



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

K.G.H.Abeywardhana
(Reg. No.: MS21902666)

A THESIS
SUBMITTED TO
SRI LANKA INSTITUTE OF INFORMATION TECHNOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

December 2025

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.



Dr Lakmini Abeywardhana

Approved for MSc. Research Project:



MSc in IT Programme Co-ordinator, SLIIT

Approved for MSc:

Head of Graduate Studies, FoC, SLIIT

DECLARATION

This is to certify that the work is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources). The work has not previously been submitted in any form to the Sri Lanka Institute of Information Technology or to any other institution for assessment for any other purpose.

Sign: *Hasara*

K.G.H.Abeywardhana

Date: 2025/12/30

ABSTRACT

Hasara Abeywardhane Kankanam Gamage

MSc. in Information Technology

Supervisor: Dr Lakmini Abeywardhana

December 2025

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

ACKNOWLEDGEMENT

My sincere thanks go to the authority of the Sri Lanka Institute of Information Technology (SLIIT) for pointing me to the correct path to conduct and complete this research project in my final year degree program M.Sc. in Information Technology. I would pay my gratitude to the supervisor Dr. Lakmini Abeywardhana for her valuable guidance and assistance in this work. Your feedback helped me narrow down my research topic, select one particular area, and develop it as the major part. It pushed me to think deeply about one major area, make a clear path to reach my research objectives, and complete the research successfully.

Always my parents were a big strength to me to finish my research up to this level. They were a giant shadow for me and helped me overcome the challenges that came my way. Also, my husband, my sister, who was my closest friends, helped me to successfully overcome those difficulties by sharing their knowledge with me when I faced difficulties.

Thank You very much my friends who are following the same course for helping me every time I am stuck on some issues. Your feedback also has been very important to me to succeed in this work.

TABLE OF CONTENTS

DECLARATION	ii
ABSTRACT.....	iii
ACKNOWLEDGEMENT	iv
TABLE OF CONTENTS.....	v
List of Figures.....	vii
List Of Tables	ix
Chapter 1 Introduction	1
1.1 Background of the Study.....	1
1.2 Research Question.....	2
1.3 Objectives.....	4
1.4 Gaps in Existing Literature	7
Chapter 2 Literature Review	9
Chapter 3 Research Design and Methodology.....	20
3.1 Flow Diagram for the Proposed Model.....	20
3.2 Conceptual Framework	20
3.3 Data Processing & Feature Engineering	22
3.4 Machine Learning Model Development	22
3.5 Use case Diagram.....	23
3.6 Data Gathering	23
3.7 Data Preprocessing.....	29
3.8 Feature Engineering and Feature scaling	40
3.7 Building the Model.....	41
3.7.1 Machine Learning.....	41
3.7.2 Tools and Techniques	43
3.8 Experiment and Results.....	45
3.8.1 Analyzing the dataset.....	45
3.8.2 Predictive Analysis	63
3.8.2 Feature Importance – Random Forest.....	70
3.9 Main User Interfaces	74
3.10 Improve predicted output	80
3.11 Testing and Evaluation.....	82
3.11.1 Cross Validation	82
3.11.2 K-Fold Validation	82

3.11.3 Execution Time for final model.....	86
3.11.4 Evaluation of the User friendless of the Model	86
Chapter 4 Conclusion.....	89
References.....	90
Appendix	93
Appendix 1: Circulars 2008/1.....	93
Appendix 2: Circulars 2016-13s.....	96
Appendix 3 – Questionnaire.....	99
Appendix 4 – Codes.....	101

List of Figures

Figure 3.1 Research Methodology	20
Figure 3.2 Use case Diagram	23
Figure 3.3 Data Collecting Methods.....	24
Figure 3.4 Satisfaction Graph & dividing streams.....	28
Figure 3.5 Reasons for not satisfied.....	29
Figure 3.7 Subject grade categorization graph 1	35
Figure 3.8 Subject grade categorization graph 2	36
Figure 3.9 Subject grade categorization graph 3	36
Figure 3.10 Extra curriculum activities type 1.....	37
Figure 3.11 Extra curriculum activities type 2.....	37
Figure 3.12 Extra curriculum activities type 3.....	38
Figure 3.14 Graph of inborn talents.....	39
Figure 3.15 Graph of Job areas	40
Figure 3.16 Results of Before pre-processing.....	41
Figure 3.17 Results of After pre-processing	41
Figure 3.18 Inborn talents of arts stream	45
Figure 3.19 Inborn talents of Biology stream	46
Figure 3.20 Inborn talents of Commerce stream	47
Figure 3.21 Inborn talents of Mathematics stream	48
Figure 3.22 Extra curriculum activities type 1 of Mathematics stream	49
Figure 3.23 Extra Curriculum activities type 1 Biology stream	50
Figure 3.24 Extra Curriculum activities type 1 Commerce stream	51
Figure 3.25 Extra Curriculum activities type 1 Mathematics stream	52
Figure 3.26 Extra Curriculum activities type 2 Arts stream	53
Figure 3.27 Extra Curriculum activities type 2 Biology stream	54
Figure 3.28 Extra Curriculum activities type 2 Commerce stream	55
Figure 3.29 Extra Curriculum activities type 2 Mathematics stream	56
Figure 3.30 Extra Curriculum activities type 3 streams vice categorization	57
Figure 3.31 Extra Curriculum activities type 4 streams vice categorization	58
Figure 3.32 Distribution for Home Location	61
Figure 3.33 Distribution for Family Income.....	61
Figure 3.34 Distribution for Mother Education	61
Figure 3.35 Distribution for Father Education.....	62
Figure 3.36 Random Forest model training results.....	65
Figure 3.37 SVM Model training.....	65
Figure 3.38 Decision Tree model training	66
Figure 3.39 KNN Model training.....	67
Figure 3.40 Accuracy level of machine learning models	68
Figure 3.41 Feature importance	69
Figure 3.43 Web App.....	70
Table 3.12 Description of main subjects in ordinary level	71
Figure 3.42 5-fold cross validation of random forest	73
Figure 3.43 Web App.....	74

Figure 3.44 Interface 1	74
Figure 3.45 Interface 2	76
Figure 3.46 Interface 3	77
Figure 3.47 Interface 4	77
Figure 3.48 Interface 5	78
Figure 3.49 Interface 6	78
Figure 3.50 Increasing output	81

List Of Tables

Table 2.1 Literature Review	19
Table 3.1 Inborn Talents Categorization	30
Table 3.2 Extra curriculum activities type 1 categorization	30
Table 3.3 Extra curriculum activities type 2 categorization	30
Table 3.4 Extra curriculum activities type 3 categorization	31
Table 3.5 Extra curriculum activities type 4 categorization	31
Table 3.6 Job area categorization.....	31
Table 3.7 Stream vice activities categorization	32
Table 3.8 Main Subject Description	33
Table 3.9 Categorization of subject grades in O/L examination	34
Table 3.1 Subject grade categorization.....	35
Table 3.13 Extra curriculum activities type 4.....	38
Table 3.11 Table Stream levels.....	60
Table 3.12 Results of Randomized search CV	65
Table 3.12 Description of main subjects in ordinary level	71
Table 3.13 Table for testing output.....	88