Paper No: SE-10 **Systems Engineering**

Predicting examination performance using machine learning approach: A case study of the Grade 5 scholarship examination in Sri Lanka

U. M. Ranawaka*
Department of Industrial Management
Faculty of Science
University of Kelaniya, Sri Lanka
umranawaka@gmail.com

Chathura Rajapakse
Department of Industrial Management
Faculty of Science
University of Kelaniya, Sri Lanka
chathura@kln.ac.lk

Abstract: **Universal primary school education is a must requirement and one of the criteria that should be fulfilled by the developing countries according to the International development goals which are also recognized as "Eight Millennium Development Goals". In the context of Sri Lanka, Government is mostly involved in primary education through government-controlled schools. The success of primary education is measured by conducting a scholarship examination. Those who are getting higher results are given opportunities to attend well-facilitated schools for secondary education. Due to that case, there is a massive competition for passing the examination. Limitless pressure for examination provides lots of issues to students. This paper uses data to investigate a model of academic performance as measured by past results of school tests of Grade 4 and Grade 5. 500 students from eight primary schools in the Gampaha district have been selected for collecting data. The Data on the above-mentioned students have been collected by conducting questionnaires to the teachers who in-charged the classes. Then the Logistic Regression model and Multiple Linear Regression model have been applied to predict students' performances at the examination. The model depicts the likelihood of a student passing or failing the grade 5 scholarship examination and predicts the range of results that students will obtain in the examination. The accuracy of predictive models is measured using the results of students who have already faced the Grade 5 examination. Revealing the potential of students at the grade 5 examination is heavily benefited by teachers because they can provide personalized education for talented students and provide opportunities to other students to improve their talents. The initial architecture of the Grade 5 examination results' predictive model is being discussed in this paper.**

Keywords*: Academic Performance Predicting, Logistic Regression, Machine Learning, Multiple Linear Regression*

I. INTRODUCTION

This paper focuses on the capability of measuring academic performance to incorporate children's primary education system with the possibility of reducing the pressure that applied to children's minds and increasing the efficiency of primary education. The objective of this paper is to predict the results of the students before the examination and whether the student can pass or fail in the examination. Section (1) of this paper is an introduction about the current education in Sri Lanka, Grade 5 scholarship examination and Predictive analysis. Section (II) discusses deep insight into the current environment with related studies and identifies problems in current education environment. Section (III) considers about how research happens and model is going to build. Data Collection, Data Preprocessing, Parameter Selection Method and proposed model are described here. The results and the accuracy of the model is discussed under the section (IV). Finally, in the section (V) it manifests the new implementations to the field along with the recommendations and improvements of the current model.

A. Education in Sri Lanka

Because of the free education approach (1947) and offering education through Sinhala, Tamil and English mediums to stand as the mechanism of training, Sri Lanka accomplished universal primary education in 1964 and eliminated the disparities in enrolment in education. This is a huge opportunity for talented students who are undergoing financial difficulties to stand out from the crowd as a well educated person and shape the future of Sri Lanka. Free education provides equity and equality of opportunity to every person. Social integration is influenced by education in terms of relevant resources in the field, opportunities, and consequences [1].

Schools in the Sri Lankan education system have been conducted and provisioned under five categories known as National, 1AB, 1C, Type 2 and Type 3 schools. Funds and facilities are provided differently for those school types. That reason brings some schools more popular than other schools. When schools become more popular, the demand for getting into school goes high.

Government spends nearly 4% of the Gross Domestic Product (GDP) on education. As a result, there is a significant development in literacy and school enrolment in both male and female students. Also, literacy rate is high in Sri Lanka (2012- Male 92.6% and Female 90.0%) [2].

B. Grade 5 scholarship examination

Primary education is one of the key stages in the Education Framework in Sri Lanka. Because, children between 5-10 years of age belong to primary education. Primary education is the first step of the whole education, first touch point and first impression of the society, teachers, and friends of a child. Primary education is the base of the children's life and directly affects the growth and further education. If primary education becomes the worst nightmare of a child, that will affect the entire life of a student.

At the end of the primary education Department of Examinations conducts a scholarship examination to measure the education of students and provides a scholarship to go to

another school for further education. In order to obtain a scholarship, students should score better results than average results (more than 75%). There are a limited number of opportunities to enroll in well-facilitated schools, Therefore, Grade five scholarship examination is one of the highly competitive examinations in Sri Lanka.

There are arguments whether the Grade 5 examination is good or bad for a young student because it brings on undue pressure on them to obtain a good result. The President of Sri Lanka pointed out the pressure that students are facing at a young age because of exams [3]. The Ministry of Education suggests converting Grade 5 scholarship examination as non-compulsory [4]. Even though the Grade 5 examination brings difficulties to students, the government can't abolish it. Because it provides scholarships for talented students and there are no other alternatives to measure talent of students other than a written examination. Increasing the number of schools which have better facilities is the best solution. But countries like Sri Lanka will process many years to fulfil the target and cost a huge amount of money. Therefore, Education policy makers try to figure out alternate solutions for this problem.

C. Grade 5 examination at Gampaha District

Gampaha as one of the three districts in the western province, Sri Lanka consists of the second largest population in Sri Lanka. Colombo district is an adjacent district of Gampaha and both districts have many national schools and 1AB schools which have highest popularity among Sri Lankans like Royal College, Ananda College, Vishakha College, Devi Balika College and so on [5]. Students in Gampaha district have the opportunity to attend those schools by passing Grade 5 examination. Because of the popularity of those few best schools in Sri Lanka there are huge competition among students, parents of students as well as teachers and principals in primary schools. Parents want to attend their students to high rating schools because it improves their dignity in society. Teachers like to have more passed students in their class. It can measure the quality of primary teaching. So, there is a competition among teachers to be the best teacher who passed most students. Principals and other administrative staff consider school-wide passed students in exams. They want to be the best school who passed most students in Sri Lanka. It will increase the popularity of school and it will increase parents' demand to attend their child to that school for primary education. Therefore, all entities in the education of children have competition to pass the Grade 5 examination with good results.

Due to the competition, students must follow a tough schedule to get prepare for the examination. Other than schools there are private tuition classes, paper classes, and individual training to prepare students for examinations. When examination is near many parents go to supernatural beliefs to support their students in examination. Parents go to temples, Gods, Astrology services and follow so many worship methods to give blessing to their children in examination. Because the examination is difficult, a typical student prepares for Grade 5 examination since Grade 3, two years before examination.

Typical Students in Gampaha district try their best to get a good result, because there are lots of good schools near them. But the reality is every student can't pass the exam. Every student doesn't have the same talent in academics. If teachers can identify students with other skills, teachers can prevent wasting time on unrealistic goals and guide them to improve their other skills. This will help to get rid of feelings that will hurt a student's mind like Toxic looser Syndrome, pressure, stress and depression. If not, students will lose their trust in their own talents. Students should be allowed to enjoy their childhood as much as possible.

D. Predictive analysis

Predictive models and analytics are the practices of extracting information from datasets to forecast future probabilities and outcomes. It is used to determine patterns and special behaviors in datasets in order to make decisions easily in future. Predictive analytics does not tell you what is going to happen within the future. Instead, it can give an idea of what may happen within the future with an appropriate level of reliability and includes what-if situations and risk assessment. The knowledge obtained through analysis can contribute to resolve many problems in the current system and improving the current system to a better level.

Predictive analysis in education and educational data is referred to as Educational Data Mining (EDM). EDM is defined as a process that transforms raw data generated by the education system into beneficial information that could be used by any relevant authority to take accurate actions [6].

II. LITERATURE REVIEW

A. Education in Sri Lanka

Education system of Sri Lanka is free for all students, but it is controlled by the Government. Therefore, there are both positives and negatives to may affect students' academic life. According to research paper [7], Sri Lanka had been able to register almost all the primary aged students into schools marking the net primary enrolment rate as 96%, and the primary completion rate as 95%. Gender equity also prevailed; male and female enrollment rates were equal at all educational levels. Unlike many countries in the world, Sri Lankan education hadn't gender discrimination, ethnic discrimination or other discriminations. Every child had equal rights to get education from schools. The teacher student ratio was 1:18.5 in 2010 in Sri Lankan education [8]. At the same time, the dropout rate had also been recorded as low. Among the students who started schooling in grade one, nearly 1.2% of pupils failed to complete grade 5 as the final year of primary level [9]. According to those facts, Sri Lankan education provides good opportunities to Sri Lankan students to get a quality education.

But child abuse and violence against in government schools are one of the key factors influenced student's performance negatively. Pressurizing students to pass examinations is a kind of abuse and violence of children's freedom. A circular had been issued by the Ministry of Education in restricting corporal punishments in schools, but still there were several incidents reported on the continuation of such violence [10]. According to a survey done by Business Times [11], it showed that "69.25% of Sri Lankans believed that country's education system was given the tough syllabus' and examinations and children were over-burdened with work leading up to the government examinations and private tuition had become essential rather than an option because of the present system (67.36%)".

Apart from the government public schools there are private tuition classes which give education by certified instructor, Undergraduate students or Advanced level students. In the research [12] pointed out that private education became an essential item to Sri Lankan economy. According to his survey, "More than 80% of students who study in grade 6 joined private tuition classes". In Sri Lankan education culture, private tuition is considered as an added advantage for examination. In that research paper said higher income and higher educational levels of the parents had been recorded as the factors which led to spend more on their children's' private tutoring. He also mentioned that the education level of parents had a positive relationship on household private tuition expenditure and employment status. A research paper [13] had observed the factors that led to the expenses of private tutoring and how it would affect the academic performance of the students in Vietnam. His results also suggested that there was an enhancement of the academic performance of the students because of the private classes.

Private tuition being a necessary item is not good for a country which has a free education system in government schools. Due to the competitiveness of examinations, Students were forced to participate in more additional classes to get more advantage in examinations. But private tuition classes might enhance the social differences and damage the free education policy [12]. The paper done by Daniel & Wang [14] demonstrated that school-provided after-school classes could reduce the need for private tuition and raise academic outcomes without harming student mental health in Korea's high school.

B. Grade 5 scholarship Examination

Grade 5 scholarship examination is introduced for Selecting highly potential and but underprivileged students for scholarships and to select the best students for the schools which are categorized as highly rated. Grade 5 examination is a great opportunity for children who have low-income families, because it gives opportunity to attend popular and well-facilitated schools and provide bursaries to the economically disadvantaged students.

Unfortunately, the Grade 5 scholarship examination has become a very competitive examination which has a negative influence on the students now. The Minister of Education pointed out that the Grade 5 scholarship examination caused several mental issues to students. A special parliamentary advisory committee for education had been appointed to investigate this matter. According to the educationalists and child psychologists, this exam had created an unwanted competition among the students without considering the variations in brain volume and growth in children. The time and energy of the students that could be used to explore much new knowledge had been limited by the examinations and cause to have schools which were not much productive, creative or innovative.

Grade 5 examination presently is a torment to kids, having supported unhealthy competition and over-ambitious parents. Mothers even more especially, got used to slave-drive their young children of Grades 4 and 5 to get more and more study. This was not so healthy for an impressionable kid. The students were forced to go to private tuition classes and to cram – even from the Grade one [15]. In order to pass the examination student should score at least 150 marks out of 200 (75%). In 2019, the Grade 5 Examination pass mark was 159 out of 200 (79%). If a student wanted to attend a good school with good facilities, he should score at least 170 marks (85%). In 2018 79.09% students obtained more than 70 marks and mean value was 117.98 out of 200 results in Grade 5 scholarship examination [16]. But the cutoff mark of Gampaha district for schools was 168 marks (84%). Therefore, many students who didn't get more than 168 marks couldn't go to a better school than their current school.

In Annual Performance Report [16] of the Ministry of Education said that ministry pays attention to find a way to minimize the pressure caused to the students due to making Grade 5 Scholarship examination a competitive examination. Also, the ministry paid attention to provide ample opportunities to the students of rural schools and the students of low-income families in examinations. Pressurizing students for grade 5 examination is a national level problem and many authorities in Sri Lanka has paid attention to mitigate that problem.

C. Educational data mining and Predictive models

Data mining is scanning for specific patterns within huge sets of data which makes a lot of possibilities for decision makers. By analyzing those patterns, better choices can be made in order to uplift the process of learning and assessment. To analyze patterns, various research papers use various data mining techniques.

In a research paper done by Mamcenko J. [17] summarized the level of student success in achieving the corresponding outcomes and objectives by incorporating clustering and association rules in data mining to analyze the data extracted from the electronic exam. This paper used unsupervised Machine Learning for analyzing. Researchers gave the easy exam questions and difficult exam questions to students and analyzed the relationship between time spent and correct answers during first pass. Using association techniques, he obtained students' knowledge level. In research paper done by Agasisti [18] used classification methods for their supervised machine learning for analyzing Schools and Higher educational data such as course scheduling, budgets, student transportation cost-benefit analysis and so on. In the journal Tanna [19] analyzed data using Decision Rules. Results were calculated from which a student could choose which stream and college he/she could select for further education based on Entrance Exam marks he/she had scored. This analysis used past inputs which enabled the user to get a more realistic result as compared to a system generating results based purely on the thresholds assigned.

Kumar V. [20] applied the CARM (Conversation Analytic Role-play Method) methods using a tool called Tanagra to improve university students' performance. Apriori algorithm was used to associate students' performances in general courses at the undergraduate and postgraduate levels, recognized associations, and then identified the factors that influence the students' opportunity to get success or improve, their failure, preferences, syllabus designing, and teaching and evaluation techniques. Baradwaj and Pal S. [21] performed a study using Bayesian classification on students from Awadh University in India. Generally, they applied the process through selecting three hundred students from several levels. The authors identified that the students' performance was strongly correlated with other elements, other than students'

effort such as students' daily habits, family income and different factors were mentioned in the paper. Kasih J. [22] suggested a model to predict students' final grade in the programming course. The prediction had been categorized in three options: "Extraordinary", "very satisfactory", and "satisfactory". A solution was granted based on the Apriori algorithm in order to understand the relationship between a programming course and other subjects' students took during the first four-semester of their study period. As a result, they discovered a high correlation between computer courses and math courses. Abdullah Z. [23] proposed a model to recognize appropriate programs for the students considering their interest rather than on availability. It was evident from the study that was conducted by using Computer Science students from Malaysia University, suggested that even though many students were learning computer science programs that were not under their field of interests.

G. Badr used a classification based on an association rules algorithm to develop a classifier to assess the student's performance in the programming course in research titled "Predicting Student's Performance in University Courses' ' [6]. These predictions would help the students to get to know how well they would perform in a course even before they register in it to avoid having to drop out. G. Badr tried to predict student performance in a programming course based on their performance in English and mathematics courses. He used two approaches. The first method used "students' grades in two English courses and two mathematics courses, which produced four rules with accuracy of 62.75%. The second method used only two English course results, which had accuracy of 67.33%". The results showed English courses had a significant effect on predicting performance in the programming course. In another research [24] said that integration of many noisy variables reduced accuracy of the prediction and would lead to overfitting of data.

There are many educational data mining applications in the world and many researchers' tries to improve the quality of education using data analysis.

D. Research Objectives

This paper introduces a model to predict performance at academics of students before the examination. Talented students in academics can be identified though this model. Teachers can give personalized education for talented students in academics. Also, teachers and parents can identify students who have other talents. Teachers can give essential education and give more time to develop their own skills like art, singing, sports skills and so on. Those students can become personalities that they deserved to be one day. Especially those students can spend their childhood very happily without unnecessary competition or pressure.

III. METHODOLOGY

Research approach of this development-oriented research could be defined under three major phases. In the first phase , there were interviews conducted with the respective authorities to understand the processes of the Sri Lankan education system. Also, a broad and extensive study of literature was done along with this. During the second phase, data from eight schools and 500 students were collected. This stratified sampling represented 1AB ("Maha vidyala"), Type 2 ("Kanishta vidyala") and Type 3 ("Primary") schools. Clustered sampling was used to select classes from a selected school. In each school one or two classes selected as clustered sampling. In a class, every students' data was collected for research. If sampling method was applied in this step, there would be a discrimination of student talent. By collecting all students in class, best students, normal students and weak students could be selected without any bias. In the final phase, model was developed using those data, tested and validated with collected test sample.

A. Data Collection

Data was collected using a questionnaire that was given to teachers of students. This survey was done for teachers because teachers spent 30 hours per week with students and they had good knowledge in the academic environment of students. Teachers and parents nearly had equal knowledge of the academic talent of students. So, teachers were selected for the data collection. This questionnaire was based on research "The impact of school quality, socioeconomic factors, and child health on students' academic performance: evidence from Sri Lankan primary schools, 2011" done by Aturupane H., Glewwe P [7]. In that research, researchers used questionnaires to collect data of grade 4 students and their environment. With suitable modifications, the same questionnaire was used for this research.

Grade 5 students who sat for scholarship examination in 2019 were selected for the research. Their academic results in school tests, final grade and final marks were collected. Data consisted of 500 students who studied in eight different schools in Gampaha district. Both urban and rural schools were selected to reduce the biases of facilitated and non-facilitated school environments. Principals of those schools in Gampaha district were well-qualified, good experienced and they had sound knowledge in primary education. Principals provided maximum effort to give better education to students.

In selected schools, one or two classes randomly selected for data collection. There were well-experienced and well-talented teachers for Grade 5 and Grade 4 in primary sections in the context of Gampaha district. Teachers' experience played a major impact for students' education. Therefore, well-experienced teachers were selected for data collection in selected schools.

In the questionnaire, following data collected from selected entities.

- Gender
- Grade 4, 1st term test – Buddhism, Sinhala, English, Mathematics, Science (Environment)
- Grade 4, 2nd term test – Buddhism, Sinhala, English, Mathematics, Science (Environment)
- Grade 4, 3rd term test – Buddhism, Sinhala, English, Mathematics, Science (Environment)
- Grade 5, 1st term test – Buddhism, Sinhala, English, Mathematics, Science (Environment)
- Grade 5, 2nd term test – Buddhism, Sinhala, English, Mathematics, Science (Environment)

These subjects are taught in every school at grade 5. The Grade 5 scholarship examination normally consisted of knowledge of Mathematics, Science (Environment), Sinhala and a little bit of English.

Other than the above data, teachers' experience and principals' experience were considered as control factors of research. In Gampaha district typically there were well-experienced teachers and principals for primary education who have more than 5 years of experience.

B. Data preprocessing

Data pre-processing has been performed to remove noise and to gain an accurate representation of data. This step is crucial in building the model since uncleaned data would affect the performance of the model. As the 1st step of the preprocessing, outliers of the dataset were removed. In the 2nd stage, uncompleted or unknown data entries were changed in the dataset. When teachers were filling out the questionnaire, they could mark some attributes that were unknown and left uncompleted. Mean value of selected subject had been used for treating uncompleted data entries.

C. Parameter selection method

There are 26 main attributes or columns in the data set. If all attributes were analyzed and used to create predictive models, it would reduce the performance of models and deliver overfit or underfit results. Therefore, it was vital to identify the most important features that would affect overall results of the model. Two methods were used to identify important parameters.

1. Principal Component Analysis (PCA)

PCA is a method of developing new variables (principal components), which are linear composites of the original variables. The values of principal components created by PCA were known as principal component scores (Table 1).

TABLE I. PRINCIPAL COMPONENT SCORES

Name of the Component	PCA values
PCA 1	0.71 (71%)
PCA 2	0.04 (04%)
PCA 3	0.03 (03%)
PCA 4	0.03 (03%)

Table 1 showed the first four PCA values out of many PCA values. PCA explained the variance of the dataset. PCA 1 held 71% of variance in the dataset. For this research first four PCA values were selected which hold 81% of variance.

1. Backward Elimination (Stepwise regression)

Backward Elimination is a method which includes beginning with all candidate variables, testing the deletion of each variable using a chosen model fit criterion, removing the variable whose loss generates the most statistically insignificant deterioration of the model fit, and reiterating this process until no more variables can be removed without a statistically significant loss of fit. In this research P-values were calculated of each independent variable and compared with significant level. If the highest P-value was greater than the significance level, variable/feature was removed. This was an iterative process and the process happened until there was no highest P-values greater than significance level. After calculations, significant variables which affected results mostly were identified (Table II).

TABLE II. BACKWARD ELIMINATION RESULTS

Significance Level = 0.05	
Subject	***P-Value***
Grade 4, Term 2 – Sinhala	0.031
Grade 4, Term 2 – Mathematics	0.000
Grade 4, Term 3 – Science (Environment)	0.002
Grade 5, Term 1 – English	0.001
Grade 5, Term 1 – Mathematics	0.045
Grade 5, Term 2 – Sinhala	0.000
Grade 5, Term 2 – Mathematics	0.000
Grade 5, Term 2 – Science (Environment)	0.032

After filtering out above features/parameters were selected for further developments of model.

D. Model Selection and Development

All independent variables are quantitative data. Independent variables consist of term test results. Dependent variables have two categories.

- Grade 5 examination marks – Quantitative
- Grade 5 examination grade – Qualitative

Two separate methods were introduced in this research in order to predict both marks and grade.

1. Method 1 – Predicting Marks

In order to predict marks that students will get at the examination, Linear regression would be used. Linear regression is a linear approach to developing the relationship between a dependent variable and one or more independent variables. For more than one independent variable, the model is called multiple linear regression. Independent variables are features that are extracted from parameter selection methods.

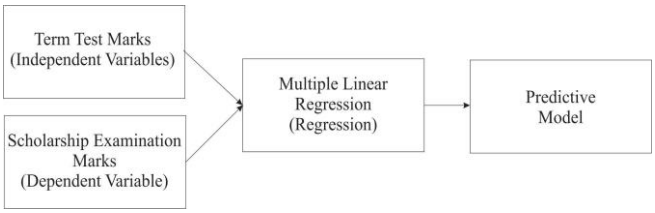

Fig 1 Predicting Marks – Method 1

The outcome of the model has predicted marks at given independent variable values.

2. Method 2 – Predicting Grade

In Grade 5 scholarship examination results consist of student grade at the examination. It is Pass the examination or Fail the examination. This is a dichotomous dependent variable. Logistic Regression is a supervised machine learning algorithm used in binary classification. Binomial Logistic regression was used to develop a model to predict whether the student would pass or fail at the upcoming examination.

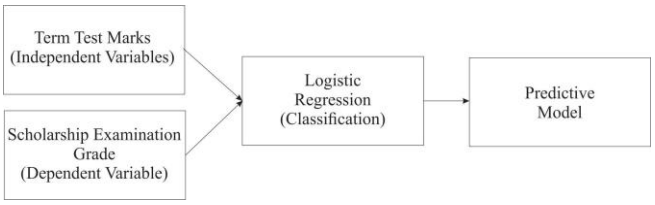

Fig 2 Predicting Grade – Method 2

The outcome of this model is whether students pass or fail at the examination. The threshold point was defined at 159 marks out of 200 marks. If students got below 159, they would be considered a failure. If students scored 159 or above, they would be considered as pass. Pass marks of the 2019 examination (159 marks) were used as binary classification.

IV. TESTING & VALIDATION

After developing the models, they should be tested and get an idea of how much applicability they derived and the performance of them in selected tasks. Accuracy of the model was a good measurement of the model's success in predicting student's performance at the exam.

In order to create a training data set and testing data set, the percentage split technique was used. Data set divided into 80% training data set and 20% testing data set. The model would train with training data and updated models' algorithms according to data. Then models' learning success is measured using testing data. All independent variables were scaled or standardized to minimize errors generated from high weighted values.

A. Method 1 – Predicting Marks

Multiple Linear regression was used to predict marks of Grade 5 scholarship examination. After successful training and testing data plots were developed to show the success of prediction (Figure 3).

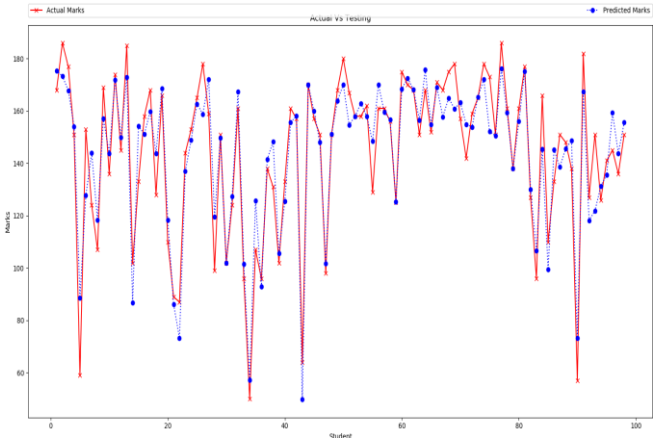

Fig 3 Multiple Linear Regression Results – Actual vs Predicted

Test data consists of 98 students and their actual marks and predicted marks which predicted using model shows in this graph. Solid line style and "x" marker are actual marks of test data set and dotted line style and dot marker are predicted marks predicted by the regression model.

TABLE III. STATISTICS

Intercept	22.90602
Root mean squared error (RMSE)	115.43913
Sum of Square of Residuals	11313.03512
R2 Score	0.87602

The mean absolute percentage error (MAPE) is used as a loss function for regression problems in machine learning. Using MAPE, Accuracy of the regression model can be extracted.

TABLE IV. ACCURACY OF REGRESSION MODEL

Total Error	6.33117
MAPE % (Mean Absolute Percentage Error)	6.46038
Accuracy %	93.53961

B. Method 2 – Predicting Grade

Predicting the grade that will be achieved by students was a classification problem. Therefore, logistic regression was used for developing this method. Confusion matrix can be used to calculate the accuracy of the model (Table V)

TABLE V. CONFUSION MATRIX

	Predicted Fail	**Predicted Pass**
Actual Fail	53	7
Actual Pass	6	32

Accuracy shows how the classifier model correctly predicted results. Error rate shows how often the prediction is wrong. Sensitivity shows when a student passes, how the classification model correctly identified it as passed. Specificity shows when a student fails, how the model identified it as fail. Precision shows the accuracy of the predicted pass. Prevalence shows How often the pass condition occurs in our dataset. (Table VI)

TABLE VI. ACCURACY OF CLASSIFICATION MODEL

Accuracy	0.86734
Error Rate	0.13265
Sensitivity / Recall	0.84210
Specificity	0.88333
Precision	0.82051
Prevalence	0.38775

A 10-fold cross-validation method used as a resampling method has been used to improve the accuracy of the model. The 10-fold cross-validation dataset was divided into 10 parts and tested each part with a model. The remaining nine parts would train the model. This process repeated for nine more iterations and more accurate results could be extracted. 10-fold cross-validation gave 10 accuracies for their 10 iterations. Table VII shows mean accuracy and Standard deviation of the resulting accuracy.

TABLE VII. 10-FOLD CROSS-VALIDATION RESULTS

Mean Accuracy	0.844079
Standard Deviation	0.047012

V. DISCUSSION

Techniques used in data mining allows the extraction of knowledge from student result data, presenting interesting

possibilities for the education domain. In this study, two models were created based on term test result variables gathered through a questionnaire approach. Some of the most influencing variables/factors were recognized and taken to predict the examination result and grade of a candidate. For predicting marks, Multiple Linear Regression was used (method 1). For predicting grade, the Logistic Regression method was used (method 2).

The Predicting model - method 1 achieved an accuracy of 94% for predicting student's marks. It shows that predicting model gives less error in prediction like the Mean Absolute Percentage Error is 6%. In simple terms, our prediction model can predict nearly correct marks

Actual Marks = Predicted Marks ± 5 Marks

The actual vs Predicted graph clearly shows that this equation is valid for most cases. Therefore, the obtained results from method 1 are good. Because predicted values are not much overfitting or underfit for actual values. In the educational environment, this model can suggest a range of results that will be taken by students. The range of results gave an idea about performance at the examination of students.

In the prediction model - method 2 which used to predict the grade of the examination had been achieved 84% accuracy using a 10-fold cross-validation method. For determining the grade, the threshold point/cut-off mark is defined as 159 marks. Using this model, a prediction can be done on whether a student can get more than 159 or not. Overall accuracy, Sensitivity, Specificity, and other measures are relatively good. But there is some noise in those models.

Even if those preliminary results derive satisfactory results, there is much room to improve accuracy, reduce the noise. Those two models were developed on 500 students' data. If there are more students' data accuracy of models will increase. Selecting sample schools all over the Gampaha district, increasing the number of sample schools and increasing the number of sample classrooms are good ideas to develop the model. As well as collecting data from many years is essential. In this research, all students are 2019 grade 5 students. It is essential to continue research for several years and collect students' data for several years. Those methods will increase the accuracy of prediction. Artificial Neural Network, Support vector machine, random forest algorithms can be used to develop separate models for these scenarios. Then the accuracy of all developed models can be compared and can figure out the most accurate model. Those models can be done fine-tuning to improve accuracy. As an example, changing classification algorithms, changing threshold values may increase the accuracy of the model.

It is better to consider social background, family background and mental background for developing predicting performance of the student. Parents' education level, parents' income, number of extra tuitions attended, extra-curricular activities and so many other factors have an impact on students' performance. As future work, those factors should be studied, and the prediction model should be recreated to make predictions more accurately.

Using educational data mining techniques and models developed in this research, teachers can get a sound knowledge of their students. Teachers can get an idea about how teachers are going to improve students' skills effectively. This research doesn't mean to abandon students who don't have the academic talent and give focus only to talented students. But teachers should show the correct way to both talented and weak students in academics. Passing grade 5 examination is not the only way to create a bright future and get a better life. Teachers should encourage students to develop their talents rather than blindly studying for an examination and to be multi-talented students. With the help of these techniques, students can learn whatever they like and be talented happily.

REFERENCES

[1] P. Colenso, *Education, and social cohesion: developing a framework for education sector reform in Sri Lanka*, vol. 35, 2005, pp. 411–428.

[2] Research and Reports, "Sri Lankan Education" [Online], Available: www.unicef.org /infobycountry /sri_lanka_statistics.html.

[3] Hiru news, "Government decides to abolish the grade 5 scholarship exam, says President" [Online], Available: http://www.hirunews.lk/213156/government-decides-to-abolish-the-grade-5-scholarship-exam-says-president.

[4] I. Mudugamuwa, "Government issues circular making Grade 5 Scholarship Exam non-compulsory" [Online], Available: http:// www .dailynews.lk/2019/04/08/local/182549/government-issues-circular-making-grade-5-scholarship-exam-non-compulsory.

[5] Statical branch Ministry of education, *School Census Report_2017*, 2017.

[6] G. Badra, A. Algobaila, H. Almutairia, M. Almuterya, Predicting Students' Performance in University Courses*: A Case Study and Tool in KSU Mathematics Department*, 2016.

[7] H. Aturupane, P. Glewwe, S. Wisniewski, The impact of school quality, socioeconomic factors, and child health on students' academic performance: evidence from Sri Lankan primary schools, 2011.

[8] United Nations Development Programme, *Sri Lanka Human Development report*, 2012, pp.60.

[9] World Bank, Treasures of the education system in Sri Lanka: Restoring performance, expanding opportunities and enhancing prospects, South Asia Region, Washington, DC, 2005.

[10] S.Jayaweera, C.Gunawardena, Social inclusion: gender and equity in education swaps in south asia sri lanka case study, 2009.

[11] Business Times, "Parents, administrators, civil society slam Lankan education system" [Online], Available: http://www.sundaytimes.lk/110306/index.html.

[12] A. Pallegedara, Demand for private tuition classes under the free education policy. Evidence based on Sri Lanka, 2011.

[13] H.A. Dang, *The determinants and impact of private tutoring classes in Vietnam. Economics of Education Review*, 2007, vol. 26, pp. 684–699.

[14] C. Daniel. and L. C. Wang, The effect of after-school classes on private tuition, mental health, and academic outcomes*: evidence from Korea*, Monash University

[15] I. M. K. Liyanage, Education System of Sri Lanka: Strengths and Weaknesses, 2014.

[16] Ministry of Education Sri Lanka, *Annual Performance Report*, 2018.

[17] J. Mamcenko, I. Sileikiene, J. Lieponiene, R. Kulvietiene, Vilnius, Analysis Of E-Exam Data Using Data Mining Techniques, *Gediminas Technical University*.

[18] T. Agasisti, A.J. Bowers, Data Analytics and Decision-Making in Education*: Towards the Educational Data Scientist as a Key Actor in Schools and Higher Education Institutions*, 2017.

[19] M. Tanna, Decision Support System for Admission in Engineering Colleges based on Entrance Exam Marks, *International Journal of Computer Applications*, August 2012, vol. 52, no. 11, pp. 0975 – 8887.

[20] V. Kumar, A. Chadha, "Mining association rules in student's assessment data", in 2012 *IJCSI International Journal of Computer Science*, vol 9, pp. 211-216.

[21] B. Bhardwaj, and S. Pal, "Data Mining: A prediction for performance improvement using classification", in 2011 *International Journal of Computer Science and Information Security (IJCSIS),* vol. 9, no. 4, pp. 136-140.

[22] J. Kasih, M. Ayub, and S. Susanto, "Predicting students' final passing results using the Apriori algorithm", in 2013 *World Transactions on Engineering and Technology Education,* vol. 11, no. 4, pp. 517-520.

[23] Z. Abdullah, T. Herawan, N. Ahmad, and M. Deris, "Extracting highly positive association rules from students' enrollment data", in 2011, *Procedia - Social and Behavioral Sciences,* vol. 28, pp. 107-111.

[24] Ransohoff, *Rules of evidence for cancer molecular marker discovery and validation. Nature Reviews/Cancer*, 2014, vol. 4, pp. 309–313