Republic of the Philippines

PALAWAN STATE UNIVERSITY

College of Sciences

Puerto Princesa City

"A Machine Learning Approach for a Best College Program Recommendation System Based on Cognitive Abilities and Skill Sets"

In Partial Fulfillment of the

Requirements for the Degree of

Bachelor of Science in Computer Science

Written By:

GATCHALIAN, JOHN REX

RECARZE, SPLEDELYN CRISTINE

JAMION, ANEZA

NAVARRO, LANCE ARMSTRONG

Adviser:

ADONIS C. AMPONGAN

Table of Contents

CHAPTER I

INTRODUCTION

Background of the study

Education is being emphasized and hailed as the key to having a successful future (Ankit et al., 2023). In the Philippines, drastic transformation in the education field has made it a distinct system among its neighboring countries in recent years. The basic education was extended to 12 or 13 years; this is composed of 6 years in elementary, 4 years in junior high school, and 2 years in senior high school to provide sufficient time for learning and mastering the fundamental academic concepts and lifelong skills preparing them for tertiary education (RA 10533, section 7; Cariaga, 2023).

Aside from reforming the structure of basic education, laws have been passed affecting the education opportunities of the student, especially those in higher education. Republic Act 10931, or the Universal Access to Quality Tertiary Education Act was passed and signed in 2016, amending the free tuition for tertiary education. The law provides a subsidy for students enrolled and who will enroll in undergraduate programs of SUCs, LUCs, and private higher education institutions, as well as technical-vocational institutions (RA 10931, section 7; ABS-CBN News, 2017). Although these changes reshape the field in the country, higher education is found struggling to find a solid footing. According to the EDCOM II report in 2024, the higher education participation rate is at 34.89 percent in the country, which is significantly lower than the ASEAN average of 41.10 percent. Employment is the top deterrent, at 44.17 percent, followed by lack of personal interest, which sits at 24.94 percent (Gonzales, 2025, para. 9).

However, there are other factors that affect these numbers. In the recent PSA survey for 2024 Functional Literacy, Education, and Mass Media Survey, or FLEMMS, which aims

to determine the literacy rate in the country, it reflected that a total of 18.96 million Filipino students who graduated from junior and senior high school in 2024 are considered functionally illiterate. Functional literacy is defined as the ability to read, write, and comprehend (PSA, 2024). The attrition rate in the country is disturbingly high, with a 39 percent national dropout rate (EDCOM 2, 2024).

This is why the selection of an appropriate college program is significant to be able to survive the chosen program. There are several factors that can influence the choice of the students. This includes parental influence, social inference, affordability of the program, practicality, and personal preference or interests (Sadjail et al., 2018).

Moving to higher education, most high school students are unsure about their specific university major. In the country, traditional recommendations are the domineering methods, which are limited and general. Guidance counselor assessment is the primary method used by many students as a basis for decision-making. Although counselors use students' data such as transcripts and extracurricular activities, their recommendations are limited to their personal opinion and insights. Another method is personal assessment and interests of the students. Most students are self-assessing their interests, growth, and strengths to decide their career. Not only is this method subjective but also influenced by bias from surroundings, family pressure, peer pressure, and time pressure.

This inaccuracy in the recommendation system leads to higher attrition rates among universities and higher educational institutions. The recent data from the Commission on Higher Education (CHED) reveals the national dropout rate of approximately 39.9 percent due to various reasons (PSA, 2024). One of the contributing factors of this problem is the mismatch between cognitive abilities, visible skills, and choice of programs. The option of choosing a specific college program is not restricted by their senior high school strand as per

directive of CMO No. 105 s 2017 (CHED, 2017). Due to these factors, many students struggle in their academics, leading to dropout from university.

Machine learning (ML) is a branch of artificial intelligence (AI) that explores the capabilities of computers and machines to perform tasks autonomously and learn by imitation of the human learning process. Machine learning performance and accuracy are improved through experience and exposure to enormous amounts of relevant data (IBM, 2021). It is applied in many fields of study, such as health, economics, computer vision, autonomy, and even education (Geekforgeeks, 2023). In the study of Yurtkan et al. (2023), they used the FFNN to predict students' success in a higher education setting. They used the framework to model the relationships in student profiling and program matching.

Although machine learning has been applied in some educational fields for predicting performance outcomes, such as the work of Yurtkan et al. (2023), there is a significant gap in its application in optimal college program recommendations, especially in the country. Current practices in guidance remain subjective and opinion-driven, lacking the precision and data-driven decision-making that machine learning can provide.

With machine learning, this study aims to develop a prediction model that can provide best-fit college programs suitable for students' cognitive and skill capabilities. The prediction model will recommend the appropriate program with increased survival chances for a student.

Statement of the Problem

This study seeks to enhance the academic decision-making of the students by offering analysis on their cognitive abilities and skill sets, reducing the possibilities of mismatch between the students and their chosen field of study. This results in better academic

performance and career readiness. Specifically, this study seeks to answer the following questions:

1. What is the projected attrition rate of College of Sciences students in the next five years based on current enrollment and graduation data?

2. How can machine learning models utilize cognitive abilities and skills in determining student success rate in chosen programs?

3. To what extent can we compare the result of our machine learning model accuracy compared to the traditional method, and what percentage improvement in accuracy can be observed?

Objectives of the Study

The purpose of this study is to create a predictive analysis with machine learning (ML) to recommend a best-fit college program based on a student's cognitive abilities and skill sets.

This research aims to:

1. Project the attrition rate of College of Sciences students for the next five years using historical data to imply the contribution of cognitive abilities and skill misalignment to the college attrition rate.

2. Ensemble machine learning models to provide a higher level of accuracy to predict the best recommendation of a college program based on cognition and skills.

3. Evaluate the accuracy of the model compared to traditional methods to improve the outcome of prediction.

Scope

The scope of this study focuses on ensembling a prediction model using machine learning to recommend the best-fit college program to the students based on their cognitive abilities and skill sets. The objective is to provide personalized and data-driven recommendations that align students with college programs best suited to their strengths and potential.

This study will also analyze the current attrition rate at the College of Sciences, Palawan State University–Main Campus, and determine the level of program indecisiveness among the students to assess the degree of misalignment between students and their chosen programs. These findings will support the development of more effective academic interventions and enhance student survivability.

Specifically, this study will:

1. It will collect data from college students to determine program indecisiveness and examine the level of mismatch between students and their current program.
2. Designing and testing various machine learning models to determine the most suitable model to use for specific phases of the system and identify the most effective approach in predicting optimal college programs.
3. The study is limited to college students, particularly those who are currently enrolled at Palawan State University–Main Campus.

Limitations

Despite the intended contributions of the system, the study has limitations that are necessary to acknowledge. The factors identified that will affect the predictive analysis are limited only to cognitive abilities, skill sets, and academic status. There's also a limitation in

terms of the breadth and depth of the dataset of skills that will be utilized for training and testing of machine learning models.

Significance of the Study

The study on prediction analysis using Feedforward Neural Network (FFNN), Naive Bayes (NB), Random Forest (RF), ensemble,aims to help students in choosing their preferred career path that suits their skill set. It will be evaluated based on the analysis of their cognitive abilities and skill assessment of the students.

This study will be beneficial for:

Students

The predictive system will be beneficial to the career path as a guide for incoming college students in choosing their program as well as expanding their awareness of other feasible programs. It will also enhance their chances of academic success and reduce the number of shifting programs or dropouts.

Universities and Higher Institutions

The prediction model can reduce the overpopulation of college programs and reduce the number of due and retention due to the undecided choice of college program. This can lead to improved student satisfaction and higher retention rates.

Guidance Counselors

The prediction model can help the guidance counselors to give efficient student counseling through the data-driven insights that will support their advisory role.

Future researchers

This system will contribute to the existing body of knowledge in the field. The findings, methodologies, and development system can be used as references for future research relating to predictive machine learning and cases relating to collegiate track.

Definition of Terms and Formulas

For clearer understanding of the terms that will be used in this study, the following terms are defined theoretically and operationally:

Attrition Rate: It refers to the rate of students who delayed their completion and graduated with their degree. This is classified into two types: normal attrition rate and adjusted attrition rate (TCSI, n.d.).

In this study, attrition rate will serve as a tool to measure the retention rate within the university. It will be used to assess student survivability, or the number of students who successfully completed their program requirements. The formula below will be used to calculate the attrition rate**.**

$$\text{Attrition Rate (\%)} = \frac{\text{Total Enrollees (Admission Year) - Total Graduates (After 4 years)}}{\text{Total Enrollees (Admission Year)}}$$

Source: CMO No. 46, s. 2012

Cognitive ability: It is the mental prowess and competence of an individual that enables them to think, learn, reason, and solve problems efficiently and effectively. It is classified into four categories: attention, memory, logic and reasoning, and processing (Indeed, 2025). While these abilities occur naturally in the brain, they can be further improved by taking up consistent mental challenges and actively engaging in learning tasks.

In this study, cognitive abilities will be tested from the admission test result data which covers numerical reasoning, abstract reasoning, verbal reasoning, spelling, and

language usage, which covers core subjects such as Math, Science, and English. These aim to identify the strengths and weaknesses of students' cognitive abilities as a factor in the recommendation system.

Skills: These are the individual's abilities, knowledge, and expertise that are needed to do a specific task or activity effectively. According to Zhang (2019), skill is an expertise that has been developed through training and experience. Understanding and managing skills effectively can contribute to improving academic and professional performance.

In this study, skills will be identified as a key factor in identifying and recommending the most suitable college program. The system aims to match individual competencies with the demands of various academic programs, increasing the likelihood of program satisfaction and success.

College Program: A title of a program consisting of courses and related curriculum to achieve specific learning outcomes. It consists of the degree (e.g., BS) and the field of study (e.g., computer science).

In this study, college programs serve as the primary prediction output of the recommendation system. Based on the students' cognitive and skills assessment, the system aims to recommend the most suitable college program that aligns with their strengths, increasing the likelihood of survival in a certain program.

Indecisiveness: The state or quality of being unable or having difficulty making a decision to resolve something. It is defined as a tendency to experience problems with decision-making across situations or domains. It can stem from different factors such as lack of confidence, fear of failure, information overload, or cognitive overload. (Merriam Dictionary, 2025).

In this study, indecisiveness is recognized as a potential barrier for students when choosing an appropriate college program. It is a contributing factor to program mismatch and dropout, as students who are uncertain or misinformed make uninformed decisions. The machine learning-driven recommendation system aims to mitigate the effects of indecisiveness by providing data-driven guidance based on a student's cognitive abilities and skill sets.

Feedforward Neural Network: A type of artificial neural network in which information flows in a single direction. It is used for pattern recognition tasks like predictive analytics. (GeeksforGeeks, 2025)

In the study, Feedforward Neural Network is used to capture the non-linear and multi-dimensional interactions between input data.

Naive Bayes: A classifier that is a supervised machine learning algorithm that is used for classification tasks. It uses the principle of probability to perform its classification tasks. (IBM, 2025)

In the study, Naive Bayes is the baseline model to be applied first and will calculate the probability of each category in the system.

Random Forest: A commonly used machine learning algorithm that combines the output of multiple decision trees to reach a single result. It is commonly used because of its flexibility that can handle both classification and regression problems. (IBM, 2025)

In the study, Random Forest is used to build many decision trees for the input data, and the output will be used to provide a ranking feature.

Machine Learning: It is a branch of artificial intelligence (AI) focused on enabling computers or machines to mimic humans' way of learning and processing information. It is programmed to perform tasks autonomously and improve performance and accuracy through experience and exposure to more relevant information. (IBM, 2021)

In this study, machine learning is used to analyze and interpret cognitive and skill-based data to generate an accurate, data-driven college program recommendation.

Artificial Intelligence: It is a branch of computer science that enables computers and machines to simulate human intelligence on computer systems. These processes include learning, reasoning, problem-solving, comprehension, decision-making, and self-correction.

In this study, AI serves as the overarching discipline in which machine learning techniques are applied to guide academic decision-making for students.

Predictive Analysis: or Predictive Analytics, is a branch of advanced analytics that creates predictions using the historical data combined with statistical modeling, data mining techniques, and machine learning. (IBM, 2022)

In the study, predictive analysis is used to determine the most suitable college program for a student by examining their cognitive and skill profiles, hence increasing the chances of academic success.

Predictive Model: Predictive modeling is a process used in data science to create a mathematical model that predicts the outcome based on the input data. It involves statistical algorithms and machine learning techniques to analyze the data and make predictions. (Geekforgeeks, 2024)

In this study, the predictive model integrates the cognitive assessment and skill selection into two-stage machine learning frameworks to recommend optimal college programs.

Chapter II
Review of Related Literature and Studies

Foreign Literature

Recent advancements in predictive analytics for academic performance have demonstrated the effectiveness of integrating deep learning with traditional machine learning techniques. A study by Kukkar et al. (2024) highlighted the superior accuracy (97%) of a hybrid model combining Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Random Forest (RF) in forecasting student pass-or-fail outcomes. This hybrid model outperformed other combinations such as RNN + LSTM + Support Vector Machine, Naïve Bayes, and Decision Trees. The results emphasize the importance of capturing both temporal dependencies in student data and complex feature interactions, offering a powerful tool for early intervention strategies.

Rohman et al. (2025) conducted a study on predicting student academic performance by addressing challenges with imbalanced datasets in machine learning. The researchers proposed a hybrid model combining Logistic Regression (LR) and Random Forest (RF), enhanced with SMOTE oversampling and hyperparameter tuning (GridSearch and RandomSearch). Their findings showed that the hybrid model significantly outperformed traditional LR and RF, achieving an accuracy of 95.74%, with higher precision, recall, and F1-scores. The study emphasizes that integrating LR and RF provides both interpretability and robustness, while SMOTE helps balance data distribution, leading to more reliable predictions. This approach contributes to a practical solution for student performance prediction in higher education, enabling early detection of at-risk students and supporting timely academic interventions.

According to Kord et al. from the study “” mining is used for extracting hidden knowledge in educational data. It was stated that students often encounter difficulties in choosing appropriate courses and suitable programs, which is considered the most important factor in avoiding career failures. So in this case, their study used machine learning algorithms like Random Forest, SVM, and F-1 scores. The results show that the Support Vector Classification (SVC) model outperformed others, achieving a 78.04% multi-classification accuracy and a 75.37% F1-Score. Also, the prediction model was developed to predict students' academic grades in their upcoming courses based on their past performance, which then will recommend a model for the students in choosing their suitable courses and programs.

The study by Trujillo et al. (2024) titled: *ARTIFICIAL INTELLIGENCE IN EDUCATION: A SYSTEMATIC LITERATURE REVIEW OF MACHINE LEARNING APPROACHES IN STUDENT CAREER PREDICTION” is a* systematic literature review of using machine learning techniques in higher education career recommendation. Their study shows that in the times of leveraging artificial intelligence for personalized academic guidance. They analyzed 38 studies from an initial pool of 1,296 articles from a custom-built web leveraging the CrossRef API that are based on machine learning techniques, data types, and validation metrics. Their findings revealed that Random Forest, Support Vector Machines and Neural Networks are the most frequently employed models to have a high accuracy in career recommendations for higher education. This article also highlights the key validation, such as precision, recall, and F-1 scores that reflect the effectiveness of these models. However, some limitations were discovered, like the lack of access to open datasets and scarcity of studies with relevant data that evaluate the long-term impact of the study’s recommendation. This literature review provides a solid foundation for enhancing career

recommendation systems using machine learning techniques to modernize academic guidance and alignment to their career goals.

In the study, "*RECOMMENDING COLLEGE PROGRAMS TO STUDENTS USING MACHINE LEARNING*" by Ayesha et al. (2022), students often make decisions without careful thought, mainly because of a lack of proper advice and support. In this study, they used four methods in utilizing students' data, particularly their performance in highschool, college placement test, and the standardized IELTS exam for college program recommendation, and to predict the students' GPA in the certain programs. Each data were evaluated and compared using Decision Trees (DT), Neural Networks (NN), K-Nearest neighbor (KNN), and Linear Regression (LR). Which in this case, linear regression was highlighted in selecting a program to recommend on students. This study also suggests that the more records, the more accurate a system recommendation could be.

Synthesis of Foreign Literature

The reviewed studies collectively emphasize the increasing effectiveness of hybrid deep learning models integrated with traditional machine learning techniques in predicting student academic performance. Kukkar et al. (2024) demonstrate that combining RNN, LSTM, and RF significantly improves prediction accuracy by leveraging both temporal dependencies and complex feature interactions. Focusing on the advantage of the deep learning architectures with the traditional machine learning combination technique for academic performance of students.

In addition, Rohman et al. (2025) find a solution for the imbalance in educational datasets by proposing a hybrid logistic regression and random forest model enhanced with SMOTE oversampling and hyperparameter tuning. The results of their study showed that this

approach achieved a higher accuracy and improved precision, recall, and F-1 scores compared to other models. In general, these studies support the use of hybrid machine learning approaches as an effective tool in predicting student academic outcomes and identifying at-risk learners. By leveraging complementary strengths of hybrid models and different algorithms. Which contributes to academic decision-making and timely intervention strategies in educational settings.

Local Literature

Factors considered among the University of the East-Manila students in their College Program Preference Title Page Mendoza et al... University of the East - Manila. This research is a quantitative study that investigated relationships among variables of the factors, which include the National Career Assessment Examination (NCAE), employability, and financial capability and its relationship to college program preference of the University of the East Manila students in Academic Year 2022-2023. The main objective of the study was to prove that there was a significant relationship between the three factors and college program preference. Aptitude tests, which are evaluated through grades, are significant data points used to know a student's career strength. The researchers used survey methods via Google Forms. Overall, the three factors, namely the National Career Assessment Examination, employability, and financial capability, have shown a significant relationship from the perception of the respondents in terms of college program preference. Moreover, students aim to pursue a career inclined with their college program even though rising tuition fees have been one of their major concerns. Determining the mentioned factors' relationship with college program preference could contribute to one-on-one consultation sessions, which can be used to articulate the results of the career assessment to the student, inform them of the

benefits and drawbacks of pursuing each path, and provide them with the clarity and understanding they need to make the right decisions

Complementing this, Bravo (2023) conducted a data visualization-based study to investigate student dropout trends at a state university in Northern Philippines from 2019 to 2023. The study used secondary data from the Registrar's Office to reveal the impact of socioeconomic factors on student retention, emphasizing that financial hardship remains a key contributor to dropouts even under policies like the Free Tuition Act. The study suggests that incorporating qualitative methods such as interviews could deepen understanding of the underlying reasons for student attrition. Furthermore, longitudinal tracking and regional comparisons are recommended for identifying long-term patterns and systemic causes of dropouts.

The Philippines shifted from a 10-year to a K-12 basic education program, adding two years to the students' basic education. Where students from the grade 10 level must choose a career path that will give them skills to be future-ready. Selecting the SHS track according to their preference and capabilities is crucial in their life decisions; this study will serve as a recommender system in order to avoid a strand or track mismatch. The researcher collected related data from public schools in the city of Valencia. These data are family background, the Big Five personalities, grades, IQ tests, and some tools from the Department of Education. These data were filtered and classified using machine learning algorithms such as Decision Trees (DT), K-Nearest Neighbor (KNN), Naïve Bayes (NB), and Support Vector Machine (SVM).

Based on the study of *"DALAN: A COURSE RECOMMENDER FOR FRESHMEN STUDENTS USING A MULTIPLE REGRESSION MODEL"* their study was conducted to

develop a tool that uses multiple regression to forecast incoming college students courses or programs to pursue in college. The prediction model was based on the identified predictors, such as the Cumulative Semestral Grade Point Average of all College of Information Technology and Engineering students from S.Y. 2013 to S.Y. 2016. The data needed are entrance exam results, high school grade point average, and grade point average (GPA). As a result, their predictive model was considered appropriate for course recommendation. Which contributes to the collected data of the current researchers in terms of SHS GWA, entrance exam results in each category, and the correlation in the pointing system of both the high school and college university.

Synthesis of Local Literature

Local research emphasizes how socioeconomic background and cognitive profiling both influence students' educational decisions and retention. According to Mendoza et al. (2022), financial capacity, employability concerns, and the National Career Assessment Examination (NCAE) all have a considerable impact on students' preferences for college programs. Bravo (2023) found that even with free tuition schemes, financial hardship remains a significant factor in dropout rates. Additionally, local studies that use machine learning techniques show how academic records, cognitive indicators, and behavioral data can be combined to predict student performance and support education decision-making, reinforcing the role of predictive analysis in improving education outcomes.

Schools can more effectively support students' decision-making and perseverance in higher education by integrating cognitive assessment data (such as aptitude tests and numerical reasoning profiles) with institutional interventions and socioeconomic factors. These studies emphasize the value of data-driven advising systems by using Decision Trees, KNN, Naive Bayes, and regression models, to avoid program mismatch and increase the

potential for preparation of students in higher education. Overall, these studies support the development of an intelligent recommender system that uses cognitive, academic, and socioeconomic factors for guiding students into suitable academic pathways.

Related Foreign Studies

A relevant study by Dzamihuddin et al. (2024) titled *"Predicting Academic Success of College Students Using Machine Learning Techniques"* (Data, 9(4), 60) explored the application of machine learning models such as XGBoost and decision trees to forecast student success based on academic and socioeconomic data. The study followed the CRISP-DM methodology and emphasized the role of preprocessed academic records in predicting outcomes like graduation rates and dropout risks. While their research provides valuable insights into student performance prediction, it focuses primarily on retrospective academic data.

In contrast, the present study emphasizes a forward-looking, personalized approach by analyzing cognitive and skill assessments to recommend optimal college courses. This introduces a more individual-centered strategy that not only predicts success but also actively guides students toward suitable academic pathways, thus filling the gap left by prior research focused purely on outcome prediction rather than personalized course guidance.

Beaulac et al. (2019) conducted a comprehensive analysis using random forest algorithms to predict undergraduate students' academic success and major choices based on their course enrollment data from the first two semesters. Their study utilized a decade-long dataset from a Canadian university, encompassing every course taken by students over ten years. The researchers developed two classifiers: one to predict degree completion and another to forecast students' major selections. The random forest models demonstrated high

accuracy and provided insights into variable importance, aiding university administrations in understanding factors influencing student outcomes.

While their research offers valuable predictive tools based on early academic performance, it primarily focuses on retrospective academic data. This forward-looking approach seeks to guide students proactively, aligning their inherent strengths with suitable academic pathways, thereby addressing the gap in personalized course recommendation systems that consider individual cognitive and skill profiles.

Gil et al. (2020) conducted a comprehensive study employing data mining techniques to predict the academic success of first-year bachelor's degree students in a Portuguese higher education institution. Analyzing data from 9,652 students over ten academic years, the researchers identified 68 features encompassing socio-demographic information, social origin, prior education, special status, and educational pathways. They developed predictive models at three distinct stages: entrance, end of the first semester, and end of the second semester. Among the models tested, the Support Vector Machines (SVM) model demonstrated the best overall performance. The study highlighted that factors such as prior academic performance, study gaps, and age were significant predictors at the entrance stage, while current evaluation performance became more influential in later stages. The authors suggested implementing study support groups for at-risk profiles and creating monitoring frameworks to promote academic success.

While this study provides valuable insights into predicting academic outcomes using a broad range of features, it primarily focuses on retrospective data.

This foreign study by Abo-Al-Ez et al. (2021) compared multiple machine learning algorithms, including Random Forest, Logistic Regression, and Support Vector Machines in terms of their ability to predict academic success using diverse student datasets.

The emphasis was placed on evaluating model performance metrics such as accuracy and precision. Although the study did not concentrate on identifying the most influential features, it revealed that Random Forest and Gradient Boosting provided the highest prediction accuracy, which involves evaluating and comparing the effectiveness of various predictive models after applying feature selection techniques (Abo-Al-Ez et al., 2021).

A foreign study by Paula, Nogueira, Nonato, and Ariovaldo (2025) revealed that age, male gender, state or municipal school background, and long working hours were associated with the high dropout rates; some traditional indicators of disadvantage, such as low financial status , race, and limited parental support, did not consistently predict attrition. Furthermore, students admitted through affirmative action and those from less-educated families were observed to have lower probabilities of student dropout. The study was conducted on student dropout at the Federal University of Minas Gerais (UFMG) from 2012 to 2019 with the use of descriptive statistics and multilevel regression models to examine the extent to which dropout is correlated with the academic programs they are enrolled in and their socioeconomic backgrounds.

Most of the predictors consistently were connected to program characteristics, especially course selectivity. Students who are less selective and in less prestigious programs exhibited substantially higher dropout probabilities, ranging from 19 to 28 percentage points higher than those who are in highly selective programs. This study suggests that student dropout is not always the result of socioeconomic or academic disadvantage but also the student's strategic decisions to leave less valued courses in search of more prestigious and economically rewarding degrees. The study highlights the importance of considering both individual profiles and institutional characteristics in understanding persistence in higher education.

Synthesis of Related Foreign Studies

Studies from other countries agree that machine learning has the ability to revolutionize the prediction of student outcomes. Beaulac et al. (2019) used random forests to predict major choice and degree completion, Dzamihuddin et al. (2024) used XGBoost and decision trees to forecast dropout risks, and Gil et al. (2020) used SVM models to evaluate academic success based on a variety of factors. Also, Abo-Al-Ez (2021) proves that ensemble-based models, particularly Random Forest and Gradient Boosting, achieved the highest precision accuracy. In terms of educational applications, this shows the importance of model evaluation and comparison since accuracy and precision are the ones measured in performance when it comes to the prediction systems.

Paula et al. (2025) offered a distinctive viewpoint by demonstrating that student dropout frequently represents strategic mobility, with students switching from less selected programs to more prestigious ones, rather than being only the result of subpar performance or socioeconomic hardship.

By demonstrating that predictive analytics, when paired with cognitive and skill-based assessments, can go beyond risk detection and actively suggest the best academic programs, matching students' strengths with the most appropriate career pathways, these studies collectively offer a solid basis for this research. This strengthens the foundation for developing intelligent recommender systems that assist students in making well-informed, personalized choices, not only the academic risks.

Related Local Studies

This study explored the cognitive ability and career choices of 151 graduating senior high school students enrolled for the 2024-2025 school year in schools under the Special Geographic Area, BARMM, Philippines. Personal factors influencing students' career decisions were also identified. Student cognitive ability was measured using the Filipino

Intelligence Test (F.I.T.). Data was collected via a standardized test and a researcher-made questionnaire. It was analyzed using mean, Pearson product-moment correlation, and Spearman's rho. Descriptive statistics showed students' cognitive abilities in vocabulary, analogy, and numerical reasoning were generally moderate, indicating adequate but not exceptional proficiency. The most common career choices were teaching (knowledge-based), nursing (skill-based), and business management (entrepreneurial). Interestingly, no statistically significant correlations were found between knowledge-based careers and any cognitive ability domains. However, students with higher numerical reasoning scores showed a greater inclination toward skill-based professions.

The article titled "Forecasting Students' Success to Graduate Using Predictive Analytics" by Almonteros et al. (2024) explores the use of machine learning algorithms and feature selection techniques to predict the likelihood of students graduating on time. The study utilizes data from Caraga State University, incorporating demographic profiles, academic achievements, and college admission results. Various algorithms, including Decision Tree, Random Forest, Ensemble, KNN, Logistic Regression, SVM, and Naïve Bayes, were tested, with feature selection methods like LASSO Regression, Ridge Regression, and Genetic Algorithm applied to enhance predictive accuracy.

"Predicting Academic Performance of First-Year College Students in the Philippines Using Path Analysis" by Alipio, M. (2020), this study developed a model to predict academic performance based on factors such as the Senior High School (SHS) strand taken by students. It concluded that academic adjustment and performance are influenced by the SHS strand, aligning with the variables considered by the previous study.

A study by Kumar et al. (2025) titled “Predictive *Modeling of Student Learning Outcomes Through Cognitive and Emotional Skill Integration” states that* cognitive and non-cognitive skills both play a significant role in the patterns of learning of students. Which highlighted that cognitive abilities are the most used factors in terms of data evaluation to predict students' performance. But, this study aims to bridge the gap between cognitive and noncognitive abilities on learning capacities of students. ance, learning preference, and socio-emotional aspects. The approach adopted makes use of predictive analytics. It is deployed here as machine learning algorithms in the form of Logistic Regression (LR), Naive Bayes, k-Nearest Neighbors (k-NN), Decision Trees (DT), and Support Vector Machines (SVM) to classify the learners into very fast, fast, average, and slow learners. The algorithm of k-NN also achieved the highest accuracy classification and showed good robustness for learning the students' learning rates. This study proves that large datasets combined with advanced algorithms can be applied to a range of educational fields to support tailored learning strategies.

Synthesis of Related Local Studies

Research from the Philippines shows how psychometric evaluation and predictive analytics can complement one another to help students succeed. According to the BARMM (2024–2025) study, a preference for skill-based occupations is correlated with higher numerical reasoning scores, indicating that cognitive profiles might inform recommendations for career paths.

In order to increase accuracy, Almonteros et al. (2024) used a variety of algorithms (Decision Tree, Random Forest, KNN, SVM, and Logistic Regression) to further illustrate the prediction of machine learning in predicting graduation success. Additionally, senior high school strands have a major impact on college academic success, serving as an early program placement indicator, according to Alipio (2020).

Kumar et al. study highlights the cognitive and non- cognitive integration in prediction study, while the Machine Learning application applies an effective academic guidance. In the Philippine setting, these studies collectively demonstrate that combining cognitive and skill-based profiles with machine learning models can proactively direct students into academic programs that play to their strengths, enhancing retention and graduation rates.

Synthesis of both Local and Foreign Studies

The collective body of both foreign and local literature demonstrates a growing consensus on the crucial role of machine learning and deep learning in enhancing academic performance prediction and educational decision-making. Across various international studies, advanced models, particularly hybrid architectures like RNN-LSTM-RF and specialized deep learning models such as GRU, have shown superior predictive accuracy. These approaches not only capture temporal and nonlinear patterns in student data but also support early interventions and personalized academic guidance.

Additionally, both foreign and local related studies reinforce the value of feature selection techniques and algorithm comparisons in optimizing model accuracy. These works collectively validate the methodological foundation of the current research: using students' cognitive abilities and skills assessments to recommend tailored academic paths. This forward-looking, personalized strategy addresses existing gaps in traditional outcome-focused models, offering a proactive approach to student success.

In summary, the synthesis of all reviewed literature supports the application of integrated, ethical, and context-aware machine learning approaches to forecast academic performance and recommend individualized educational trajectories. These insights guide the

development of robust, student-centered systems designed to foster academic achievement and reduce dropout rates through intelligent, data-driven interventions.

Foreign Review of Related Systems

The use of machine learning in predictive analytics has been extensively explored in various domains. Bokonda et al., in their work *"Predictive Analytics and Machine Learning for Real-Time Supply Chain Risk Mitigation and Agility,"* reviewed 30 peer-reviewed papers over the last five years. The study highlighted techniques such as artificial neural networks (ANN), decision trees (DT), random forests, and logistic regression, which are commonly applied in education, engineering, and medical fields. Supervised learning emerged as the most utilized predictive analytics approach. The study emphasizes the effectiveness of ANN and DT in student achievement prediction. These findings align with the current study's aim of recommending college programs using machine learning based on individual cognitive and skill assessments.

In terms of educational data mining, the feedforward network (FFNN) is widely acknowledged as an architecture for modeling structured, unstructured, and non-temporal data. According to Merceron & Tato (2023), FFNNs are usually used in predicting the students' performance tasks due to the reason that they accept a fixed set of variables and propagate information forward through hidden layers. Since the study uses cognitive ability and skill-set indicators rather than time-series or sequential data, the FFNN is more suitable as a framework.

According to recent research, Feedforward Neural Networks (FFNN) are very good at forecasting academic achievement and suggesting educational paths based on static, structured data, like cognitive ability and demographic characteristics (Al-Emran et al., 2020; Huang & Fang, 2022). Without requiring temporal dependencies, FFNNs are excellent at

capturing intricate nonlinear interactions among independent input variables, in contrast to recurrent models (RNN, LSTM), which are intended for time-series data. Because of this, FFNNs are a viable model for the current study, which examines students' cognitive and skill profiles in order to suggest acceptable college programs.

Using student behavioral and educational backgrounds, Yurtkan et al. (2023) used a feed-forward neural network (FFNN) to predict students' success in a higher education setting. They discovered that FFNN is better suited for educational outcome prediction tasks with structured student profile data. In this study, the FFNN architecture is used to model the relationships in student profiling and program matching, since it targets program recommendation while also predicting students' ability to survive by mapping the students' cognitive and skill sets to a categorical recommendation of the program.

Synthesis of Foreign-Related System

All of the analyzed research demonstrates how well Feedforward Neural Networks (FFNN) model and forecast educational results using structured, non-temporal data. According to Huang & Fang (2022) and Al-Emran et al. (2020), FFNNs outperform conventional machine learning algorithms in predicting academic achievement and suggesting educational pathways because they are highly effective at capturing intricate nonlinear relationships among cognitive, demographic, and behavioral variables. The capacity of FFNNs to map student qualities, like cognitive talents and learning behavior, to academic achievement was further empirically proven by Yurtkan et al. (2023), who highlighted the models' suitability for educational prediction tasks involving structured student profiles. The theoretical foundations of FFNNs in educational data mining were developed by Merceron and Tato (2023) to support these findings. They claimed that because FFNNs propagate information in a single direction through hidden layers without the need for temporal dependencies, they perform better with static datasets than sequential ones. All of

these results lend credence to the selection of FFNN as the optimal architecture for the current study.

Local Review of Related Systems

In the Philippine educational landscape, several technology-driven systems have been developed to support academic decision-making and performance prediction using machine learning and data analytics. These systems reflect the country's increasing focus on evidence-based educational interventions.

One notable system is the web-based career track recommender developed by Robert et al., which utilizes a Deep Neural Network (DNN) to assist guidance counselors in helping junior high school students choose suitable Senior High School (SHS) tracks. The system analyzed academic records and socio-demographic data from 1,500 students, achieving an accuracy rate of 83.11%. This recommender system aimed to minimize mismatches in track selection, reduce dropout rates, and ensure more efficient allocation of educational resources. It demonstrated how integrating machine learning into guidance counseling could improve personalization and student outcomes.

Another locally developed system is presented in Bravo's (2023) study, which used a data visualization platform to analyze dropout trends at a state university in Northern Philippines. By leveraging enrollment and demographic data from 2019 to 2023, the system identified financial hardship as a primary driver of student attrition, even with the implementation of the Free Tuition Act. Though primarily quantitative, the system's visual analytics capabilities enabled institutions to detect and respond to dropout trends more effectively, supporting data-driven policymaking. This study mainly supports decision-making indirectly for the institution following the given data.

Another study conducted by Asor et al. (2022) with the title “Prediction of Senior High School Students’ Performance in a State University: An Educational Data Mining Approach” aims to predict the performance of senior high school students of Laguna State Polytechnic University–Los Baños (LSPU–LB) through data mining. In developing the prediction model, the researchers utilized well-known machine learning algorithms such as : Decision Tree, Naive Bayes, Random Forest, Neural Networks, and Linear Regression. Upon the development of their model, they found out that Naive Bayes attained the highest accuracy among the other remaining algorithms, which also indicates that it has a high probability of a satisfactory rating in Grade-11 semester subjects. Additionally, Neural Network shown a promising result in the prediction of students' performance. This study will be used in predicting the student program recommendation, expecting a high accuracy and probability just like in this previous study.

These local systems collectively underscore the role of machine learning and data analytics in refining educational strategies. Whether through career guidance, dropout monitoring, or performance prediction, the integration of intelligent systems enhances institutional capacity to deliver timely, personalized, and effective academic interventions tailored to Filipino learners.

Synthesis of Local Related System

In the Philippine educational setting, a growing number of technology-driven systems have been developed to enhance academic decision-making through machine learning and data analytics. These systems exemplify the country's shift toward data-informed educational strategies, aiming to provide personalized, timely, and effective interventions.

A common theme among the reviewed systems is the integration of machine learning algorithms such as decision trees, random forests, SVM, and deep neural networks. These

approaches have proven effective in improving the accuracy and efficiency of models designed to predict student performance, forecast graduation outcomes, and support academic guidance.

For instance, systems like Robert et al.'s career track recommender leverage deep learning to personalize senior high school track recommendations, while platforms such as those developed by Bravo (2023) focus on dropout trend visualization for institutional policy-making. Similarly, predictive models by Asor et al. (2023) showed multiple machine learning algorithms, including neural networks and linear regression, can achieve accurate predictions of the performance of senior highschool students. These approaches highlight the importance of selecting the proper algorithms and student variables to improve the prediction accuracy and educational planning.

Collectively, these local systems highlight the transformative potential of artificial intelligence and data analytics in Philippine education. By addressing issues such as track mismatches, dropout risks, and academic underperformance, they empower institutions to make proactive, evidence-based decisions that cater to the diverse needs of Filipino learners.

Overall Synthesis

The review of both domestic and international literature together creates a clear path toward individualized, research-based educational interventions and implemented educational systems. Philippine studies demonstrate the significant impact of cognitive profiling and socioeconomic environment on student judgments, while international research machine learning models, particularly those that utilize Feedforward Neural Networks , regression techniques, and Random Forest models achieve greater predictive accuracy when applied in predicting the academic performance of students, career recommendations, and program matching systems.

Incorporating these insights, the current study suggests a personalized recommendation system and predictive and classification-based models for college programs that match individuals with the best academic courses based on an analysis of their cognitive capacities and skill sets. The present deficiencies are filled by this method, which goes beyond passive risk prediction to provide proactive, tailored counsel, promote prompt interventions, lower dropout rates, institutional decision-making and encourage academic achievement.

Author & Year	Approach	Key Contributions	Gaps
Kukkar et al., 2024	Hybrid ML (RNN + LSTM + RF)	Achieved 97% accuracy in predicting pass/fail using student data; captured temporal and feature-based dependencies.	Focused on performance prediction,not on recommending specific academic programs based on cognitive abilities.
Rohman et al.	SMOTE + Logistic Regression + Random Forest	Achieved 95.74% accuracy, with higher precision, recall, and F-1 scores.	The prediction focused on predicting the student's academic performance and not cognitive profiling or program recommendation.
Dzamihuddin et al., 2024	XGBoost, Decision Trees	Forecasted graduation/dropout risks by factors affecting academic success.	Retrospective analysis only; no forward-looking course guidance based on cognitive/skill data.
	National Career Achievement Examination	College program preference of students and the type of professional they are really equipped for have a relationship with taking the NCAE test.	NCAE measures general aptitude but does not provide fine-grained cognitive profiling that could guide personalized course recommendations.

Gil et al., 2020	SVM, data mining	Classified students' likelihood of academic success based on input features.	No course/program recommendations based on traditional academic predictors.
Abo-Al-Ez et al., 2021	Multiple ML algorithms (RF, LR, SVM)	Identified top-performing algorithms (RF, GB) for academic success prediction.	Did not analyze feature importance or apply personalization in recommendations.
Paula et al. 2025	Quantitative analysis of dropout causes	Most of the factors affecting the student dropout rate are the less valued courses, economically rewarding degrees, and socioeconomic background.	Needed for early personalized academic and career guidance are the highlights.
Robert et al	DNN-based recommender system	Recommended SHS tracks with 83.11% accuracy using academic and socio-demographic data; improved guidance counseling.	Focused only on SHS track selection, not college-level program decisions or cognitive/skills alignment.
BARMM (2024-2025)	Descriptive research on academic proficiency and cognitive testing.	Student's performance and cognitive abilities based on numerical reasoning have greater correlation with the cognitive profiles and career decisions.	The results were not translated into a system for course matching.
Almonteros et al., 2024	ML ensemble models with feature selection (LASSO, Ridge, GA)	Predicted on-time graduation using diverse data; tested multiple algorithms for optimal accuracy. Used	Did not consider cognitive or skill-based variables for proactive academic pathing. Focused on graduation prediction

		demographic profiles, academic achievements, and college admission results.	only.
Alipio, 2020	Path Analysis	Showed SHS strand influence on college success	Limited to correlation, not a recommendation system developed
Bokonda et al.	Review of ML techniques (ANN, RF, DT, LR) in multiple domains	Supported GRU, ANN, and DT use in education; reviewed 30 studies; emphasized effectiveness of supervised learning for predictive education analytics.	Cross-domain focus; general findings, not deeply tailored to educational settings.
Aljohani	Real-time ML analytics (RNN, anomaly detection, classification)	Promoted adaptive systems for early intervention and dynamic alerts, applicable to educational tracking.	Primarily logistics/supply chain context; not tested in educational deployment.
Beaulac et al., 2019	Random Forest and regression models	Predicted degree completion and major choice based on sociodemographic data and academic data.	Focused on completed course data, excluding cognitive ability measures
Bravo, 2023	Data visualization	Analyzed dropout trends, highlighted financial hardship as a major factor, and informed educational policy.	Quantitative only; no ML modeling or predictive recommendation.
Al-Emran et al. (2020); Huang & Fang (2022)	Used **FFNN** to predict academic performance from cognitive and demographic data.	Proved FFNN's high accuracy in modeling nonlinear relationships for educational prediction.	Focused on performance forecasting, not program recommendation.

Yurtkan et al. (2023)	Applied **FFNN** for student success prediction using behavioral and academic data.	Confirmed FFNN's suitability for structured student profiles and improved prediction accuracy.	Limited to outcome prediction; lacks personalized recommendation aspect.
Merceron & Tato (2023)	Provided a **conceptual framework** on FFNN use in educational data mining.	Explained FFNN's strength in handling static, non-temporal data for student modeling.	Theoretical only; no empirical system implementation.

Chapter III

Theoretical Background

Theoretical Framework

This study is grounded in establishing several frameworks, theories, and models that will collectively evaluate the factors influencing academic decision-making and evaluate the development of a predictive model using machine learning.

Expectancy-Value Theory

Students establish their decisions based on the implied outcomes. Expectancy-Value Theory (Leaper, 2011) states individuals make decisions based on the value they expect from the outcome and belief in their ability to succeed. In essence, people's choices are rooted in (1) their expected performance and (2) seeing a meaningful value in the future result. The decision is motivated with an achievement in mind, and the options are mainly impacted by the ability to succeed or future value. (Wang & Xue, 2022)

In this study, the theory explains that students are likely to choose a college program that adheres to their prior experiences or future goals. For example, a student who once actively participated in a volunteer role in a hospital or medical outreach is likely to see a potential value and confidence in entering a health-related program such as nursing or medicine. Their expectancy of success and value of the outcome guide their academic decisions.

This theory provides a psychological foundation for understanding why students select certain college programs, making it a crucial variable in modeling decision-making behavior. The predictive model developed in this research incorporates cognitive abilities (linked to expectation of success) and skill set (as indicators of value and interests), making

expectancy-value theory fits as a conceptual basis. It reinforces the rationale for matching students' cognition and skill with recommended programs.

Information Overload Theory

Due to oversaturation and information bombing, decisions can be impaired, leading to uncertainty and apprehensive emotions towards their decisions. Many conclude doubtful outcomes of their choices, eventually leading to poor performance or low satisfaction with their final choice. In the 15th century, the overwhelming information formed in the advent of the printing press; however, "information overload" was first coined in 1964 (Jones, n.d.). Information overload suggests exposure to a plethora of choices can cause overwhelming stress from choosing, decreasing a person's decision-making ability. The ability to process information exceeds its capacity; any information after that point will be ignored by the cognition.

Information overload underscores the importance of filtering information during decision-making. It justifies the need for a data-driven recommendation system that processes complex data and narrows it down into a manageable set of options, reducing the cognitive burden.

The predictive model aims to alleviate the effects of information overload by using machine learning to generate college program suggestions based on a student's unique cognitive skill profiles. Through the statistically-limited options, the system minimizes the overwhelming choice and improves decision confidence and address a real-world problem by the theory

Together, Expectancy-Value Theory and Information Overload Theory, this study offers a balanced psychological foundation. By combining these theories, it explores the potential of machine learning as a career guide to undecided students by recommending the

most potent option as a suggestion to help them in their academic decision-making and career path. Machine learning models will serve as potential model for evaluating the best option

Zero Rule Accuracy

The ZeroR classification is one of the fundamental baselines for validation in predictive modeling. As noted by Termmedi et al. (2023), this method relies solely on the target, disregarding other independent predictors (features). Theoretically, ZeroR is beneficial in establishing the null accuracy floor. By defining the performance of a model that only predicts the majority class, researchers can rigorously evaluate whether advanced algorithms (such as Random Forest or Naive Bayes models) are genuinely learning the underlying patterns of the dataset or merely guessing based on the class imbalance.

Stratified K-Fold Cross-Validation

A standard cross-validation can produce bias especially in imbalance datasets where one class dominates the other in terms of number by generating a training and validation fold that lacks the minority class.

To ensure robustness and unbiased evaluation, the system utilized a Stratified K-Fold Cross-validation. According to Prusty et. al. (2022), K-Fold CV ignores data division, it ensures that the data point appears exactly once but in the training set, it could appear in ‘k-1’ times. This could lead to repeated appearance and extended evaluation. Tanimu et al. (2021), introduce the stratified strategy. This method guarantees that each fold of the dataset contains the same proportion of observations with each label. The folds guarantees of keeping the fraction of sample with each class constant on a stratified class distribution

FeedForward Neural Network

Neural networks are a machine learning model derived from the human brain's neural structure, with biological neurons functioning as nodes and terminals. Neural networks replicated the neurological networks of the human brain; hence, they are generally referred to as artificial neural networks. This computational model has been used for contemporary applications such as artificial intelligence and deep learning for various tasks, such as classification, pattern recognition, and prediction.

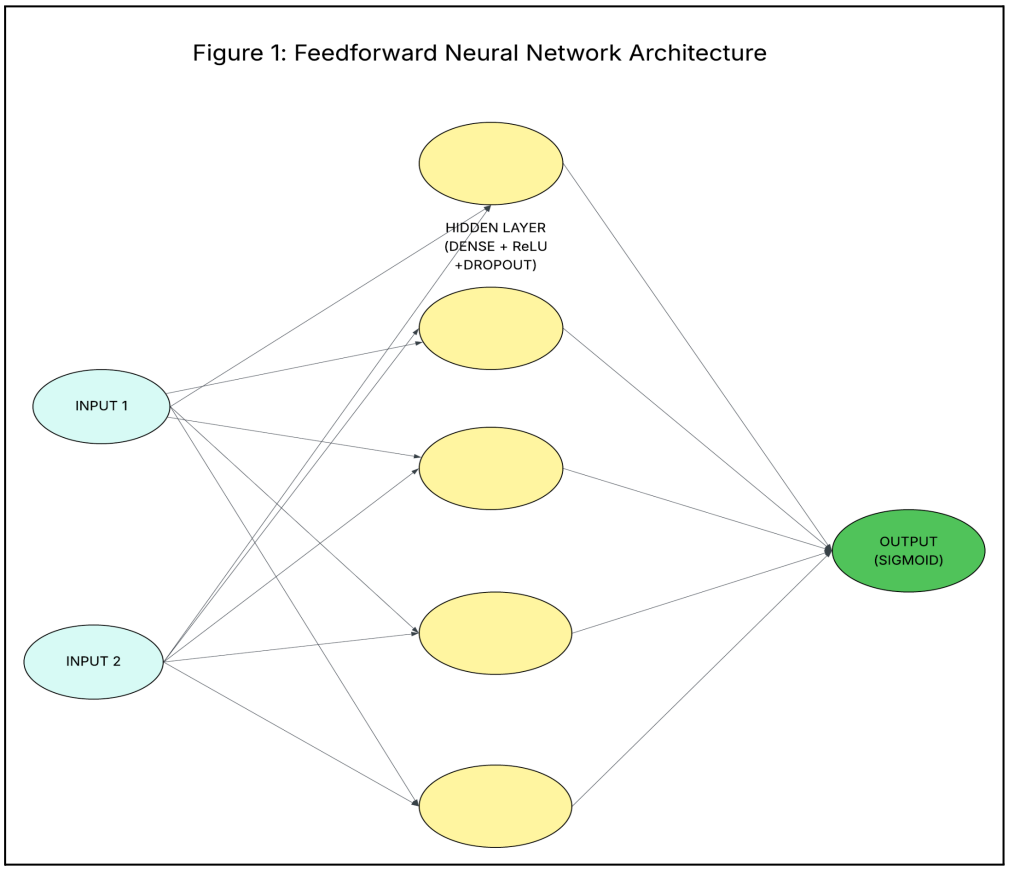


Figure 1. Feedforward architecture

A feedforward neural network is a type of artificial neural network that moves in only one direction—forward. This model has been selected for its ability to handle mixed data, complexity learning, and nonlinear relationships among multiple feature inputs. The FFNN model is effective in learning nonlinear patterns and interactions between variables and

influence that is helpful for prediction. Its reliance on large datasets necessitates the use of proper scaling and dropout regularization to reduce overfitting. Despite being less descriptive compared to rule-based models, it provides high accuracy and adaptability, aligning well with the study's objective of a personalized college program.

Random Forest Algorithm

Random Forest is a machine learning algorithm that uses a collection of decision trees and combines the output to create a prediction. The algorithm is based on a set of decision trees on training data where outputs are based on the higher accuracy tree and majority vote of decision trees.

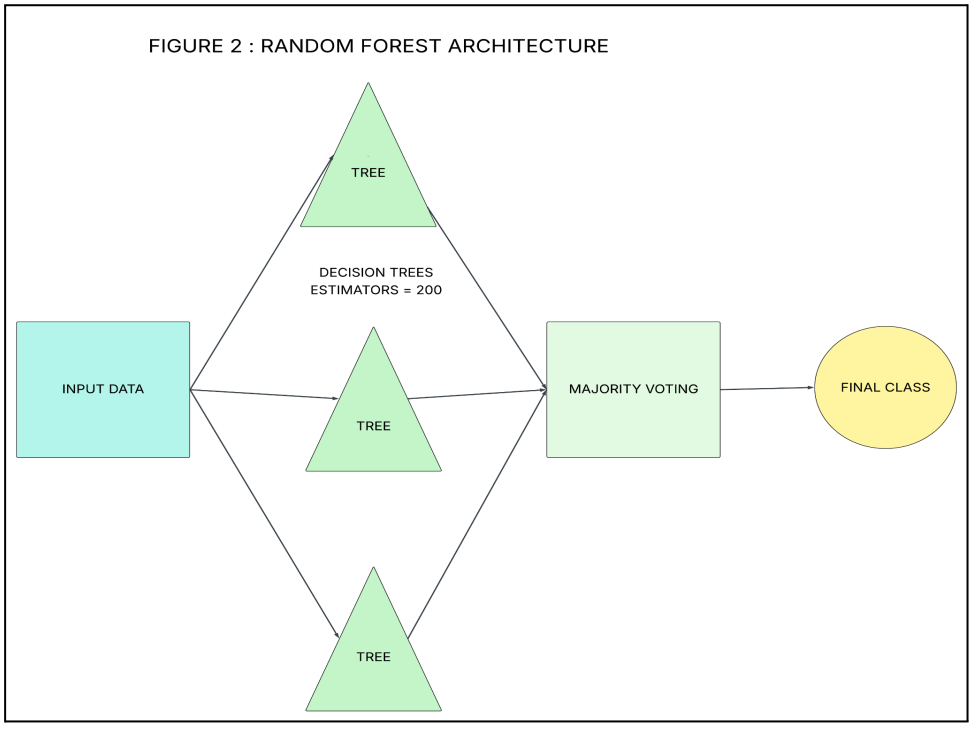


Figure 2. Random Forest Architecture

Random forest is selected as one of the predicting models for its versatility and consistency, as it has a high classification rate and can handle noise and outliers, and overfitting has a lesser chance to accumulate. The random forest algorithm has criterion for

its hyperparameters, which is the split quality measurement function; max_depth, the maximum depth for the tree; min_sample_leafs, which is the minimum sample number on the leaf to decide; min_samples_split, the minimum sample to decide a split; and lastly, n_estimators, the number of decision trees to be built on the random forest

Naive-Bayes Classification Algorithm

Naive Bayes is a machine learning algorithm based on Bayes' Theorem that predicts probabilistic categorical data points, assuming all features are independent of each other, hence the term "naive."

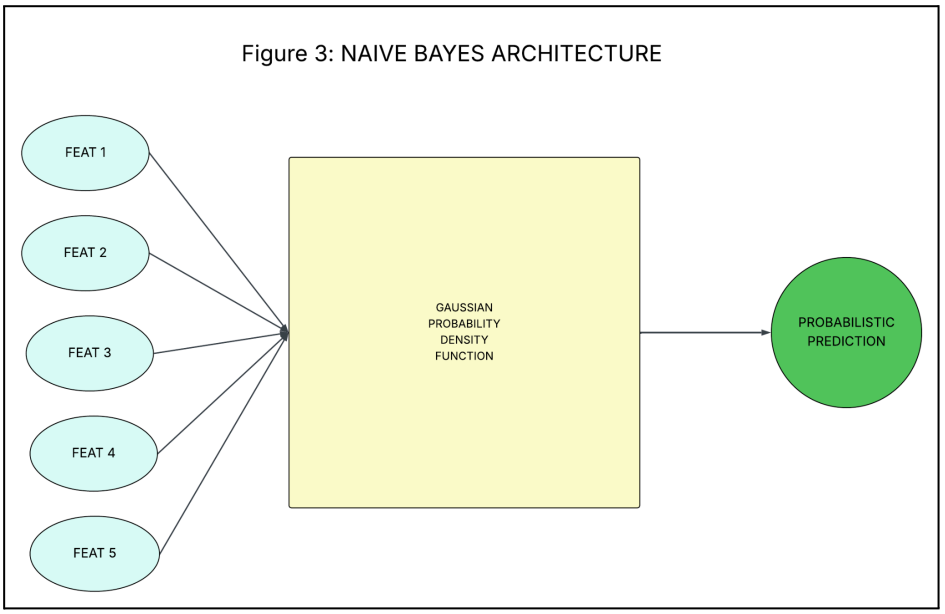


Figure 3. Naive Bayes Architecture

Naive Bayes is selected as a predicting model for the independence of the features, treating them equally with no relation to each other; thus, each feature contributes to the prediction. Naive Bayes is proven effective for high-dimensional tect classifications and predicts faster compared to other classification algorithms.

System Architecture

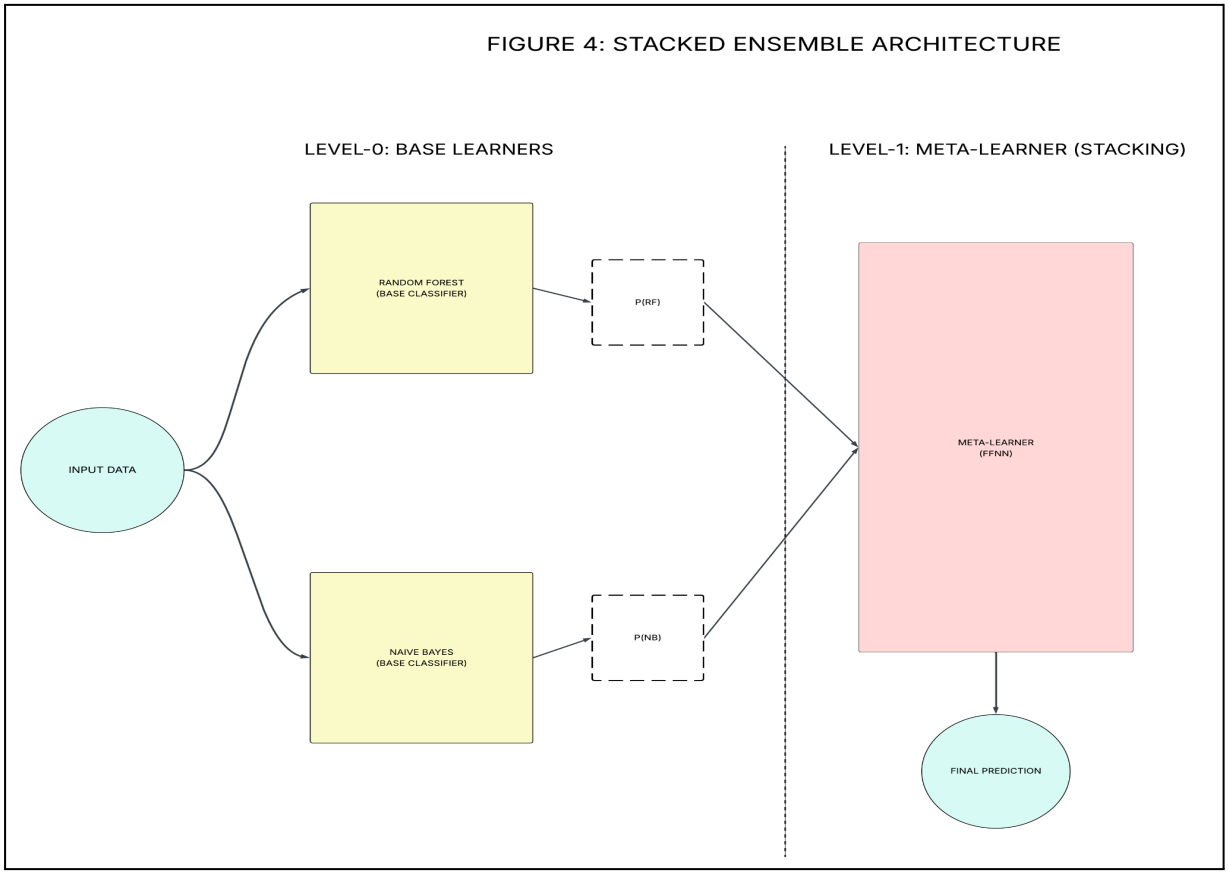


Figure 4. Stacked Ensemble Architecture

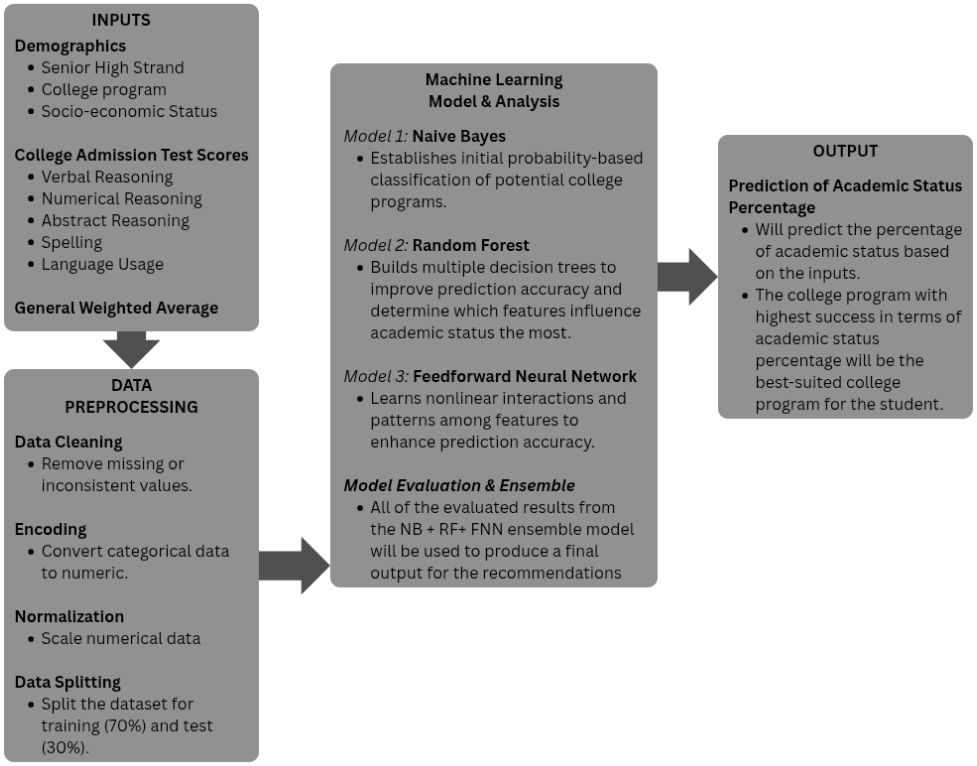
Figure 4 shows the two-tiered Stacking Ensemble architecture designed for the recommendation system. The architecture consists of two (2) layers: a Level-0 layer consisting of the Random Forest classifier, which utilizes a voting strategy across multiple decision trees, and a Gaussian Naive Bayes mode operating on the assumption of Gaussian distribution. The Level-1 layer consists of Feedforward Neural Network, which functions as a meta-learner by synthesizing the predictive outputs from Level-0 models to generate the final classification.

Chapter IV
Methodology

Conceptual Framework

This study will be adapting an Input-Process-Output (IPO) framework to design a predictive system that will recommend a best-fit college program that will best suit the students' cognitive abilities and skill sets.

With the help of this conceptual framework, we will be able to identify the focus of our study by identifying key input variables and establishing their relationships. Also, it will serve as a guiding structure that will keep our research aligned with its objectives throughout the study.



Variables

In conducting this study, it is essential to clearly identify the variables that will be used for analysis and to draw meaningful conclusions. This section presents key variables relevant to the study, distinguishing between the independent and dependent variables.

In this study, the independent variables will be the collection of four types of data: **demographics** *(senior high strand, college program)*, **test results from the College Admission Examination** *(verbal, numerical, abstract, spelling, language usage)***, General Weighted Average** *(GWA),* **and Soft and Hard Skills.** These variables will be used to provide a comprehensive profile of each student.

The dependent variable in this study is the outcome once all of the independent variables have been processed. The result will be the **recommendation of a best-fit college program with the highest percentage result of success in terms of academic status**. This recommendation is derived from a comprehensive analysis of the input data, ensuring the alignment with the students' profile.

Research Design

This study adapts an applied research approach, aiming to develop a machine learning-based system that will assist in the prediction analysis of best-fit college program selection. The system is designed to help students in making informed decisions based on the results of their demographics, college admission tests, general weighted average, and soft and hard skills.

To achieve this objective, this study will be using multi-machine learning models that can be utilized in processing the data independently.

Naive Bayes (NB) will be the baseline classifier with its simple, interpretable, and fast baseline to be used for comparison. This model treats all features independently. The Random

Forest (RF) that will be used for a better prediction through classification of the data and averaging the regression that helps in improving accuracy and reducing error. The RF will be trained to classify the students into their potential academic field by analyzing the extracted result from the collected data.

The other machine learning model that will be used in this study is the Feedforward Neural Network (FNN) that can also be used in classification, regression, and prediction analysis of the input data.

This study will also be utilizing a quantitative method for collecting and analyzing numerical data from students, including their college admission test scores, demographics, and general weighted average.

Participants

In this study, the researchers will be employing a strategic sampling technique to randomly select the participants for this study. For the purpose of this research, the following categories will be considered with the specific qualification criteria:

- 300 randomly selected fourth-year students that are currently enrolled at the College of Sciences from Palawan State University-Main Campus.
- Another set of participants will be included in this research; the participants will be students from the College of Sciences at Palawan State University - Main Campus who are currently enrolled within the academic year. These students are targeted to collect their data about their decision-making upon enrolling in college.

Data Collection Instrument

In this study, the researchers will be needing research instruments to gather relevant information for developing a machine learning-based college program recommendation system.

For the primary data collection, the researchers will utilize a structured questionnaire administered through Google Forms. This approach ensures that the data gathered will be both credible and reliable for use in the study. This set of questionnaires is specifically designed to gather personal information like demographics, college admission test results, and general weighted average, and soft and hard skills.

Data Collection Procedures

The data collection for this study will be conducted in a systematic and structured manner to ensure the integrity and accuracy of the data collected in developing a machine learning-based college program recommendation system.

First, we will coordinate with the University President's Office, where formal letters will be sent to request the approval of conducting a data collection procedure to the targeted respondents of this study. Once the request is approved, the questionnaires will be disseminated to the respondents. After collecting all of the data, it will be sorted according to the different categories that will be fed into the system.

The researchers will be present during the physical survey distribution to ensure a high response rate and to provide assistance and clarification if there is a question from the students. The data collection will span three weeks.

Lastly, for the data compilation and verification, the completed questionnaires will be checked for accuracy and completeness. The incomplete or invalid responses will then be

discarded. Data collected will be encoded into the database for further analysis. Sensitive data will be anonymized to protect participant privacy.

Statistical Treatment of Data

To provide a valid and reliable interpretation of the gathered data, the researchers used the following statistical tool:

1. Attrition Rate

Formula:

$$\text{Attrition Rate (\%)} = \frac{\text{Total Enrollees (Admission Year) - Total Graduates (After 4 years)}}{\text{Total Enrollees (Admission Year)}}$$

Source: CMO No. 46, s. 2012

Where:

- **Total Enrollees (Admission Year):** the total number of students who first enrolled for a specific program in a given starting academic year.
- **Total Graduates (After 4 years):** the number of students from the specific starting academic year who successfully completed all degree requirements and graduated.

2. Linear Regression

Formula:

$$y = mx + b$$

Where:

- **y:** is the predicted attrition rate.
- **x:** is the academic year of the input attrition rate
- **m:** average yearly change in attrition rate
- **b:** the starting attrition rate at year zero

3. Zero Rule Baseline

Formula:

$$\text{ZeroR Accuracy} = \frac{\text{Frequency of Majority Class}}{\text{Total Number of Instances}} \times 100$$

Data Analysis Plan

This study employs a multi-stage data analysis plan to ensure a systematic approach in analyzing the collected data. The process involves an integration of both statistical and machine learning techniques that will support the study's predictive goals.

The data gathered through the Google Forms questionnaire will undergo a structured and systematic data analysis process to ensure accuracy, consistency, and reliability of the results. The gathered data, which includes demographic profiles, college admission test results, general weighted average (GWA), and soft and hard skills will be first cleaned and preprocessed in order to prepare the dataset to be reliable. This process involves data cleaning to assess any missing and inconsistent entries using standard data cleaning techniques. Also, it involves the encoding and normalization where the responses in the category will be encoded and formatted appropriately for machine learning analysis.

The core of the analysis will involve the implementation of three machine learning models—Naive Bayes (NB), Random Forest (RF), and Feedforward Neural Network (FNN). These models will serve as the primary computational tools for predicting the students' best-fit college program that have a high percentage in academic status, classified as either regular or irregular. The random forest and Naive Bayes model will be trained and tested using the same dataset but will function independently, allowing each to analyze the data according to its own computational mechanism and learning principles. After their prediction, the feedforward model acting as a meta-learner will take their input and make the final prediction.

The Naive Bayes will treat each variable independently, and it will calculate the probability of each category. As for the numerical data, such as college admission test results and general weighted average, the NB will assume a distribution and calculate probabilities of a student belonging to a specific academic status category.

Another model to be used is the Random Forest, which will build many decision trees for the input data and average their prediction to improve accuracy and reduce overfitting. This approach allows for the identification of complex decision rules and provides a ranking feature importance; this is to highlight which factors contribute significantly to classification outcomes.

Lastly, the Feedforward Neural Network (FNN) will be implemented to capture the non-linear and multi-dimensional interactions between input data. Through this multi-layered approach, the FNN is capable of modeling subtle and non-linear relationships, giving significant insights into the academic performance patterns.

After each model generates its individual predictions, the study will employ an ensemble strategy to determine the final classification output. In this ensemble approach, the results from the two models will be utilized as input for the meta-learner model, and the college program with academic status that receives the highest percentage outcome will be considered as the final system output. This method enhances predictive reliability by combining the strengths of each model.

The performance of each model will be evaluated using standard machine learning metrics such as accuracy, precision, recall, and F1-score. Confusion matrices will be generated to visualize the classification results and assess the correctness of predictions for both categories. These following metrics will be used to quantify how effectively the models classify students into regular or irregular categories. The formulas used are as follows:

- ***Accuracy*** , which is used to measure the proportion of correct predictions. It is expressed as:

  $$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

- ***Precision***, which measures the number of positive predictions made that are actually correct. This is given by:

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

- ***F1 Score,*** is used when we need to get the harmonic mean of precision and recall. It is calculated as:

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

- ***Recall***, which evaluates how many of the actual positive cases were correctly identified. It is computed as:

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

  Where:

  TP = True Positive,

  TN = True Negative,

  FP = False Positive,

  FN = False Negative.

Then the statistical and graphical analysis will also be performed to interpret model performance and validate the robustness of the ensemble's prediction outcomes.

Lastly, for the **interpretation and discussion phase,** the findings will be synthesized within the broader context of the research objectives. Trends, patterns, and outliers will be explained, and the findings will be contextualized with existing literature to offer a theoretical foundation.

Ethical Considerations

In conducting this research study, it is essential to maintain ethical standards throughout the data collection process.

First, we need to provide **informed consent** to the participants with the complete information on the purpose, goal, procedure, and likely benefits and risks of this study. Second, this research study is **voluntary participation,** and students can withdraw from the study at any moment without any penalty or consequence. Third, this study will impose **confidentiality and anonymity,** where the personal data and answers of participants will be maintained confidentially and identifiable data will be anonymized.

Fourth, when it comes to data **privacy and security,** all of the collected data will be safely stored in password-protected devices and files to which only the researchers have access. Printed documents will also be kept in locked containers to ensure that there's no unauthorized access. Also, the collected data will be employed exclusively for this research and academic reportage. It will be neither revealed to third parties nor for commercial use.

Lastly, for the **compliance with legal and institutional guidelines,** this research will strictly follow the university's set of ethical standards and be in accordance with the Data Privacy Act of 2012 (Republic Act No. 10173) of the Philippines.

Chapter V

Results and Discussion

Overview

The results of the testing and training of the Random Forest, Naive Bayes, and FeedForward Neural Network (RF+NB→FFNN) machine learning ensemble for the diagnostic and prescriptive prediction for college programs are tabulated and discussed in this chapter. The primary objective of this study is to develop an ensemble of machine learning models with high accuracy in predicting the best college program for incoming college students.

Data Description and Preparation

To train and test the model for the best-fit recommendation system, the demographic data and academic and skills data were obtained from participants. The process gathered three hundred (300) respondents from five (5) programs under the College of Sciences. The dataset comprises a total of 300 training sets, consisting of 213 regular and 87 irregular students that are the basis of the academic success rate.

Throughout the data collection process, researchers uphold ethical standards and ensure the informed consent, anonymity, and confidentiality of the participants. Proper procedures are followed, safeguarding the rights and privacy of the participants.

After the collection process, the researchers tabulate and clean the collected data as part of the data preprocessing phase of the study. By manually encoding and analyzing the data collected, insights regarding the students' attributes were established. However, to prepare the raw dataset for the Stacked Ensemble model, a rigorous data cleaning was implemented. This involved detecting and removing non-numeric entries from the gwa column to ensure data integrity. Categorical variables such as “strand”, “soft skills”,

and “hard skills” were transformed into numerical format using a Label Encoder module for it to be interpreted by the algorithms. Ultimately, the numeric features were scaled properly to normalize the range of values, ensuring the model would not lean into bias to a variable with larger magnitude values.

In line with the Philippines Data Privacy Act of 2012, researchers ensure the confidentiality, integrity, and availability of the data collected during the survey. The processing of personal data was carried out in a fair, transparent manner and served a specific purpose. Participants were provided with clear written consent and informed of the nature and purpose of the data collection.

Linear Regression for Attrition Rate Projection

Increasing attrition rate is one of the major challenges in higher education institutions. In the College of Sciences at Palawan State University–Main Campus, it has exhibited a growing trend in the recent academic years. Attrition is defined as the percentage of the students who discontinue, shift, drop out, or fail to complete their academic program within the expected time period, i.e., four-year courses. It is a crucial indicator of institutional effectiveness and student success. According to the Commission on Higher Education (CMO No. 46, s. 2012), retention and completion rates are the key performance indicators that display both education quality and students’ academic placement.

In the context of this study, the attrition rate is analyzed and projected to determine its relationship with students’ cognitive and skill alignment to their chosen programs. Steady or consistent growth of the attrition rate may signify the student-program misalignment or indecisiveness in program selection.

Computation of Cohort Attrition Rate

The researchers obtained enrollment and graduation data from the College of Sciences for the last three (5) years. Using this data, the cohort attrition rate from each respective academic year was calculated.

Historical Cohort Attrition Rate Trend from 2019 to 2021

Enrollees		Graduates		Cohort Attrition Rate
Academic Year	**Total Number**	**Academic Year**	**Total Number**	
2017-2018	154	2020-2021	63	59.09%
2018-2019	406	2021-2022	159	60.83%
2019-2020	350	2022-2023	229	34.571%
2020-2021	379	2023-2024	180	52.50%
2021-2022	288	2024-2025	247	14.13%

Table 5.1 Historical Enrollment, Graduates, and Computed Attrition rate

This table presents the historical data for enrollees, graduates, and the derived attrition rate for the College of Sciences programs, including the academic years 2017-2018, 2018-2019, 2019-2020, 2020-2021, and 2021-2022. This data will serve as the basis for the survivability of the students who successfully finished the program requirements to completion

The first two (2) years, academic years 2017-2018 and 2018-2019, exhibit a high cohort attrition rate at 59.09 percent and 60.63 percent, respectively. For the attrition rate of the academic year 2019-2020, the attrition rate drops to 34.571 percent. The two initial academic years were pre-pandemic, having the normal mode of learning. Although their final years coincided with the early stages of the pandemic, many students had already dropped out, shifted programs, or were retained, resulting in an inability to complete the program within four years.

Meanwhile, the 2019-2020 academic year significantly experienced the impact of the pandemic, requiring students to complete their programs under a flexible mode of learning. During these years, the Commission on Higher Education (CHED) issued orders encouraging leniency to HEIs to ensure the completion of students. The flexible learning modalities, the encouraged academic leniency, and adaptive grading system resulted to lowering the cohort attrition rate to 34.571 percent

The succeeding year, the 2020-2021 academic year, experienced a sharp increase of cohort attrition, peaking at 52.50 percent. A primary reason for this high attrition rate was the COVID-19 pandemic, which required a sudden shift in learning modalities. The first two years coincided with an adaptive learning environment and lenient grading systems. However, toward the end of their program, the learning environment shifted back to face-to-face classes as well as the reinstatement of strict policies and grading systems. This left many students struggling to cope with the sudden increase of academic rigor.

Lastly, the cohort attrition rate of the academic year 2021-2022 shows a data challenge, as the number of graduates is almost equivalent to the number of enrollees. The cohort attrition rate was 14.23 percent, which is remarkably a sudden dip in the attrition trend compared to the preceding years. Upon further examination, it was found that the graduate figures include the delayed students. The apparent high attrition rate actually reflects a large number of students retained in their program but eventually graduating in the same year. Although the researchers attempted to isolate specific cohorts to rectify this, no disaggregated records were available. Therefore, the attrition rate for this academic year must be considered invalid.

Projection of Future Attrition Rate Using Linear Regression

A linear regression was employed to project the attrition trends of students within the College of Sciences for the five-year period from 2021 to 2026, utilizing historical data from the academic years 2017-2018 up to 2020-2021. Notably, the academic year 2019-2020 was excluded from the regression analysis as a statistical outlier. This period was treated as an anomaly due to external interventions, i.e., academic leniency that artificially suppressed attrition, requiring its exclusion to ensure the regression accurately reflects the standard attrition trajectory.

Linear regression values tabulation

x (academic year)	**y** (attrition rate)	**xy**	**x^2**
1	60	60	1
2	61	122	4
This academic year is excluded			
4	53	212	16
Total: 7	**174**	**394**	**21**

Table 5.2 Tabulation of Linear Regression Values

The application of this linear regression to the historical data (excluding the 2019-2020 outlier) provides a quantitative model to project student attrition trends within the College of Sciences. The resulting equation provides a data-driven insight into the relationship between academic years and the college attrition rate.

Interpretation of the Slope (m)

The slope is the most important result, as it indicates the trajectory of student success. By using this formula:

Formula:

$$m = \frac{n(\sum xy) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

Where:

- **∑xy:** sum of x (academic year) and y (attrition rate)
- **∑x:** sum of x (academic year)
- **∑x²:** the sum of **x²**

- **n:** number of x (academic year) in the table

Computation of Slope (m):

$$m = \frac{3(394) - (7)(174)}{3(21) - (7)^2}$$

$$m = \frac{-18}{7}$$

m = -2.57

The slope indicates that for every one-unit increase in the academic year, the value of the attrition rate is predicted to decrease by approximately **2.57%.**

Interpretation of b (y)

The y-intercept will also be computed to identify the projected value of the attrition rate when the academic year is zero. By having this formula:

Formula:

$$\mathbf{b = \frac{\sum y - m(\sum x)}{m}}$$

Where:

- **∑y:** sum of y (attrition rate)

- **m**: result of the slope

- **∑x:** sum of x (academic year)

Computation of Y-intercept:

$$b = \frac{174 - (-2.57)(7)}{-2.57}$$

b = 63. 99

b = 64.00

This y-intercept represents the value of the attrition rate when the academic year is theoretically zero. It shows that the baseline attrition rate at the starting point of the student's trend would have been **64.00%.**

This linear regression analysis establishes a relationship between the academic year (x) and the attrition rate (y) in the College of Sciences. As for the result of the equation $\hat{y} = 64.00 - 2.57x$ indicating a strong negative slope (m = -2.57). This shows that there's an inverse and dependent relationship between the academic year and attrition rate in the College of Sciences. As the College of Sciences progressed through successive academic years, the attrition rate would significantly decrease, as shown in the computation of the trend.

Linear regression

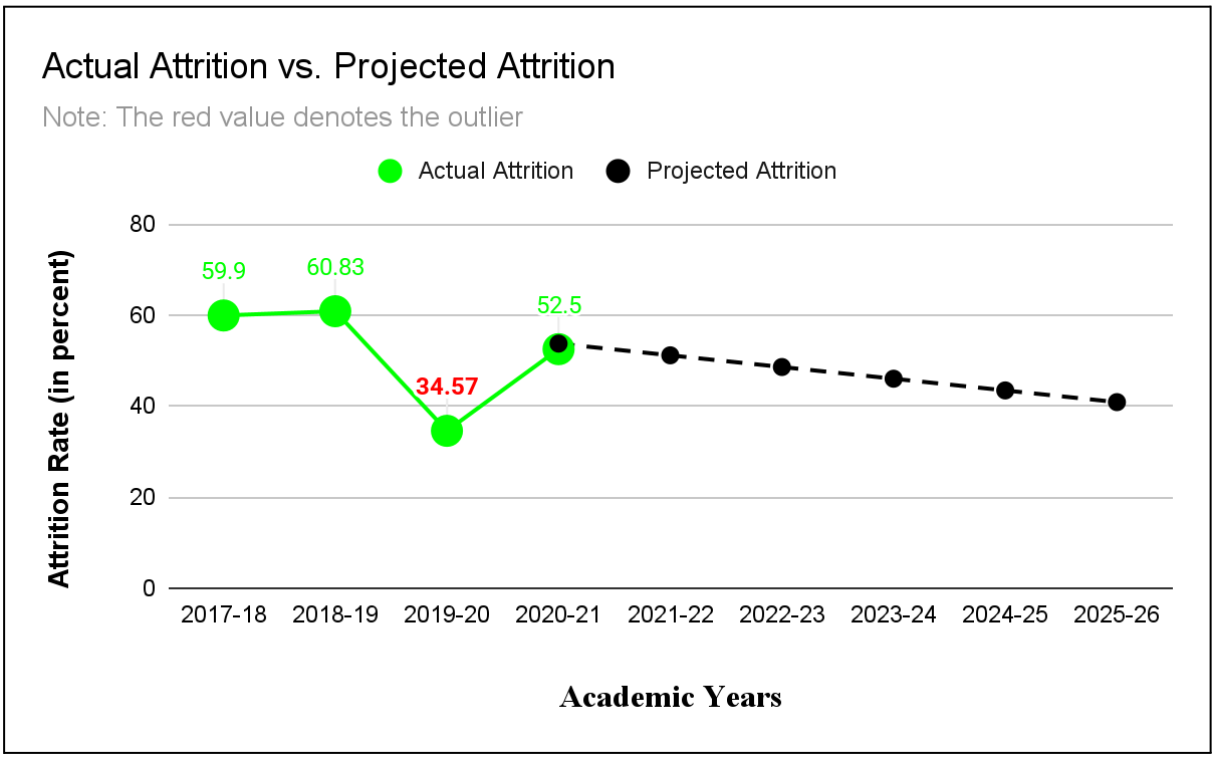


Chart 5.1 Attrition Rate Projection Using Linear Regression

This chart visually shows the data presented in Table 5.2, including the x (academic year) and ŷ (projected attrition rate) expressed as percentages. The chart illustrates the negative linear relationship found in the regression analysis; the attrition rate is projected to decrease as time progresses.

x (academic year)	y (y-intercept)	m (slope)	ŷ (projected attrition)
5	64.00	-2.57	51.15%
6	64.00	-2.57	48.58%
7	64.00	-2.57	46.01%
8	64.00	-2.57	43.44%
9	64.00	-2.57	40.87%

Table 5.3 Projected Attrition Rate Using Linear Regression

Based on the projection, in year 5, the projected attrition is at 51.15 percent. By year 9, the attrition rate is expected to drop to 40.87 percent. While the linear regression projects a gradual decrease in attrition, this trajectory assumes a stable environment that counters the historical volatility of the dataset. The attrition decay of -2.57% per year is merely theoretical given the historical spike, e.g., the surge from 34.571% to 52.5%; the attrition remains unstable.

The institution risks not only stagnation at this high baseline but a potential resurgence of attrition spikes, as seen in historical data, a volatility that the linear model cannot predict. Without a proactive mechanism, such as a recommendation system, to identify at-risk students on certain programs early, the institution remains vulnerable to these recurring patterns.

Indecisiveness Rate

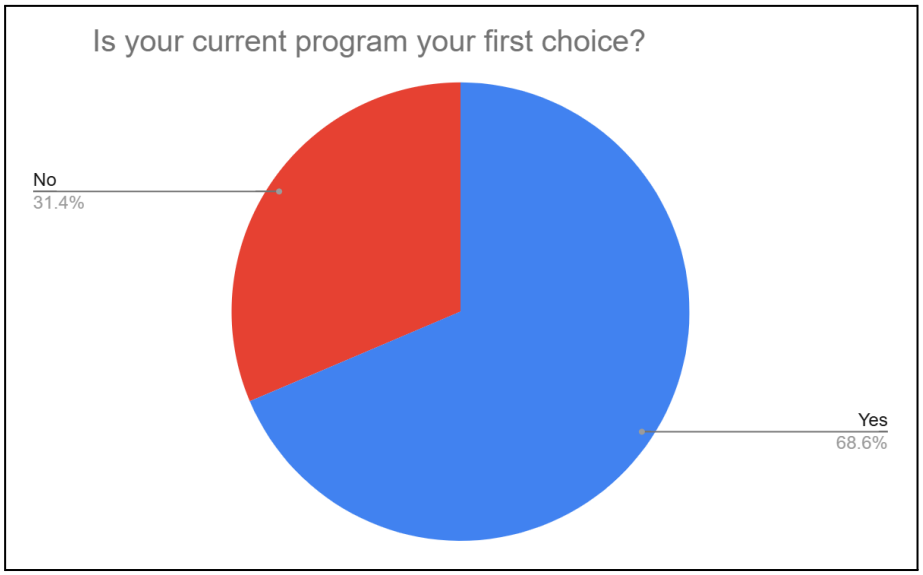


Chart 5.2 Indecisiveness Rate

This chart illustrates the indecisiveness rate of the students from the College of Sciences by answering whether the students' current program was their first choice. It shows here that a significant majority of the respondents, 68.6 percent, are currently enrolled in their first-choice program implying a high level of academic alignment and suggests that most students had a clear direction for their preferred field of study.

However, 31.4 percent of the respondents answered otherwise. Though it shows a minority, it still represents approximately one-third of the respondents. This indecisiveness can attribute to various factors such as drop outs, transfers, and shifting resulting to a higher attrition rate. This data indicates that while the majority of the respondents are satisfied with their initial choice, a substantial portion of the respondents may be at a higher risk of drop outs, and shifting due to being in a second-choice program and lack of commitment.

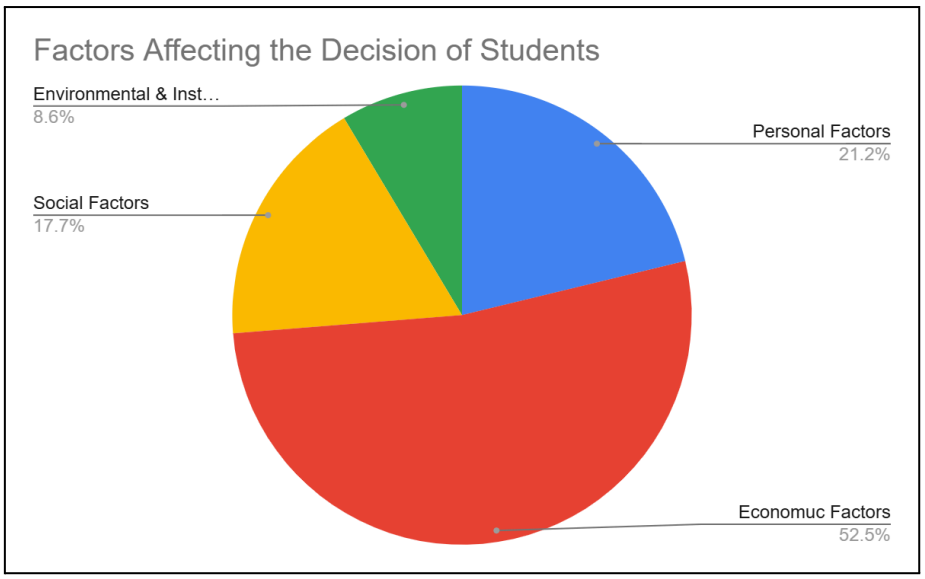


Chart 5.3 Factors affecting the decision in choosing a program

This chart illustrates the breakdown of factors that influence how students choose their college program. It shows that *Economic factors (52.5 %)* are the most influential driver. This indicates that students prioritize financial outcomes such as miscellaneous fees, scholarship availability, and the return on investment or job market demand after graduation. Followed by *Personal factors (21.2%)* and *Social factors (17.7%)* that when combined will make up approximately 40% of the influence. This highlights the role of individual interests and the pressure or guidance from a family or peers. The relative low impact of *Environmental & Institutional Factors (8.6 %)* indicate that the specific brand of school or the physical location is less important to students than the financial and social outcomes of the degree.

Implication of Indecisiveness Rate to the Attrition Rate Trend

The indecisiveness rate among students serves as a critical underlying factor that influences and sustains the high baseline of the attrition rate trend. The survey reveals that 31.4 percent of students are enrolled in programs that were not their first choice. This

indecisiveness rate represents a large cohort of students with potentially lower initial commitment, making them highly susceptible to attrition.

Survey data highlights the external pressures contributing to this vulnerability. Chart 5.3 identifies economic factors (52.5%) and personal factors (21.2%) as the primary factors affecting student decision-making in choosing their college program. For indecisive students, these external pressures act as catalysts that turn their initial lack of program commitment into a final decision to drop out, shift, or transfer.

While linear regression analysis indicates a favorable downward trend in attrition, the projected attrition rate in year 9 remains critically high at 40.87%. The 31.4% indecisiveness rate indicates a substantial portion of this high attrition, driven by students who may eventually leave their current program. Although the regression results show that the college is successfully reducing student loss by approximately 2.57 points annually, the structural indecisiveness rate imposes a roof for potential improvements. Consequently, targeted interventions, such as career counseling for the 31.4 percent of non-first-choice students, are necessary to break through the high baseline attrition.

By addressing these specific areas, the College of Sciences can potentially break the 40.87% projection. Successfully converting the 31.4% indecisive students into committed students, the annual reduction in attrition could exceed the current 2.57 percent slope., shifting the trend from a gradual decline to a significant recovery.

Student Data Analysis

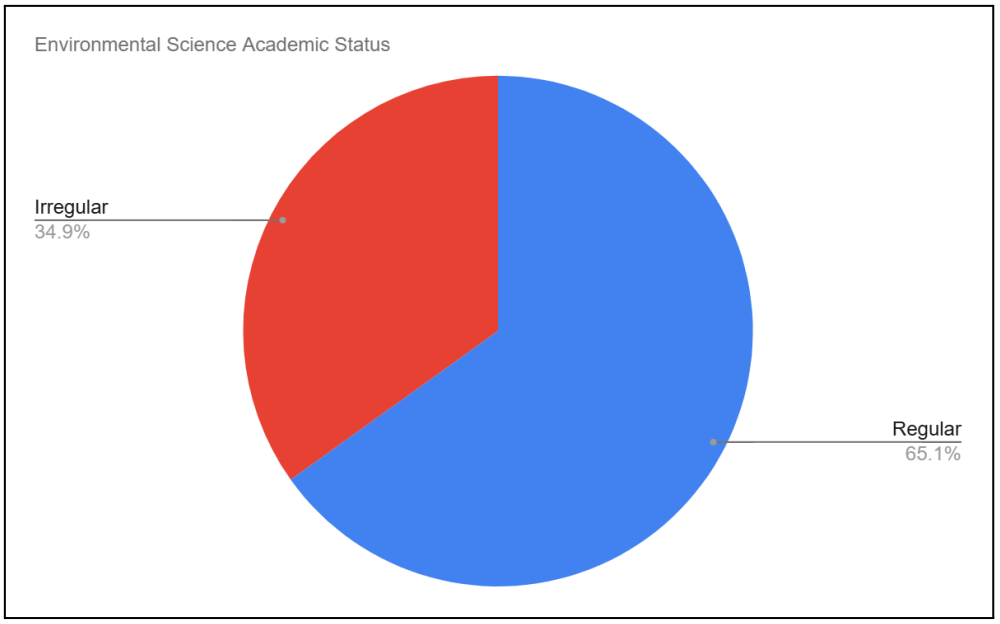


Chart 5.4 Environmental Science Academic Status

The College of Science comprises five programs. Chart 5.4 indicates that Environmental Science demonstrated a moderate level of student retention, with twenty-eight (28) regular and fifteen (15) irregular students. While the overall retention within the program remains satisfactory, the prevailing student irregularity often precedes attrition. The observed disparity emphasizes the need for targeted academic monitoring and timely intervention

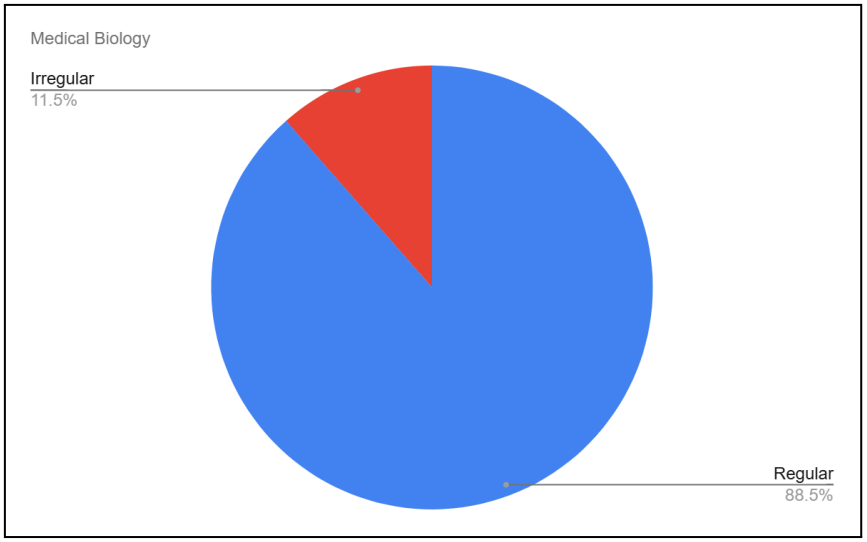


Chart 5.5 Medical Biology Academic Status

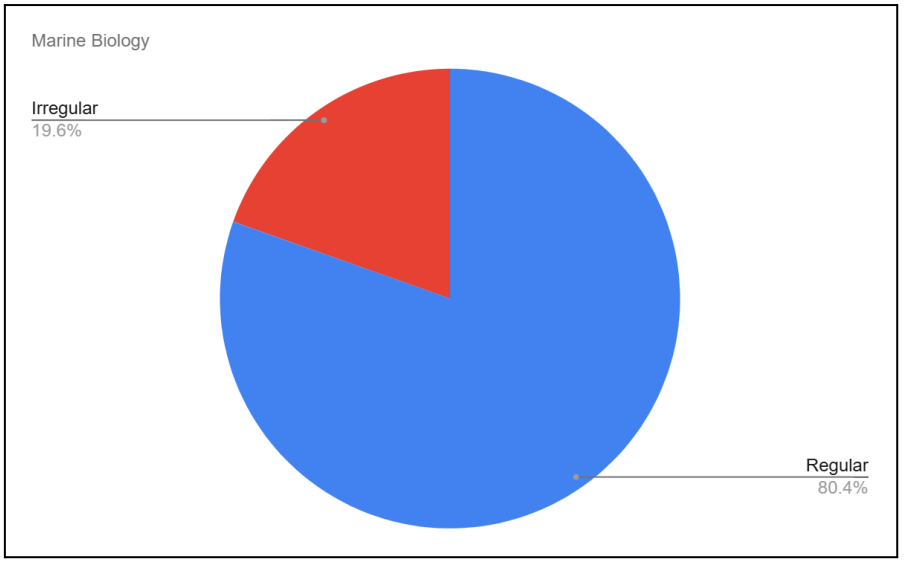


Chart 5.6 Marine Biology Academic Status

Charts 5.5 and 5.6, from Biological Sciences, indicate the highest levels of academic continuity. The Medical Biology program recorded seventy-seven (77) regular students compared to ten (10) irregular students, followed by Marine Biology with thirty-seven (37) regular and nine (9) irregular students. These strong retention outcomes are attributed to structured degree plans and the availability of extensive academic support systems, which help minimize deviations from prescribed program requirements.

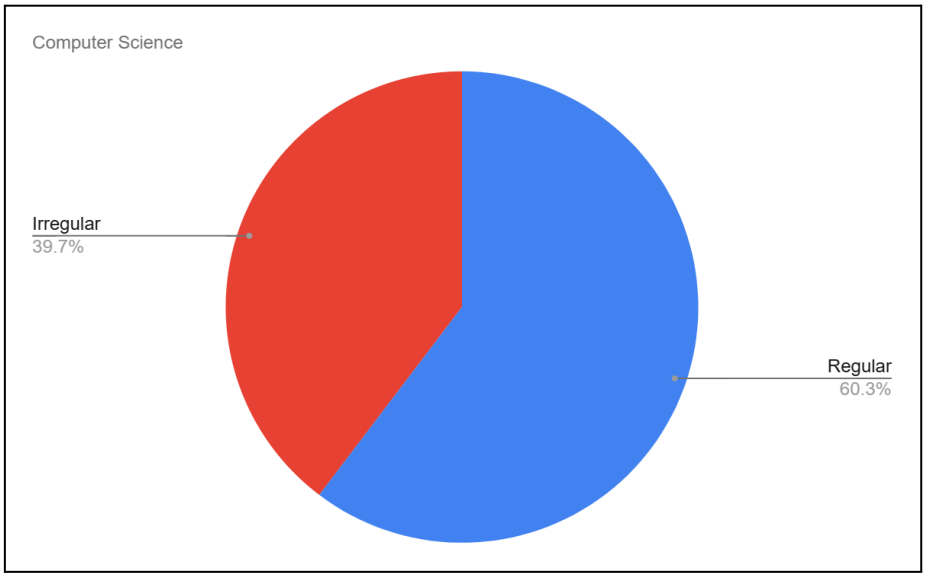


Chart 5.7 Computer Science Academic Status

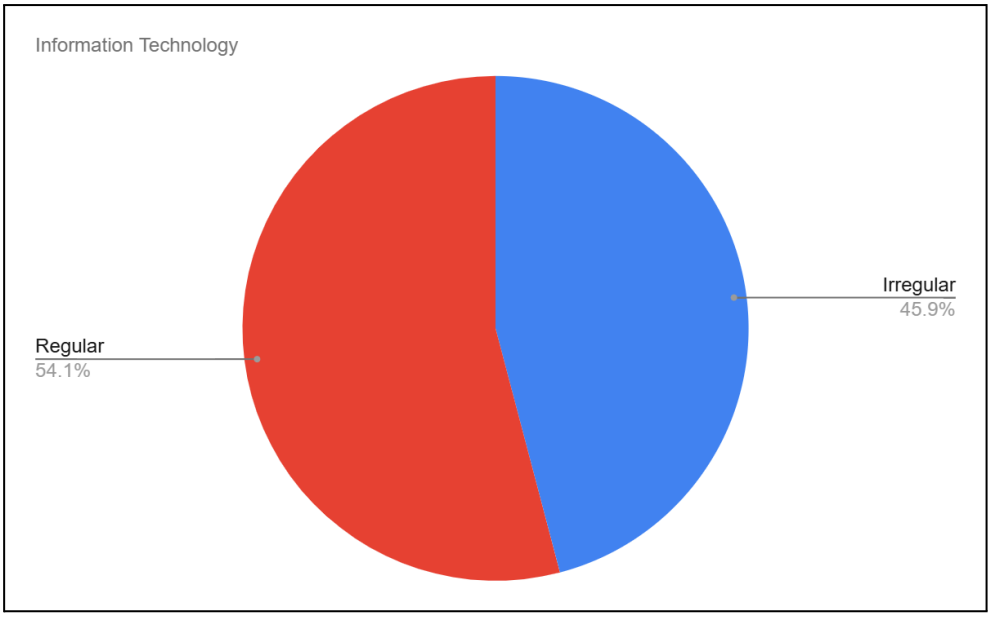


Chart 5.8 Information Technology Academic Status

In contrast, technology-oriented disciplines exhibited a narrower gap between regular and irregular students. Computer Science reported thirty-eight (38) regular and twenty-five (25) irregular students, while Information Technology recorded thirty-three (33) regular and twenty-eight (28) irregular students. The relatively higher incidence of irregularity in these programs may reflect the demanding nature of prerequisite-intensive coursework and the cumulative challenges of technical subject progression.

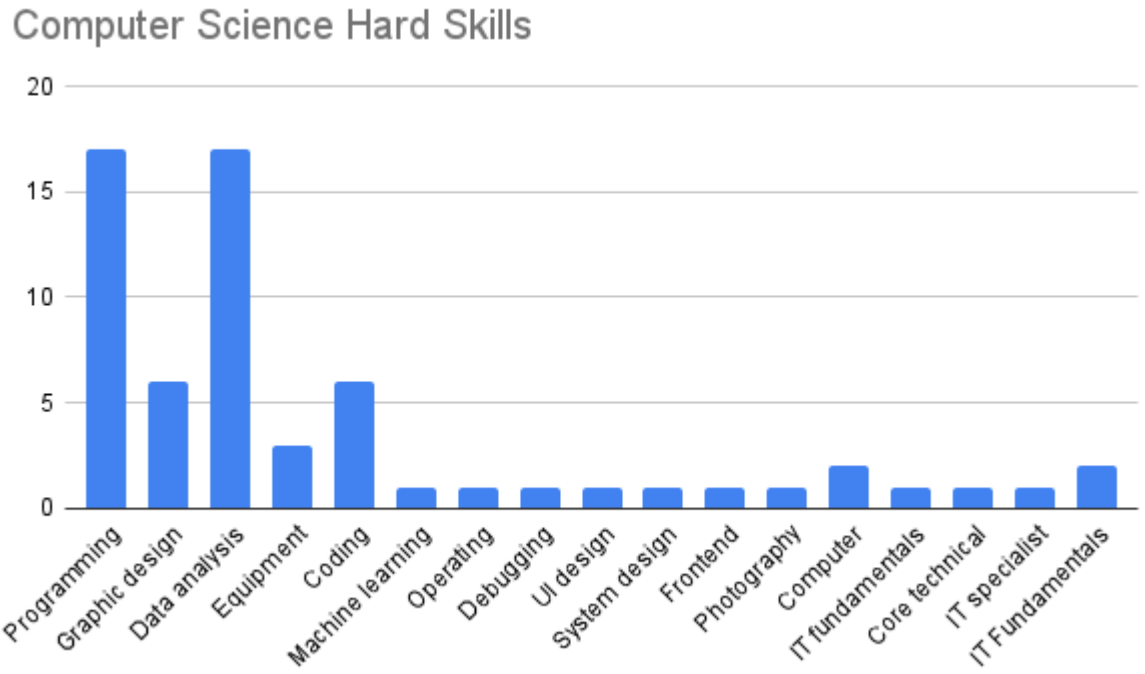


Chart 5.9 Computer Science most valued hard skills

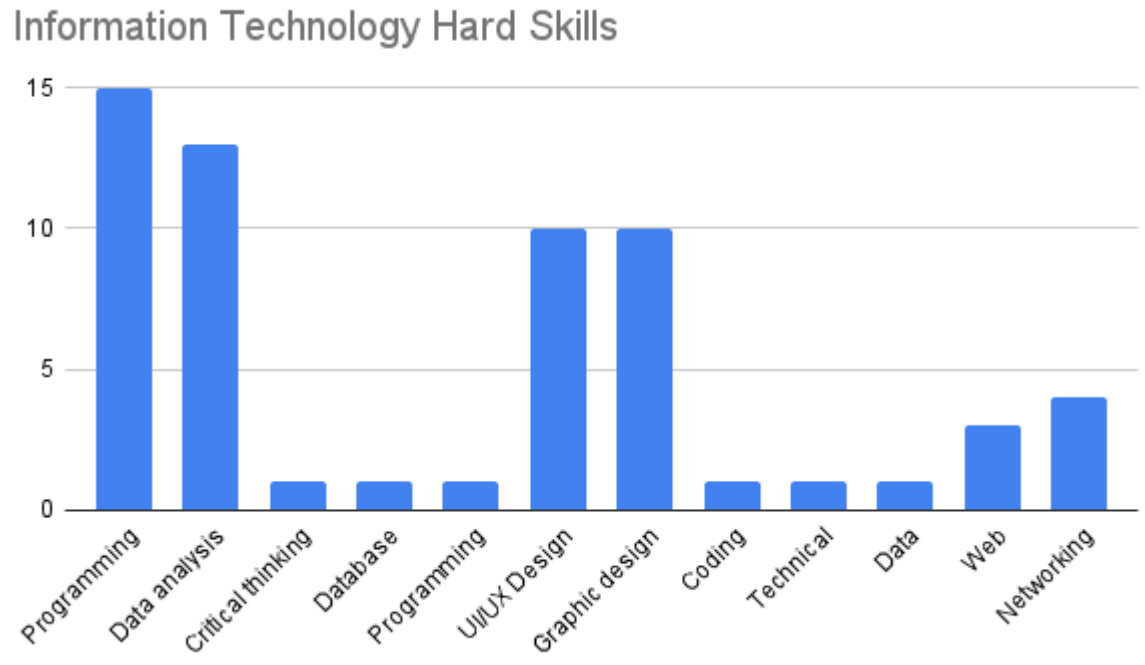


Chart 5.10 Information Technology most valued hard skills

With respect to skill proficiency, students across disciplines demonstrated a tendency to prioritize analytical and discipline-specific competencies rather than general skills. In the technology sector, particularly in Computer Science and Information Technology, programming and data analysis emerged as the most valued hard skills. Computer Science students reported seventeen (17) selections for both programming and data analysis, while Information Technology students recorded fifteen (15) selections for programming.

Student skill proficiency trends reveal a preference for discipline-specific analytical competencies over broad general skills. Within the technology sector, Computer Science and Information Technology students prioritize programming and data analysis as their primary hard skills—a reflection of a curriculum rooted in software development and algorithmic logic.

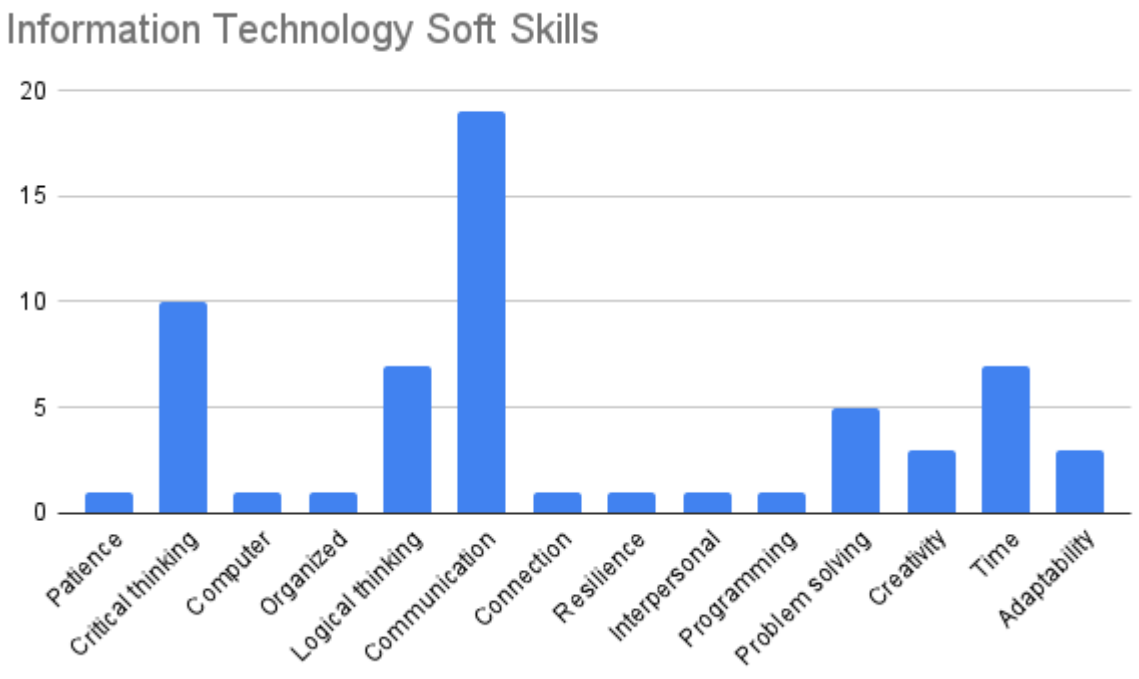


Chart 5.11 Information Technology most valued soft skills

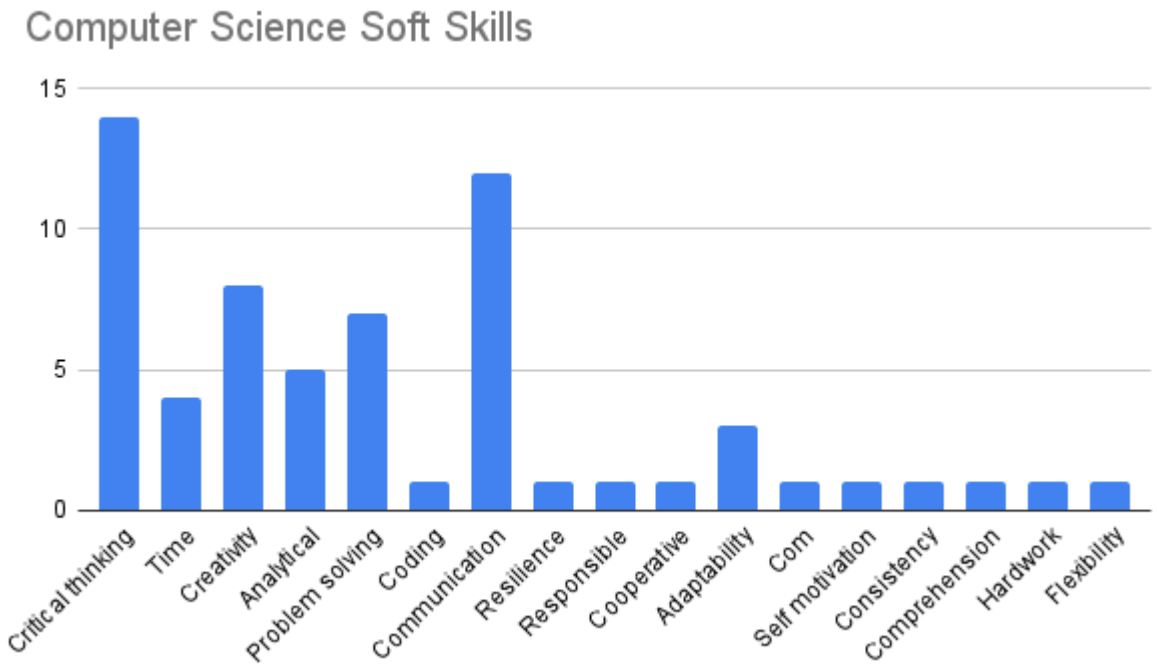


Chart 5.12 Computer Science most valued soft skills

Despite this technical focus, a "dual prioritization" is evident: IT students highly value communication (19 students select), and CS students prioritize critical thinking (14 students select), signaling an awareness that modern industry success requires a blend of technical mastery and collaborative problem-solving.

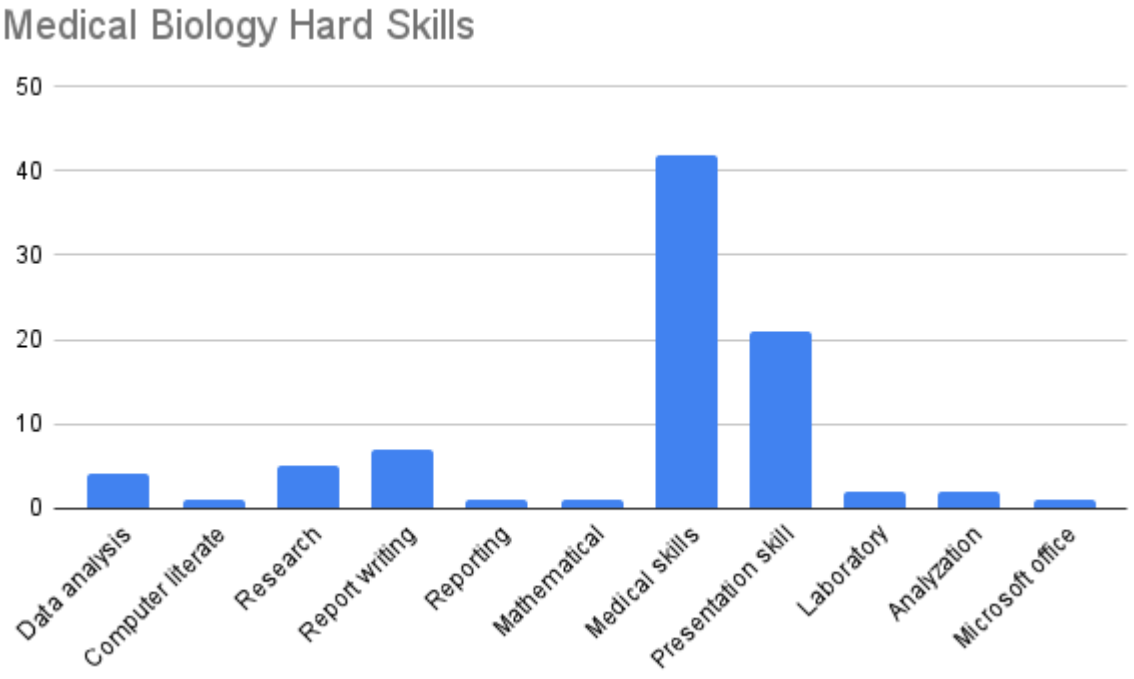


Chart 5.13 Medical Biology most valued hard skills

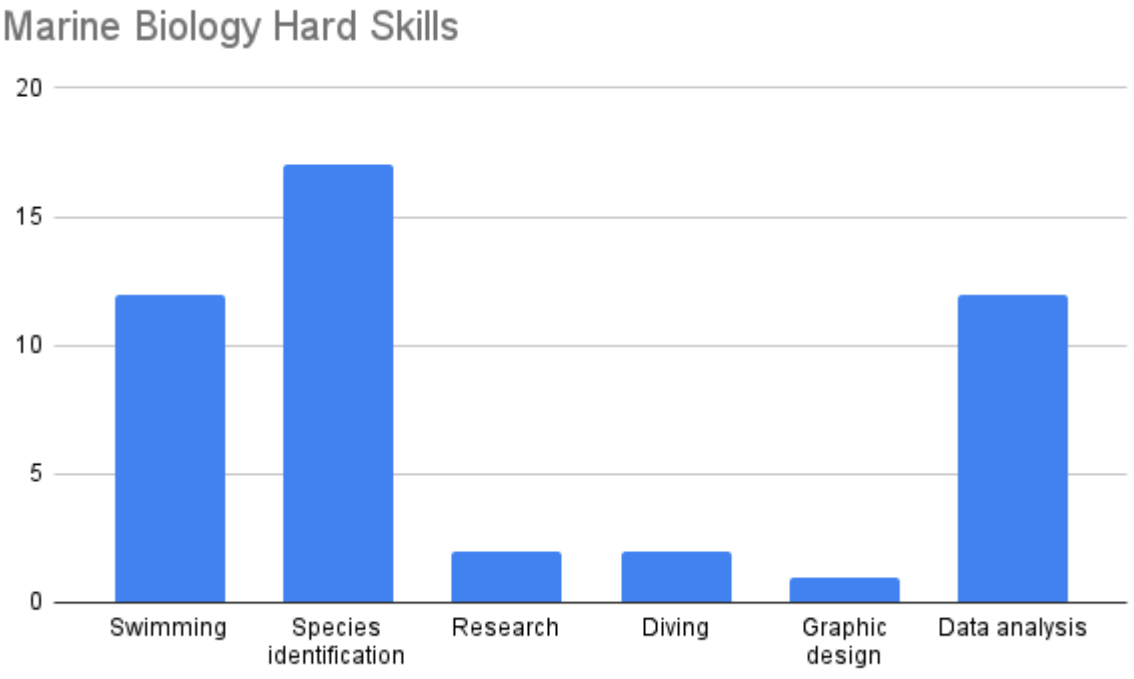


Chart 5.14 Marine Biology most valued hard skills

Conversely, students in the biological sciences showed a strong preference for practical, field-specific hard skills. Medical Biology students overwhelmingly selected Practical Medical Skills (42 students select), while Marine Biology students prioritized Species Identification (17 students select). Differences in preferred soft skills further reflect the distinct academic demands of these disciplines.

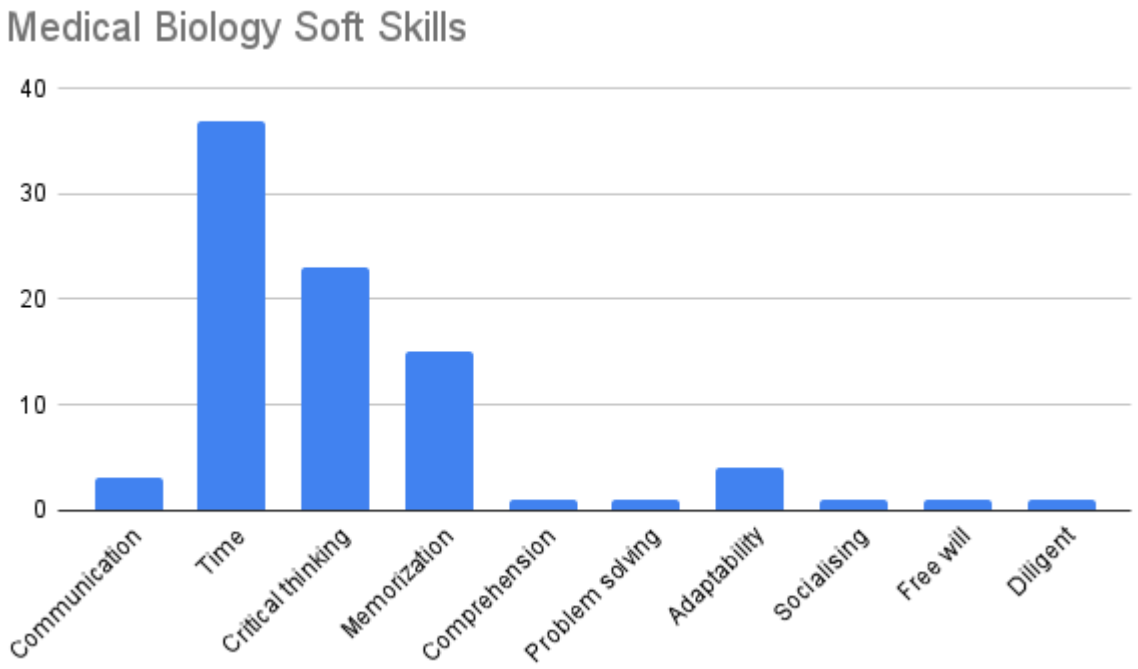


Chart 5.15 Medical Biology most valued soft skills

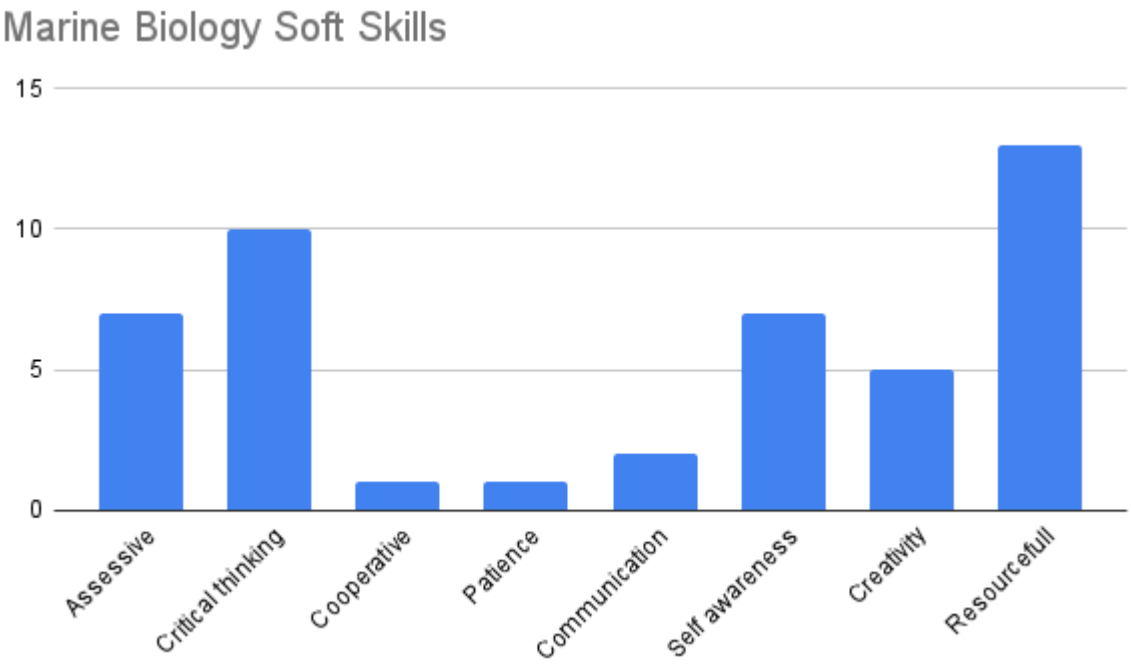


Chart 5.16 Marine Biology most valued soft skills

Medical Biology students identified Time Management as their most important soft skill (37 students selected), likely due to the intensive and content-heavy nature of pre-medical coursework. In comparison, Marine Biology students emphasized Resourcefulness (13 students selected), a skill essential for conducting fieldwork and adapting to variable environmental conditions.

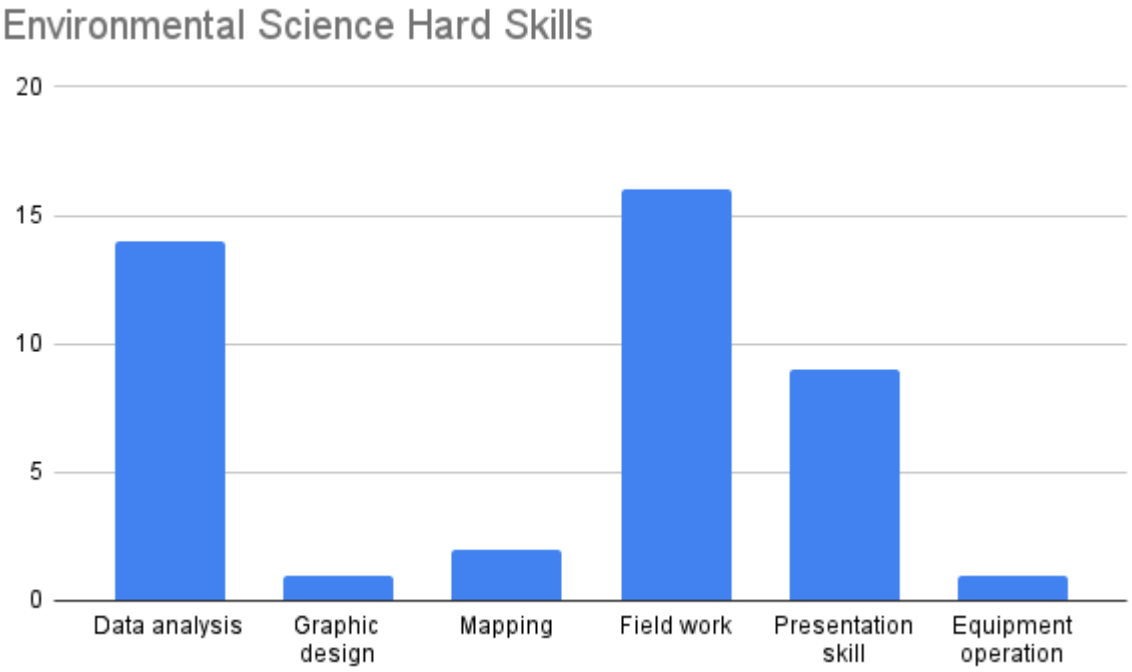


Chart 5.17 Environmental Science most valued hard skills

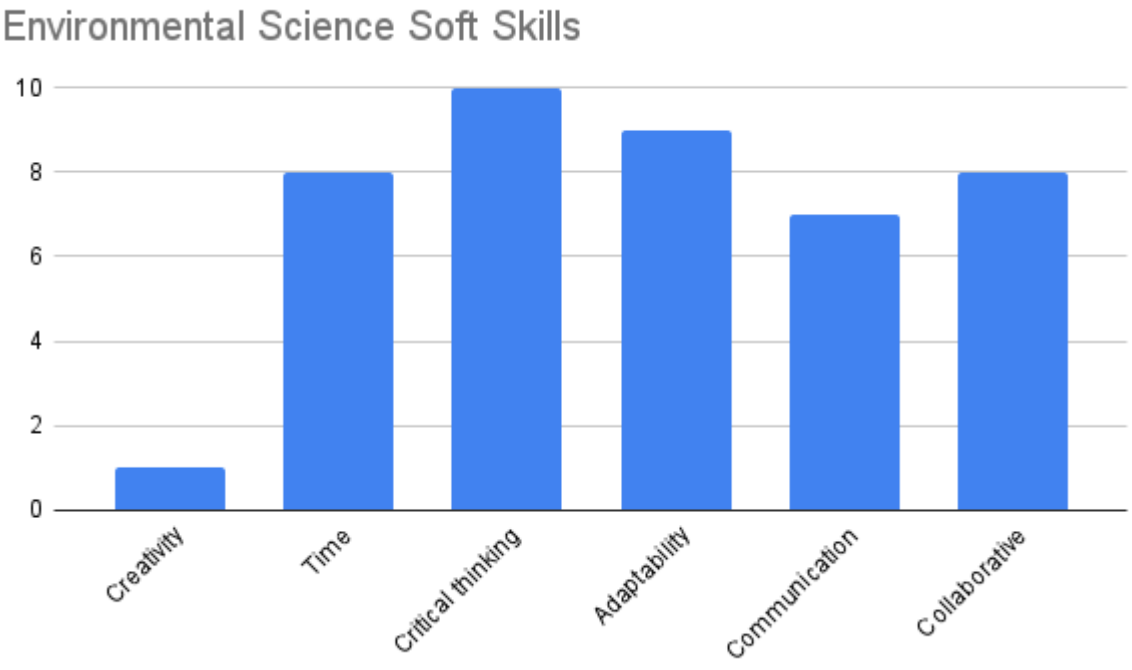


Chart 5.18 Environmental Science's most valued soft skills

Environmental Science students occupied an intermediate position between the technology and biological sciences. Field Work emerged as their primary hard skill preference (16 students selected), followed closely by Data Analysis. Critical Thinking was identified as their most important soft skill (10 students selected). This combination reflects the interdisciplinary character of Environmental Science, which requires students to collect empirical data in field settings while simultaneously engaging in analytical evaluation to address complex ecological and environmental challenges.

Model Training and Testing

Random Forest and Naive Bayes Stacking

The Level-0 of the stacking ensemble employed two distinct base classifiers: the Random Forest and Gaussian Naive Bayes models. The Random Forest was configured with 200 estimators and a maximum depth of 10 to balance the model's generalization and complexity. The Gaussian Naive Bayes Model was utilized to capture probabilistic characteristics of the features, assuming a Gaussian distribution.

The training was conducted on a 70-30 split, with 210 training samples and 90 testing samples. A Stratified K-Fold Cross-Validation was incorporated to maintain class consistency across all training folds. After the individual training and testing, the Random Forest model exhibits a superior performance with an accuracy of 84.44%, significantly outperforming the Naive Bayes model, which demonstrates a 78.89% accuracy.

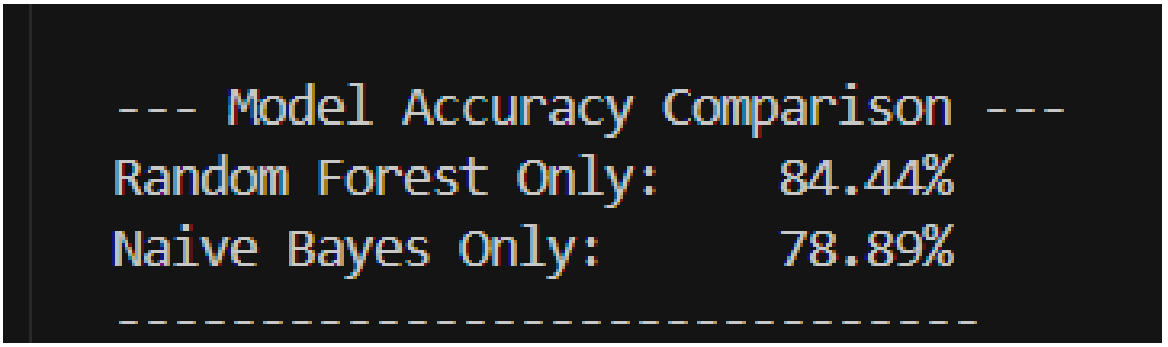


Figure 5.1.1 Individual Accuracy of Random Forest and Gaussian Naive Bayes Models

While this comparison proves that the Random Forest model performs better as a standalone predictor, the inclusion of the inferior Naive Bayes model remains critical. Its probabilistic approach provides a disparate decision boundary compared to the decision-tree-based Random Forest model. The diversity in prediction logic gives the Level-1 Meta-Learner a generous set of inputs, averting its reliance on a single algorithmic perspective.

Feedforward Neural Network as Meta Learner

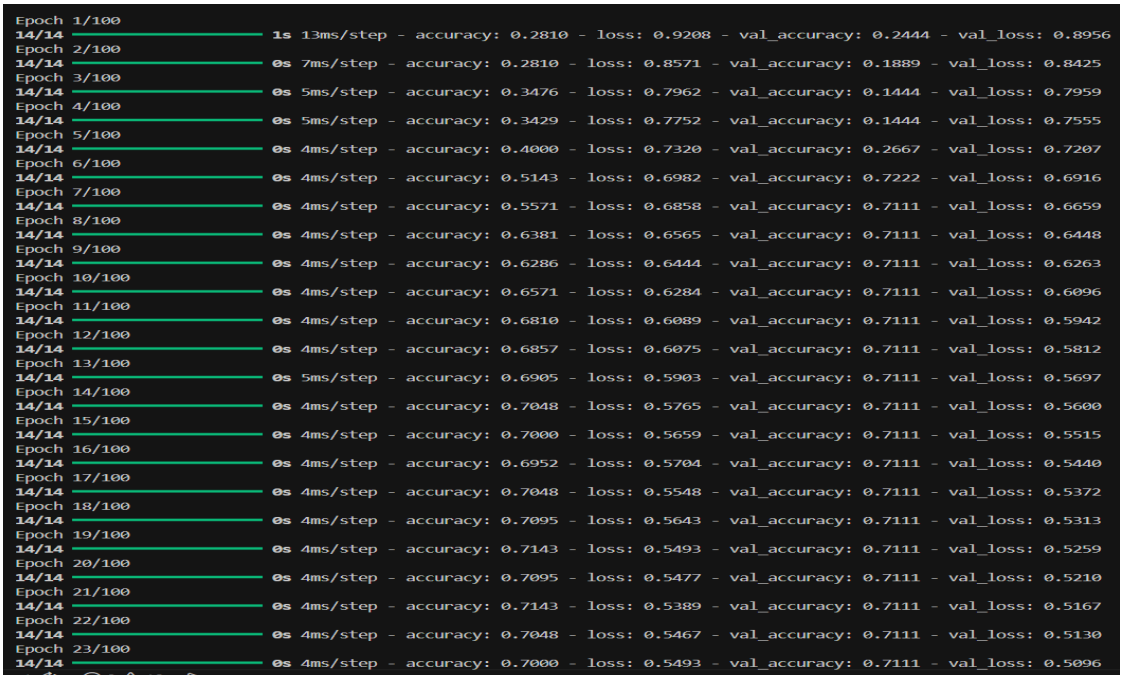
The Level-1 of the stacking ensemble, a Feedforward Neural Network (FFNN), was utilized as the Meta-Learner model. Designed to integrate the predicted outputs of the Random Forest and Gaussian Naive Bayes base classifiers, the FFNN served as the final decision-maker, learning to weigh the results of base classifiers dynamically

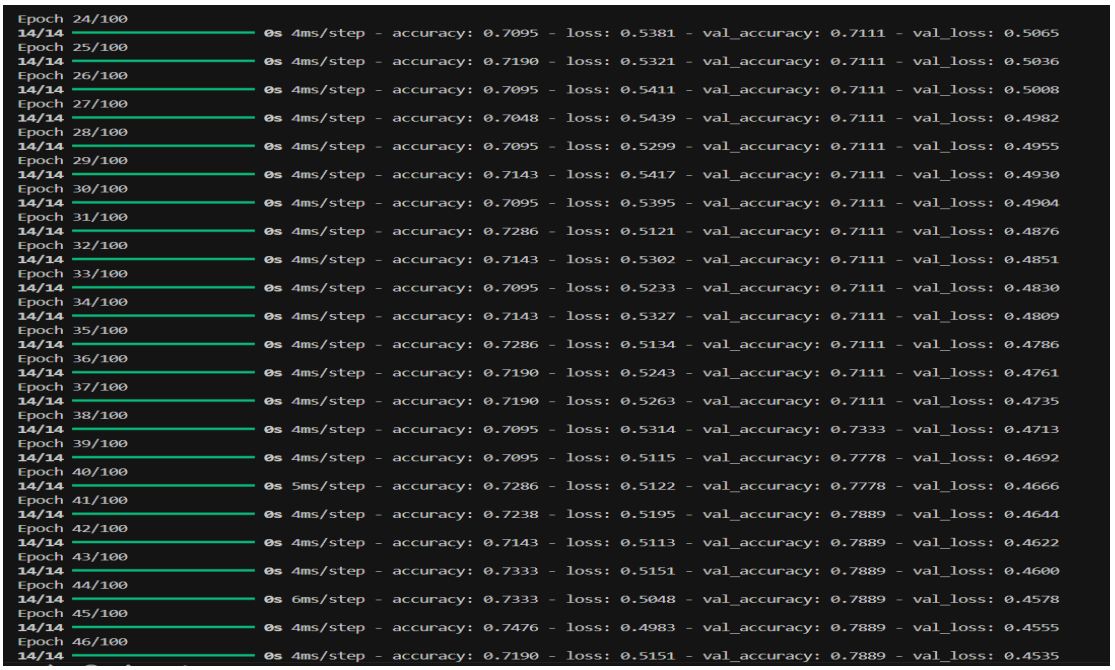
The network input consists of meta-features generated by the Level-0 models. The predicted scores of the Random Forest and the probabilistic prediction of the Gaussian Naive Bayes model were processed through a dense hidden layer containing 16 activated neurons activated by the Rectified Linear Unit (ReLU) function. Dropout was implemented to enhance the model generalization and mitigate the risk of overfitting. The dropout regularization layer with a rate of 0.2 was applied before the final output. The network ends in a single neuron using the Sigmoid activation function to produce a binary classification (regular = 1, irregular = 0).

The model was compiled using the Adam optimizer with a learning rate of 0.001 and binary cross-entropy as the loss function. The training was configured with a maximum of

100 epochs; however, an Early Stoping mechanism, with patience at 10, was implemented to automatically stop training if the validation loss fails to improve to ensure the model retained its optimal weights.

Training logs of the ensemble model





```
Epoch 47/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7333 - loss: 0.5030 - val_accuracy: 0.8000 - val_loss: 0.4517
Epoch 48/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 5ms/step - accuracy: 0.7429 - loss: 0.4922 - val_accuracy: 0.8000 - val_loss: 0.4493
Epoch 49/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7524 - loss: 0.4960 - val_accuracy: 0.8111 - val_loss: 0.4470
Epoch 50/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7476 - loss: 0.4922 - val_accuracy: 0.8111 - val_loss: 0.4450
Epoch 51/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7333 - loss: 0.4927 - val_accuracy: 0.8333 - val_loss: 0.4428
Epoch 52/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7619 - loss: 0.4829 - val_accuracy: 0.8333 - val_loss: 0.4407
Epoch 53/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7571 - loss: 0.4706 - val_accuracy: 0.8333 - val_loss: 0.4387
Epoch 54/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7381 - loss: 0.4869 - val_accuracy: 0.8444 - val_loss: 0.4368
Epoch 55/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7381 - loss: 0.4949 - val_accuracy: 0.8444 - val_loss: 0.4343
Epoch 56/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7333 - loss: 0.4841 - val_accuracy: 0.8444 - val_loss: 0.4320
Epoch 57/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7667 - loss: 0.4822 - val_accuracy: 0.8444 - val_loss: 0.4299
Epoch 58/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7571 - loss: 0.4733 - val_accuracy: 0.8444 - val_loss: 0.4280
Epoch 59/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7476 - loss: 0.4802 - val_accuracy: 0.8444 - val_loss: 0.4263
Epoch 60/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7762 - loss: 0.4713 - val_accuracy: 0.8444 - val_loss: 0.4246
Epoch 61/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7571 - loss: 0.4627 - val_accuracy: 0.8556 - val_loss: 0.4229
Epoch 62/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 5ms/step - accuracy: 0.7524 - loss: 0.4757 - val_accuracy: 0.8556 - val_loss: 0.4207
Epoch 63/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7619 - loss: 0.4873 - val_accuracy: 0.8667 - val_loss: 0.4189
Epoch 64/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7714 - loss: 0.4789 - val_accuracy: 0.8667 - val_loss: 0.4172
Epoch 65/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7810 - loss: 0.4672 - val_accuracy: 0.8667 - val_loss: 0.4157
Epoch 66/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7905 - loss: 0.4506 - val_accuracy: 0.8667 - val_loss: 0.4143
Epoch 67/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7667 - loss: 0.4677 - val_accuracy: 0.8667 - val_loss: 0.4127
Epoch 68/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 5ms/step - accuracy: 0.7524 - loss: 0.4639 - val_accuracy: 0.8667 - val_loss: 0.4112
Epoch 69/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7524 - loss: 0.4570 - val_accuracy: 0.8667 - val_loss: 0.4097
```

```
Epoch 70/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 5ms/step - accuracy: 0.7571 - loss: 0.4665 - val_accuracy: 0.8667 - val_loss: 0.4084
Epoch 71/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7571 - loss: 0.4862 - val_accuracy: 0.8667 - val_loss: 0.4073
Epoch 72/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 5ms/step - accuracy: 0.7714 - loss: 0.4538 - val_accuracy: 0.8667 - val_loss: 0.4057
Epoch 73/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7619 - loss: 0.4848 - val_accuracy: 0.8667 - val_loss: 0.4042
Epoch 74/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7667 - loss: 0.4555 - val_accuracy: 0.8667 - val_loss: 0.4026
Epoch 75/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7667 - loss: 0.4444 - val_accuracy: 0.8667 - val_loss: 0.4012
Epoch 76/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7714 - loss: 0.4713 - val_accuracy: 0.8667 - val_loss: 0.4001
Epoch 77/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7714 - loss: 0.4480 - val_accuracy: 0.8667 - val_loss: 0.3992
Epoch 78/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7905 - loss: 0.4306 - val_accuracy: 0.8556 - val_loss: 0.3978
Epoch 79/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7905 - loss: 0.4632 - val_accuracy: 0.8556 - val_loss: 0.3963
Epoch 80/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7810 - loss: 0.4492 - val_accuracy: 0.8556 - val_loss: 0.3949
Epoch 81/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7905 - loss: 0.4575 - val_accuracy: 0.8556 - val_loss: 0.3937
Epoch 82/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7905 - loss: 0.4399 - val_accuracy: 0.8556 - val_loss: 0.3925
Epoch 83/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7857 - loss: 0.4501 - val_accuracy: 0.8556 - val_loss: 0.3918
Epoch 84/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7667 - loss: 0.4620 - val_accuracy: 0.8556 - val_loss: 0.3911
Epoch 85/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.8048 - loss: 0.4317 - val_accuracy: 0.8556 - val_loss: 0.3895
Epoch 86/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7429 - loss: 0.4579 - val_accuracy: 0.8556 - val_loss: 0.3881
Epoch 87/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7762 - loss: 0.4358 - val_accuracy: 0.8556 - val_loss: 0.3871
Epoch 88/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7857 - loss: 0.4549 - val_accuracy: 0.8556 - val_loss: 0.3859
Epoch 89/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7524 - loss: 0.4701 - val_accuracy: 0.8778 - val_loss: 0.3856
Epoch 90/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7714 - loss: 0.4493 - val_accuracy: 0.8889 - val_loss: 0.3851
Epoch 91/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7905 - loss: 0.4590 - val_accuracy: 0.8889 - val_loss: 0.3842
Epoch 92/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7810 - loss: 0.4352 - val_accuracy: 0.8889 - val_loss: 0.3838
```

```
Epoch 93/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 5ms/step - accuracy: 0.7714 - loss: 0.4494 - val_accuracy: 0.8889 - val_loss: 0.3828
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7857 - loss: 0.4296 - val_accuracy: 0.8889 - val_loss: 0.3811
Epoch 97/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7905 - loss: 0.4309 - val_accuracy: 0.8889 - val_loss: 0.3799
Epoch 98/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7857 - loss: 0.4364 - val_accuracy: 0.8889 - val_loss: 0.3792
Epoch 99/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7857 - loss: 0.4370 - val_accuracy: 0.8889 - val_loss: 0.3786
Epoch 100/100
14/14 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - accuracy: 0.7619 - loss: 0.4575 - val_accuracy: 0.8889 - val_loss: 0.3777
```

Figure 5.1.2 Epoch Result of FeedForward Neural Network

In the training logs, the model exhibited a robust learning curve, with the validation accuracy consistently outperforming the training accuracy. This is attributed to the dropout layers forcing the network to learn resilient patterns rather than memorizing the training data. The ensemble model gives a training accuracy of 77.19% and a validation accuracy of 88.89% after 100 epochs.

Model Evaluation Metrics

```
--- Model Accuracy Comparison ---
Random Forest Only:    84.44%
Naive Bayes Only:      78.89%
------------------------------
Meta Learner Accuracy: 88.89%

Stacking improved accuracy by 4.44%

--- Confusion Matrix (Stacked Model) ---
Confusion Matrix:
[[17  9]
 [ 1 63]]

Classification Report (1=Regular, 0=Irregular):
              precision    recall  f1-score   support

           0       0.94      0.65      0.77        26
           1       0.88      0.98      0.93        64

    accuracy                           0.89        90
   macro avg       0.91      0.82      0.85        90
weighted avg       0.90      0.89      0.88        90
```

Figure 5.1.3 Accuracy, Precision, Confusion matrix, F1, Recall

To assess the effectiveness of the Stacking Ensemble model, its performance was evaluated using the standard confusion matrix and key classification metrics: Accuracy, Precision, Recall, F1-score.

Confusion Matrix Analysis

Figure 5.4.1.3 provides a detailed breakdown of the model's prediction accuracy compared to the actual student status. True positives (63) and true negatives (17) mean the model correctly predicted 63 students who were actually "regular" and 17 students who were "irregular," demonstrating the system's robustness in recognizing actual student profiles. False positives (9) were the instances where "irregular" students were predicted as "regular." Even though this is an error, in an educational context, it is less critical than discouraging a capable student. False negatives (1): the model misclassified 1 student into "irregular"; however, this extremely low false-negative rate indicates that the model is safe, as it almost never discourages students who have potential.

Classification Metrics (Precision, Recall, F1-Score)

The classification report provides deeper insights into the reliability of the model for each classification.

The model achieved high accuracy (88.89%), indicating that the Stacking Ensemble model correctly predicts approximately 9 out of 10 students. This performance validates the use of the meta-learner to culminate the strengths of the base classifiers (Random Forest and Naive Bayes). The following classification metrics evaluated the model's prediction accuracy:

Precision (prediction's quality): Irregular (0.94) and regular (0.88) indicate that when the model predicts a student will be irregular, it is at least 94% correct, and 88% accurate when predicting regular students. This quality of prediction makes the model robust in predicting at-risk students and potential students with high confidence.

Recall (sensitivity): Regular (0.98) and irregular (0.65) indicate the model achieves extremely high recall for regular students, which is a critical factor for a recommendation system, implying the model is capable of capturing almost all students who are aligned for the program, ensuring no missed opportunities. However, it struggles in recalling irregular students, which is at a moderate rate of 65%, suggesting some students with passing traits may struggle due to unmeasured external factors.

The weighted F1-score of 0.88 proves that the model maintains a strong balance between precision and recall. Not only does it provide high-confidence prediction by maximizing accuracy by guessing the majority class, but it also provides meaningful predictive power for both stable and at-risk students.

Recommender System Implementation

To evaluate the practical application of the model, the researchers simulated two distinct student profiles to test the system's decision-making logic.

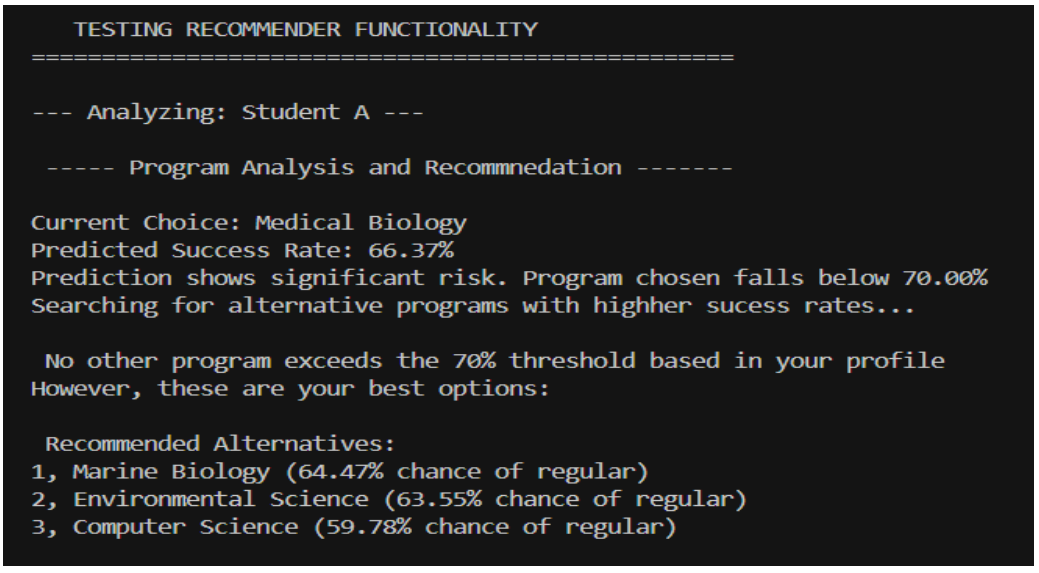


Figure 5.1.4 Recommendation System Simulation of a Mismatch Student

First, the test profile (Student A) is created with a general weighted average and admission test scores set at average-to-low. The hard skill feature was deliberately

mismatched, while the soft skill was aligned to the chosen program (Medical Biology). The model predicted the student success probability of 66.37%, which falls below the provided safety threshold of 70%. With this, the system recommended alternative programs, e.g., Marine Biology, Environmental Science, where the student experiences almost the same amount of success probability. This demonstrates the system's capability to detect academic misalignment early.

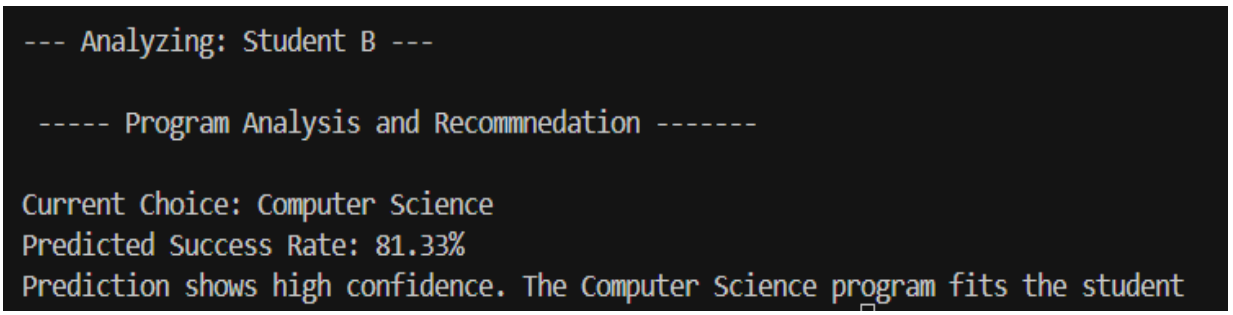


Figure 5.1.5 Recommendation System Simulation of a Aligned Student

In the second scenario, a profile (Student B) was constructed with high academic metrics and skills that strictly align to the chosen program (Computer Science). The model returned a student success probability of 81.33%, significantly exceeding the safety threshold.

Based on this result, the system concluded that the student has a high probability of maintaining "regular" status throughout the four-year duration of the program. This validation confirms the model's ability to recognize and endorse qualified applicants, providing the guided recommendation system with data-driven confidence in college program recommendations.

Comparison with Traditional Method

To evaluate the practical utility of the proposed system, the Stacking Ensemble model was benchmarked against the ZeroR (Zero Rule) algorithm. In educational data mining, ZeroR serves as the standard "Traditional Method" baseline, representing the

performance achievable by only predicting the majority class. The dataset was split into 70-30 for training and testing; the testing data (N = 90) consists of sixty-four (64) "regular" students, and twenty-six (26) are "irregular" students. To calculate the baseline, the ZeroR Accuracy formula is used:

$$\text{ZeroR Accuracy} = \frac{\text{Frequency of Majority Class}}{\text{Total Number of Instances}} * 100$$

$$\text{ZeroR} = \frac{64}{90} * 100$$

$$\text{ZeroR} = 0.711111 * 100$$

$$\text{ZeroR} = 71.1111$$

ZeroR = 71.11%

The Zero Rule baseline assumes that all students are "regular" because of its majority class. As shown in the calculation, this traditional probabilistic guess yields an accuracy of 71.11%.

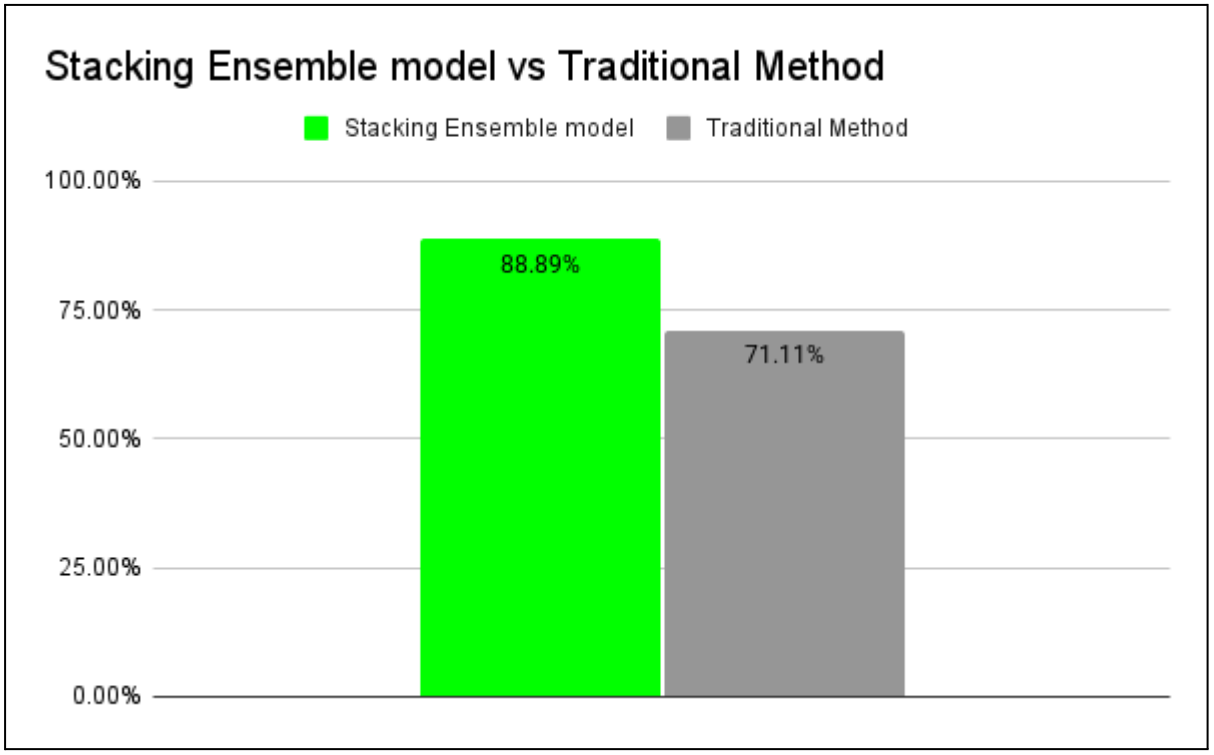


Chart 5.19 Accuracy Comparison of ML Model vs. Traditional Method

In contrast, the Stacking Ensemble model achieved a validation accuracy of 88.89%. Comparing the two methods, the ML model significantly provided a performance lift of 17.78% over the baseline. This margin confirms that the stacking ensemble is successful in extracting complex patterns from the student data to identify “irregular” students that the traditional method would miss

CHAPTER VI

Summary, Conclusions, and Recommendations

Overview

In this chapter, the researchers present the conclusions derived from the findings and analyses from the previous chapter. The conclusions are formulated in accordance with the stated research objectives and are based on the interpretation of the results obtained from the data analysis.

The recommendations are then provided as insights and proposed further actions that can address the identified problem, enhance the accuracy and applicability of the model, and guide future studies related to the topic.

Conclusion

This study aims to enhance the academic decision-making of the students by offering analysis on their cognitive abilities and skill sets. To achieve this, we developed a machine learning approach using the following models such as Naive Bayes + Random Forest → Feedforward Neural Network. It is concluded that the stacking ensemble models can evaluate the data of students that were collected by the researchers in the College of Sciences with a validation accuracy of 88.89 percent.

With all the objectives in this study, the following are greatly achieved:

1. **The researchers will project the attrition rate of College of Sciences students for the next five years using historical data to imply the contribution of cognitive abilities and skill misalignment to the college attrition rate.**

   By using the linear regression for the analysis of historical data, it revealed a consistent negative relationship between the academic year and the attrition rate defined by the equation $\hat{y} = 64.00 - 2.57x$. It also identified the significant downward trend in student attrition. By having this result of the equation, it indicates that the

college is successfully reducing attrition rate by approximately 2.57 percentage points per year. By having a projected attrition rate of 40.87 percent by year 9, this high rate can contribute to the 31.4 percent indecisiveness rate discovered among the student population.

2. **The researchers will ensemble machine learning models to provide a higher level of accuracy to predict the best recommendation of a college program based on cognition and skills.**

   Based on rigorous selection of the models with high accuracy based on related literature and studies, a Stacking Ensemble architecture was developed. By integrating the decision-tree-based voting mechanism of the Random Forest algorithm and probabilistic approach of Gaussian Naive Bayes model, the system provided an effective diverse predictive strategies. While the individual base learners achieved accuracies of 78.89% (Naive Bayes) and 84.44% (Random Forest), the Feedforward Neural Network meta-learner successfully synthesized the inputs to surpass their individual capabilities. The meta-learner was trained through 100 epochs with a dropout rate of 0.2 to ensure generalization; the final stacking ensemble model achieved a superior 88.89% validation accuracy. The comprehensive evaluation via classification metrics and confusion matrix analysis both confirms the model is robust and reliable, particularly in its ability to identify suitable academic programs with high precision.

   Furthermore, this predictive prowess is operationalized through a recommender system that evaluates the student's profile against a 70% safety threshold. By analyzing the success probability for different academic tracks, the system can provide a statistical recommendation: validating qualified student applicants while

automatically suggesting alternative programs for students flagged as at-risk on a particular program.

3. **The researchers will evaluate the accuracy of the model compared to traditional methods to improve the outcome of prediction.**

   The Zero Rule baseline was used to evaluate the accuracy of the traditional method. In comparison, the traditional method achieved an accuracy of 71.11%. Meanwhile, the stacking ensemble model achieved an 88.89% validation accuracy. This significant performance margin of 17.78% demonstrates the machine learning model successfully transcends simple heuristic guessing, offering a statistically superior and more reliable mechanism compared to conventional baseline approaches.

Recommendations

This study establishes a foundational framework for predicting student attrition using ensemble learning. To guide the future researchers in further enhancing the model's accuracy and practical utility, the researchers propose these recommendations for future development, which are listed below:

1. Implementing Longitudinal Analysis (LSTM)

   Given that attrition rate is a temporal phenomenon, it is recommended to transition from static models to Long Short-Term Memory (LSTM) Networks. Unlike standard classifiers, LSTMs are better suited for sequential data, allowing the model to analyze a student's performance patterns and trajectory over multiple semesters. This will enable the model to detect declining engagement and predict attrition risks preemptively before the effects become irreversible.

2. Enhancement of Ensemble Diversity and Robustness

To further improve the classification and predictive accuracy of the ensemble model, future iterations of the model should incorporate wide variations of base learners, such as K-Nearest Neighbors (KNN), to capture local data structures and similarities between student profiles. Additionally, the meta-learner component, currently a simple Feedforward Neural Network (FFNN), could also be upgraded to more sophisticated algorithms to better synthesize the predictions of the base models.

3. Development of a Deployed Interface

In the future development of the ensemble prediction model, the focus can extend beyond the model itself to the development of interfaces; it is recommended to develop a user-friendly graphical interface (GUI) or web-based application. Currently, the model exists as a backend script, but the integration to the deployed interface will allow users to input real-time student data and receive an immediate assessment without requiring technical programming knowledge. This promotes accessibility and ensures the tool can be integrated into the existing Student Information System (SIS).

4. Expansion of Institutional Scope

The current dataset is limited to the College of Sciences, which may bias the model towards STEM-centric programs. To improve the model's generalizability, it is strongly recommended to include more programs from other colleges (e.g., College of Arts and Humanities, College of Engineering, College of Business and Accountancy, College of Teacher Education). Training the model on a more diverse dataset will allow it to learn broader academic trends and provide accurate recommendations for the university's population and not only limited to science majors.

5. Granularity of Skill Attributes

The skill attributes used in this study were specifically derived from the College of Sciences programs. To enhance the recommender system's precision, the feature pool should be expanded to include a comprehensive taxonomy of hard and soft skills relevant to all academic disciplines. Adding a more granular and diverse skill dataset will enable the model to make nuanced distinctions between programs and reduce the likelihood of false positives in student placements.

REFERENCES

Asor, J. R., Catedrilla, G. M. B., Buama, C. A. C., Malabayabas, M. E., & Malabayabas, C. E. (2023). Prediction of Senior High School Students' Performance in a State University: An Educational Data Mining Approach. International Journal of Information and Education Technology, 13(6), 925–931. https://doi.org/10.18178/ijiet.2023.13.6.1888

Bucad, A. T. (2024). Towards the Development of a Career Path Recommender System for Senior High School in Selected Public Schools using Multi-Label Classification. International Journal of Research and Innovation in Applied Science, IX(III), 374–382. https://doi.org/10.51584/ijrias.2024.90334

Chau, H. K. (2023). Explainable Course Recommendation: Connecting College Education to Knowledge and Careers through Skills. In ProQuest LLC. ProQuest LLC. 789 East Eisenhower Parkway, P.O. Box 1346, Ann Arbor, MI 48106. Tel: 800-521-0600; Website: https://eric.ed.gov/?q=Course+Exploration+system&id=ED656914

Commission on Higher Education. (2012). Policy-standard to enhance quality assurance (QA) in Philippine higher education through an outcomes-based and typology-based QA (CHED Memorandum Order No. 46, Series of 2012). Website: https://www.pacu.org.ph/wordpress/wp-content/uploads/2017/03/CMO-No.46-s2012.pdf

Commission on Higher Education. (2017). Policy on the admission of senior high school graduates to the higher education institutions effective academic year 2018-2019 (CHED Memorandum Order No. 105, Series of 2017). Website: https://depedpines.com/2018/05/indorsement-ched-memorandum-order-cmo-no-105-s-2017-policy-on-the-admission-of-senior-high-school-graduates-to-the-higher-education-institutions-effective-academic-year-2018-2019/

GeeksforGeeks. (2020, September). Evaluation Metrics in Machine Learning. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/metrics-for-machine-learning-model/

Ghazal, A., Allah, F., & Bilquise, G. (2022). RECOMMENDING COLLEGE PROGRAMS TO STUDENTS USING MACHINE LEARNING. 100(19). https://www.jatit.org/volumes/Vol100No19/33Vol100No19.pdf

Islam, M. S., & Hosen, A. S. M. S. (2025). Personalized Course Recommendation System: A Multi-Model Machine Learning Framework for Academic Success. Digital, 5(2), 17. https://doi.org/10.3390/digital5020017

Jeremiah Tanimu, J., Hamada, M., Hassan, M., & Yusuf Ilu, S. (2021). A Contemporary Machine Learning Method for Accurate Prediction of Cervical Cancer. SHS Web of Conferences, 102, 04004. https://doi.org/10.1051/shsconf/202110204004

Kord, A., Aboelfetouh, A., & Shohieb, S. M. (2025). Academic course planning recommendation and students' performance prediction multi-modal based on educational data mining techniques. Journal of Computing in Higher Education. https://doi.org/10.1007/s12528-024-09426-0

Kumar, A., Tak, T. K., Ali, S. M. S., Haque, M., Paralkar, T. A., Kshirsagar R, P., & Upreti, K. (2025). Predictive Modeling of Student Learning Outcomes Through Cognitive and Emotional Skill Integration. International Research Journal of Multidisciplinary Scope, 06(01), 892–910. https://doi.org/10.47857/irjms.2025.v06i01.02895

Maloshonok, N., & Terentev, E. (2017). The mismatch between student educational expectations and realities: prevalence, causes, and consequences. European Journal of Higher Education, 7(4), 356–372. https://doi.org/10.1080/21568235.2017.1348238

Paula, G. B. de, Nogueira, C. M. M., Nonato, B. F., & Ariovaldo, T. C. de C. (2025). STUDENT DROPOUT AT UFMG: ANALYSIS OF THE INFLUENCE OF STUDENTS' SOCIOECONOMIC PROFILES AND COURSE CHARACTERISTICS. Avaliação: Revista Da Avaliação Da Educação Superior (Campinas), 30. https://doi.org/10.1590/1982-57652025v30id2806184

Prusty, S., Patnaik, S., & Dash, S. K. (2022). SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. Frontiers in Nanotechnology, 4. https://doi.org/10.3389/fnano.2022.972421

Rebelo Marcolino, M., Reis Porto, T., Thompsen Primo, T., Targino, R., Ramos, V., Marques Queiroga, E., Munoz, R., & Cechinel, C. (2025). Student dropout prediction through machine learning optimization: insights from moodle log data. Scientific Reports, 15(1), 9840. https://doi.org/10.1038/s41598-025-93918-1

Rohman, M. G., Abdullah, Z., Kasim, S., & Rasyidah. (2025). Hybrid Logistic Regression Random Forest on Predicting Student Performance. JOIV International Journal on Informatics Visualization, 9(2), 852–852. https://doi.org/10.62527/joiv.9.2.3972

Serrano, M. R., Hontiveros, N. L., Ryle, E., Riza, & Bein, N. (2022). DALAN: A Course Recommender for Freshmen Students using a Multiple Regression Model. International Journal of Computer Science & Engineering Survey, 13(5/6), 09-21. https://doi.org/10.5121/ijcses.2022.13602

Termedi, M. I., Aini Marina Ma'rof, Ab., & Ishak, I. (2023). Utilizing Educational Data Mining for Enhanced Student Performance Analysis in Malaysian STEM Education. International Journal of Academic Research in Progressive Education and Development, 12(4). https://ijarped.com/index.php/journal/article/view/793

Trujillo, F., Pozo, M., & Suntaxi, G. (2025). Artificial intelligence in education: A systematic literature review of machine learning approaches in student career prediction. Journal of Technology and Science Education, 15(1), 162–162. https://doi.org/10.3926/jotse.3124

Trujillo, F., Pozo, M., & Suntaxi, G. (2025). Artificial intelligence in education: A systematic literature review of machine learning approaches in student career prediction. Journal of Technology and Science Education, 15(1), 162–185. https://www.jotse.org/index.php/jotse/article/view/3124/937?utm_source

Yurtkan, K., Adalier, A., & Tekgüç, U. (2023). Student success prediction using feedforward neural networks. Romanian Journal of Information Science and Technology, 26(2), 121–136. https://doi.org/10.59277/ROMJIST.2023.2.01

Zayed, Y., Salman, Y., & Hasasneh, A. (2022). A Recommendation System for Selecting the Appropriate Undergraduate Program at Higher Education Institutions Using Graduate Student Data. Applied Sciences, 12(24), 12525. https://doi.org/10.3390/app122412525

ZUMAIRA NAWAL MANALONDONG, & MARVIEN MARAGON BARRIOS. (2025). Cognitive ability and career choice among graduating senior high school students. International Journal of Science and Research Archive, 15(3), 1664–1670. https://doi.org/10.30574/ijsra.2025.15.3.1911

APPENDICES

APPENDIX A - 1

Attachments

Attachment A. THESIS/DISSERTATION MENTORING PROGRESS REPORT

THESIS/DISSERTATION MENTORING PROGRESS REPORT

1st Semester, A.Y. 2025-2026

Department: Computer Science College: College of Sciences

Seq. Number	
Name of Student/s	Gatchalian, John Rex O.
	Jamion, Aneza H.
	Navarro, Lance Armstrong T.
	Recarze, Spledelyn Cristine P.
Course/Year/Section	Bachelor of Science in Computer Science/ 4th Year/ Block 2
Thesis Title	"A Machine Learning Approach for a Best College Program Recommendation System Based on Cognitive Abilities and Skill Sets"
Adviser/Rank	Mr. Adonis C. Ampongan
Date	August 18, 2025

To be filled by the adviser

Please answer the following:

1. Describe current status/stage of the thesis/dissertation. Relate with the Workplan.

After the proposal defense done by the group, the revision was made based on the panel's comments, suggestions and recommendations. The researcher implemented all of the applied suggestions to the paper in order to continue the data gathering process.

2. Identify problems encountered. Discuss how the problem was resolved.

Some of the problems encountered by the researchers after the proposal defense were: the scope of the respondents for the study is wide causing a problem in gathering all necessary data for the input of the system, and the statement of the problem is not aligned with the objectives of the study. The conceptual framework needs to be more specific with the criteria of each data processing, also the proposed models need to be specified based on its use in each data type.

To resolve these problems, the researchers first narrowed down the population of respondents from all colleges at Palawan State University to College of Sciences students', next the researchers modified the statement of the problem to align with the proposed objectives of the study. This was followed by looking at the specific and best suited model to be used in a system that can provide a higher accuracy for the input.

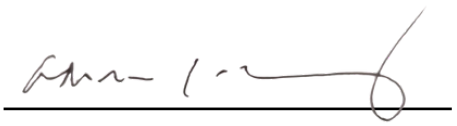
3. If the problem is unresolved this is due to (specify reasons):

N/A

4. What modifications or changes do you think will still allow you to address/answer the problems you have set for your objectives?

N/A

This is to certify that the above-mentioned student(s) has undertaken a consultation with their adviser.

Signature of the Thesis Adviser: 

APPENDIX A - 2

Attachments

Attachment A. THESIS/DISSERTATION MENTORING PROGRESS REPORT

THESIS/DISSERTATION MENTORING PROGRESS REPORT

1st Semester, A.Y. 2025-2026

Department: Computer Science College: College of Sciences

Seq. Number	
Name of Student/s	Gatchalian, John Rex O.
	Jamion, Aneza H.
	Navarro, Lance Armstrong T.
	Recarze, Spledelyn Cristine P.
Course/Year/Section	Bachelor of Science in Computer Science/ 4th Year/ Block 2
Thesis Title	"A Machine Learning Approach for a Best College Program Recommendation System Based on Cognitive Abilities and Skill Sets"
Adviser/Rank	Mr. Adonis C. Ampongan
Date	November 11, 2025

To be filled by the adviser

Please answer the following:

1. Describe current status/stage of the thesis/dissertation. Relate with the Workplan.

The current data gathering is still on-going with their respective respondents within the College of Sciences and the system is at 20 percent of its creation.

2. Identify problems encountered. Discuss how the problem was resolved.

The problems that the researchers encountered in this stage of the study is the collection of data due to the conflict from the specific office within the Palawan State University. To resolve this problem, the researchers did an alternative method to gather all necessary data through Google Forms and disseminate it to their respective respondents.

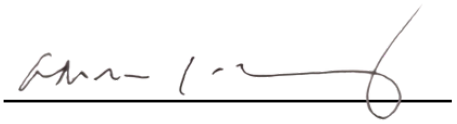
3. If the problem is unresolved, this is due to (specify reasons):

N/A

4. What modifications or changes do you think will still allow you to address/answer the problems you have set for your objectives?

N/A

This is to certify that the above-mentioned student/s have undertaken a consultation with their adviser.

Signature of the Thesis Adviser: 

APPENDIX A - 3

Attachments

Attachment A. THESIS/DISSERTATION MENTORING PROGRESS REPORT

THESIS/DISSERTATION MENTORING PROGRESS REPORT

1st Semester, A.Y. 2025-2026

Department: Computer Science College: College of Sciences

Seq. Number	
Name of Student/s	Gatchalian, John Rex O.
	Jamion, Aneza H.
	Navarro, Lance Armstrong T.
	Recarze, Spledelyn Cristine P.
Course/Year/Section	Bachelor of Science in Computer Science/ 4th Year/ Block 2
Thesis Title	"A Machine Learning Approach for a Best College Program Recommendation System Based on Cognitive Abilities and Skill Sets"
Adviser/Rank	Mr. Adonis C. Ampongan
Date	December 9, 2025

To be filled by the adviser

Please answer the following:

1. Describe current status/stage of the thesis/dissertation. Relate with the Workplan.

At this stage of the study, the researchers already analyzed and interpreted the result of the study. The researchers write the results and discussion of the study alongside with the graphs as a supporting document for the output. Together with this, the researchers write the conclusions and recommendations for this study. As for the progress of the system it is already 100 percent finished and functioning accordingly with the desired result and accuracy of the researchers.

2. Identify problems encountered. Discuss how the problem was resolved.

The minor problems that the researchers encountered in this stage of the study is the manual computation for the attrition rate and projection of attrition rate, as well as the manual encoding of the researchers for their source code of the system.

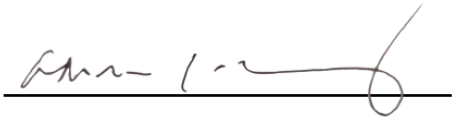
3. If the problem is unresolved this is due to (specify reasons):

N/A

4. What modifications or changes do you think will still allow you to address/answer the problems you have set for your objectives?

N/A

This is to certify that the above-mentioned student/s have undertaken a consultation with their adviser.

Signature of the Thesis Adviser: 

READINESS FORM FOR FINAL DEFENSE

PSU-ACA-008

Republic of the Philippines
PALAWAN STATE UNIVERSITY
Puerto Princesa City
College ofSciences
Department of Computer Studies

Certification of Readiness for Proposal Defense

RONALDEDILBERTO A. ONA
Dean
This College
Sir:

This is to certify that Gatchalian, John Rex O., Jamion, Aneza H., Navarro Lance Armstrong T., Recarze, Spledelyn Cristine P. are ready to defend their research study "A Machine Learning Approach for a Best College Program Recommendation System Basedon Cognitive Abilities and SkillSets".
Itis requestedthat oralproposal defense bescheduled.

ADONIS C. AMPONGAN
Research Adviser

Schedule for Oral FinalDefense:

Date & Time: January13, 2026

Venue: _____

RecommendingApproval:

MENCHIEL. LOPEZ
Research Instructor

Approved

RONALD EDILBERTO A. ONA
Dean

RELEVANT SOURCE CODE

```
import pandas as pd

import numpy as np

import sys



from sklearn.model_selection import train_test_split, StratifiedKFold

from sklearn.preprocessing import StandardScaler, LabelEncoder

from sklearn.metrics import accuracy_score, confusion_matrix,
classification_report, precision_score, recall_score, f1_score

from sklearn.ensemble import RandomForestClassifier

from sklearn.naive_bayes import GaussianNB



from tensorflow.keras.models import Sequential

from tensorflow.keras.layers import Dense, Dropout, Input

from tensorflow.keras.callbacks import EarlyStopping

from tensorflow.keras.optimizers import Adam

from tensorflow.keras.utils import to_categorical

import tensorflow as tf



np.random.seed(42)

tf.random.set_seed(42)



class RecommenderSystem:

    def __init__(self, college_data):

        self.college_data = college_data

        self.scalers = StandardScaler()

        self.encoders = {}

        self.models = {}

        self.features = ['strand', 'gwa', 'verbal', 'numerical', 'abstract',
```

```
                            'spelling', 'usage', 'program',
                            'soft_skills', 'hard_skills']



    def loading_data(self):
        print("-----Loading Data-----")
        try:
            self.data = pd.read_csv(self.college_data)
        except FileNotFoundError:
            print("Error: File not found")
            return


        # initial_count = len(self.data)
        # self.data = self.data.dropna()
          # print(f"Data Cleaning: Dropped {initial_count - len(self.data)} rows
with missing values.")


        categorical = ["strand", "program", "soft_skills", "hard_skills"]


        for col in categorical:
            le = LabelEncoder()
            if col in self.data.columns:
                self.data[col] = le.fit_transform(self.data[col])
                self.encoders[col] = le
            else:
                print(f"Warning:  Column {col} is not found.")


        target = 'acad_status'
        if target not in self.data.columns:
            print(f"Error: Target column '{target}' not found.")
```

```
            return


        self.data['target'] = self.data[target].astype(str).str.lower()

        binary_map = {'regular':1, 'irregular':0}


        #create a new column on the file with binary values

        self.data['target_col'] = self.data['target'].map(binary_map)


        #data catch for typo in csv file

        self.data = self.data.dropna(subset=['target_col'])


        #feature validation

        valid_features = [c for c in self.features if c in self.data.columns]

        X = self.data[valid_features]

        y = self.data["target_col"]


        #data splitting

        self.X_train, self.X_test, self.y_train, self.y_test = train_test_split(

            X, y, test_size=0.3, random_state=42, stratify=y

        )


        #converting pandas to numpy

        self.y_train = self.y_train.values

        self.y_test = self.y_test.values


        #scaling

        self.X_train = self.scalers.fit_transform(self.X_train)

        self.X_test = self.scalers.transform(self.X_test)
```

```
        print("Data is successfully loaded and preprocessed")


    #out-of-fold training for rf+nb
    def oofprediction(self, model, X, y):


        oof_prediction = np.zeros((X.shape[0], 1))
        skf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42) #reminder
to add in rrl


        for train_index, val_index in skf.split(X, y):
            X_fold_train, X_fold_val = X[train_index], X[val_index]
            y_fold_train = y[train_index]


            from sklearn.base import clone
            clf  =clone(model)
            clf.fit(X_fold_train, y_fold_train)


            preds = clf.predict_proba(X_fold_val)[:, 1]
            oof_prediction[val_index] = preds.reshape(-1, 1)


        return oof_prediction



    def train_stacking(self):
        print("\n ----- Level 0 Stacking Model Training -----")


        rf = RandomForestClassifier(
            n_estimators=200, max_depth=10, random_state=42
        )
        nb = GaussianNB()
```

```
#generating input for ffnn from rf+nb stacking
print("Generating RF prediction")
rf_oof = self.oofprediction(rf, self.X_train, self.y_train)

print("Generating NB prediction")
nb_oof = self.oofprediction(nb, self.X_train, self.y_train)

self.X_train_meta = np.hstack([rf_oof, nb_oof])

rf.fit(self.X_train, self.y_train)
nb.fit(self.X_train, self.y_train)
self.models['rf'] = rf
self.models['nb'] = nb

rf_test = rf.predict_proba(self.X_test)[:, 1].reshape(-1,1)
nb_test = nb.predict_proba(self.X_test)[:, 1].reshape(-1,1)
self.X_test_meta = np.hstack([rf_test, nb_test])

print("\n ----- Level 1 Meta-Learner Model Training -----")

meta_dim = self.X_train_meta.shape[1]

ffnn = Sequential([
    Input(shape=(meta_dim,)),
    Dense(16, activation='relu'),
    Dropout(0.2),
    Dense(1, activation='sigmoid')
])
```

```
        #ffnn compilation
        ffnn.compile(optimizer=Adam(learning_rate=0.001),
                     loss='binary_crossentropy',
                     metrics=['accuracy'])

        #early stopping
                  early_stop   =   EarlyStopping(monitor='val_loss',   patience=10,
restore_best_weights=True)

        ffnn.fit(self.X_train_meta, self.y_train,
                 validation_data=(self.X_test_meta, self.y_test),
                 epochs=100, batch_size=16, verbose=1, callbacks=[early_stop])

        self.models['meta'] = ffnn

    def evaluation(self):
        print("\n ----- Final Performance Evaluation -----")

        final_probs = self.models['meta'].predict(self.X_test_meta)

        #random forest evaluation
        rf_preds = self.models['rf'].predict(self.X_test)
        rf_acc = accuracy_score(self.y_test, rf_preds)

        #naive bayes evaluation
        nb_preds = self.models['nb'].predict(self.X_test)
        nb_acc = accuracy_score(self.y_test, nb_preds)

        #meta learner (ffnn) evaluation
```

```
        meta_preds = (final_probs >= 0.5).astype(int)
        meta_acc = accuracy_score(self.y_test, meta_preds)

        #comparison table
        print("\n--- Model Accuracy Comparison ---")
        print(f"Random Forest Only:    {rf_acc:.2%}")
        print(f"Naive Bayes Only:      {nb_acc:.2%}")
        print("-" * 30)
        print(f"Meta Learner Accuracy: {meta_acc:.2%}")

        #result context
        best_base = max(rf_acc, nb_acc)
        if meta_acc > best_base:
            print(f"\nStacking improved accuracy by {meta_acc - best_base:.2%}")
        elif meta_acc == best_base:
            print(f"\nStacking matched the best base model.")
        else:
            print(f"\nStacking slightly underperformed (Overfitting risk).")


        #confusion matrix
        print("\n--- Confusion Matrix (Stacked Model) ---")

        conf = confusion_matrix(self.y_test, meta_preds)
        print(f"Confusion Matrix: \n{conf}")

        print("\nClassification Report (1=Regular, 0=Irregular):")
        print(classification_report(self.y_test, meta_preds))
```

```
    def recommend_program(self, student_dict, threshold=0.70):
        print("\n ----- Program Analysis and Recommnedation -------")

        current_program = student_dict.get('program', 'Unknown')

        def get_prediction(prog_name):
            try:
                temp_df = pd.DataFrame([student_dict])
                temp_df['program'] = prog_name

                row_vals = []

                for col in self.features:
                    if col not in temp_df.columns: continue
                    val = temp_df.iloc[0][col]

                    if col in self.encoders:
                        try:
                            val = self.encoders[col].transform([val])[0]
                        except:
                            val = 0
                    row_vals.append(val)

                X_input_df = pd.DataFrame([row_vals], columns=self.features)
                X_input = self.scalers.transform(X_input_df)

                 p_rf = self.models['rf'].predict_proba(X_input)[:, 1].reshape(-1,
1)
                 p_nb = self.models['nb'].predict_proba(X_input)[:, 1].reshape(-1,
1)
```

```
                meta_input = np.hstack([p_rf, p_nb])


                return self.models['meta'].predict(meta_input, verbose=0)[0][0]
            except Exception as e:
                print(f"Skipping {prog_name} due to data error: {e}")
                return 0.0


        #program evaluation if success rate is high
        current_score = get_prediction(current_program)
        print(f"\nCurrent Choice: {current_program}")
        print(f"Predicted Success Rate: {current_score:.2%}")


        if current_score >= threshold:
                 print(f"Prediction shows high confidence. The {current_program}
program fits the student")
            return


        else:
             print(f"Prediction shows significant risk. Program chosen falls below
{threshold:.2%}")
                 print(f"Searching for alternative programs with highher sucess
rates... ")


            available_programs = self.encoders['program'].classes_
            recommendations = []


            for prog in available_programs:
                if prog == current_program: continue


                score = get_prediction(prog)
```

```
                if score >= threshold:
                    recommendations.append((prog, score))

            recommendations.sort(key=lambda x: x[1], reverse=True)

            if not recommendations:
                print("\n No other program exceeds the 70% threshold based in your
profile")
                print("However, these are your best options:")

                all_prog = []
                for prog in available_programs:
                    if prog ==  current_program: continue
                    all_prog.append((prog, get_prediction(prog)))
                all_prog.sort(key=lambda x: x[1], reverse=True)
                recommendations = all_prog[:3]

                print(f"\n Recommended Alternatives:")
                for i, (prog, score) in enumerate(recommendations[:3], 1):
                    print(f"{i}, {prog} ({score:.2%} chance of regular)")

if __name__ == "__main__":

    #initializing the class
                            stacker       =       RecommenderSystem(r"C:\Users\Rex
Gatchalian\Desktop\Recommendation_Model\college_data.csv")

    #Loading and training the models
    stacker.loading_data()
    stacker.train_stacking()
```

```
    stacker.evaluation()



    # print("\n" + "="*50)

    # print("   TESTING RECOMMENDER FUNCTIONALITY   ")

    # print("="*50)



    # #debugger

    # program_0 = stacker.encoders['program'].inverse_transform([0])[0]

    # skill_0 = stacker.encoders['hard_skills'].inverse_transform([0])[0]

        #  print(f"\n[debug]  ID  0  maps  to  ->  Program:  '{program_0}',  Skill:
'{skill_0}'")



    # test_student_1= {#mismatch student, should prompt recommendations

    #     'student_name': 'Student A',

    #     'strand': 'HUMSS',

    #     'gwa': 80,

    #     'verbal': 2,

    #     'numerical': 1,

    #     'abstract': 2,

    #     'spelling': 1,

    #     'usage': 2,

    #     'program': 'Medical Biology',

    #     'soft_skills': 'Critical Thinking',

    #     'hard_skills': 'IT Fundamentals'

    # }

    # print(f"\n--- Analyzing: {test_student_1['student_name']} ---")

    # stacker.recommend_program(test_student_1)



    # print(f"\n--- Testing {test_student_1['student_name']} ---")
```

```
    # Debug checker: Manually check encoding before running the full function

    # try:

                              #                                       prog_id       =
stacker.encoders['program'].transform([test_student_debug['program']])[0]

    #     print(f"Program '{test_student_debug['program']}' found! ID: {prog_id}")

    # except:

      #       print(f"ERROR:  Program  '{test_student_debug['program']}'  NOT  FOUND.
Defaults to 0.")



    # try:

                              #                                       skill_id       =
stacker.encoders['hard_skills'].transform([test_student_debug['hard_skills']])[0]

       #         print(f"Skill  '{test_student_debug['hard_skills']}'  found!  ID:
{skill_id}")

    # except:

     #      print(f"ERROR: Skill '{test_student_debug['hard_skills']}' NOT FOUND.
Defaults to 0.")



    # test_student_2= {  #match student

    #     'student_name': 'Student B',

    #     'strand': 'STEM',

    #     'gwa': 96,

    #     'verbal': 3,

    #     'numerical': 3,

    #     'abstract': 2,

    #     'spelling': 2,

    #     'usage': 3,

    #     'program': 'Computer Science',

    #     'soft_skills': 'Critical Thinking',

    #     'hard_skills': 'Programming'

    # }
```

```
# print(f"\n--- Analyzing: {test_student_2['student_name']} ---")

# stacker.recommend_program(test_student_2)
```

A Machine Learning Approach for a Best College Program Recommendation System Based on Cognitive and Skill

Good day!

We are the 4th-year students from the Bachelor of Science in Computer Science currently conducting our thesis study entitled:

"A Machine Learning Approach for a Best College Program Recommendation System Based on Cognitive and Skill"

This survey serves as a primary data collection instrument for our study. The purpose of this study is to collect relevant data on students' cognitive abilities, academic performance, and skill sets to support the development of a recommendation model that can suggest best college programs tailored to individual strengths and interests.

Participants will be asked to provide information regarding their demographic background, General Weighted Average (GWA), admission test result, and skill sets. The data gathered will be used solely for academic and research purposes, specifically to train and evaluate machine learning models that aim to improve the accuracy and personalization of program recommendations. Participation in this survey is voluntary, and respondents have the right to withdraw their participation at any stage without any negative consequences.

By completing the survey, participants provide informed consent for their anonymized data to be used for academic research purposes only. **This survey will take 3 - 5 minutes to complete.**

All data collected from the respondents will be treated with the highest level of confidentiality and used strictly for academic and research purposes. In compliance with **Republic Act No. 10173, also known as the Data Privacy Act of 2012,** respondents' identities will not be disclosed, and all data will be anonymized prior to analysis.

Researchers:
Gatchalian, John Rex O,
Jamion, Aneza H.
Navarro, Lance Armstrong T.
Recarze, Spledelyn Cristine P.

Steps to find your admission test results:
1. Go to the gmail account you used to apply for admission
2. Navigate to "Important"
3. Click the filter "Attachment" and check PDFs box

4. Find the "PSU MAIN - Admission Result"

* Indicates required question

1. I have read and understood the above information and voluntarily agree to participate in this survey. *

Check all that apply.

- [ ] Yes

Demographic Background

This section aims to gather general background information about the respondents. The data collected will help the researchers analyze patterns and trends across various student groups.

2. **Name** *(optional)*

_____

3. **Sex**

Mark only one oval.

- ◯ Female
- ◯ Male
- ◯ Other: _____

4. **Senior High School Strand** *

Mark only one oval.

- ◯ STEM (Science, Technology, Engineering, and Mathematics)
- ◯ ABM (Accountancy, Business, and Management)
- ◯ HUMSS (Humanities and Social Sciences)
- ◯ GAS (General Academic Strand)
- ◯ TVL (Technical-Vocational Livelihood)
- ◯ Other: _____

5. **General Weighted Average** *(Senior High Final Grade - Grade 12)* *

_____

6. College Admission Test Result/Rating for **Verbal Reasoning** *(based on your College Admission Test result from the Admin)* *

Mark only one oval.

- ◯ Low
- ◯ Average
- ◯ High

7. College Admission Test Result/Rating for **Numerical Reasoning** *(based on your College Admission Test result from Admin)* *

Mark only one oval.

- ◯ Low
- ◯ Average
- ◯ High

8. College Admission Test Result/Rating for **Abstract Reasoning** *(based on your College Admission Test result from Admin)* *

Mark only one oval.

- ◯ Low
- ◯ Average
- ◯ High

9. College Admission Test Result/Rating for **Spelling** *(based on your College Admission Test result from Admin)* *

Mark only one oval.

- ◯ Low
- ◯ Average
- ◯ High

10. College Admission Test Result/Rating for **Language Usage** *(based on your College Admission Test result from Admin)* *

Mark only one oval.

- ◯ Low
- ◯ Average
- ◯ High

11. **Name of College** *

Mark only one oval.

- ◯ CS (College of Sciences)

12. **Name of Program** *(E.g. BS Computer Science)* *

Mark only one oval.

- BS Computer Science
- BS Environmental Science
- BS Information Technology
- BS Medical Biology
- BS Marine Biology

13. **Academic Status** *

Mark only one oval.

- Regular
- Irregular

14. **Socio-economic Status** *(Monthly family income)* *

Mark only one oval.

- 5, 000 - below
- 5,001 - 10,000
- 10, 001 - 20, 000
- 20, 001 - 30, 000
- 30, 001 - 40, 000
- 40, 001 - 50, 000
- 50, 000 - above

Skip to question 15

Skill Assessment

This section aims to assess the respondents' skill sets that may influence college program suitability.

15. What three soft skills do you have that you think help you succeed in your current program? List at least three. *(E.g. communication, creativity, etc.)* *

16. Among the three listed soft skills, which is the most important skill that help you succeed in your program? *

17. What three hard skills do you have that you think help you succeed in your current program? List at least three. *(E.g. data analysis, graphic design, etc.)* *

18. Among the three listed hard skills, which is the most important skill that help you succeed in your program?

Skip to section 4 (Submission and Confirmation)

Submission and Confirmation

Thank you for taking the time to complete this survey. Your honest and thoughtful responses are appreciated to the success of this thesis study.

By submitting this form, you confirm that:

- All the information and answers you have provided are **true and accurate.**
- You **voluntarily participated** in this study and understand that your responses will remain **anonymous and confidential.**
- You **consent** the inclusion of your anonymized data in the analysis and reporting of research findings.

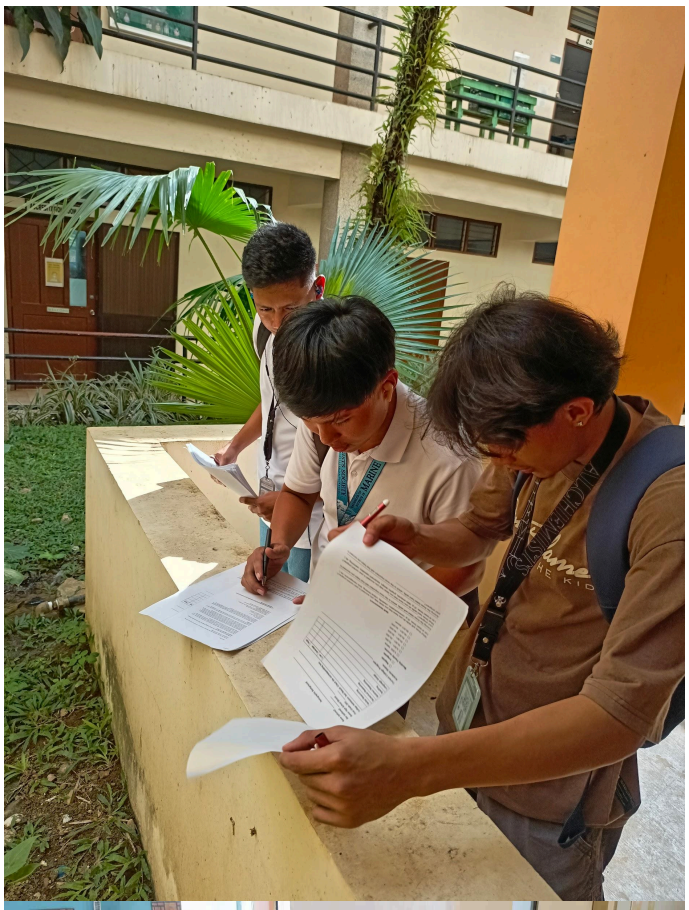
SAMPLE GENERATED OUTPUTS

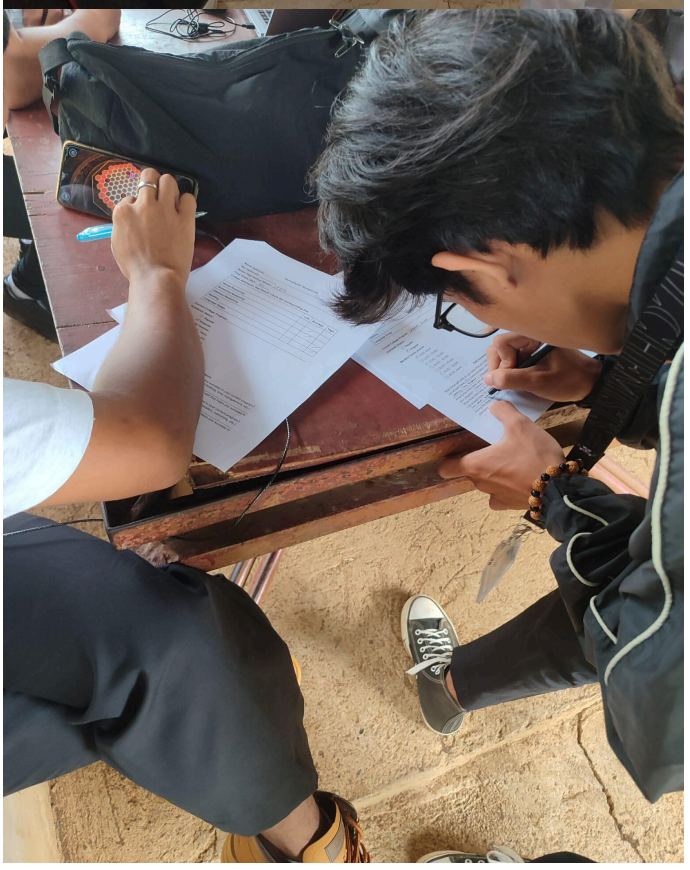
```
--- Model Accuracy Comparison ---
Random Forest Only:     84.44%
Naive Bayes Only:       78.89%
--------------------------------
Meta Learner Accuracy: 88.89%

Stacking improved accuracy by 4.44%

--- Confusion Matrix (Stacked Model) ---
Confusion Matrix:
[[17  9]
 [ 1 63]]

Classification Report (1=Regular, 0=Irregular):
              precision    recall  f1-score   support

           0       0.94      0.65      0.77        26
           1       0.88      0.98      0.93        64

    accuracy                           0.89        90
   macro avg       0.91      0.82      0.85        90
weighted avg       0.90      0.89      0.88        90
```

```
   TESTING RECOMMENDER FUNCTIONALITY
==================================================

--- Analyzing: Student A ---

 ----- Program Analysis and Recommnedation -------

Current Choice: Medical Biology
Predicted Success Rate: 66.37%
Prediction shows significant risk. Program chosen falls below 70.00%
Searching for alternative programs with highher sucess rates...

 No other program exceeds the 70% threshold based in your profile
However, these are your best options:

 Recommended Alternatives:
1, Marine Biology (64.47% chance of regular)
2, Environmental Science (63.55% chance of regular)
3, Computer Science (59.78% chance of regular)
```

```
--- Analyzing: Student B ---

 ----- Program Analysis and Recommnedation -------

Current Choice: Computer Science
Predicted Success Rate: 81.33%
Prediction shows high confidence. The Computer Science program fits the student
```

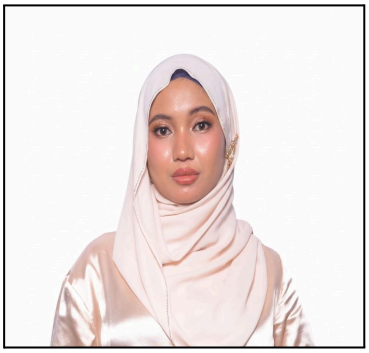
DATA GATHERING





Attachment G. BIONOTES OF THE RESEARCH PROPONENTS

UNDERGRADUATE RESEARCHERS

John Rex O. Gatchalian (Researcher 1) 	*John Rex Gatchalian is a Bachelor of Science in Computer Science student at Palawan State University. He is proficient in Data Analysis, Machine Learning and artificial intelligence and has a growing interest in Cybersecurity. Currently, his research involves developing a machine learning system to reduce the attrition rate in higher education.*
Aneza H. Jamion (Researcher 2) 	*Aneza Jamion is a Bachelor of Science in Computer Science student at Palawan State University, Puerto Princesa City. She was born on February 26, 2004 in Sofronio Española, Palawan. She completed her elementary education at Pulot Shore Elementary School, and secondary education at Pulot National High School.* *She is currently completing her undergraduate thesis entitled "A Machine Learning Approach for a Best College Program Recommendation System Based on Cognitive Abilities and Skill Sets" under the supervision of Mr. Adonis C Ampongan. Her academic interests include Human–Computer Interaction (HCI) and Web Application Development.* *She may be contacted at 202280050@psu.palawan.edu.ph.*
Lance Armstrong T. Navarro (Researcher 3) 	*Lance Armstrong T. Navarro is a dedicated 4th year student at Palawan State University, taking a degree of Bachelor of Science in Computer Science. Beyond his academic career, Navarro is committed to continuous learning. Upon completing his degree, he aims to pursue the field of application development to contribute to the tech industry.*

Spledelyn Cristine P. Recarze (Researcher 4) 	*Spledelyn Cristine P. Recarze is a 4th year student at Palawan State University, pursuing a Bachelor of Science in Computer Science. Throughout her academic career, Recarze has demonstrated a strong aptitude for communication and analysis. Looking forward, she aims to pursue a career of Web Development. Her career is driven by passion for innovation and academic excellence.*

THESIS/DISSERTATION ADVISER

Mr. Adonis C. Ampongan (Adviser) 	*Insert bionote*