

Model for career pathway prediction for Kenya's competency-based education curriculum

Innocent Wamalwa Barasa¹
Collins Odoyo²
Raphael Angulu³

¹ibwamalwa254@gmail.com (+254 741 209 149)

²codoyo@mmust.ac.ke (+254 725 961 657)

³rangulu@mmust.ac.ke (+254 729 260 641)

^{1,2,3}Masinde Muliro University of Science and Technology, Kenya

<https://doi.org/10.51867/ajernet.7.3.139>

ABSTRACT

Kenya's Competency-Based Education (CBE) curriculum puts an emphasis on learning based on project work, co-curricular activities involvement and continuous evaluation of the learner to influence learners into the most suitable career pathway. However, the choice of career pathway of learners in junior secondary schools in Kenya greatly relies on the manual judgment of the teacher and parent, which may be subjective, lack consistency and inconsiderate of the learning context. Therefore, there needs a consistent data-driven tool that combines both academic and non-academic pointers, based on actual school data, to advise on career pathway selection and that utilizes Machine Learning to inform the choice of the learner, teacher and parent. The main objective of the paper was to develop a Machine Learning model that can help select the most suitable career pathway for the learner under the Competency-Based Education curriculum. In specific the study aimed to: review existing literature to recognize the factors considered in the determination of the career pathway choice of the student; to analyze data to determine the most consequential factors that influence the career pathway choice of the student; develop a model to predict the career pathway choice using Random Forest algorithm; and to assess and validate the model's ability to predict the individual's career pathway. The study used a quantitative experimental research design, with data of 1,083 students gathered from databases in junior secondary schools in the county of Trans-Nzoia, involving both public and private schools. The data included the academic scores, project marks, extra-curricular participation, leadership involvement, gender and the school type, functioning as the mediating variable. The data was cleaned, standardized, split into 80:20 training-testing set. Random Forest feature-importance recognized 7 most influential features as Integrated Science (29.89%), Agriculture (16.60%), Pre-Technical Studies (10.47%), Mathematics (8.63%), Sports (7.38%), Music and Drama (5.12%) and the type of school (4.84%). The features together accounted for 82.92% of all feature importance. The dataset was trained using three Machine Learning algorithms. Random Forest was the best performing model with an accuracy of 89%, compared to Decision Trees with 81.46% and 78% for Neural Networks. However, Decision Trees and Neural Networks had a better stability on ten-fold cross-validation, where Random Forest had an accuracy score of 86%, 82% for Decision Tree and 78.62% for Neural Network. The Random Forest model recorded AUC score of 0.90, 0.97 and 0.98 for Arts, Social Science and STEM in that order, which illustrates very good discriminative ability. The study in conclusion states that academic performance, extra-curricular participation and school environment are collectively useful in forecasting career pathway choice of students. The final result is a dependable evidence-driven decision-support system – the Random Forest career pathway prediction model, which enhances objectivity and equality in selection of a clear pathway for the student under the CBE curriculum. The tool serves as a complimentary system for career pathway selection and not as a replacement of professional career guidance. The paper suggests future studies to be done in other counties and inclusion of the interests of the student, parents and other primary statistics considered.

Keywords: Competency-Based Education, Career Pathway Selection, Machine Learning, Random Forest

I. INTRODUCTION

Education is globally accepted as the foundation of social, economic and personal development. In Kenya, great efforts have been made to increase access to education through Free Primary Education and Free Day Secondary Education policies. These efforts have resulted in the improvement in rates of enrolment and a significant increase in educational breakthroughs across the nation (Ministry of Education, 2019). One of the most outstanding reforms in the education sector is the entrance of the Competency-Based Education (CBE), which comes in replacement of the 8-4-4 system. CBE stresses on holistic learning, continuous assessment and competency growth, contrary to the 8-4-4 system, which is criticized for being overly exam-oriented (Ministry of Education, 2019).

One of the most conspicuous innovations of the CBE curriculum is introducing the career pathways to learners in an early age. These pathways which are categorized broadly as Social Sciences, Arts and Sports, and STEM (Science, Technology, Engineering and Mathematics) – are all intended to nurture the talent, interests and

abilities of learners, from an early stage (Kenya Institute of Curriculum Development, 2017). By the identification of the strength of the learners and alignment to the best suited pathways, CBE aims to channel out not only academically performing students but also persons armed with practical and creative skills to prosper in diverse professional and entrepreneurial settings.

Career choice is a complex phenomenon. Social Career Theory (SCCT) suggests that interest in career, choices and performance result from relations individual credence, expected results, goals and the nature of the environment (Lent et al., 1994). Holland's theory additionally elaborates career selection in terms of the interaction between individual preferences and the environment which is represented through Realistic, Investigative, Artistic, Social, Enterprising and Conventional settings (Holland, 1997). These theories imply that academic achievement (subject scores in Mathematics, English, Kiswahili, Integrated Science, Pre-Technical Studies, Agriculture, Religious Education & Social Studies) only, cannot be the reason for a certain pathway selection. The involvement of the learner in extra-curricular activities and the learning environment (type of school) also influence the career pathway choice of the student.

Educational Data Mining (EDM) offers a realistic technique to analyze such complex data. EDM employs calculative and statistical methods to uncover patterns in educational statistics and to boost on forecasting and decision-making (Romero & Ventura, 2013, 2020). In school environments, forecasting models can be employed to complement career pathway guidance by discovering patterns that may be hard to comprehend manually and with huge student numbers. Machine Learning techniques are particularly useful in solving this problem because pathway choice is a multiclass forecasting problem with school data frequently being both categorical as well as numerical characters. Random Forest is an ensemble technique that uses several decision trees combined and can understand both discrete as well as continuous data patterns (Breiman, 2001). Newer studies on career forecasting show increased use of Machine Learning by combining academic scores, project marks and co-curricular information for career pathway prediction (Berlikozha et al., 2025; Urlamma et al., 2025).

In Kenya, EDM has shown the possibility of using school records to make predictive models. Mgala (2016) used academic records to forecast the outcome of performance in Kenyan schools. Such studies show the importance of using school records to model the future. However, in the context of Kenya's CBE curriculum, there exists a contextual gap. Few studies have been done, pertaining Machine Learning in CBE and a gap exist due to lack of a career guidance model, tailored under CBE environment. This paper hence uses stored student records to model career pathway tool under the CBE curriculum in the county of Trans-Nzoia.

1.1 Statement of the Problem

The Competency-Based Education (CBE) in Kenya stipulates that students are guided into career pathways depending on their aptitudes, interests and competencies. However, reports by the Ministry of Education (2019) and Oduor (2022) show glaring challenges in the implementation of CBE, including selection of the career pathway. Career pathway selection is heavily reliant on the manual teacher and parental judgement. Parents are not sufficiently informed and hence have the tendency to over-emphasize on academic scores while assuming other key considerations like project-based evaluations, co-curricular involvement and leadership duties. On the other hand, teachers tend to be subjective, biased or inconsistent. Moreover, contextual disparities between public and private schools are easily overlooked. It also requires a lot of time for teachers to manually analyze all the factors accurately and hence unsuitable in large number of students. Currently available systems do not incorporate extracurricular activities, school context and mainly used in grading the students' performance and not recommending the pathways using Machine Learning. This results to many students risking getting mis-profiled to career pathways that do not reflect their complete potential.

There therefore necessitated a need to develop a Machine Learning tool that uses the past student records to provide just, evidence-driven and dependable career pathway recommendation, which lines with the goals of the CBE curriculum. This paper hence attempts to bridge the gap by the development of an interpretable Random Forest model that integrates different dimensions of student data to boost on fairness and consistency in the selection of the career pathway of the learner.

1.2 Research Objectives

- i. To review existing literature to recognize the factors considered in the determination of the career pathway choice of the student;
- ii. To analyze data to find the most consequential factors that influence the career pathway of the student;
- iii. To develop a model to predict the career pathway choice using Random Forest algorithm.
- iv. To assess and validate the model's ability to predict the individual's career pathway.

II. LITERATURE REVIEW

2.1 Theoretical Review

2.1.1 Social Cognitive Career Theory (SCCT)

Social Cognitive Career Theory is one of the most important and widely researched frameworks that can help in the comprehension of career decisions, growth and performance. It is built on the work of psychologist Albert Bandura's broader Social Cognitive Theory. Social Cognitive Career Theory (SCCT) (Lent et al., 1994), elaborates career interests, selections and performance in the face of obstacles. At its engine, SCCT uses the concept of triadic reciprocity initially developed by Bandura; that behavior (career explorations, choice actions, performance, persistence), individual factors (self-efficacy, results expectations, goals), and the environment (social support, barriers and economic state) consistently interact and affect each other. It acknowledges that a factor doesn't singlehandedly determine the results but an interchange among the three influence decisions. Its key constructs are self-efficacy, outcome expectations and personal goals. Self-efficacy is an individual's conviction that they can perform a task successfully. It is the main driver of SCCT because higher self-efficacy results to a higher interest, persistence and resilience. Self-efficacy is influenced by four main sources: mastery experiences, learning by observing others, social influence and persuasion, and psychological state of the individual as either anxious, stressed or excited. Outcome expectation is the belief about the result of a given behavior e.g. a learner believes if they become a doctor in future, they will help the sick, have a stable job or work in a high stress environment. Personal goal is the intent to participate in a specific activity to get a certain result. Goals influence interests and intentions into action. SCCT is better understood in its three development models, one is the Interest Development Model which elucidates that human beings are interested in actions that they feel self-efficacious and want good results. A learner who is confident in science (efficacy) and is convicted that scientists e.g. doctors solve important issues, is likely to be interested in Science as a learning area, and consequently choose STEM career pathway choice under the CBE curriculum. Second is the Choice Model, that elaborates how people make educational and career choices; in that interests affect choice goals but the channel from goal to action is influenced by background aspects like support from parents, scholarships as well as barriers like financial incapacitation. In a real-life example, persons interested in being a pilot (interest), sets their goal to become one (choice goal) despite the barriers like the expensive amounts it costs to train to be a pilot, sturdy self-efficacy and key supports like mentors and supportive parents lead them to the pilot career (choice action). The performance model forecasts the level of performance e.g. success in the STEM pathway or students selecting Arts and Sports pathway. it explains that past successes influence performance indirectly by influencing self-efficacy. This explains why the research is built on the past records of the data to influence the pathway choice the student chooses to pursue. In the settings of CBE, strong performance in subjects, activities or projects develops self-efficacy and optimistic beliefs for related pathways. SCCT additionally puts an emphasis on contextual support barriers, which in this research is embodied by type of school (public/private) as a mediating factor.

2.1.2 Holland's RIASEC Model

Holland (1977) suggested that career fulfillment is a result of a fit between personality and environment, and that human beings work in the environments that are aligned to their personal preferences. It is catalogued into six work environment types: Realistic, Investigative, Artistic, Social, Enterprising and Conventional (RIASEC). These classes lie in line with CBE pathways: STEM (Investigative and Realistic), Arts & Sports (Artistic and Realistic), Social Science and Humanities (Social and Enterprising) and Vocational (Realistic and Conventional). Variables like co-curricular involvement and leadership activities function as representations of interests and character fitting.

2.1.3 Bronfenbrenner's Ecological Systems Theory

Bronfenbrenner (1979) doubled down on the effect of environment on development. His core idea was that a person's development is influenced by series of interlinked environmental systems, from immediate environment to broader cultural settings. The individual is at the heart of linked interacting systems that affects development with time. The five systems of nested levels include the immediate environment - called the microsystems, like family, school, peers, which affect the learner directly. In this research, type of school represents school resources. For example, a score of 60 may represent a higher competency in public school than a private one. The environmental interpretation validates adjustment of forecasting for school setting to enhance fairness. Private schools are more likely to have better infrastructure as previously stated. This subsequently influences the academic outcomes, with more private school students heading to STEM pathway because of availability of STEM infrastructure like science labs and computer labs, mesosystems which is the connection between microsystems like student-teacher interaction, home-work relationships and that the core principle here is that the (in)consistency between the environment affects progression. Private schools have better teacher: student ratio as compared to public schools which results to student performing better in subjects that require intensive nursing like Mathematics and Science, exo-systems which are the indirect environmental influences like the school board and mass media affect the person, macrosystem which is the

bigger context in the culture, society and ideology like political ideologies, economic systems and social practices can influence the career of the person. Students who come from settings that value doctors are more likely to pursue STEM while sporting cultures nurture sportsmen, and chronosystem or the temporal dimension that elaborates that new job and past events also influence career choices – bad experiences influence negatively while positive influences influence the direction towards the path e.g. if a engineers lacked employment opportunities, many of students would not choose STEM career. If a country had successful sportsmen and women, many would resort to choosing the pathway under the CBE curriculum.

2.2 Empirical Review

2.2.1 Factors Considered in the Selection of Career Pathway

Empirical and theoretical evidences suggest that career choice is as a result of several other factors than only academic performance. SCCT wovens interests and choices with self-efficacy, expected results and the type of environment (Lent et al., 1994), while Holland's theory insists interaction between students and the environment (Holland, 1997). These theories give a background for combining academic performance with student interest and the environment the learning takes on. EDM research demonstrate the importance of fusing several student attributes together. For instance, Mgala (2016), showed that educational records in Kenyan schools can be used in forecasting modelling of the student results. Berlikozha et al.(2025a) analyzed feature identification for forecasting of career pathway and insisted on the need of a career advisory platform and determination of the most important factors considered during prediction. Urlamma et al. (2025) modelled a Machine Learning career prediction system using academic performance, student abilities and interest. These studies combined emphasize a multidimensional technique but leave a contextual gap in regard to pathway choice for CBE setting in Kenya.

2.2.2 Factors Embodied in the Secondary Data

The second objective pertains the features found in the actual school data which are only wanted in the theory. The records included the subject performance, project scores, extra-curricular involvement, gender, leadership abilities and the type of school. This is in agreement with EDM literature which accepts archived educational data as a source of patterns for forecasting (Romero & Ventura, 2013, 2020). Feature-selection is specifically useful because a model for prediction can have many features without all features contributing equally towards prediction. Berlikozha et al.(2025a) shows the relevance of selecting the most important factors in making of career prediction model. The paper stretches this line of research to junior secondary schools' data in Kenya with diverse characteristics.

2.2.3 Development of the Prediction Model

Random Forest is recommended where the links between the predicting factors and results may be discrete as well as continuous. Breiman (2001) demonstrated that putting together several decision trees by ensemble learning can support predictive performance and decrease the variance linked with specific trees. In educational forecasting, data-mining techniques have been employed to uncover patterns from student data and aid in decision making (Baradwaj & Pal, 2011; Romero & Ventura, 2020). The gap is not only on the use of Random Forest model itself. The contribution of the paper and its applicability to junior secondary schools in Kenya under CBE curriculum, using a combination of academic, extra-curricular, project and contextual aspects. comparison of Random Forest with Decision Trees and Neural Networks also gives the justification for the selection of Random Forest as the best available classifier algorithm.

2.2.4 Validation of the Model

Model validation is important is important because good performance on a single train-test split doesn't necessarily mean the model is perfect. Stratified cross-validation disclosures a model to several divisions while it maintains the class representation across folds. ROC-AUC gives an extra style of evaluating the model especially in multi-class environment when one-vs-rest curves are employed (Hanley & McNeil, 1982). The present paper hence combines test-set metrics with stratified ten-fold cross validation, ROC-AUC and Random Forest learning curve. This shares stronger evidence than only measuring the accuracy from one split.

2.3 Research Gap

The paper emphasizes the use of machine learning and educational data for career forecasting, and therefore, three gaps exist. First, lack of Kenya-specific Machine Learning in CBE. Despite CBE emphasizing career pathway choice, few researches have applied ML to influence these decisions. Second is, lack of ML-based career guidance models, tailored under Kenyan settings (junior secondary school) for CBE curriculum. Many studies done on Educational Data Mining cover the older 8-4-4 syllabus and lastly, Contextual factors, that few models account for differences in public and private schools.

III. METHODOLOGY

3.1 Study Design

The paper employed a quantitative experimental research design, that focuses on collection of archived records, preprocessing, training, testing and validation of a supervised Machine Learning model. This type of research is best suited because it pertains manipulation of input variables, under specific conditions (type of school) - the mediating variables, to test their effect on the dependent or outcome variable- career pathway selection. The model will be trained on past data and tested on its ability to forecast new information, therefore giving contributing factors into the prediction influence of the independent variables (Breiman, 2001). Similar methods have been widely employed in data mining in educational sector where predictive models are trained and validation done with actual datasets (Romero & Ventura, 2020; Liu & Tan, 2022).

3.2 Area of Study

The research study area was Trans-Nzoia County. It is approximately 360km from Nairobi, Kenya and located in the formerly Rift Valley Province. The secondary schools are located in the rural, semi-urban and urban areas. The county has got a diverse school type - public and private schools, which is pivotal in the assessment and generalization of the model. The county selection was as a result of its embodiment of several other counties in Kenya, accessibility to educational statistics and schools having diverse environmental settings. (Ministry of Education, 2019).

3.3 Target Population

The research was concerned with junior secondary school students in Kenya under the Competency-Based Education (CBE) curriculum, precisely those who already have chosen their career pathways. The justification for using Grade 9 is that they've finished with deciding on their career pathway, allowing the researcher obtain historical data from the class that influenced their selection (independent variables) and their decisions (dependent variables).

3.4 Data Inclusion Method and Dataset Depiction

The research relied on secondary records gotten from databases in selected 11 junior secondary schools in the county of Trans-Nzoia. Schools with a large population of students, complete records and willing to share data were the most preferred. Most populous schools have at least 100 students per class, which results to the availability of more than 1000 student records. Consequently, sampling methods that are traditionally used during research weren't be employed. In its place, the research made use of every student data that meets set criteria for inclusion from the granted school databases.

After getting cleared by the School of Post Graduate Studies of MMUST and getting the necessary document (research permission letter), that authorizes procession of the data collection stage, the researcher visited the Ministry of Education County headquarters, based in Kitale town, Trans-Nzoia county. He enquired for stored data of student records from schools, through the MoE, who asked willing JSS schools to share the stored data - both hardcopy and softcopy data, about their school academic performance data that was considered in selection of career pathway, extra-curricular activity participation and school type. Hardcopy data that was stored in papers was later converted to softcopy by recording the data in excel worksheets. The schools with the biggest population, more complete data of students from the shared data were selected. 9 public schools and 2 private schools met the criterion.

Records that were inconsistent, missing values, or lacked pathway outcomes were be dealt with during data preprocessing. This method guarantees that the dataset that will be used for developing the model is complete, dependable and mirrors the actual-world school statistics. The final dataset had 1,083 records. This figure is aligned with established Machine Learning heuristics, such as the '10-times rule,' which requires a minimum of 10 samples per feature to reduce overfitting and ensure model generalizability (Hastie et al., 2009). With the set of independent variables, this rule suggests a minimum of 50-70 samples. Our target of 1,000 records significantly surpassed this minimum, providing a robust dataset for training and validation of the Random Forest model, therefore increasing the reliability and stability of the results.

3.5 Data Sources and Collection

The research was exclusively reliant on secondary data, obtained from school databases. The data comprised the academic performance that was used while selecting the career pathway of the learner i.e. scores in Mathematics, English, Kiswahili, Integrated Science, Pre-Technical Studies, Agriculture, Religious Education and Creative Arts & Sports. It also included the scoring from CBE projects assessments (Science, Social Studies, & Creative Arts project scores). The study also used co-curricular participation in sports, drama & music, prefect positions or club membership and scouts' involvement. The school type as either public or private and career pathway choice made by the student were also variables of the study.

3.6 Research Instruments

Data extraction templates – Standardized Microsoft Excel forms, with all the variables labelled on the title, were initially employed to record raw data. All the data from the different school excel forms were copied into one main excel form, called the ‘STUDENT_PATHWAYS’. It was later transferred to python for preprocessing and training of the model.

3.7 Research Variables

Independent Variables – Academic performance, project marks, co-curricular involvement, leadership roles.

Mediating Variables - school type as either public or private.

Dependent Variable - Career pathway - STEM, Arts & Sports Science, Social Science

Table 1

Tabular Representation of Variables

Type of Variable	Variable	Indicator	Measurement Scale
Independent	Academic Performance	Subject scores (Math, Science, English, Kiswahili, Pre-Technical Studies, Religious Education, Agriculture, Performing Arts)	Interval
Independent	Project Scores	CBE project marks in Science, Social Studies, C/Arts	Interval
Independent	Co-curricular Involvement	Membership in clubs	Nominal (0=No, 1= Yes)
Independent	Leadership roles	Prefect, scouts, debate	Nominal (0=No, 1= Yes)
Mediating	Type of School	Public/Private	Nominal
Dependent	Career pathway selection	STEM, Arts & Sports, Social Sciences, Technical/Vocational	Nominal

3.8 Data Preparation and Preprocessing

The records were examined for inconsistent data, missing figures and scores that are out of range. Continuous subject marks were checked against their maximum and minimum scorable points. In instances where numerical figures needed treatment, class-level techniques were employed. Categorical data that missed were treated using the modal category. Representation like gender were treated as ‘1’ male and ‘0’ female. The same was applied to other categorical variables (extra-curricular participation) which had yes and no data, yes represented by ‘1’ and no, ‘0’. The three pathways were encoded as 0 = Arts & Sport Science, 1= Social Science and 2=STEM. The dataset was divided into training and testing subgroups. A split of 80:20 was adopted in the research, where 80% of the records was used to train the model and 20% reserved for evaluating the model. The testing set served as an unseen data that examines the performance of the model. By the assessment of the model on data that was unfamiliar with-it during training, it enabled unbiased evaluation of the model’s ability to predict. The study employed Synthetic Minority Over-Sampling Technique (SMOTE) because the data was imbalanced. SMOTE is useful in that it reduces bias of the model in favoring the most represented class.

3.9 Model Development

Before development of the model, Random Forest feature analysis was done to rank the predictor factors. The scores were converted to percentages to allow understanding and ranked in descending order. Three models were developed - Random Forest, Decision Trees and Neural Networks, trained using the three different algorithms. Random Forest was the chief model because it can handle discrete as well as continuous data as well as its ability to explain feature importance (Breiman, 2001). Decision Tree gave an individual tree benchmark while Neural Network provided non-linear one. The model was evaluated using accuracy, precision, recall and F-1 metrics.

3.10 Model Validation

Of the three developed models, Random Forest was the main model of the research. The paper employed ten-fold cross-validation to examine the generalization in different groups of the data while it preserves the representation of the classes of the pathways. Additionally, ROC curves were used and the AUC values of the three classes (Arts, Sport Science and STEM) noted. A learning curve was additionally employed to understand the model performance as the sample size grew.

3.11 Ethical and Data Privacy Considerations

Permission to get data from school archives was granted by the relevant institutions and educational management. Student names were made anonymous before the data was analyzed. The paper intends to come up with a tool that acts as a support system for, parents, teachers, career counsellors and stakeholders and not as a replacement.

IV. FINDINGS & DISCUSSION

4.1 Findings

4.1.1 Considered during Selection of Career Pathway

The literature found that career pathway choice is dependent on several dimensions. Academic performance gives evidence on the displayed subjects while project scores give the results on the student creativity and applied knowledge. Extra-curricular involvement shares visible proof interest and abilities school gives the environment the student is. The above are in line with SCCT's and Holland's theory which insist on the interconnection between personal traits, preferences and settings (Holland, 1997; Lent et al., 1994). The empirical evidence additionally agrees with the use of educational statistics and Machine Learning to discover hidden patterns (Mgala, 2016; Romero & Ventura 2020).

4.1.2 Most Impactful Features Represented in the Data

The dataset had 1,083 student records which the information contained linked to career pathway selection was catalogued widely as academic performance, project scores, extra-curricular participation and context of the learning environment as either based in public or private schools. The available records give a wider set of characters other than performance in the learning areas only. The features in academic performance include subject scores in Mathematics, English, Kiswahili, Integrated Science, Pre-Technical Studies, Agriculture, Religious Education, Creative Arts and Social Studies. Project scores involved Science project scores, Creative Arts and Agriculture project. Extra-curricular activities included participation in Music & Drama, Sports, prefect and scouts. Contextual factors like gender of the learner and type of school were all considered in the dataset.

The research ranked the most important features from the dataset using Random Forest feature-importance analysis. Feature-importance evaluations are acknowledged as the capability of Random Forest models to be interpretive (Breiman, 2001), while EDM studies elaborate the importance of recognizing important patterns and predictors within student data (Romero & Ventura, 2013, 2020). The paper found substantive differences in contribution to career pathway selection; with strongest factors predicting pathway selection to be subject performance in Integrated Science responsible for 29.89% of the combined feature importance. Agriculture was second (16.60%), Pre-Technical Studies (10.47%), Mathematics (8.63%), Sports participation (7.38%), Music & Drama (5.12%) and the type of school (4.84%) as seven the most influential features. The seven highest ranking factors contributed to 82.92% of the feature importance. The other remaining variables though scored lower in feature importance ranking were Social Studies (3.06%), English (3.05%), Creative Arts (3.01%), Religious Education (2.97%), Science Project (2.80%), Kiswahili (2.47%), Creative Arts (2.09%), Agriculture Project (1.95%), Gender (0.43%) and Prefect and Scouts (0.009%). Nevertheless, it is important to note that feature importance should be understood as the relative contribution of factors to the prediction of the model, and not as indication that these features causally determine student career pathway choices.

4.1.3 Model Development and Performance

The study developed three classifier models. Random Forest had the best accuracy score of 89.27%, Decision Tree with 81.46% and 78.04% for Neural Networks. Random Forest also had the best scores in precision, recall and F-1 as compared to the other 2 models.

Table 2

Model Performance Comparison

Model	Accuracy	CV Accuracy	Precision	Recall	F1-Score
Random Forest	89.27%	86%	90.23%	89.27%	89.57%
Decision Trees	81.46%	82%	85.23%	81.26%	82.84%
Neural Networks	78.04%	78.62%	79.96%	78.05%	78.54%

The above demonstrates Random Forest as the best model on the test data that was available. It performs better than Decision Trees which is in agreement with the ensemble-learning theory that suggests combining several decision trees together can boost performance of the prediction model while it minimizes variance (Breiman, 2001). The least performance by Neural Network should not be misinterpreted as Neural Networks being inferior because they best perform with massive amounts of data unlike our dataset of only 1,083 records.

4.1.4 Model Validation

Ten-fold cross validation gave an accuracy of 86%, 82% and 78.62% for Random Forest, Decision Trees and Neural Networks respectively. It is notable that difference in accuracy registered under Random Forest during cross validation was higher compared to the decision tree model, hence Decision Tree was more stable and improved its

accuracy score to 82% from 81.46%. The Decision Tree model consistency suggests that it has better generalizability stability. However Random Forest had a better overall accuracy of 86%, despite dropping from 89%, hence more favorable in forecasting because of its better classifying accuracy. This drop suggests that despite the model performing well on a certain train-test split, the performance is a little bit sensitive to data variations. The reduction in accuracy suggests that a slight degree of overfitting – in that, the model captures patterns that are specific to the training data. The demonstrated behavior is aligned with the ensemble learning theory. Same characteristics have been noted in research such as Breiman (2001), where models that use ensemble techniques show sturdy forecasting power - which slightly lowers when resampled. Contrastingly, Decision Trees, known to be sensitive to data variation, gain from cross validation because it exposes the model to different subgroups of the dataset hence improving model stability and lowering variance (Hastie et al., 2009). Similarly, Neural Network gain from iterated training in different folds, which improves its ability to generalize by learning from different data samples and reducing the random initialization effects (Goodfellow et al., 2016). The research findings propose that despite Random Forest having the best accuracy overall, DT and NN models are demonstratively more consistent.

The Receiver Operating Characteristic (ROC) curve is a pictorial assessment tool used in the assessment of the classification performance of a model by analyzing the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) at different cataloguing thresholds. It shares the comprehensive view of the ability of the model to separate groups. For this study, which is a multi-class classification problem, where the outcomes are STEM, Arts and Social Science pathways, the ROC curve is always extended using a one-vs-rest (OVR) technique. In this approach, a separate ROC curve is made for each class by treating it as the positive class while the rest catalogued as negative. This facilitates the examining of how good the model separates each class individually.

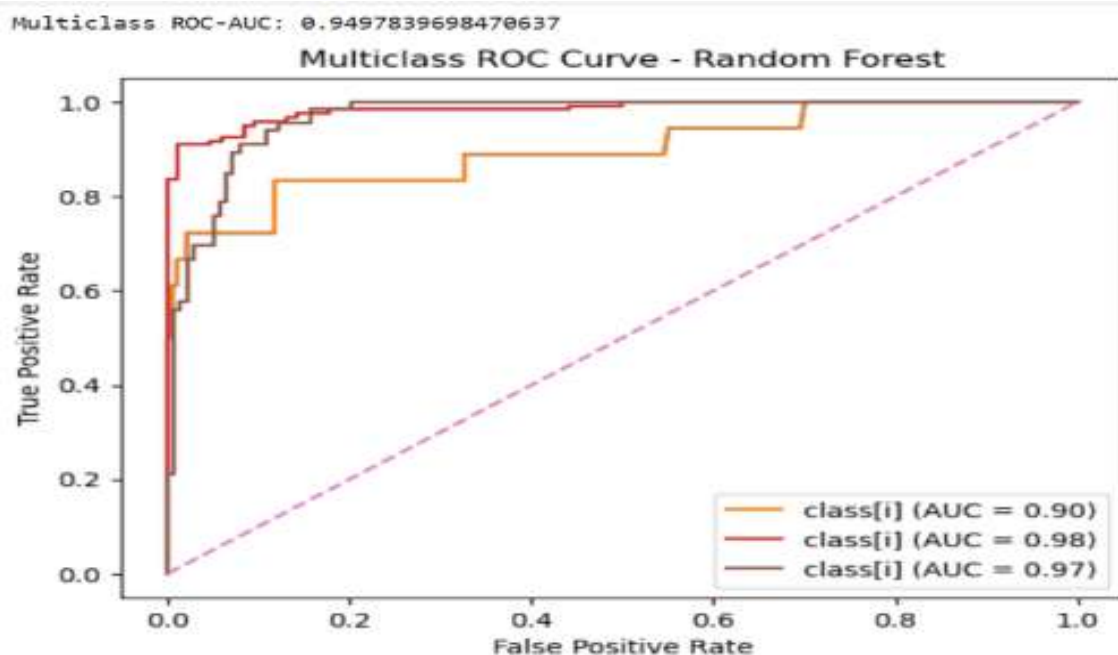


Figure 1
Receiver Operating Curve

Random Forest model reached a high AUC value of 0.90, 0.97 and 0.98 for Arts, Social Science and STEM respectively, thereby showing the model's ability to distinguish between different academic pathways, because the values are more than the threshold (0.80) hence indicates a strong performance. The AUC for the Arts pathway (0.90), which is the minority class is the most complex to classify. This demonstrates that the model has a good ability to discriminate even on the least represented class.



Figure 2
Model Learning Curve Source

In this research, the Random Forest learning curve demonstrated that the accuracy of the training improves as more data is added until it finally converges towards the training curve. The insignificant gap seen between the two means the model has low variance and satisfactory generalization ability. Furthermore, the curve convergence proposes that adding more data (past the current 866 training size) may only lead to minimum improvement meaning the model has achieved optimal learning volume in consideration to the available factors.

4.2 Discussions

The study finding that career pathway forecasting is a multifaced problem concurs with SCCT and Holland's theoretical propositions. SCCT postulates that career interests and decisions come from interrelationship between individual and environmental aspects and not wholly dependent on a single factor (Lent et al., 1994). Likewise, Holland's theory relates career choice to the interaction between personal factors and environment (Holland, 1997). The paper worked on these ideas using the factors that were already available in archived school records. This is fundamental because the model the predictive model does not suggest that a certain subject mark or activity singlehandedly influences a career choice but it is dependent on a set of factors combined to guesstimate the career pathway choice.

The accessibility of 1,083 student records from junior secondary schools additionally agrees with argument that stored educational statistics can act as an actual basis for forecasting modelling. This is in agreement with Mgala (2016), who showed the viability of educational forecasting using Kenyan school data and with Romero and Ventura (2013 & 2020) who explain EDM as a method of understanding patterns and making prediction models from educational statistics. The Random Forest feature importance found that academic performance features ranked highly in influencing career pathway prediction which is in agreement with earlier EDM studies which showed the essence of student academic performance in academic prediction models (Romero & Ventura, 2013). Nevertheless, the current research goes past this evidence by demonstrating that best predicting factors were not exclusively academic performance. The essence of Sports (7.38%) and Music & Drama (5.12%) show that extra-curricular involvement greatly contributed to the prediction. This finding agrees with the career development theories (Holland 1997; Lent et al., 1994) which appreciate interests and experiences as important to career choice. The type of school had a feature importance of 4.84%, seventh, implies that school settings may have important effect on career pathway choice. This agrees with ecological perspective of development, which appreciates that students grow within intermingling environmental arrangements (Bronfenbrenner, 1979). Disparity in resources could result in the difference in patterns observed.

Random Forest's high cataloguing accuracy of 89.27% is in agreement with Breiman's (2001) analysis of ensemble learning and is also widely agrees with recent career recommendation studies that have successfully employed Random Forest (Berlikohza et al., 2025b; Urlamma et al., 2025). The paper findings should however be synthesized in context. The dataset had 1,083 records obtained from only a single county and with an imbalanced pathway class distribution. The study therefore evidences the usability of Random Forest in career pathway prediction but does not guarantee exact accuracy in other Kenyan counties.

The decrease in accuracy of the Random Forest model from 89.27% to 86% when cross validated shows some level of sensitivity to the specific train-test share illustrating why model validation is important. The improvement of the Decision Tree from 81.46% to 82% postulates better consistency of the model its low ability to forecast means that, its stability only isn't a sure measure of picking the most suitable model. The ROC-AUC findings give additional

support for selection of Random Forest model. AUC value scores of 0.90, 0.97 and 0.98 imply powerful ability to discriminate between classes. The relatively low AUC for Arts pathway may be as a result of the smaller representation of the Arts class. this demonstrates why it is important to examine the several class instead of being sorely reliant on accuracy metrics. Using one-vs-rest ROC analysis is in agreement with the role of ROC-AUC as an evaluation of discrimination (Hanley & McNeil, 1982).

V. CONCLUSION & RECOMMENDATIONS

5.1 Conclusion

The study illustrates that machine learning can be used in stored student data in junior secondary schools in Kenya under the CBE curriculum to predict the most suitable career pathway of the student. The literature study found that career pathway selection has several dimensions (multi-dimensional), while the dataset illustrated actual availability of academic performance, project scores, extra-curricular activity as well contextual variables in school records. The Random Forest feature importance ranked academic factors as important variables influencing the model prediction as well as well extra-curricular activities involvement. The study appreciates the role of environment in influencing academic prediction because type of school ranked highly in the feature-importance score. The paper factored out the best models that work with mixed data, robust and reliable. Among the three models developed, Random Forest registered the best performance of 89.26% accuracy which dropped to 86% on cross-validation, due to associated variance, while Decision Tree was 81.43% accurate on development and cross-validation and similarly Neural Network models had an accuracy of 78% after training and cross-validating. Random Forest had An AUC curve with values 0.90, 0.97 and 0.98 for Arts, Social Science and STEM which implies it has good ability to discriminate across different classes.

Random Forest was therefore selected for having a better cataloguing accuracy as compared to the two previous models. It is also easy to visualize the model up to certain depths as it is built on decision trees. Additionally, Random Forest offers feature importance aspect, that helps rank the best factors that influence the outcome of a model. The study findings do not mean the Random Forest is a tool to replace professional career guidance but only acts as a tool to support and compliment career guidance in junior secondary schools in Kenya.

5.2 Recommendations

Teachers and career counsellors in school should consider adopting Machine Learning support tools for making decisions as a complimentary tool while observing professional career counselling and deliberations with the student. Teachers are advised to approach career choice wholly, by considering all factors such as project performance, co-curricular involvement and the type of school instead of traditional methods of only observing the student's academic scores. The Kenya Institute of Curriculum Development should consider piloting studies in other counties in Kenya to understand diverse school settings before further deployment of the model. These pilot studies should conduct public and private school separately to minimize bias. Parents should also be given model explanation and feature explanation to understand the most impactful variables the model analyzed to reach the career pathway choice it suggested. This transparency helps parents understand and embrace Machine Learning decision-making support tools for career pathway choice.

Declaration of Interest

The authors declare that they do not have any known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Declaration

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

- Baradwaj, B. K., & Pal, S. (2011). Mining educational data to analyze students' performance. *International Journal of Advanced Computer Science and Applications*, 2(6), 63–69.
- Berlikozha, B., Serek, A., Amirgaliyev, B., Baishemirov, Z., & Zhamanov, A. (2025a). Identification of key features for career prediction through recursive feature elimination. *Bulletin of Abai KazNPU. Series of Physical and Mathematical Sciences*, 90(2), 189–199. <https://doi.org/10.51889/2959-5894.2025.90.2.016>
- Berlikozha, B., Serek, A., Zhukabayeva, T., Zhamanov, A., & Dias, O. (2025b). Development of method to predict career choice of IT students in Kazakhstan by applying machine learning methods. *Journal of Robotics and Control*, 6(1), 426–436. <https://doi.org/10.18196/jrc.v6i1.25558>



- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Harvard University Press.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29–36. <https://doi.org/10.1148/radiology.143.1.7063747>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
- Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Psychological Assessment Resources.
- Kenya Institute of Curriculum Development. (2017). *Basic education curriculum framework: Nurturing every learner's potential*. <https://kicd.ac.ke/wp-content/uploads/2017/10/CURRICULUMFRAMEWORK.pdf>
- Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior*, 45(1), 79–122. <https://doi.org/10.1006/jvbe.1994.1027>
- Liu, R., & Tan, A. (2020). Towards interpretable automated machine learning for STEM career prediction. *Journal of Educational Data Mining*, 12(2), 19–32. <https://doi.org/10.5281/zenodo.4008073>
- Mgala, M. (2016). *Investigating prediction modelling of academic performance for students in rural schools in Kenya* [Doctoral dissertation, University of Cape Town]. <https://hdl.handle.net/11427/23463>
- Ministry of Education. (2019). *National education sector strategic plan 2018–2022*. Government of Kenya.
- Oduor, A. (2022, April 25). *We are ready for CBC shift, ministry says*. *The Standard*.
- Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), 12–27. <https://doi.org/10.1002/widm.1075>
- Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3), e1355. <https://doi.org/10.1002/widm.1355>
- Urlamma, D., Naga Vasavi, T., Sucharitha, N., Supriya, P., & Aaisha, S. (2025). Smart career advisor using machine learning. *International Journal for Research in Applied Science and Engineering Technology*, 13(4), 5308–5312. <https://doi.org/10.22214/ijraset.2025.69537>