

**Department of Computer Science and Engineering**  
**University of Moratuwa**



**CS4202 - Research and Development Project**  
**Data Driven Instruction Strategies for Sri Lankan**  
**Schools**

**Group Members**

140097M D.G.V.M. Dayasiri

140253N S.A.N. Jayasooriya

140381E R.B. Mahanama

140392M W.S. Mendis

**Supervisors**

Dr. Uthayasanker Thayasivam, Prof. Umashanger Thayasivam

**Coordinated by**

Dr. Charith Chitraranjan

THIS REPORT IS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR  
THE AWARD OF THE DEGREE OF BACHELOR OF SCIENCE OF ENGINEERING AT  
UNIVERSITY OF MORATUWA, SRI LANKA.

December 28, 2018

## Declaration

We declare that this is our own work and this report does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgment is made in the text. Also, we hereby grant to the University of Moratuwa the non-exclusive right to reproduce and distribute our thesis, in whole or in part in print, electronic or any other medium. We retain the right to use this content in whole or part in future works (such as articles or books).

Signatures of the candidates:

.....

D.G.V.M Dayasiri - 140097M

.....

S.A.N. Jayasooriya - 140253N

.....

R.B. Mahanama - 140381E

.....

W.S. Mendis - 140392M

Supervisors:

.....

(Signature and Date)

Dr. Uthayasanker Thayasivam

Prof. Umashanger Thayasivam

Coordinator:

.....

(Signature and Date)

Dr. Charith Chitraranjan

## **Abstract**

Data-driven instruction strategies platform is capable of providing unique, individual-based instructions to high school students in order to improve their learning experiences while providing strategic instructions for the teachers and administrators. Overall, this platform has the potential to enhance the level of the whole Sri Lankan educational system by providing strategic instructions to all the main personals within the system. This platform even has the potential to be even extended to e-learning platforms and private educational institutes. These institutes also require personalized instructions for their students because the student to teacher ratio is much high. Hence this platform facilitates and covers up most of the aspects of the education sector by providing strategic instructions.

The research was based on the collected dataset [20] from the Sri Lankan students in grade 6, 7 and 8. It consists of data from over 700 students covering various demographics of the country. The project addresses mainly four research areas and they are described in the separate chapters in the report.

1. Student Performance Prediction
2. Learning Style Analysis
3. Impact of Assistive Learning
4. Subject Level Relationships

The research problem, proposed methodology and research outcomes are identified for each four key research areas. And finally the results of these research problems integrated to a software application which can be used in providing data-driven instructions and insights for main key stakeholders of the education sector in Sri Lanka.

## **Acknowledgment**

We would like to take this opportunity to express our gratitude to all the people who helped and guided us in making this project a success. Without their contribution and guidance, this project might not be successful.

First of all, we would like to thank our project supervisor, Dr. Uthayasanker Thayasivam for providing this project idea and guiding us throughout the year to make this project successful. His knowledge, instructions, and suggestions helped us immensely throughout the time period of the project to identify relevant research areas and address them effectively.

Then, we are extremely grateful to our co-project supervisor, Prof. Umashanger Thayasivam for guiding and mentoring us throughout the duration of the project by sharing his expertise and insights with us.

We would also like to thank our evaluation panel, Prof. Gihan Dias, Dr. Shantha Fernando for their reviews and valuable feedback on the project which were immensely useful in identifying issues in the project and research areas which were required to be addressed. We would also like to thank Dr. Charith Chithranjan for coordinating the final year projects of the Department of Computer Science and Engineering, University of Moratuwa for providing the support for us to successfully complete the project.

Finally, we would like to especially thank the principals, teachers and other staff members along with the students of the schools which were participated in the data collection process of our project.

# Table of Contents

1 INTRODUCTION .....	1
1.1 Problem .....	1
1.2 Background .....	1
1.3 Motivation .....	2
2 LITERATURE REVIEW.....	3
3 METHODOLOGY.....	4
3.1 Section 1 - Data Collection .....	4
3.1.1 Introduction .....	4
3.1.2 Literature Review .....	7
3.1.3 Procedure.....	8
3.1.4 Discussion .....	10
3.1.5 Conclusions .....	14
3.2 Section 2 - Student Performance Prediction .....	15
3.2.1 Introduction .....	15
3.2.2 Literature Review .....	16
3.2.3 Procedure.....	18
3.2.4 Experiments.....	20
3.2.5 Results .....	21
3.2.6 Discussion .....	22
3.2.7 Conclusions .....	24

3.3 Section 3 - Learning Style Analysis.....	26
3.3.1 Introduction.....	26
3.3.2 Literature Review.....	27
3.3.3 Experiments.....	29
3.3.4 Discussion.....	30
3.3.5 Conclusions.....	31
3.4 Section 4 - Impact of Assistive Learning.....	33
3.4.1 Introduction.....	33
3.4.2 Literature Review.....	34
3.4.3 Experiments.....	35
3.4.4 Results and Discussion.....	39
3.4.5 Conclusions.....	42
3.5 Section 5 - Subject Level Relationships.....	43
3.5.1 Introduction.....	43
3.5.2 Literature Review.....	44
3.5.3 Experiment.....	45
3.5.4 Discussion.....	47
3.5.5 Conclusions.....	52
4 DATA DRIVEN INSTRUCTION STRATEGY PLATFORM - SOFIA.....	53
4.1 Student Interface.....	53
4.2 Teacher Interface.....	56

5 CONCLUSIONS.....	61
6 REFERENCES.....	63

## List of Figures

Figure 1: Skewed Responses Distribution .....	11
Figure 2: Non-skewed Responses Distribution.....	11
Figure 3: Pearson's Correlation Coefficient among the Questions .....	13
Figure 4 : Final Model Diagram Representation .....	20
Figure 5: Pearson Correlation among all the Subjects for Grade 6 and 7.....	47
Figure 6: Pearson Correlation among all the Subjects for Grade 8 and Spearman Correlation among all Subjects for Grade 6.....	47
Figure 7: Spearman Correlation among all the Subjects for Grade 7 and 8 .....	48
Figure 8: Kendall Correlation among all the Subjects for Grade 6 and 7.....	48
Figure 9: Kendall Correlation among all the Subjects for Grade 8.....	49
Figure 10: Hierarchical Cluster Dendrogram for Pearson's Correlation Matrix with Cosine Distance & Average Cluster Method .....	51
Figure 11: Sofia Login Interface .....	53
Figure 12: Student Data Collection Interface.....	54
Figure 13: Student Subject Overview .....	55
Figure 14: Student Subject Specific View with Learning Instructions.....	56
Figure 15: Teacher Dashboard .....	57
Figure 16: Individual Student View .....	57
Figure 17: Student's Grade View.....	58
Figure 18: Student Learning Styles Overview .....	59
Figure 19: Subject Correlation Map.....	60



## List of Tables

Table 1: Collected Data Types .....	6
Table 2: Learning Background Data .....	7
Table 3: Average of Answers Option Count.....	12
Table 4: Parent Education Level Percentages .....	14
Table 5: Gender Percentages.....	14
Table 6: Grade Classes.....	19
Table 7: Prediction Results .....	21
Table 8: Confusion Matrix for Random Forest (Left) and AdaBoost (Right).....	22
Table 9: Regression Models: Impact of Learning Style Incorporation.....	23
Table 10: Classification Models: Impact of Learning Style Incorporation.....	24
Table 11: Cluster Evaluation.....	30
Table 12: Sub Cluster Evaluation .....	31
Table 13: Marks Variation Plot due to Assistive Learning – Maths, Grade 8.....	36
Table 14: Marks Variation Plot due to Assistive Learning – Maths, Grade 7, 8.....	36
Table 15: Marks Variation Plot due to Assistive Learning –English, Grade 8.....	37
Table 16: T-Test Results Comparing Two Groups for each Year .....	40
Table 17: T-Test Results Comparing Two Group for each Year.....	40
Table 18: T-Test Results Comparing Two Groups for each Year .....	41
Table 19: Descriptive Statistics for Subjects by Year.....	46
Table 20: Higher Correlated Subjects in Grade 6 .....	50

Table 21: Comparison of Spearman correlation coefficients in grade 6, 7 and 8..... 51

# **1 INTRODUCTION**

## **1.1 Problem**

The education system in Sri Lanka lacks a proper methodology for providing strategic instructions [1] [2] to improve the learning experience of students. Furthermore, the teachers are unable to improve the teaching process due to lack of proper individual student evaluations. Additionally, authorities are having difficulties with calibrating Sri Lankan syllabi at the fine-grained level with an analytics-driven design of courses [3].

The project intends for enhancing the educational experience for students, teachers and the authorities through providing data-driven instruction strategies. In the case of students, they would be provided with the instructions that optimize their learning process according to the contents of the subject. On the other hand, the teachers would be able to receive instructions that optimize the teaching process to match with the types of learners in the class while providing instructions for the authorities to calibrate the syllabi to match with the evaluation criteria of the examination.

## **1.2 Background**

There exist a significant level of uncertainty in the teaching and learning process on whether a course will be able to help students to achieve the grading criteria expected at the examination. In many cases, the evaluation of the suitability is done at the end of examination and the measures will be reactive. The methodologies that have been identified in such evaluations will not be suitable for the next set of students who will be taking the course. Thus the measures are required to be proactive rather than reactive.

Further, these evaluations do not consider the differences in the unique learning styles and sociological backgrounds among students. For instance, some students might prefer their own ways of approaching issues while some might follow the traditional methodology. Further such differences require each student to be taught in

a different manner for retaining the education. With no consideration given for such information, the designing and teaching process would not provide the optimum course structure for the students.

The research is about how this information can be incorporated into building a model that is capable of providing strategic instructions for students, teacher, and administrators for improving the learning experience.

### **1.3 Motivation**

The primary motivation of this research project is enhancing the personalized learning experience of the Sri Lankan school students through an analytical platform. The generation of personalized insights based on learning styles profiling of individual students is useful for students to improve their learning styles based on strengths and weaknesses. Student profiling is also useful for teachers in identifying unique educational requirements of each individual students in the classroom. Facilitations for this kind of an individual evaluation of the students lack in the current Sri Lankan education context. Further, high-level analysis on collected data regarding subject level relationships and impact of external factors for education is useful for educational administrators in policy making and calibrating the syllabi at a fine-grained level.

## **2 LITERATURE REVIEW**

The related researches are analyzed in detail in the related sections of the Methodology chapter. The literature reviews of the identified research problems are also included in the aforementioned subsections. The methodology chapter consists of the following main five subsections.

1. Data Collection
2. Student Performance Prediction
3. Learning Style Analysis
4. Impact of Assistive Learning
5. Subject Level Correlation

The data collection section addresses the related educational data mining researches which carried out similar data collection processes. The suggested data collection procedures in the literature are critically analyzed and compared with the proposed methodology. In the student performance prediction section, related performance prediction models and algorithms are evaluated and analyzed. The proposed prediction models in the literature are comparatively analyzed with the developed prediction model to evaluate the prediction accuracy. The learning style analysis section analyzes various existing methodologies to evaluate student learning styles. It also addresses the importance of developing a proper learning style evaluation methodology for Sri Lankan educational context. In the impact of assistive learning section, related local and international researches and literature were critically analyzed. The resulted conclusion of the research problem is compared with related literature results for validation. Finally, the literature review of the subject-level correlation analyzes the literature related to the research topic. It emphasizes the advantage of the holistic approach of the proposed methodology as compared to other suggested research procedures.

## **3 METHODOLOGY**

### **3.1 Section 1 - Data Collection**

#### **3.1.1 Introduction**

Educational data mining relates with discovering novel and potentially useful information from large amounts of data. Even though there has been significant work on educational data mining itself [4], there is a substantial amount of work related to data collection for educational data mining. Data collection for educational data mining has varying limitations depending on the education system that data collection is focused on. Furthermore, data collection plays an important role in educational data mining as data collection sets the foundation for educational data mining and related work.

Among the key challenges in educational data mining is that data being problematic to measure and collect. The main reason for the problematic situation is that there are a significant number of factors influencing student behavior and performance in examinations [6] [8]. For instance, the school condition, teaching methodologies, student learning patterns, student demographics are few of many factors influencing the education performance. Increasing the number of dimensions in consideration in an educational data mining project would complicate the data mining process.

In addition to the complication of the required data sets, the higher number of factors influencing academic performance would require a significant number of data to eliminate any bias that might be introduced to the dataset. In the case of a typical data mining research on educational data would include process being applied to a classroom or several classrooms in a single institution. For instance, the locality of the student can be a factor in determining performance and such research would not produce accurate results due to the bias in the data set.

Despite numerous researches being carried out to identify the relationship of different student aspects on the student performance, no common ground has been developed for data collection for educational data mining projects.

A framework for data collection in educational data mining was developed with the learnings from educational data mining projects. The proposed framework discusses the educational data collection for educational data mining under three main dimensions as performance, student learning background and learning style data. Furthermore, this framework also includes various means of data collection and the factors that should be considered for means of data collection.

Unlike the various researches carried out in the literature to correlate the student performance with different metrics, the framework provides an effective, comprehensive and complete methodology for data collection for educational data mining research projects along with valuable validation techniques.

<b>Data Type</b>	<b>Objective</b>
Performance Data	<p>Consists of examination marks for all the subjects over a three year period (grade 6, 7, 8)</p> <p>Useful in assessing weak and strong subject areas of each student.</p> <p>Can be used in the development of a mark prediction model</p>
Learning Background Data	<p>Consists of economic, sociological background, participation in assistive teaching programs and extra-curricular activities of the students.</p> <p>Useful in evaluating the impact of sociological background for the student performances.</p> <p>Can be used to evaluate the impact of assistive teaching programs for a certain subject</p>
Learning Style Data	<p>Evaluated based on student responses for the created questionnaire.</p> <p>Inspired by the LCI model [9]</p>

	Useful in evaluating the correlation between student learning styles and their learning preferences.
--	--

**Table 1: Collected Data Types**

<b>Feature</b>	<b>Type</b>	<b>Description / Values</b>
Father's Education Level	Categorical	1 - Below O/L 2 - O/L
Mother's Education Level	Categorical	3 - A/L 4 - Graduate
Sibling Education Level	Categorical	For each sibling,  1 - Below O/L 2 - O/L 3 - A/L 4 - Graduate  And calculate the total score for every sibling.  Usually 1-20
Number of Siblings	Categorical	1-10
Grade 5 Scholarship Marks	Numerical	Scholarship marks out of 200
Tuition	Categorical	Indicates whether the student participated in tuition for the considered subject in grade 6,7 and 8  Only in grade 6 - 100



		<p>Only in grade 7 - 010</p> <p>Only in grade 8 - 001</p> <p>In all grades - 111</p> <p>Likewise from binary values between 000 to 111</p>
--	--	--

**Table 2: Learning Background Data**

### 3.1.2 Literature Review

Types of data play a major role in the context of educational data collection. According to the research carried out in [9], the teachers and other educational staff are much interested in outcome data such as test results, assignment marks. However, usage of only outcome data is not sufficient for the data-driven decision making in education. Outcome data should be combined with non-educational data such as demographics of the student to create a much more enhanced data model. The data model which is represented in this framework includes several types of data including input, process, outcome and satisfaction data. The framework also suggests that the combination of this data model leads to actionable knowledge which is eventually used in decision making. These decisions include assessing progress, addressing individual needs and evaluating effectiveness. These decisions lead to more types of data which can be continuously used in this cyclic framework. Moreover, Bharadwaj and Pal [6], the study shows that the academic performances of the students are not always depending on their own effort. Living location, the medium of teaching, parent's qualification, students other habits, family annual income, and student's family status were highly correlated with the student academic performance.

A similar type of educational data mining research was carried out in [5]. One important factor discussed in the paper was that the collected data should be transformed into a comparable model. The selected domain variables in the research were categorized into defined ranges rather than using the raw value data. For an

instance data like the semester marks, are categorized into first, second, third and fail where each category has a defined range of marks.

The importance of identifying objectives, data and techniques in an educational data mining activity is emphasized in [10]. It also highlights that these data mining objectives are different for various stakeholders in the educational sector. For instance, students may require data mining to enhance their learning experience and get personalized instructions and recommendations. Whereas, the teachers require it to analyze student insights and decide which students require more attention and support.

### **3.1.3 Procedure**

#### **3.1.3.1 Process**

The data collection process refers to how the data is collected from the sample of the population. The data collection process involves various data collection techniques such as questionnaires, in-class activities, interviews etc. During the selection of the methodology, factors such as practicality, the impact of the method on responses needs to be considered. For instance, interviewing would not be practical for a large group of students.

In addition to questionnaires, in-class activities are an alternative methodology for a project of this scale. For instance, in-class activities can be compulsory activities thus providing a complete set of data relating to a sample of the population. On the other hand, the students' enthusiasm will depend on the subject the in-class activity is related with and as a result might not provide a proper image on the sample of the population.

#### **3.1.3.2 Questionnaire Format**

The questionnaires consist of mainly two sections to evaluate the demographics and learning styles of the students. As described in the following sections the questionnaire has been prepared and evaluated for validity. Furthermore, the educational level of the students should be considered when preparing the

questionnaire since it should be properly understood by the students. It is best suited to prepare the questionnaire

### **3.1.3.3 Data Digitization**

Data digitization is one of the major challenges which is involved in educational data collection. Most of the schools maintain written documentation to keep performance data of the students such as term wise examination marks. Since these are mostly handwritten documents technologies such as Optical Character Recognition (OCR) impossible be used for data digitization process. The only available option is to manually enter the collected data into a spreadsheet or a database.

However, the overhead of manual data entry can be eliminated for student learning background and learning style information through the use of a simple software solution. The questionnaire can be developed as a simple software which can directly save the responses from the students to a centralized database. However, this approach may have certain limitations due to technical inabilities and restrictions in certain schools and students. Another advantage of a software-based questionnaire over the written medium is that it can maximize the student interaction with the data collection process.

### **3.1.3.4 Medium**

The medium of data collection describes the mean of educational data collection. The medium of data collection can include mainly digital and physical mediums. The selection of medium of data collection needs a high focus on the familiarity of the sample population with the medium of data collection. A familiar medium of data collection will enhance the confidence of the user and as a result, the participants would provide more information. On the other hand, unfamiliar mediums of data collection will result in the user in providing the minimum amount of data expected by the question.

In the context of local schools, when the number of students is high, it is suitable to go with a physical medium such as the printed version of the questionnaire. It can be

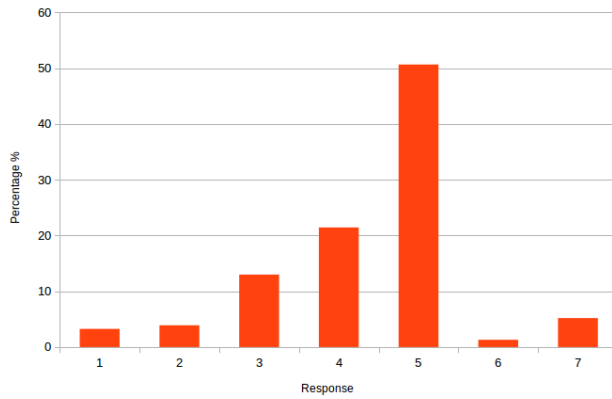
distributed among the student simultaneously and completed quickly within a higher number of students. However, the disadvantage of this method is that it has the added overhead of digitizing collected data.

The digital medium method which was used as a simple software application which consists of the questionnaire. The students can select the answers and submit the responses in the application. This approach reduces the overhead of digitizing data since the student responses are saved in a database. However, the students should have enough computer literacy to interact with a software application to use this approach. Moreover, most of the local schools do not have the required IT infrastructure to facilitate a sufficient number of students.

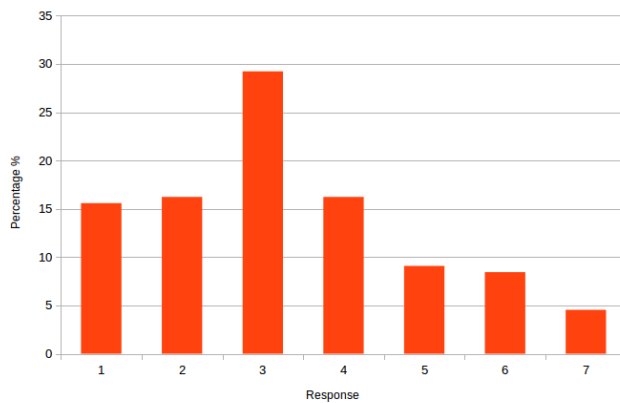
### **3.1.4 Discussion**

#### **3.1.4.1 Learning Style Questionnaire Analysis**

The simplest approach for identifying the skewness in data is mainly analyzing the distribution pattern of the answers to the data. If a certain question has a distribution pattern with a significantly high number of responses corresponding to the same answer option, the question is not identifying a feature that distinguishes each sample. For example, consider the following distributions corresponding to answers provided by students during a survey conducted for the research. Figure 1 Illustrates responses received from a selected sample of students which indicates a significant skewness. On the other hand, Figure 2 illustrates the response distribution corresponding to a set of non-skewed responses for another question on the same set of students.



**Figure 1: Skewed Responses Distribution**



**Figure 2: Non-skewed Responses Distribution**

The suggested approach is highly suitable for a liker scale based questions as the responses are provided from a very limited set of answers. In the case of a written question, a similar approach can be followed through the identification of keywords related to the answer provided by the student. In a deeper level, the associations of the entities can be identified and treated as mentioned previously.

However, when eliminating any data representing skewness based on the answers, the considered data should represent the total population fairly. For instance, the likelihood of a student in preferring practical for learning might depend on the region and the income level. Further, the skewness can be a result of the question being prompted to answer being vague or ambiguous about the answer expected from the student.

In addition to the nature of the question, the medium of surveying also has an impact on the data collection. For instance, a typical student who is comfortable facing

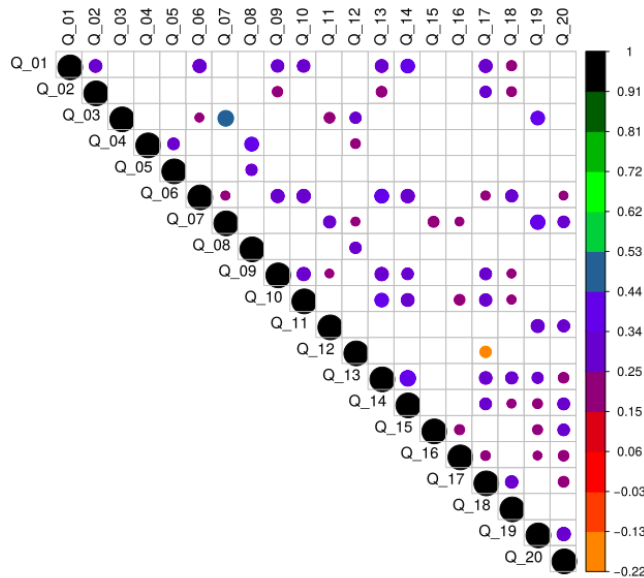
written examinations tends to understand the question more in the case of a written questionnaire rather than a computer-based questionnaire.

Following table 2 indicates the average number of answers provided for extra tuitions and sporting activities participated by the students. A higher number of responses were provided by the students on printed media compared to the computer media. The main reason for the behavior is that the students were more familiar with the written medium than the digital medium. Thus, the nature of the responses can depend on the medium of data collection.

	<b>Written</b>	<b>Computerized</b>
Extra Tuition	3.75	3.5
Sports	1.75	0.2

**Table 3: Average of Answers Option Count**

In addition to the question-related aspects, there also can be questionnaire related aspects such as the association between types of responses between two questions as well. The approach here is to identify such cases from the calculation of the correlation between the responses for the answers provided to the questions. For the purpose, the Pearson's correlation coefficient can be used along with the responses. Figure 3 Illustrates the correlation coefficients calculated from the data collection survey conducted for the research. The empty boxes in the figure represent the p-value of Pearson's correlation coefficient is higher than the significance level of 0.01.



**Figure 3: Pearson's Correlation Coefficient among the Questions**

As the chart illustrates, there is no clear association between questions highlighted by the coefficient of correlation as all the values lies below 0.5 and above -0.2, indicating no significant linear relationship.

Cronbach's Alpha value for this questionnaire is 0.68 which is not high but acceptable reliability and internal consistency. Table 3 represents the parameters related to the internal consistency of the questionnaire.

### 3.1.4.2 Learning Background Questionnaire Analysis

The validity and the reliability of the learning background questionnaire cannot be directly measured from a mathematical procedure since it contains a various different type of questions. The validity of these collected data tends to increase with the higher diversity of the considered samples. The student samples considered for this research includes students from various geographical regions and localities.

Both the male and female genders are considered for the student samples. The gender percentages were equally distributed as represented in the second table. The first table represents the student percentages for various educational backgrounds of the parents. This kind of data will be useful in studying the impact of the educational background of the parents on the academic performances of the student.

	<b>Father Education</b>	<b>Mother Education</b>
Less than O/L	4.25 %	7.98 %
O/L	5.59 %	20.48 %
A/L	47.34 %	42.82 %
University	18.09 %	20.21 %
Unknown	24.73 %	8.51 %

**Table 4: Parent Education Level Percentages**

<b>Gender</b>	<b>Percentages</b>
Male	46.96 %
Female	53.04 %

**Table 5: Gender Percentages**

### **3.1.5 Conclusions**

The major problems that are associated with educational data collection for educational data mining are identified and some possible methodologies are proposed to overcome the issues. Even though the development of an optimal methodology is challenging, the proposed methodology provides proofs of efficiency, comprehensiveness, and completeness for data collection. These facts are highlighted through questionnaires being minimally redundant and internally consistent indicated by the analyses discussed earlier. Further, the proposals of the research can be used for the development of an advanced framework for educational data collection.



## **3.2 Section 2 - Student Performance Prediction**

### **3.2.1 Introduction**

Student marks prediction is one of the major applications of Educational Data Mining (EDM). It allows both students and teachers to identify the students who are on the verge of failing a certain subject prior to the examination. This enables the opportunity for students to adjust their learnings to avoid the failure while the teachers can provide the required instructions and teachings to overcome the situation. This is a classic application of Data Driven Instruction provision in EDM.

Many prior types of research have been carried out regarding this application area of EDM. The main focal point of those researches was enhancing the current data model by combining various data related to the student along with the performance metrics of the student. This mainly includes sociological data related to the student such as their economic status, parent education level, living location status etc. And these researches conclude that choosing these sociological data as features for the prediction model along with performance metrics enhances the overall precision and accuracy of the prediction model.

In prior researches, various prediction models have been tried out such as tree-based models and regressive models. However, most of these researches stuck with a one particular prediction model when obtaining results rather than comparing it with other potential prediction models and selecting the best model which is capable of providing optimum prediction results. Moreover, student performance prediction problem can be approached in the main two different methods.

1. Regression Approach
2. Classification Approach

In the regression approach, regressive prediction models are used to predict the student marks to the nearest whole number. In the classification approach, the student marks are discretized into grading levels and predict the most probable grading level using classification models. However, most of the prior researches

stuck with either one of those approaches with the most favored one being classification approach.

This prediction model is based on over 700 student data [20] which were collected around Sri Lanka covering different demographic regions of the country. The novelty of this research can be addressed under the following main three points.

1. Prediction model consists of both regression and classification models
2. Usage of boosting algorithms and other prediction models with parameter tuning
3. Enhancement of the data model with learning style data, subject correlation data, and assistive learning data

As mentioned above the prior researches were also focused on enhancing the data model by combining additional features such as sociological data of the students to improve the prediction accuracy. In this approach, further features were added which can directly affect performances of the students. One of them is student learning style data where the impact of it on student performance is explained in the study in [16]. There are various defined ways of evaluating learning style of the students. In this approach, a modified version of the LCI (Learning Connections Inventory) model is used to evaluate the learning styles of the students. Further, data regarding assistive learning or tuition participation of the students also considered in this research. Subject correlation data were calculated by performing correlation analysis techniques on the collected performance data for various subjects over a period of three years from the students. Boosting algorithms and ensemble models with optimum parameter tuning were used for prediction model development in both regression and classification approaches.

### **3.2.2 Literature Review**

The research in [11] identifies student performance prediction as a major application of EDM. And it also recognizes students, teachers, and administrators as the parties who are facilitated by EDM. The research is based on predicting student performances in an engineering college in India. The considered student sample is

346 and the decision tree induction is used to predict whether the student is going to pass or fail. The considered data includes past performance data along with several sociological data such as living location condition and parents education level. The developed model predicted the fail students with a good true positive rate of 0.907 while predicting pass students with a higher false positive rate of 0.617 for 10-fold cross-validation evaluation.

The research in [12] based on a model to predict secondary school student performances in Portugal. The considered data includes a relatively high amount of sociological data such as travel time to school and internet access along with past performance data. The prediction is carried out in three different methods; binary (pass or fail), 5-level classification and regression. The results conclude that tree-based algorithms such as decision tree and random forest outperform the nonlinear function methods such as support vector machine (SVM) and neural networks. The research also identifies that the past performance data have a higher impact on student performances than the sociological data.

The research in [13] used a CHAID prediction model to predict academic performances of 772 secondary school students in India. Similar to other researches both past performances and sociological data were considered and the model prediction accuracy was about 44.69%. The chi-square calculation was used to identify the high potential variables for the prediction model.

The research in [15] used an ensemble model to predict student performance and identified that the best results are achieved by SVM. The lack of the number of failed students in the dataset was addressed in the research and it concluded that considering social behavior data improves the prediction accuracy of the model.

The research in [2] used past student performance data in a decision tree model to predict student performances. The examination marks were discretized and categorized into few classes in order to approach the problem in a classification method. The ID3 algorithm was used in the decision tree induction and if-then rules are obtained as the final outcome. Another research in [3] indicates that sociological

factors such as living location and parent education level have a higher potential in influencing student performance.

In summary, all of the above researches are focused on enhancing the data model by combining sociological data with the student performance data in order to improve the prediction accuracy. Following researches focused on identifying the effect of student learning styles for their academic performances.

The study in [19] presents the results of academic performance and learning style self-predicting mechanism for undergraduates of the University of California. The research hypothesis is students who show the strongest self-prediction of individual learning style (VARK score) having the highest course grades. But the results could not significantly show that there was a direct correlation between self-predicting of learning styles and academic performances.

Even though these researches explore the idea of student learning styles and how it correlates with student academic performances, they do not directly use it in a data model to predict the student performances. Our approach is combining both student sociological/demographic data and student learning style data with student academic performances to create a better prediction model using machine learning techniques.

### **3.2.3 Procedure**

#### **3.2.3.1 Prediction Model Development**

Supervised learning refers to the construction of a general hypothesis which is used to make predictions based on reasoning on externally supplied instances.[21] Regression and classification are the main two modeling approaches when developing a prediction model using supervised learning. Regression modeling refers to the approximation of a mapping function from input variables (features) to a continuous output variable. In classification modeling, the mapping function from the set of input features is approximated to discrete output variables. In this research, the output (target) variable is the examination marks obtained by a student. The output variable can be represented in both continuous and discretized formats. In the

Sri Lankan education system, examination marks for a regular term test are provided out of a hundred. Therefore this value can be used as the continuous output/target variable value when training a regression model. And there are also five defined grade categories based on the examination marks obtained by the students.

Grade	Mark Range
A	Above 75
B	65 to 75
C	55 to 65
S	40 to 55
F	Below 40

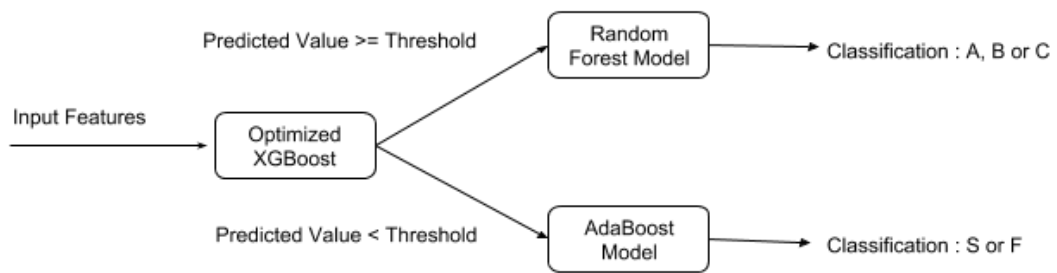
**Table 6: Grade Classes**

This categorization can be used in the output/target variable when training a classification model. In this research, a regression model is incorporated with classification models to generate a hybrid model to improve the overall prediction accuracy.

### **3.2.3.2 Regression and Classification Model Combination**

The final prediction model is constructed using both classification and regression models. It consists of optimized XGBoost regressor with AdaBoost and Random Forest classifiers. The AdaBoost is optimized for classifying lower grades (F, S) with much higher prediction accuracy while Random Forest is better in classifying higher grades (A, B, C). This was analyzed further using confusion matrices in the discussion section. This characteristic was utilized when optimizing classifying model using combined models. The input features for these models were pre-analyzed and categorized in order to classify through one of these models. An optimized XGBoost regression model was used for this categorization of inputs. If

the predicted value of XGBoost model for a given input is positioned in the range of A, B, C grades, that input classified through the Random Forest model whereas other inputs are classified through AdaBoost model. Therefore a threshold (value of 55) is defined for the output of the optimized XGBoost model to select the most optimized classifier to a given set of input features.



**Figure 4 : Final Model Diagram Representation**

### 3.2.4 Experiments

The prediction accuracy of the final model was comparatively analyzed with other classification models which were suggested in related researches in predicting student academic performances. For this, dataset described in the study [20] was used to train the models and evaluate the prediction accuracy with 10-fold cross-validation.

#### 3.2.4.1 Dataset

The chosen dataset consists of a complete set of data with the above features for about 700 students in Sri Lanka. The detailed description about the dataset is provided in the data collection section.

#### 3.2.4.2 Feature Selection

Performance prediction of the students was carried out for each subject separately. Therefore a prediction model is created for each subject separately. The student performances of the other subjects may irrelevant for the performance prediction in a particular subject. Therefore performances of only highly correlated subjects were considered for the performance prediction process. Since data were collected over

three years, nine examination marks (three per year) for each subject is available. These are used as the main basis features for the prediction process and other features are incorporated with it to create a better data model.

When incorporating learning background data to the data model, features like sibling and parent education levels, number of siblings, participation in assistive learning and extra-curricular activities are mainly considered. Learning style data are collected based on the responses student provided for the created questionnaire.

### 3.2.5 Results

<b>Model</b>	<b>Area Under the Receiver Operating Characteristic Curve</b>	<b>Precision</b>	<b>Recall</b>
Random Forest	0.845	0.723	0.719
AdaBoost	0.832	0.717	0.711
Neural Network	0.747	0.634	0.639
Logistic Regression	0.770	0.649	0.651
Naive Bayes	0.733	0.613	0.623
Decision Tree	0.723	0.601	0.612
Final Model	0.881	0.791	0.754

**Table 7: Prediction Results**

According to the above results, the optimized regression-based classifier was capable of outperforming predictions of all the other models. The interesting inspection here

is the prediction accuracies of both Random Forest and AdaBoost classifier models were improved when optimized with the combination of an XGBoost regressor.

### 3.2.6 Discussion

#### 3.2.6.1 Impact of Model Combination

The following results were obtained based on the above experiments carried out in the dataset described in [20]. When the prediction results of the Random Forest model and AdaBoost models are analyzed using confusion matrices it is evident that the Random Forest model is better in predicting higher grades A, B, C, and AdaBoost model is better in predicting lower grade F and S.

		Predicted													
		Random Forest							AdaBoost						
A c t u a l		A	B	C	F	S	$\Sigma$	A	B	C	F	S	$\Sigma$		
	A	158	2	1	2	13	176	113	29	9	5	20	176		
	B	24	38	3	2	16	83	27	29	11	2	14	83		
	C	12	4	27	2	17	62	22	18	16	1	5	62		
	F	17	1	0	29	20	67	1	7	3	42	13	67		
	S	31	5	9	10	62	117	7	9	14	11	76	117		
	$\Sigma$	242	50	40	45	128	505	170	93	53	61	128	505		

**Table 8: Confusion Matrix for Random Forest (Left) and AdaBoost (Right)**

The confusion matrix reveals that the Random Forest model have a tendency in predicting higher grades even for actual lower grades like F and S. 17 out of 67 F grades and 31 out of 117 S grades were classified as A grades in the Random Forest model. This drastically reduces the prediction accuracy and the precision of the model for lower grades. Since the Random Forest model is an ensemble model based on bagging, the same data point may be used to train different subtrees in the model.



Due to the high percentage of higher grades A, B, C (63%) within the dataset, the probability of retraining with a higher graded data point is high. Therefore the above characteristic can be discovered in the Random Forest model. Since the AdaBoost model is based on boosting, the wrongly predicted data points are retrained again. Therefore, the lack of lower grades data points in the dataset affects less for AdaBoost model. However, the combination of these models using a regressive model allows overcoming the above weaknesses of the models. And the XGBoost regressor is also optimized with hyper-parameter tuning to select the best classification model for a given set of input features.

### 3.2.6.2 Incorporation of Learning Style and Assistive Learning Data

One of the novelties of this research is the incorporation of student learning style data and student assistive learning data for the data model. The impact of the incorporation of these features was comparatively analyzed for the regression and classification models.

Model	RMSE with all features	RMSE without learning style	RMSE without assistive learning	RMSE without both features
XGBoost	10.96	13.09	12.86	13.71

**Table 9: Regression Models: Impact of Learning Style Incorporation**

Model	AUC ROC with all features	AUC ROC without learning style	AUC ROC without assistive learning	AUC ROC without both features
-------	---------------------------	--------------------------------	------------------------------------	-------------------------------

AdaBoost	0.829	0.786	0.801	0.723
Random Forest	0.846	0.790	0.798	0.754
Final Model	0.882	0.801	0.817	0.787

**Table 10: Classification Models: Impact of Learning Style Incorporation**

For both regression and classification models, the highest prediction accuracy was obtained when learning style data and assistive learning data are incorporated into the data model. Therefore it is comparatively proven that incorporation of learning style data and assistive learning data improved the prediction performances. The main reason for this could be the impact of the student learning styles towards their academic performances as identified in the study in [16].

### 3.2.7 Conclusions

The methodology addresses the potential machine learning approach for student performance prediction. Many of the related work stuck with one particular approach and comparatively analyze the prediction results with the actual results. But in this research a novel prediction model which consists of XGBoost regressor and Random Forest, AdaBoost classifiers is presented with a comparative analysis. And hyper-parameter tuning was performed on these sub-models to generate the most optimized prediction model.

The data model was also improved by incorporating learning styles data. Previous researches only focused on incorporating demographic data into the data model. But the analysis proved that incorporation of learning style data improved the prediction accuracy. And finally, the combination of regression and classification models also improved the prediction accuracy. The combined model consists of an optimized XGboost model for choosing the optimal classification model based on the threshold. This model manages to overcome the weaknesses of each classification model by providing a set of inputs which are optimum for each model. Currently, the model is trained with over 500 training samples collected in the data collection phase [20].

Model prediction accuracy can be further improved by collecting more data to train the model.

### **3.3 Section 3 - Learning Style Analysis**

#### **3.3.1 Introduction**

Educational data mining (EDM) relates with discovering novel and potentially useful information from large amounts of data comes from educational settings. Student profiling identifies different types of students existing in the student population. Even though there has been significant work on the field of educational data mining, there has been relatively less work has been carried out in the field of student profiling using the learning styles and abilities of the students.

VARAK [22] and LCI [23] models can be considered as the most widely used methods for student profiling based on the learning styles and abilities. However, the key limitation of these models is that they perform poorly when applied to out of domain data. Since the majority of the prior works are performed in developed countries and western countries, there arises a need to build such systems for other parts of the globe.

The research identifies different student profiles existing among Sri Lankan students based on the responses provided to the questionnaire provided. Further, the paper also presents different the implications of the learning profiles for education stakeholders based on the findings of the EDM project carried out. Finally, the research discusses the possible methodologies of improving the findings of the research and the future directions for a similar kind of researches on EDM.

Development of a successful model for modeling student learning preference can assist in educational data mining related researches in numerous means such as allowing the teachers and other stakeholders to identify, assist and customize the teaching methodologies to match with the requirements and capabilities of the students. Further, the modeling of learning preference could assist in mining different association rules and correlations among different aspects of the student performance. For example, association rules can be derived between the type of student learning preference with the performance in different subjects.

The research sample included over 600 grade 9 students with data collected through a questionnaire developed by Bhanuka et al [24] as a part of the research conducted for providing data-driven instructions for Sri Lankan students.

### **3.3.2 Literature Review**

The learning style of an individual can be distinguished as the natural style of the learning process in an individual. According to Keefe [25], the learning style can be described “characteristic cognitive, effective and physiological behaviors that serve as relatively stable indicators of how learners perceive, interact with and respond to the learning environment”. However, the explanation provided by Keefe does not provide with a methodology to identify different types of learning styles and preferences existing in a given student population.

Thus for the purpose of identifying different learning preferences and styles, complete and comprehensive model is required. LCI and VARK models can be considered as the most widely used models for identifying different learning styles. As indicated by Fleming et al [22] elaborating the VARK model, the learning styles of the students can be mainly classified under 4 modalities for learners as visual, auditory, read/write and kinesthetic. According to the model, a student may prefer one or many types of learning modality depending on the environment. In addition, the learning type is dependent on the age category for instance matured students tend to be less kinesthetic since they have been able to develop their auditory and visual skills.

Compared to the VARK model, the LCI model [23] provides a similar form of metrics to evaluate the learning type of the student using student behavior related with learning tasks. However, unlike the VARK model which is concerned more toward perception, the LCI model concentrates on the brain-mind interface. The uses the patterns or filter that exist between the brain-mind interfaces in order to determine the learning profile of an individual. The filters presented under the model can be identified as confluence, technical, sequence, and precision.

In addition to the mentioned methodologies, independent researches have been carried out using different student populations in order to identify different student learning styles and preferences. Many of the research work involved with the identification of different student learning profiles based on academic performance.

However, Bouchet et al [26] have conducted independent research on student profiling based on the interaction of the students with a tutoring platform. The research has been conducted using the og-files, facial expressions, diagrams drawn and notes were taken on paper and eye-tracking data gathered during a tutoring session on Human Circulatory system. The methodology proposed by Buchet et al highly depends on the context in terms of the lesson or subject of the tutoring system as there can be an existence of students in different learning profiles interacting in a similar manner.

Vellido et al. [27] used clustering of multivariate data regarding students' behaviors in a virtual course in order to identify and characterize atypical students (outliers) and to estimate the relevance of available data features. The approach proposed during the research is quite similar to the approach proposed by Bouchet et al, and bounded by the limitation of the context.

Tian et al. [28] used both the learning strategies employed by the students as well as information regarding their personality to cluster them. Their methodology is also in two steps since they validate their clusters definition through an analysis of frequent patterns. Manikandan et al. [29] provided an interesting example of a virtual classroom system grouping students by performance.

However, the key issue considering the applicability of the methodologies is that the high dependency of the context for the purpose of segmentation leading the segmentation methodology to become unaware of features due to similarities that could exist between similar student segments of the population.

### **3.3.3 Experiments**

The dataset gathered through from the study [24] was used for performing the experiments conducted in the section and the experiments are based on the learning style data section of the dataset. The experiment section includes how the features have been selected and improved for performing the student segmentation.

#### **3.3.3.1 Dataset**

The dataset used for the study comprised of the responses gathered for the learning style questions in the study [24]. The learning style questionnaire consisted of 20 questions with Likert scale responses describing the applicability of the actions with the actions of the learner. Each statement or question in the questionnaire described a possible style of study or type of behavior of a student and the aspects covered were inspired by the LCI model [23].

#### **3.3.3.2 Feature Selection**

Prior to the feature selection and improvements, the data had to initially treated for missing values since some of the responses had no responses or marked the question to be not understandable. For the purpose, the missing data points were filled by averaging the responses provided to the questions in the category of the question.

Subsequently, the scores were summed across each category for each student. As the feature selection, the categorical percentage of the students in each category was calculated in order to determine the relative strength of the student in each category compared to the other categories.

#### **3.3.3.3 Methodology**

In order to determine the different student profiles, the categorical percentage scores were clustered under different clustering mechanisms in order to identify the naturally forming learning styles among the students. In the research, hierarchical clustering was performed using the cosine similarity as the affinity measure. Through

the cosine affinity, the students with similar relative strengths are expected to be classified together as required by a naturally forming segment.

### 3.3.4 Discussion

#### 3.3.4.1 Clustering

The clustering of the data set was performed with the intention of maximizing both Dunn index and the Silhouette index. As the initial step, the Dunn indices and Silhouette indices were calculated for a different number of clusters ranging from 1 to 7 clusters.

Number of Clusters	Silhouette Score	Dunn Index
2	0.606	1.451
3	0.256	1.299
4	0.269	1.333
5	0.293	1.056
6	0.289	1.481
7	0.245	1.017

**Table 11: Cluster Evaluation**

As identified by the values obtained for the metrics, the initial number of clusters was determined to be 2 based on the implications. Profiling of the clusters reveals that the minority cluster is having the characteristic of high technical and the confluence with low precision and sequence and the majority cluster to have mixed characteristics.

#### 3.3.4.2 Sub-Clustering

The majority cluster identified in the earlier stage was then further clustered with the intention of identifying segments with notable characteristics. Similar to the previous



stage, the objective of the clustering was to maximize both Dunn index and the Silhouette score.

Number of Clusters	Silhouette Score	Dunn Index
2	0.316	1.451
3	0.282	1.299
4	0.303	1.333
5	0.292	1.056
6	0.245	1.481
7	0.258	1.016

**Table 12: Sub Cluster Evaluation**

As the metrics suggest, the majority cluster was sub-clustered further into two clusters. Upon inspection of the characteristics, one of the sub-clusters had high sequence and precision with low confluence and technicality. The other sub-cluster had mixed characteristics with no notably distinguishable characteristic in the scores.

### **3.3.5 Conclusions**

The hierarchical clustering on the categorical percentage scores of the students revealed the existence of 3 major student segments based in terms of the learning style and preference. The three student groups can be identified as follows,

1. High Technical and confluence with low precision and sequence
2. High precision and sequence with low technical and confluence
3. Mixed characterized students

Compared with the other researches conducted with the purpose of identifying different student segments, the derived student characteristics can be considered as the optimal for the context of the Sri Lankan education system. Further, the results

obtained are independent of the subjects the students follow at the school, thus the findings could be applied across different student samples across the locality.

## **3.4 Section 4 - Impact of Assistive Learning**

### **3.4.1 Introduction**

Private tutoring is a trending topic when it is considering the school level education. This part of the project is conducted to measure the effectiveness or impact of the tutoring also known as assistive learning. Not only Sri Lanka, but many other countries have experienced an increment of tuitions attendance of the students. Reasons like the competition of learning, the hardness of the examinations, lack of individual evaluations in the public schools' teachers, quality of learning experience in public schools, lack of exam directed teaching may have caused this increment of tuition attendance. It doesn't have any clear study that discusses the impact of the tuition and the student can be wasting their time in private classes. This research is going to discuss this problem.

Tuition is also known as 'shadow learning' refers to the teachers who are teaching without following professional standards. Those teachers can have a proper educational background but it is not necessary, if they can express the concepts and theories well, they will be a good tuition master. They assist and support the learning of others in an interactive, purposeful, systematic and efficient way. Tuition can be categorized by the participants count. Some are individual one-to-one basis tuitions, few numbers of students groups and large scale such as more than two hundred students. Sometimes the parents or siblings of the students themselves can be tutors.

It is more challenging to estimate the effectiveness of the private tuition because the tutored and non-tutored students are different. In [30] R. Cole has been given a set of statistics comparing both groups of student. In general, it shows the tutored student get more advantages than the other group. There is some sort of socioeconomic difference between the two groups as well. The parents of the student who having tuition, are somewhat better educated, have higher income and help their student a bit more than the other.

This study will be more helpful for the people who are thinking as tuition is the only option to get educated without considering the public school academic. In some

cases, grade 5 students take about 3 or 4 tuition classes. It is a serious factor that little children do not get their childhood experience, free time, on behalf of a small examination.

### **3.4.2 Literature Review**

There are number of research carried out on measuring the impact of Assistive learning (also known as shadow education or private tutoring). These research are based on different methods. Paper [30] is written about the Sri Lankan context. Rachel Cole used Inter-Sectoral Study on Health and Education dataset, conducted by Sri Lanka's NEC to show the effectiveness of the tuition. In this study, Cole was considering about the children who are in grade 5 and the performance of scholarship examinations with a various number of other factors like family background, teachers education levels, time and money spent on tuition by the students, the scale of the tuition classes. This paper uses bivariate regressions and multivariate regressions to estimate the impact. They conclude the study as the impact of the tuition is not statistically significant for a small duration of tutoring such as five months but a significant impact for a longer duration of teaching.

The research [32] is done in Pakistan region considering only mathematics marks of the students. Qaiser Suleman and Ishtiaq Hussain in [32] have tested the impact of the tuition by using fifty students from a school in Pakistan and used only 4 topics of the mathematics syllabus, there for the research is not seems to be a more generic experiment. Using experimental and control group of students it conducted three tests called pre, post and retention tests. To get the impact of tutoring they perform their own tuition sessions only for the experimental group by using some experienced teachers. Then the hypothesis is proved by the results of these test results.

In research [31], Pallegedara has done a survey for evaluate the demand of the private tutoring in Sri Lanka and experienced that the private tuition expenditure has change form luxury good in 1995/96 to a necessity good in 2006/07. Smyth did [33] research, by using the data from a survey with school leavers. Their conclusion was

there is no significant difference between the two groups of tuition takers and non-takers.

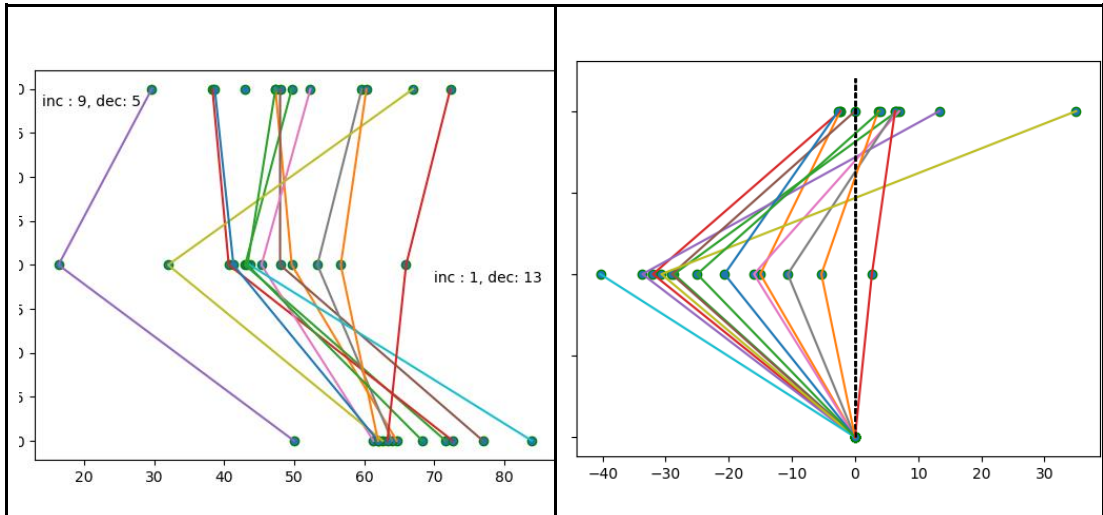
### 3.4.3 Experiments

The objectives of this study are to explore the impact of the tuition on the academic performance of the student who is in grade 6, 7, 8. This study is focused on different subjects like mathematics, science, and English.

In order to measure the impact of assistive learning on student performance, the average marks are taken for all three years for each student for the subject in consideration. A simple plot as illustrated below can provide insights into the variation of the performance with assistive learning. The marks variation plot would provide the variation of the scores in each year whereas the differential plot would provide how the marks have been varying relative to the benchmarking year of grade 6.

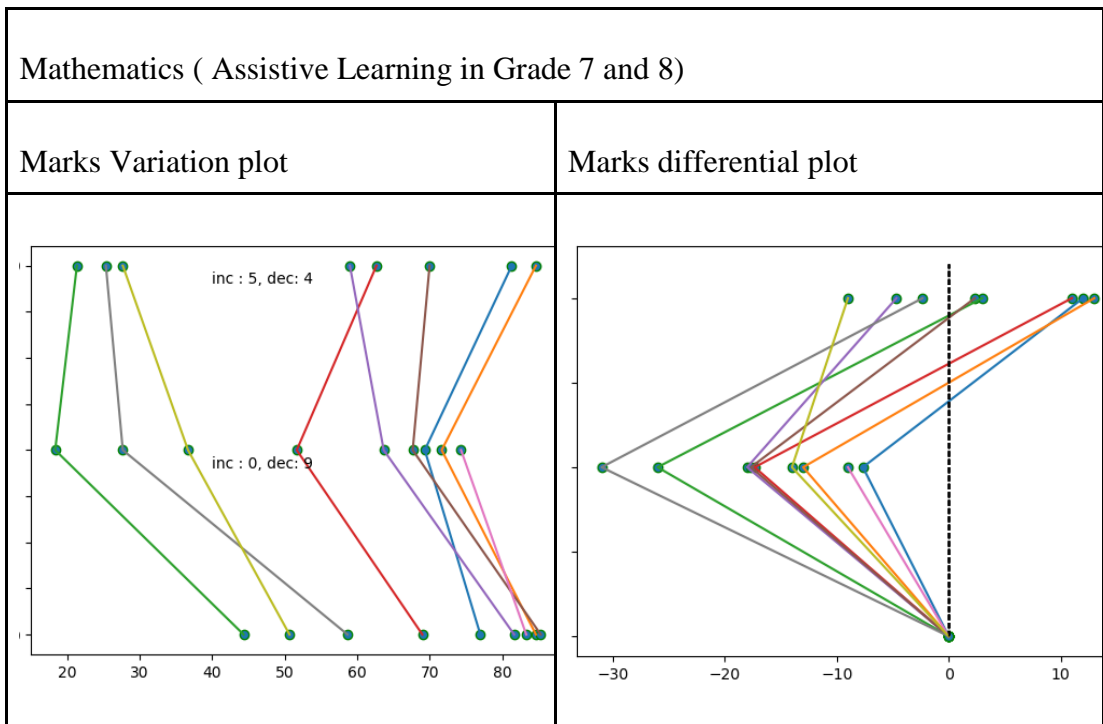
For instance, the following plots illustrate the impact of assistive learning for students who took tuition in grade 8 but not in grade 6 or 7. The marks differential plot would indicate the variation in terms of increments or decrements. It can be clearly observed that the students who took tuition in grade 8 have a tendency of improving the performance.

Mathematics (Assistive Learning only in Grade 8)	
Marks variation plot	Marks differential plot



**Table 13: Marks Variation Plot due to Assistive Learning – Maths, Grade 8**

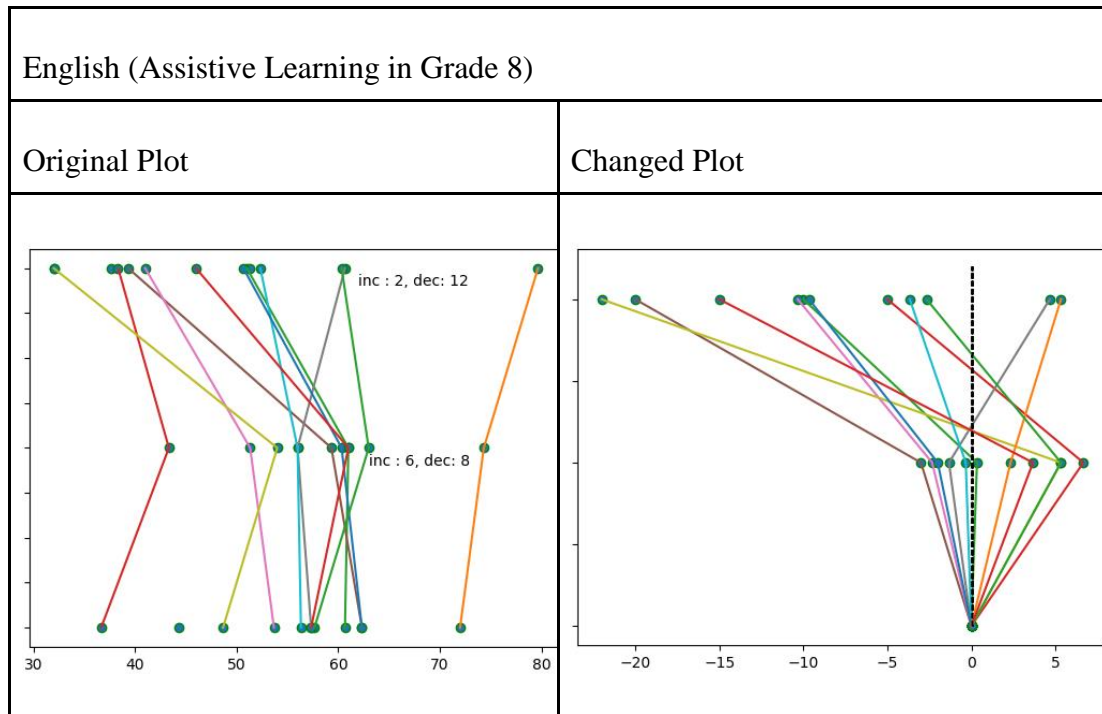
Similar to the graphs illustrated above, following graphs indicate the performance variation of for mathematics of the students who started following assistive learning in grade 7 and 8. It can be noted that the performance has improved in grade 8 despite the drop in grade 7.



**Table 14: Marks Variation Plot due to Assistive Learning – Maths, Grade 7, 8**

However, in the case of subjects such as English, a clear relationship cannot be distinguished between assistive learning and performance of the student. The

following graphical illustrations can be used to indicate the non-existence of such a relationship.



**Table 15: Marks Variation Plot due to Assistive Learning –English, Grade 8**

As a further analysis, a statistical experiment method is used to measure the effectiveness. In this method, T-Test is used to compare the average marks of the two student groups which take tuitions on a particular subject or not taking tuition.

### 3.4.3.1 Hypotheses of the Study

The followings are the three null hypotheses and alternative hypotheses were developed for the test the above mention objectives.

#### 1. For English

H0 - There is a significant difference between the group which attends to the tuition and the group which does not attend to the tuition for English.

H1 - There is no a significant difference between the group which attends to the tuition and the group which does not attend to the tuition for English.

## 2. For Mathematics

H0 - There is a significant difference between the group which attends to the tuition and the group which does not attend to the tuition for mathematics.

H1 - There is no a significant difference between the group which attends to the tuition and the group which does not attend to the tuition for mathematics.

## 3. For Science

H0 - There is a significant difference between the group which attends to the tuition and the group which does not attend to the tuition for mathematics.

H1 - There is a no significant difference between the group which attends to the tuition and the group which does not attend to the tuition for mathematics.

### **3.4.3.2 Significant level**

5% (0.05)

Because of the hypothesis are considering both increasing and decreasing events, this test can be taken as a two-tailed test. Therefore the significant interval is reduced to 0.025.

### **3.4.3.3 Population**

All student in grade 6, 7 and 8 in Sri Lanka.

### **3.4.3.4 Sampling and Sampling Technique**

More than 500 grade 9 students from different demographic areas in Sri Lanka. The students are categorized into two groups by the tuition attendance for that particular subject. (In the result section they are called as 'A' and 'B')



### 3.4.3.5 Delimitation of the Study

This experiment measures only whether a student has attended the tuition for the subject or not in a particular year. How much time they spend on tuition is not counted.

### 3.4.3.6 Data Collection

The data is collected by using the survey type application and the printed form. The student examination marks were directly taken from the schools as hard copy and then digitized. Statistical tools i.e. mean, standard deviation and differences of means were calculated for both groups of the student who attends to tuition and others.

### 3.4.4 Results and Discussion

A - Those who did not go to tuition on that year

B - Those who went to tuition on that year

#### 3.4.4.1 English

year	group	count	mean	std	T value	P value
6	A	107	68.34959	18.24401	2.04458	0.04164
	B	247	72.5356	17.44621		
7	A	83	65.05223	21.31249	1.13019	0.25916
	B	272	68.07446	21.32852		
8	A	62	67.03399	20.70118	1.47254	0.14177

	B	293	70.79944	17.74764		
--	---	-----	----------	----------	--	--

**Table 16: T-Test Results Comparing Two Groups for each Year**

For all above three cases, the magnitudes of t-value are smaller which is statistically significant as it is greater than the significant level of 0.025. Hence the null hypothesis that “There is a significant difference between the group which attends to the tuition and the group which does not attend to the tuition for English” is rejected. It clearly indicates that the students who are attending tuition and who are not are not significantly different.

#### 3.4.4.2 Mathematics

year	group	count	mean	std	T value	P value
6	A	100	65.53928	17.72656	1.7172	0.08682
	B	255	69.12692	17.69938		
7	A	86	63.61661	19.32385	0.77347	0.43976
	B	269	65.32044	17.26505		
8	A	54	71.03886	18.69922	-0.1396	0.88906
	B	301	70.68977	16.58736		

**Table 17: T-Test Results Comparing Two Group for each Year**

For all above three cases, the magnitudes of t-value are smaller which is statistically significant as it is greater than the significant level of 0.025. Hence the null hypothesis that “There is a significant difference between the group which attends to the tuition and the group which does not attend to the tuition for mathematics” is

rejected. It clearly indicates that the students who are attending tuition and who are not are not significantly different.

### 3.4.4.3 Science

year	group	count	mean	std	T value	P value
6	A	130	71.80148	16.38461	3.51893	0.00049
	B	225	77.96044	15.59344		
7	A	96	71.34829	15.88555	2.16568	0.03101
	B	258	75.32167	15.14238		
8	A	68	71.18545	15.664	0.65782	0.51108
	B	287	72.526	14.97681		

**Table 18: T-Test Results Comparing Two Groups for each Year**

In the last two cases except the first one, the magnitudes of t-value are smaller which is statistically significant as it is greater than the significant level of 0.025. Hence the null hypothesis that “There is a significant difference between the group which attends to the tuition and the group which does not attend to the tuition for Science” is rejected. It clearly indicates that the students who are attending tuition and who are not are not significantly different. For the first case, the null hypothesis cannot be rejected and the alternative one is accepted. Therefore in grade 6, those who are attending tuition have more probability to take different marks than the others.

### **3.4.5 Conclusions**

Considering all three cases, the experiment conclusion is there is no statistically significant difference of performance between the two groups who having tuition or not.

## **3.5 Section 5 - Subject Level Relationships**

### **3.5.1 Introduction**

Exploring an association between subjects is one of the major research areas in educational data mining. Correlation analysis is widely used to explore an association between subjects. This analysis uses a holistic approach to identify the correlation among subjects. A holistic approach means thinking about the big picture. In educational perspective holistic approach refers to the teaching method that without avoiding excluding any significant aspects of the human experience [1]. Therefore this research is introduced subject correlation analysis on all subjects without excluding any to explore a better way to improve learning experience on students in secondary school level. Among the main challenges in a holistic approach for the number of dimensions is high due to the number of subjects are increased. Moreover, it's hard to cluster the subjects properly without having a high number of student sample. The current study useful to alter the educational policies to achieve the best learning experience to the students. Moreover, this study useful to predict the student marks in the future. Even though there has been a significant level of work relating to subject level correlations [2], there is only limited work which focuses on the analysis with a holistic approach. Majority of the researches focuses on analyzing the correlation between several specifically selected subjects. Since an academic term (4 months period) can contain any number of subjects, there can be influences from other subjects are not taken for consideration on the information taken for the study. Thus, these studies and discoveries might not reflect the true nature of the sample of the study.

The paper presents a framework and a methodology for analyzing correlations between different subjects with a holistic approach in order to gather insights from the data. The research addresses the identification of patterns using basic data mining approaches. Compared to other researches which are confined to a part of syllabus [2], [3] and [4], the proposed approach of the paper would provide an effective, comprehensive and complete methodology. Further, the learnings of the study can

provide useful insights into student performance analysis using a holistic approach which can be applied universally in order to perform a similar analysis.

### **3.5.2 Literature Review**

Various researches have been carried out in the area of correlation analysis between different subjects. The main purpose of these researches was the identification of the impact of a specific subject towards the student's performance of another subject. These researches consist with different sample sizes and different age ranges of the students. In addition, each research focuses on analyzing the correlation between several specifically selected subjects, contrary to this research which considers the correlation between all the subject in the curriculum, in order to identify the highly correlated subjects and categories.

One of the highly focused research problems in this domain is correlation analysis between science and language skills. This problem is addressed in [3] and [4] under different student samples and circumstances. [3] and [4] both suggest that relevant language skills are required to make necessary inferences in a subject like science. Hence the language skills directly affect the student's performance in science. Moreover [3] propose that having the knowledge available does not guarantee that the student will use the knowledge; in order to effectively use the knowledge and provide proper answers in an examination situation students require proper language skills

Another research problem in this domain is analyzing the correlation between science and mathematics. [5] addresses this problem for the students in one particular age group (grade 8) whereas [4] addresses the problem for an extended range of ages. [5] concludes that there is a linear correlation between mathematics and science performances of the students in the TIMSS and TIMMS-R datasets which were considered in the research. [4] has extended the research carried out in [5] by considering the language and reading skills of the students as well. It also considered the performance over several years of the students and calculated the correlation coefficients for each instance. It concludes that reading (language) achievement

significantly mediate the relationship between mathematics and science across time. Similar research is carried out in [6] which concludes that reading achievement had a larger effect on science achievement than mathematics achievement. More similar researches were carried out in [2] and [8] too.

[8] is the only research which analyzed the correlation between multiple subjects. It concludes that students with higher performances and lower performances have higher correlation across all the subjects and students with moderate performances have inconsistent correlations. However, none of these researches deeply analyze the correlation between all the subjects in the curriculum and provide a framework to do so. The next section addresses the key steps of this framework in detail.

### 3.5.3 Experiment

The data collection process was carried out in several schools in Sri Lanka covering over 700 students from urban, suburban and rural regions of the country []. The student sample consists of different demographics with approximately equal percentages of male and female students.

All the data was collected from grade 9 students and it includes performance data of the students over 3 years. Hence end term examinations marks for all the subjects were collected for the considered student sample. Reliability measurement for the collected sample has a higher value of 0.98 which implies the sample has a high internal consistency. Table 1 provides the statistical measurements of the collected sample.

	Grade 6		Grade 7		Grade 8	
Subject	mean	sd	mean	sd	mean	sd
Mathematics	68	20	57	25	60	26
Science	71	19	65	24	69	22

Sinhala	69	19	70	22	69	24
English Language	66	22	63	24	59	26
Religion	74	19	65	24	69	22
History	65	20	69	27	67	23
Health	71	18	70	23	69	21
Citizenship Education	75	19	76	22	78	21
Geography	77	18	64	22	69	25
PTS(Practical Technical Studies)	55	33	65	27	71	23
Second Language	70	23	60	28	53	24
Art	74	18	64	23	60	25

**Table 19: Descriptive Statistics for Subjects by Year**

The performance of the student in a specific subject in a specific grade corresponds to the scores obtained at the term examinations of the grade. The process of identifying similar subjects is mainly carried by considering two approaches. The first approach is based on the average term marks for every academic year for each student and the other is using the first principal component corresponding to the performance score vector. Although principal component analytics seems to be a good solution to dimensions reduction the explained variance for each subject differs from 0.6 to 0.8. Therefore the first approach which average term marks for every academic year selected.



### 3.5.4 Discussion

The discussion consists of the main two part. One is correlation analysis on subjects and other is hierarchical cluster dendrogram analysis. Correlation analysis show year by year analysis of subjects and hierarchical clustering show overall analysis of subjects.

#### 3.5.4.1 Correlation Analysis

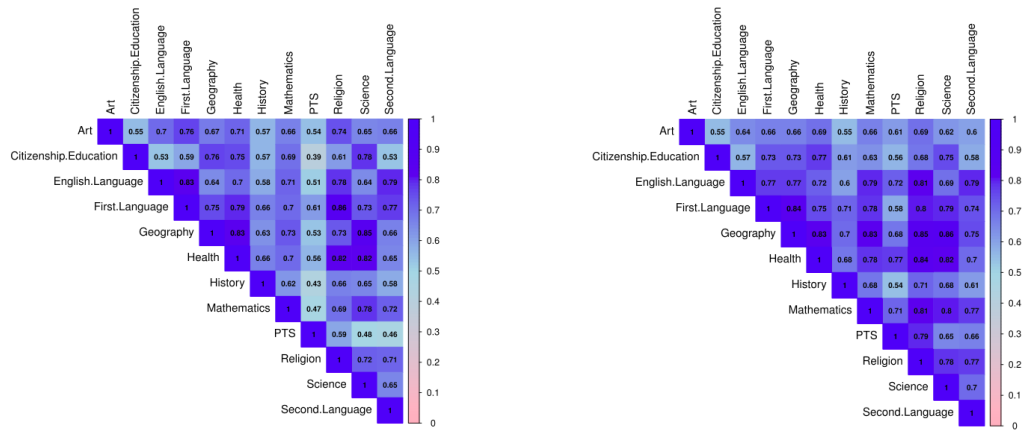


Figure 5: Pearson Correlation among all the Subjects for Grade 6 and 7

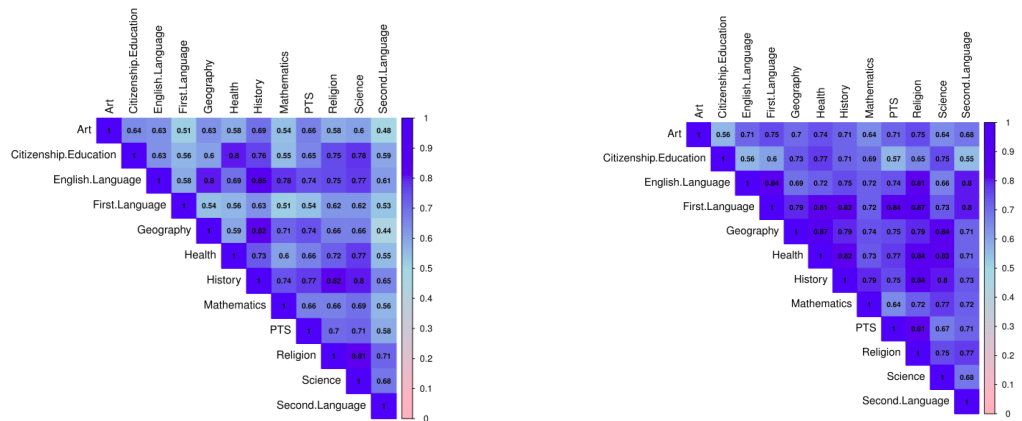


Figure 6: Pearson Correlation among all the Subjects for Grade 8 and Spearman Correlation among all Subjects for Grade 6

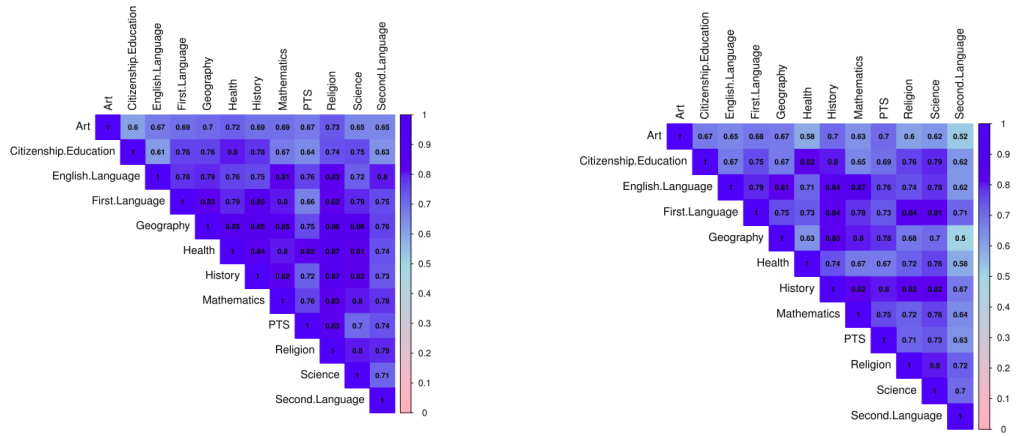


Figure 7: Spearman Correlation among all the Subjects for Grade 7 and 8

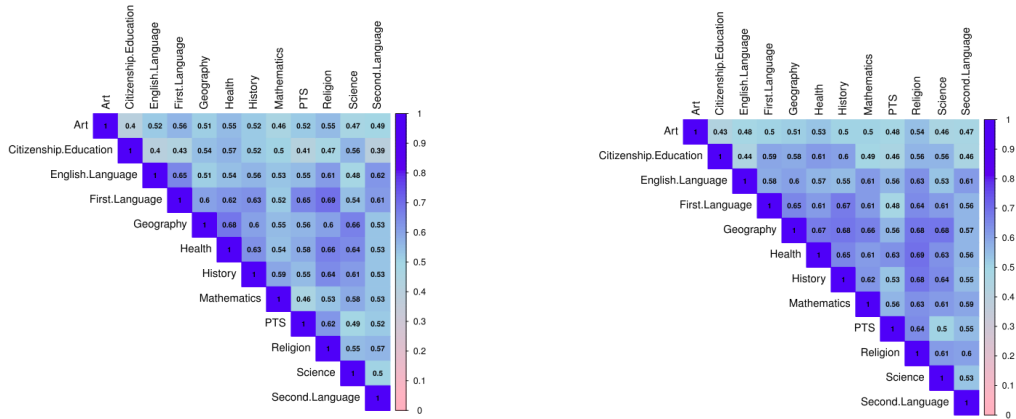
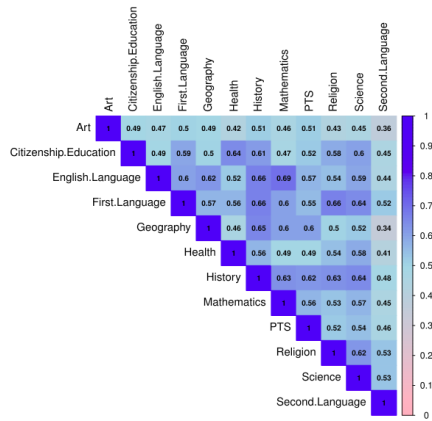


Figure 8: Kendall Correlation among all the Subjects for Grade 6 and 7



**Figure 9: Kendall Correlation among all the Subjects for Grade 8**

Under correlation analysis, we perform basic three correlation techniques. Kendall, Pearson, Spearman. In our results, Spearman and Pearson show higher correlation scores to subjects and Kendall show medium scores to the subjects (Fig. 1-9). But all three methods are showing a similar distribution of correlation among subjects. Specially Spearman and Kendall methods are showing similar distributions. As an example, Spearman and Kendall show similar higher correlated subject pairs in table 2. while Pearson shows a subset of that pairs.

Method	Highly correlated subject pairs (coefficient up to X)	X
Pearson	First Language - English Language Geography-Health Religion - Health Science - Health Science - Geography	0.8
Spearman	First Language - English Language Geography-Health History-First Language Health - History First Language - PTS Religion - Health Religion - History Science - Health Science - Geography	0.81

Kendall	First Language - English Language Geography-Health First Language - PTS Religion - Health Health - History Science - Health Science - Geography	0.63
---------	---	------

**Table 20: Higher Correlated Subjects in Grade 6**

The table 2 shows that Spearman's and Kendall's methods identified more relationships between subjects which Pearson method did not identify. This explains subjects do not have a completely linear relationship but a monotonic relationship. Table 3 shows the comparison of Spearman correlation coefficients of grade 6, 7 and 8.

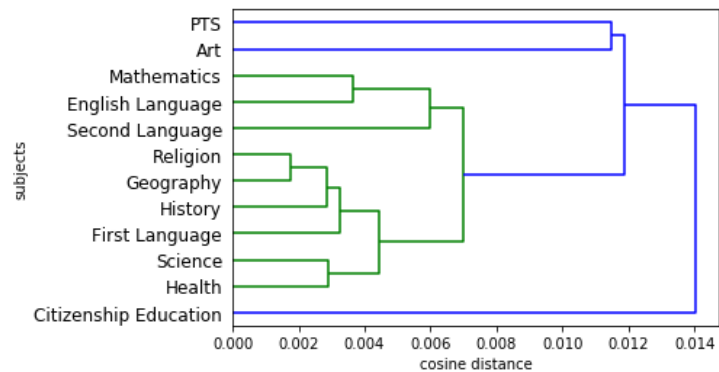
	Grade 6	Grade 7	Grade 8
Highest correlation	Religion - First Language	Religion - Health, Religion - History	Mathematics - English
Lowest correlation	CE vs Second Language	CE - Art	Geography - Second Language
Range	$0.87 - 0.55 = 0.32$	$0.87 - 0.60 = 0.27$	$0.87 - 0.50 = 0.37$
Special relations	Citizenship Education, Art, Mathematics compared to others showing low correlation with other subjects.	Art compared to others showing low correlation with other subjects.	Second Language, Art compared to others showing low correlation with other subjects.
Highly correlated subjects	First Language - Foreign Language Geography-Health History-First Language Health - History First Language - PTS Religion - Health Religion - History Science - Health	Geography - First Language Health - Geography First Language - History Geography - History History - Health Geography - Mathematics Mathematics - History PTS - Health	CE - Health History - English First Language - History History - Geography Mathematics - History Religion - History Religion - First Language History - Science

	Science - Geography	Religion - First Language	
		Religion - Foreign Language	
		Religion - Geography	
		Religion - PTS	
		Religion - Mathematics	
		Science - Geography	
		Science - History	

**Table 21: Comparison of Spearman correlation coefficients in grade 6, 7 and 8**

### 3.5.4.2 Hierarchical Cluster analysis

Hierarchical clustering is applied for identification of the similar subjects in the syllabus. For the purpose of identification of similar subjects, each subject is represented by a vector of 12 dimensions. A value corresponding to each dimension of the representation is calculated by considering the Pearson's correlation coefficient of each subject against other subjects. Cosine similarity is used for the calculation of the similarity between the subjects. The dendrograms obtained by the approach will provide an overview of the grouping of the subjects. Due to the use of the cosine similarity, subjects in the same cluster will require similar skill sets.



**Figure 10: Hierarchical Cluster Dendrogram for Pearson's Correlation Matrix with Cosine Distance & Average Cluster Method**

Other than PTS, Art and Citizenship Education which having larger distance with other subjects hierarchical cluster dendrogram identified clearly two clusters among subjects. In figure 10 mathematics and English language in the same cluster which have the most failure rates in ordinary level exam [9]. Moreover in [10] states that

mathematics and foreign languages associates with each other. Furthermore in [11] explains mathematics can be thought of like a foreign language, with its own unique terminology and symbol system. Therefore the results in figure 10 confirm the students in Sri Lanka is also show a similar attitude towards secondary languages and mathematics.

Religion, Geography, History, and Sinhala have low distances with each other subjects and all these subjects are related to reading and memorizing. The next closest distances with the above subjects are Science and Health. Which is also related to reading skills as described in the literature. Cluster dendrogram confirms that with science has the closest distance with the above-identified reading cluster.

### **3.5.5 Conclusions**

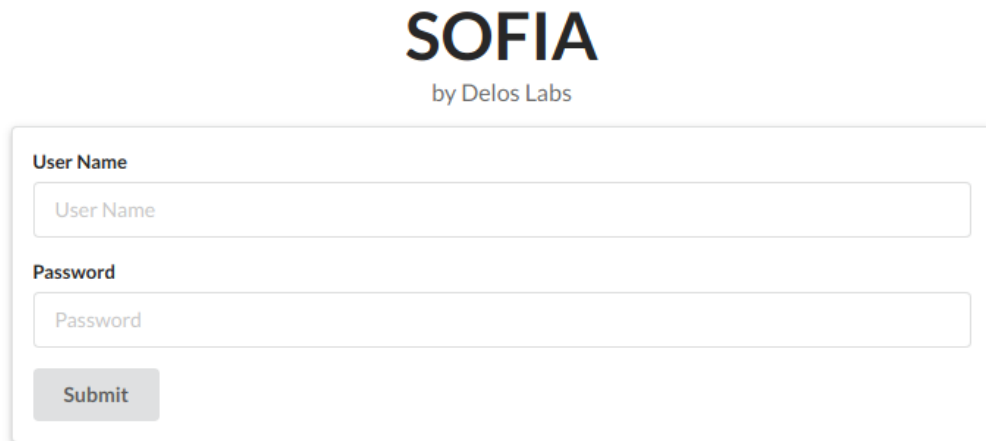
Present findings are consistent with the existing literature that supports high associations between reading and science achievement [3], [5] and mathematics and foreign languages achievements [10], [11]. Moreover reading and science has a higher relationship than the relationship between science and mathematics is consistent with the existing literature [6], [7]. According to our analysis science and mathematics have a moderate association. But the current analysis did not include data beyond eighth grade. Therefore mathematics and science may have a higher association between them in grades beyond eight.

In comparing different correlation techniques, the Spearman's correlation is more suitable for the data set because the Spearman correlation coefficient is identified more relationships between the subjects rather than Pearson's correlation coefficient.

In hierarchical cluster dendrogram Citizenship education, Art subjects and PTS have higher distance with other subjects. Therefore these subjects do not have a high impact on other subjects and they may solely evaluate students' born skills.

## 4 DATA DRIVEN INSTRUCTION STRATEGY PLATFORM - SOFIA

Based on the results of the identified research problems, a software tool is developed mainly targeting the two key stakeholders of Sri Lankan education; teachers and students. Sofia is the digital platform which provides useful insights and instructions to both teachers and students in order to enhance the individual learning and teaching experience of a classroom. A user can be a teacher or a student where each type of users are provided with different interfaces and functionalities in order to provide various insights based on the collected data.



The image shows the login interface for the SOFIA platform. At the top, the word "SOFIA" is displayed in a large, bold, black font, with "by Delos Labs" in a smaller, grey font underneath. Below the logo is a white rectangular form with a thin grey border. Inside the form, there are two input fields: the first is labeled "User Name" and contains the placeholder text "User Name"; the second is labeled "Password" and contains the placeholder text "Password". Below these fields is a grey "Submit" button.

**Figure 11: Sofia Login Interface**

### 4.1 Student Interface

Unique learning instructions are provided to the students based on the entered data. The initial step of a student is to complete their profile by filling up the required data field for the application. These data fields consist of learning background data, participation in extra-curricular activities and assistive learning, a questionnaire to evaluate learning style of the students and student performance data for all the subjects for grade 6, 7 and 8.

**Subject** 1 2 Sin Eng Tam

What are your favourite subjects?  
 Sinhala  Maths  Science  Art  Dance  Buddhism  History  Tamil  English  Geology

What are your favourite lessons?  
 What are your favourite lessons?

What are the hardest subjects for you?  
 Sinhala  Maths  Science  Art  Dance  Buddhism  History  Tamil  English  Geology

What are the hardest lessons for you?  
 What are the hardest lessons for you?

What is your ambition?  
 What is your ambition?

What is your grade 5 scholarship marks?  
 What is your grade 5 scholarship marks?

**Next**

---

**Subject** 1 2 Sin Eng Tam

For which subjects do/ did you participate tuition classes?  
 Sinhala  Maths  Science  Art  Dance  Buddhism  History  Tamil  English  Geology

Additional

	Grade 6	Grade 7	Grade 8
Science	✓	✓	✓
Dance	✓	✓	✓

**Finish**

---

Footer

### Term Test Marks

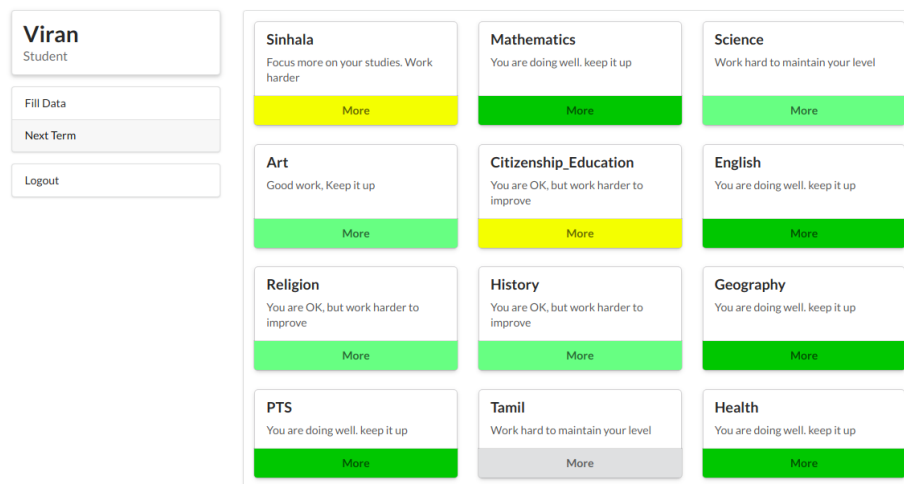
Subject	Grade 6			Grade 7			Grade 8	
	I	II	III	I	II	III	I	II
Sinhala	90	90	90	90	5	09	8	6
Mathematics	75	87	98	96	95	75	88	98
Science	98	88	97	92	65	35	67	80
Art	18	58	55	87	84	6	88	48
Citizenship_Education	45	45	46	45	41	65	45	65
English	80	90	95	99	97	96	80	90
Religion	15	65	48	65	53	46	48	68
History	25	65	36	26	21	65	9	69
Geography	75	80	85	70	79	80	88	85
PTS	98	98	99	89	98	92	97	93
Tamil	25	68	56	69	69	59	25	78
Health	75	84	56	58	68	69	45	87

**SUBMIT** **Fill**

**Figure 12: Student Data Collection Interface**



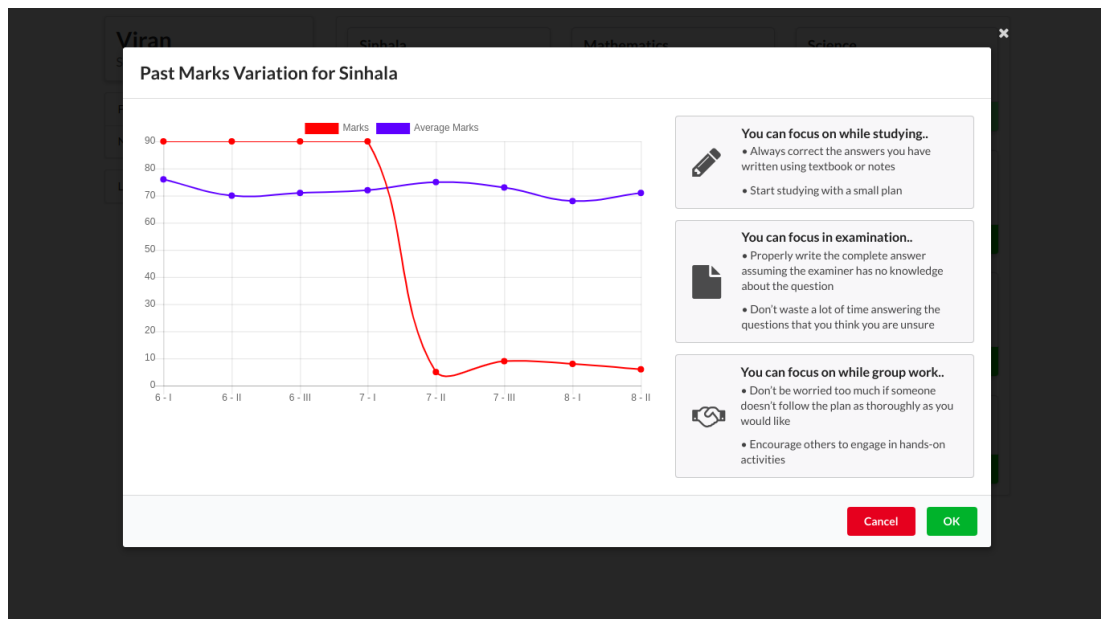
The collected data are used in analytics to infer results from them and to train the machine learning models developed for student academic prediction. The machine learning model is capable of taking provided data as input and predict the upcoming grades for a given subject. If the grade is predicted to be going down, learning instructions are provided to the student based on the learning style. Since there are mainly three types of students according to the learning style-based student segmentation research outcome, the instructions are provided separately for these three student types; methodological, experimental and composite.



localhost:4200/student/std-marks

**Figure 13: Student Subject Overview**

The above interface provides an overview of student's current situation for all the subjects. Here the student's upcoming examination mark for each subject is predicted but it is not displayed to the student directly since that can affect their mentality. Instead, general instructions are provided to work hard or focus more on subjects which are predicted to attain lower grades. (E.g. Sinhala and Citizenship Education in Fig 6.2)

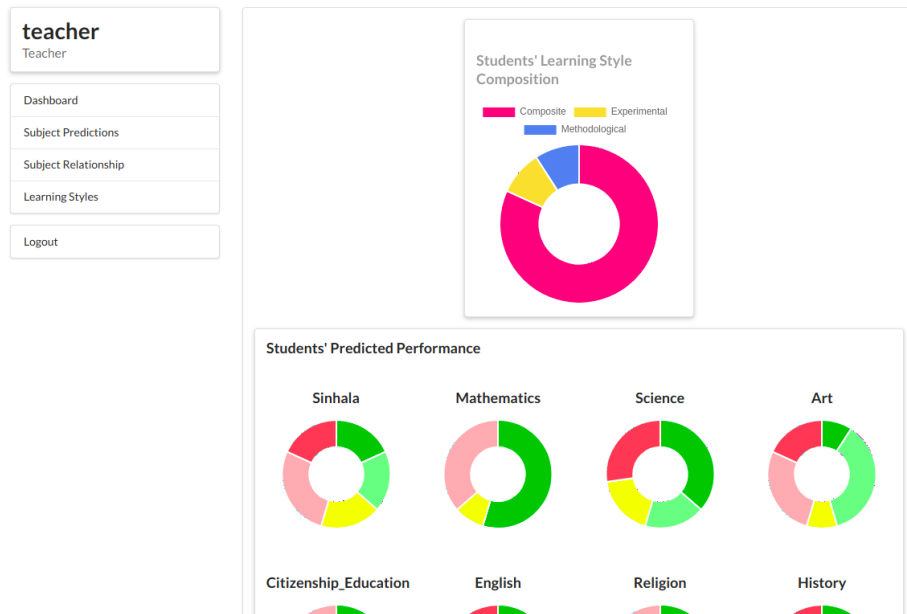


**Figure 14: Student Subject Specific View with Learning Instructions**

This is the interface provided for students to get a more detailed view of a selected subject. The graph shows the marks variation of the student (Red) along with the average mark for that subject by other students. (Blue) That would provide a comparative view for the student to assess where his/her mark is located compared to other students. The right side of the view the instructions that the student have to follow in order to overcome the situation. The instructions are provided under three main scenarios. They are instructions to be followed while studying, during examinations and during group work. And as mentioned above, these instructions are provided based on the learning style of the student in order to obtain the optimum effect of following the instructions properly.

## 4.2 Teacher Interface

An overview of the students in a classroom is provided to the teachers. The view consists of learning style composition of the classroom along with the predicted grade percentages for each of the subjects.



**Figure 15: Teacher Dashboard**

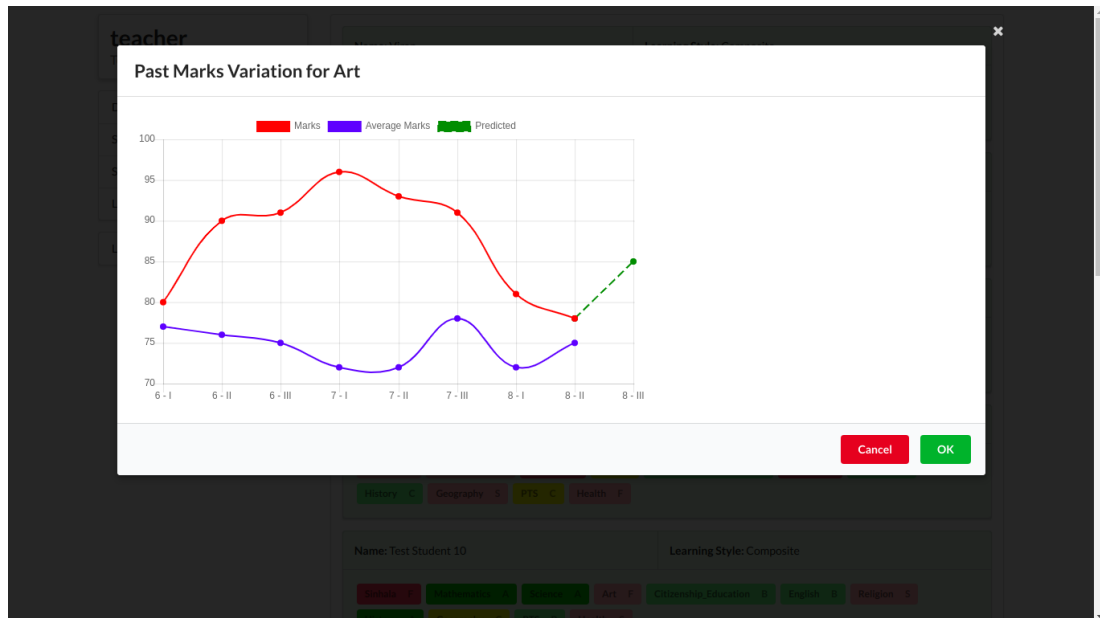
According to the above Fig 6.4, for science, a higher failure rate is expected which is about 30%. Then, the teacher can provide more attention to such subjects in order to overcome the higher failure rates. The teacher also can check the status of the students in a classroom separately as well.



**Figure 16: Individual Student View**

In this view, the teacher can identify the students who are in danger of failing a subject in the upcoming examination. According to Fig 6.5, Student 4 is in danger of

failing Science and English subjects. Therefore the teacher can pay more attention for these students, especially on those subjects. Further, the teacher can also get the exactly predicted mark for a student.



**Figure 17: Student's Grade View**

Here the green dotted line indicates the expected grade for the students for the selected subject in the next examination. This view is only visible to the teachers.

Further, an overview of student profiling based on learning styles is also provided to the teachers. Here, the main characteristics of the students in each student segments along with suitable instruction types are provided for the teachers.

**teacher**  
Teacher

---

Dashboard

---

Subject Predictions

---

Subject Relationship

---

Learning Styles

---

Logout

Characteristics    Instructions

Type	Characteristics
Methodological	<ul style="list-style-type: none"> <li>Working to a plan</li> <li>Follows examples</li> <li>Cares about neatness</li> <li>Explore information</li> <li>Cares about constant feedback</li> </ul>
Experimental	<ul style="list-style-type: none"> <li>Likes to perform hands-on activities</li> <li>Likes to experiments</li> <li>Willing to take risks</li> <li>Performs well independently</li> </ul>
Mixed	<ul style="list-style-type: none"> <li>Combination of these characteristics</li> </ul>

**teacher**  
Teacher

---

Dashboard

---

Subject Predictions

---

Subject Relationship

---

Learning Styles

---

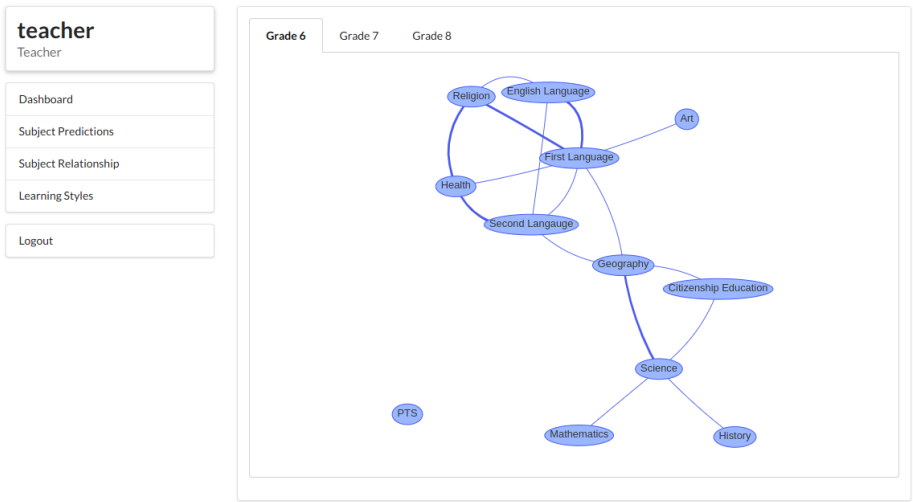
Logout

Characteristics    Instructions

Type	Experimental		Methodological	
Studying	Studying alone may be much more effective than group studying	Study by answering questions in the textbook and assignments	Start studying with a small plan	Refer to external sources like books, papers, and internet when studying
	Explore practical ways to study where you can perform hands-on learning	Try brainstorming the content that you are planning to study	Always follow the notes and textbook in the order of syllabus	Learn by making questions from the lessons and answering them by yourself
	Focus on writing answers to questions while studying	Always correct the answers you have written using textbook or notes	Follow example questions when studying before you try answering	Do not worry too much about certain theorems and how they are proved. Only the application is important
	Take a short break periodically while studying for a long time	Take down short notes in point form while studying	Do not worry too much about neatness when studying	Only focus study on the given scope of the syllabus. Do not overanalyze external sources.
		Revise your studies once in a while	Practice to do questions on time	
			Do not consume a lot of time when making a study plan	
	Properly write the complete answer assuming the	Use the scrap paper for	If you are struggling to answer a question,	Don't try to provide the long descriptive

**Figure 18: Student Learning Styles Overview**

And finally, a subject correlation map is provided to the teachers, based on the research outcomes of subject-level relationship analysis. This might be useful for teachers to identify highly correlated subjects in the curriculum.



**Figure 19: Subject Correlation Map**

## 5 CONCLUSIONS

This research project was carried out under five main research areas. They are,

1. Student Performance Prediction
2. Learning Style Analysis
3. Impact of Assistive Learning
4. Subject Level Relationships

Each research area specifies a research problem and proposes a methodology to find the solution and finally, the research outcomes are analyzed and compared with existing or related methodologies in order to validate. Other than these research areas, the data collection process also played an important role in this project.

The proposed student academic performance prediction model was capable of producing the best prediction accuracy for the considered dataset compared to other prediction models. The learning style analytics results in a complete methodology to profile students into three main identified student groups based on their learning style preferences. The statistical testing used in determining the effect of assistive learnings for the secondary education suggests that a significant impact does not exist in the Sri Lanka educational context. The holistic approach to subject relationship analysis gives consistent results with existing literature which high association between science and reading achievements and mathematics and foreign languages achievements. Moreover, when comparing different correlation techniques, the Spearman's correlation is more suitable for the data set because the Spearman correlation coefficient is identified more relationships between the subjects rather than Pearson's correlation coefficient.

And finally, these research outcomes are integrated and used in the development process of Data Driven Instruction Strategy Platform - Sofia. The platform was mainly targeted to teachers and students and it is capable of enhancing both teaching and learning experiences. This platform is capable of collecting user data which are really important in training the developed machine learning models. On the other

hand, the platform provides efficient and required functionalities and services to both the user types; students and teachers.



## 6 REFERENCES

- [1] E. Mandinach, M. Honey and D. Light, "A Theoretical Framework for Data-Driven Decision Making", in annual meeting of AERA, San Francisco, 2006.
- [2] R. Molina-Carmona, C. Villagra-Arnedo and F. Gallego-Duran, "Analytics-driven redesign of a instructional course", in TEEM 2017, Cadiz, Spain, 2017.
- [3] S. Chow, K. Yacef, I. Koprinska and J. Curran, "Automated Data-Driven Hints for Computer Programming Students", in UMAP '17 Adjunct Publication of the 25th Conference on User Modeling, Bratislava, 2017.
- [4] Ryan S J D Baker and Kalina Yacef. The State of Educational Data Mining in 2009 : A Review and Future Visions. *Journal of Educational Data Mining*, 1(1):3–16, 2009.
- [5] Brijesh Kumar Baradwaj. Mining Educational Data to Analyze Students ” Performance. *IJACSA) International Journal of Advanced Computer Science and Applications*, 2(6), 2011.
- [6] Brijesh Kumar Bhardwaj and Saurabh Pal. Data Mining: A prediction for performance improvement using classification. (*IJCSIS) International Journal of Computer Science and Information Security*, 9(4), 2011.
- [7] A Boden, B Nett, T von Rekowski, and V Wulf. *Strategic Learning. . . . of the Sciences of Learning*, 2012.
- [8] Shiv Kumar Gupta, Sonal Gupta, and Ritu Vijay. Prediction of Student Success that are going to enroll in the Higher Technical Education. *International Journal of Computer Science Engineering and Information Technology Research*, 3(1):95–108, 2013.
- [9] Julie A. Marsh, John F. Pane, and Laura S. Hamilton. *Making Sense of Data-Driven Decision Making in Education. Evidence from Recent RAND Research*. Santa Monica, CA: RAND Corporation, 2006.

- [10] Cristbal Romero and Sebastin Ventura. Educational data mining: A review of the state of the art, 2010.
- [11] R. Kabra and R. Bichkar, "Performance Prediction of Engineering Students using Decision Trees", in International Journal of Computer Applications, 2011.
- [12] P. Cortez and A. Silva, "Using Data Mining To Predict Secondary School Student Performance", 2018.
- [13] M. Ramaswami and R. Bhaskaran, "A CHAID Based Performance Prediction Model in Educational Data Mining", in International Journal of Computer Science Issues, 2010.
- [14] N. Thai-Nghe, L. Drumond, A. Krohn-Grimberghe, and L. Schmidt-Thieme, "Recommender system for predicting student performance," *Procedia Comput. Sci.*, vol. 1, no. 2, pp. 2811–2819, 2010.
- [15] Nguyen Thai Nghe, N. T. Nghe, P. Janecek, and P. Haddawy, "A comparative analysis of techniques for predicting academic performance," in 2007 37th annual frontiers in education conference - global engineering: knowledge without borders, opportunities without passports, 2007.
- [16] B. Brunton, "Learning Styles and Student Performance in Introductory Economics," *Journal of Education for Business*, vol. 90, no. 2, pp. 89–95, 2014.
- [17] B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch, "Predicting student performance: an application of data mining methods with an educational web-based system," in 33rd Annual Frontiers in Education, 2003. FIE 2003.
- [18] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "PREDICTING STUDENTS' PERFORMANCE IN DISTANCE LEARNING USING MACHINE LEARNING TECHNIQUES," *Appl. Artif. Intell.*, vol. 18, no. 5, pp. 411–426, 2004.
- [19] J. Breckler, C. Teoh and K. Role, "Academic Performance and Learning Style Self-Predictions by Second Language Students in an Introductory Biology Course", in *Journal of the Scholarship of Teaching and Learning*, 2011.

- [20] B. Mahanama, W. Mendis, A. Jayasooriya and M. Dayasiri, "Educational Data Mining: A Review on Data Collection Process", in ICTeR, Colombo, 2018.
- [21] S. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", in Informatica, 2007.
- [22] Fleming, N.D. and Mills, C. (1992), Not Another Inventory, Rather a Catalyst for Reflection, To Improve the Academy, Vol. 11, 1992., page 137.
- [23] A Boden, B Nett, T von Rekowski, and V Wulf. Strategic Learning of the Sciences of Learning, 2012.
- [24] Educational Data Mining: A Review on Data Collection Process Bhanuka Mahanama, Wishmitha Mendis, Adeesha Jayasooriya, Viran Malaka, Uthayasanker Thayasivam, Thayasivam Umashanger 2018 International Conference on Advances in ICT for Emerging Regions (ICTer)
- [25] Keefe JW 1987 Learning style theory and practice Reston, Virginia
- [26] Clustering and Profiling Students According to their Interactions with an Intelligent Tutoring System Fostering Self-Regulated Learning FranÇOis Bouchet, Jason M. Harley, Gregory J. Trevors, And Roger Azevedo, Journal of Educational Data Mining, Volume 5, Issue 1, April 2013
- [27] Vellido, A., Castro, F., Nebot, A., And Mugica, F. 2006. Characterization of atypical virtual campus usage behavior through robust generative relevance analysis. In Proceedings of the 5th IASTED international conference on Web-based education. Anaheim, CA, USA: ACTA Press, 183–188.
- [28] Tian, F., Wang, S., Zheng, C., And Zheng, Q. 2008. Research on e-learner personality grouping based on fuzzy clustering analysis. In Proceedings of 12th International Conference on Computer Supported Cooperative Work in Design (CSCWD'2008), Xi'an, China: IEEE, 1035-1040.
- [29] Manikandan, C., Sundaram, M.A.S., And Mahesh, B.M. 2006. Collaborative E-Learning for Remote Education; An Approach For Realizing Pervasive Learning

Environments. In Proceedings of the 2nd International Conference on Information and Automation (ICIA'2006), Colombo, Sri Lanka: IEEE, 274-278

[30] R. Cole, "Estimating the impact of private tutoring on academic performance: primary students in Sri Lanka", *Education Economics*, vol. 25, no. 2, pp. 142-157, 2016. Available: [10.1080/09645292.2016.1196163](https://doi.org/10.1080/09645292.2016.1196163).

[31] A. Pallegedara, "Demand for Private Tutoring in a Free Education Country: The Case of Sri Lanka", *SSRN Electronic Journal*, 2012. Available: [10.2139/ssrn.2156145](https://doi.org/10.2139/ssrn.2156145).

[32] Q. Suleman and I. Hussain, "Effects of Private Tuition on the Academic Achievement of Secondary School Students in Subject of Mathematics in Kohat Division, Pakistan", *Journal of Education and Learning (EduLearn)*, vol. 8, no. 1, p. 29, 2014. Available: [10.11591/edulearn.v8i1.203](https://doi.org/10.11591/edulearn.v8i1.203).

[33] E. Smyth, "Buying your way into college? Private tuition and the transition to higher education in Ireland", *Oxford Review of Education*, vol. 35, no. 1, pp. 1-22, 2009. Available: [10.1080/03054980801981426](https://doi.org/10.1080/03054980801981426).

[34] S. Mahmoudi, E. Jafari, H. Nasrabadi and M. Liaghatdar, "Holistic Education: An Approach for 21 Century", *International Education Studies*, vol. 5, no. 3, 2012. Available: [10.5539/ies.v5n3p178](https://doi.org/10.5539/ies.v5n3p178).

[35] J. Wang, "Relationship Between Mathematics and Science Achievement at the 8th Grade", *International Journal of Science and Mathematics Education*, vol. 5, pp. 1-17, 2005.

[36] T. O'Reilly and D. McNamara, "The Impact of Science Knowledge, Reading Skill, and Reading Strategy Knowledge on More Traditional "High-Stakes" Measures of High School Students' Science Achievement", *American Educational Research Journal*, vol. 44, no. 1, pp. 161-196, 2007. Available: [10.3102/0002831206298171](https://doi.org/10.3102/0002831206298171).

- [37] J. Maerten-Rivera, N. Myers, O. Lee and R. Penfield, "Student and school predictors of high-stakes assessment in science", *Science Education*, vol. 94, no. 6, pp. 937-962, 2010. Available: 10.1002/sce.20408.
- [38] L. Barnard-Brak, T. Stevens and W. Ritter, "Reading and mathematics equally important to science achievement: Results from nationally-representative data", *Learning and Individual Differences*, vol. 58, pp. 1-9, 2017. Available: 10.1016/j.lindif.2017.07.001.
- [39] E. Kumtepe, S. Kaya and A. Kumtepe, "The Effects of Kindergarten Experiences on Children's Elementary Science Achievement", *Elementary Education Online*, vol. 8, 2009.
- [40] W. Gustin and L. Corazza, "Mathematical and verbal reasoning as predictors of science achievement", *Roeper Review*, vol. 16, no. 3, pp. 160-162, 1994. Available: 10.1080/02783199409553564.
- [41] L. Frailey and C. Crain, "Correlation of excellence in different school subjects based on a study of school grades.", *Journal of Educational Psychology*, vol. 5, no. 3, pp. 141-154, 1914. Available: 10.1037/h0072639.
- [42] G.C.E (O.L) Examination 2017 Performance of Candidates. Department of Examinations - Sri Lanka, 2018.
- [43] H. Oh, "A STUDY ON THE RELATION BETWEEN MATHEMATICS AND FOREIGN LANGUAGE", *Korean J. Math*, vol. 18, no. 4, pp. 409–424, 2010.
- [44] S. Bergen, *Mathematics and Foreign Language: Authentic Texts in Mathematics*. The Ohio State University, 2017.