Scientific Journal of Computer Science
futuristech PT. Teknologi Futuristik Indonesia
Date of publication March 11, 2026, date of current version March 11, 2026.
Digital Object Identifier 10.64539/sjcs.v2i1.2026.397

Article

Smart Choice: Machine Learning Insight into Factors Influencing Students' Programme Selection at the Tertiary Institution

Abubakar Mahami Yakubu[1], **Elijah Ofori**[1,*]

[1] Department of ICT Education, University of Education, Winneba, P. O. Box 25, Winneba, Ghana.
* Correspondence: Elijah Ofori, elijah.ofori@yahoo.com

The authors received no financial support for the research, authorship, and/or publication of this article.

Abstract: Understanding the factors influencing students' choice of programme of study is increasingly important for tertiary institutions in Ghana, particularly amid rising enrolment rates and growing competition. While prior studies have applied machine learning to predict academic performance, limited research has examined programme selection behaviour at the senior high school level using mixed-type clustering techniques. This study addresses this gap by applying the K-prototype clustering algorithm and supervised classification models to survey data collected from 1,042 final-year Business and Home Economics students across ten senior high schools in Northern Ghana. The clustering process identified three behavioural segments comprising 423, 382, and 237 students, respectively, with the majority aged 16–20 years. Internal validation metrics indicated modest cluster separation. Subsequent classification modelling using Naïve Bayes, Logistic Regression, Decision Tree (J48), Random Forest, and Support Vector Machine (SVM) showed that SVM achieved the highest predictive performance (Accuracy = 99%) when predicting cluster membership. Key influencing factors included parental education, parental occupation, counselling exposure, socio-cultural beliefs, and peer influence. The findings highlight the need for strengthened, context-sensitive guidance and counselling frameworks at the pre-tertiary level to support informed and independent programme selection decisions.

Keywords: Educational data mining (EDM); Machine learning (ML); Artificial Intelligence (AI); K-prototype; Students; Programme Choice.

1. Introduction

The rapid expansion of tertiary education in Ghana has contributed to a steady increase in student enrolment across institutions [1]. In 2022, tertiary enrolment reached approximately 635,000 students, representing a 9.34% increase compared to the previous year [2]. This growth intensifies competition among institutions and underscores the need to understand the factors influencing students' programme selection decisions. Strategic educational decisions shape institutional policies, economic development, and workforce readiness [3], [4].

Identifying the determinants of programme choice is essential for developing a skilled workforce capable of driving technological innovation and national development [5]. While numerous studies have explored student enrolment trends and academic performance prediction, limited research has focused on modelling programme selection behaviour at the senior high school (SHS) level prior to tertiary entry. Existing studies often rely on structured institutional databases and predominantly numerical features, overlooking mixed categorical and behavioural attributes that influence decision-making.

Furthermore, traditional clustering methods such as K-means are not well suited for mixed-type datasets, limiting interpretability when both categorical and numerical variables are involved. This gap restricts the ability to generate behavioural segmentation insights that are contextually meaningful in pre-tertiary environments.

To address this limitation, this study applies the K-prototype clustering algorithm, which accommodates mixed categorical and numerical data, to segment SHS students based on socio-economic and behavioural attributes. The study further integrates supervised classification modelling to evaluate predictive patterns across derived clusters. By combining mixed-type clustering and classification analysis, this research provides a data-driven framework for understanding programme selection behaviour in Ghanaian senior high schools.

The following research questions guide the study:

1) How many appropriate and distinct clusters are generated from the mixed-type dataset?
2) Which classification algorithm demonstrates the highest predictive performance?
3) Which attributes most strongly influence students' programme selection decisions?

The remainder of this paper is structured as follows: Section 2 reviews related literature on programme selection and educational data mining. Section 3 describes the dataset and methodology. Section 4 presents clustering and classification results. Section 5 discusses findings and implications. Section 6 concludes the study and outlines directions for future research.

2. Literature Review

Machine Learning (ML) is a sub-field of Artificial Intelligence (AI) which adds the ability for machines to automatically learn and improve from experience without setting up any rules or explicitly programming systems. Alan Turning studied the question "can machines think?"[6]. In his work. Since then, the field has experienced its major evolution in the last two decades and the availability of more learning data is made possible by digitization of the society. ML focuses on the development of computer programmes that can access data and use it to learn for themselves [7].

In educational data mining (EDM), predicting student decision-making behaviour with machine learning (ML) algorithms has taken the stage. Large volumes of data are collected from sources by educational institutions [8], [9]. One of the main goals of educational research is to understand how students make decisions. These choices have a big impact on the student's general wellbeing, career development, and academic performance. Finding trends that influence students' academic and behavioural decisions has become more important as educational institutions, policymaker, and educators increasingly turn to data-driven approaches.

Robust algorithms that can analyse large datasets have been made possible by advances in machine learning, which can help uncover important information about how students make decisions. Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees, and Ensemble methods were used in this study. The Ensemble method outperformed the traditional statistical models in predicting student decision-making [10]. Educational Data Mining (EDM) constitutes an emerging research field, which has gained popularity in developing methods for exploring the unique types of data that originate from the contemporary educational era because of its potential to improve the quality of the educational institutions and their policies. This area of research has grown exponentially due to the fact that it enables educational stakeholders to discover new and useful knowledge about students and has the potential to improve some aspects of quality education.

In this study a proposed hybrid predicting system incorporating a number of possible machine learning algorithms was used: Naive Bayes (NB), Back Propagation (BP), Repeated Incremental Pruning to Produce Error Reduction (RIPPER), 3NN, C4.5 and Sequential Minimal Optimization (SMO). The objectives of this study are to support student admission procedures and strengthen the service system in educational institutions. The 3NN algorithm presents the best performance with 59% [11]. In order to choose the best career path, the students in this study need to identify their area of interest to determine their academic interests. In order to assist engineering students in decision-making, this research uses machine learning classification algorithms to predict the career they can choose after graduation. The algorithms compared and examined the classifier results that were produced. Accuracy score, confusion matrix, heatmap, and classification report are the criteria utilized to develop the classifiers. With an accuracy of 63.64%, the KNN was the most effective [12]. Predicting college placement based on academic performance is critical to supporting educational institutions and students in making informed future career decisions.

This research employs machine learning techniques to facilitate student placement based on academic performance data. This research uses data from grades, test scores, and extracurricular activities obtained from a varied sample of students. The study used machine learning algorithms, including LR, NB, RF, SVM, and KNN, to create the predictive model. The model is evaluated using the following performance criteria, such as accuracy, precision, recall, and F1-score. With an accuracy of 96%, RF performed the best, demonstrating that variables like grades and test scores significantly affect prediction accuracy [13].

Tertiary institutions of learning are constantly exploring factors that maximize enrollment of students into their various programmes of study. These factors provide management with information on the potential applicants who are likely to enroll into their institution. This research explores the effects of the various pre-admission factors, such as WAEC grades and JAMB scores, that may influence the enrollment of students in South-west-Nigeria. The study adopted machine learning techniques to evaluate the correlation of the different factors on students enrollment. Algorithms such as Artificial Neural Network (ANN), Logistic Regression (LR), Decision Tree

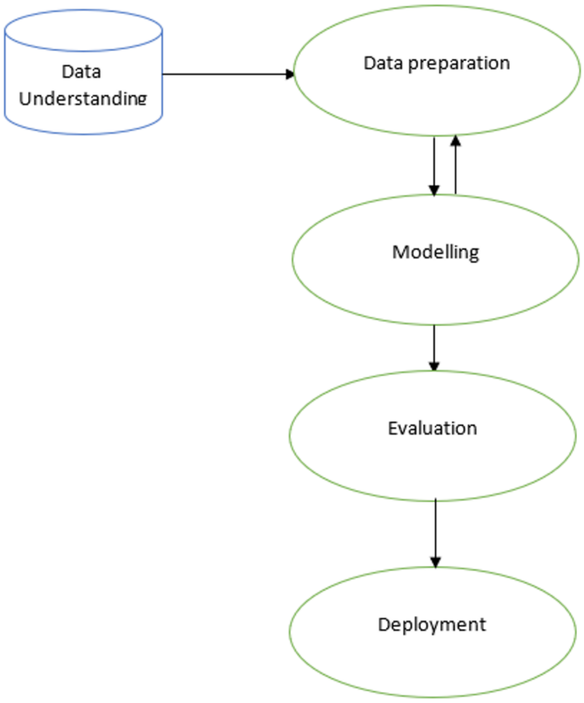

Figure 1. CRISP-DM Framework.

(ID3), Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Naïve Bayes (NB) were used. With a 97% accuracy rate, the Decision Tree is the most accurate [14].

3. Methodology

As illustrated in Figure 1, this study adopted the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology [15] as a comprehensive framework for guiding the data mining and machine learning workflow. CRISP-DM is a structured data mining framework that provides a systematic process for transforming raw data into validated models and actionable insights, and it is applicable to datasets of varying scale, including questionnaire-based survey datasets. In this study, CRISP-DM was operationalized through five stages: data understanding, data preparation, modelling, evaluation, and deployment.

The research data was collected from final-year senior high school students in the home economics and business classes in Northern region. The study utilized the convenient sampling approach, a nonprobability method for data collection. The convenient sampling method was adopted due to the respondents' immediate availability and accessibility. Google form was administered to students, and the research objectives were clearly stated. In addition, the respondents were mandated to agree to an ethics consent form before filling out the questionnaire. Throughout the CRISP-DM process, the nondisclosure and privacy of respondents' data were adhered to strictly. In adhering to the confidentiality of data, no information in the questionnaire could be traced back to the respondents. A total of 1042 responses were received.

3.1. The Dataset

As shown in Table 1, the dataset comprises twenty-nine (29) analytical attributes. Under the biodata, 31.57% of the respondents are males, whiles 68.43% are females, 10.9% are between 10-15years, 80.9% are between 16-20years and 8.3% are between 21-25years. Under the SHS information, 100% of the respondents attended mixed schools, with 63.6% boarding status and 36.4%-day status, 66% are home economics students and 34% are business students. According to the university tracker, 92.6% responded they intend to continue to tertiary institutions, and 52.4% responded they are not counselled in school.

3.2. Data Preparation

In the data preparation and cleaning phase, the dataset was screened to ensure completeness and consistency prior to modelling. Records with incomplete or inconsistent entries were excluded during screening, and the final dataset used for analysis contained 1,042 complete responses across all analytical variables. Since the modelling dataset contained no missing values, missing-value imputation was not required. Where applicable, numerical values were scaled to support classification performance, while categorical variables were retained in their original form for mixed-type clustering using the K-prototypes algorithm. As a result, the final dataset used for model training and evaluation was fully valid and suitable for classification.

3.3. K-Prototype Algorithm

Huang [16] proposed the k-prototypes algorithm for clustering mixed-type data, which combines the ideas of the k-means algorithm [17] and the k-modes algorithm [18]. The k-prototypes algorithm divides the dataset into $K(K \in N^{+})$ different subclusters to minimize the value of the cost function. The cost function is shown in the following formula:

$$F(U,Q) = \sum_{l=1}^{K}\sum_{i=1}^{n} u_{il} d(x_i, q_l) \quad (1)$$

The k-prototypes algorithm combines the "means" of the numerical part and the "modes" of the categorical part to build a new hybrid Cluster Centre "prototype." Based on the "prototype," it builds a dissimilarity coefficient formula and the cost function applicable to the mixed-type data. The parameter γ is introduced to control the influence of the categorical feature and the numerical feature on the clustering process. It is assumed that the mixed-type dataset has ρ numerical features and $m - \rho$ categorical features [16].

The questionnaire administered to the students contained both numeric and categorical attributes. The study therefore employed the K-prototype algorithm on the

Table 1. Dataset Attributes and Response Options.

Section	Item	Response
Biodata	Gender	Male, Female
	Age	10–15 years, 16–20 years, 21–25 years
	Form	Form 3
SHS Information	Programme of study at Senior High School	Business, Home Economics *(General Science appears in dataset: n=2)*
	Senior High School category	Mixed, Single
	Residential status in Senior High School	Day, Boarding
University Tracker	Do you intend to continue your education at a tertiary institution?	Yes, No
	Counselling service (in school)	Yes, No
	Does your current course relate to future tertiary studies?	Yes, No
	Which tertiary institution do you intend to go?	University, Polytechnic, Nursing Training, Teacher Training
	Indicate your purpose of being in school	For academic purposes, To find a job, To make friends, Because of my parents
	Consider quality and prestige of programme	Yes, No
	Consider job opportunities	Yes, No
	I have always wanted to study this programme	Yes, No
	Can your choice of programme be influenced by external factors?	Yes, No
	Are you choosing this programme because your friends are choosing it?	Yes, No
	Will your choice be based on your cultural beliefs?	Yes, No
	Will your choice be based on your gender?	Yes, No
	Is there demand in the job market for this programme?	Yes, No
	How often do you go out with friends?	Very low, Very high
Socio-economic Background	Student guardian	Father, Mother, Other
	Social class of family	Lower, Middle, Upper
	Mother's educational level	Uneducated, Moderately educated, Highly educated
	Father's educational level	Uneducated, Moderately educated, Highly educated
	Guardian educational level	Uneducated, Moderately educated, Highly educated
	Mother's work	Unemployed, Self-employed, Employed
	Father's work	Unemployed, Self-employed, Employed
	Guardian work	Unemployed, Self-employed, Employed
Dominant Influence Factor (CLASS)	CLASS (key factor affecting programme choice)	Personal Interest, Career Aspiration, Parent Occupation, Friends Influence, Cultural Background, Gender, Stereotyping, Cost of Programme, Industrial Expectation, Recognition of programme

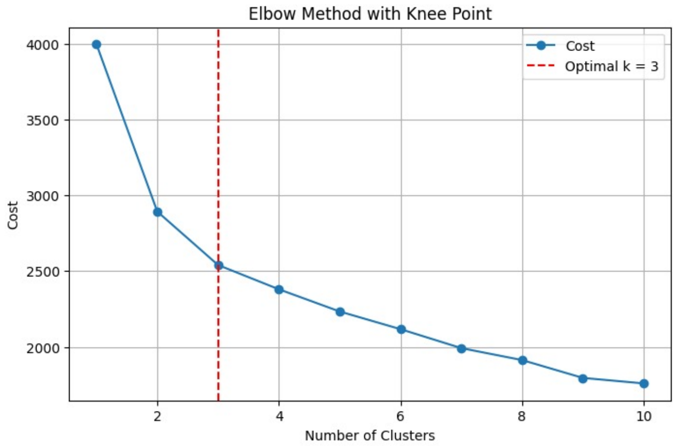

Figure 2. Elbow Method with Knee Point.

data set. The K-prototype algorithm integrates the K-means and the K-mode algorithm for mixed-type data. The K-prototype uses the Euclidean distance of K-means and the dissimilarity measurement of K-mode to cluster the data points.

3.4. Classification Algorithm

The researchers implemented classification algorithms, Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), and Logistic Regression (LR) and clustering (K-means, K-modes, and K-prototype) algorithms in analyzing the numerical and categorical datasets.

3.5. Evaluation

After building the model, the researcher evaluated the model through the training and testing of the dataset. We used accuracy, precision, recall and F1-Score.

The accuracy characterizes the degree to which a predicted value agrees with an actual value [19].

$$Accuracy = \frac{True\ positive + True\ Negative}{True\ Positive + False\ Positive + False\ Negative} \quad (2)$$

$$Precision \ = \frac{True\ positive}{True\ Positive + False\ Positive} \quad (3)$$

$$Recall = \frac{True\ positive}{True\ Positive + False\ Negative} \quad (4)$$

$$F - measure = \frac{2\mathrm{PR}}{\mathrm{P} + \mathrm{R}} \quad (5)$$

Where P represents precision and R represents recall.

3.6. Feature Selection Procedure

Feature selection was performed using a Random Forest–based importance ranking approach. A Random Forest classifier was trained on the encoded dataset to estimate feature importance scores based on impurity reduction (Gini importance). Features were ranked according to their importance values, and the top-ranking features were selected for classification modelling.

To mitigate potential information leakage and dominance of highly cluster-defining variables, selected features were further refined by excluding variables directly related to programme identity and counselling services in adjusted experiments. This two-stage feature selection process enhanced model interpretability while reducing overfitting risk.

3.7. Model Validation and Robustness Assessment

The classification task involved predicting cluster membership derived from the K-prototype algorithm. To evaluate predictive performance, the dataset was partitioned using an 80/20 train–test split. All models were trained on the training subset and evaluated on the held-out test subset.

To reduce potential information leakage, highly cluster-defining variables (e.g., programme of study and counselling service) were excluded from adjusted model evaluations. However, because cluster labels were derived from the same feature space, high classification accuracy reflects strong internal separability within the dataset rather than external predictive generalization.

4. Results

4.1. Cluster Characteristics

- *RQ1. How many appropriate and distinct clusters were generated by the model from the numeric and categorical dataset based on a student's programme selection?*

In response to RQ1, The K-prototype uses the cost function that combines numeric and categorical variables to determine the number of clusters. The elbow method for optimal selection was implemented. As shown in Figure 2, the cluster formation starts from 0, and the elbow function produces 3 clusters. Cluster zero (0), cluster one (1) and cluster two (2). The elbow graph shows the within-cluster sum-of-square (WCSS) values on the y-axis corresponding to different values of K (x-axis); the optimal k-value is the point at which the graph forms an elbow and from Figure 2, the optimal number of clusters at the point the elbow formed with a change in the cost function is at k = 3. After applying the elbow method and K-prototypes clustering, the dataset was grouped into three distinct student clusters (k = 3). Each cluster reveals unique behavioural, educational, and socio-economic characteristics that offer valuable insights into student orientation, motivation, and decision-making regarding academic programmes. The k-prototype algorithm is used to cluster mixed data types; it uses the idea of the K-means and K-mode algorithm to divide the dataset into different subclusters.

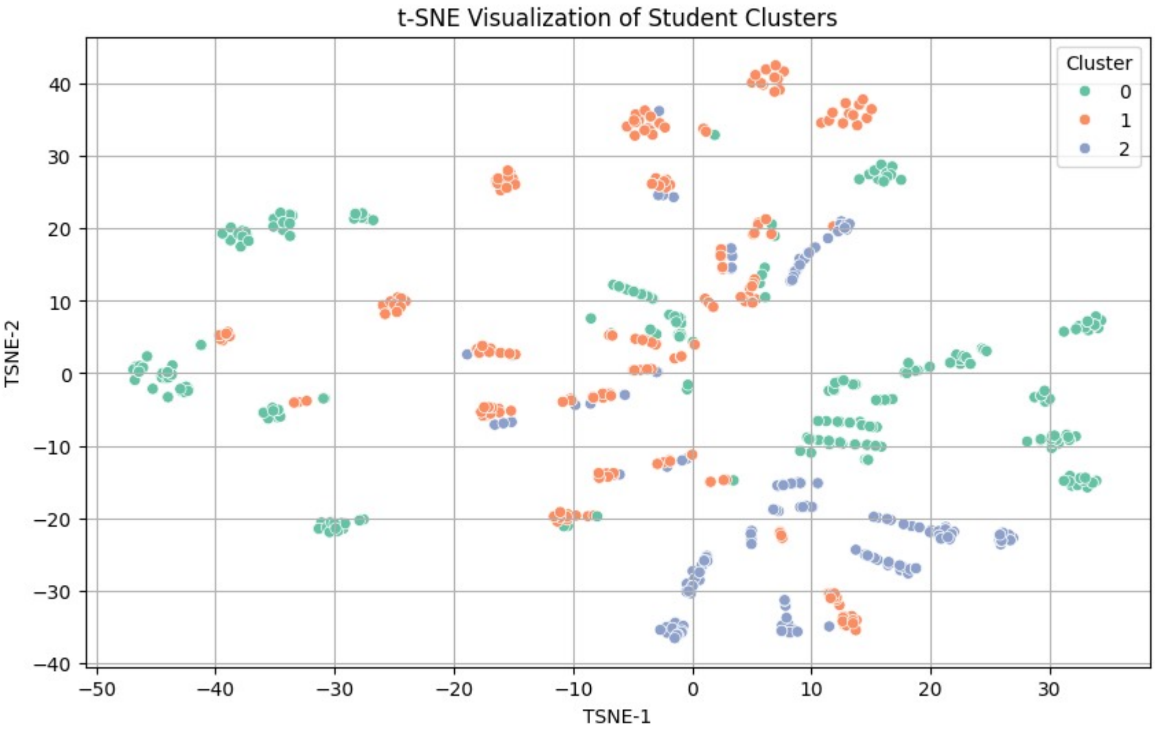

Figure 3. t-SNE Cluster Plot.

Table 2. Cross-Cluster Observations.

Aspect	Cluster 0	Cluster 1	Cluster 2
Core Driver	Academics	Peer/Cultural Influence	Social Belonging
Family Education	Low	Low	High
Parental Employment	Self-employed/employed	Unemployed	Employed
Social Activity	Very Low	Very High	Very High
Counselling Received	No	Yes	Yes
Tertiary Aspiration	Yes	Yes	Yes
Influence of Gender/Culture	No	Yes	Yes

In Figure 2, the number of clusters is determined by the K-prototype using the cost function in analyzing the numerical and categorical data, the questionnaire had mixed data types namely categorical data and numerical data. The elbow method for optimal cluster selection is used to determine the number of clusters. The cost was plotted against the number of clusters using the elbow method. When the data points are assigned to a cluster and the number of clusters is set to zero (0), the (WCSS) distance is frequently high. The sum of squared distances decreases as the number of clusters increases and eventually starts to decline. The elbow is the point where a spot is formed, and in this instant, it is at k = 3. And it indicates that the WCSS distance decreased when the number of clusters increased.

To quantitatively assess the cluster structure, internal validation metrics were computed. The Silhouette Score was 0.077, indicating modest separation between clusters. The Calinski–Harabasz Index was 69.91, while the Davies–Bouldin Index was 3.25. These values suggest that although the three-cluster solution provides interpretable behavioural segmentation, overlap exists among clusters, which is expected in self-reported behavioural survey data. Therefore, the clusters represent tendencies rather than strictly separable categories.

Cluster 0; Academic Independents: Cluster 0 comprises 423 students, primarily male and aged between 16–20 years, all enrolled in business classes. These students are most often day students attending mixed-gender SHS. They demonstrate a clear academic orientation, with all respondents in this group indicating that their primary purpose for being in school is for academic advancement. Additionally, they report a strong sense of career alignment with their current course of study, as well as recognition of job market demand for their chosen fields. Notably, these students' decisions appear to be internally motivated, they are unaffected by external influences such as peer pressure, gender norms, or cultural expectations and they do not benefit from career-focused guidance and counselling. The socio-economic background of this cluster reflects modest means. A majority of the students' parents and guardians are uneducated and work in self-employment or informal sectors. Despite these limitations, students in this group display independence in thought and behavior. Cluster 0 can therefore be characterized as self-determined, academically focused students who rely on personal motivation and long-term goals to navigate their educational paths.

Cluster 1; Influenced Scholars: This cluster had 382 participants with balanced gender distribution, mostly boarders in home economics classes aged 16-20 years. The cluster reflects socially influenced, moderately resourced students. While they show academic motivation, they are significantly shaped by peer groups, gender expectations, and cultural norms. They benefit from guidance and counselling, and career awareness programmes to ensure their decisions align with long-term goals rather than short-term social conformity.

Cluster 2; Privileged Socialites: This group had 237 students and embodies a socio-economically privileged student archetype, largely females aged 16-20, with high representation from home economics classes. Their decisions are interpersonally and culturally motivated, potentially at the expense of long-term educational alignment. While resources are abundant, they benefit from programmes that bridge social orientation with career-focused guidance and counselling and help balance autonomy with responsibility, Table 2 represents a summary of both clusters.

Cluster 0 (green) appears widely distributed but grouped into distinct tight pockets. This reflects strong internal consistency among academic independents suggesting subtypes within this group but overall clear identity. Cluster 1 (orange) also forms distinct regions, though more interspersed with other clusters, indicating socially influenced students share characteristics with both independent and privileged types. Cluster 2 (blue) generally, occupies a compact region, showing cohesion among privileged students. This supports their well-defined socio-economic and behavioral profiles.

In Figure 3, the t-SNE confirms that clusters are not random, and also confirms that, there is some inter-cluster overlap, especially between Clusters 1 and 2, consistent with their shared use of counseling, social activity, and future aspirations. Cluster 0 shows more isolation, aligning with their self-reliant, career-driven traits. This cluster group do not benefit from career-focused guidance.

4.2. Classification Algorithms

- *RQ2. Which classification algorithm has the highest performance metric for the future prediction of students' choice of programme of study?*

In response to RQ2**,** in determining the algorithm with the highest performance metrics, NB, LR, SVM, J48-DT and RF, a supervised machine learning algorithm were used to model the dataset as depicted in Table 3 and Figure 4–8, each algorithm perform excellently. The percentage split techniques were implemented to ascertain the performance of the classification algorithms. The SVM (RBF Kernel) algorithm performed the best with accuracy of 0.99%, recall value of 0.99% and also recorded the best F-Score value of 0.99% and, Precision value of 0.99%. and the least performing algorithm is Naïve Bayes with an accuracy, recall and F1 Score of 89% with better precision of 90%.

4.3. Student Choice Of Programme

- *RQ3. What are the most dominant features likely to influence a student choice of programme study?*

In response to RQ3 as shown in Figure 9**,** Programme of Study (Highest importance), this is the strongest predictor of cluster membership and also likely distinguishes between academically driven students and those selecting based on social or external motivations. It reinforces that students in different clusters are enrolled in different academic tracks with clear motivational differences. Counselling Service, High importance suggests a significant divide between students who receive focused-career guidance Cluster 1 and 2, where Cluster 0 do not receive focused-career guidance and counselling. This aligns with observed behavioral dependence on external support. Cultural beliefs influence plays a major role in shaping student decision-making, particularly within Clusters 1 and 2. Its presence confirms a cultural lens on academic choices for certain groups. Parental education (Mother and Father), Reflects how family educational background correlates with access, orientation, and social positioning. Cluster 2 (Privileged Socialites) heavily benefits from this. Residential status, boarding vs. day schooling reveals structural and environmental influences. Correlates with both autonomy and institutional exposure. Purpose of schooling is highly discriminative it separates those with academic intent from those attending for social engagement. In cluster 0, all students indicate they are in school for academic purposes, and they are not influenced by external factors such as gender, friends or cultural beliefs. In cluster 1, decisions are influenced by external factors, including friends, gender roles and cultural beliefs. Cluster 2, programmes choices are strongly influenced by cultural belief, gender-based and peer factors.

The feature selection mechanism is primarily implemented to remove weaker attributes and maintain more vital features to improve classification accuracy. From Table 4, this study considered the following most effective features: programme of study, counselling service, cultural belief, parents' education (mother and father), residential status, purpose of schooling,

Table 3. Classification Algorithms Performance.

Model	Accuracy	Precision	Recall	F1-Score
NB	0.89	0.90	0.89	0.89
LR	0.88	0.88	0.88	0.88
SVM-(RBF Kenel)	**0.99**	**0.99**	**0.99**	**0.99**
DT-(J48)	0.93	0.93	0.93	0.93
RF	0.94	0.95	0.94	0.94

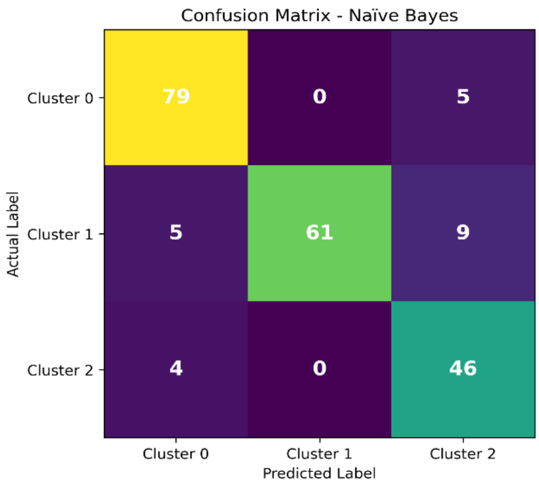

Figure 4. Confusion for Naïve Bayes.

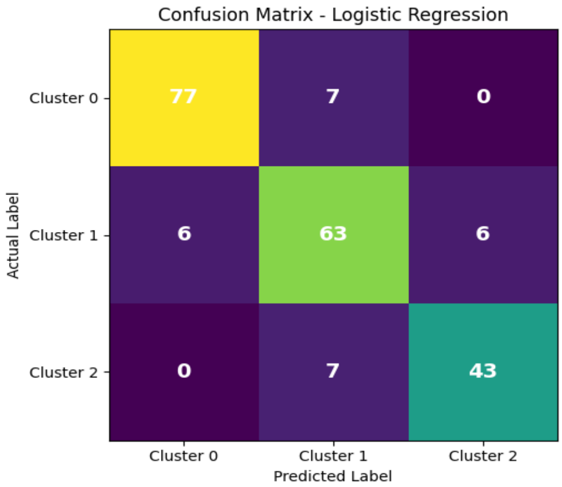

Figure 5. Confusion Matrix for Logistic Regression.

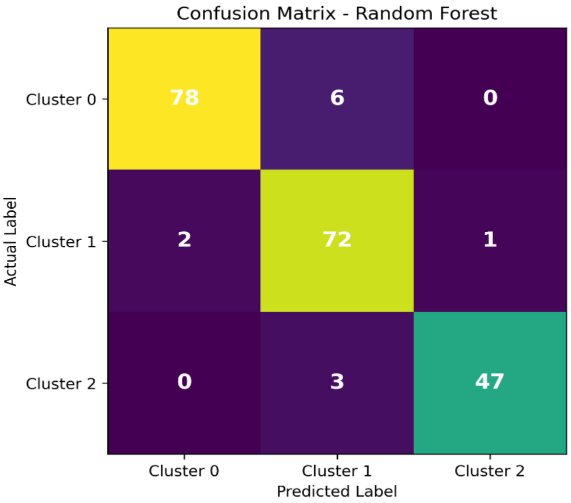

Figure 6. Confusion Matrix for Random Forest.

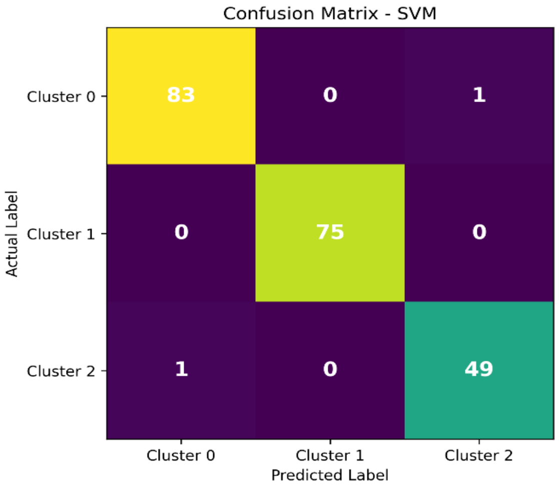

Figure 7. Confusion Matrix for Support Vector Machine.

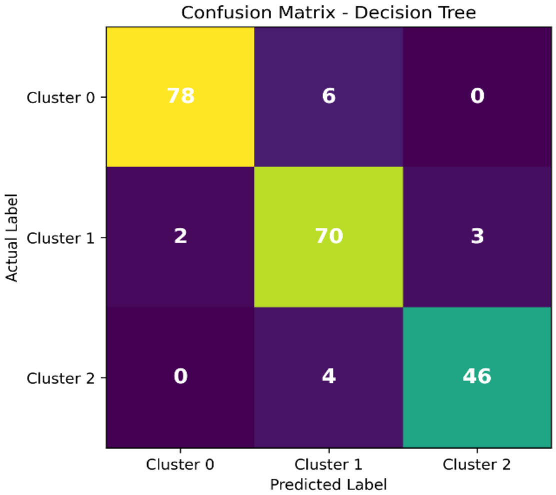

Figure 8. Confusion Matrix for Decision Tree.

with peers, students targeted institutions, gender-driven decision, parents work (mother and father), guardian work, and guardian educational level. The remaining features had minimal or no impact on the prediction process. Therefore, they were eliminated.

5. Discussion

The study applied the K-prototype algorithm to segment senior high school students into three distinct behavioural clusters based on mixed-type survey data. The majority of respondents were aged 16–20 years, reflecting the typical age group of final-year SHS students. Notably, 40.5% of participants reported not benefiting from structured guidance and counselling services. This finding highlights a critical gap in pre-tertiary academic

Table 4. Key Features Influencing Cluster Membership.

Feature	Description	Why It Matters	Cluster Impact
1. What is your programme of study at the Senior High School?	The SHS program each student is enrolled in (e.g., Home Economics, Business)	Reflects academic direction and future career orientation	Strongly differentiates academic-focused (Cluster 0) vs socially influenced (Clusters 1 & 2)
2. Counselling Service	Whether the student receives counselling at school	Indicates institutional guidance and external support availability	Cluster 1 & 2 receive more counselling; Cluster 0 mostly self-guided
3. Will your choice be based on your cultural beliefs?	Whether cultural values influence the student's academic choices	Highlights socio-cultural factors shaping decisions	Clusters 1 & 2 show higher cultural influence
4. Mother's educational level	Level of education attained by the student's mother	Reflects home academic environment and awareness	Higher in Cluster 2 (Privileged Socialites)
5. Residential status in Senior High School	Whether the student is a day student or boarder	Affects peer exposure, autonomy, and institutional influence	Cluster 0 = mostly day; Clusters 1 & 2 = mostly boarders
6. Indicate your purpose of being in school.	Student's self-reported reason for attending school	Reveals personal motivation (academic vs. social)	Cluster 0 = academic; Cluster 2 = social
7. Father's educational level	Education level of the student's father	Reinforces home academic orientation and support level	Higher in Cluster 2
8. How often do you go out with friends?	Frequency of social outings	Measures social involvement and peer alignment	Cluster 0 = low; Cluster 1 & 2 = high
9. Which tertiary institution do you intend to go?	Student's targeted institution type (e.g., university, college)	Tied to aspirations and program alignment	Similar across clusters but nuanced by background
10. Will your choice be based on your gender?	Whether gender roles influence choice	Gender-driven decision markers	Clusters 1 & 2 more affected
11. Can your choice be influenced by external factors?	If decisions are shaped by things like family, society, etc.	Measures susceptibility to influence	Cluster 0 = independent; Clusters 1 & 2 = more susceptible
12. Father's work?	Father's occupation	Socioeconomic proxy	Cluster 0 = informal/self-employed; Cluster 2 = employed
13. Mother's work?	Mother's occupation	Socioeconomic and gender role indicator	Similar to above, impacts outlook
14. Guardian work?	Guardian's occupation	Support network for decision-making	Varies with Cluster 2 having higher formal employment
15. Guardian educational level	Guardian's academic attainment	Academic environment at home	Highest in Cluster 2

support systems and aligns with prior studies emphasizing the importance of early academic counselling in shaping tertiary programme decisions [5].

The clustering results reveal meaningful differences in motivational drivers, socio-economic background, and decision-making influences. Cluster 0 ("Academic Independents") demonstrated strong internal motivation and career alignment, largely unaffected by peer, gender, or cultural pressures. In contrast, Clusters 1 and 2 exhibited greater susceptibility to external influences, including cultural beliefs, gender norms, peer dynamics, and family background. These findings are consistent with research highlighting the role of socio-cultural and peer influences in shaping educational choices [20], [21]. By applying mixed-type clustering at the senior high school level, this study extends existing literature that often focuses primarily on tertiary-level academic performance prediction.

Internal cluster validation metrics indicated modest separation among clusters (Silhouette Score = 0.077; Davies–Bouldin Index = 3.25), suggesting overlap in behavioural tendencies. This reflects the multidimensional and

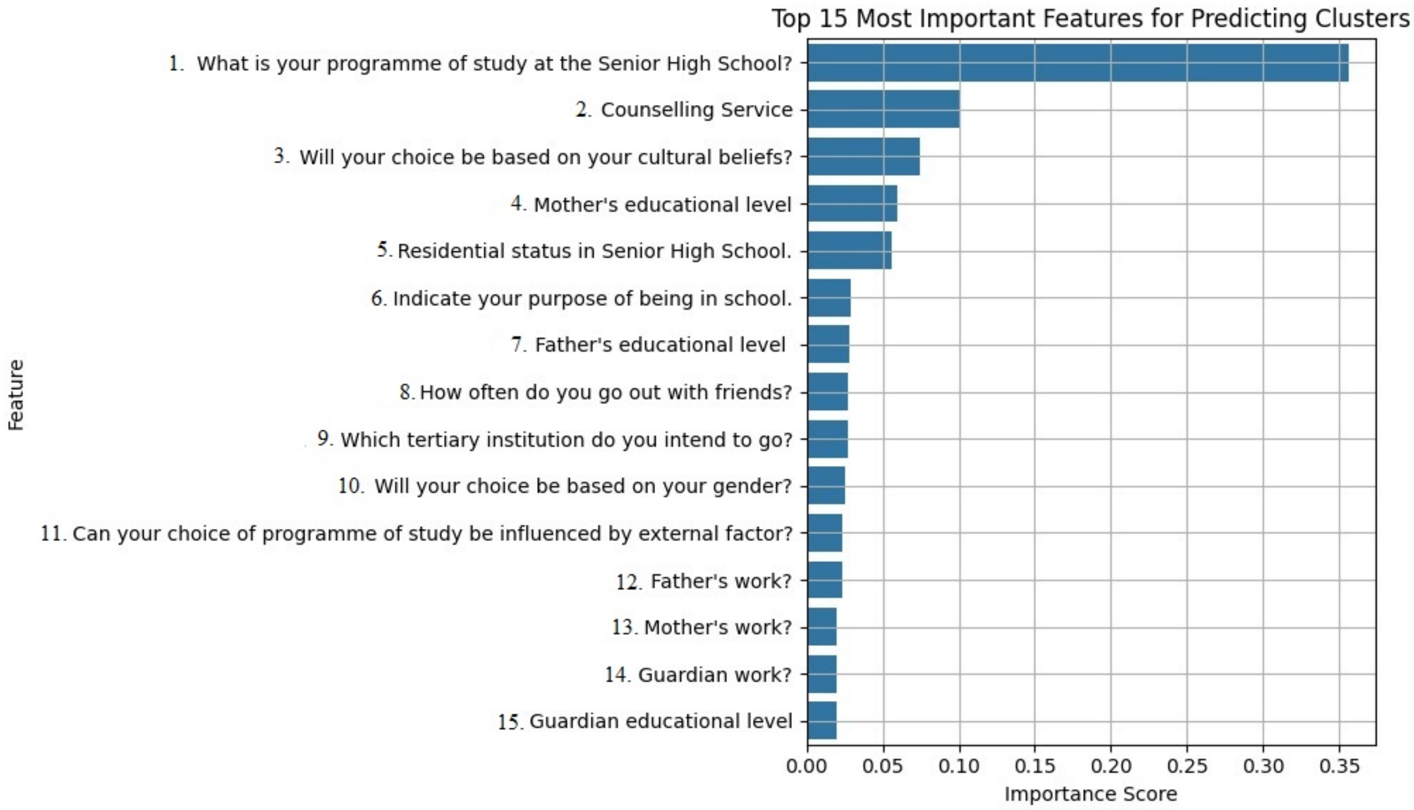

Figure 9. Key Features Driving Cluster Differentiation.

intersecting nature of student decision-making processes rather than sharply separable categories. Therefore, the cluster labels should be interpreted as behavioural tendencies rather than rigid classifications.

The classification results indicate that the Support Vector Machine (SVM) achieved the highest predictive performance among the evaluated models. This aligns with previous findings reporting strong SVM performance in educational data mining applications [22]. However, high predictive accuracy should be interpreted cautiously and in light of the dataset characteristics and validation procedures. Although the SVM achieved high predictive accuracy (99%), this result should be interpreted within the context of cluster-based classification. Since cluster labels were derived from the same dataset, high separability may naturally lead to elevated predictive performance. Therefore, the results demonstrate internal consistency of cluster structure rather than independent outcome prediction.

Overall, the findings underscore the need for strengthened career guidance frameworks within second-cycle institutions to support informed and independent programme selection decisions.

6. Limitations

Although the aim of the research was achieved, some limitations should be acknowledged. First, the study focused on final-year Business and Home Economics students from ten senior high schools in the Northern Region of Ghana; therefore, the findings may not fully generalize to students in other regions, school types, or academic programmes. Second, the study relied on self-reported questionnaire responses, which may be affected by response bias or social desirability effects. Third, the use of purposive/convenience sampling may limit the representativeness of the sample. Future studies should expand the dataset to include additional regions, school categories, and programme tracks, and evaluate more modelling approaches and validation strategies to strengthen generalizability. This study can serve as a blueprint for further research on data-driven guidance and counselling interventions. Future work should incorporate k-fold cross-validation and external dataset validation to further assess robustness and generalizability.

7. Conclusion

The study analyzed three research questions, and in response to RQ1, three clusters were formed when the K-prototype algorithms were implemented. In RQ2, the classification algorithm with the highest performance metric prediction of students' choice of programme of study is the Support Vector Machine. And final, in response to RQ3, the dominant attributes likely to influence a student's choice of programme of study at the tertiary level is programme of study, counselling service, cultural belief, parents' education, residential status, and purpose of schooling.

8. Declarations

8.1. Author Contributions

Yakubu Abubakar Mahami: Conceptualization, Methodology, Data Collection, Data Curation, Software Implementation, Formal Analysis, Visualization, Writing – Original Draft; **Ofori Elijah:** Supervision, Validation, Writing – Review & Editing, Critical Revision of Manuscript, Project Administration.

8.2. Institutional Review Board Statement

Not applicable.

8.3. Informed Consent Statement

Not applicable.

8.4. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

8.5. Acknowledgment

Not applicable.

8.6. Conflicts of Interest

The authors declare no conflicts of interest.

9. References

[1] K. M. Badau, "Factors influencing the choice of tertiary education institutions in Nigeria," *Journal of Resourcefulness and Distinction*, vol. 6, no. 1, pp. 1–13, 2013. https://www.globalacademicgroup.com/journals/resourcefulness/Factors%20Influencing%20the%20Choice%20of%20Tertiary%20Education.pdf.

[2] R. Faek, "International student mobility in Sub-Saharan Africa, Part 3: Trends in Ghana," World Education News & Reviews. https://wenr.wes.org/2024/10/international-student-mobility-in-sub-saharan-africa-part-3-trends-in-ghana.

[3] S. Bawakyillenuo, I. O. Akoto, C. Ahiadeke, E. B. D. Aryeetey, and E. K. Agbe, "Tertiary education and industrial development in Ghana," *Policy Brief*, vol. 33012, pp. 1–13, 2013. https://www.theigc.org/sites/default/files/2015/02/Bawakyillenuo-Et-Al-2013-Working-Paper.pdf.

[4] Y. Nieto, V. Gacia-Diaz, C. Montenegro, C. C. Gonzalez, and R. G. Crespo, "Usage of Machine Learning for Strategic Decision Making at Higher Educational Institutions," *IEEE Access*, vol. 7, pp. 75007–75017, 2019, https://doi.org/10.1109/ACCESS.2019.2919343.

[5] H. M. Ibrahim, A. N. Yousif, and R. D. Resen, "Determining the Relative Importance of Factors Affecting the Selection of High School Students for University Colleges Using Machine Learning Algorithms," *International Journal of Computer Science and Mobile Computing*, vol. 12, no. 3, pp. 40–48, Mar. 2023, https://doi.org/10.47760/ijcsmc.2023.v12i03.005.

[6] A. M. TURING, "I.—Computing Machinery and Intelligence," *Mind*, vol. LIX, no. 236, pp. 433–460, Oct. 1950, https://doi.org/10.1093/mind/LIX.236.433.

[7] A. Thorat and P. Mohite Rohan Hanbar, "MACHINE LEARNING AND ITS APPLICATIONS," *Journal of Emerging Technologies and Innovative Research*, vol. 10, no. 5, 2023. https://www.jetir.org/papers/JETIR2305274.pdf.

[8] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601–618, Nov. 2010, https://doi.org/10.1109/TSMCC.2010.2053532.

[9] R. S. Baker and P. S. Inventado, "Educational Data Mining and Learning Analytics," in *Learning Analytics*, New York, NY: Springer New York, 2014, pp. 61–75. https://doi.org/10.1007/978-1-4614-3305-7_4.

[10] H. Luo, "Prediction of Student Decision-Making Behaviour based on Machine Learning Algorithms," *Pakistan Journal of Life and Social Sciences (PJLSS)*, vol. 22, no. 2, 2024, https://doi.org/10.57239/PJLSS-2024-22.2.001188.

[11] I. E. Livieris, T. A. Mikropoulos, and P. Pintelas, "A decision support system for predicting students' performance," *Themes in Science and Technology Education*, vol. 9, no.1, pp. 43-57, 2016. https://www.learntechlib.org/p/174254/.

[12] A. Pandey and A. Jain, "Comparative Analysis of KNN Algorithm using Various Normalization Techniques," *International Journal of Computer Network and Information Security*, vol. 9, no. 11, pp. 36–42, Nov. 2017, https://doi.org/10.5815/ijcnis.2017.11.04.

[13] K. O. Adefemi, M. B. Mutanga, and V. Jugoo, "Hybrid Deep Learning Models for Predicting Student Academic Performance," *Mathematical and Computational Applications*, vol. 30, no. 3, p. 59, May 2025, https://doi.org/10.3390/mca30030059.

[14] J. Arruarte, M. Larrañaga, A. Arruarte, and J. A. Elorriaga, "Measuring the Quality of Test-based Exercises Based on the Performance of Students," *Int. J. Artif. Intell. Educ.*, vol. 31, no. 3, pp. 585–602, Sep. 2021, https://doi.org/10.1007/s40593-020-00208-0.

[15] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, pp. 29–39, 2000. https://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf.

[16] Z. Huang, "Clustering large data sets with mixed numeric and categorical values.," in *In Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 1997, pp. 21–34.

[17] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–291, 1967. https://books.google.co.id/books?id=IC4Ku_7dBFUC.

[18] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values," *Data Min. Knowl. Discov.*, vol. 2, no. 3, pp. 283–304, Sep. 1998, https://doi.org/10.1023/A:1009769707641.

[19] T. Devasia, Vinushree T P, and V. Hegde, "Prediction of students performance using Educational Data Mining," in *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, IEEE, Mar. 2016, pp. 91–95. https://doi.org/10.1109/SAPIENCE.2016.7684167.

[20] M. Shah, C. Sid Nair, and L. Bennett, "Factors influencing student choice to study at private higher education institutions," *Quality Assurance in Education*, vol. 21, no. 4, pp. 402–416, Sep. 2013, https://doi.org/10.1108/QAE-04-2012-0019.

[21] N. A. Sarkodie, A. Asare, and D. Asare, "Factors influencing students' choice of tertiary education," *Africa Development and Resources Research Institute Journal*, vol. 28, no. (11)(5), pp. 58–92, 2020. https://www.researchgate.net/publication/343318494.

[22] S. K. Yadav and S. Pal, "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification," *arXiv preprint*, arXiv:1203.3832, 2012. https://doi.org/10.48550/arXiv.1203.3832.