# Performance Prediction using Data Mining Techniques: A Study

**Mohammed Adnan[1] Umar Farooq[2]**
[1,2]Department of Computer Science and Engineering
[1,2]P.E.S Institute of Technology and Management, Shivamogga, Karnataka, India

*Abstract*— Predicting student performance is important to make at university to prevent student failure. The drop-out number is one of the parameters that can be used to measure student performance and one important point that should be considered in the accreditation of a university. Data Mining is widely used to predict student performance, as well as data mining used in the field commonly referred to as Educational Data Mining. This study enabled Feature Selection to select high-quality attributes for student performance in the Department of Engineering in various universities. Subsequently, algorithms of two popular categories were used, the Bayesian Network and the Decision Tree, and they were compared to know the best predictive effect. The result revealed that the student attendance and GPA in the first semester were the highest in all Feature Selection methods, and the Bayesian Network is more successful than Decision Tree because it has a higher level of accuracy.

***Keywords:*** Student Performance, Prediction, Feature Selection, Data Mining

## I. INTRODUCTION

Data mining is the process of automatically retrieving useful information from large data repositories [1]. Data Mining techniques are used to obtain logical patterns and rules from large amounts of anonymous data. There are many strategies, methods and rules in Mining Data for extracting a specific database. Recently, the data mine has been widely used in many applications, such as customer relations, engineering, medicine, marketing, education, etc. There is a growing number of studies that have used data mining in education, and the data mining process used in this file is called Education Data Mining. Education Data Mining can be done in many ways such as Decision Tree, Neural Network, Naïve Bayes, Ne-Nearest neighborhood, and many more [2].

Education Data Mining has been widely used in many universities. Implementing Mining Data can help universities predict their student performance. Universities today operate in a highly competitive environment and have a very complex system. It is very important that today's universities critically evaluate their performance in order to develop future strategies and actions. University administrators should view their student performance as one of the key keys to a marketing campaign and look at the promising power of the university [3]. Guessing student performance can help predict a student's vulnerability and help university administrators make strategies and decisions related to student performance improvement.

University administrators can take the initiative to prevent student failure or dropout. After being able to predict, it is expected of weak students that they can improve their performance and gain better points. Student performance can be used depending on a number of factors such as personal, psychological, social, and other factors [4] [5]. This type of analysis can help to identify a weak student and help university administrators to make strategies and decisions related to student performance improvement. University administrators can take the initiative to prevent student failure or dropout. After being able to predict, it is expected of weak students that they can improve their performance and gain better points.

Among the major data mining techniques, one of the oldest and most widely used methods of classification is [6] [7]. Tan et al [1] define segregation as the function of distributing objects in one of the aforementioned categories. It uses a set of pre-made models to create a model that can distinguish a large number of records [8]. Separation methods are often used to predict student performance [2] [3] [5] [6]. The Bayesian Network and Decision Tree are two popular planning methods that have been used by many experts to make predictions for student performance [3] [9] [10] [11] [12] [13] [14]. Those algorithms are powerful class algorithms and the rules are easy to interpret.

The current study will compare the Bayesian Network with the Decision Tree and whether the student will resign (DO) or not from the Department of Industrial Engineering Universitas Islam Indonesia in terms of specific characteristics. Feature selection is done first before using the separation algorithm to select the most powerful attributes. This study aims to determine the best student performance profile in the Department of Industrial Engineering Islam Indonesia by conducting a feature section and comparing the Bayesian Network with the Decision Tree. The rest of the paper is structured as follows, Section 2 presents the literature review, Section 3 presents the research methods related to this study, Section 4 presents the results and discussion. The conclusion is finally made in Section 5.

## II. RELATED STUDIES

### A. Decision Tree

Tree choice is another popular method of separation. Decision tree, as shown in Fig. 1, is like the shape of a tree that flows like a stream. The tree consists of a root node (the first oval node shown), the inner nodes can have two or more types (the middle node is oval-shaped), and the leaf nodes (shown in rectangle). All internal nodes contain partitions, which check the number of attribute adjectives.

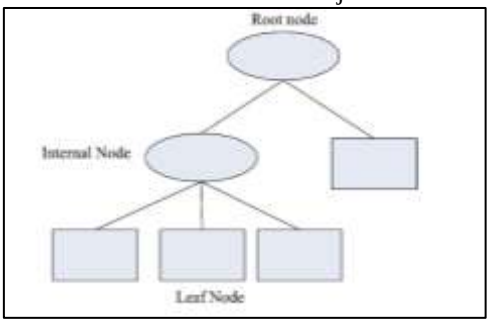


Fig. 1: Decision Tree [1]

Yadav & Pal [2] used the techniques of Educational Data Mining Techniques to create a predictable model for

 

student performance in engineering. CART, ID3, and C4.5 drug decisions are made to predict their final test performance. The result of the decision tree can be used to predict that students will be able to pass, fail, or be promoted next year.

The data for this study was compiled for a 2010 session at VBS Purvanchal University in Jaunpur (Uttar Pradesh), Institute of Engineering and Technology. Initial data contains 90 students. The variables used in this study were 13 and used class results for the final test. Details are made in the WEKA software. The result showed that although Classification and Regression Tree (CART), ID3, and C4.5 algorithms provided an acceptable level of accuracy, the C4.5 process outperformed the other by 67.77% accuracy compared to other methods.

Yadav et al [12] also used decision-making medicine (CART, ID3, and C4.5) to predict the performance of students in the Computer Applications department actually Master of Computer Applications from the 2008 to 2011 session at VBS Purvanchal University Jaunpur (Uttar Pradesh), India. The original size has 48 records. The variables used were previous semester marks, class test scores, meeting performance, distribution, presence, web function, and phase flexibility, using end semester marks (first, second, third, failed). The result showed that CART was leading with 56.25% accuracy compared to other strategies, but the ID3 algorithm also showed an acceptable level of accuracy [2] [5] [14].

Along with the previous author, Ahmed & Elaraby [8] have also attempted to use Educational Data Mining to conduct student performance analysis. The decision tree (using the ID3 algorithm method) was used in this study. The database was collected from the Department of Information student database from session 2005 to 2010, and the initial data contained 1547 records. The variables included in this study were student departments, high school degrees, intermediate marks, laboratory distance, conference performance, distribution, student participation rate, homework, and the target class. The Midterm brand has great benefits; then, it was used as the original root zone. The study concluded that conducting a student analysis can help students improve student performance. In addition, it can identify students who need special care to reduce the failure rate and take appropriate action at the right time.

### *B. Bayesian Network*

The Bayesian Network is like a mathematical system that combines prior class knowledge with new data obtained from data [1]. The numerical value of this method is shown in figure (1).

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

X and Y are a few random attributes. The probability of joining X and Y, P (X = x, Y = y), is likely that variable X will receive a certain value x, while variable Y will receive value y. The conditional probability is the probability that a random variable will take a certain amount given that the effect of another random variable is known.

There have been many studies using the Bayesian Network as a means of differentiation. Baradwaj and Pal [5] worked on the study to predict student performance. The subject of this study is 600 students from five different Computer Application colleges for Bachelor's degree courses, Dr. R. M. L. Awadh University Faizabad, India. This study uses the Bayesian Network as a classification and includes 17 attributes. The results show that the student's grade in high school examinations, location, teaching media, mother's qualifications, other student practice, annual family income, and student family status was highly correlated with student performance.

Along with Baradwaj and Pal [5], Pandey and Pal [15] also conducted the study using the Bayesian Network section, language, and background. The study concluded whether new emerging students would sing or not. Millian et al [12] also attempt to use the Bayesian Network to model student engineering. Ramaswami and Rathinasabapathy [16] also used the Bayesian Network approach in the field of education to predict student performance.

They performed data analysis of high school students including 5650 records with 35 marks, and class data varied in five cases. In this study, they used feature selection before using predictive models. Data analysis for this study was performed using the WEKA software. This study used a selected simple guessing algorithm, such as Hill Climbing, K2, LAGD Hill Climbing, Repeated Hill Climbing, Tabu Search, and Network Augmented with Tree, to find conditional tables for Bayes Network. From the result, it can be concluded that Bayesian Network models with Network Augmented with Tree search algorithm surpassed other types of search algorithms for all types of class values. The result also found that the flexibility of the HScGrade category is closely related to the level obtained at the second level, the type of transportation to school, between teaching, sibling structure (number of brothers and sisters), and family economic status.

### *C. Bayesian Network Contrast with Decision Tree*

There have been many experts trying to compare the performance of the Bayesian Network with the Decision Tree in predicting student performance. Al-Radaideh et al [9] attempted to predict the performance of the Technology & Computer Science Faculty, Yarmouk University, Jordan students who have studied Programming I (C ++). This paper also attempted to compare ID3 algorithms, C4.5 algorithms with Naïve Bayes.

There were 12 attributes, including gender, age, department, high school, high school grade, type of education, sponsorship, accommodation, teacher qualifications, teacher's gender, faculty department, repetition rate and C ++ (A) grade, B, C, D). WEKA has been used to analyze data. The high school level was the most profitable, so this quality was considered the root of the decision tree. From the result it can be concluded that the precision of the division of the three different division algorithms was not so high. It was found that the samples and symbols used in this study were insufficient to create a high-quality classification model.

Kabakchieva [3] presented the first results of a study conducted at the University of Bulgaria. The main purpose of the research conducted by Kabakchieva [3] was to determine whether there were enough data patterns to predict student

 

performance at a university based on individual and pre-university characteristics and symptoms. Nearby neighborhoods, decision tree segments, and Bayes dividers were used to analyze the data. The analysis is done on the WEKA software because it is widely used for research purposes in data mining and is freely available to the public.

Preliminary data include 10330 students and 20 qualifications. The final results showed that the division of the decision tree (J48) was superior to other divisions because it was more accurate, followed by the law (JRip) and the kNN editor, while the Bayes dividers were less accurate than the others. However, all classified items were shown to have an absolute accuracy of less than 70% which means that the error rate was high and speculation could not be trusted.

## III. Analysis Method

This study consists of three steps. The first step is to clean the data, the second step is to select a feature and the last step is to analyze it in stages. Data purification is an important step that must be done before analyzing data to identify lost, noisy and inconsistent data.

The second step is the selection of the feature. The feature selection algorithm can be determined as a combination of a search method to elevate the subsets of a new feature, as well as a test rating that finds different subsets of the feature [17]. The main purpose of feature selection is to select the lower set of input features by removing the features, which are unimportant or that provide little predictive detail [16]. This step needs to be used to define high variables that influence student performance predictions.

Anuradha and Velmurugan [17] and Ramaswami and Bhaskaran [16] used a variety of methods to select feature selection in their study, and the five most popular filtering methods with Rank setting or Greedy search method are a test based on Correlation-based Attribute, Gain-Rating of adjective test, Assessment of Qualifications of Information, Analysis of the Auxiliary Perspective, and the Comprehensive Feature of Uncertainty. Those methods have been used in the present study.

The final step was a split analysis using the Bayesian Network and the Decision Tree. The student performance parameter used in this study is whether the student will be a DO or not. The performance parameter for comparing both algorithms used in this study is the Accuracy Rate. For experimental purposes, WEKA is a software that opens open data sources freely used to use the algorithm.

## IV. Result and Discussions

### A. Data Cleaning

Details of this study were collected from a student database that can be accessed through the Islamic Indonesian University (UNISYS) information system. Preliminary data contained 178 student data for the 2007 academic year and after refining the data, only 104 set data was obtained with 12 marks (including class attribute, exit or not) while 13 data (12.5%) was categorized as quit school, with 91 details (87.5%) included in the category. The attributes are divided into four categories and are shown in Table 1.

| Category | Attributes |
|---|---|
| Personal Information | Gender (GE)<br>Origin (OR) |
| Family information | Father education (FE)<br>Father occupation (FO)<br>Mother education (ME)<br>Mother occupation (MO) |
| Pre-university characteristic | Senior high school type (ST)<br>Senior high school department (SD)<br>Senior high school final grade (SF) |
| University features | first semester Attendance (AT)<br>first semester GPA (GPA)<br>Drop out (DO) or Not (as the Class attribute) |

Table I: Student Data Set

### B. Feature Selection

Feature Selection has identified 12 attributes that have a significant impact on student performance. Feature Selection results from the five paths are shown in Table 2.

| Feature Selection Method | Selected Attributes |
|---|---|
| Correlation-based Attribute evaluation | MO, **AT, GPA** |
| Gain-Ratio Attribute evaluation | **GPA, AT**, SD, GE, FO, MO, ME, OR, **ST, FE, SF** |
| Information-Gain Attribute evaluation | **GPA, AT**, SD, GE, FO, MO, ME, OR, **ST, FE, SF** |
| Relief Attribute evaluation | **AT, GPA**, FO, SD, **FE**, MO, GE, ME, OR, **ST, SF** |
| Symmetrical Uncertainty Attribute | **GPA**, AT, SD, GE, FO, MO, ME, OR, **ST, FE, SF** |

Table II: Feature Selection Result

Combined attribute analysis is a greedy search method that selects only the most influential attribute of the predictive attribute. Based on the Attribute contribution test, MO, AT, GPA, have a high impact on student performance. Other options are the Rank search method that sets attributes ranging from high influences to attributes with low impact. The Gain-Ratio, Information-Gain Attribute, and Symmetrical Unriberty Attribute tests have exactly the same results, while the Relief Attribute test has a slightly different effect. From the whole result, it can be concluded that the two attributes from the first list in almost all selection methods are GPA and AT. It means that the first semester of GPA and meeting attendance have a major impact on whether a student will do it or not. By finding the lowest influence the four search methods of Rank show similar results. ST, FE, SF is below the third rank from Gain-Ratio Attribute, Information-Gain Attribute, and Symmetrical Uncertainty Attribute, while Relief Attribute Assessment FE is moderate. Depending on this result, the attributes to be used in further analysis of variance are GPA, AT, SD, GE, FO, MO, ME, OR because many methods show the same effect. This classification analysis will be compared to an analysis that includes FE attributes as indicated in the assistive attribute test results.

### C. Classification Analysis

In the analysis of the three-dimensional performance comparison comparisons between the Bayesian Network and the Decision Tree algorithm to predict whether a DO has been performed or not. The first case used all attributes, the second case compared both algorithms using selected attributes from the Feature Selection result including FE (GPA, AT, SD, GE, FO, MO, ME, OR, FE), and the last state used only attributes highly selected (GPA, AT, SD, GE, FO, MO, ME, OR). Table 3 shows the Separation Effect from certain conditions.

| Scenario | Attributes Used | Classification Algorithm | Accuracy Rate | Number of Incorrectly Classified Instance |
|---|---|---|---|---|
| 1 | All | Bayesian Network | 95,19% | 5 |
| | | Decision Tree | 93,27% | 7 |
| 2 | GPA, AT, SD, GE, FO, MO, ME, OR, FE | Bayesian Network | 97,11% | 3 |
| | | Decision Tree | 94,23% | 6 |
| 3 | GPA, AT, SD, GE, FO, MO, ME, OR | Bayesian Network | **98,08%** | **2** |
| | | Decision Tree | 94,23% | 6 |

Table III: Classification Result

From the test result, it can be seen that within the same algorithm, 1 condition that includes all attributes has a lower accuracy level than condition 2 and 3 which only includes selected attributes from the feature selection. After that, it can be proved that Feature Selection can improve the accuracy of the prediction model. And it can be seen that condition 2 has a lower degree of accuracy than condition 3. It means that it is better to exclude the FE attribute in the analysis. In the comparison of the division algorithm, it can be seen that in all cases the Bayesian Network shows a violation of the Decision, and from the rest of the study, predicting the Bayesian Network using selected attributes GPA, AT, SD, GE, FO, MO, ME, OR to others.

## V. CONCLUSION AND SUGGESTION

Predicting student performance is important to make at university to prevent student failure. The drop-out number is one of the parameters that can be used to measure student performance and one important point that should be considered in the accreditation of Indonesian universities. Of course, this number should be reduced. Other factors that have a high impact on student prediction that can be considered by university administrators to reduce student outcomes based on the results of this study are First Semester GPA, First Semester, Department of High School, Gender, Father's Work, Mother's Work, Mother's Education, and Origin. It can also be proved in this study that the selection of a feature can increase the accuracy of the prediction, and the Bayesian Network is implementing the decision because it has a high degree of accuracy. After all, it is a good idea to speculate whether a student will resign from the Department of Industrial Engineering in Islamic Indonesian University exerts a high influence in selecting attributes from feature selection and using the Bayesian Network as a classification algorithm.

## REFERENCES

[1] Tan, P. N., Steinbcah, M, Kumar, V., (2006) Introduction to Data mining, Introduction to Data mining. MA: Pearson Education.
[2] Yadav, S. K. and Pal, S., (2012), Data Mining : A Prediction for Performance Improvement of Engineering Students using Classification, World of Computer Science and Information Technology Journal WCSIT, 2(2), 51–56.
[3] Kovatchev, D., (2013), Predicting student performance by using data mining methods for classification, Cybernetics and Information Technologies, 13(1), 61–72.
[4] Hijazi, S. and Raza-Naqvi, S. M., (2006), Factors Affecting Students' Performance, 3(1), 1–10.
[5] Bhardwaj, B. K. and Pal, S., (2011). Data Mining: A prediction for performance improvement using classification, (IJCSIS) International Journal of Computer Science and Information Security, 9(4).
[6] Romero, C., Ventura, S., Espejo, P. G. and Hervás, C., (2008), Data Mining Algorithms to Classify Students, EDM, 8–17.
[7] López, M. and Luna, J. M., (2012), Classification via clustering for predicting final marks based on student participation in forums, Proceedings of the 5th International Conference on Educational Data Mining, pp. 4–7.
[8] Ahmed, A. B. E. D. and Elaraby, I. S., (2014), Data Mining: A prediction for Student's Performance Using Classification Method, World Journal of Computer Application and Technology, 2(2), 43–47.
[9] Al-Radaideh, Q. A., Al-Shawakfa, E. M. and Al-Najjar, M. I., (2006), Mining Student Data Using Decision Trees, International Arab Conference on Information Technology.
[10] Garcia, P., Amandi, A., Schiaffino, S. and Campo, M., (2007), Evaluating Bayesian networks' precision for detecting students' learning styles, Computers and Education, 49(3), 794–808.
[11] Dekker, G. W., Pechenizkiy, M. and Vleeshouwers, J. M., (2009), Predicting students drop out: A case study, EDM'09 - Educational Data Mining 2009: 2nd International Conference on Educational Data Mining, 41–50.
[12] Millán, E., Loboda, T. and Pérez-De-La-Cruz, J. L., (2010), Bayesian networks for student model engineering, Computers and Education. Elsevier Ltd, 55(4), 1663–1683.
[13] Kumar, S. A. and Vijayalakshmi, M. D., (2011), Efficiency of Decision Trees in Predicting Student' Academic Performance, © Cs & It-Cscp 2011, 335–343.
[14] Yadav, S., Bharadwaj, B. and Pal, S., (2012), Data mining applications: A comparative study for predicting student's performance, International Journal of Innovative Technology & Creative Engineering, 1(12), 13–19.
[15] Pandey, U.K. and Pal, S., (2011), Data Mining: A prediction of performer or underperformer using classification, International Journal of Computer Science and Information Technologies, 2 (2), 686-690
[16] Ramaswamy, M. and Bhaskaran, R., (2009), A Study on Feature Selection Techniques in Educational Data Mining, 1(1), 7–11.
[17] Anuradha, C. and Velmurugan, T., (2016), Feature Selection Techniques To Analyse Student Acadamic Performance Using Naïve, The 3rd International Conference on Small & Medium Business, 345–350.

 