English School Mate

Madushani N.G.H

IT17027670

BSc (Hons) in Information Technology

Specializing in Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

September 2020

English School Mate

Madushani N.G.H

IT17027670

Dissertation submitted in partial fulfillment of the requirement for the

B.Sc Special Honors Degree in Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

September 2020

1. Declaration

Declaration Declaration of Candidates

“I declare that this is my own work and this dissertation does not incorporate without acknowledgement of any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. Also, I hereby grant to Sri Lanka Institute of Information Technology, the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic, or other media. I retain the right to use this content in whole or part in future works (such as articles or books).”

Signature: Date:

Declaration of the Supervisor

“The above candidate has carried out research for the bachelor’s degree dissertation under my supervision.”

Signature of the supervisor: Date:

2. Abstract

Predicting the results, performance of a student is a great concern to the education management. As one of the most challenging courses for grade 6 students. many students perform poorly or even fail because they haven’t proper guide for learning english The scope of this paper is to identify the factors influencing the performance of students in examinations and find out a suitable data mining algorithm to predict the grade of students so that reason teachers, parents or a guide can give timely and an appropriate warning to students those who are at risk. In the present investigation, a survey cum experimental methodology was adopted to generate a database and it was constructed from a primary and a secondary source. The obtained results from hypothesis testing reveals that type of school is not influence student performance and parents' occupation plays a major role in predicting grades. A valid model for predicting student academic performance in english is helpful in developing their knowledge and coping with the society.This enhances the teaching and learning process easily. Parents and teachers can easily identify the state of their students. Here there is a teacher dashboard , teacher, parent or guide can easily access it and see the performance by graph or scores.This work will help the educational institutions to identify the students who are at risk and to provide better additional training for the weak students.

3. Acknowledgement

First and foremost I express my deepest sense of gratitude towards my internal supervisor Mr. Dhammika de Silva Senior Lecturer, Faculty of computing , Sri lanka institute of information Technology , Malabe, Sri Lanka. for his excellent advice, immense guidance and encouragement initiation and developing this investigation.

I am also grateful to my internal co-supervisor Mrs. Anjali Gamage. Senior Lecturer, Faculty of computing , Sri Lanka institute of information Technology , Malabe, for her great guidance, assistance throughout the research period for the successful completion.

Further, I wish to extend my sincere with a deep sense of respect to the English teachers who help me to provide all the details of the students and give us the way of teaching English. They provide guidelines of English. I am also thankful to all other academic staff in all the school and English classes for assisting me in every possible way.

I give a dissertation work to my loving parents whose words of encouragement and push for tenacity ring in my ears , friends who have supported me throughout the process, all the organizations around the world who are under numerous information thefts everyday .and lecturers who have helping develop my technology skills.I will always appreciate all they have done.we believe that our research will help for the students to improve their English knowledge and cope with the society well manner

My special thanks go to above mention peoples who gave me their support a lot in finishing this project within this limited time.

Table of Contents

List of Tables

List of Figures

1. Introduction

Since online learning can generate large amounts of records in students' learning process, it provides an effective way to get a deep understanding of students' learning behaviors and predict their academic performance. Due to the benefits of online learning, more and more schools combine with online education to achieve better teaching results. In this study, data is collected from anonymous students. The dataset records the students details like student term test marks their preference for the English activities, students background details like family, parents occupation, school their online learning And entertaining activities. Due to the absence of face-to-face meetings, web-based systems is most important.. Other factors were the student's financial status, which was captured according to the area they live in and whether they commuted to school or not.A set of additional institutional factors such as program of study and fee type and other factors such as address and gender were used at the early stage of the analysis, but then removed due to redundancy or weak correlation to the prediction model. The students enrolled at the school come from various states with different education profiles and have also had different levels of success measured consistent with average grades at schools or state examinations. This, along with their current engagement,probably affects their success within the early phase of their studies. Predicting the success of students within the early phase of their studies helps faculties in directing more activities to less performing students so on improve their success.analyzing academic success is important for education , on condition that the strategic planning of study programs implies expanding or reducing the scope or depth of the curriculum further as modifying the English language depending on student achievements. A lot of research observes academic success generally, like success in individual courses or groups of courses or within the individual phases of studying, bushed terms of current variables like commitment to studying, fulfillment of obligations, quality of delivered educational processes, perceived difficulty of the curriculum and different socio-demographic variables (place of residence, gender, income,habits). Rarely have we undertake scientific observation of success in highschool, especially in individual subjects, or success within the state exam and completion of the school curriculum. it's our opinion that these factors can have an enormous influence on students' success within the early phase of upper education because they contain acquired knowledge, work habits and attitudes towards studying.

Therefore, success in highschool is included within the suggested model so as to investigate its influence on the output variable.

With the wide usage of computers and internet, there has recently been a huge increase in publicly available data that can be analyzed. Be it online sales information, website traffic, or user habits, data is generated everyday. Such a large amount of data present both a problem and an opportunity. The problem is that it is difficult for humans to analyze such large data. The opportunity is that this type of data is ideal for computers to process, because it is stored digitally in a well-formatted way, and computers can process data much faster than humans. The concept of machine learning is something born out of this environment. Computers can analyze digital data to find patterns and laws in ways that is too complex for a human to do. The basic idea of machine learning is that a computer can automatically learn from experience (Mitchell, 1997). Although machine learning applications vary, its general function is similar throughout its applications. The computer analyzes a large amount of data, and finds patterns and rules hidden in the data. These patterns and rules are mathematical in nature, and they can be easily defined and processed by a computer. The computer can then use those rules to meaningfully characterize new data. The creation of rules from data is an automatic process, and it is something that continuously improves with newly presented data. Applications of machine learning cover a wide range of areas. Search engines use machine learning to better construct relations between search phrases and web pages. By analyzing the content of the websites, search engines can define which words and phrases are the most important in defining a certain web page, and they can use this information to return the most relevant results for a given search phrase (Witten et al., 2016). Image recognition technologies also use machine learning to identify particular objects in an image, such as faces (Alpaydin, 2004). First, the machine learning algorithm analyzes images that contain a certain object. If given enough images to process, the algorithm is able to determine whether an image contains that object or not (Watt et al., 2016). In addition, machine learning can be used to understand the kind of products a customer might be interested in. By analyzing the past products that a user has bought, the computer can make suggestions about the new products that the customer might want to buy (Witten et al., 2016). All these examples have the same basic principle. The computer processes data and learns to identify this data, and then 2 uses this knowledge to make decisions about future data. The increase in data has made these applications more effective, and thus more common in use. Depending on the type of input data, machine learning algorithms can be divided into supervised and unsupervised

learning. In supervised learning, input data comes with a known class structure (Mohri et al., 2012; Mitchell, 1997). This input data is known as training data. The algorithm is usually tasked with creating a model that can predict one of the properties by using other properties. After a model is created, it is used to process data that has the same class structure as input data. In unsupervised learning, input data does not have a known class structure, and the task of the algorithm is to reveal a structure in the data (Sugiyama, 2015; Mitchell, 1997). This thesis focuses on supervised learning, more specifically predictive analytics, which is the process of using machine learning to predict future outcomes (Nyce, 2007). Predictive analytics has a wide range of applications, such as fraud detection, analyzing population trends, or understanding user behavior (Sas, 2017). The specific focus of this thesis is education. The aim is to predict student performance. Data about students is used to create a model that can predict whether the student is successful or not, based on other properties. First, the training data set is taken as input. There are two different data sets, containing different types of information. These data sets are in tabular format, where each row represents a student and each column, or variable, contains certain information about a student, such as age, gender, family background or medical information. In addition, a column representing the success of the student is used as the variable that the algorithm is trying to predict. The algorithm creates a model, which is a function that outputs success or failure of the student, using other variables as input. This thesis evaluates the effectiveness of different machine learning algorithms and methods. While algorithms that are used in creating predictive models are numerous, this thesis focuses on three of them, which are linear regression, decision trees, and naïve Bayes classification. The thesis also measures the improvement made by feature engineering, which refers to modifying the data to make it more suitable for machine learning. 3 There are widely used indicators for evaluating the effectiveness of machine learning algorithms, such as precision, recall and F-measure (Powers, 2011). These are covered in detail in further chapters. These indicators can also be used in evaluating the predictive models. Algorithms were compared to each other in terms indicator values, to determine which algorithm provides the best results. In addition to the algorithm choice, the importance of feature engineering was also tested. To improve the prediction performance, the data sets were modified by variable selection and custom variable creation. Finally, improvements made by feature engineering were compared to improvements made by algorithm choice, to see if one is a more determinant factor than the other.

Results of the comparison indicates that feature engineering provides better improvements than method selection.

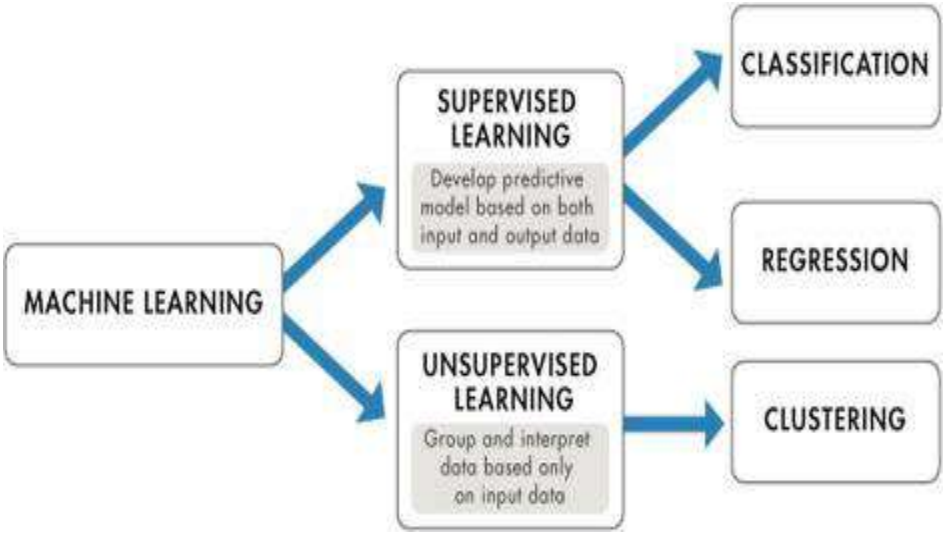


Figure 1: Supervised learning versus unsupervised learning

1.1 background literature

Most of the institution and schools using final examination grade of the student as the student.Academic performance criteria. The final grades of any student depend on assessment and test. The performance of the student depends upon how many grades a student score in the final examination. Norlida Buniyamin, Pauziah Mohd Arsad et al. (2013) stated that what are the significance of academic analytics for an educational institution and how they work for the improvement of education. They also proposed an intelligent recommendation intervention system to improve the student's performance and achievement in education

Computers have become ubiquitous, especially in the last three decades, and are significantly widespread. This has led to the collection of vast volumes of heterogeneous data, which can be utilized for discovering unknown patterns and trends (Han et al., 2011), as well as hidden relationships (Sumathi & Sivanandam, 2006), using data mining techniques and tools (Fayyad & Stolorz, 1997). The analysis methods of data mining can be roughly categorized as: 1) classical statistics methods (e.g. regression analysis, discriminant analysis, and cluster analysis) (Hand, 1998), 2) artificial intelligence (Zawacki-Richter, Marín, Bond, & Gouverneur, 2019) (e.g. genetic algorithms, neural computing, and fuzzy logic), and 3) machine learning (e.g. neural networks, symbolic learning, and swarm optimization) (Kononenko & Kukar, 2007). The latter consists of a combination of advanced statistical methods and AI heuristics. These techniques can benefit various fields through different objectives, such as extracting patterns, predicting behavior, or describing trends. A standard data mining process starts by integrating raw data – from different data sources – which is cleaned to remove noise, duplicated or inconsistent data. After that, the cleaned data is transformed into a concise format that can be understood by data mining tools, through filtering and aggregation techniques. Then, the analysis step identifies the existing interesting patterns, which can be displayed for a better visualization (Han et al., 2011) .

To accomplish our goals, we developed a predictive analytic model utilizing machine learning (ML) algorithms. ,e most appropriate ML predictive model was selected for analyzing student interactions in VLE learning activities and determining students' levels of engagement in VLE courses given that a lack of student engagement results in a high dropout rate . Predictive models are currently used in many educational institutions A predictive model can help instructors guide students in succeeding in a course, and be used to determine which activities and materials are more important to the course assessment. Such models also enable instructors to engage students in different activities through the VLE, thereby encouraging the students to participate in the VLE course. Instructors must invest time discerning why student engagement in particular course activities and material is attenuated. Our models can easily be integrated into VLE systems and can enable teachers to identify low-engagement students through different assessments, the use of different course materials, and the number of times VLE activities (e.g., dataplus, forumng, glossary, resources, URL, homepage, oucollaborate, and subpages) are accessed. Teachers can also spend more time on assessments and materials that are difficult for a particular group of

students, enabling them to discover why an assessment is easy or difficult and providing supplementary intervention to students who need it. A predictive system enables an instructor to automatically identify low-engagement students during a course based on activities from that online course. Given such detection, the instructor can then motivate (e.g., send an e-mail reminder) or identify difficulties during the course . When a student receives an advisory e-mail from an instructor (i.e., an e-mail asking about any difficulty), on a weekly basis, the student is more likely to work hard and increase their engagement. Such communication is important because it assesses student workloads and addresses issues at an early stage of the course . Apt advice will also improve student retention and decrease the course dropout rate

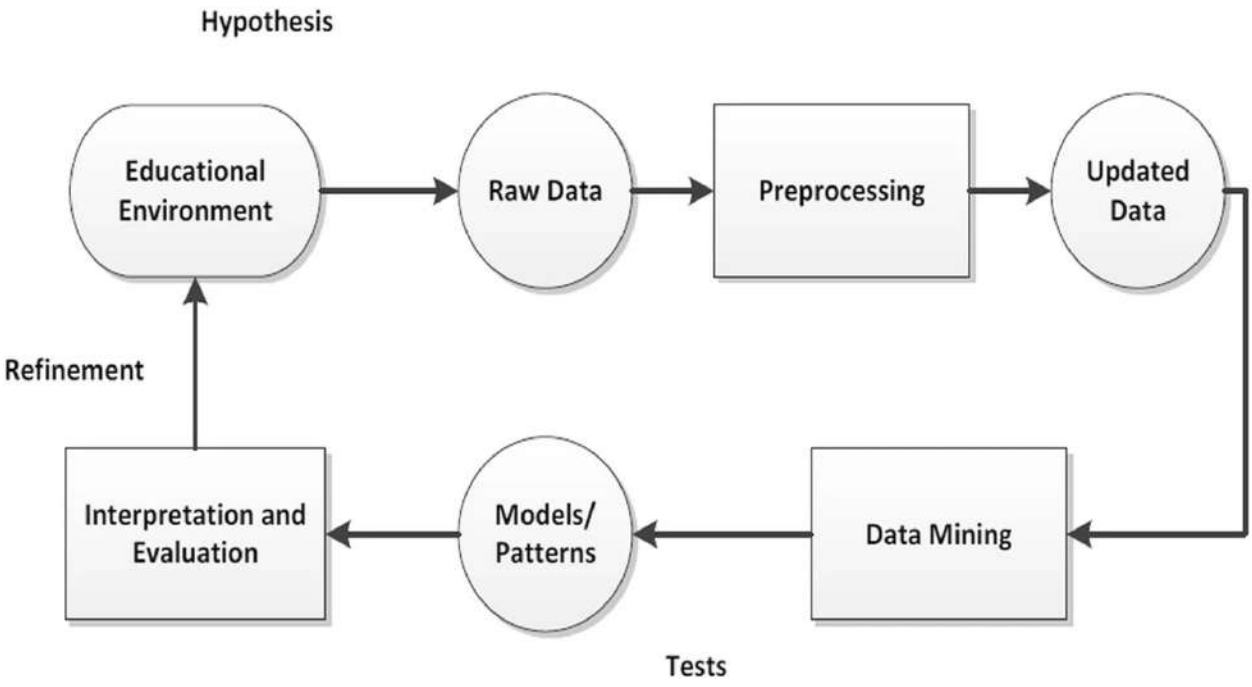


Figure 1.1.1 data flow

Figure 1.1.2 data flow chart

Little research has been devoted to the relationship between academic support and retention to graduation in both the literatures on retention and academic support. Many authors of years of data from the various university Resources for Academic Achievement unit (REACH) to test the hypotheses that a larger quantity of time spent engaged in academic support services is associated with a higher likelihood of graduation and that assesments mediates the relationship between hours spent using academic support . The findings support these hypotheses, suggesting a relationship between academic support and retention to graduation that should be given serious consideration by scholars and administrators and the learners.

Prediction in Research examples:

One of the earliest examples of this was the Muslim scholar, Al-Razi. He was asked to find the best location to build a hospital, in the city of Baghdad.

Cleverly, he hung a piece of meat and predicted that the place where the meat took longest to rot would be the best place to build a hospital. Although he knew little about the exact processes behind the transmission of illness, he realized that some environments were unhealthier than others, especially in a hot climate where gangrene was a problem. His idea was used for many years, until the bacterial processes behind illness were uncovered.

Sticking with medicine, a later example of prediction in research can be found in the wonderful work of Semmelweiss, a scientist responsible for saving countless thousands of lives. In 1847, the Hungarian Dr. Ignaz Semmelweis's close friend, Jakob Kolletschka, cut his finger during an autopsy and contracted a nasty disease known as puerperal fever. Semmelweis also observed that puerperal fever killed 13 percent of women giving birth in his hospital, whilst the nearby hospital, run entirely by midwives, lost only two percent of its birthing mothers.

Semmelweis noticed that students moved between the autopsy room and the delivery room without washing their hands, and predicted that this was the reason for the higher death rate in the teaching hospital. He informed students that they had to wash their hands in a chlorine solution when entering the maternity wing and mortality rates from puerperal fever promptly dropped to two percent.

Unfortunately, the unfortunate Semmelweis became a victim of politics and the director of the hospital, livid that the young doctor was indirectly blaming him for the high rates of mortality, made sure that Semmelweis never worked in Vienna again. Eventually, he returned to Budapest and used his methods there, eventually publishing a book of his findings.

Sadly, the medical establishment rejected his ideas and the disillusioned Semmelweis died in a mental institution after being severely beaten by guards. The autopsy revealed extensive internal injuries - the cause of death was blood poisoning.

1.2 Research Gap

Due to development of technology, over past few years researchers have introduced different kind of mobile applications and web applications to develop English learning abilities. Some applications mainly focused on English learning. But our application is focused on not only the English learning but also entertainment and brain improvement.this app focus to improve english knowledge in many students.

1.3 Research problem

Nowadays English has become a very competitive matter for getting a good job in our country. Those who have a good command in English can easily get a good job. Today various types of organizations often ask for a good working knowledge of English. These organizations give priority to those employees who can speak fluently and write a standard form of English. moreover, we need to learn English to communicate with other countries. An English knowing person can easily exchange his ideas, thoughts, and views with the foreigners. english is a language.lt is an international language .lt is spoken all over the world.the people of the world communicate with each other by this language .this language is used in everyday life-at home,in the market,in the office,at school,in college ,etc.lt is also the medicion of communication through email and internet.moreover, english has been the medium of education in all parts of the world.

In school also teaching english is most valuable thing and important subjects like english,science,economics,history,geography,medicine,engineering,etc.are taught in english.without understanding english the international student cannot get scholarship in america,britain,japan,china or other rich countries. english language is also important to understand literature like poems,stories,essays,dramas or films.The most important world literature is written in english. the politicians should understand english lf they cannot speak english,they cannot know world politica,they cannot communicate their ideas to the foreigners. also,english is used in industries, factories and commerce.therefore,the value and importance of english language is increasing day by day.lt is used in different aspects of life.in our country, those who do not speak,write or understand english language,they seem to be backward.thus,this language has become the language of human progress and prosperity.

In srilanka there is a poor family , medium class family,high class family. In poor students cant go to private tuition because of financial matters. In using this app the also develop their knowledge at the home also.there is a huge problem for many students and parents. In here we try to make a system easy to handle,easy to learn , low cost , easy to access. In rural areas school there is a huge problem that is a lack of teachers. So that reason students can not be able to learn and improve their knowledge.i used Prediction and Visualization Dashboard for using this teachers , parents and guide can easy to access the child's examination results and their information. And also using child's information teachers, parents,and guides can easily know the results they must score. And then if the child can not reach the marks the system shows their states and inform he or she good or bad.

1.4 Research Objectives

English language plays an important role in our lives because it makes communication between different countries the only common language across the globe. English books are the common available medium of literature and information that is accessible to everyone.
English in Sri Lanka is fluently spoken by approximately 23.8% of the population, and widely used for official and commercial purposes. It is the native language of approximately 74,000 people, mainly in urban areas. Nowadays English is one of the main subjects of school syllabus in Sri Lanka. There are three categories of grammar, spoken and listening lessons in the school English pupils' book. So most of the time teaching methods are different from urban province schools than the rural province schools. According to that reason those students' knowledge was different.

With the rapid enchantment of technology, the use of mobile phones has increased a lot. Therefore the modern society tend to use mobile phones to make their lives easier rather than sticking in to traditional methods. The most important fact is that this is followed not only by adults but also by teenagers and children. The best example is that most of the parents use online learning platforms such as Youtube to teach students using videos. Therefore generally students who are in grade 06 are not interested to study using books but by using online methods specifically related mobile technologies. When compared to the global academia most of the developed countries use this method which has brought them positive results when used within the limits. But in Sri Lanka the assets available for the students, specifically those who are starting there secondary education is very much limited.

1.4.1 Main Objective

The main objective of this research is to develop a hybrid solution to improve and evaluate the Spoken English, Written English, English Listening and English Reading abilities of the grade 06 students. To bring out this main objective in a more creative and an effective manner, the complete solution is divided into four major components. The first component focuses on improving and evaluating the Written English knowledge of students while the second component focuses on improving and evaluating the Spoken English knowledge of the students. The third component is for the convenience of the mentor where the mentor can view the results in an organized dashboard and also it is possible for the mentor to predict the marks for each exercise depending on external factors. The fourth component is a hybrid game which includes a brain development and an interactive vocabulary development game. As a whole the product was developed by taking the above mentioned four research areas as the main objectives. In addition to the main objective several specific objectives were designed to increase the productivity and efficiency of the product while maintaining the commercial quality.

1.4.2 Specific Objectives

The ability to predict individual success in exams and courses has been researched this. Accurately predicting students' exam or course grades has the potential to help students in various ways; by using accurate predictions we can detect early on students who have difficulties with the course materials and help them to improve. Moreover, using this kind of prediction technique can help in several other education-related areas .

- Collection of student details relevant to the grade 06 .
- Refer all the lessons of grade 6
- Categorization of lessons.
- Maintaining the functional independence of the component to make sure that the final product can be switched accordingly.
- Implementation of a user-friendly environment to allow the users to operate the system with minimum knowledge to gain maximum Performance.
- Ensure that the designed solution will not make any difference to the cost estimation of the final product.
- Development of the prediction and teacher dash board component to function based on minimum resources consumption but maximum efficiency.
- Ensure that the solution is easily accessible and portable while maintaining the security.
- Adoption of a commercially valuable development structure and a sustainable outcome.

One common approach for solving this type of prediction problem is to extract as many attributes as possible, sometimes as many as hundreds. By evaluatinthe value of each attribute, researchers can attempt to predict exam grades or other variables using linear regression or multiple regression methods. Usually, when using regression, one tries to predict the dependent variables' values using independent attributes of different types. The number of independent variables is very large and includes age and gender marks of the term test, educational level of parents , emotional and social factors , and even the complexity measure of teachers' notes . Other methods used in tackling the grade prediction problem are the factor analysis or other classification schemes with statistical analysis student's exam grade. We also demonstrate - by using multiple regression and machine learning that other social parameters are also influential in determining a student's grade. Moreover, they can help predict which students are likely to fail the test.

SYSTEM ENVIRONMENT AND DESIGN

The objective of this work is to create a web service for the prediction of students results based on their criteria such as their attendance, illness, previous academic scores. Additionally Python and R programming tool are used for the complete execution of the work.

The steps to be followed for the students result prediction are given as follows:

A. Data set preparation:

A data set is prepared in the form of csv file to give training to the machine and testing it.

B. Data splitting:

The given data has to be split for data transformation. First 70% of the data can be used for training the model and the remaining 30% of the data is used for testing the model.

C. Training model :

Train the machine model based on the data set. The train model has two inputs: 1) ML algorithm, 2) 70% of the data split provided by the user.

D. Scoring model:

We have to use score model. This model has two inputs of data: 1) train model, 2) 30% of the data split provided by the user.

E. Evaluation model:

This model evaluates the score results and calculates the machine learning parameters: true positive, true negative, false positive, false negative, receiver operating characteristic (ROC), precision, recall, accuracy, and F1 score.

F. Web service deployment and publishing:

The entire model is deployed in the form of web service so that anyone can access this system. For that, Cortana Intelligence gallery is used here.

IV. SYSTEM EVALUATION

Microsoft azure machine learning studio publishes models as web services that can easily be consumed by custom apps or business intelligence (BI) tools such as Excel. To develop a predictive analysis model, we can typically use data from one or more resources transform and analyse that data through various data manipulation and statistical functions and generate a set of results [10].

Let A be true positive, B be false negative, C be false positive and D be true negative. Then accuracy, precision, recall and F1 score can be calculated using the formulae given below:

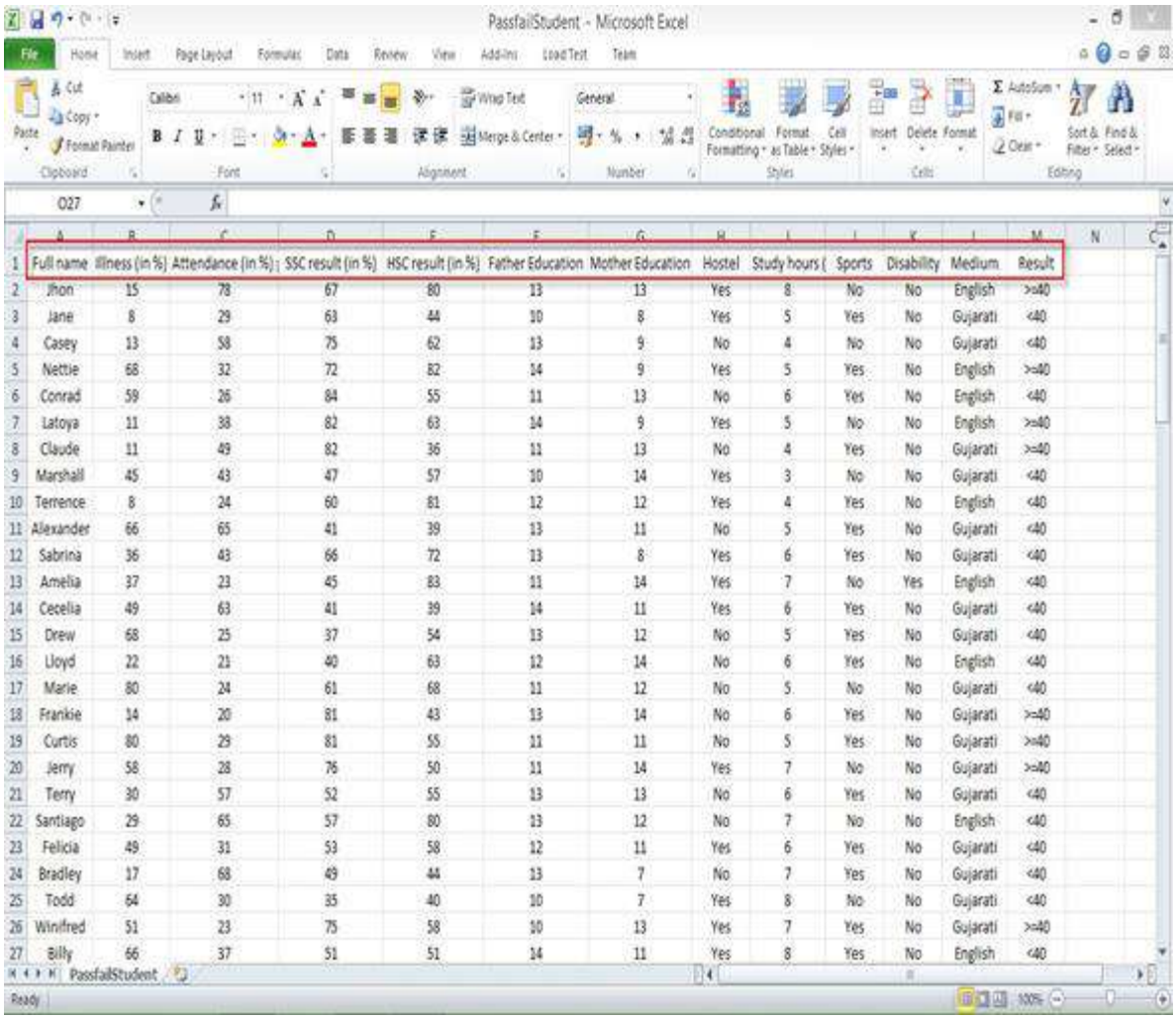
Accuracy = (A+D)/(A+B+C+D)-------- (1)

Precision = (2)

Recall = (3)

F1 score = 2 * (4)

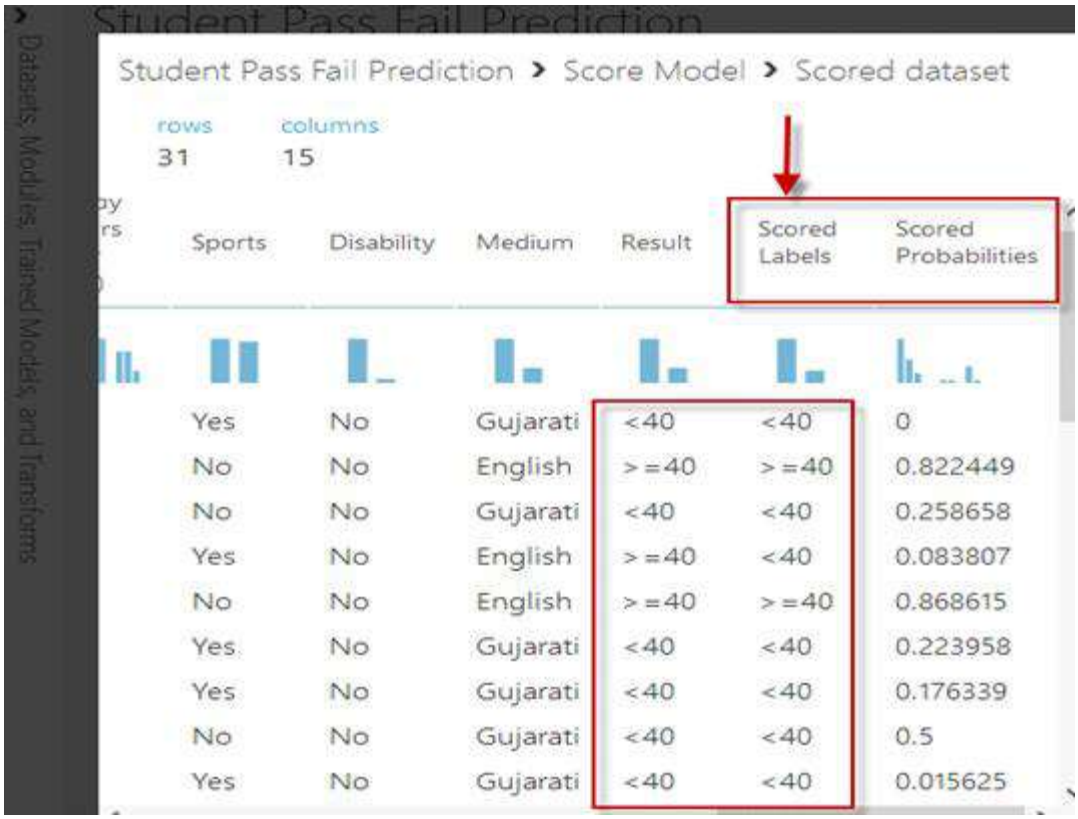
Initially a csv file is created with the fields: 1) Student full name, 2) Illness in %, 3) Attendance in %, 4) SSC result in %,
5) HSC result in %, 6) Fathers education, 7) Mothers education, 8) Hostel staying, 9) Study Hours, 10) Sports, 11) Disability, 12) Medium studied, and 13) Result of score. The result of score can be >=40 or <40 based on teh other criteria which has to be manually entered. The figure 1 shows a sample csv file for student data entry.



PassfailStudent - Microsoft Excel

Full name	Illness (in %)	Attendance (in %)	SSC result (in %)	HSC result (in %)	Father Education	Mother Education	Hostel	Study hours (	Sports	Disability	Medium	Result
Jhon	15	78	67	80	13	13	Yes	8	No	No	English	>=40
Jane	8	29	63	44	10	8	Yes	5	Yes	No	Gujarati	<40
Casey	13	58	75	62	13	9	No	4	No	No	Gujarati	<40
Nettie	68	32	72	82	14	9	Yes	5	Yes	No	English	>=40
Conrad	59	26	84	55	11	13	No	6	Yes	No	English	<40
Latoya	11	38	82	63	14	9	Yes	5	No	No	English	>=40
Claude	11	49	82	36	11	13	No	4	Yes	No	Gujarati	>=40
Marshall	45	43	47	57	10	14	Yes	3	No	No	Gujarati	<40
Terrence	8	24	60	81	12	12	Yes	4	Yes	No	English	<40
Alexander	66	65	41	39	13	11	No	5	Yes	No	Gujarati	<40
Sabrina	36	43	66	72	13	8	Yes	6	Yes	No	Gujarati	<40
Amelia	37	23	45	83	11	14	Yes	7	No	Yes	English	<40
Cecelia	49	63	41	39	14	11	Yes	6	Yes	No	Gujarati	<40
Drew	68	25	37	54	13	12	No	5	Yes	No	Gujarati	<40
Lloyd	22	21	40	63	12	14	No	6	Yes	No	English	<40
Marie	80	24	61	68	11	12	No	5	No	No	Gujarati	<40
Frankie	14	20	81	43	13	14	No	6	Yes	No	Gujarati	>=40
Curtis	80	29	81	55	11	11	No	5	Yes	No	Gujarati	>=40
Jerry	58	28	76	50	11	14	Yes	7	No	No	Gujarati	>=40
Terry	30	57	52	55	13	13	No	6	Yes	No	Gujarati	<40
Santiago	29	65	57	80	13	12	No	7	No	No	English	<40
Felicia	49	31	53	58	12	11	Yes	6	Yes	No	Gujarati	<40
Bradley	17	68	49	44	13	7	No	7	Yes	No	Gujarati	<40
Todd	64	30	35	40	10	7	Yes	8	No	No	Gujarati	<40
Winifred	51	23	75	58	10	13	Yes	7	Yes	No	Gujarati	>=40
Billy	66	37	51	51	14	11	Yes	8	Yes	No	English	<40

1.4.2.1. Data entry table



1.4.2.1. ***Output of the score model with the box highlighting the score from the user and the score generated by the score model with score probabilities***

1.5.Audience

The factors that are explained in the previous sections prove that one of the major reasons for not using eLearning in the educational system of Sri Lanka is the lack of focus on developing a solution that is fixed to a specific user group. As an example Duolingo application which is a vocabulary development solution is useful only to those who are keen on speaking irrespective to the objective they use it. Therefore this research is developed with the objective of expanding the audience to motivate the users to use these types of solutions. As a result the solution mainly focuses the development and evaluation of English language of grade 06 students.

2.Methodology

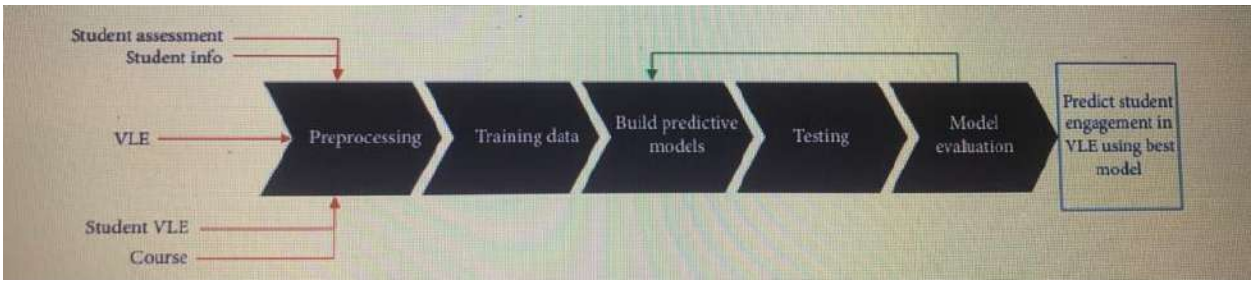
The solution proposed under the research topic is broken down into 4 major components in order to fulfill all the objectives required to cover the prevailing research problems.

1. Written English Module
2. Spoken English Module
3. Prediction and Visualization mentor Dashboard
4. Brain Development and Vocabulary Improvement Game

Nowadays knowing English language increases your chances of getting a good job in a multinational company within your home country or for finding work abroad. It is also the language of international communication, the media, and the internet, so learning English is important for socializing and entertainment as well as work. So for that reason, English learning can be considered so important.

In this study, we utilized various ML techniques to study student engagement in different VLE activities. ,e selected techniques were suitable for both domain and categorical educational attributes. Brief details of the ML, training, testing, and input data are provided below. Machine Learning Technique. Various types of ML techniques have been used as predictive models. ,e ML techniques tested as predictive models in the current study are described below. Decision Tree (DT). A DT has a tree-like structure with internal nodes represented by rectangles and leaves represented by ovals. An internal node has two or more child nodes. ,e internal nodes represent dataset features, and the branches represent the values of these features. Each leaf contains a class related to the dataset . e DT is trained with a training set containing tuples. Finally, the DT is used to classify a dataset with unknown class labels . DTs are primarily used to process information for decision-making ,e tree is constructed from the dataset by determining which attributes best split the input features at the child nodes. In this case, we used the concept of information gain which is dependent on information theory. When a node has minimum entropy (highest information gain), that node is used as a split node . A DT is important when a study seeks to determine which features are important in a student prediction model . ,e rules for DTs are easy to understand and interpret, and we know exactly which classifier leads to a decision. A J48 decision tree belongs to the DT family; it both produces rules and creates the tree from a dataset. ,e J48 algorithm is an improved version algorithm It is a sample-predictive ML model that predicts the target values of an unseen database based on the different values of input features in the current dataset. ,e rules of this approach are easily interpreted. Moreover, this method is an implementation of the ID3 (interactive dichotomize) algorithm and is a supervised ML algorithm used primarily for classification problems. ,e internal nodes of a J48 decision tree represent the input features (attributes), and the branches of the tree represent the possible values of the input features in the new dataset. Finally, the terminal nodes (leaves) display the final values of target variables ,e attribute-selection process is based on the information gain method (gain ratio) . ,e J48 decision tree works for both numeric and categorical variables; moreover, it determines the variables that are best at splitting the dataset ,e attribute with the highest gain ratio reflects the best split point.

Figure 2.1. Data preprocessing



A plethora of pattern recognition methods have been applied to problems in bioinformatics including rule based, statistical methods and machine learning -based methodologies. The goal of machine learning is to train a computer system to distinguish classify cases based on known examples. Machine learning methods include sev-eral widely differing approaches such as support vector machines, neural networks, Bayesian classifiers, random forests and decision trees.In the following discussion we concentrate on machine learning methods as they are nowadays widely used to tackle complex phenomena, which would be otherwisedifficult to handle. Successful machine learning methoddevelopment requires good quality training set. The data-set should represent the space of possible cases. Thisspace is huge for genetic variations as they can have somany different effects and underlying mechanisms. Another aspect is the choice of the machine learning approach. There is not a superior architecture among them. Third, the quality of the predictor depends on how the training has been done, which features are used to and insufficient to adequately describe the pattern in the feature space. Another problem is over-fitting, which means that the learner, due to sparse data,complex model or excessivelearning procedure, describes noise or random features in the training dataset, instead ofthe real phenomenon. It is crucial to avoid overfitting as itleads to decreased performance on real cases. Many predictors provide a measure for the probability of prediction, in this domain a measure of how likely the variation is pathogenic. This information can be used forranking the investigated cases. A more advanced version is to obtain e.g. by bootstrapping an estimate of the standard error of the prediction indicative of the prediction.

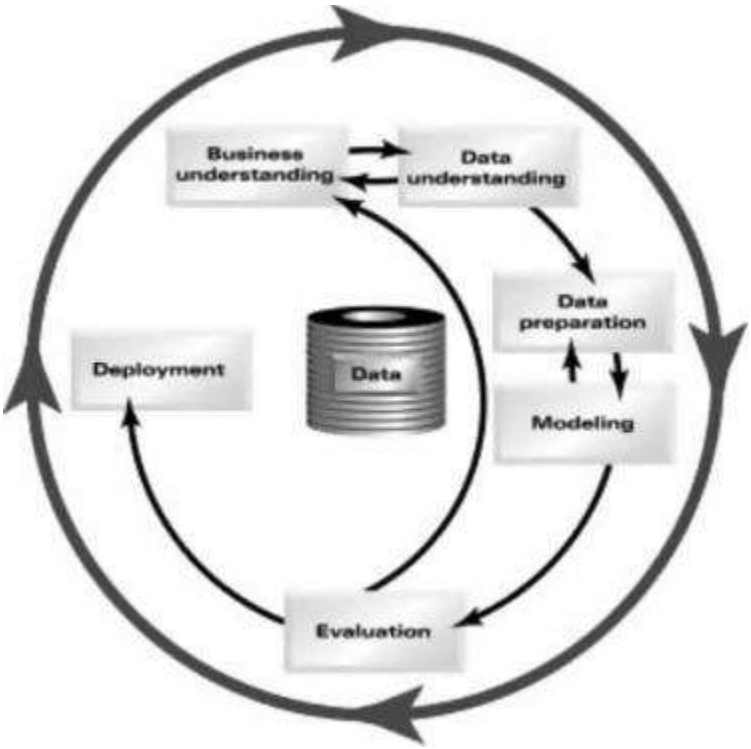
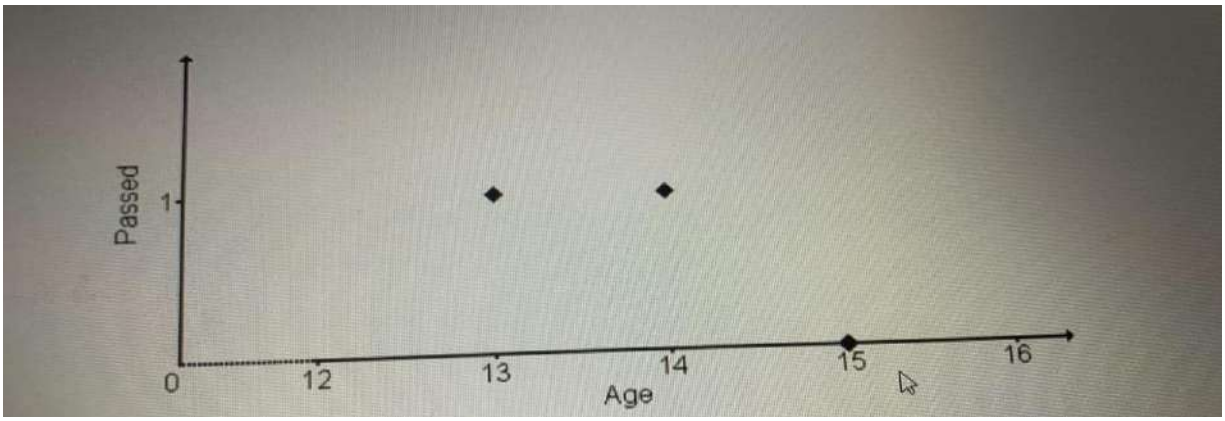


Figure 2.2 data processing cycle

A common definition of machine learning is (Mitchell, 1997): “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.” Basically, machine learning is the ability of a computer to learn from experience (Mitchell, 1997). Experience is usually given in the form of input data. Looking at this data, the computer can find dependencies in the data that are too complex for a human to form. Machine learning can be used to reveal a hidden class structure in an unstructured data, or it can be used to find dependencies in a structured data to make predictions. Latter is the main focus of the thesis. 3.1.2. Predictive analytics Predictive analytics is the act of predicting future events and behaviors present in previously unseen data, using a model built from similar past data (Nyce, 2007; Shmueli, 2011). It has a wide range of applications in different fields, such as finance, education, healthcare, and law (Sas, 2017). The method of application in all these fields is similar. Using previously collected data, a machine learning algorithm finds the relations between different properties of the data. The resulting model is able to predict one of the properties of future data based on properties (Eckerson, 2007)

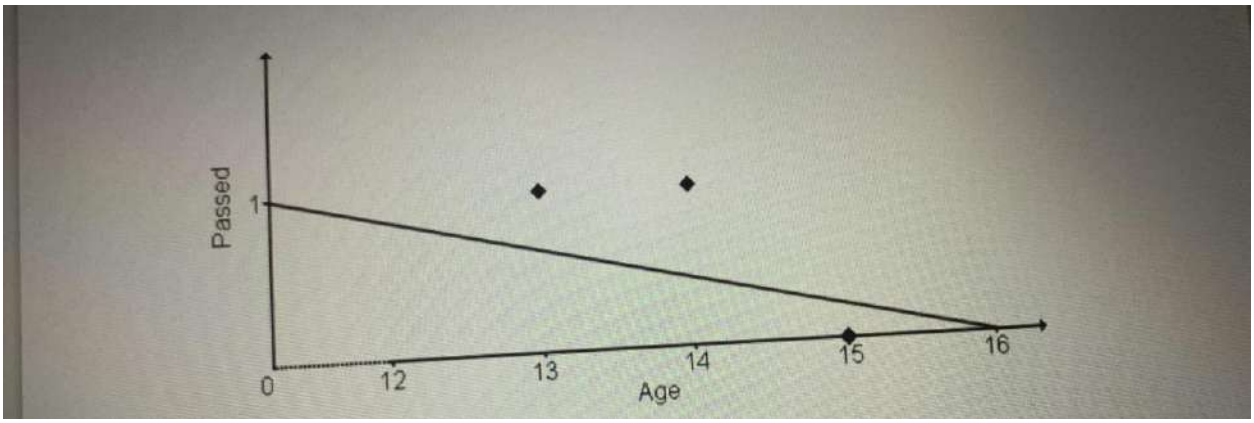
For the sake of simplicity, the data has only one independent variable. Figure 1 depicts a two dimensional graph that shows the relation between the student age and the dependent variable indicating whether they have passed the exam or not.

Figure 2.3 Graph representation of data.



Depending on the type of regression method, regression creates a straight line or a curve that fits the best to the data. Figure 2 shows the graph after the regression.

Figure 2.4Graph after the regression.



2.1 Commercialization Aspects of the Product

Commercialization Factor	Previous Solution	English School Mate
Academic relavance	Acceptable	High
Audience	Not Specific	Specific
Lower resource Consumption	No	Yes
Portable	No	Yes
Cast effective	No	Yes
Evaluatable	No	Yes
Vishualizable	No	Yes
Covers every Subject areas	No	Yes
Low operational knowlage	No	Yes

2.2Testing & Implementation

An appropriate training-test split needs to be decided on. This may depend on the amount of data that the researcher has on hand – whether they have only one season of data, or multiple seasons. Usually professional sport competitions are organized in rounds, with teams playing matches over the weekend. Teams usually play one match in each round unless they have a 'bye'. In the case where one season of data is on hand, the number of rounds that will be used for training the model, and the number of rounds that will be used for testing the model needs to be determined. For example, in a data set with 10 rounds of data, the first 7 rounds of the competition could be used for training the model and the last 3 rounds of the competition could be used for testing the model. However, to obtain a more realistic measure of model performance, round 1 could be used as training to test on round 2, round 1 & 2 could be used as training to test on round 3, round 1–3 could then be used as training to test on round 4, and so on. So, within a season which contains a certain number of competition rounds, we use rounds 1 to $n - 1$ to train our model, and use round n as the test data set, for each round n in N, where N is the total number of rounds in the competition. We thus obtain a classification accuracy for each of these training/test splits, and take an average of the accuracies to give an overall measure of model performance.

Rather than round-by-round prediction, another possibility is to update the training data set after every match has been played. In this case, all past matches up to the current match as training data, and the upcoming match as the training data (i.e. only having that one record as the training data). This is essentially like order-preserved leave-one-out cross-validation. This match-by-match approach is probably not necessary unless teams play more than one match over the same competition round.

Some papers have used multiple seasons of data.

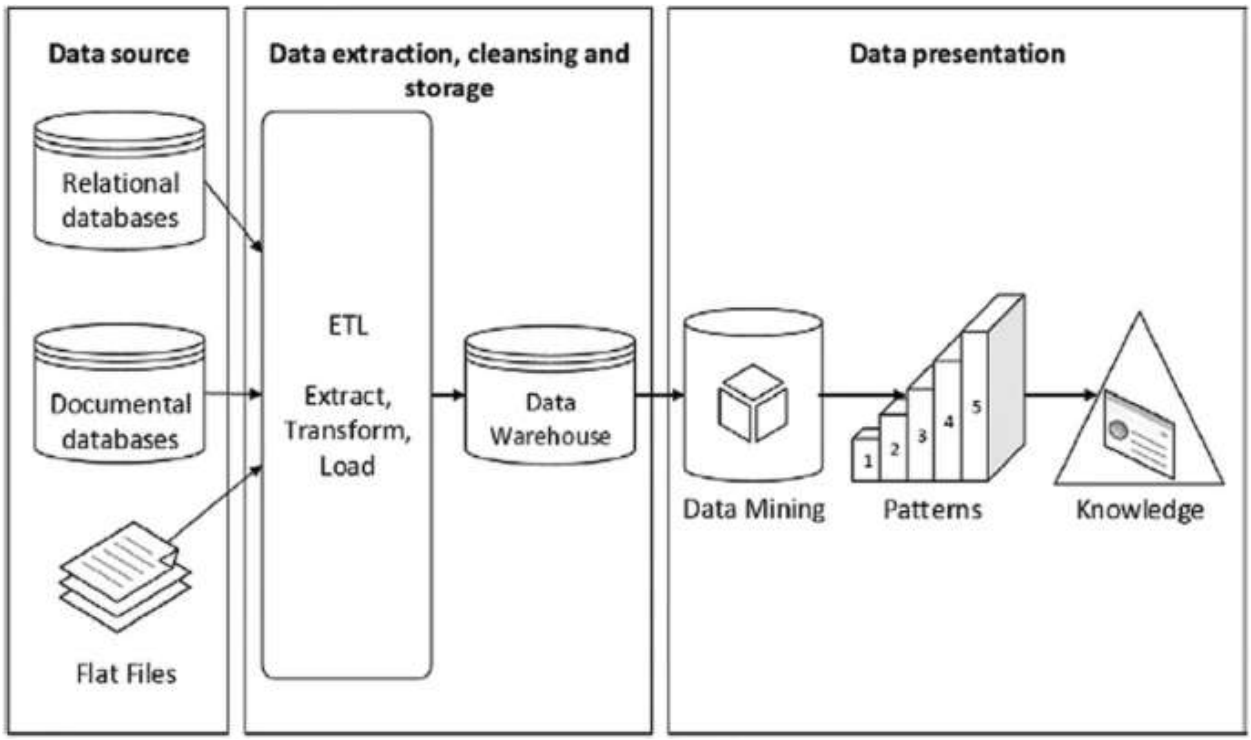


Figure 2.2.1 data presentation architecture

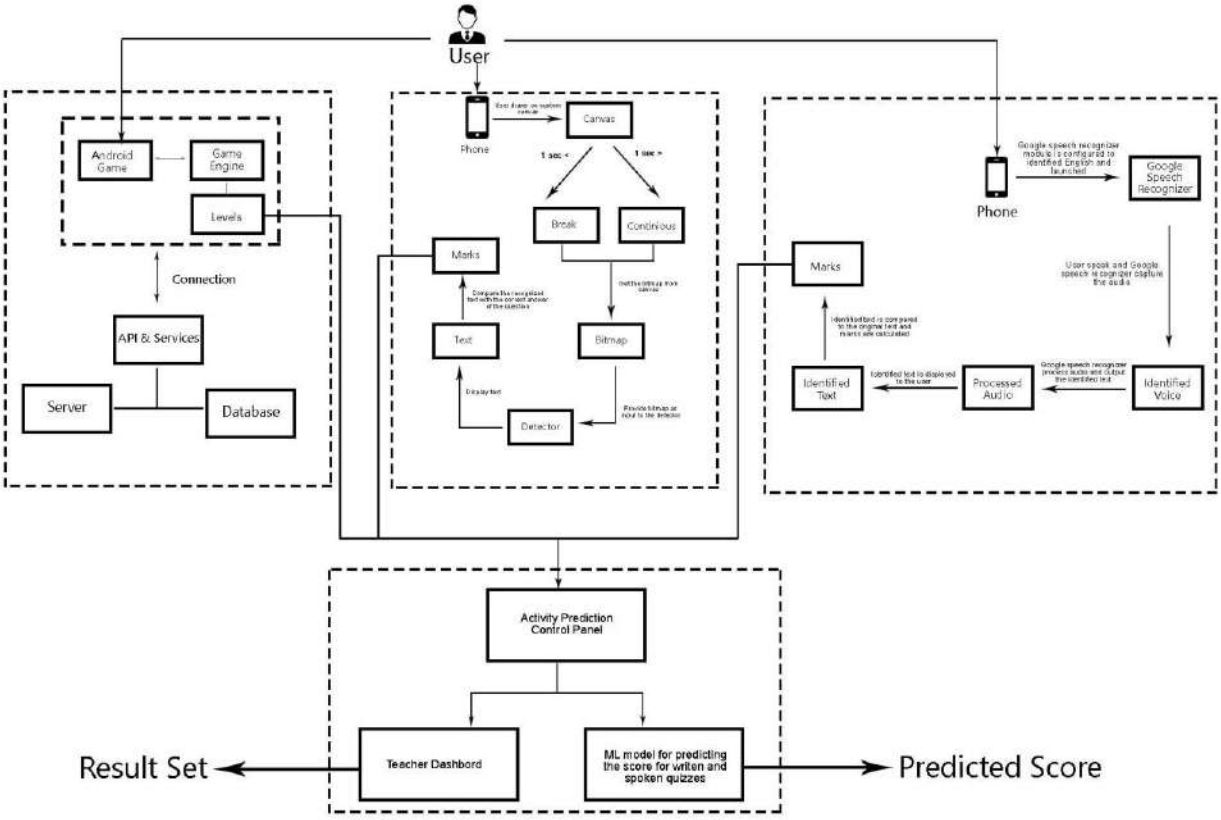


Figure 2.2.2 overall architecture

The main implementation of the Smart Student system has 2 major aspects as the Authentication segment and the Backend and Database. The Authentication segment is implemented using Firebase Authentication as a service. This is composed of a Sign-in screen and a Sign-Up functionality. Further,

Teacher dashboard

- Implemented using Reactjs, bootstrap and axios
- Shows statistics on students' marks and participation for quizzes

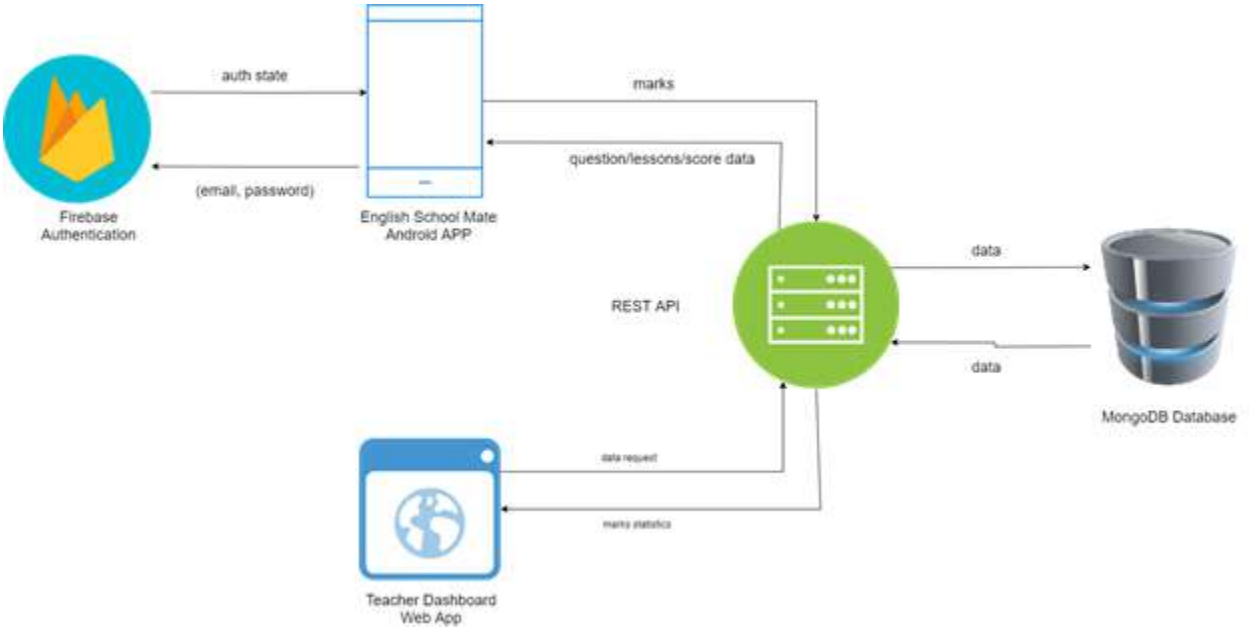


Figure 2.2.3 data flow

ML model for predicting the score

- Data collected through Google Forms
- Data preprocessing and developing a model using Jupyter notebook as the IDE
- Sklearn and pandas libraries will be used for models and utilities needed for data preprocessing, training models, model evaluation

3. Evaluating the User Experience of the Interfaces

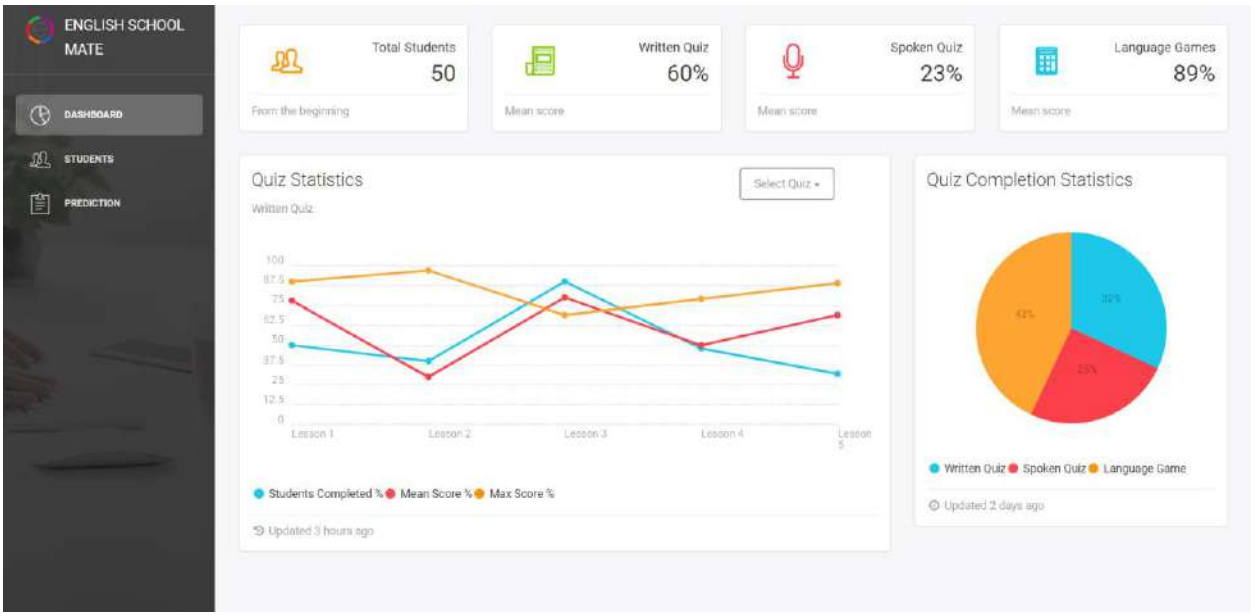


Figure 3.1. Interface 1

This interface shows the students performance.it shows the students all students writing scores, spoken scores, gaming development.
This also have a quiz statistic chart and quiz completion statistic chart.

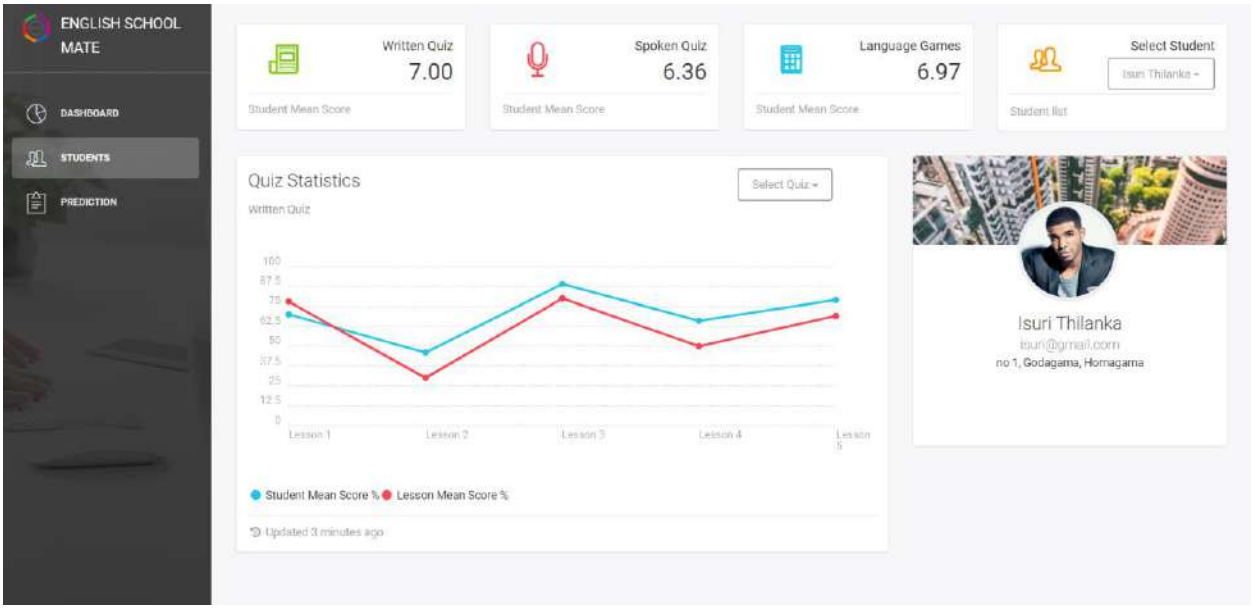


Figure 3.2 interface 2

This shows the student wise performance of each students.

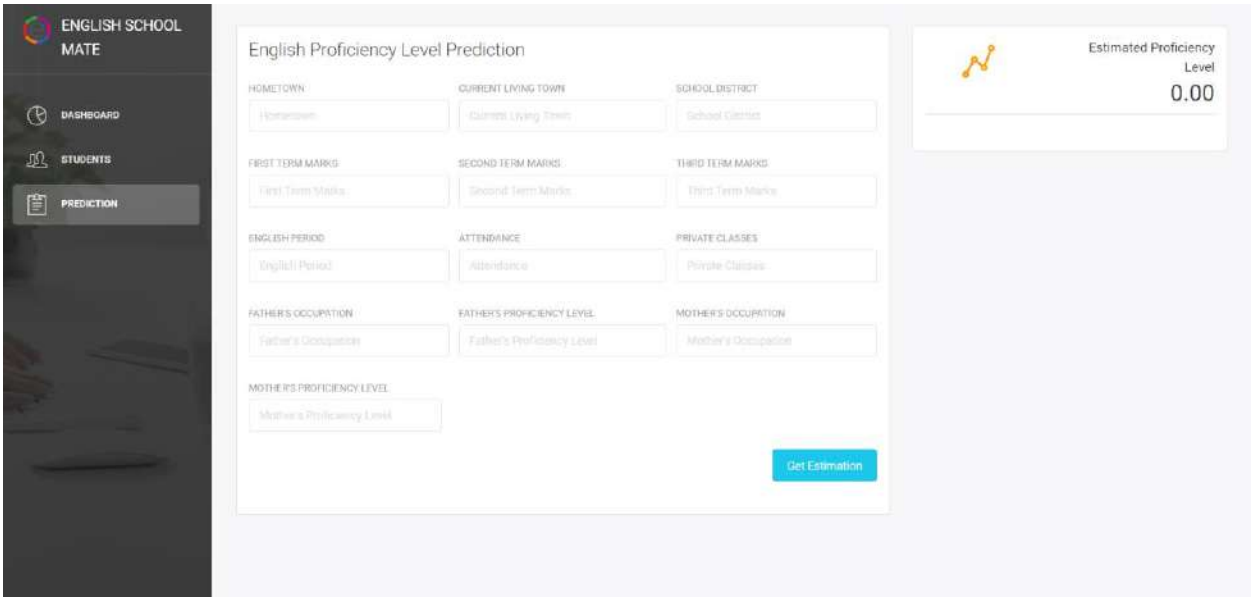


Figure 3.2 interface 3

4.Results and discussion

In this part of the study, we predicted the numbers of students from the different activities of a VLE course using features related to student activity. To answer the research questions of the current study, we performed several experiments. We used the ML algorithms and the Rapid Miner tool to build the learning models, Data Visualization and Statistical Analysis of the Data. To understand the student data, we performed statistical analyses of the number of times students clicked on activities and the student engagement level. We also visualized the dataset. ,is step is important in ML studies because the performance of a predictive model sometimes decreases when the data quality is poor We visualized the input variables (student clicks on VLE activities) of the OU course to illustrate how important the input variables are in predicting.

1.1 Research Findings

· Use the backend part to connect the application and the web portal to improve the scalability of the application

· new way to get predictions.

· In the predictting model using the panda libraries and use many techniques all are mention above.

1.2 Discussion

The focus of this section is to explain how the implemented solution and the results generated are fulfilling the objectives discussed above. As discussed under the sections above, the system was developed as a web and android application to improve the spoken knowledge of the grade 6 students english knowledge and teacher dashboard. The results generated proved that the

accuracy of the solution is 90% which can be considered as the best outcome expected by the research team.

2 Summary

Component	Task
	Designing the user interfaces
	Implemented web application
	Collect data
	Gather data
	Predict results
	Implement teacher dashboard
	Finalize the application

6 Reference

https://www.researchgate.net/publication/332133485_Students_E-Learning_Performance_Improvement_and_Predicting_the_Students_Learning_Interest_Using_Data_Mining