DSC1109 - Machine Learning

Bayquen, Christopher Gilbert A.
Billones Cristel Kaye

A College Program Predictive Model

Machine Learning Final Project – Written Report

I. DATA SET THEME

This dataset aims to predict which college programs (e.g., engineering, business, arts, etc.) incoming university students are likely to choose based on academic performance, environmental factors, and personal interests. This data could help universities provide better academic counseling, optimize marketing for different programs, and streamline resource allocation for popular courses.

Target Attribute: College Program (Engineering, Business, Nursing, etc.)

Attribute Contexts

Environmental Factors These are external conditions related to students' surroundings, family backgrounds, and school settings that can impact their career choices.	B. High School Type (public, private, science highschool, technical) - *What type of high school did you attend?* D. Parent's Occupational Field (technology, business, arts, healthcare, etc.) - *In which field(s) do your parent(s) or guardian(s) work? Select all that apply. (to include both parents and/or multiple profession)* C. Average Daily Commute time to campus (mins) - *Approximately how many minutes do you expect to commute daily to reach your campus on average?* C.Parental Education Level (No formal education, High School Graduate, College Graduate, Postgraduate)

	- *What is the highest level of education attained by your parents/guardians? (Please select the highest level achieved by either parent or guardian)*
Cultural Factors These are societal norms, familial expectations, and traditional values that influence decisions, especially in the Philippines.	A. Gender - *What is your gender?* B. Influence of Family on Career Choice (low, moderate, high, very high) - *To what extent has your family influenced your decision to choose this program?*
Financial Factors These attributes assess the financial background and readiness of students, impacting their ability to pursue certain programs or institutions.	D. Primary Source of Financial Support (parents, self-funded, scholarships, government aid) - *What is your main source of financial support for college?* D. Family Income Level (Monthly household Income) - *What is your household's average monthly income?* D. Financial Preparedness (not ready, somewhat ready, ready, very ready) - *How financially prepared were you coming into college?*
Academic Factors These attributes reflect academic performance, preparation, and subject strengths, which can influence eligibility and interest in specific programs.	A. High School GPA (Grade Point Average upon graduating HS) - *What was your Grade Point Average (GPA) upon graduating high school?*

	B. Average Study Hours per day - *On average, how many hours per day did you spend studying in high school?* C. College Preparedness Level (unprepared, somewhat prepared, prepared, highly prepared) - *How prepared did you feel going into college?* E. Perceived Strength in Science, Technology, Engineering, and Mathematics subjects (Very Weak, Weak, Moderate, Strong, Very Strong) F. Perceived Strength with Accounting, Business, and Management subjects (Very Weak, Weak, Moderate, Strong, Very Strong) G. Perceived Strength with Technical-Vocational-Livelihood subjects (Very Weak, Weak, Moderate, Strong, Very Strong) H. Perceived Strength in Humanities and Communications subjects (Very Weak, Weak, Moderate, Strong, Very Strong) - *How would you rate your strengths in the following subjects?*
Personal Factors These reflect individual preferences, passions, and extracurricular involvement, which are often driven by personal aspirations rather than external pressures.	A. Passion for Chosen Program (very low, low, moderate, high, very high)

	- *How passionate are you about pursuing this program?* A. Preferred Career Paths/Industry (technology, business, arts, healthcare, etc.) - *Which career paths or industries are you most interested in? (Select all that apply)* B. Subjects of Interest (math, science, history, literature, technology) - *Which subjects do you enjoy the most? (Select all that apply)* C. Hobbies/Extracurricular Activities (reading, sports, gaming, art, etc.) - *What are your hobbies or extracurricular activities? (Select all that apply)*

II. INTRODUCTION

Choosing a college program is a pivotal decision for incoming university students. Understanding the factors that influence these decisions can help universities provide tailored academic counseling, optimize marketing efforts, and allocate resources efficiently for popular courses. This project explores machine learning techniques to predict which college program a student is likely to choose based on academic performance, environmental factors, and personal interests.

The data used for this project were collected through survey forms hosted via Google Forms for easy accessibility and to extend the reach of the project. This will allow the researchers to be able to obtain necessary data to train the model from different parts of the country, ensuring diversity in student background and opportunities which may affect their personal data. Having this variety in locality of the data, will help capture a broader understanding of the patterns and behaviors of students and their environment which may influence their career decision making as they reach their college level education. A total of 100 respondents from Filipino students all over the country were collected through this survey.

The goal is to develop a robust predictive model using a structured dataset and evaluate its performance using metrics like accuracy, precision, recall, and F1 score. This report provides an overview of the dataset, feature engineering, machine learning models used, evaluation metrics, and results

III. EXPLORATORY DATA ANALYSIS REPORT

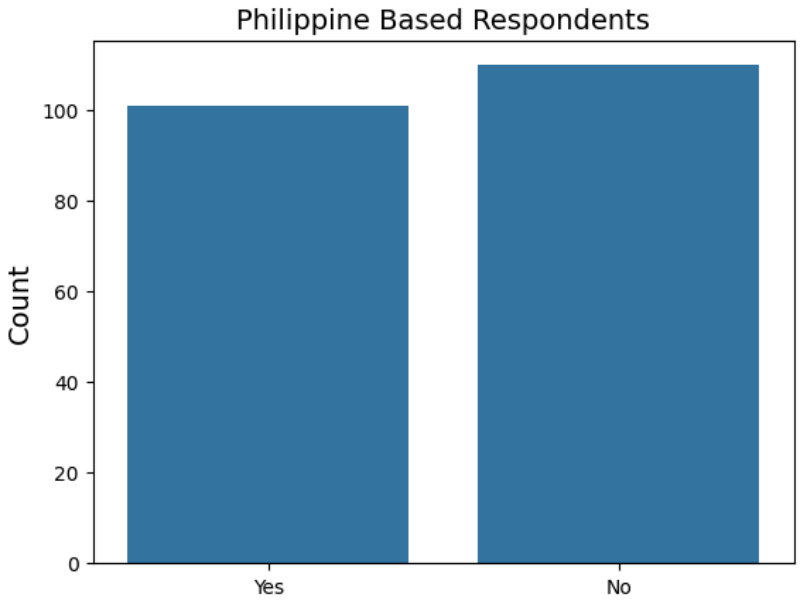

Figure 1. Philippine Based Respondents Distribution

Due to the nature of how the survey was disseminated, the data we collected reached respondents outside of the scope of the study– foreign students. A total number 211 responses were collected throughout the data collection process, having 100 Philippine Based respondents and 111 Non-Philippine Based respondents. This was then filtered to only have Philippine based respondents to ensure the integrity of the data.

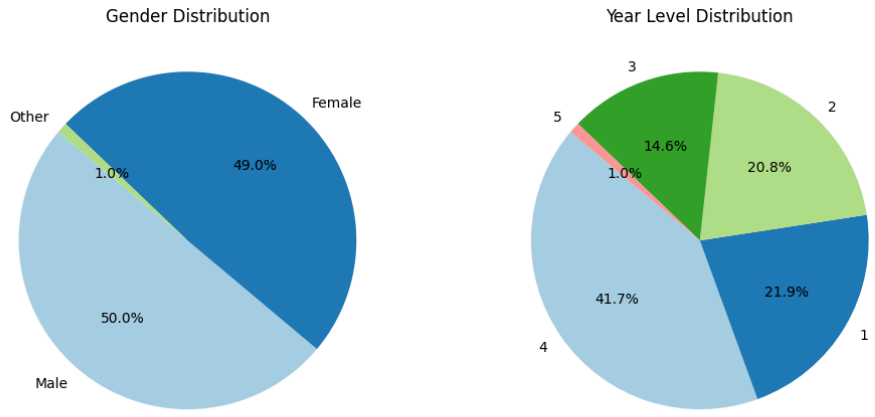

Figure 2. Gender and Year Level Distribution

The gender distribution showed a near 50/50 distribution between male and female, having only a 1% difference due to an “Other” response in the data. Majority of the college students that were surveyed were year level 4 students at 41.7%, followed by 1st year students at 21.9%, then 2nd year students at 20.8%, then 3rd year students at 14.6%, and lastly, 1% of the respondents were 5th year students.

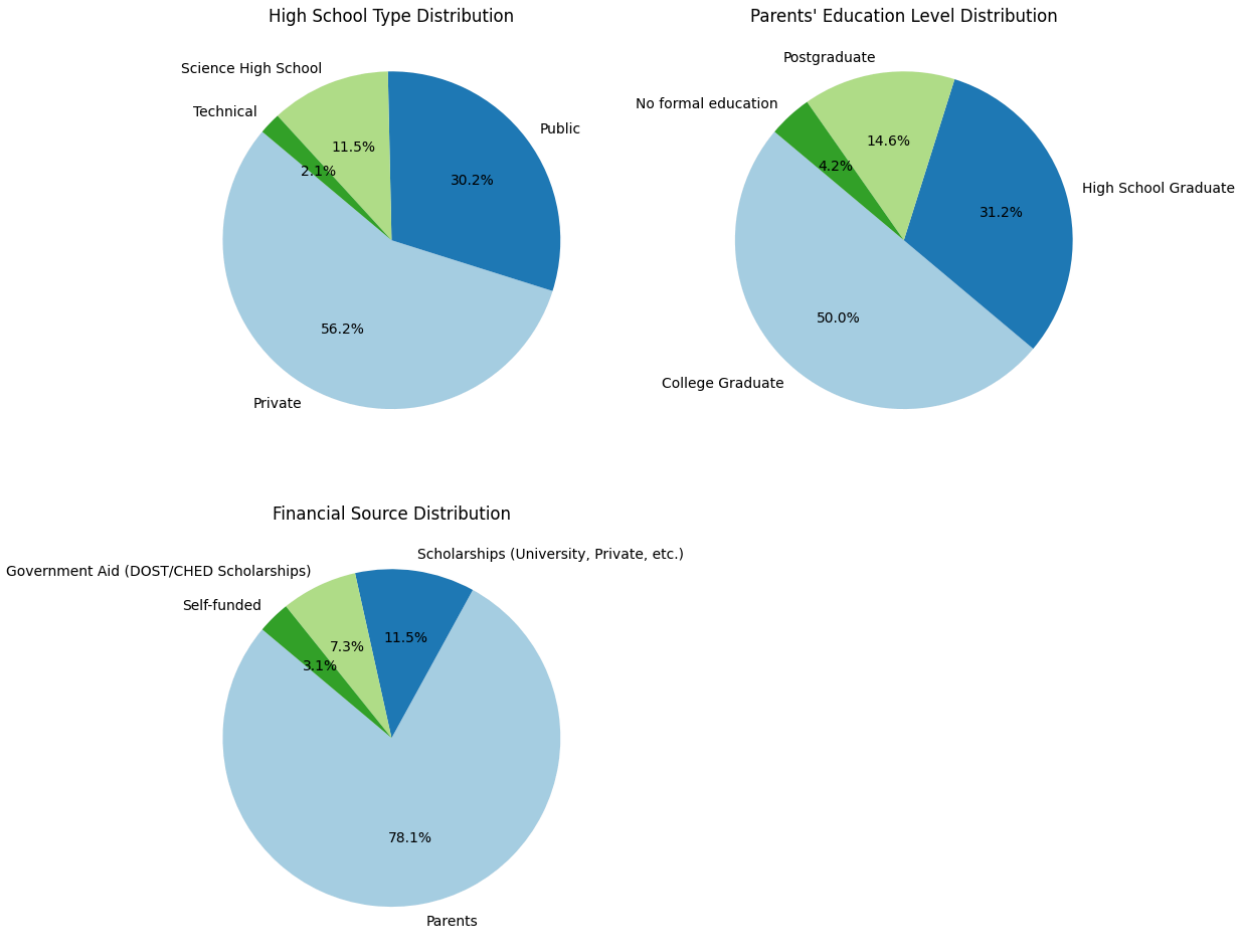

Figure 3. High School Type, Parent’s Education, and Financial Source Distribution

For the High School Type Distribution, most of the respondents came from private schools (consisting of 56.2% of the entire population), then 30.2% of the people came from public high schools, 11.5% coming from science high schools, and the remaining 2.1% coming from technical high schools.

The parent’s education level distribution showed that 50% of the respondent’s parents graduated college, meanwhile 31.2% of the respondents answered high school graduate as the highest education level of their parents, 14.6% of the students’ parents achieved a postgraduate degree, and lastly 4.2% having no formal education.

The financial source distribution shows that 78.1% of the students rely on their parents for their main financial support, while 11.5% rely on scholarships, followed by 7.3% of the students relying on government aid, and lastly, 3.1% of the students are self-funded.

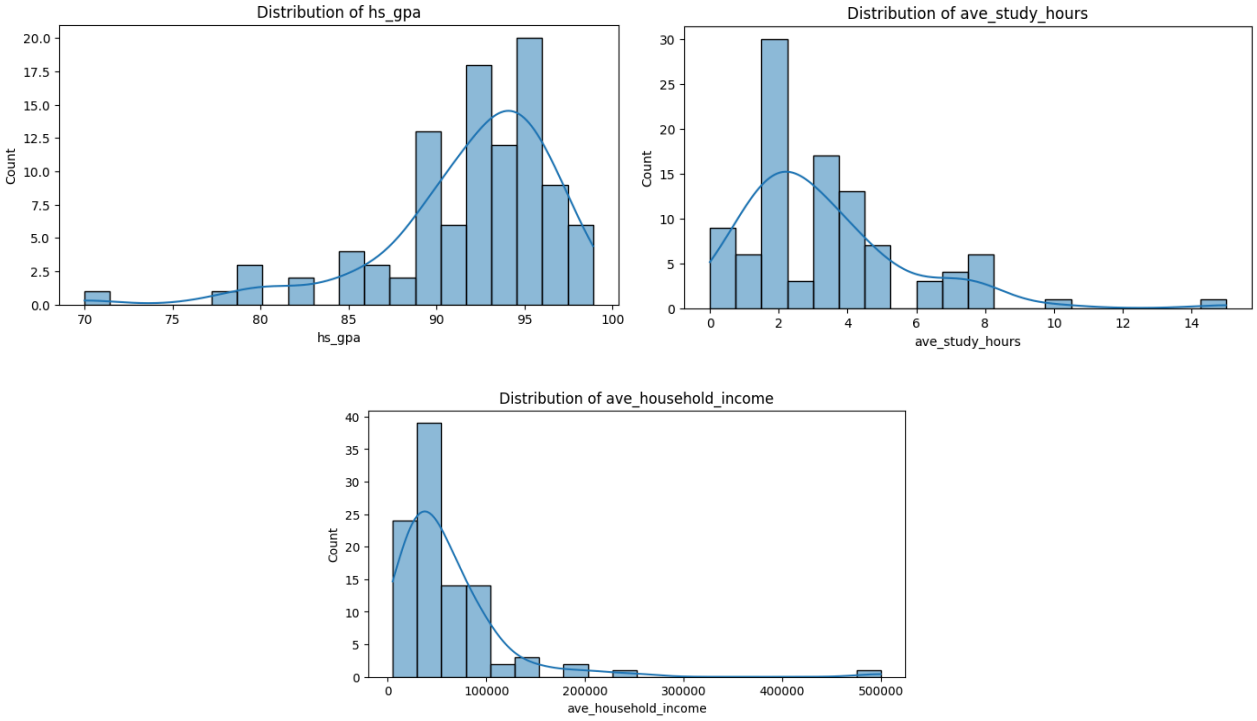

Figure 4. GPA, Average Study Hours, Household Income Distribution

In the plots above we can observe the distribution of the highschool GPA, average study hours, and average household monthly income– here we can observe that the data is skewed due to extreme outliers which will be verified and handled. These outliers where checked using the interquartile range approach which resulted to the following:

'hs_gpa'	**‘ave_study_hours’**	**‘ave_household_income’**
Index: 79, Value: 80.0	Index: 8, Value: 8.0	Index: 0, Value: 150000.0
Index: 85, Value: 82.0	Index: 17, Value: 8.0	Index: 7, Value: 200000.0
Index: 91, Value: 80.0	Index: 32, Value: 7.5	Index: 16, Value: 200000.0
Index: 92, Value: 70.0	Index: 64, Value: 10.0	Index: 53, Value: 150000.0
Index: 95, Value: 80.0	Index: 76, Value: 8.0	Index: 61, Value: 500000.0
Index: 98, Value: 77.45	Index: 80, Value: 8.0	Index: 66, Value: 150000.0
	Index: 91, Value: 15.0	Index: 73, Value: 250000.0
	Index: 98, Value: 7.5	

Table 1. Outliers

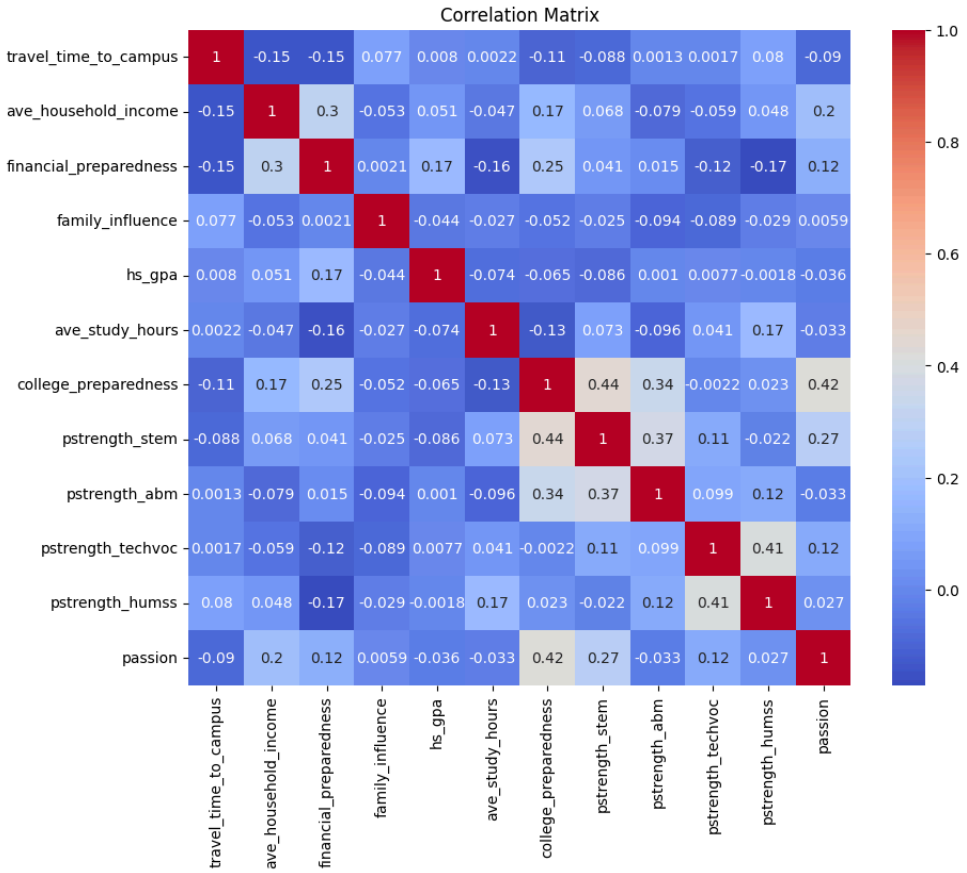

Figure 5. Numerical Features Correlation Heatmap

The figure above shows the correlation heatmap of the numerical features in the dataset. This correlation heatmap shows that most of the numerical features have nearing correlations with each other, and since the features that we have selected to be part of the survey were thoroughly researched, we will be utilizing all of the numerical features when creating the different machine learning models. Additional features may be added through feature engineering which will be discussed in Section V.

IV. DATA PREPROCESSING REPORT

As for the data preprocessing of the dataset that we have collected, minimal cleaning is necessary since we have formatted the survey to return data that is already structured and formatted in the data type that we needed. Only a few columns are required to be cleaned further, while handling missing values and outliers will also be implemented in this stage.

Checking for missing values

The missing values were checked using the *.isnull()* function of the pandas library which returned the number of cells in which there is a null value in it. The function was used to return a number of null values for each of the columns, and for the dataset that we have collected, only the ave_household_income column contained 1 missing value. This was handled by imputing the median value of the entire ave_household_income column.

Handling Outliers

Since the dataset that we are using only consists of 100 rows of data, and most of the outliers that we have detected does not differ too much from the range of values in in each column, we have decided to just drop the *extreme outliers* which will significantly affect the integrity of the data having either too high of a value in ‘ave_household_income’ (Index: 61, Value: 500000.0) or to low of a value in 'hs_gpa' (Index: 92, Value: 70.0). By dropping the significant extreme outliers, the total number of rows left in the data set for our modelling was 96 rows.

Handling Inconsistent Entries

Since the current_program field in our survey was set up to be user input, the values collected in this column contain inconsistent formatting for the same expected values. Some inconsistencies show different spelling of the same program and the tracks/specializations of each program were typed differently. Through extensive domain knowledge, the researchers automated the cleaning process of these inconsistencies by normalizing the character capitalizations, removing unnecessary prefixes and suffixes, then manually grouping the same values written in different formats.

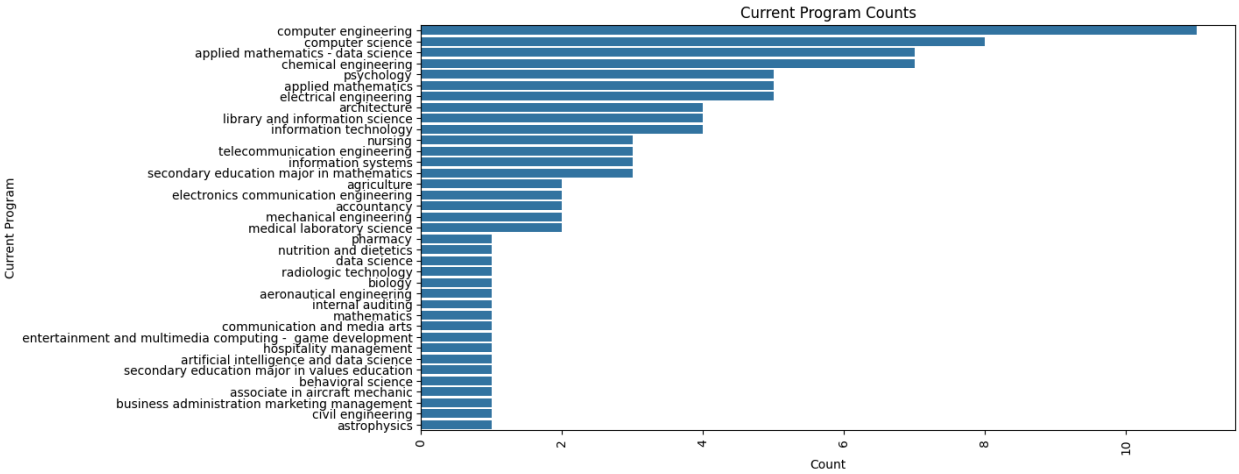

Figure 6. Current Program Distribution

After Cleaning and preprocessing, we have the following distribution of our target attribute. In this figure we can already observe a critical issue with the data that we gathered– and that is, the current programs that we have gathered are skewed to STEM careers and programs. As we can observe, there are multiple programs that also only has a frequency of 1, this imbalance in the classes will affect the models that we are going to train.

V. FEATURE ENGINEERING REPORT

Feature engineering is a crucial step in machine learning, as it involves creating new features or modifying existing ones to better represent the underlying patterns in the data. For this project, the following features were engineered to improve the predictive power of the model:

STEM Background Check

This binary feature indicates whether at least one parent works in a STEM-related field (e.g., science, technology, engineering, or mathematics). A parent's occupation in a STEM field can influence a student's academic interests, career aspirations, and likelihood of choosing STEM-related programs in college. For instance, children of STEM professionals may have greater exposure to technical fields and resources.

This feature serves as a proxy for early academic influence and role modeling, which is particularly relevant for predicting choices related to engineering, IT, and science programs.

Count of Extracurricular Activities

This feature tallies the number of extracurricular activities a student participates in, such as sports, music, art, or gaming. Extracurricular activities provide insight into a student's interests, time management skills, and alignment with specific college programs

This feature enriches the model's ability to identify behavioral and personality traits that align with specific academic tracks.

Number of Interested Subjects

This feature counts the total number of subjects (e.g., math, literature, science) that a student has expressed interest in; The breadth of academic interests reflects the student's curiosity and versatility. For example, A wide range of interests may indicate a preference for interdisciplinary or general studies programs or a narrow focus might suggest specialization in technical fields.

This feature helps to capture the diversity and depth of a student's academic preferences, enabling the model to better understand program alignment.

Average Academic Strength

This numerical feature is calculated as the average perceived strength across academic areas (e.g., STEM, business, humanities, etc.). Students self-assess their capabilities in each area, and the scores are aggregated to produce an overall average. This feature reflects the student's academic confidence and aptitude. By summarizing a student's self-perception of academic strengths, this feature acts as a key indicator of their readiness and interest in specific domains.

VI. MODEL PERFORMANCE COMPARISONS

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.10	0.575000	0.10	0.083333
KNN	0.30	0.728333	0.30	0.211905
Decision Trees	0.05	0.575000	0.05	0.060000
Naive Bayes (Gaussian)	0.10	0.600000	0.10	0.066667
Linear Discriminant Analysis (LDA)	0.05	0.656667	0.05	0.011765
Support Vector Machine (SVM)	0.05	0.516667	0.05	0.025000
Random Forest	0.25	0.757500	0.25	0.181905
Gradient Boosting Machine (GBM)	0.05	0.616667	0.05	0.025000
XGBoost	0.05	0.516667	0.05	025000

The table summarizes the performance of various machine learning models on a classification problem using four key metrics: Accuracy, Precision, Recall, and F1 Score. These metrics provide different perspectives on the models' effectiveness:

- **Accuracy**: The proportion of correctly classified samples.
- **Precision**: The proportion of correctly predicted positive observations out of all predicted positives.
- **Recall**: The proportion of correctly predicted positive observations out of all actual positives.
- **F1 Score**: The harmonic mean of Precision and Recall, balancing these two metrics.

Accuracy
KNN has the highest accuracy (0.30), followed by Random Forest (0.25). Other models performed poorly, with accuracies of 0.10 or lower.

Precision
Random Forest achieves the highest precision (0.757500), indicating it minimizes false positives better than others. KNN also has high precision (0.728333).

Recall
KNN leads in recall (0.30), suggesting it detects the highest proportion of true positives among all actual positives.Random Forest follows with a recall of 0.25.

F1 Score

KNN has the highest F1 Score (0.211905), balancing precision and recall. Random Forest comes second with an F1 Score of 0.181905.

In this project, KNN is the best choice for this dataset, offering the most balanced performance across all metrics. Its high recall and F1 Score suggest it is effective for detecting positives and handling class imbalances.

VII. ASSOCIATION RULE MINING REPORT

Association Rule Mining is a machine learning technique that helps uncover hidden relationships between items in large datasets. In this project, we utilized the Apriori Algorithm to implement the association rule mining for our multivalued attributes.

careerpath_interest

Antecedents	Consequents	Support	Confidence	Rank
Education and Academia	STEM	0.20	0.952381	1
Business and Management	STEM	0.27	0.870968	2
Health and Medicine	STEM	0.26	0.866667	3
Law and Public Service	STEM	0.12	0.800000	4
Arts and Humanities	STEM	0.13	0.684211	5

hs_favesub

Antecedents	Consequents	Support	Confidence	Rank
Physical Education, Science	Technology	0.10	0.833333	1
Technology, Mathematics	Science	0.22	0.785714	2
Physical Education, Technology	Science	0.10	0.769231	3
History, Science	Technology	0.13	0.722222	4

Physical Education	Technology	0.33	0.687500	5

hs_extracurr

Antecedents	**Consequents**	**Support**	**Confidence**	**Rank**
Sports, Coding and Technology	Gaming	0.12	0.923077	1
Gaming, Science Clubs	Coding and Technology	0.12	0.857143	2
Science Clubs, Coding and Technology	Gaming	0.12	0.800000	3
Coding and Technology	Gaming	0.20	0.645161	4
Photography	Music and Dancing	0.15	0.625000	5

Parents_fields

This is due to the nature of the data having only more or less, two values for each respondent.

Antecedents	Consequents	Support	Confidence	Rank

VIII. NEURAL NETWORK SUMMARY

The network was designed as a multi-layer perceptron (MLP) with an architecture tailored for multi-class classification. It consisted of an input layer followed by two hidden layers and an output layer. The first hidden layer included 64 neurons with ReLU activation and L2 regularization to prevent overfitting. A dropout layer with a rate of 0.3 was applied to randomly deactivate 30% of neurons during training, enhancing generalization. The second hidden layer had 32 neurons with ReLU activation and similar L2 regularization, followed by another dropout layer with a rate of 0.2. The output layer used a softmax activation function to generate probabilities for each class (college program). The model was compiled using the Adam optimizer with a learning rate of 0.001, categorical cross-entropy loss function, and accuracy as the performance metric.

The data preparation process included standardizing input features to ensure compatibility with the neural network and encoding the target variable (college program). LabelEncoder was used to map categorical labels to numerical values, which were then converted to one-hot encoded format for training. The dataset was split into training and testing sets, with 10% of the training data reserved for validation. Early stopping was employed to monitor validation loss, stopping training after 10 epochs of no improvement and restoring the best model weights. The model was trained for a maximum of 1,000 epochs with a batch size of 32.

The neural network achieved promising results, effectively capturing non-linear relationships within the data. Key strategies such as L2 regularization and dropout layers helped mitigate overfitting, while the network's adaptability allowed it to model complex patterns in student preferences. Evaluation on the test set demonstrated strong generalization capabilities, with the model achieving a test accuracy of approximately **20%.**