000      0.667    0.800      0.749    0.849     0.860     No
               1.000    0.333    0.842      1.000    0.914      0.749    0.849     0.847     Yes
Weighted Avg.  0.880    0.213    0.899      0.880    0.873      0.749    0.849     0.852

=== Confusion Matrix ===

  a  b   <-- classified as
 12  6 |  a = No
  0 32 |  b = Yes
```

Fig2: Results Obtained for Naive Bayesian Classifier

Based on the Naïve Bayes probability estimate in Figure 2, the posterior probability from equation (1.1) can be calculated.

(**PFE** = Yes) = 0.63	Pr (**PFE** = No) = 0.37	
Pr (**Attendance** = Good | PFE = Yes) = 0.74	Pr (**Attendance** = Average | PFE = Yes) = 0.2	Pr (**Attendance** = Poor | PFE = Yes) = 0.06
Pr (Attendance = Good | PFE = No) = 0.76	Pr (Attendance = Average | PFE = No) = 0.14	Pr (Attendance = Poor | PFE = No) = 0.09
Pr (**Class Parti** = Normal | PFE = Yes) = 0.74	Pr (**Class Parti** = High | PFE = Yes) = 0.26	
Pr (Class Parti = Normal | PFE = No) = 0.7	Pr (Class Parti = High | PFE = No) = 0.3	
Pr (**Gender** = Male | PFE = Yes) = 0.85	Pr (**Gender** = Female | PFE = Yes) = 0.15	
Pr (Gender = Male | PFE = No) = 0.8	Pr (Gender = Female | PFE = No) = 0.2	
Pr (**Hostel Proxi** = Near | PFE = Yes) = 0.71	Pr (**Hostel Proxi** = Far | PFE = Yes) = 0.29	

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 10, October 2017

Pr (Hostel Proxi = Near | PFE = No) = 0.6	Pr (Hostel Proxi = Far | PFE = No) = 0.4	
Pr (**Assignment** = Average | PFE = Yes) = 0.37	Pr (**Assignment** = Excellent | PFE = Yes) = 0.6	Pr (**Assignment** = Weak | PFE = Yes) = 0.03
Pr (Assignment = Average | PFE = No) = 0.57	Pr (Assignment = Excellent | PFE = No) = 0.19	Pr (Assignment = Weak | PFE = No) = 0.24
Pr (**Test Score** = Weak | PFE = Yes) = 0.03	Pr (**Test Score** = Average | PFE = Yes) = 0.6	Pr (**Test Score** = Excellent | PFE = Yes) = 0.37
Pr (Test Score = Weak | PFE = No) = 0.57	Pr (Test Score = Average | PFE = No) = 0.29	Pr (Test Score = Excellent | PFE = No) = 0.14

Table 3: Calculating the probability of each of the attributes on the Class Predictor using Training Set Data

Based on the Naïve Bayes Algorithm deployed on the Training Data Set, we can now predict whether a new student will Pass Final Exam (PFE). This becomes a Test Data for the Naïve Bayes Algorithm

Instance	Attendance	Class Participation	Gender	Hostel Proximity	Assignment	Test Score	Pass Final Exam?
1	Average	Normal	Male	Far	Average	Average	? = Yes
2	Good	High	Female	Near	Average	Weak	? = No

Table 4: Using Naïve Bayes Classifier to Predict a Test Data

Scenario One

Based on Naïve Bayesian Classification using Table 3for *Test Data Instance 1*of Table 4:

P(PFE = Yes) = P(Attendance = Normal | Yes) * P(Class Participation = Normal | Yes) * P(Gender = Male | Yes) * P(Hostel Proximity = Far | Yes) * P(Assignment = Average | Yes) * P(Test Score = Average | Yes) * P(PFE = Yes)
= 0.63*0.2*0.74*0.85*0.29*0.37*0.6
= 0.05

P(PFE = No) = P(Attendance = Normal | No) * P(Class Participation = Normal | No) * P(Gender = Male | No) * P(Hostel Proximity = Far | No) * P(Assignment = Average | No) * P(Test Score = Average | No) * P(PFE = No)
= 0.37*0.14*0.7*0.8*0.4*0.57*0.29
= 0.02

From the probability ratio and the Naïve Based Algorithm a student with such data will have a higher probability (0.05) of**passing the final exam**.

Scenario Two

Based on Naïve Bayesian Classification using Table 3 for *Test Data Instance 2* of Table 4:

P(PFE = Yes) = P(Attendance = Good | Yes) * P(Class Participation = High | Yes) * P(Gender = Female | Yes) * P(Hostel Proximity = Near | Yes) * P(Assignment = Average | Yes) * P(Test Score = Weak | Yes) * P(PFE = Yes)
= 0.74*0.26*0.15*0.71*0.37*0.03*0.63
= 0.00014

P(PFE = No) = P(Attendance = Good | No) * P(Class Participation = High | No)) * P(Gender = Female | No)) * P(Hostel Proximity = Near | No) * P(Assignment = Average | No) * P(Test Score = Weak | No)) * P(PFE = No)
= 0.76*0.3*0.2*0.6*0.57*0.57*0.37
= 0.0033

From the probability ratio and the Naïve Based Algorithm a student with such data will have a higher probability (0.0033) of **failing the final exam**.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 10, October 2017

VI. CONCLUSION

In this paper, we used the Naïve Bayes Classifier on a training set data to predict whether a student will Pass or Fail the final exam. The classifier gave a high accuracy rate of 88% with incorrectly classified instances of 12%. With the help of the classifier, the probability of each attribute in determining the Final Grade was calculated. It was possible then to use a test data to predict whether a student mid-semester will Pass or Fail the final exam.

REFERENCES

[1] Undavia, J. N., Dolia, P. M.; Shah, N. P. "Prediction of Graduate Students for Master Degree based on Their Past Performance using Decision Tree in Weka Environment". International Journal of Computer Applications; Volume 74 (21), (2013).

[2] Kumar, S. A., &Vijayalakshmi, M. N. (2011, October). 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, pp. 335-343).

[3] Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. *arXiv preprint arXiv:1201.3417*.

[4] Quadri, M. M., &Kalyankar, N. V. (2010). Drop out feature of student data for academic performance using decision tree techniques. *Global Journal of Computer Science and Technology*, *10*(2).

[5] Nghe, N. T., Janecek, P., &Haddawy, P. (2007, October). A comparative analysis of techniques for predicting academic performance. In *Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual* (pp. T2G-7). IEEE.

[6] Kabra, R. R., &Bichkar, R. S. (2011). Performance prediction of engineering students using decision trees. *International Journal of Computer Applications*, *36*(11), 8-12.

[7] Bresfelean, V. P. (2007, June). Analysis and predictions on students' behavior using decision trees in Weka environment. In *Proceedings of the ITI* (pp. 25-28).

[8] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

[9] Murphy, K. P. (2006). Naive bayes classifiers. *University of British Columbia*.

[10] Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41-46). IBM

[11] Hongbo Deng, Yizhou Sun, Yi Chang, and Jiawei Han "Probabilistic Models for Classification," in Data Classification Algorithms and Applications, IBM T.J. Watson Research Center Yorktown Heights, New York, USA, 2015, pp. 66-69

[12] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10-18.