

## **Prediction of Student Academic Performance and Social Behaviour Using Data Mining**

**Dr. FLORENCE VIJILA S,**  
HOD of Computer Science,  
CSI Ewart Women's Christian College,  
Melrosapuram, Chengalpet District,  
Tamil Nadu, India.  
[florencevijila@yahoo.com](mailto:florencevijila@yahoo.com)

### **Address for Correspondence**

**Dr. FLORENCE VIJILA S,**  
HOD of Computer Science,  
CSI Ewart Women's Christian College,  
Melrosapuram, Chengalpet District,  
Tamil Nadu, India.  
[florencevijila@yahoo.com](mailto:florencevijila@yahoo.com)

### **Abstract**

Education is used to reach new heights in the world. Educational Data Mining is used to extract useful information from previously acquired knowledge. Educational data mining is the process of analysing and visualising data from an educational institution using various data mining tools and techniques in order to discover a unique pattern of students' academic performance and behaviour. The purpose of this paper is to improve students' academic performance by utilising data mining techniques. The Naive Bayesian algorithm can be used to predict the academic performance and behaviour of students. The training and testing phases are involved in classifying students into two groups, pass and fail. The Naive Bayes classifier is built during the training phase, and it is used to make predictions during the testing phase. The WEKA tool is used to calculate the classifier's accuracy. The obtained classifier accuracy is 87%, which can be improved further by selecting appropriate attributes. Developing Classification algorithms in this manner aids in the development of a more efficient student performance predictor tool using other data mining algorithms, as well as in the improvement of educational quality in institutions.

**Keywords**— WEKA, Naive Bayes Classifier, Educational Data Mining

## **LINTRODUCTION**

Education is essential for the growth of the country's economic progress. The literacy statistics show the percentage of failure rates due to students dropping out of school and failing a specific subject. Over the last few years, there has been an increase in the global literacy rate. An educational institution must keep track of all of its students' records, which results in a large database. For example, how many students will give equal importance to all subjects, and what types of courses can be used to attract students? Is it possible to forecast the performance of students? What factors influence students' performance? etc. can be extracted from the stored records collected. Due to the exponential growth of databases, an interest in business intelligence and data mining techniques arose to mine such type of information.

The database contains valuable information such as trends and patterns that can be used to improve decision making and increase success rates. Data mining is a technique for extracting relevant information from large databases in order to gain knowledge [1]. To improve student performance, various data mining techniques for analysis, classification, and prediction are available [2]. Classification is a type of supervised learning that is used to categorise a data set based on predefined class labels. Data mining techniques such as SVM, Nave Bayes, and Decision Tree, among others, can be used to build classification models [3]. The alternative is to use automated tools to analyse raw data and extract knowledge for decision makers.

## **II. REVIEW OF LITERATURE**

Several studies have addressed input variables such as gender, age, and performance in the past years, and the proposed system has outperformed traditional allocation procedures. They used a variety of approaches, including neural networks and decision trees (94% combined accuracy) and binary classification (72% accuracy) [4].

The Naive Bayes classification produced the best results. The authors used a regression approach to predict math skills based on individual scores. The majority of students enrol in public schools to receive free education. Like in other countries, there were some core courses that shared a common language. The grading point is scaled from 0 to 20, with 0 being the lowest and 20 being the highest.

Students were evaluated three times during the school year, with the final evaluation coming at the end. [12] employs Nave Bayes Classification to construct a model in which the

probability distribution function is computed to handle continuous data. To improve model accuracy, optimal equal width binning for discretization is introduced. Furthermore, the model classes are balanced to improve accuracy.

Employs [13] two classifiers, namely Nave Bayes and J48, with data from the UCI Machine Learning Repository. The WEKA tool is used to analyse these algorithms, and the accuracy of the models is increased by discretizing continuous features. Many classification algorithms are used in [14], one of which is the NB classifier. The students are divided into four classes: A, B, C, and F, which are labelled. The entire data set is used to construct the classifier, and the Bootstrap method is then used to improve the accuracy of each classifier. The Bayesian Network is used in [15] to classify students based on their grades. The model is built using training data, and the test data is used to compare relative performance. For model evaluation, 10-fold cross validation is used. In [16], a technique is used to predict student performance by combining three classification algorithms, Naive Bayes, 1-NN, and WINNOW, and employing voting methodology.

### **III.MOTIVATION**

This study aims to first implement an automated system that only requires a data set of students and then automatically classifies the students into two classes: pass and fail. Second, developing classification algorithms for educational environments aids in identifying students who require individualised tutoring or counselling from the school. Such classification models can be used by an institution's higher authorities to improve students' performance based on the data set. The proposed system forecasts students' academic performance as well as the factors that influence performance failure.

### **IV.DATA DESCRIPTION**

The data set under consideration contains 395 tuples and 34 attributes [1]. Each tuple represents a student's attribute values or provides information about the student's academic performance and social behaviour.

Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: <i>Gabriel Pereira</i> or <i>Moussão de Sábeira</i> )
address	student's home address type (binary: urban or rural)
Pstatus	parent's colabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4 <sup>th</sup> )
Mjob	mother's job (nominal <sup>8</sup> )
Fedu	father's education (numeric: from 0 to 4 <sup>th</sup> )
Fjob	father's job (nominal <sup>8</sup> )
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: $\leq 3$ or $> 3$ )
famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
traveltime	home to school travel time (numeric: 1 - < 15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour or 4 - > 1 hour).
studytime	weekly study time (numeric: 1 - < 2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours or 4 - > 10 hours)
failures	number of past class failures (numeric: n if $1 \leq n < 3$ , else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 - very low to 5 - very high)
goout	going out with friends (numeric: from 1 - very low to 5 - very high)
Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	current health status (numeric: from 1 - very bad to 5 - very good)
absences	number of school absences (numeric: from 0 to 30)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

Fig. 1 The details of a student which forms the data-set

V.PROPOSED SYSTEM

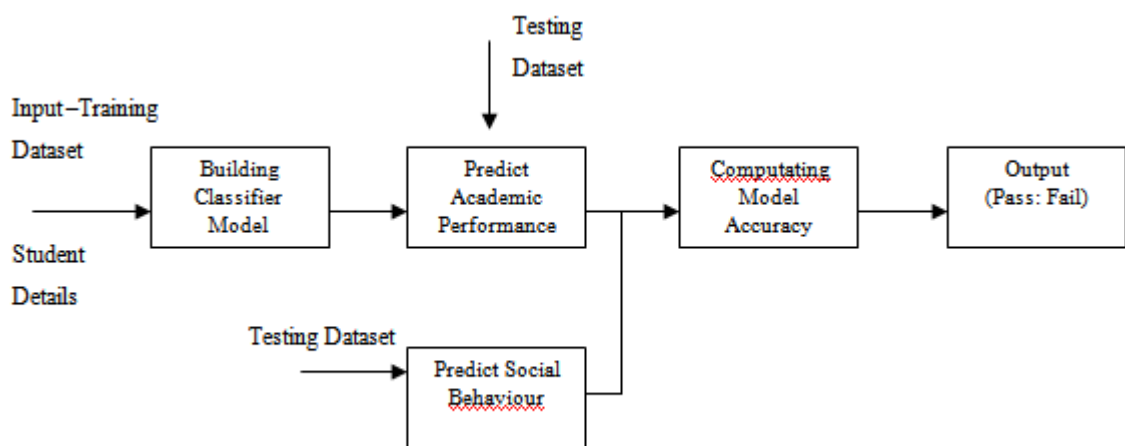


Fig 2: Proposed System

The considered data-set contains some categorical data for a specific number of attributes. All categorical information is converted into binary data as a pre-processing step, with "yes" as "1" and "no" as "0". Following the pre-processing step, the data-set contains 35

attributes for each student and only numerical data. The classifier is then built using the Nave Bayes algorithm to classify the student as pass or fail [3]. The Nave Bayes classifier is applied to the preprocessed training data-set. This data set also includes class labels with pass (class label) labelled as 1 and fail (class label) labelled as 0. All of these steps are part of the training phase. The testing data set is fed into the built classifier for prediction.

The Nave Bayes algorithm divides the students into two groups. There are two classes: pass and fail. Section VI discusses the naive Bayes classification. On the training data-set, this classification is performed. Training is performed to predict whether a student will pass or fail, as well as to predict the other attributes of a student that describe a student's social behaviour. Later, using this model, a prediction is made as to whether a student will pass or fail, as well as a prediction for any attribute given the remaining attributes (predicting social behavior). The classifier's prediction is discussed in Section VII. Furthermore, the confusion matrix is used to calculate classifier accuracy.

## VI.NAIVE BAYES CLASSIFICATION

Classification is the process of separating classes based on extracted features. Following classification, the classes formed will be distinct from one another. The patterns discovered in the training data-set are crucial in the construction of the classifier. Classification algorithms such as k-nearest neighbours, Decision tree learning, support vector machine, naive bayes, and neural networks can be used depending on the needs of an application.

The Nave Bayes classifier is used in the proposed system. Statistical classifiers are Nave Bayes classifiers. This classifier can predict which class a tuple belongs to given a tuple. A Nave Bayesian classifier is based on the Bayes theorem [6] [8]. The Bayes theorem allows us to calculate the posterior probability,  $P(c|x)$ , from  $P(c)$ ,  $P(x)$ , and  $P(x|c)$ . The naive Bayes classifier assumes that the effect of a predictor's ( $x$ ) value on a given class ( $c$ ) is independent of the values of the other predictors. This is known as class conditional independence. The Bayes theorem is stated by

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (1)$$

$$P(c|x) = P(x_1|c) \times P(x_2|c) \dots \dots \dots P(x_n|c) \times P(c)$$

where  $P(c|x)$  is the class's posterior probability for a given attribute.  $P(c)$  is the class prior probability.  $P(x|c)$  is the likelihood, which is the likelihood of the predictor given class.

$P(x)$  is the predictor's prior probability. The naive Bayes classifier assumes that the effect of a predictor's ( $x$ ) value on a given class ( $c$ ) is independent of the values of the other predictors. This is known as class conditional independence. A frequency table is constructed for each feature (attribute) against a specific class in order to calculate the posterior probability. The obtained frequency tables are then converted into probability tables, and the Nave Bayesian (Eq. 1) method is used to compute the posterior probability for each class. The class with the highest posterior probability is the result of this computation. The student with the highest posterior probability will be added to the class. All students are classified using the same criteria. To avoid computing probability values of zero, the Laplacian correction or smoothening is used [7]. Because categorical data will be converted into binary form during the pre-processing step, normal distributions are assumed. The probability density function for a normal distribution is determined by two parameters: mean and standard deviation. The following are the equations for calculating the mean, standard deviation, and normal distribution.

Mean  $(\mu) = \frac{1}{n} \sum_{i=1}^n x_i$  where  $x_i$  is the value of an attribute Deviation from the mean

Where  $x$  represents an attribute value,  $\mu$  is the mean, and  $\sigma$  is the standard deviation. All of these steps are completed prior to the testing phase. During the testing phase, a separate data-set is provided as input. Probability density is calculated using Eq. 1. Depending on the value of the posterior probability, the classifier assigns or predicts a student to the pass (class label assigned as 1 during pre-processing) or fails (class label assigned as 0 during pre-processing) class. The student is assigned to the class with the highest probability density value. This section categorises and predicts the student into two groups: pass or fail. When applied to large databases, these classifiers are also fast. This classifier's accuracy is found to be nearly comparable to that of decision trees and neural networks.

## VII.PREDICTION OF CLASSIFIER FOR STUDENT BEHAVIOR

The classifier predicts student behaviour on the test data-set after the model is built in the training phase [10] [11]. The predictor uses 34 attributes as input and predicts for any remaining attribute. The accuracy of each type of prediction is calculated separately.

**The steps are as follows:**

1. Open the data-set.
2. **Finding the mean and standard deviation:** The mean is the data's central tendency, and it serves as the middle of our normal distribution when calculating probabilities.

For each attribute in a class, the standard deviation is computed value. When calculating probabilities, the standard deviation describes the variation of data spread, which is used to characterise the expected spread of each attribute in our normal distribution.

3. **Data separation:** In the proposed system, data sets are separated by class values.
4. **Summarize the data:** The Naive Bayes model includes information about the data summary in the training data-set. When a test data set is provided as input, this summary is used to make predictions. The classifier returned the mean and standard deviation values for each attribute as a summary. These values will be used to compute the likelihood of an attribute belonging to a specific class.
5. **Making predictions:** Predictions are made in this phase based on the summaries obtained from the training data. Predictions are made by calculating the normal probability density function, which uses the attribute's mean and standard deviation from the training data.

## VIII. RESULTS AND DISCUSSION

The training data set contains 395 student records. Each student has 35 values or characteristics. The students' grades are only for the mathematics subject. The testing data-set consists of 20 students, and it is predicted whether they will pass or fail, as well as any attributes. The model's accuracy is calculated using the Weka tool's confusion matrix.

```

=== Summary ===
Correctly Classified Instances   344      87.0886 %
Incorrectly Classified Instances  51       12.9114 %
Kappa statistic                  0.3404
Mean absolute error              0.2223
Root mean squared error          0.3334
Relative absolute error          79.5622 %
Root relative squared error      89.3755 %
Coverage of cases (0.95 level)  100 %
Mean rel. region size (0.95 level) 98.481 %
Total Number of Instances       395

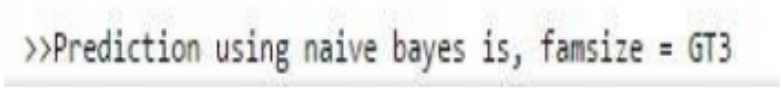
=== Confusion Matrix ===

 a  b  <-- classified as
16 50 | a = no
 1 328 | b = yes

```

**Fig. 3: Nave Bayes Classifier Results Using the WEKA Tool**

Figure 3 shows that out of 395 tuples, 344 are correctly classified and 51 are incorrectly classified. The confusion matrix in Figure 3 correctly classifies 16 students as failing. Fifty failed students are incorrectly marked as passed. 1 passed student is incorrectly labelled as failed. 328 students have successfully completed the course. The classifier model's accuracy was determined to be 87%. The classifier's accuracy can still be improved by selecting the appropriate attributes from the entire data-set as part of data pre-processing. It represents the part of the testing phase, where the user is asked to provide details of the student and the classifier makes the prediction accordingly.



```
>>Prediction using naive bayes is, famsize = GT3
```

**Fig. 4: shows the outcome of the Nave Bayes predictor**

Figure 4 depicts the predictor's output when the user selects family size to be predicted. The famsize attribute has two values: GT3 and LE3. The classifier predicts the attribute value GT3 when the user provides other attribute values.

## IX.CONCLUSION

Education is extremely important in today's generation, and methods for analysing the education system in schools and forecasting institutional advancement are critical. The proposed automated system's emphasis on making student predictions for the advancement of the institution is critical. The proposed automated system focuses on forecasting students' academic performance and social behaviour. The model's accuracy is also calculated. Further scope includes developing a classifier using Support Vector Machine (SVM) and determining which classifier is best suited for classification.

## X.FUTURE WORK

This paper employs the Naive Bayes algorithm; however, future work could employ other classification algorithms. For more accurate results, future work can be done with a large volume of data set.



**REFERENCES**

- [1] Mrinal Pandey, Vivek Kumar Sharma. “A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction”. *International Journal of Computer Applications* (0975 – 8887), Volume 61– No.13, January 2013.
- [2] Shaeela Ayesha and et al,” *Data Mining Model for Higher Education System*”, *European Journal of Scientific Research*, Vol.43 No.1 (2010), pp.24-29.
- [3] Bhardwaj, “Data Mining: A prediction for performance improvement using classification”. *International Journal of Computer Science and Information Security*. Volume 9(4). (2011). .Bekele, R., Menzel, W. “A bayesian approach to predict performance of a student (BAPPS): A Case with Ethiopian Students”. *Journal of Information Science* (2013).
- [4] Ahmed, A. B. E, Ibrahim S. E. "Data Mining: A prediction for Student's Performance Using Classification Method." *World Journal of Computer Application and Technology* Volume 2(2) (2014).
- [5] P. Cortez and A. Silva. “Using Data Mining to Predict Secondary School Student Performance”. In A. Brito and J. Teixeira Eds., *Proceedings of 5th Future Business Technology Conference*.
- [6] Meghna Khatri. “A Survey of Naïve Bayesian Algorithms for Similarity in Recommendation Systems”. *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 2, Issue 5, May 2012.
- [7] P. Domingos and M. Pazzani. “On the optimality of the simple Bayesian classifier under zero-one loss”. *Machine Learning*, 29:103–130, 1997.
- [8] Bekele, R., Menzel, W. “A bayesian approach to predict performance of a student (BAPPS): A Case with thiopian Students”. *Journal of Information Science* 2013).
- [9] Sonali Agarwal, G. N. Pandey, and M. D. Tiwari, *Data Mining in Education: Data Classification and Decision Tree Approach*, 2012.
- [10] E.Chandra and K. Nandini, ”Predicting Student Performance using Classification Techniques”, *Proceedings of SPIT-IEEE Colloquium and International Conference*, Mumbai, India, p.no, 83-87.
- [11] S. Huang, & N. Fang, *Work in Progress - Prediction of Students’ Academic Performance in an Introductory Engineering Course*, In *41st ASEE/IEEE Frontiers in Education Conference*, (2011), 11–13.

- [12] Syed Tanveer Jishan, Raisul Islam Rashu, Naheena Haque and Rashedur M Rahman. “Improving accuracy of students’ final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique”, Springer Open Journal, (2015).
- [13] Kayah, F. “Discretizing Continuous Features for Naive Bayes and C4. Classifiers”. University of Maryland publications: College Park, MD, USA.
- [14] S. Taruna, Mrinal Pandey, “ An empirical analysis of classification techniques for predicting academic performance”, IEEE Advances Computing Conference (2004).
- [15] Varsha Namdeo, Anju Singh, Divakar Singh and Dr. R.C Jain, “Result Analysis Using Classification Techniques”, International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 22. (2010)
- [16] S. Kotsiantis, K. Patriarcheas, M. Xenos, “A combinational incremental ensemble of classifiers as a technique for predicting students’ performance in distance education”, Knowledge-Based Systems 23 Elsevier, 529–535, (2010)