JOURNAL OF MACHINE LEARNING AND APPLICATIONS
July-Dec 2025, VOL. 1, ISSUE 2, pp.
https://doi.org/10.61577/jmla.2025.100006

RESEAPRO JOURNALS

ORIGINAL ARTICLE

OPEN ACCESS

Machine learning-based predictive model for optimal A-level combination selection using O-level academic performance in Rwandan high schools

Jean Joas Niyitegeka[1,4], Emmanuel Bugingo[1,2], Wilson Musoni[1], Célestin Kamanga Tshimanga[3], Leopord Hakizimana[1], and Uwineza Joseph[4]

[1]School of Computing and Information Technology, University of Kigali, KN 5 Road, Kigali, 2611, Kigali, Rwanda
[2]College of Business and Economics, University of Rwanda, KK737, Huye, 4285, South, Rwanda
[3]School of computer science, University of Kinshasa, H8J56PX, Kinshasa, Congo – Kinshasa, DRC
[4]Rwanda Polytechnic-Musanze College, Musanze, 226, North, Rwanda

ABSTRACT

The selection of A-level subject combinations in Rwandan high schools significantly influences students' academic and career trajectories. However, the existing manual selection process, influenced by subjective factors such as parental, peer pressure, and institutional preferences, often results in misalignments between students' abilities and chosen combinations. This study proposes the first machine learning-based predictive model tailored for Rwanda's education system to optimize A-level combination selection using students' O-level academic performance data. A dataset comprising 2,614 students from selected schools was examined using various classification algorithms, including Support Vector Classifier (SVC), Random Forest (RF), and Decision Tree (DT) models. To handle class imbalance, the Synthetic Minority Over-Sampling Technique (SMOTE) was employed. The models' performance was measured based on accuracy, precision, recall, and F1-score, with SVC achieving the highest performance, attaining 97.51% accuracy, 97.43% precision, 97.44% F1-score, and 97.49% recall. A k-fold cross-validation approach (k=10) was applied to validate model robustness (SVC Mean Accuracy (10-Fold CV): 97.61%, Random Forest Mean Accuracy (10-Fold CV): 84.94%, and Decision Tree Mean Accuracy (10-Fold CV): 67.15%). Statistical significance tests (ANOVA) confirmed the reliability of SVC's superior performance (ANOVA F-statistic: 260.8576, ANOVA p-value: 0.000000). SVC performance is significantly better than RF and DT ($p < 0.05$). The study demonstrates the potential of predictive analytics in enhancing student placement by providing objective, data-driven recommendations. Future improvements include incorporating deep learning models, integrating student feedback, and conducting longitudinal studies to assess the impact of machine- assisted selection over time.

KEYWORDS

Predictive modeling; Machine learning; A-level combination selection; O-level performance; Education analytics.

ARTICLE HISTORY

Received 19 June 2025;
Revised 22 July 2025;
Accepted 29 July 2025

Introduction

Transitioning from O-Level to A-Level education is crucial in Rwandan students' academic journey. A-Level Combination Selection determines future university pathways and career prospects, making it essential for students to choose combinations aligned with their abilities and aspirations. However, the current selection process lacks a structured, data-driven approach, resulting in cases where students struggle in combinations that do not match their academic strengths [1]. As education systems increasingly embrace data-driven methodologies, leveraging predictive analytics in student placement can enhance academic performance and career alignment.

The existing A-Level combination selection process in Rwandan high schools faces several challenges. Many students rely on external influences such as parental expectations, peer recommendations, and institutional policies rather than objectively assessing their abilities. Studies indicate that 40% of students in Rwanda struggle due to misaligned A- Level combinations, contributing to lower academic performance and higher dropout rates [2]. This often leads to students being assigned to subject combinations that do not align with their academic performance or career interests. The absence of an efficient selection framework increases the likelihood of misplacements, affecting long-term academic success and employability.

Traditional approaches involve teacher recommendations, aptitude tests, and manual career guidance sessions. In other contexts, predictive analytics have been successfully applied to educational decision-making, particularly in student performance prediction and career counseling [3].

Machine learning models have shown potential in optimizing A-Level Combination Selection by analyzing historical academic data and identifying patterns that can guide students toward suitable choices [4].

***Correspondence:** Dr. Emmanuel Bugingo, College of Business and Economics, University of Rwanda, KK737, Huye, 4285, South, Rwanda. e-mail: ebugingo@uok.ac.rw

Despite the advancements in predictive analytics, existing solutions face limitations. Many models are designed for higher education course selection rather than secondary school subject combinations [5].

Additionally, some models rely on limited datasets or lack adaptability to different education systems' unique grading and curriculum structures [6]. No standardized predictive model has been implemented in Rwanda to assist students in A-Level combination selection, leaving a gap in data-driven decision-making [4].

This study addresses these gaps by developing a machine learning-based predictive model tailored for A-Level combination selection in Rwandan high schools.

The key contributions of this research are:

- Developing a predictive model that utilizes O-level academic performance data recommend optimal A-level subject combinations.
- Evaluating multiple machine learning algorithms to determine the most effective model for student placement.
- Implementing a user-friendly system that provides data-driven recommendations will enhance the selection process for students, educators, and policymakers.
- Providing insights into the impact of predictive analytics on academic decision-making in Rwanda's education system.

This paper is organized as follows: Section 1 introduces the study's research problem, objectives, and significance. Section 2 reviews related literature, providing an overview of previous research on predictive analytics in education. Section 3 outlines the research methodology, covering data collection, preprocessing, and model evaluation. Section 4 presents the results, analyzing the performance of various machine learning models.

Finally, Section 5 summarizes key findings, offers recommendations, and suggests directions for future research.

Literature Review

Numerous studies have explored the use of machine learning (ML) in predicting student field specialization (majors). Previous research has applied various ML techniques and input features to analyze the relationship between student data and their chosen majors [7]. Alshaikh et al. [8], developed a recommendation system to assist KAU students in selecting suitable colleges based on their grades, college specializations, and enrollment criteria. This system was tested on a dataset of 960 KAU preparatory students from 2017. The accuracy of the k-nearest neighbor (KNN) algorithm was evaluated using two approaches. In the first, the dataset was split into 80% for training and 20% for testing, yielding an accuracy of 70.83%. The second method utilized k-fold cross-validation, dividing the dataset into K subsets and testing each.

Ezz and Elshenawy [9], introduced an adaptive recommendation system to predict suitable engineering departments for engineering preparatory year college students. They employed classification techniques such as support vector machine (SVM), KNN, linear regression (LR), quadratic discriminant analysis (QDA), and random forest (RF).

The system, which recommended one of seven engineering departments based on academic performance, achieved an average accuracy of 82.57%.

Reynaldo et al. [10], examined three machine learning (ML) algorithms: naïve Bayes (NB), RF, and sequential minimal optimization (SMO), using a dataset from three educational institutions in Bangladesh. Their study aimed to identify the most effective approach for selecting optimal educational majors, with RF demonstrating the highest performance, achieving 84.9% accuracy, 84.9% precision, 84.6% sensitivity, and an F-measure of 84.3%.

Meng and Fu [11], applied eight classification techniques: SVM, decision tree (DT), Gaussian naïve Bayes (GNB), RF, gradient boosting decision tree (GBDT), convolutional neural network (CNN), collaborative filtering (CF), and recurrent neural network (RNN) to recommend suitable college majors. RF achieved the best results, with an accuracy of 97.87% and an F-score of 96.60%.

Wei et al. [12], proposed an enhanced SVM-based model for predicting students' second primary selection. Their experimental findings showed that the improved approach performed best, attaining an accuracy of 87.36%, an AUC of 0.8735, a sensitivity of 85.37%, and a specificity of 89.33%.

Latifah et al. [13], utilized an artificial neural network (ANN) to predict student specializations using a dataset of 314 students. The data was sourced from the Gracias Integrated Academic Information System at Telkom University in 2016. Their model achieved an accuracy of 94.81%.

Thammarak's study [14], showed that the Backpropagation Neural Network exhibited high precision, reaching up to 78%, and demonstrated the best performance for the testing data. This research demonstrated that machine learning algorithms significantly improve the accuracy and efficiency of the training course recommendation.

Career guidance theories are crucial in student decision-making, helping them align their abilities and interests with suitable career paths.

The RIASEC framework developed by Holland (1997) categorizes students based on their personality traits and career preferences, offering a structured approach to subject selection [15].

Additionally, decision theory provides a mathematical basis for making optimal choices under uncertainty, which aligns well with predictive modeling [16]. Educational psychology concepts, such as self-efficacy and motivation, further influence subject selection and career aspirations [17].

Integrating these theoretical models with machine learning can enhance the accuracy and relevance of academic recommendations [18].

Empirical studies have demonstrated the effectiveness of machine learning in optimizing A-level subject selection across various educational systems. For instance, research in Western

https://doi.org/10.61577/jmla.2025.100006

and Asian contexts has shown that predictive models can analyze historical student performance data to recommend subject combinations that maximize academic success [19].

Some studies have also examined the impact of socioeconomic factors on subject choice, emphasizing the need for data-driven decision-support systems [20]. These findings underscore the practical benefits of machine learning in streamlining student guidance and reducing dropout rates [21].

Despite these advancements, limited research has focused on applying predictive analytics to subject selection in the Rwandan educational system. Most existing studies generalize findings from other regions or focus solely on higher education placement rather than secondary school subject combinations [22]. Given Rwanda's unique educational policies and curriculum structure, existing models may not apply directly without adaptation [23]. This gap highlights the need for localized research considering contextual factors such as national exam grading systems, school infrastructure, and career opportunities within the Rwandan job market.

Addressing this gap, the present study aimed to develop a machine learning- based predictive model tailored to Rwanda's secondary education system.

By leveraging student performance data from O-Level, the model recommends an optimal A-Level subject combination aligned with individual strengths and career aspirations. This approach seeks to improve academic outcomes and ensure students make informed choices, ultimately bridging the gap between education and future career success.

Methodology

Research design

The study adopts a quantitative research approach, utilizing machine learning algorithms for predictive modeling. A supervised classification framework was implemented to classify students into optimal A-level combinations based on historical O-level performance data.

Dataset

Data sources

The data set used in this study consists of 2,614 records of students from selected schools in two districts. It includes students 'scores in core subjects, such as Mathematics, Physics, Chemistry, Biology, History, Geography, Entrepreneurship, Kinyarwanda, and English. Additionally, it contains the A-Level combinations chosen by these students.

Data preprocessing

Several preprocessing techniques were applied to prepare the dataset to ensure optimal performance of the machine learning models. Data cleaning was performed to handle missing values and inconsistencies, such as misclassified A-Level subject combinations, ensuring data integrity. Normalization was implemented using the Weighted Sum Model (WSM) to maintain consistency in feature scaling, preventing any feature from dominating the model due to differences in magnitude. Additionally, to tackle the class imbalance among different A-level subject combinations, the Synthetic Minority Over-Sampling Technique (SMOTE) was utilized. This method created synthetic samples for minority classes, enhancing the model's capacity to generalize effectively across all subject combinations.

Feature selection

The model's performance was assessed using multiple metrics to ensure a thorough evaluation of its predictive capabilities. Accuracy measured the overall percentage of correctly classified A-Level subject combinations, serving as a general indicator of model effectiveness. Precision evaluated the proportion of correctly identified positive cases among all predicted positives, helping to minimize false positive classifications. Recall was used to determine the model's ability to identify all relevant instances, making it particularly useful for assessing the classification of underrepresented subject combinations. The F1- Score, which represents their harmonic mean, was computed to balance precision and recall. Additionally, the AUC-ROC (Area Under the Receiver Operating Characteristic) Curve was employed to analyze the model's classification performance across various threshold values, providing insight into its ability to differentiate between different subject combinations. These metrics collectively ensured a comprehensive assessment of the model's effectiveness.

Evaluation metrics

The model's performance was assessed using multiple metrics to ensure a thorough evaluation of its predictive capabilities. Accuracy measured the overall percentage of correctly classified A-Level subject combinations, serving as a general indicator of model effectiveness. Precision evaluated the proportion of correctly identified positive cases among all predicted positives, helping to minimize false positive classifications. Recall was used to determine the model's ability to identify all relevant instances, making it particularly useful for assessing the classification of underrepresented subject combinations. The F1- Score, which represents their harmonic mean, was computed to balance precision and recall. Additionally, the AUC-ROC (Area Under the Receiver Operating Characteristic) Curve was employed to analyze the model's classification performance across various threshold values, providing insight into its ability to differentiate between different subject combinations. These metrics collectively ensured a comprehensive assessment of the model's effectiveness.

Proposed method

Algorithm selection

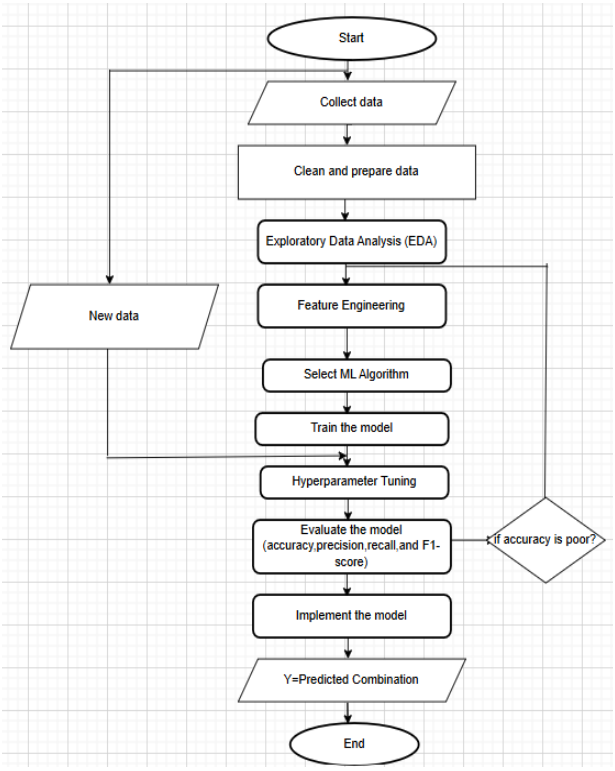
The study explored multiple machine learning algorithms, including:

1. Decision Tree Classifier – Simple but prone to overfitting.
2. Random Forest Classifier – Improved generalization through ensemble learning.
3. Support Vector Classifier (SVC) – Best overall performance, effectively handling multi-class classification

Flowchart

A structured workflow for the predictive model consists of the following steps:

https://doi.org/10.61577/jmla.2025.100006



Equations

Decision tree - Gini impurity formula

The Decision Tree algorithm was applied to classify students into the A-level combinations (PCM, MCB, PCB, MEG, and HGL) based on their O-level subject marks. The model recursively splits the dataset into smaller subsets using Gini impurity or Information Gain to determine the best feature to split at each step.

Example

If Mathematics > 75% and Physics > 70%, the model may predict PCM.

If Biology > 80%, Chemistry > 75%, and Mathematics < 70%, the model may predict MCB.

The Gini Impurity formula was used to determine the best split.

$$\text{Gini}(t) = 1 - \sum_{i=1}^{C} p_i^2 \quad (1)$$

Where:

- Gini(t) represents the level of impurity within node t.
- pi denotes the likelihood of class i occurring in node t.
- C refers to the total number of classes.

The information Gain formula was used to measure the reduction in entropy.

$$IG(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} Entropy(S_i) \quad (1)$$

Where:

- S is the dataset.
- A is the feature used for splitting.
- Si are the subsets after the split.
- |S| and |Si| are the sizes of the dataset and subsets

Random forest - prediction formula

Random Forest enhanced the performance of the Decision Tree by generating multiple trees (T trees) and averaging their predictions. Each tree was trained on a different random subset of the dataset (bootstrapping), and the final output was determined by majority voting (for classification).

$$y = \frac{1}{T} \sum_{t=1}^{T} h_t(x) \quad (3)$$

Where:

$t=1$ t

- Ht (x) represents the prediction made by the t-th decision tree.
- T denotes the total number of decision trees.

Example:

Suppose a student's O-Level marks are:

- Mathematics = 80%
- Chemistry = 78%
- Physics = 65%

Some trees might classify the students as PCM, while others might classify them as MCB. The ultimate decision is determined through a majority voting mechanism. This method helps improve the reliability of predictions by considering the outputs of multiple models rather than relying on a single one. To minimize overfitting, multiple decision trees are trained on different subsets of the dataset instead of depending on a single Decision Tree. By averaging their predictions, the model achieves better accuracy and generalization. This ensemble technique enhances overall model performance on both training data and previously unseen test data. As a result, it produces more dependable predictions for selecting optimal A-level subject combinations.

SVC - Decision Boundary Equation

$$\mathcal{W}\, x\, X + b = 0 \quad (4)$$

Where:

- w represents the weight vector.
- X denotes the feature vector.
- b is the bias term.

Support Vector Classification (SVC) was employed to determine the most effective decision boundary for categorizing students into various A-level subject combinations.

This was achieved by identifying the optimal hyperplane that maximizes the margin between student groups. By doing so, SVC enhances classification accuracy and ensures a more apparent distinction between the subject choices. This approach ensures that students with similar academic performance are classified into the most suitable combination. Additionally, a non-linear kernel (RBF) was applied to capture and handle complex relationships between subject marks and combination selection, improving the model's ability to classify students accurately.

Example

- If a student has Physics = 85%, Chemistry = 80%, and Mathematics = 90%, the SVC finds a decision boundary that places them in PCM.

https://doi.org/10.61577/jmla.2025.100006

- If a student has Biology = 88%, Chemistry = 85%, and Mathematics = 65%, they may be classified into MCB.

SMOTE-Formula

$$Xnew = Xi + \lambda\, x(Xj - Xi) \quad (5)$$

Where:

- Xi =A randomly selected minority class instance.
- Another randomly selected the nearest neighbor of Xi.
- s=A random number between 0 and 1.

In my dataset, some A-level combinations, such as PCM and MCB, were more frequent than others, like HGL, leading to an imbalance in class distribution. To address this, SMOTE was applied to generate synthetic samples for underrepresented combinations, ensuring similar representation across all classes. This approach helped prevent the model from favoring majority classes, ensuring a more balanced classification.

As a result, the model improved its ability to assign students to less common A- Level subject combinations accurately, enhancing overall fairness and predictive performance.

Example

If only 50 students chose HGL while 500 chose PCM, SMOTE generated additional synthetic HGL students by interpolating feature values.

SMOTE was helpful because it enhanced the model's ability to classify minority combinations correctly by generating synthetic data points, ensuring a balanced dataset. This prevents the classifier from being biased toward majority classes, allowing it to make fair and accurate predictions across all A-Level combinations.

Weighted sum model-formula

$$S_i = \sum_{j=1}^{n} W_j \; x \; X_{ij} \quad (6)$$

Where:
$\mathcal{W}j$
$x\, \mathcal{X}ij$ (6)

- Si=Total score for alternative i
- $\mathcal{W}$j=Weight assigned to criterion j
- $\mathcal{X}$ij=Value of criterion j for alternative i

The Weighted Sum Model (WSM) was used to account for the varying importance of different subjects in each A-level combination by assigning appropriate weights to them. This process helped normalize the scores, ensuring that all subjects contributed to the model's decision-making. By doing so, WSM improved the accuracy and relevance of the predictions, aligning students with the most suitable combinations based on their academic strengths.

For example, if a student is evaluated for the PCM (Physics-Chemistry- Mathematics) combination, the Weighted Sum Model assigns higher importance to Mathematics and Physics than other subjects. Suppose the assigned weights are Mathematics (0.4), Physics (0.3), Chemistry (0.2), and English (0.1). If the student's scores are Mathematics = 85, Physics = 80, Chemistry = 75, and English = 70, the weighted sum score for PCM is calculated as: SPCM= (0.4×85)+(0.3×80)+(0.2×75)+(0.1×70)=34+24+15+7=80.

Recall-formula

$$\text{Recall} = \frac{T\,P}{T\,P + F\,P} \quad (7)$$

Where:

TP: Count of correctly predicted positive cases.

FN: Count of actual positive cases mistakenly classified as harmful.

Precision-formula

$$\text{Precission} = \frac{T\,P}{T\,P + F\,P} \quad (8)$$

Where

- TP: is the count of correctly classified positive instances. It represents cases where the model correctly predicts a student belongs to a specific A-Level combination when they do.
- FP: is the count of incorrectly classified positive instances. It represents cases where the model predicts a student belongs to a combination, but they do not.

F1-score-formula

$$F = 2x \frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

Example

To extract the values for MCB from the confusion matrix: The True Positives (TP) for MCB represent the number of correctly predicted class instances. In this case, the model correctly classified 104 instances as MCB, corresponding to the value found at the intersection of the "MCB" row and column in the confusion matrix.

The False Positives (FP) indicates the number of times the model incorrectly predicted MCB when it should have classified a different class. There were 3 cases where HGL was misclassified as MCB, 0 cases for MEG, 1 for PCB, and 2 for PCM, resulting in 6 false positives.

On the other hand, the False Negatives (FN) refer to actual MCB instances that were wrongly classified as other subjects.

Specifically, one instance of MCB was misclassified as MEG, and another was classified as PCM, leading to 2 false negatives.

- $Recall = \frac{104}{104+2} = \frac{104}{106} = 0.98$
- $Precision = \frac{104}{104+6} = \frac{104}{110} = 0.94$
- $F1 = 2x \frac{0.94\ x\ 0.98}{0.94+0.98} = 2x \frac{0.92}{1.92} = 0.96$

Results & Discussion

Experimental setup

Hardware specifications

The experiments were conducted on a high-performance computing system to ensure efficient model training and evaluation. This system was equipped with an Intel Core i7 10th Gen processor featuring a 2.6 GHz clock speed and eight cores,

https://doi.org/10.61577/jmla.2025.100006

delivering strong computational power for handling complex tasks. It was equipped with 16 GB DDR4 RAM, allowing smooth execution of data processing and machine learning tasks. The storage was handled by a 512 GB SSD, ensuring fast read/write speeds for handling large datasets.

To accelerate computations, especially for deep learning tasks or large-scale data processing, an NVIDIA GeForce GTX 1660 Ti GPU with 6GB of VRAM was used. This enhanced the system's ability to handle complex operations efficiently, improving overall processing speed and performance and significantly improving the efficiency of model training. The system operated on Windows 10 64-bit, providing a stable and compatible environment for machine learning development.

Software and programming tools

Python 3.9 served as the primary programming language for this research, facilitating data preprocessing, model training, and evaluation. Its versatility and extensive libraries contributed to efficient implementation and analysis throughout the study. The development and analysis were conducted using Jupyter notebook, which provided an interactive environment for coding and visualization. Scikit-Learn, a widely used machine learning library, implemented classification models, ensuring efficient development.

For data manipulation and numerical computations, Pandas and NumPy played a crucial role in processing and structuring the dataset. Matplotlib and Seaborn were used to effectively present findings for data visualization, generating insightful graphs and plots that enhanced interpretability.

To address class imbalance in the dataset, SMOTE (Synthetic Minority Over- sampling Technique) was applied, ensuring that minority class instances were adequately represented to improve model performance. Furthermore, a Flask- based web interface was developed for deploying the predictive model, integrating HTML, CSS, and JavaScript to create a user-friendly and interactive experience for students and academic advisors.

Results and Visualizations

Performance metrics

The model's performance was evaluated using multiple metrics, including Accuracy, Precision, Recall, F1-score, and AUC-ROC.

Table 1. Comparison of Model Performance.

Model	Training Accuracy (%)	Testing Accuracy (%)	Precision	Recall	F1-Score
Decision Tree	66.28	57.93	0.58	0.59	0.58
Random Forest	100	83.17	0.83	0.84	0.83
Support Vector Classifier	99.55	97.51	0.97	0.97	0.97

Table 1 showcases the performance metrics of three machine learning models: Decision Tree, Random Forest, and Support Vector Classifier (SVC) in predicting the most suitable A-level subject combinations.

The Decision Tree model demonstrates relatively lower performance, achieving a training accuracy of 66.28% and a testing accuracy of 57.93%. This discrepancy suggests a moderate overfitting, where the model performs well on the training data but struggles to generalize effectively to unseen data. Its precision of 0.58, recall of 0.59, and F1-score of 0.58 suggest it struggles with generalizing predictions.

On the other hand, the Random Forest model achieves 100% training accuracy and 83.17% testing accuracy, suggesting strong learning capabilities but potential overfitting. It outperforms the Decision Tree model, achieving a precision of 0.83, a recall of 0.84, and an F1-score of 0.83. The Support Vector Classifier (SVC) model outperforms both, achieving outstanding results with a training accuracy of 99.55% and a testing accuracy of 97.51%, indicating minimal overfitting. Attaining a precision, recall, and F1-score of 0.97, the SVC model is the most reliable in accurately classifying student subject combinations, making it the best choice for deployment in real-world applications.

Confusion matrix

The confusion matrix below shows how different models classified the five A- level subject combinations: HGL, MCB, MEG, PCB, and PCM.

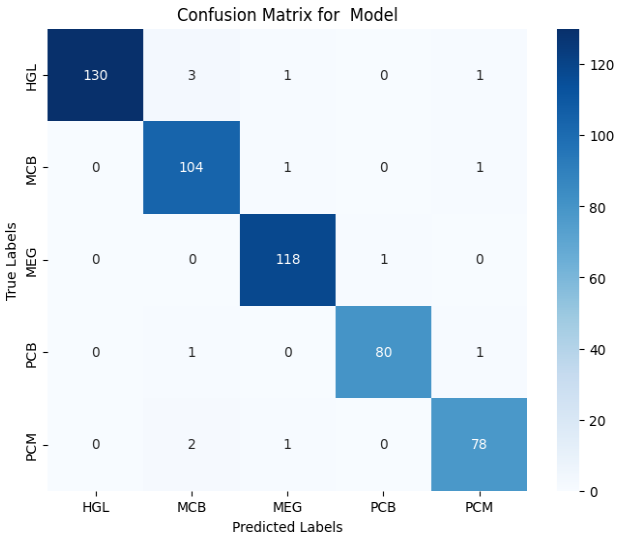


Figure 1. Confusion Matrix for best model (SVC).

The confusion matrix comprehensively assesses the model's classification performance in predicting A-Level subject combinations: HGL, MCB, MEG, PCB, and PCM. The diagonal elements denote correctly classified instances, while the off-diagonal elements highlight misclassifications.

The model demonstrates high accuracy, particularly for MEG (118 correct predictions), HGL (130 correct predictions), and MCB (104 correct predictions), indicating strong generalization. PCB and PCM also perform strongly, with 80 and 78 correct classifications, respectively. However, there are a few misclassifications: MCB was misclassified as PCM twice, PCB was misclassified as MCB once, and a small number of instances were incorrectly assigned to other categories.

https://doi.org/10.61577/jmla.2025.100006

The confusion matrix suggests that the model effectively predicts subject combinations with minimal misclassifications, reinforcing its high precision and recall values.

ROC curve

The ROC Curve below shows the performance of the SVC model across different class thresholds.

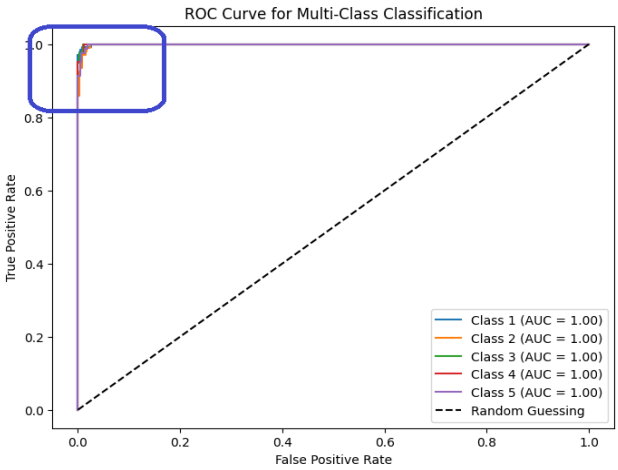


Figure 2. ROC Curve showing the performance of the SVC model across different class thresholds.

Figure 2 shows that each class has an Area Under the Curve (AUC) score of 1.00, indicating that the model has achieved perfect classification for all five classes. The curve is tightly clustered in the upper left corner, showing a very high actual positive rate with a minimal false positive rate. The dashed black line represents random guessing, but the model's performance is significantly superior, as indicated by the curves being far above this baseline.

Discussion

The findings of this study highlight the significant role predictive modeling can play in improving A-Level combination selection. The high accuracy achieved by SVC indicates that machine learning models can effectively identify patterns in student performance data, leading to more informed decision-making.

One key challenge observed is the disparity in subject availability, which prevents students from pursuing their most suitable combinations even when the model provides accurate recommendations. This suggests that predictive analytics should be integrated alongside policy changes to ensure equal access to all subject combinations across schools.

Moreover, the study highlights inconsistencies in grading scales, which can impact model performance. Standardizing grading criteria could improve prediction accuracy and ensure fairer recommendations. Additionally, while the model provides strong predictive insights, incorporating additional student preferences, such as career aspirations and personal interests, could enhance the recommendation system.

Conclusion & Recommendations

The study successfully developed a machine learning-based predictive model to assist students in selecting optimal A-level subject combinations based on their O-level performance. Among the three models tested, the Support Vector Classifier (SVC) proved to be the most effective, attaining the highest accuracy (97.51%) along with excellent precision, recall, and F1-score values (0.97 each). The Random Forest model also performed well, with an accuracy of 83.17%, while the Decision Tree model exhibited lower performance (57.93% accuracy), indicating its limited suitability for this task. The confusion matrix results further confirmed that the SVC model had minimal misclassifications, reinforcing its reliability for real-world implementation. A user-friendly web interface has also been designed, making the model accessible to students, teachers, and academic advisors.

Despite these promising results, further improvements can be made to improve the model's effectiveness and applicability. Additional student-related factors like career aspirations and teacher recommendations could refine prediction accuracy. Additionally, fine-tuning model hyperparameters may optimize performance and reduce the risk of misclassifications. To enhance interpretability, techniques such as SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-Agnostic Explanations) can offer valuable insights into the model's decision-making process. Expanding the dataset to include students from diverse schools and regions would enhance the model's generalization ability, ensuring it performs well across different academic environments. Lastly, exploring deep learning techniques could improve classification performance, particularly for complex subject combinations. By implementing these improvements, the predictive model can serve as a more robust decision-making tool for students, ultimately leading to better academic and career outcomes.

Disclosure Statement

No potential conflict of interest was reported by the authors.

References

1. Ndihokubwayo K, Ukobizaba F, Byusa E, Rukundo JC. Issues in subject combinations choice at advanced level secondary schools in Rwanda. Problems of Education in the 21st Century. 2022;80(2): 339. https://doi.org/10.33225/pec/22.80.339
2. Education_Statistical_Year_Book_2021-22. 2022. https://www.mineduc.gov.rw/index.php?eID=dumpFile&t=f&f=70247&token=97659d85ed14644f7fa6ccbd3970c418a780547d
3. Romero C, Ventura S. Educational data mining and learning analytics: An updated survey. Wiley Interdiscip Rev: Data Min Knowl Discov. 2020;10(3):1355. https://doi.org/10.1002/widm.1355
4. Kotsiantis SB, Zaharakis ID, Pintelas PE. Machine learning: a review of classification and combining techniques. Artificial Intelligence Review. 2006;26:159-190. https://doi.org/10.1007/s10462-007-9052-3
5. Kamal N, Sarker F, Rahman A, Hossain S, Mamun KA. Recommender system in academic choices of higher education: a systematic review. IEEE Access. 2024;23(12):35475-35501. https://doi.org/10.1109/ACCESS.2024.3368058
6. Taylan O, Karagözoğlu B. An adaptive neuro-fuzzy model for prediction of student's academic performance. Comput Ind Eng. 2009;57(3):732-741. https://doi.org/10.1016/j.cie.2009.01.019
7. Alsayed AO, Rahim MS, AlBidewi I, Hussain M, Jabeen SH, Alromema N, et al. Selection of the right undergraduate major by students using supervised learning techniques. Appl Sci. 2021; 11(22):10639. https://doi.org/10.3390/app112210639
8. Alshaikh K, Bahurmuz N, Torabah O, Alzahrani S, Alshingiti Z, Meccawy M. Using recommender systems for matching students

https://doi.org/10.61577/jmla.2025.100006

with suitable specialization: An exploratory study at King Abdulaziz University. Int j emerg technol learn. 2021;16(3):316-324. http://dx.doi.org/10.3991/ijet.v16i03.17829
9. Ezz M, Elshenawy A. Adaptive recommendation system using machine learning algorithms for predicting student's best academic program. Educ Inf Technol. 2020;25(4):2733-2746. https://doi.org/10.1007/s10639-019-10049-7
10. Reynaldo SJ, Kawet CR, Manoppo R, Tumimomor F. Decision support systems major selection vocational high school in using fuzzy logic android-based. InInt Conf Electr Eng Inform, Its Educ 2015.
11. Meng Y, Fu M. CMRS: Towards intelligent recommendation for choosing college majors. In Proceedings of the 4th International Conference on Advances in Image Processing 2020. 2020: 152-157. https://doi.org/10.1145/3441250.3441272
12. Wei Y, Ni N, Liu D, Chen H, Wang M, Li Q, et al. An improved grey wolf optimization strategy enhanced SVM and its application in predicting the second major. Math Probl Eng. 2017;2017(1): 9316713. https://doi.org/10.1155/2017/9316713
13. Latifah SN, Andreswari R, Hasibuan MA. Prediction analysis of Student specialization suitability using artificial neural network algorithm. In2019 International Conference on Sustainable Engineering and Creative Computing (ICSECC) 2019: 355-359. https://doi.org/10.1109/ICSECC.2019.8907173
14. Thammarak K, Kesornsit W, Sirisathitkul Y. Predictive Model for Academic Training Course Recommendations Based on Machine Learning Algorithms. Int J Mod Educ Comput Sci. 2024;16(3): 17-26. https://doi.org/10.5815/ijmecs.2024.03.02
15. Bullock-Yowell E, Reardon RC. Holland's RIASEC Hexagon: A Paradigm for Life and Work Decisions. 2024. https://doi.org/10.33009/fsop_bullock-yowell0524
16. Zainudin ZN, Rong LW, Nor AM, Yusop YM, Othman WN. The relationship of holland theory in career decision making: A systematic review of literature. J Crit Rev. 2020;7(9):884-892. http://dx.doi.org/10.31838/jcr.07.09.165
17. Schunk DH, DiBenedetto MK. Self-efficacy and human motivation. Adv Motiv Sci 202. 153-179. https://doi.org/10.1016/bs.adms.2020.10.001
18. Pliakos K, Joo SH, Park JY, Cornillie F, Vens C, Van den Noortgate W. Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems. Comp Edu. 2019;137:91-103. https://doi.org/10.1016/j.compedu.2019.04.009
19. Agyemang EF, Mensah JA, Ampomah OA, Agyekum L, Akuoko-Frimpong J, Quansah A, et al. Predicting Students' Academic Performance Via Machine Learning Algorithms: An Empirical Review and Practical Application. "Comp Eng Intell Syst. 2024;15(1):86. http://doi.org/10.7176/CEIS/15-1-09
20. Joshi BM, Khatiwada SP, Pokhrel RK. Influence of socioeconomic factors on access to digital resources for education. Rupantaran: Multidiscip J. 2024;8(01):17-33. https://doi.org/10.3126/rupantaran.v8i01.65197
21. Hassan MA, Muse AH, Nadarajah S. Predicting student dropout rates using supervised machine learning: Insights from the 2022 National Education Accessibility Survey in Somaliland. Appl Sci. 2024;14(17):7593. https://doi.org/10.3390/app14177593
22. Zoungrana O, Messan A, Nshimiyimana P, Pirotte G. The paradox around the social representations of compressed earth block building material in Burkina Faso: the material for the poor or the luxury material? Open J Soc Sci. 2021;9(1)50-65. https://doi.org/10.4236/jss.2021.91004
23. Bizimana B. Building education for sustainability: A philosophy of education analysis of competence-based curriculum implementation in Rwanda. Rwandan J educ. 2024;7(3):3-21.

https://doi.org/10.61577/jmla.2025.100006