A Machine Learning Approach to Identify the Key Factors Affecting Correct Stream Selection and to Predict Suitable Subject Streams for Advanced Level Students in Sri Lanka

Hasara Abeywardhana[1]; Dr. Lakmini Abeywardhane[2]

[1]Department of Information Technology Sri Lanka Institute of Information Technology Malabe, Sri Lanka
[2]Department of Information Technology Sri Lanka Institute of Information Technology Malabe, Sri Lanaka

Publication Date: 2025/12/11

Abstract: Education plays a vital role in shaping the economic growth and sustainable development of a nation. It is not only a measure of a country's intellectual wealth but also a determining factor in its future progress. In Sri Lanka, education is provided free of charge by the government from primary school through university, ensuring equal access for all students. Within this framework, the General Certificate of Education (Ordinary Level) – G.C.E. (O/L) and the General Certificate of Education (Advanced Level) – G.C.E. (A/L) examinations represent two critical milestones in the academic journey. The G.C.E. (A/L) examination, in particular, serves as the gateway to higher education and university admission, marking a pivotal stage in shaping students' academic and professional futures. At the end of the O/L stage, students are required to select a subject stream such as Science, Arts, Commerce, or Technology to pursue during their A/L studies. This choice has a lasting impact, as it directly determines the student's educational direction and career opportunities. However, many students make this crucial decision based on external influences, such as parental pressure, peer comparison, or limited guidance, rather than through a clear understanding of their academic strengths, personal interests, or long-term career aspirations. Consequently, this often leads to dissatisfaction, stream switching, or even discontinuation of studies. To address this issue, it is essential to adopt a data-driven approach that considers multiple factors, including students' O/L examination performance, inborn talents, extracurricular activities, and preferred professional fields. This research introduces a machine learning-based model the Subject Stream Prediction System—designed to recommend the most suitable A/L subject stream for students. The proposed system not only predicts the optimal subject stream but also provides additional guidance by suggesting potential career paths, relevant educational qualifications, and technical skills aligned with the student's profile. Four supervised machine learning algorithms K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Support Vector Machine (SVM)were trained and evaluated to develop the predictive model, ensuring the highest possible accuracy and reliability.

Keywords: *Machine Learning Algorithm, Subject Stream, Prediction System.*

How to Cite: Hasara Abeywardhana; Dr. Lakmini Abeywardhane (2025) A Machine Learning Approach to Identify the Key Factors Affecting Correct Stream Selection and to Predict Suitable Subject Streams for Advanced Level Students in Sri Lanka. *International Journal of Innovative Science and Research Technology*, 10(11), 2938-2947. https://doi.org/10.38124/ijisrt/25nov1533

I. INTRODUCTION

Education is a very important part of any society. Education makes people who can live in society very well. When the child learns to read and write they feel more confident in their abilities. Education is also a human right in all nations. [1]Therefore, it is the responsibility of every nation to solve the problems which occur in the education systems and find the appropriate solutions. I will work on this research to suggest the best subject stream and subjects to follow for the students for their senior secondary education. I plan to develop that system according to the academic circulars (2016/13) which were published by the educational ministry of Sri Lanka. According to the current education schemas in Sri Lanka, secondary education lasts eight years. It comprises two major parts. Junior secondary education is considered from grade 6 to grade 9. Senior secondary education is considered from grade ten to grade 10 to grade 13. Two types of major exams are conducted by the examination department of Sri Lanka have happened to face to complete the secondary education of a student going to the government school in Sri Lanka. General Certificate of Education Ordinary Level - G.C.E.(O/L)) exam happened to be faced at the end of grade 11. The General Certificate of Education Advanced Level - G.C.E.(A/L)) the exam is held at the end of grade 13.

[1] The student happens to select one major subject area to begin his/her senior secondary education part two and he/she happens to face the Advanced Level examination from that stream. Most of the students in Sri Lanka have completed their primary education successfully. But many students do not complete or drop their senior secondary education for many reasons. The major reason I revealed was that the students do not have enough knowledge to select the correct subject stream to continue the senior secondary education part two. I worked on this research to suggest the best subject stream 978-1-6654-0741-0/22/$31.00 ©2022 IEEE and relevant subjects to follow in secondary education in key stage two (Advanced Level-G.C.E.(A/L). [2]I aimed to target the students who have finished their Ordinary Level exams as main users of this system. According to the current education schemas in Sri Lanka, the secondary level exam results (key stage 2) have affected the career path of most students. I developed the system by using machine learning algorithms to suggest the best subject stream to the students depending on their previous ordinary-level exam results and their inborn talents, skills, and preferred area for doing jobs in the future. After implementing the system, the students happen to fill their Ordinary Level results, for main subjects, extra curriculum activities that they have done in school, and preferred working field that they expect to do their career in the future. [3]The other advantage of this implemented model is that this model suggests another suggestion with appropriate subject streams with suitable job positions relevant to users' input values and essential professional and technical qualifications needed to reach those positions.

II. BACKGROUND

Education plays a central role in shaping an individual's career trajectory and consequently influences the social and economic development of a country. In Sri Lanka, the Gen-eral Certificate of Education (G.C.E.) Ordinary Level (O/L) and Advanced Level (A/L) examinations act as two critical milestones in the school education system. The O/L examina-tion represents the completion of junior secondary education, whereas the A/L examination functions as a gateway to univer-sity admission, professional training, and skilled employment pathways. Therefore, the selection of an appropriate A/L subject stream at the end of the O/L stage is one of the most consequential academic decisions in a student's life. This choice does not merely determine performance in the A/L examination but significantly affects long-term career satisfaction, employability, and higher education opportunities. Despite its importance, many Sri Lankan students struggle with selecting the most suitable A/L stream. Research and school-level observations reveal that several students choose streams due to external pressure, influence from peers or family, prestige associated with certain subject areas, or a limited understanding of their own academic strengths and career aspirations. As a result, mismatches between students' abilities, interests, and chosen streams are frequently reported. These mismatches often contribute to poor academic per-formance, lack of motivation, stress, discontinuation of A/L studies, and in extreme cases, long-term career dissatisfaction among young professionals.

A key issue within the Sri Lankan education context is the absence of structured, data-driven guidance systems to support students during the O/L to A/L transition. Currently, decisions are mostly influenced by subjective judgments: teacher rec-ommendations, parental expectations, or students' perceptions of "popular" streams such as Biology or Physical Science. These methods do not adequately consider individual varia-tions in cognitive capabilities, inborn talents, extracurricular involvement, or long-term career interests. Furthermore, edu-cational performance alone is often insufficient for determining future success because many careers require a combination of academic ability, technical skills, behavioral competencies, and specific personality traits. Therefore, a more holistic and evidence-based decision-making framework is needed.

Internationally, machine learning has gained prominence in educational decision support systems, with various studies focusing on predicting student performance, recommending study programs, and guiding career development. Techniques such as Decision Trees, Random Forests, K-Nearest Neigh-bors, Support Vector Machines, and Gradient Boosting have been successfully applied to identify patterns in student data and support personalized educational decisions. These models demonstrate the capability to extract non-linear relationships among academic, behavioral, and demographic factors, en-abling predictions that are more accurate than traditional advising methods. However, most existing systems are context-specific, institution-specific, or country-specific and therefore cannot be directly applied to the Sri Lankan education struc-ture.

Another limitation in traditional approaches is the lack of multidimensional factor analysis. Students often possess unique combinations of inborn talents (such as creativity, an-alytical thinking, or communication skills) and extracurricular involvement (such as leadership roles, club participation, cadet activities, or sports achievements). These elements contribute significantly to shaping students' abilities and interests but remain largely unutilized in decision-making surrounding A/L stream selection. Moreover, career guidance at school level rarely incorporates real-world labor market trends or job area preferences. Consequently, students may choose streams that do not align with emerging career opportunities in fields such as information technology, engineering, biosciences, business analytics, or creative industries.

This research aims to address these gaps by developing a machine learning–powered advisory system that predicts the most suitable A/L subject stream for students using a combina-tion of academic performance, inborn talents, extracurricular engagement, and preferred job fields. The model incorporates multiple algorithms such as Random Forest, Decision Tree, Support Vector Machine, and K-

Nearest Neighbors to identify the most effective predictive approach. Additionally, the sys-tem integrates talent–stream probability mapping based on real data from working professionals. By analyzing the relationship between their school choices, abilities, extracurricular profiles, and eventual job satisfaction, the system can identify the key factors influencing correct stream selection.

Furthermore, this study incorporates career-field mapping to ensure that recommendations align not only with students' abilities but also with practical employment pathways. By connecting each A/L stream with related job areas, degree programs, and required technical skills, the proposed system offers a comprehensive guidance framework. It also generates personalized explanations, graphical visualizations, and PDF reports to enhance clarity and usability for students, teachers, and parents.

Overall, this research contributes to the Sri Lankan educa-tion system by introducing a data-driven, stakeholder-aligned, and employment-oriented decision support tool for A/L subject stream selection. It fills critical research gaps by combining predictive modeling, talent analytics, academic profiling, and career mapping into a single coherent framework. The findings and developed system can support educational institutions, policymakers, counselors, and students in making informed, personalized, and future-aligned decisions.

III. LITERATURE REVIEW

[1] Developed a machine learning–based system named CareerRec, designed to recommend suitable career paths for IT graduates. The model utilizes both technical and soft skill attributes—such as programming ability, logical reasoning, database administration, networking, and communication pro-ficiency—to generate personalized career suggestions. The dataset used for this study was collected from employees working in the IT sector in Saudi Arabia, meaning that the system's applicability is largely limited to individuals within this regional and professional context. To evaluate its performance, the researchers compared five machine learning algorithms: K-Nearest Neighbors (KNN), Decision Tree (DT), Bagging Meta-Estimator, Gradient Boosting, and XGBoost. Among these, the XGBoost algorithm produced the best performance, achieving an accuracy rate of 70.47 percent, indicating its superior capability in handling complex feature relationships and achieving higher predictive precision. In a related study, Bobadilla et al. [2] (2021) conducted a com-prehensive survey on recommender systems, emphasizing the effectiveness of collaborative filtering techniques in generating personalized suggestions. Their research outlined two primary approaches: user-based and item-based collaborative filtering. The user-based method identifies similarities between users based on shared interests or usage patterns, while the item-based approach focuses on identifying relationships among items that are frequently used together. These methods, though initially developed for e-commerce and digital services, have increasingly been adapted to educational and career recom-mendation systems to improve personalization and predictive accuracy. M.C.B. Natividad et al. (2019) [3] developed a Career Recommender System designed to assist senior high school students in making informed decisions about their future careers. The model incorporated a feature selection technique to identify the most relevant attributes and transform them into precise inputs for processing. It was built using a fuzzy logic–based inference engine, allowing the system to handle uncertainty and provide more flexible, human-like recommendations. However, the framework was specifically tailored to the Philippine educational context, meaning that its applicability is limited to students within that system.

In related work [4]introduced a web-based recommendation system that leverages fuzzy association rule mining to suggest relevant web pages to users. This method enhances recommen-dation efficiency by identifying hidden relationships between user preferences and content features, thereby offering more precise and personalized results. [1]developed a career path recommendation system aimed at guiding graduates toward achieving their desired professional goals based on their cur-rent educational qualifications. The proposed system utilized string-matching techniques combined with decision tree algo-rithms to match individuals with potential career paths. How-ever, [5]the model primarily focused on the most recent educa-tional attainment of the user, assuming that career progression is solely dependent on it. In reality, a more comprehensive approach should also incorporate specialized qualifications, certifications, and multidisciplinary education, as many pro-fessions require diverse academic and technical competencies beyond a single degree. In another study, [6]conducted a com-prehensive review of string-matching algorithms, identifying their classifications, advantages, and associated challenges. Their analysis particularly focused on exact string-matching algorithms, which are fundamental in various computational domains such as text analysis, natural language processing (NLP), speech recognition, and pattern recognition. The study emphasized the importance of optimizing these algorithms for efficiency and adaptability across different application areas. Similarly, [7] introduced a predictive model titled "Cluster Centers Based on XGBoost", designed to forecast students' career preferences. Their model incorporated four key feature categories: r [8]eading interests, mastery of professional skills, behavioral regularity, and family economic background. While this approach offers an innovative method for career predic-tion, it may not fully represent the diversity of all students. Relying on factors such as socioeconomic background could unintentionally create bias, making the system less equitable across varying social and educational contexts.The applica-tion of machine learning in subject stream selection presents a transformative approach to improving students' decision-making processes in senior secondary education.

While existing studies provide valuable [9] insights into machine learning-based educational recommendations, several gaps still exist. Most research focuses primarily on student performance prediction rather than optimizing subject selec-tion using real-world career success metrics. Additionally, there is limited integration of job market trends and satisfac-tion data from professionals who have successfully navigated their career paths. [?]Incorporating real-world career outcomes into machine learning models can significantly enhance the relevance and accuracy of subject stream recommendations. Moreover, the absence of adaptive learning systems that evolve with students' academic progress is another critical limita-tion. Many current recommendation models operate statically, failing to adjust recommendations based on students' chang-ing interests and skills over time. Reinforcement learning approaches, which continuously update recommendations as students develop new competencies, could address this issue and improve the long-term effectiveness of these systems. Another crucial research gap lies in the lack of real-time, interactive student advisory platforms that allow dynamic engagement between students, educators, and AI-powered recommendation systems. Most existing models function as offline prediction tools rather than as integrated decision-support systems. [4]Future research should focus on develop-ing interactive platforms that utilize real-time feedback mech-anisms to refine recommendations based on students' evolving preferences. [6]Additionally, psychometric assessments and personality-based profiling should be incorporated into subject stream selection models to ensure a holistic understanding of students' strengths and weaknesses. Overall, [?]machine learning has shown great potential in revolutionizing subject selection in secondary education by providing evidence-based, personalized recommendations. However, achieving higher ac-curacy and fairness in predictions requires a multidimensional approach that incorporates academic performance, [10]behav-ioral traits, cognitive abilities, career aspirations, and evolving job market demands. Addressing ethical concerns, algorithmic biases, and the interpretability of AI models is essential to gaining trust among students, educators, and policymakers. Future studies should focus on developing transparent, career-aligned, and adaptive recommendation systems that bridge the gap between academic choices and professional success. The integration of reinforcement learning, explainable [?] AI, real-time adaptive models, and career trajectory analysis will fur-ther enhance the efficacy of machine learning in subject stream selection, ensuring that students make informed decisions that align with their potential and future career paths.

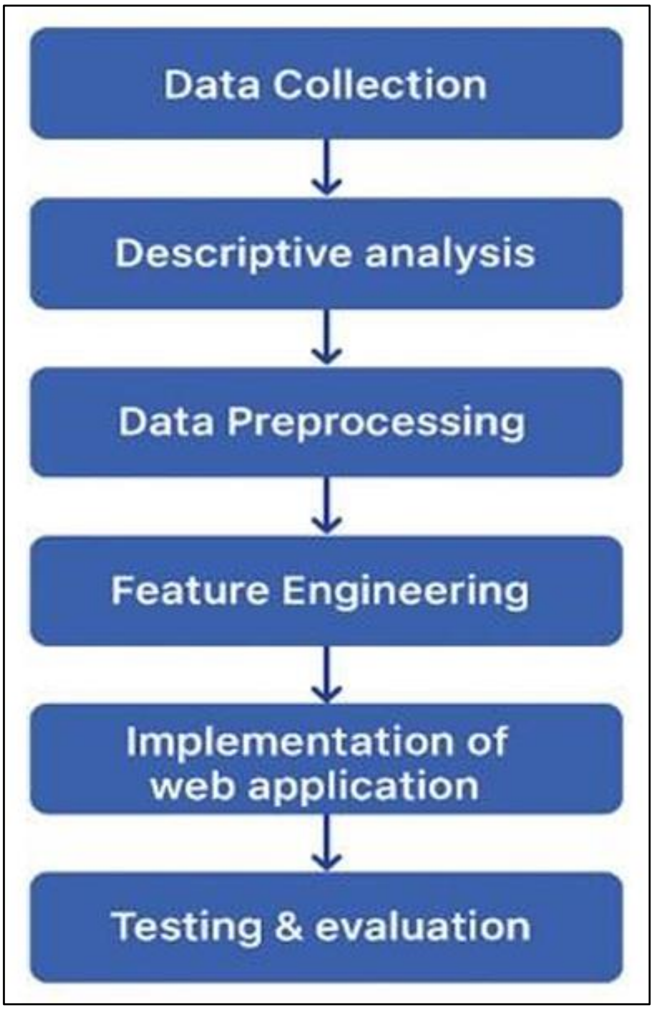


Fig 1 Methodology

IV. METHODOLOGY

This system was developed under above 6 phases. The main phases are Data collection, Descriptive analysis, Data preprocessing, Feature engineering, Implementation of web application and testing and evaluation.

- *Flow Diagram for Methodology*

- *Use Case Diagram*

- Data Collection - Gather survey or questionnaire data from respondents
- Descriptive Analysis - Perform descriptive statistics and visualization
- Data Prepossessing - Clean, encode, and prepare the dataset
- Feature Engineering - Identify and create important features for the ML model
- Implementation of Web Application - Develop and integrate ML model into a web system
- Testing Evaluation - Test application accuracy and us-ability, evaluate model

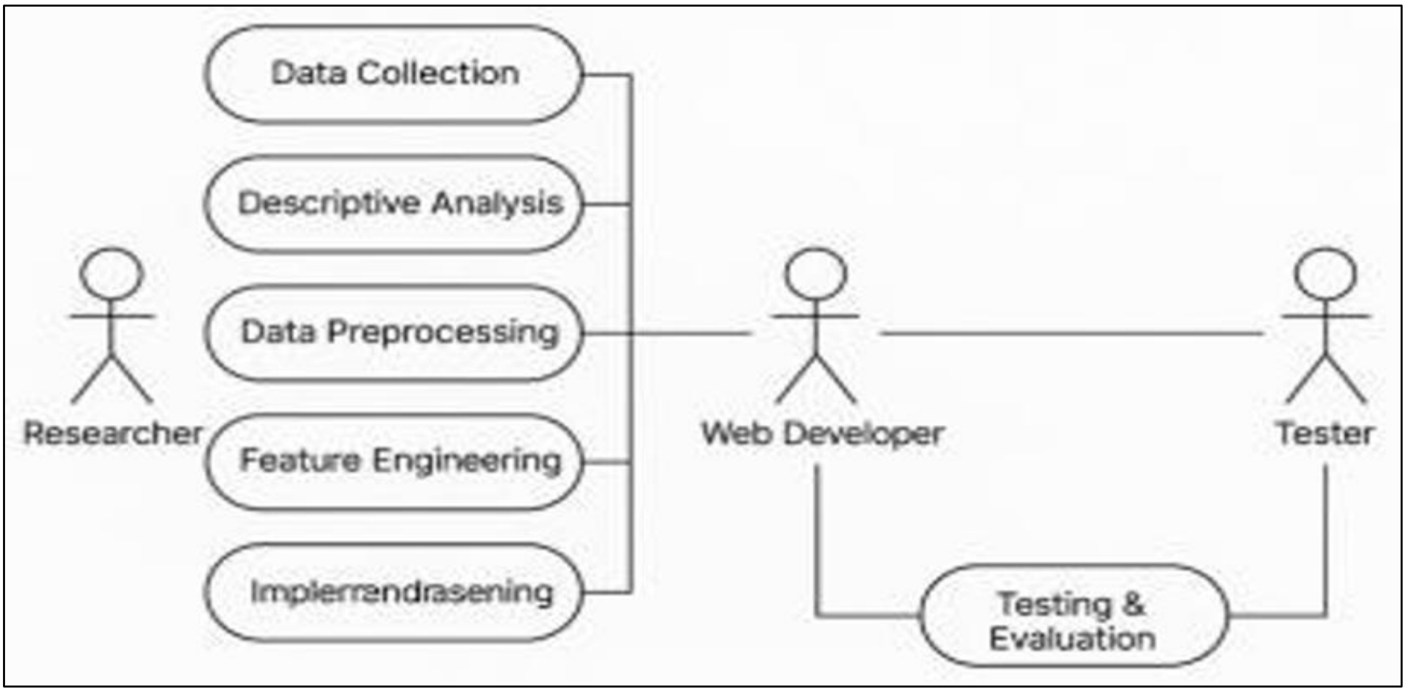


Fig 2 Use Case Diagram

- *Relationships*

- Researcher performs → *Data Collection, Descriptive Analysis, Data Preprocessing, Feature Engineering, and Testing Evaluation*
- Web Developer performs → *Implementation of Web Application*
- Tester participates in → *Testing Evaluation*

- *Data Gathering*

Because of the limitation of the time, I have used two data gathering methods to collect data from the target populations. My target population size was 1000 job holders who are doing their job satisfactorily. To cover my population size, I collected data via an online survey by issuing Google forms. The responses which were coming from the Google form was low. I issued questionnaires to target respondents and collected data from them. I collected 30 percent of relevant data from google forms and 70 percent of data from by issuing questionaries to the target population. I have collected data under 27 features. They are Gender, O/L faced year, Maths Grade, Science Grade, Religion Grade, History Grad, English Grade, Language Grade, A/L faced year, Stream, Inborn talents, Extra curriculum 1, Extra curriculum 2, Extra curriculum 3, Extra curriculum 4, Home location, Family in-come, Mother education level, Father education level, Mother having occupation, Father having occupation, Title of job, Area of job, Educational qualifications for job, Name of de-gree/diploma/MSC, Technical skills for job, Are you satisfied with job A/L stream selection ? The extracurricular activities were categorized into four main types:

- Type 1 – Activities related to membership or leadership roles in clubs, associations, or societies.
- Type 2 – Activities aimed at assessing discipline and character development, such as participation in cadets, scouting, or prefect boards.
- Type 3 – Sports involvement, focusing on the types of sports participated in during school.
- Type 4 – Achievements and recognitions obtained at school, district, or provincial levels.

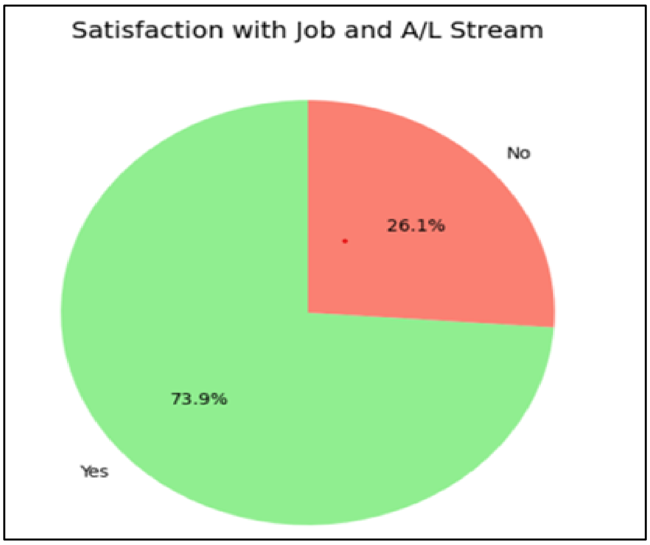


Fig 3 Satisfaction Graph

These categories were designed to evaluate each participant's skills, behavioral attributes, and personal interests. Respondents were also asked to self-assess their inborn talents using a three-level scale: high, medium, or low. Furthermore, information related to their current employment, including job title, working sector, educational background, and technical qualifications required to achieve their current position, was collected. Finally, participants were asked to indicate their level of job satisfaction and, if dissatisfied, to specify the reasons for their dissatisfaction.

According to data gathering, 26.1 percent of respondents were not satisfied without their current job position and the path that they followed to obtain their job. Therefore, their response data were not used to implement the model. 73.9 per percent of respondents were satisfied with their current job positions. I used those respondents' feedback to implement the predicted model. Percentages of user satisfaction and non-satisfaction are shown in figure 3. The ultimate objective of this research is to suggest the best subject stream to continue senior secondary education. According to Sri Lankan education schemas, there are five types of subject stream schemas: Physical Science Stream – Physical Subject Stream (Maths) Stream, Biological Stream, Commerce Stream, Art Stream, and Technological Stream. The technological subject stream was introduced in 2013.Several job holders coming from technological subject stream were less when comparing other subject streams. Therefore, I ignored the responses coming from the technology subject stream to implement the model. The percentage of each subject stream was shown in figure 4.

➢ *Data Preprocessing*

During the data preprocessing stage, the collected dataset was refined and organized to ensure it was suitable for machine learning model development. To address incomplete or in-consistent information, I personally contacted the respondents through email and mobile communication to verify details, correct inaccuracies, and fill in missing values. It was observed that participants often used different terms or naming conven-tions to describe the same features, which could potentially reduce the consistency of the dataset. Therefore, to enhance data quality and improve the predictive accuracy of the ma-chine learning model, these variations were standardized and all independent variables were systematically categorized to maintain uniformity across the dataset.

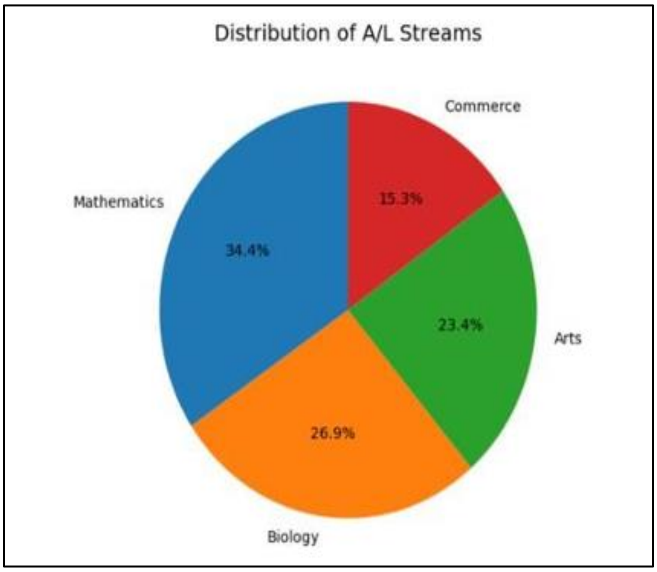


Fig 4 Stream Graph

➢ *Feature Engineering and Feature scaling*

Feature engineering techniques were applied to enhance the quality and relevance of the dataset. Irrelevant attributes such as contact information, gender, and job location were removed to ensure that only meaningful variables contributed to the model's learning process. Subsequently, label encoding was performed on all categorical variables to convert textual data into numerical form suitable for machine learning algorithms. Figure 3.13 presents the dataset prior to label encoding, while Figures 17 and 18 illustrate the transformed values of the inde-pendent variables after encoding. In addition, feature scaling was implemented across all independent variables to minimize data inconsistencies and ensure uniformity in variable ranges. This process helped to enhance model stability.

➢ *Machine Learning Model Training*

Four machine learning algorithms were used for model training.

- K-Nearest Neighbors (KNN) K-Nearest Neighbors is a simple, instance-based learning algorithm that makes predictions based on the similarity between data points. It classifies a new sample by examining the "k" closest neighbors in the training dataset. The algorithm assumes that similar inputs should have similar outputs. KNN does not build an explicit model; instead, it relies heavily on distance metrics such as Euclidean distance.
- Support Vector Machine (SVM) Support Vector Machine is a powerful supervised learning algorithm used for classification by finding the optimal hyperplane that separates different classes. It maximizes the margin between data points of different categories, improving generalization performance. SVM can handle non-linear patterns using kernel functions such as RBF or polynomial kernels.
- Decision Tree Classifier A Decision Tree classifier works by recursively splitting the dataset based on feature values to create a tree-like structure of decision rules. It selects splits that maximize information gain or minimize impurity (e.g., using Gini index or entropy).
- Random Forest Classifier Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. Each tree is trained on a randomly selected subset of data and features, introducing diversity that strengthens the overall model.

V. EXPERIMENTS AND RESULTS

➢ *Analyzing the dataset*

In this section, an analysis was carried out to determine the distribution of respondents across different A/L subject streams. To maintain the relevance of the dataset, approxi-mately 26 percent of responses were excluded — specifically, those from participants who expressed satisfaction with their current job roles and did not intend to progress further in their careers. Using the remaining data, the total number of unique responses for each independent variable was calculated according to subject stream. The corresponding percentages were also derived to identify patterns and variations among the streams. The figures presented below illustrate how these calculations were performed and highlight the comparative distribution of each variable across all subject streams.

- The findings clearly indicate that Arts stream students predominantly possess creative, communicative, and performance-based inborn talents. This pattern sup-ports the notion that students naturally drawn to the Arts are those who value expression, imagination, and crafts-manship over analytical or mathematical skills. Such insights reinforce the importance of aligning A/L stream selection with a student's innate creative strengths to improve engagement and long-term satisfaction in both education and career pathways.
- The inborn talent distribution for Biology stream students reveals a strong dominance of logical, analytical, and strategic abilities. These talents align directly with the requirements of scientific study, research-based thinking, and practical experimentation inherent to the Biology stream.
- The inborn talent profile of Commerce stream students suggests that they are primarily communication-oriented and numerically skilled individuals. Their strengths in writing, mathematical thinking, and computer literacy are well-aligned with the core expectations of the Commerce stream which emphasizes financial analysis, business communication, and digital business applications.
- The inborn talent distribution for Mathematics stream students demonstrates a clear dominance of analytical and cognitive competencies over creative or performance-based abilities. This profile is consistent with the skill de-mands of the Mathematics stream, where success depends on logical reasoning, precision, and analytical thinking.
- Overall, the Arts stream students demonstrate a balanced combination of creativity, leadership, and collaboration, reinforcing that extracurricular involvement complements and enhances their academic and personal strengths.
- The extracurricular profile of Biology stream students reflects a scientifically driven yet creatively balanced orientation. Their high participation in Physics, Science, and Web Team activities indicates a strong alignment with investigative and experimental learning, while their en-gagement in Art and Music clubs shows an appreciation for creative balance and emotional expression
- Overall, Commerce stream students exhibit a balanced development of soft skills (communication, teamwork, creativity) and technical exposure (digital tools, business clubs), indicating readiness for business, entrepreneur-ship, and marketing-related careers.
- The extracurricular involvement of Mathematics stream students highlights a strong orientation toward analytical thinking, technology, and scientific exploration. High participation in Physics, Science, IT, and Robotics clubs reflects the students' preference for logical, structured, and innovative activities, consistent with the skill require-ments of the Mathematics stream
- The extracurricular profile of Arts stream students for Type 02 activities reflects a multifaceted personality that combines creativity, leadership, and communication skills. Their high participation in Reading and Blogging reveals an affinity for literary and expressive forms, while Cadet and Scouting involvement showcases team spirit and organizational discipline.
- The extracurricular involvement of Biology stream stu-dents suggests a well-balanced personality profile that blends scientific curiosity with leadership and creativity. Their engagement in Cadets, Scouting, and Prefect Board reflects discipline, responsibility, and teamwork, while activities such as Reading, Blogging, and Web Creating reveal a growing interest in scientific communication and digital innovation.
- The analysis of Extra-Curricular Activities (Type 02) among Commerce stream students reveals a strong en-gagement in communication, leadership, and collabora-tive activities, supported by moderate participation in creative and digital domains. These patterns are consis-tent with the skill profile required in commerce-related disciplines, where organization, communication, and in-terpersonal abilities are essential.
- The analysis of Extra-Curricular Activities (Type 02) among Mathematics stream students reveals a strong pref-erence for structured, analytical, and leadership-oriented activities, balanced with moderate involvement in creative and collaborative engagements.
- The analysis of Extra-Curricular Activity Type 03, which primarily includes sports and physical activities such as Cricket, Volleyball, Athletics, Netball, and Karate, indicates that this category has minimal influence on A/L stream selection.
- Overall, Type 04 activities reveal that students' extracur-ricular choices strongly mirror their academic special-izations, confirming that academic stream selection and extracurricular involvement are interrelated dimensions of students' personal strengths and interests.
- Students in the Mathematics stream exhibit the highest overall capability, with 76.86 percent showing High Ca-pability in Mathematics and 71.04 percent in Science. This reflects strong analytical and problem-solving skills, consistent with the nature of the stream. The lower percentages of Low and Very Low Capability indicate that most students choosing Mathematics at A/L possess a solid foundational competence from O/L, especially in numeracy and logical reasoning.

For Biology stream students, the subjects most relevant are Science and Mathematics. Both show strong High Capability percentages — Science (71.04 percent) and Maths (76.86 percent) indicating readiness for advanced scientific study Students in the Commerce stream rely heavily on Mathematics, English, and History. Their High Capability in Maths (76.86 percent) and History (74.42 percent) sup-ports this choice, but a relatively lower Medium and Low Capability in English (13.53 percent and 9.07 percent) signals a moderate communication challenge.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25nov1533

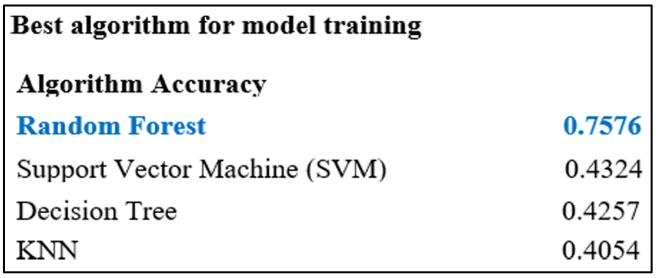
For the Arts stream, core foundation subjects include Lan-guage, History, and Religion. The data show strong High Capability in Religion (75.91 percent), History (74.42 percent), and Language (71.18 percent), confirming a good match between linguistic and humanities strengths.

- The socio-economic background of respondents was an-alyzed across A/L streams using six variables, home location, family income, mother's and father's education levels, and parents' occupations. The distribution results show only minor variations between streams, suggesting that these factors do not have a substantial influence on the choice of A/L stream.

➢ *Predictive Analysis*

All four machine learning algorithms were trained using a dataset consisting of 745 records. To evaluate their per-formance, a confusion matrix was employed as the primary assessment tool. The confusion matrix results for each al-gorithm are presented in the subsequent tables. This matrix is a widely used method for analyzing the effectiveness of classification models, as it provides a detailed comparison between the actual class labels and the predicted outputs generated by the model. Additionally, it enables the identi-fication and quantification of the error rate, offering insight into how accurately the model distinguishes between different classes. The confusion matrix thus serves as a valuable tool for understanding the classification accuracy, precision, recall, and overall performance of each applied algorithm.

This indicates that Random Forest effectively captured complex nonlinear relationships between student features such as O/L subject grades, inborn talents, extracurricular involve-ment, and job area and their correct A/L stream.



Best algorithm for model training

Algorithm	Accuracy
Random Forest	**0.7576**
Support Vector Machine (SVM)	0.4324
Decision Tree	0.4257
KNN	0.4054

Fig 5 Comparison Algorithms

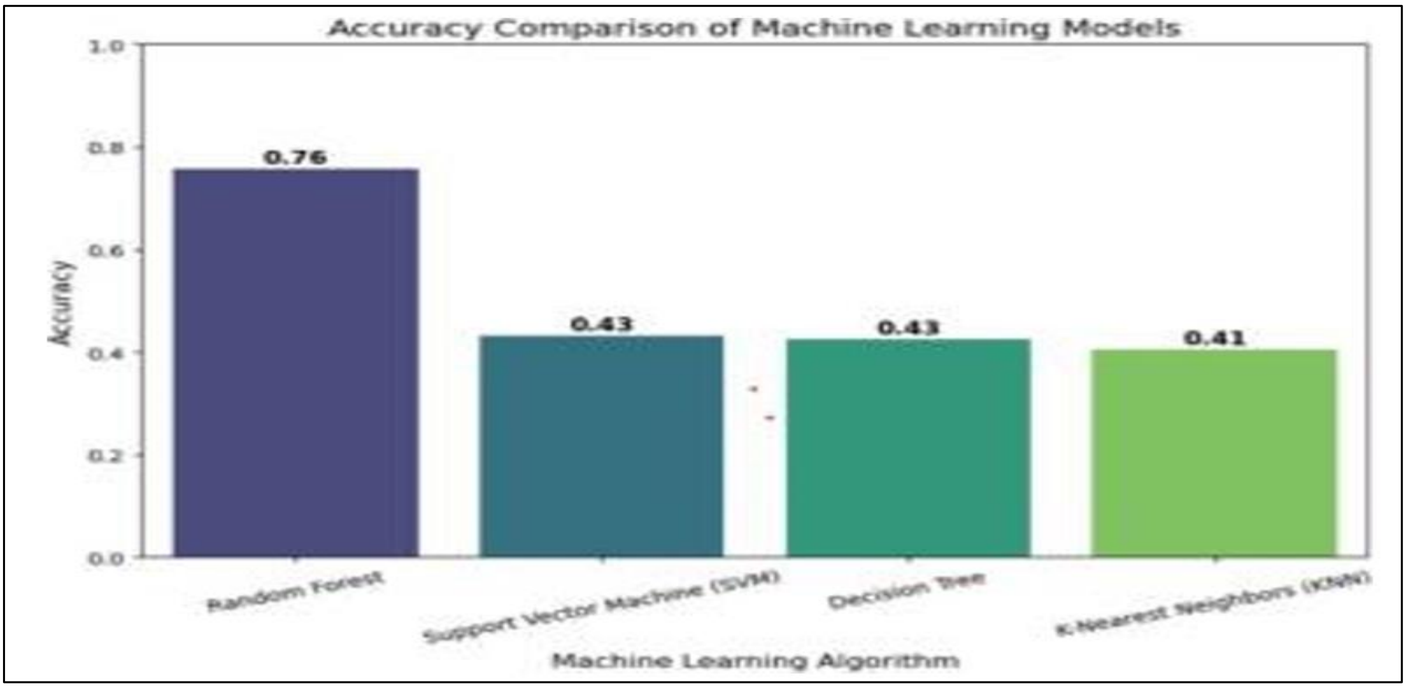


Fig 6 Graph for Comparison Algorithms

➢ *Feature Importance – Random Forest*

This uses for which features most strongly influence the correct A/L stream selection. The Random Forest algorithm provides a built-in measure of feature importance, which quan-tifies how much each input variable contributes to the overall model's prediction performance. The higher the importance score, the more influence that feature has in determining the correct A/L subject stream.

➢ *Cross-Validation Analysis of Random Forest Mode*

The cross-validation results confirm that the Random Forest model maintains stable and reliable predictive performance. While the mean cross-validation accuracy (62.5 percent) is slightly lower than the final test-set accuracy (75.76 per-cent), this difference is expected because cross-validation is a more conservative estimate of model generalization. The low variance between folds demonstrates that the Random Forest model's decision boundaries are robust, and the model is not overfitting to specific subsets of data.

➢ *Main User Interface*

I have developed my web application by adding user in-terfaces to get user details and show the predicted output to the user. A custom-built web application was developed to operationalize the machine learning model and provide an interactive platform for users. The system allows students to input their O/L grades, inborn talents, extracurricular activities, and preferred job areas. Using these inputs, the applica-tion generates predictions for the most suitable A/L subject streams by combining model probabilities, academic scoring, talent–stream relationships, and job-area matching. The in-terface visualizes results through ranked stream suggestions, probability charts, and score breakdowns, providing users with transparent reasoning. Additionally, the application offers personalized career paths and educational recommendations, and enables users to download a complete PDF report sum-marizing predictions, explanations, and guidance. To enhance the performance and practicality of the Random Forest algo-rithm, an additional feature was incorporated into the model to generate more comprehensive predictive outcomes. The standard Random Forest Classifier typically produces a single predicted output; however, in the context of subject stream and career guidance, providing only one recommendation is insufficient, as such a decision can significantly influence a student's future academic and professional trajectory. There-fore, a custom scoring mechanism was developed to extend the model's functionality, enabling it to generate multiple relevant suggestions rather than a single output.

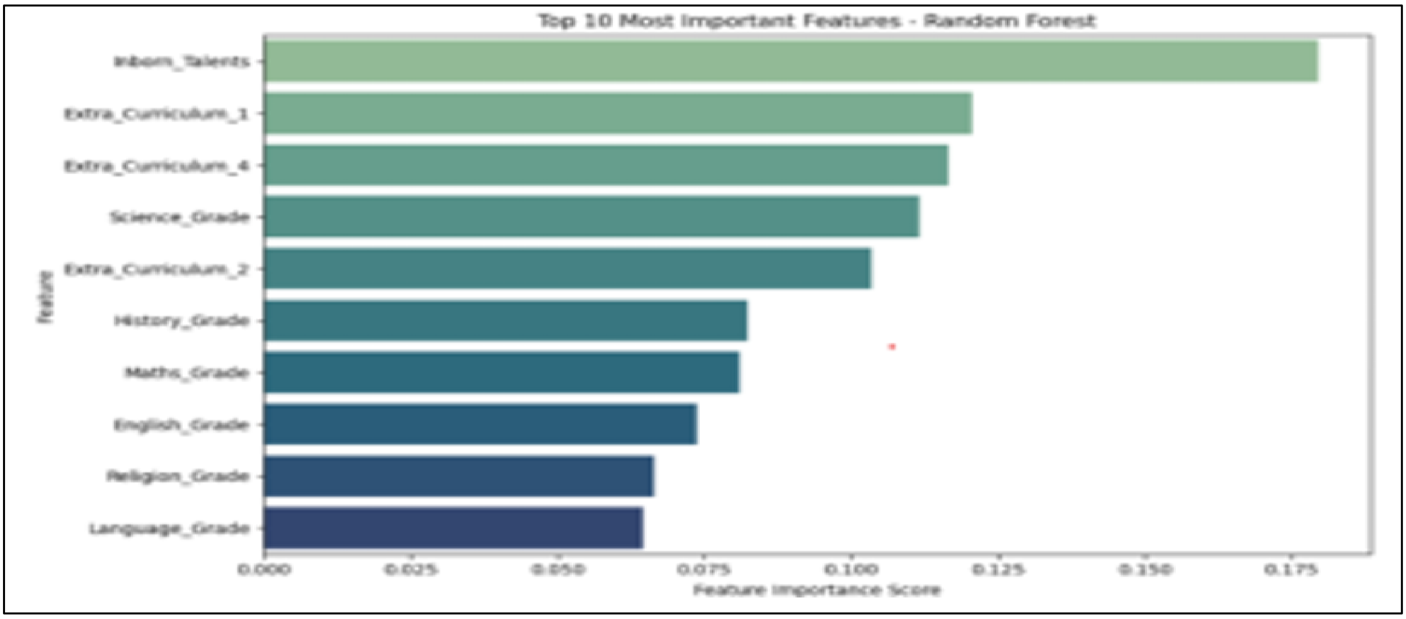
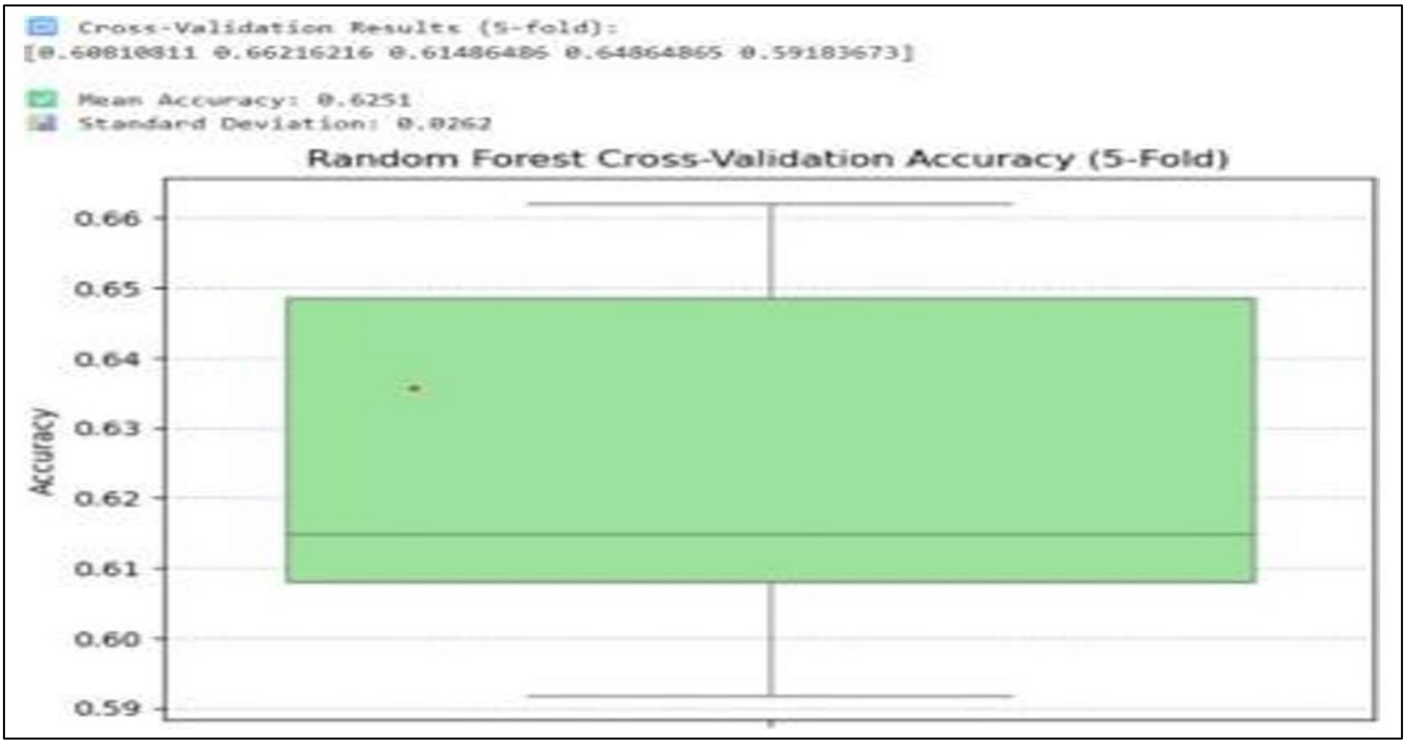


Fig 7 Feature Importance - Random Forest



```
Cross-Validation Results (5-fold):
[0.60810811 0.66216216 0.61486486 0.64864865 0.59183673]

✅ Mean Accuracy: 0.6251
📊 Standard Deviation: 0.0262
```

Fig 8 Cross Validation Results

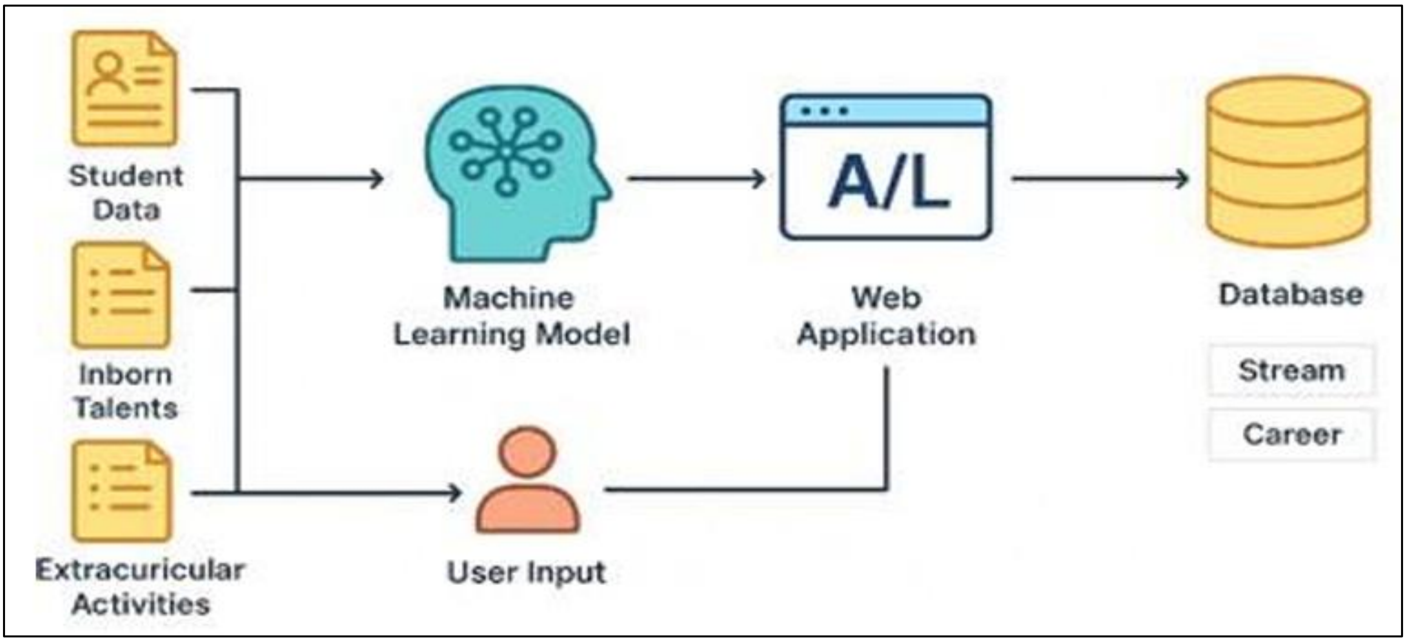


Fig 9 Web App

VI. CONCLUSION

This research successfully demonstrated a machine learning-based approach to identify the key factors influencing the correct selection of A/L subject streams among Sri Lankan students. By integrating academic, personal, and socio-economic variables, the study analyzed how individual characteristics and performance contribute to suitable stream selection. The Random Forest algorithm outperformed other models such as Support Vector Machine, Decision Tree, and K-Nearest Neighbors, achieving an accuracy of approximately 75In addition to the predictive model, a scoring and ranking mechanism was developed to generate multiple personalized recommendations rather than a single categorical output. This enhancement enabled the system to provide the top five most suitable stream and career path combinations based on academic results, talents, extracurricular involvement, and job preferences, ensuring more realistic and ethical guidance for students. The results also indicated that academic and personal factors, particularly subject performance, inborn talents, and extracurricular activities, had a stronger influence on stream selection than socio-economic variables. The proposed web-based decision support system, developed using Streamlit, demonstrates the practical applicability of the model by enabling users to input their data and receive data-driven recommendations for both A/L stream selection and future career development. Overall, this research contributes to the improvement of career guidance in Sri Lankan schools through a data-driven, transparent, and student-centered approach, sup-porting more informed and meaningful educational decisions.

REFERENCES

[1]. H. N. F. Al-Dossari, Z. A. M., A.-Q., and Others, "A machine learning approach to career path choice for information technology graduates," *Engineering, Technology & Applied Science Research*, 2020.

[2]. J. Bobadilla, F. Ortega, and A. Hernando, "Recommender systems survey," *ACM Computing Surveys*, March 2013.

[3]. A. Nagpal and S. P., "Career path suggestion using string matching and decision trees," *International Journal of Computer Applications*, vol. 117, no. 7, 2015.

[4]. T. P. A. Kumar, "Collaborative web recommendation systems based on an effective fuzzy association rule mining algorithm (farm)," *Indian Journal of Computer Science and Engineering*, 2019.

[5]. M. Department of Census and Statistics, *Statistical Pocket Book 2024*. Department of Census and Statistics, Sri Lanka, 2024.

[6]. I. M. Sahib, K. A., P. S., G. G., and K. W., "Exact string matching algorithms: Survey, issues, and future research directions," *IEEE Access*, pp. 1–1, 2019.

[7]. Y. G. Nie, Z. M., Z. R., W. D., and G. Y., "Career choice prediction based on campus big data—mining the potential behavior of college students," *Applied Sciences*, vol. 10, p. 2841, 2020.

[8]. J. Kim and L. K. Kim, "Determinants of academic stream choice among korean secondary students: An empirical study on performance, interest and career alignment," *Korean Journal of Educational Research*, vol. 61, no. 4, pp. 223–240, 2023.

[9]. P. E. Illukkumbura, "Factors affecting students' selection of g.c.e. advanced level science subjects: A case study of sinhala medium students in nuwara-eliya education zone," Master's thesis, University of Peradeniya, 2016.

[10]. Department of Census and Statistics, *Statistical Pocket Book 2023*. Department of Census and Statistics, Sri Lanka, 2023.