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

Delali Kwasi Dake¹, Esther Gyimah²
Lecturer, Department of Information and Communication Technology, University of Education, Winneba, Ghana¹
Lecturer, Department of Information and Communication Technology, University of Education, Winneba, Ghana²

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’s failure 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].
Problem Statement
Grading students using class score, assignment, project score and final exam is one of the main academic focuses 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 underperforming 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.
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 being 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 \(X = (x_1, \ldots, x_d)\) respectively. It is desirable to determine the posterior probability that the class \(Y(T) = i\) of the test instance \(T\). To determine the posterior probability

\[
P(Y(T = i) | x_1, \ldots, x_d) = \frac{P(Y(T = i) \cdot P(x_1, \ldots, x_d | Y(T = i))}{P(x_1, \ldots, x_d)}
\]

(1.1)

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:

\[
P(Y(T) = i | x_1, \ldots, x_d) \propto P(Y(T) = i) \cdot P(x_1, \ldots, x_d | Y(T) = i)
\]

(1.2)

IV. DATA DESCRIPTION

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

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
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].

Table 2: Training Dataset viewed in ARFF format of Weka
The Training Set Data was subjected to Naive Bayesian Classifier. The results obtained from the Classifier gave 88% accuracy rate for correctly classified instances.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Class</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attendance</td>
<td></td>
<td>0.37</td>
<td>0.63</td>
</tr>
<tr>
<td>Good</td>
<td>16.0</td>
<td>26.0</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>3.0</td>
<td>7.0</td>
<td></td>
</tr>
<tr>
<td>Poor</td>
<td>2.0</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>[total]</td>
<td>21.0</td>
<td>35.0</td>
<td></td>
</tr>
<tr>
<td>Class Participation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>14.0</td>
<td>25.0</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>6.0</td>
<td>9.0</td>
<td></td>
</tr>
<tr>
<td>[total]</td>
<td>20.0</td>
<td>34.0</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>16.0</td>
<td>29.0</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>4.0</td>
<td>5.0</td>
<td></td>
</tr>
<tr>
<td>[total]</td>
<td>20.0</td>
<td>34.0</td>
<td></td>
</tr>
<tr>
<td>Hostel Proximity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Near</td>
<td>12.0</td>
<td>24.0</td>
<td></td>
</tr>
<tr>
<td>Far</td>
<td>8.0</td>
<td>10.0</td>
<td></td>
</tr>
<tr>
<td>[total]</td>
<td>20.0</td>
<td>34.0</td>
<td></td>
</tr>
<tr>
<td>Assignment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>12.0</td>
<td>13.0</td>
<td></td>
</tr>
<tr>
<td>Excellent</td>
<td>4.0</td>
<td>21.0</td>
<td></td>
</tr>
<tr>
<td>Weak</td>
<td>5.0</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>[total]</td>
<td>21.0</td>
<td>35.0</td>
<td></td>
</tr>
<tr>
<td>Test Score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weak</td>
<td>12.0</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>6.0</td>
<td>21.0</td>
<td></td>
</tr>
<tr>
<td>Excellent</td>
<td>3.0</td>
<td>13.0</td>
<td></td>
</tr>
<tr>
<td>[total]</td>
<td>21.0</td>
<td>35.0</td>
<td></td>
</tr>
</tbody>
</table>
Fig2: Results Obtained for Naive Bayesian Classifier

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

<table>
<thead>
<tr>
<th>Event</th>
<th>Pr (Event)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(PFE = Yes)</td>
<td>0.63</td>
</tr>
<tr>
<td>(PFE = No)</td>
<td>0.37</td>
</tr>
<tr>
<td>(PFE = Yes)</td>
<td>0.74</td>
</tr>
<tr>
<td>(PFE = No)</td>
<td>0.76</td>
</tr>
<tr>
<td>(PFE = Yes)</td>
<td>0.74</td>
</tr>
<tr>
<td>(PFE = No)</td>
<td>0.7</td>
</tr>
<tr>
<td>(PFE = Yes)</td>
<td>0.85</td>
</tr>
<tr>
<td>(PFE = No)</td>
<td>0.8</td>
</tr>
<tr>
<td>(PFE = Yes)</td>
<td>0.71</td>
</tr>
<tr>
<td>(PFE = No)</td>
<td>0.7</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\text{Pr (Attendance = Good | PFE = Yes) } &= 0.74 \\
\text{Pr (Attendance = Average | PFE = Yes) } &= 0.2 \\
\text{Pr (Attendance = Poor | PFE = Yes) } &= 0.06 \\
\text{Pr (Attendance = Good | PFE = No) } &= 0.76 \\
\text{Pr (Attendance = Average | PFE = No) } &= 0.14 \\
\text{Pr (Attendance = Poor | PFE = No) } &= 0.09 \\
\text{Pr (Class PArt = Normal | PFE = Yes) } &= 0.74 \\
\text{Pr (Class PArt = High | PFE = Yes) } &= 0.26 \\
\text{Pr (Class PArt = Normal | PFE = No) } &= 0.7 \\
\text{Pr (Class PArt = High | PFE = No) } &= 0.3 \\
\text{Pr (Gender = Male | PFE = Yes) } &= 0.85 \\
\text{Pr (Gender = Female | PFE = Yes) } &= 0.15 \\
\text{Pr (Gender = Male | PFE = No) } &= 0.7 \\
\text{Pr (Gender = Female | PFE = No) } &= 0.2 \\
\text{Pr (Hostel Proxi = Near | PFE = Yes) } &= 0.71 \\
\text{Pr (Hostel Proxi = Far | PFE = Yes) } &= 0.29
\end{align*}
\]
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.

**Scenario One**

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

\[
P(PFE = Yes) = P(Attendance = Normal | Yes) \times P(Class Participation = Normal | Yes) \times P(Gender = Male | Yes) \times P(Hostel Proximity = Far | Yes) \times P(Assignment = Average | Yes) \times P(Test Score = Average | Yes) \times P(PFE = Yes)
\]

\[
= 0.63 \times 0.74 \times 0.29 \times 0.37 \times 0.6
\]

\[
= 0.05
\]

\[
P(PFE = No) = P(Attendance = Good | No) \times P(Class Participation = Normal | No) \times P(Gender = Male | No) \times P(Hostel Proximity = Near | No) \times P(Assignment = Average | No) \times P(Test Score = Weak | No) \times P(PFE = No)
\]

\[
= 0.37 \times 0.14 \times 0.7 \times 0.57 \times 0.57
\]

\[
= 0.0033
\]

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:

\[
P(PFE = Yes) = P(Attendance = Good | Yes) \times P(Class Participation = High | Yes) \times P(Gender = Female | Yes) \times P(Hostel Proximity = Near | Yes) \times P(Assignment = Average | Yes) \times P(Test Score = Weak | Yes) \times P(PFE = Yes)
\]

\[
= 0.74 \times 0.71 \times 0.37 \times 0.37 \times 0.63
\]

\[
= 0.00014
\]

\[
P(PFE = No) = P(Attendance = Good | No) \times P(Class Participation = High | No) \times P(Gender = Female | No) \times P(Hostel Proximity = Near | No) \times P(Assignment = Average | No) \times P(Test Score = Weak | No) \times P(PFE = No)
\]

\[
= 0.76 \times 0.6 \times 0.57 \times 0.57
\]

\[
= 0.0033
\]

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.
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