ISSN 2278-3091
Volume 9, No.1, January – February 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse108912020.pdf
https://doi.org/10.30534/ijatcse/2020/108912020

IT Students Selection and Admission Analysis using Naïve Bayes and C4.5 Algorithm

Jovanne C. Alejandrino[1*], Allemar Jhone P. Delima[2], Ramcis N. Vilchez[3]
[1,2]Professional Schools, University of Mindanao, Matina Davao City, Philippines,
[3]College of Computing Education, University of Mindanao, Matina Davao City, Philippines,
jovanne_alejandrino@umindanao.edu.ph, allemardelima@umindanao.edu.ph,
ramcis_vilchez@umindanao.edu.ph

ABSTRACT

Admission to college and selection of applications have probably become an integral part of some colleges and universities in their enrolment process, yet it is girded by controversy and skepticism. A new area of research that uses techniques of data mining is known as Educational Data Mining. It incorporates machine learning algorithms and statistical methods to help for the interpretation of student's learning habits, academic performances, and further improvements- if needed. This paper focuses on the predictive values of certain academic variables, admission tests, high school academic records as related to the performance of Information Technology (IT) students at the end of the first year. For this reason, 221 data were used, and C4.5 and Naive Bayes algorithms are applied to generate a prediction on the students' performance. The C4.5 classification gained 98.64% in 10-folds cross-validation and 96.97% in the 70% training and 30% testing percentage split compared to Naïve Bayes which only gained 89.14% and 86.36% for both 10-folds cross-validation and 70% training and 30% testing percentage split respectively. The comparative analysis of the result shows that senior high school track and academic data and admission test results are the influential attributes to the performance of IT students in their first year. This paper recommends for future studies to add different data from different years to increase the accuracy of the prediction.

Key words: Academic Performance, Admission Test, Educational Data, Naïve Bayes

1. INTRODUCTION

Twentieth-century is called the "Age of Human Capital," where education, skills, and acquisition of knowledge become the basis of a country's progress [1]. Tertiary education is one of the driving forces that produce human capital and, therefore, one of the contributors to a country's sustainable economic growth[2]. The vital goal of higher education institutions is to provide the best and quality education to its students [3].

In most of the developed countries, they have established their educational system where the education is not limited to the four-cornered room but even goes online. The online-based education system has been widely used, like the Online Education System, the Massive Online Open Course system, the Project-based learning, and many others [4]. A lot of data ranging from student admission, student family data and student academic data has been generated. These data must be analyzed to extract valuable information to provide quality education [3][4].

As the data grow larger, it is impossible to obtain useful information from the data manually [5]. To utilize the rapid growth of data collected, knowledge discovery in databases (KDD), often called data mining, is used to analyze educational data in the sighting of useful information that will be useful to produce decisions and knowledge to the institution [3][6]. Every educational institution strives to maintain an accurate, helpful, and efficient educational process for the improvement of its services and the quality of the education they provide to the clientele [4].

Data mining is a process used to obtain useful information from the historical data in the database. It has got a productive focus due to its significance in decision making, and it has become an essential component in several industries, including the education sector [3]. Data mining in education is called Educational Data Mining (EDM). EDM, as defined by the International Educational Data Mining Society, "an emerging discipline concerned with developing methods for exploring the unique types of data that come from educational settings and using those methods to understand students better and the settings in which they learn in" [7].

Data Mining is one of the best computer-based smart tools used to check the performance of the students. Data mining fills the knowledge gaps in the higher education system [4]. Much knowledge results in the use of data mining techniques from the students' findings. These facts will be about student behavior, student's interest both in curricular and extracurricular and will serve as a guide to improve student academic performance. These improvements further are useful, such as maximizing the student retention rate, improvement ratio, learning outcome, and minimizing the students' dropout and failure rate that could result in the reduction of cost in the education system [4].

Recently, the Institute of Information Technology (IIT) of Davao del Norte State College (DNSC), from open admission, screened the students to enter the course based on their admission test result. Aside from the admission test result, students were also screened according to their general high school average, national certifications, senior high school academic track, and family economic status. With the newly established admission policy of the institute, questions have arisen about the admission criteria of the program. Is IIT selecting the cream of the crop of applicants, and to what extent does an admission's decision variable predict the student performance?

The objective of this paper is to determine the relationship between the students' attributes such as senior high school track and academic data, and admission test result and their performance in the BSIT program during their freshman year. Also, this study aims to predict student academic performance based on the afore-mentioned admission criteria. It will serve as a guide for the institute to accept the student to the program or advise to take other courses using some data mining techniques.

2. RELATED LITERATURE

Data are symbols that describe the attributes of objects and events. Processing the data to increase its usefulness is called information. Knowledge, on the other hand, is carried by instructions out from the information given [8]. Knowledge Discovery from data (KDD) or referred to as data mining, is the process of discovering and extracting relevant information out from the bulk of data collected [9]. Data mining in the education sector is called Educational Data Mining (EDM). EDM, as described by the International Educational Data Mining Society, "an emerging discipline concerned with developing methods for exploring the unique types of data that come from educational settings and using those methods to understand students better and the settings in which they learn in" [7].

A study on academic performance conducted by [10] using the students' educational and demographic data of 932 students. Decision tree classification was used, and it was found out that socio-demographic variables like marital status, social stratum, whether the student takes day or night classes, gender, and the number of siblings influence the academic performance of a student.

In another study conducted [7] on predicting low academic performance at a specific enrolment using data mining, the data of 3200 students admitted to the two programs from the Universidad Nacional de Colombia, Bogotá campus are used. The admission dataset fields are grouped into three categories. The first group, is the initial academic information, which includes the high school type and course choice upon admission. Second, the demographic data such as age at admission, gender, city of origin, socio-economic classification, and ethnicity. Lastly, the academic potential that includes the admission test score in five modules (i.e., Sciences, Math, Image, Text, and Social studies) and classification levels for Basic Math and Literacy. Using both Naive Bayes and Decision Tree classifier, it was found out that Naive Bayes results are better on predicting the test set.

There are three hundred sixty datasets from the Learning Management System used in the study of [11]. Data such as gender, nationality, educational stage, grade level, section, and psychological features are used in the study. The study is all about the Analysis of Education Data Mining using Classification, where J48, Support Vector, Random Forest, Naïve Bayes, Multilayer Perceptron machine algorithms are used as a comparative experiment to determine the best classifier. The study found out that Multilayer Perceptron has the best performance among other classifiers.

The study of [12] found out that among the general pointed average of the five subjects namely, Math, Physics, Computer, and English, Math's general pointed average (GPA) has the most influential for academic performance in computer programming subjects using the C4.5 algorithm. There were 474 students' data used in the study consisting of the second, third, fourth year, and fifth-year students studying in the computer education field at Nakhon Ratchasima Rajabhat University, Thailand.

A related study conducted by [13] about the Student Academic Evaluation using Naïve Bayes indicated that accuracy(AC) of the algorithm used in the study is AC of 76.79% with true positive rate (TP) of 44.62% by using quality training data of 80% and 90% have a good performance accuracy value. They used 279 samples of students' academic performance evaluation variables including age, place of birth, gender, high school status (public or private), department in high school, organization activeness, age at the start of high school level, and progress GPA and total GPA from semester 1-4 in the study.

The table below shows the summary of the literature review findings, which includes the problem, used machine learning algorithms, and the accuracy of the algorithm used.

Table 1: Related Works

Research Authors	Problem Studied	Used Data Mining Techniques	Accuracy
Sandra Rubiano and Jorge Garcia [10]	Formulation of a productive model for academic performance based on students' academic and demographic data	• Decision Tree	• 66%

Camilo E. López, Elizabeth León Guzmán, and Fabio A. González [7]	A Model to Predict Low Academic Performance at a Specific Enrolment Using Data Mining	• Decision Tree • Naïve Bayes	• 75% • 85 %
Chitra Jalota and Rashmi Agrawal [11]	Analysis of Educational Data Mining using Classification	• Random Forest • Naïve Bayes • Multilayer Perceptron Support • Vector Machine • DT - J48	• 7.40% • 64.40% • 76.07% • 75.40% • 73.60%
Pensri Amornsinlaphachai [12]	Efficiency of data mining models to predict academic performance and a cooperative learning model	• Artificial Neural Network, • K-Nearest Neighbor • Naive Bayes • Bayesian Belief Network • JRIP • ID3 • C4.5	• 65.30% • 68.80% • 71.10% • 72.80% • 73.5% • 66.2% • 74%
Haviluddin, Nataniel Dengen, Edy Budiman Masna Wati, Ummul Hairah [13]	Student Academic Evaluation	• Naïve Bayes	• 76.79%

3. METHODOLOGY

Predicting the academic performance of a student needs many factors to be measured. Prediction models that include personal, socioeconomic, demographic and previous academic variables are needed for the effective prediction of the performance of the students.

3.1 Data Preparation and Pre-processing

The dataset used in this paper was obtained from the Office of the Registrar and Guidance and Testing Office of Davao Del Norte State College. The data were from the first-year Bachelor of Information Technology (BSIT) students of the academic year 2018-2019. The students were the first graduates of the K-12 curriculum and who completed the entire year in the BSIT program. There were 285 first-year students enrolled in the second semester. However, 36 records were removed for those students who did not undergo senior high school and those students who were transferees. Only 257 were prepared for data mining, and 28 students were eliminated from the data set for those students who have incomplete grades, which left only 221 clean records.

3.2 Data Selection and Transformation

In this step, only those required fields were selected that are needed in data mining. The data was collected through the cumulative form filled by the student at the time of admission, the senior high school report card, and the grades in the major subjects for the entire first year. The data collected in the study undergo a pre-processing activity aiming to guarantee that data is appropriate, well-formatted, and complete. All the predictor and class variables used in the study are given in table 2 for reference.

Table 2: Student Variables

Variables	Variable Description	Possible Values
sex	Student Sex	1- Male 2- Female
status	Student Civil Status	1- Single 2- Married 3- Widowed 4- Divorced
address	Student Permanent Address	1- Within the city (where the school is situated) 2- Outside the City
age	Student Age	Numeric values
citizenship	Student Citizenship	1- Filipino 2- Foreigner
Economic Status	Family economic status	1- Low 2- Average 3- High
Shs Strand	Student Senior High School Strand	1- IT-Related 2- Non-IT Related
Course FirstChoice	Student first course choice	1. BSIT 2. Other course
Shs GPA	Student senior high school General Weighted Average	{1.0 – 100, 97, 98, 1.25- 97, 96, 95, 1.50- 94, 93, 92, 1.75- 91, 90, 89, 2.0- 88, 87, 86, 2.25- 85, 84, 83, 2.50- 82, 81, 80, 2.75- 79, 78, 77, 3.0- 76, 75, 5.0- 75 below}
Verbal	Student Otis-Lennon School Ability Test (OLSAT) verbal score.	Numeric {1-9}
Non-Verbal	OLSAT non-verbal score	Numeric {1-9}
Stanine	OLSAT overall score	Numeric {1-9}
Shs Math	The average grade of the Math-related subjects in senior high school (General mathematics and Statistics and probability).	{1.0 – 100, 97, 98, 1.25- 97, 96, 95, 1.50- 94, 93, 92, 1.75- 91, 90, 89, 2.0- 88, 87, 86, 2.25- 85, 84, 83, 2.50- 82, 81, 80, 2.75- 79, 78, 77, 3.0- 76, 75, 5.0- 75 below}
Shs Science	The average grade of the Science-related subjects in senior high school (Physical Science and Earth Science)	{1.0 – 100, 97, 98, 1.25- 97, 96, 95, 1.50- 94, 93, 92, 1.75- 91, 90, 89, 2.0- 88, 87, 86, 2.25- 85, 84, 83, 2.50- 82, 81, 80, 2.75- 79, 78, 77, 3.0- 76, 75, 5.0- 75 below}
Shs English	The average grade of the	{1.0 – 100, 97, 98,

	English-related subjects in senior high school (Oral Communication and Reading and Writing)	1.25- 97, 96, 95, 1.50- 94, 93, 92, 1.75- 91, 90, 89, 2.0- 88, 87, 86, 2.25- 85, 84, 83, 2.50- 82, 81, 80, 2.75- 79, 78, 77, 3.0- 76, 75, 5.0- 75 below}
Standing	Student Standing	Regular Irregular

3.3 Implementation of Data Mining

Classification is a technique where data is categorized into a given number of classes, and its primary goal is to find a class that has similarity to the categorized data [14]. Waikato Environment for Knowledge Analysis (WEKA) is one of the tools to implement classification. WEKA is a fre e software that is widely used in the machine learning platform. It has a collection of machine learning algorithms that can be used in the application of data mining [15]. The data then loaded to the WEKA Explorer. Naïve Bayes and C4.5 algorithms are used in the classification to assess the accuracy of the resulting predictive model and to visualize erroneous predictions. The validation technique used in this paper is 10-fold cross-validation and 70% training and 30% testing percentage split validation models.

3.4 Naïve Bayes Classifier

The student performance was predicted using the data mining method named classification rules. Naïve Bayes classifier (NBC) was used to predict student performance of the first year BSIT students based on the senior high school academic record, admission test result, and students' personal, socioeconomic, and demographic data. NBC algorithm is a probabilistic classifier that applies Bayes theorem- a theory predicts the future instances based on past experiences. Each attribute is dependent on each other and will contribute to the decision-making with equally weight attributes [13].

The formula (1) is Naïve Bayes rule, which is all about posterior and before one attribute, which posterior checks how one feature like xi has the probability that comes underneath the class of any labels (hi) [14].

$$P(h_i|x_i) = P(x_i|h_i)\ P(h_i)\ /\ P(x_i|h_i)+P(x_i|h_2)\ P(h_2) \quad (1)$$

Naïve Bayes theorem determines the independence between the values of attributes. NB can be both predictive and descriptive algorithm which means, the probability is descriptive and then use to predict target is called a predictive algorithm [16].

3.5 C4.5 Classifier

C4.5 algorithm is a popular algorithm used to generate decision trees. C4.5 algorithm is called a statistical classifier because of the decision trees generated by the algorithm can be used for classification [17].

4. RESULT AND DISCUSSION

In this paper, C4.5 and Naïve Bayes algorithm classification models are used for the training and testing data using WEKA. The comparison of the two algorithms is tested on the data set, and the result shows which algorithm performed better through the performance table below. Table 3 shows the algorithm test performance results with the accuracy, precision for regular and irregular students, area under the curve (AUC), Recall, and F-Measure using 10-folds cross-validation and 70% training and 30% testing percentage split.

Table 3: Results of C4.5 and Naïve Bayes Algorithm

Criteria	Cross-validation (10 folds)		Percentage Split (70%)	
	C4.5	NBC	C4.5	NBC
Accuracy	98.64%	89.14 %	96.97%	86.36%
Precision for Regular	0.981	0.946	0.951	0.917
Precision for Irregular	1.0	0.781	1.0	0.800
AUC	0.958	0.962	0.963	0.958
Recall	0.986	0.891	0.970	0.864
F-Measure	0.987	0.893	0.969	0.864

The accuracy depicts the total number of correct predictions. Naïve Bayes has 89.14% and 86.36% using both 10-fold cross-validation and 70-30 percentage split, respectively. On the other hand, C4.5 garnered 98.64 and 96.97% accuracy for 10-fold cross-validation and 70-30 percentage split, respectively. The result shows that the C4.5 classification generated more accurate predictions than the Naïve Bayes classification. Figure 1 shows the True Positive (TP) rate for two classification algorithms, whereas. Figure 2 describes the graphical representations of the accuracy of the two algorithms.

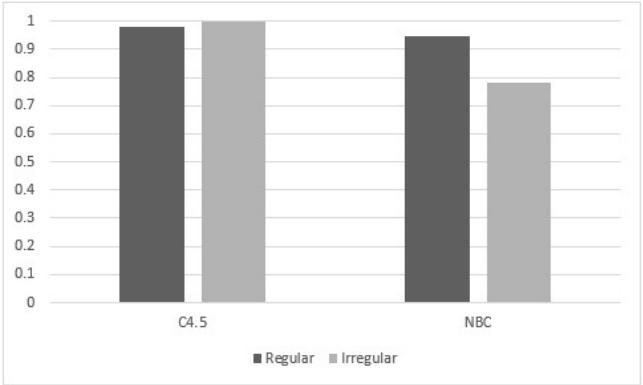

Figure 1: True Positive rate of the two algorithms

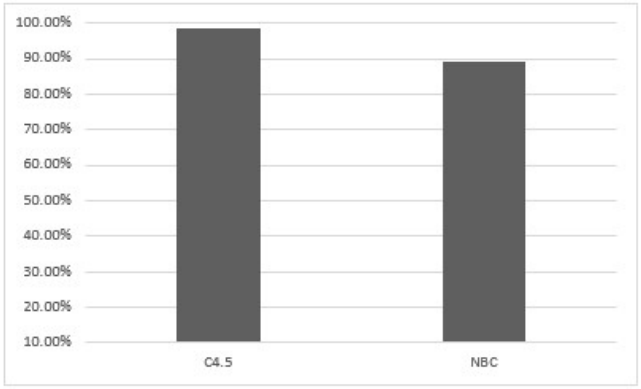

Figure 2: Accuracy of the two algorithms

Area under ROC curve is used to measure the quality of a probabilistic classifier [18]. Receiver Operating Characteristic (ROC) graph assess the machine learning algorithms. The ROC graph consists of the X-axis and Y-axis coordinates that are plotted with the false-positive rate and the true positive rate, respectively [19]. The ROC curve can be used to visualize and test the performance of the classification based on their performance [20].A perfect classifier has a measure of 1 and classifiers used in practice should preferably close to 1 [18]. If the classifier gains 0.5 AUC or lower, the prediction is random [18][19]. Figure 3 and Figure 4 show the ROC curve for regular and irregular classes of C4.5. The figures describe the C4.5 classification model is capable of distinguishing higher prediction accuracy where the ROC for both regular and irregular classes that has a higher measure close to 1. The AUC for both classes is 0.9578. Moreover, Figures 5 and 6 show the ROC curve for regular and irregular cases of Naïve Bayes classifier, where AUC measures for both classes gained 0.963- a bit higher compared to C.45.

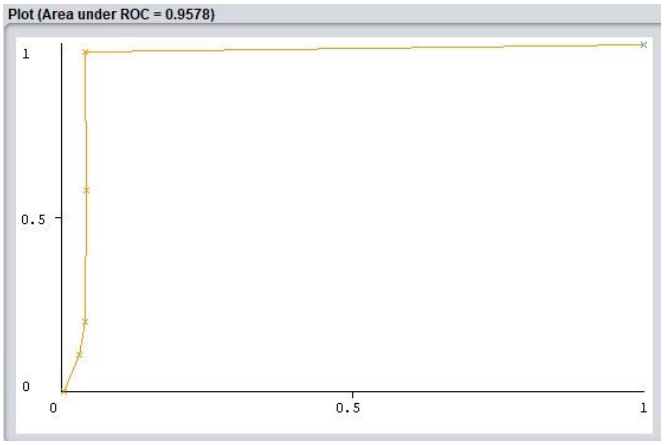

Figure 3: C4.5 ROC for Regular Students

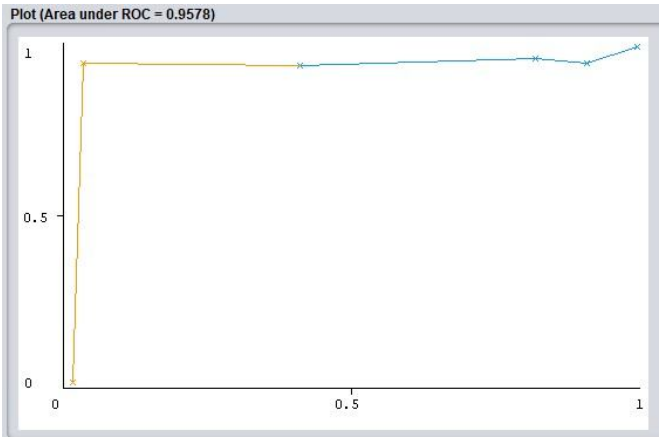

Figure 4: C4.5 ROC for Irregular Students

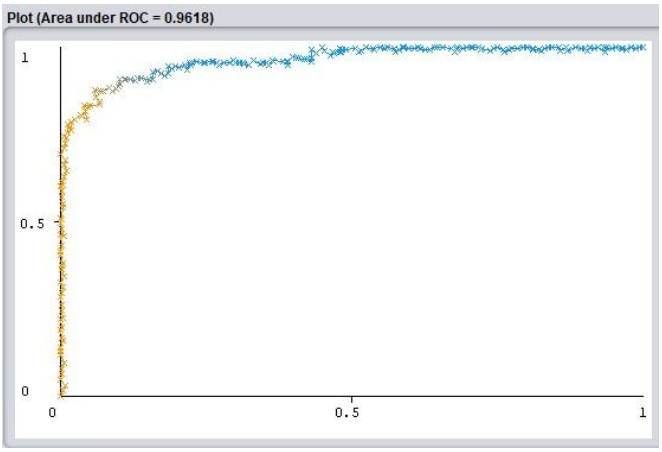

Figure 5: Naïve Bayes ROC for Regular Students

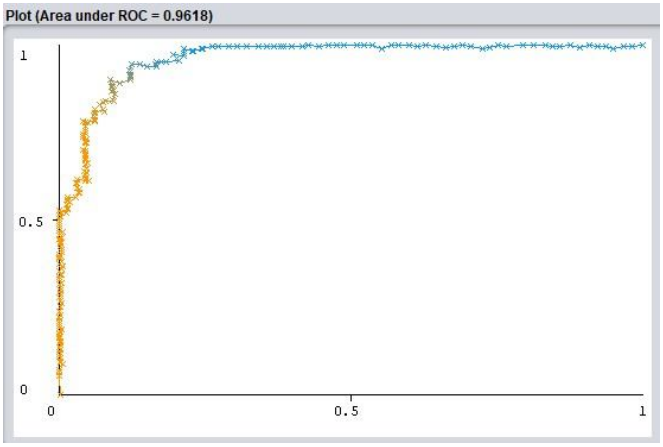

Figure 6: Naïve Bayes ROC for Irregular Students

Furthermore, this paper also targets to identify the influential attributes that affect the students 'performance. The Gain Ratio value defines the determination of these attributes. Table 3 shows that the average grade of math subjects from senior high school has the most influence on the performance of BSIT first-year students. Students' admission test results, course choice, senior high school general pointed average, and senior high school track is also among the attributes which have influences in the students' academic performance.

Table 4: Gain Ratio Attribute Evaluation

Attribute	Value
Math GPA	0.9073
OLSAT Overall Result	0.1421
OLSAT Non-Verbal Result	0.1223
Course Choice	0.1116
General Pointed Average	0.0965
Senior High School Track	0.096
OLSAT Verbal Result	0.0834

5. CONCLUSION AND RECOMMENDATION

The objective of this paper is to determine the relationship between the students' attributes such as senior high school track and academic data, and admission test result and their performance in the BSIT program during their freshman year. The most consistent attribute that affects the students' performance is the Math-related grades from senior high school. Prior research [12] explained that Math GPA is the most influential for academic performance in computer programming subjects. The study of [12] was corroborated by our findings.

This study has reinforced that success in first-year BSIT students' courses is associated with their performance from their senior high school, senior high school track, and admission test result. The program head should monitor these signals as they serve as a warning for potential failure. Demographic and socio-economic factors had no significant relation to student performance.

Furthermore, this paper analyzes the strength of data mining techniques, particularly on the performance of C4.5 and Naïve Bayes algorithms to achieve the model of academic performance. With various measure metrics, C4.5 classification gained 98.64% in 10-folds cross-validation and 96.97% in the 70% training and 30% testing percentage split. The C4.5 results were more consistent and more reliable when testing the data compared to Naïve Bayes. However, this paper only used a small number of instances, and it is recommended for future studies to add different data from different years or programs to increase the accuracy of the prediction.

ACKNOWLEDGEMENT

The authors wish to thank the Office of the Registrar and the Guidance and Testing Office of Davao del Norte State College for providing the pertinent data used in the study. The authors are also grateful to the University of Mindanao for extending financial support of this study.

REFERENCES

[1] I. Ozturk, "The Role of Education in Economic Development: A Theoretical Perspective," *SSRN Electron. J.*, vol. XXXIII, no. 1, pp. 1–7, 2011, doi: 10.2139/ssrn.1137541.

[2] F. G. and J. P. Miguel St. Aubyn, Álvaro Pina, "Study on the efficiency and effectiveness of public spending on tertiary education," in *Directorate-General for Economic and Financial Affairs Publications*, 2009.

[3] S. P. Brijesh Kumar Baradwaj, "Mining Educational Data to Analyze Students‟ Performance," *Int. J. Adv. Comput. Sci. Appl.*, vol. 2, no. 6, pp. 64–69, 2011, doi: 10.1177/039463200201500108.

[4] M. Kumar and A. J. Singh, "Evaluation of Data Mining Techniques for Predicting Student's Performance," *Int. J. Mod. Educ. Comput. Sci.*, vol. 8, no. 4, pp. 25–31, 2017, doi: 10.5815/ijmecs.2017.08.04.

[5] O. R. Devi, "A Deep Analysis on Aspect based Sentiment Text Classification Approaches," vol. 8, no. 5, pp. 1795–1801, 2019, doi: https://doi.org/10.30534/ijatcse/2019/01852019.

[6] T. M. Christian and M. Ayub, "Exploration of classification using NBTree for predicting students' performance," *Proc. 2014 Int. Conf. Data Softw. Eng. ICODSE 2014*, pp. 0–5, 2014, doi: 10.1109/ICODSE.2014.7062654.

[7] C. E. Lopez Guarin, E. L. Guzman, and F. A. Gonzalez, "A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining," *Rev. Iberoam. Tecnol. del Aprendiz.*, vol. 10, no. 3, pp. 119–125, 2015, doi: 10.1109/RITA.2015.2452632.

[8] A. Targowski, *From Data to Wisdom*, vol. 15, no. 5. 2005.

[9] S. Agarwal, *Data mining: Data mining concepts and techniques*. 2014.

[10] S. M. M. Rubiano and J. A. D. García, "Formulation of a predictive model for academic performance based on students' academic and demographic data," *Proc. - Front. Educ. Conf. FIE*, vol. 2014, 2015, doi: 10.1109/FIE.2015.7344047.

[11] C. Jalota and R. Agrawal, "Analysis of Educational Data Mining using Classification," *2019 Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput.*, pp. 243–247, 2019.

[12] P. Amornsinlaphachai, "Efficiency of data mining models to predict academic performance and a cooperative learning model," pp. 66–71, 2014.

[13] N. Dengen, E. Budiman, M. Wati, and U. Hairah, "Student Academic Evaluation using Naïve Bayes Classifier Algorithm," *2018 2nd East Indones. Conf. Comput. Inf. Technol.*, pp. 104–107, 2018.

[14] Anchal and P. Mittal, "Data Mining Techniques for IoT enabled Smart Parking Environment : Survey," vol. 8, no. 4, pp. 1688–1697, 2019, doi: https://doi.org/10.30534/ijatcse/2019/96842019.

[15] S. K. Yadav and S. Pal, "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification," vol. 2, no. 2, pp. 51–56, 2012.

[16] M. Mohammadi, M. Dawodi, W. Tomohisa, and N. Ahmadi, "Comparative study of supervised learning algorithms for student performance prediction," *2019*

Int. Conf. Artif. Intell. Inf. Commun., pp. 124–127, 2019.

[17] A. Goyal and R. Mehta, "Performance comparison of Naïve Bayes and J48 classification algorithms," *Int. J. Appl. Eng. Res.*, vol. 7, no. 11 SUPPL., pp. 1389–1393, 2012.

[18] M. Vuk and T. Curk, "ROC curve, lift chart and calibration plot," *Metod. Zv.*, vol. 1, no. 3, pp. 89–108, 2006.

[19] M. Karim and R. M. Rahman, "Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing," vol. 2013, no. April, pp. 196–206, 2013.

[20] T. Mauritsius, A. S. Braza, and Fransisca, "Bank Marketing Data Mining using CRISP-DM Approach," vol. 8, no. 5, pp. 2322–2329, 2019, doi: https://doi.org/10.30534/ijatcse/2019/71852019.