Technologique: A Global Journal on Technological Developments and Scientific Innovations
2024, Vol. 2, No.1, 49 – 60 | ISSN Online: 3028-1415 | ISSN Print: 3028-1407
DOI: https://doi.org/10.62718/vmca.tech-gjtdsi.2.1.SC-0624-031

Enhancing Student Outcomes: A Dual Methodology Using Naive Bayesian and Rule-Based Models

Severino M. Bedis, Jr.[1], Jonnifer C. Mandigma[1], Maria Leonila C. Amata[1], Aleta C. Fabregas[1]

[1]Graduate School, Polytechnic University of the Philippines, Sta. Mesa, Manila, Philippines

Article History:
Received: 15 May 2024
First Decision: 10 June 2024
Published: 29 July 2024

Abstract
This study investigates the efficacy of Naive Bayesian algorithms combined with a rule-based classifier to predict and map the academic performance of Bachelor of Science in Business Administration major in Marketing Management (BSBAMM) students at the Polytechnic University of the Philippines (PUP). Amidst disruptions caused by the COVID-19 pandemic, data-driven approaches are increasingly vital in higher education to enhance student outcomes. Historical data encompassing student demographics and academic records were analyzed to develop a predictive model, achieving 95% accuracy in forecasting student performance. This research underscores the potential of machine learning in identifying at-risk students early, facilitating timely interventions and personalized learning paths. The findings contribute valuable insights for educators and institutions seeking to optimize resource allocation and improve graduation rates.

Keywords: Predicting academic performance, higher education students, Naive Bayesian algorithm, Rule-based algorithm, historical data, student demographics, student success, student failure, machine learning algorithms, resource allocation, timely interventions, student outcomes

INTRODUCTION

The Philippines has witnessed a significant surge in higher education enrollment, growing from 2.9 million students in 2017 to 4.1 million by 2022. Academic success is crucial for career opportunities, with employers in the Philippines placing high importance on scholastic achievements like Grade Point Average (GPA) and specific course grades.

On the other hand, the COVID-19 pandemic has disrupted global higher education, highlighting challenges in delivering quality education and supporting student achievement. This has spurred interest in data-driven approaches to predict and enhance student outcomes, crucial for identifying at-risk students and providing timely interventions.

This study focuses on developing a predictive model using Naive Bayesian algorithms and rule-based classifiers at the Polytechnic University of the Philippines (PUP). By analyzing historical data of BSBAMM students, including demographics and academic performance, the study aims to understand factors influencing student success amidst the pandemic's impact.

The objective of this study is to construct a predictive model designed to distinguish students who are inclined to thrive or encounter challenges at the Polytechnic University of the Philippines (PUP). This endeavor employs a dual-pronged approach, combining the Naive Bayesian algorithm and a rule-based classifier. The Naive Bayesian algorithm, a probabilistic technique grounded in Bayes' theorem, is harnessed to estimate the likelihood of a student's success or failure by considering an array of factors that impact academic performance. Simultaneously, a rule-based classifier is employed to further refine and interpret these predictions.

This study focuses on predicting student success at the Polytechnic University of the Philippines (PUP) using a combination of Naive Bayesian algorithms and rule-based classifiers. By analyzing demographic data and academic performance, the study aims to identify factors influencing student success amidst the pandemic's impact.

The significance of this research lies in its potential to inform interventions that enhance educational resilience and outcomes at PUP and similar institutions. By understanding these

dynamics, educators and policymakers can better support students and strengthen higher education systems in times of crisis.

LITERATURES

Anticipating students' academic performance has become increasingly complex due to the vast data available in educational databases. In Malaysia, there remains a critical lack of a comprehensive system to analyze and track student progress and achievements. This issue arises primarily from two factors: insufficient research on predictive methodologies suitable for forecasting student performance in the Malaysian educational context, and a lack of investigation into the determinants influencing attainment levels in specific courses (Shahiria et al., n.d.).

A structured literature review is essential to address these gaps and enhance academic achievements through predictive analytics. The utilization of the Naïve Bayes algorithm in academic data mining has proven beneficial for evaluating and improving undergraduate academic performance by analyzing factors such as previous semester grades, attendance, assignments, discussions, and lab work (Razaque et al., 2017).

Recent evaluations indicate that predictive models incorporating distributional and hierarchical factors outperform traditional models. This suggests that Bayesian hierarchical models, which account for individual student differences, significantly enhance accuracy in predicting post-test performance, particularly in block-based programming environments (Emerson et al., 2020).

Furthermore, data mining techniques, such as the Naïve Bayes algorithm, have been successfully applied to predict graduating cumulative Grade Point Average based on applicant data from the University of Tuzla, Faculty of Economics. This approach underscores the practicality of using predictive models to enhance educational outcomes in diverse environments (Dole & Rajurkar, 2015).

Implementing a customized rule-based system (RBS) has also proven effective in identifying at-risk students early in their coursework. Through tools like the Risk Flag (RF), instructors can visualize and monitor student performance across various coursework components, facilitating timely interventions and improving overall student success rates (Dole & Rajurkar, 2015).

In conclusion, integrating advanced data mining techniques and rule-based systems offers significant potential to enhance educational practices, support educators, and improve student outcomes in higher education settings.

METHODS

The researchers employed a dual approach, combining survey and experimental methodologies. This research is founded on a comprehensive dataset derived from historical data collected through questionnaires administered to Bachelor of Science in Business Administration major in Marketing Management (BSBAMM) students at the Polytechnic University of the Philippines (PUP). This dataset encompasses a wide array of student-related information, serving as a means to collect relevant information and variables necessary for the ensuing experimental stage. The primary objective of this research endeavor was to discern and elucidate the factors or variables that possess the capacity to impact the dependent variables subject to investigation.

For data collection, the researchers utilized survey methodologies to collect a broad spectrum of student-related information. A survey was administered to 153 students enrolled in the Applied Marketing subject among BSBAMM students at PUP. The survey questionnaire was meticulously designed to capture various dimensions, including demographic characteristics, academic performance indicators, and other relevant attributes. The questionnaire included questions on gender, place of residence, parental educational background, relationship status, submission status of academic projects,

participation in class activities, attendance during lectures, and completion of assessment tasks. Academic performance indicators such as preliminary, midterm, and final grades were also included.

Data preprocessing aimed to ensure that the raw data was processed and analyzed accurately and consistently, thereby ensuring the validity of the results. This involved data cleaning to handle missing values and correct inconsistencies, categorical encoding to convert categorical data into numerical values for algorithm compatibility, and normalization to standardize numerical attributes and ensure uniformity in data range. Ethical considerations, such as informed consent and privacy safeguards, were meticulously observed throughout this phase.

Feature selection involved identifying the most relevant attributes that significantly impact the prediction of student performance. The selected attributes included gender, present address, parent's educational background, class participation, relationship status, class performance, attendance, and exam score. These attributes were used as features to build the model, as they were determined to be essential for the subsequent utilization of a naive Bayesian algorithm.

For model training, the researchers chose the Naive Bayesian classifier due to its simplicity and effectiveness, particularly for large datasets. The classifier is based on Bayes' theorem, assuming predictor independence. The components of Bayes theorem include the posterior probability of class given predictor (P(c|x)), the prior probability of class (P(c)), the likelihood or probability of the predictor given the class (P(x|c)), and the prior probability of the predictor (P(x)). The Gaussian Naive Bayes (GaussianNB) algorithm was utilized for model development due to its effectiveness in handling continuous data and its simplicity in implementation.

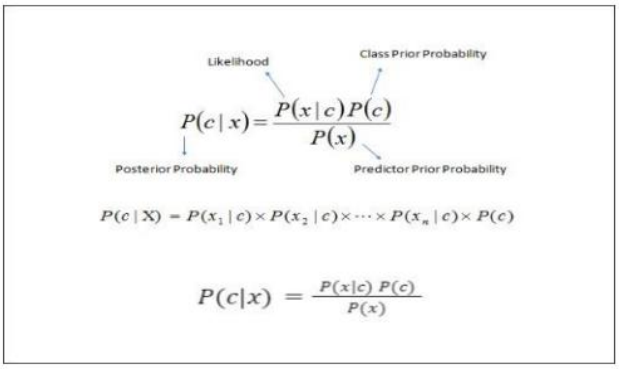


Figure 1
Naive Bayesian - Posterior Probability

- P(c|x) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of the predictor given the class.
- P(x) is the prior probability of the predictor.

Bayesian networks classifiers are popular machine learning classifiers, which have been receiving considerable attention from scientists and engineers across several fields such as computer science, medical applications, military applications, cognitive science, statistics, and philosophy.

The Bayesian network also is known as causal probabilistic network, Bayesian belief network, or simply Bayes net. The Bayesian network is defined as a directed acyclic graph over which is defined a probability distribution. Each node in the graph represents a random variable or event, while the arcs or edges between the nodes represent association or causal relationship between them. In Bayesian network the relationship between events is defined as a conditional probability, which is the probability of the event Y conditional on a given outcome of event X. Hence, a network of events connected by probabilistic dependencies is formed, called the Bayesian network. The probabilistic dependency is maintained by the conditional probability table (CPT), which is attached to the corresponding event.

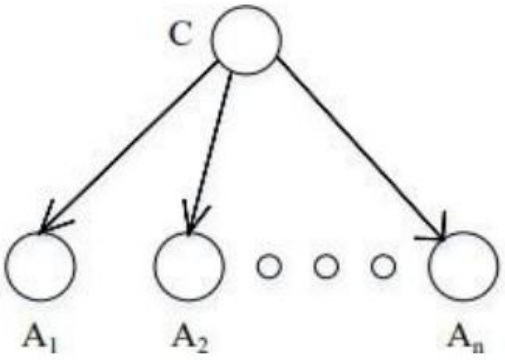


Figure 2
Structure of Naive Bayesian Network

Rule-based classification in data mining is a technique in which class decisions are taken based on various "if...then... else" rules. Thus, we define it as a classification type governed by a set of IF-THEN rules. We write an IF-THEN rule as:

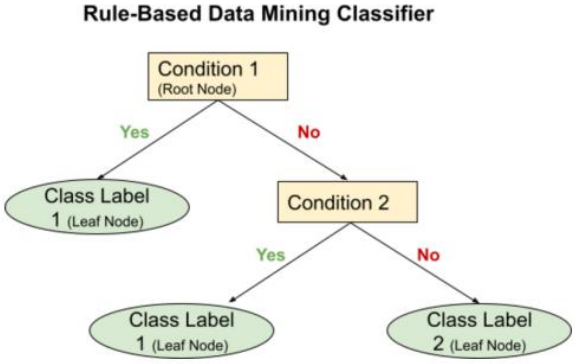
"IF condition THEN conclusion."

IF-THEN Rule

To define the IF-THEN rule, we can split it into two parts:

- Rule Antecedent: This is the "if condition" part of the rule. This part is present in the LHS (Left Hand Side). The antecedent can have one or more attributes as conditions, with logic AND operator.

- Rule Consequent: This is present in the rule's RHS (Right Hand Side). The rule consequent consists of the class prediction.



	Gender	Present Address	nts' Educational Backgr	Class Engagemen	elationshi	Attendance	ss Performar	Exam
2	Female	Metro Manila	Secondary Level	Often	Yes	100	95	95
3	Female	Outside Metro Manila	Secondary Graduate	Always	Yes	100	92.5	92.5
4	Male	Metro Manila	Elementary Graduate	Often	Yes	100	92.5	92.5
5	Female	Outside Metro Manila	Vocational	Always	Yes	100	92.5	92.5
6	Female	Outside Metro Manila	College Undergrad	Always	Yes	100	92.5	92.5
7	Female	Outside Metro Manila	Secondary Level	Always	Yes	100	95	95
8	Female	Outside Metro Manila	Secondary Graduate	Always	Yes	100	92.5	92.5
9	Female	Metro Manila	Secondary Level	Rarely	No	82.65	50	50
10	Male	Metro Manila	Elementary Graduate	Always	No	100	90	90
11	Male	Metro Manila	Secondary Graduate	Always	Yes	50	50	50
12	Female	Outside Metro Manila	Elementary Level	Often	Yes	100	92.5	92.5
13	Female	Outside Metro Manila	College Undergrad	Sometimes	Yes	83.33	90	90
14	Female	Outside Metro Manila	Secondary Level	Always	Yes	100	92.5	92.5
15	Female	Metro Manila	Vocational	Always	Yes	100	95	95
16	Male	Metro Manila	Elementary Level	Sometimes	Yes	83.33	95	95
17	Female	Metro Manila	College Undergrad	Always	Yes	100	92.5	92.5
18	Female	Outside Metro Manila	College Undergrad	Sometimes	Yes	100	90	90
19	Female	Outside Metro Manila	College Undergrad	Always	Yes	100	92.5	92.5
20	Female	Metro Manila	Vocational	Often	Yes	100	92.5	92.5
21	Male	Metro Manila	Elementary Graduate	Always	No	100	95	95
22	Male	Metro Manila	Elementary Level	Always	Yes	100	92.5	92.5
23	Female	Metro Manila	Bachelor's degree	Always	Not Appli	100	92.5	92.5
24	Female	Metro Manila	College Undergrad	Always	Yes	100	92.5	92.5
25	Female	Metro Manila	Secondary Level	Always	Yes	100	92.5	92.5
26	Male	Metro Manila	College Undergrad	Sometimes	Yes	100	92.5	92.5
27	Male	Outside Metro Manila	College Undergrad	Often	Yes	100	92.5	92.5
28	Female	Outside Metro Manila	Elementary Level	Always	Yes	100	95	95
29	Male	Outside Metro Manila	Bachelor's degree	Sometimes	Yes	100	50	50
30	Male	Outside Metro Manila	Secondary Level	Sometimes	No	100	92.5	92.5
31	Female	Outside Metro Manila	Secondary Level	Always	Yes	73.33	92.5	92.5
32	Male	Outside Metro Manila	Secondary Graduate	Sometimes	Yes	100	92.5	92.5
33	Female	Outside Metro Manila	Bachelor's degree	Rarely	Yes	100	92.5	92.5
34	Female	Outside Metro Manila	Bachelor's degree	Always	No	100	92.5	92.5
35	Female	Metro Manila	College Undergrad	Often	Yes	100	92.5	92.5
36	Male	Metro Manila	Elementary Level	Always	Yes	100	92.5	92.5
37	Female	Metro Manila	Secondary Graduate	Always	Yes	100	92.5	92.5
38	Female	Metro Manila	College Undergrad	Always	Yes	100	92.5	92.5
39	Male	Metro Manila	Elementary Level	Sometimes	No	100	92.5	92.5
40	Female	Metro Manila	College Undergrad	Always	Yes	100	92.5	92.5

Figure 3
Historical Data

Demonstrating a strong commitment to ethical considerations, the researchers gave paramount importance to ensuring data privacy and maintaining the anonymity of participants. To achieve this, personal identifiers such as names were intentionally excluded from the amassed dataset. This precautionary measure was driven by a profound sense of responsibility to safeguard confidentiality, underscoring the researchers' recognition that the compiled data would be utilized exclusively for the explicit purpose of this scholarly investigation.

The researcher determined that one of the primary goals of data preparation is to ensure that the raw data is processed and analyzed accurately and consistently, ensuring the validity of the results. As previously mentioned, the foundation of attributes for prognosticating student performance within higher education is rooted in the administered survey

questionnaires. These attributes serve as fundamental components for the subsequent utilization of a naive Bayesian algorithm, which endeavors to delineate a discernible trajectory in students' academic trajectories.

These attributes are used as features to build the model:

Table 1
Data Definition

Attribute	Descriptions
Gender	This will categorize the students as Male or Female. The researchers believe that gender is one of the important attributes as it plays a factor for the findings.
Present Address	The researchers categorize it as either Manila or residing outside of Manila.
Parent's Educational Background	Ranging from No Formal Education, Elementary Level/Graduate, Secondary Level/Graduate, College Undergraduate, Vocational Training, Associate's Degree, Bachelor's Degree, Master's Degree, to Doctorate Degree.
Class Participation	Gradations encompassing never, rarely, sometimes, always, and often.
Relationship Status	Evidenced through a binary response of Yes or No.
Class Performance	Numerical Value
Attendance	Numerical Value
Exam Score	Numerical Value

The researchers intend to adapt the posterior probabilities for all conditions of the aforementioned attributes and produce a systematic output that will be useful for the school in developing a strategy to achieve student performance. Furthermore, once the researchers have successfully developed a system that utilizes the Student Performance dataset, it will provide a reference and guidance to those in the Education Industry in developing a software/application that is demanded as well as reflected based on the student's demographics and grades.

F	G	H	J
Class Performanc	Attendance Score	Exam Score	pass_or_not
0.8	1	0.95	Pass
1	1	0.92	Pass
1	1	0.92	Pass
0.9	1	0.92	Pass
1	1	0.92	Pass
1	1	0.95	Pass
1	1	0.92	Pass
0.5	0.82	0.5	Fail
0.86	1	0.9	Pass
0.7	0.5	0.5	Fail
1	1	0.92	Pass
0.8	0.83	0.9	Pass
1	1	0.92	Pass
1	1	0.95	Pass
1	0.83	0.95	Pass
1	1	0.92	Pass
0.8	1	0.9	Pass
1	1	0.92	Pass
1	1	0.92	Pass
0.8	1	0.95	Pass
1	1	0.92	Pass
0.8	1	0.92	Pass
1	1	0.92	Pass
0.9	1	0.92	Pass
1	1	0.92	Pass
0.9	1	0.92	Pass
0.9	1	0.95	Pass
1	1	0.5	Pass
1	1	0.92	Pass
1	0.7333	0.92	Pass
1	1	0.92	Pass
1	1	0.92	Pass
1	1	0.92	Pass
1	1	0.92	Pass
1	1	0.92	Pass

Figure 4
Training Data Set

Figure 4 illustrates the graphical representation of the data attribute mapping process. This mapping constitutes the training dataset employed for the Python programming environment. It serves as the foundational dataset utilized for the generation of a comprehensive report file. This report file serves as the ultimate output, encapsulating the outcomes of the predictive analysis, specifically discerning whether a given student is anticipated to succeed or face challenges.

The model's performance was evaluated using accuracy metrics, and the Gaussian Naive Bayes algorithm achieved an accuracy of 95%, demonstrating its effectiveness in predicting student performance based on the selected attributes. The software and tools used for data analysis and model development included

Python as the primary programming environment, Pandas for data manipulation and analysis, Scikit-learn for implementing the Gaussian Naive Bayes algorithm and other machine learning tasks, and Matplotlib and Seaborn for data visualization and graphical representation of results.

```
pass_or_not = []
for i in range(len(df)):
    if df["score_mean"][i] >= 0.75:
        pass_or_not.append("Pass")
    else:
        pass_or_not.append("Fail")
df["pass_or_not"] = pass_or_not

df.to_csv( path_or_buf: 'Reports.csv', index=False)
print(df.head())
```

Figure 5
Program snippet that will tag if Passed or Fail MODULE OF THE SYSTEM

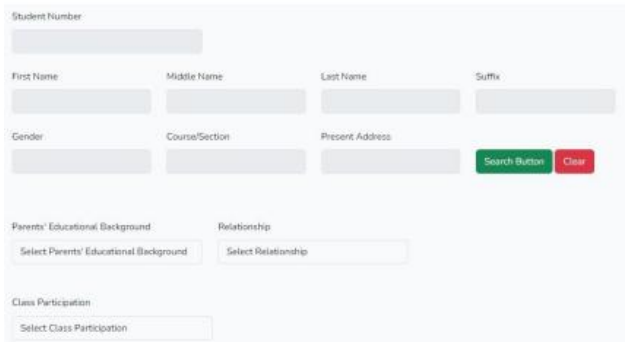


Figure 6
Student Performance Application

Depicted in Figure 6 is the intricately crafted user interface designed by the researchers. This interface functions as an access point for professors, acting as the intermediary for the input of student-specific information. These inputs are subsequently funneled into the designated database. Subsequent to this data entry process, a sequence of transformations is initiated, culminating in the generation of CSV files. These resultant files find purpose within a distinct system, meticulously tailored to facilitate the intricate task of mapping students' academic trajectories. This mapping endeavor stands as a pivotal instrument for professors, empowering them to discern and address the needs of students grappling with challenges. Furthermore, this tool equips educators with the means to customize instructional methodologies in alignment with the inclinations of contemporary learners, thereby fostering an educational milieu in harmony with the evolving demands of the contemporary landscape. This concerted endeavor holds the potential to augment the global competitiveness of graduates, aligning them aptly with the exacting requisites of prospective employer.

RESULTS AND DISCUSSION

The prototype's objective is to predict whether students will succeed or fail in a subject based on their demographic profile and grades. The program achieved an accuracy rate of 95% when tested with a dataset containing 153 students. Researchers selected 66 students at random from the dataset and employed them as input for the program.

Remarkably, all 66 records matched the data from the original dataset, yielding consistent results.

The confusion matrix is a method that is usually used to perform accuracy calculations on data mining concepts or decision support systems. In using the Confusion matrix method there are several terms, shown in the following table.

Table 2
Confusion Matrix

		TRUE VALUE	
		FALSE	*TRUE*
Prediction	FALSE	TN	*FP*
	TRUE	*FN*	*TP*

Description:

1. "True Positive" (TP), is the amount of positive data that is correctly classified by the system.
2. The amount of negative data correctly classified by the system is referred to as True Negative (TN).
3. The amount of negative data classified as positive by the system is known as false positive (FP).
4. The amount of positive data classified negatively by the system is known as false negative (FN).

Precision is the level of accuracy between the information requested by the user and the answer provided by the system. Precision can be calculated with the following formula:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall is the removal of data taken from the relevant data queried and recall is also known as sensitivity. Recall is formulated as follows:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 Measure is a combination calculation between recall and precision. The F1 Measure value range is 0 to 1, if the value is close to 0 then the prediction model is not good and vice versa if the value is close to 1 then the prediction model is good. To get the value in percentage, the value is multiplied by 100. F1 Score is to evaluate how well the hybrid metric is used for unbalanced classes. The F1 Score value range is 0 to 1, if the value is close to 0 then the prediction model is not good and vice versa if the value is close to 1 then the prediction model is good. To get the value in percentage, the value is multiplied by 100.

$$\text{F1-Score} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

As well as in calculating the percentage of accuracy with the following formula:

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

In this figure, the researcher analyzes the performance of our classification model based on the provided confusion matrix and classification report. This evaluation helps in understanding the strengths and weaknesses of the model, guiding further optimization and application in real-world scenarios.

Table 3
Detailed Classification Report and Confusion Matrix Metrics

METRIC	TRUE (CLASS 0)	FALSE (CLASS 1)	WEIGHTED AVERAGE
PRECISION	0.25	1.00	0.99
RECALL	1.00	0.95	0.95
F1-SCORE	0.40	0.98	0.97
ACCURACY	0.95 (95%)	-	-

The key metrics from our classification model reveal several important insights. The model correctly predicted the positive class (class 0) in 1 instance, known as True Positives (TP), and correctly predicted the negative class (class 1) in 62 instances, known as True Negatives (TN). There were no False Positives (FP), where the model incorrectly predicted the positive class when it was actually negative. However, the model made 3 False Negative (FN) errors, incorrectly predicting the negative class when it was actually positive.

Precision, a crucial metric that indicates the proportion of correct positive identifications, was 0.25 for the true class (class 0). This low value suggests a high number of false positives, as only 25% of the predictions for class 0 were correct. In contrast, the precision for the false class (class 1) was perfect at 1.00, indicating no false positives. Recall, which measures the proportion of actual positives correctly identified, was 1.00 for the true class, meaning all actual instances of class 0 were correctly identified. For the false class, the recall was high at 0.95, indicating that 95% of the actual instances were correctly identified, with a small number of false negatives.

The F1-Score, which balances precision and recall, was 0.40 for the true class due to the poor precision despite perfect recall. For the false class, the F1-Score was 0.98, indicating an excellent balance between precision and recall, reflecting strong model performance. Support, the number of actual occurrences of each class in the dataset, was 1 for the true class and 65 for the false class.

Overall, the model's accuracy was 0.95, demonstrating robust performance by correctly predicting 95% of the instances. The weighted average metrics, which account for class distribution imbalance, showed a precision of 0.99, recall of 0.95, and F1-Score of 0.97, providing a comprehensive view of the model's performance. These insights highlight the model's strengths and areas for improvement, particularly in increasing precision for the minority class (class 0) to enhance overall effectiveness.

Table 4
Probability of the Passed or Fail by Gender

		Probability			Probability
Male	17		Female	49	
Pass	16	94%	Pass	47	96%
Fail	1	6%	Fail	2	4%

In correlation of Performance to the Gender, Female dominance in successfully finishing the course with 96% Passing Probability, then the Male with 94% Passing Probability. In correlation of Failing Performance to Gender, Male is the highest with 6% Failing Probability than Female with 4% Failing Probability.

Table 5
Probability of the Passed or Fail by Address

		Probability			Probability
Metro Manila	34		Outside Metro Manila	32	
Pass	31	91%	Pass	32	100%
Fail	3	10%	Fail	0	0%

In correlation of Performance to the Address, Outside Metro Manila is more dominant in successfully Passing the course with 100% Passing Probability, than within Metro Manila with 91% Passing Probability. In correlation of Failing Performance to the Address, within Metro Manila is the highest Failing Performance with 10% Failing Probability than Outside Metro Manila with 0% Failing Probability.

Table 6
Probability of the Passed or Fail by Parent Educational Background

			Probability	
	Pass	Fail	Pass	Fail
Secondary Level	11	1	92%	8%
Secondary Graduate	15	1	94%	6%
Elementary Graduate	5	0	100%	0%
Vocational	4	0	100%	0%
College Undergrad	13	1	93%	7%
Elementary Level	8	0	100%	0%
Bachelor's degree	7	0	100%	0%

In correlation of Performance to the Parents Education, Elementary Graduate, Vocational, Elementary Level and bachelor's degree are the highest in successfully Passing the course with 100% Passing Probability, then the Secondary Graduate with 94% Passing Probability, College Undergrad with 93% Passing Probability, and Secondary Level with 92% Passing Probability. In correlation of Failing Performance to the Parents Education, Secondary Level is the highest Failing Performance with 8% Failing Probability than College Undergrad with 7% Failing Probability, Secondary Graduate with 6% Failing Probability, and Elementary Graduate, Vocational, Elementary Level and bachelor's degree with 0% Failing Probability.

Table 7
Probability of the Passed or Fail by Class Participation

	Often	Always	Rarely	Sometimes
Pass	9	34	6	14
Fail	0	1	1	1
Probability				
Pass	100%	97%	86%	93%
Fail	0%	3%	14%	7%

In correlation of Performance to the Class Participation, often is the highest in successfully Passing the course with 100% Passing Probability, then the Always Participation with 97% Passing Probability, sometimes with 93% Passing Probability, and

Rarely with 86% Passing Probability. In correlation of Failing Performance to Class Participation, rarely is the highest Failing Performance with 14% Failing Probability than Sometimes with 7% Failing Probability, always with 3% Failing Probability, and Often with 0% Failing Probability.

Table 8
Probability of the Passed or Fail by Relationship

	Yes	No
Pass	54	9
Fail	1	2
Probability		
Pass	98%	82%
Fail	2%	18%

In correlation of Performance to the Relationship, With Relationship is the highest in successfully Passing the course with 98% Passing Probability, then without Relationship with 82% Passing Probability. In correlation of Failing Performance to the Relationship, without Relationship is the highest Failing Performance with 18% Failing Probability than with Relationship with 2% Failing Probability.

The researcher used a rule-based classifier to identify if the attributes have correlated for the passing and failure of students in the subject.

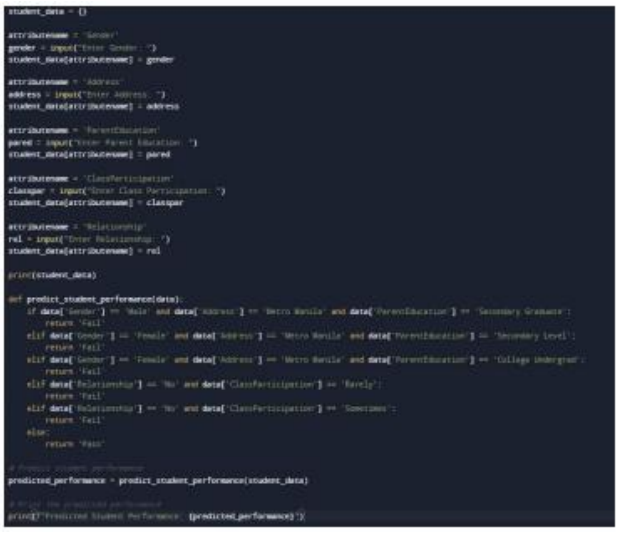


Figure 7
Rule-based Snippet Code

The researchers randomly picked 10 records from the dataset and used it as a feed file for the program. 8 out of 10 records matched the records from the dataset and results.

Table 9
Testing Data of Rule-Based

	Gender	Address	Parent Education	Class Participation	Relationship	Pass/ Fail	Rule Based
Group 1	Female	Metro Manila	Secondary Level	Rarely	No	Fail	Fail
Group 2	Male	Metro Manila	Secondary Graduate	Always	Yes	Pass	Pass
Group 3	Male	Metro Manila	Elementary Graduate	Always	No	Pass	Pass
Group 4	Male	Metro Manila	College Undergrad	Sometimes	Yes	Pass	Pass
Group 5	Male	Outside Metro Manila	Secondary Graduate	Sometimes	Yes	Pass	Pass
Group 6	Female	Outside Metro Manila	Bachelor's Degree	Rarely	Yes	Pass	Pass
Group 7	Female	Metro Manila	College Undergrad	Sometimes	No	Fail	Fail
Group 8	Female	Outside Metro Manila	Elementary Level	Always	Yes	Pass	Pass
Group 9	Male	Metro Manila	College Undergrad	Rarely	No	Pass	Fail
Group 10	Male	Metro Manila	Secondary Level	Sometimes	No	Pass	Fail

The primary objective of this study was to develop and evaluate a predictive model to determine whether students will succeed or fail in a subject based on their demographic profile and grades. The study aimed to understand the impact of various factors such as gender, address, parental education, class participation, and relationship status on students' academic performance. The model achieved an overall accuracy rate of 95% with the complete dataset of 153 students and 95% with a subset of 66 students, indicating high effectiveness in classifying student performance.

Additionally, a rule-based classifier was used to identify correlations between attributes and student performance. Ten records were randomly selected from the dataset and used as a feed file for the program, with 8 out of 10 records matching the dataset results. This further validated the model's robustness, highlighting the significant role of gender, address, parental education, class participation, and relationship status in predicting student success.

These findings confirm that demographic and behavioral factors significantly impact students' academic performance, supporting the study's hypotheses. The insights provided are valuable for educational institutions to tailor interventions and support strategies to improve student outcomes effectively.

Conclusions and Recommendations. The findings of this study have several important implications for educational institutions and policymakers. The high accuracy of the predictive model indicates that demographic profiles and behavioral factors are significant determinants of student success. This suggests that interventions tailored to specific student characteristics can enhance educational outcomes. For instance, the higher academic achievement of female students implies that male students may benefit from targeted support programs, such as mentorship, counseling, and academic workshops designed to address their specific challenges. The strong correlation between parental education and student performance highlights the importance of parental involvement, suggesting that educational institutions could develop programs to engage parents, especially those with lower educational attainment, providing them with the resources to support their children's education. The disparity in academic performance between students from Metro Manila and those from outside the region suggests a need for regional-specific interventions. Institutions could implement tailored support services, such as localized tutoring programs and remote learning resources, to address the unique challenges faced by students in different regions. The finding that students not in relationships tend to perform better academically suggests that relationship dynamics can impact academic focus and time management. Institutions might consider offering relationship counseling and time management workshops to help students balance personal and academic commitments. The strong correlation between performance across different subjects indicates the benefits of holistic educational strategies. Encouraging interdisciplinary learning and collaboration can help students develop a well-rounded skill set, promoting overall academic growth.

The findings of this study are consistent with previous research that highlights the importance of demographic and behavioral factors in academic performance. For instance, previous studies have also found that female students tend to outperform male students academically, reinforcing the need for gender-specific educational interventions. Similarly, this study confirms that parental education significantly influences student performance, aligning with earlier findings that parental involvement and higher educational attainment contribute positively to student outcomes. The study's results also align with existing literature indicating regional disparities in educational outcomes, often attributed to factors such as access to resources, quality of instruction, and socioeconomic conditions.

While the study provides valuable insights, there are several limitations that should be addressed in future research. The study focused on students from a specific program (BSBAMM) at the Polytechnic University of the Philippines. Future research should include a broader range of programs and institutions to enhance the generalizability of the findings. Additionally, the study primarily considered demographic and behavioral factors. Future research should incorporate a wider range of variables, including psychosocial and cognitive factors, to provide a more comprehensive understanding of student performance. Longitudinal studies tracking students' academic journeys over time could offer deeper insights into the long-term impacts of the identified factors. Incorporating qualitative data, such as interviews and focus groups, could provide richer context and understanding of the quantitative findings, uncovering underlying reasons behind the observed correlations and informing more targeted interventions. Lastly, while the Naive Bayesian method showed promising results, future research should explore the efficacy of other machine learning algorithms. Comparing the performance of various algorithms could identify the most effective approaches for predicting academic success.

Comparing the results of this study to those of other authors highlights various perspectives on improving educational outcomes through predictive analytics and tailored interventions. Chao and Symaco (2021) emphasize systemic challenges in the Philippine education system, such as inadequate funding and policy inconsistencies, proposing broad policy reforms rather than specific, targeted interventions. On the other hand, the current study's results point to the high accuracy of predictive models in identifying significant determinants of student success, such as demographic profiles and behavioral factors, suggesting specific actions like gender-specific support programs and parental education initiatives. Dole and Rajurkar (2015) and Razaque et al. (2017) similarly identify the effectiveness of decision trees, neural networks, and Naïve Bayes algorithms in predicting student performance. However, their studies do not propose specific intervention strategies based on these predictions. In contrast, the current study extends beyond technical evaluations to suggest actionable measures tailored to the identified determinants, such as mentorship programs for male students and support services for different regions. Emerson et al. (2020) and Shahiria, Husaina, and Rashida (2015) focus on the technical aspects of predictive modeling and data mining techniques, emphasizing their potential in educational settings. Emerson et al. demonstrate the effectiveness of Bayesian hierarchical models in specific programming environments, while Shahiria et al. provide a comprehensive review of various data mining techniques. Both studies align with the current study in recognizing the importance of predictive analytics but do not provide detailed, practical interventions based on demographic and behavioral insights. Overall, while previous studies highlight the utility of predictive models and data mining techniques, the current study uniquely contributes by translating these findings into specific, practical strategies for educational institutions and policymakers. This includes interventions like gender-specific support, parental engagement programs, regional-specific support services, relationship counseling, and interdisciplinary learning initiatives, offering a more comprehensive approach to enhancing student success.

The study's findings underscore the significant impact of demographic and behavioral factors on student academic performance. By tailoring interventions to these factors, educational institutions can better support their students and enhance overall educational outcomes. Future research should address the identified limitations and expand the scope to provide a more comprehensive understanding of the determinants of academic success. This continued research will contribute to more effective educational strategies and policies, ultimately helping students achieve their academic and career goals.

REFERENCES

Chao, Jr., R. Y., & Symaco, L. P. (2021, April 21). Higher education in the Philippines: Prospects and challenges - Digested by The HEAD Foundation. Science Technology & Innovation. The HEAD Foundation. https://digest.headfoundation.org/2021/04/06/higher-education-in-the-philippines-prospects-and-challenges/

Dole, L., & Rajurkar, J. (2015). A decision support system for predicting student performance. International Journal of Innovative Research in Computer and Communication Engineering, 2(12), 7232-7237. https://doi.org/10.15680/IJIRCCE.2014.0212015

Emerson, A., Geden, M., Smith, A., Wiebe, E., Mott, B., Boyer, K., & Lester, J. (2020). Predictive student modeling in block-based programming environments with Bayesian hierarchical models. In Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization. https://doi.org/10.1145/3340631.3394853

Razaque, F., Soomro, N., Shaikh, S., Soomro, S., Samo, J., Kumar, N., & Dharejo, H. (2017). Using naïve bayes algorithm to students' bachelor academic performances analysis. In 2017 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS) (pp. 1-5). https://doi.org/10.1109/ICETAS.2017.8277884

Shahiria, A. M., Husaina, W., & Rashida, N. A. (2015). A review on predicting student's performance using data mining techniques. The Third Information Systems International Conference. https://www.sciencedirect.com/science/article/pii/S1877050915036182