Smart Learning Guidance System for University Students

Pasan Kamburugamuwa
Department of Software Engineering
Faculty of Computing
Sri Lankan Institute of Information Technology
Malabe, Sri Lanka

Walakulu Gamage A.I.S
Department of Information Technology
Faculty of Computing
Sri Lankan Institute of Information Technology
Malabe, Sri Lanka

Dissanayaka W. R
Department of Information Technology
Faculty of Computing
Sri Lankan Institute of Information Technology
Malabe, Sri Lanka

Ahamed M. M. I
Department of Information Technology
Faculty of Computing
Sri Lankan Institute of Information Technology
Malabe, Sri Lanka

Ms. Lasantha Abesiri
Department of Information Technology
Faculty of Computing
Sri Lankan Institute of Information Technology
Malabe, Sri Lanka

Ms. Nadeesa Premadasa
Department of Information Technology
Faculty of Computing
Sri Lankan Institute of Information Technology
Malabe, Sri Lanka

***Abstract*- The motivation of this research paper is to find solutions for student problems faced during their academic years in college. In this paper, we used machine learning algorithms, feed-forward neural networks, and regression models to predict student's final GPA based on their grades in previous semesters, and calculate the estimated workload for each student based on the past academic results. The main contribution of the paper is that we have used various models to predict and analyze each function and get an accuracy of almost above 75% for each module. To preprocess data, we have used NLP techniques, applied the linear regression algorithms and the bag of words model to discover classification rules. We extracted useful knowledge for final GPA and identify the most important courses in the student's study plan based on their grades in the mandatory courses. Another main reason having this research paper is that the exponential growth in using the mobile devices market over the last decade and chatbots are becoming an increasingly popular option while this can be used for educational purposes as well. This study examined educational chat-bots for helping students to achieve the expected GPA by giving real-time answers to student questions, predict the final GPA of students based on past results, and calculate the student workload.**

Keywords—GPA, Educational data mining, classification, NLP (Natural Language Processing)

I. INTRODUCTION

The availability of educational data has been growing rapidly, and there is a need to analyze huge amounts of data generated from this educational ecosystem for something special and useful. Machine Learning and Natural Language processing can be used to analyze those data and make predictions about future. Using these techniques we can predict quite accurately the outcome of the student's final CGPA, student estimated workload for each subject, help students to identify the lecture that the questions came from and answer to real-time questions from the student perspective.

Most of the problems faced by undergraduates because of a lack of knowledge about lecture content, unable to communicate with other people about academic questions and not having proper time management for each subject [1]. Lack of having motivation towards the academic career may also lead to a lower mark of undergraduates. Why Grade Point Average (GPA) really important in student university life? The answer depends on what students seeking to do after university life. If someone plans to continue in academia by seeking a professorship, Grade Point Average is a huge effect on him/herself [2]. Other than that most of the companies try to get the most bright student abroad as they have a good recommendation from the university also. According to past research about 60% job recruiters, Grade Point Average (GPA) is the major role for the initial screening of graduates [3]. On the other hand, continuous failure of exams by university students is the main problem for both academic staff and the students[4]. Machine learning techniques can be used to forecast the performance of the students and identifying the at-risk students as early as possible so appropriate actions can be taken to enhance their performance [5]. One of the most important steps when using these techniques is choosing the attributes or the descriptive features which used as input to the machine learning algorithm. The attributes can categorize into GPA and grades, academic progress, and educational background.

According to the findings from the research, time spends for studying influence on Grade Point Average (GPA) of third year commerce and management students of eastern university, Sri Lanka than the language barrier. Therefore, time management on studies leads to high level of Grade Point Average (GPA) [6]. To find the strengths and weaknesses of students' perspectives in E-learning and Higher education was conducted by Neuza Pedro [7]. In this research, they have been

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

able to find the key features which are affecting the student's academic performances.

These are the main reasons that the Smart Learning concept is gaining momentum. There are several valuable types of research are conducting with the ambition of upgrading the student's academic results by moving from traditionally based analysis to computer-based analysis. Such studies pointed out that, provide smart services to the student using Machine Learning will improve the academic result of the students [5]. Therefore, we found that it is important and urging consequent to find a smart method to provide proper guidance to undergraduates from the student's perspective and make the system useful for them [6]. The main motivation for this research is to uplift the academic performance of undergraduates by using the academic results of undergraduates. So throughout this research paper, we will discuss these four components and how they have implemented, and what are the impacts of them to the university students.

II. RESEARCH OBJECTIVES

The main objective of this research is to guide students through the assessments and enhance the GPA accordingly by identifying the difficulties faced by them with how much effort they should put in order to achieve the target. To implement this, there will be a mobile application along with the web application to get data and display the processed data. The proposed system will use the machine learning algorithms along with the artificial intelligence to get the results.

1. Finding which lecture related to the question, which is effective in uplifting the academic performances of university students.
2. Predict the academic results regarding the past results using the machine learning algorithms.
3. Number of hours he/she must be put in effort in order to achieve the target weekly.
4. Give real-time answers to students questions based on the lecture content and other academic activities.

Showing weaknesses in each module will help to get an overview of what needs to pay more attention. The GPA prediction is very important to students as they can work early on the semester to achieve the expected results at the very end. Another main reason and the main objective of this research is to show how many hours they should work to achieve academic excellence to come to their goal. Humans like interactive introduced in order to get real-time answers to questions they are facing during academics. This will save the time and effort to find solutions to them.

So these are the main objectives of why this research is important to university students. Through these, they will be able to get solutions to the main problems they face during the academic period at the university. Early prediction of results by analyzing the current progress of the academics and past results of other will students will help to make adjustments to their study plans.

III. METHODOLOGY

The system focuses on uplifting student performance by using different aspects of student interaction with the academic process. AI-Based Mark Advisor and Lecture content identifier is the main approach that students can find academic progress of each student. In the AI-Based Mark Advisor module, we have used the linear regressions to analyze the data set which is taken from 4th-year students of faculty of computing, Sri Lanka Institute of Information Technology. We have taken 6 semesters results and each semester has subjects interrelated to each other. By getting around 200 records of each student's GPA, the predicted results were calculated. To that, there was some conversion of data columns into meaningful words. One of them converts the student's grade into range (Ex- If a student gets A for one module, then the mark range is 75-90). This allows predicting the results based on value rather than a single alphabetic character. The final results are the prediction of results (predict marks) in one subject based on the inter-related modules throughout the semester.

The other module, lecture content identifier used the natural language processing to identify the keywords in both the questions and lecture content. So this will help to identify where each question come and it will be a great help to refer these facts later on without going through the lectures every-time. This is mainly focused on the mid examination of the university as in there the students most probably get the MCQ questions. Students will be given the calculated results on each lecture content based on how he/she faced the examination.

The last module is the AI-based human-like interactive university chat-bot. To build this module, we have selected questions and answers from the object-oriented programming module from the 2nd year module of faculty of computing, Sri Lankan Institute of Information Technology. The Chabot can give real-time and most accurate answers to student questions by using the technology of natural language processing and machine learning. The models used here are a bag of word models and the feed-forward model. To train this module, we have used around 300 questions and answers related to the module object-oriented concepts.

To conclude, the entire system is supported by these four components which may help the undergraduates to achieve the expected GPA. 1) AI-Based Mark Advisor 2) AI-Based Lectures Identifier 3) AI-Based Human-Like Interactive University Chat-Bot and 4) Estimated Student Workload Calculator. In this section, there will be a brief introduction of what we have done to implement those components and also the machine learning modules we have used to build these components.

A. *AI-Based Mark Advisor*

When students are enrolled in a Bachelor's or Master's Program, Student GPA is the key factor which measure the student performance during academic years. The below table 3.1 showing the grade point values and their related grades. This is taken to consideration when predicting upcoming semester GPA based on the past results.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

TABLE I

Marks	Grade	GPV
90 % and Above	A +	4.00
Between 80% and 89%	A	4.00
Between 75% and 79%	A -	3.70
Between 70% and 74%	B +	3.30
Between 65% and 69%	B	3.00
Between 60% and 64%	B -	2.70
Between 55% and 59%	C +	2.30
Between 45% and 54%	C	2.00
Between 40% and 44%	C -	1.70
Between 35% and 39%	D	1.30 (Fail)
Between 30% and 34%	D+	1.00 (Fail)
Between 0% and 29%	E	0.00 (Fail)

Table Showing the grade points average

Calculating the GPA of each student is a major task when giving the GPA for each student. The below figure 3.1 shows the final grade point value calculated based on the past results of students' subjects. This is taken into consideration when predicting the upcoming semester GPA based on past results.

$$GPA = \frac{\sum \left(Course\ GP * Course\ Credits \right)}{\sum Gradable\ Credits}$$

Figure 3.1 GPA Calculate Example

To build this module, we have taken 300 records of student past data and converted those grades (ex – A grade in to 75-90 range of marks) in meaningful way. Then apply the linear regression model which is capable of analyze these results to predict the GPA of students. This is done using the only the matching subject modules and get the inter-connection between them.

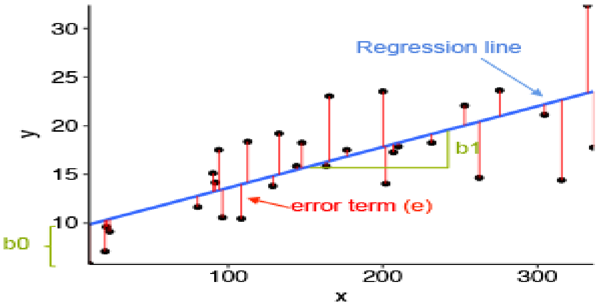


Figure 3.2 Linear regression model data representation

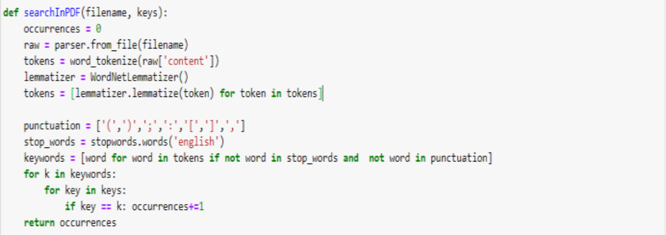
The algorithm is capable of predict grades for each subject using past similar results. There is a 89% accuracy of the predicted GPA when comparing with the real value GPA get from the actual dataset.

B. *AI-Based Lecture Identifier*

The biggest problem faced by the students is that they are unable to get the marks they expect within the exam*[10]*. Students do not identify their weaknesses, the lectures, or areas they are weak at. So there should be a proper way to mention the questions are coming from which lecture and what percentage is weak or strong in that particular lecture. Students have to be well organized and be ready to answer any question they come across. If a student can identify the areas that they should pay more attention, they will be able to score more in the exam. So in through this component, the main objective is to help the students to identify and find wherefrom the questions coming from by analyzing questions from the Natural language processing. Self-learning bots are the ones that use some Machine Learning-based approaches and are more efficient than rule-based bots. So we are focusing on building the chat-bot as a self-learning bot by using machine learning algorithms. To build the interaction between the human language and computers, we use Natural language Processing(NLP).

NLTK(Natural language toolkit) is the platform we are using in order to work with human language. This provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.

In below figure 3.3 shows the clearing of data and apply the natural language processing in order to get the most probable lecture the question which comes.



```
def searchInPDF(filename, keys):
    occurrences = 0
    raw = parser.from_file(filename)
    tokens = word_tokenize(raw['content'])
    lemmatizer = WordNetLemmatizer()
    tokens = [lemmatizer.lemmatize(token) for token in tokens]

    punctuation = ['(',')',';',':','[',']',',']
    stop_words = stopwords.words('english')
    keywords = [word for word in tokens if not word in stop_words and  not word in punctuation]
    for k in keywords:
        for key in keys:
            if key == k: occurrences+=1
    return occurrences
```

Figure 3.3 showing how the analysis done

So at the end of the there is 90% accuracy for finding the which lectures are related to the questions coming from.

C. AI Based Human-Like Interactive University Chat-Bot

This component will be able to answer the student questions based on both academic and non-academic. Natural language processing is used to build this model along with machine learning techniques. Bag of word model and feed forward model are used to build the model which is capable of giving the most probable answer to student question based on the lecture content. Below figure 3.4 shows the feed forward model which is used to build this unitbot.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

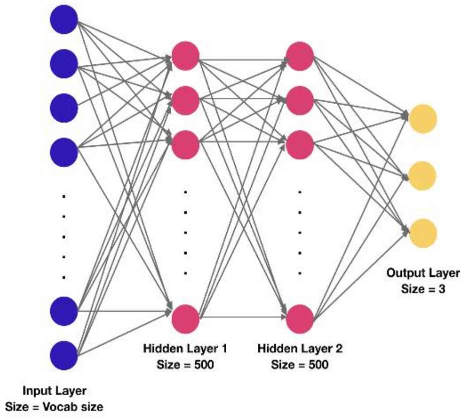
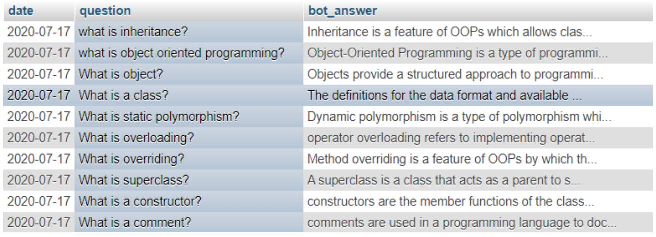


Figure 3.4 Process of the feedforward model

To feed-forward model, the input data is coming from the bag of word model which keeps the count of the total occurrences of most frequently used words. There are two hidden layers and the output size is the number of classes which is having the probabilities for each class. The most having probability will get as the output class and randomly select the answer within that class.

The output is very accurate in this process and the data is stored in the database and the below show the questions and answers getting from different questions asking from the uni-bot.



date	question	bot_answer
2020-07-17	what is inheritance?	Inheritance is a feature of OOPs which allows clas...
2020-07-17	what is object oriented programming?	Object-Oriented Programming is a type of programmi...
2020-07-17	What is object?	Objects provide a structured approach to programmi...
2020-07-17	What is a class?	The definitions for the data format and available ...
2020-07-17	What is static polymorphism?	Dynamic polymorphism is a type of polymorphism whi...
2020-07-17	What is overloading?	operator overloading refers to implementing operat...
2020-07-17	What is overriding?	Method overriding is a feature of OOPs by which th...
2020-07-17	What is superclass?	A superclass is a class that acts as a parent to s...
2020-07-17	What is a constructor?	constructors are the member functions of the class...
2020-07-17	What is a comment?	comments are used in a programming language to doc...

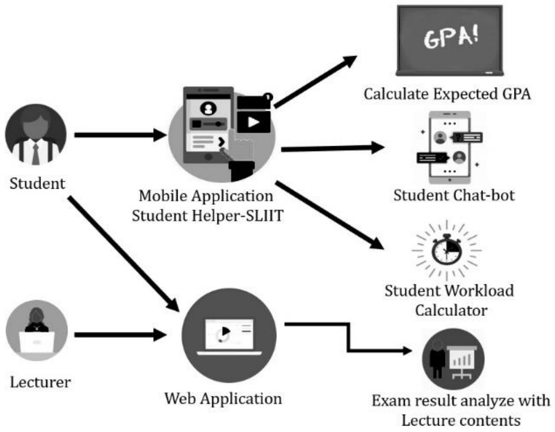
Figure 3.5 showing the requests made by students and answers given through chatbot

D. Estimated Student Workload Calculator

The estimated student workload will be calculated by considering the number of hours that should student focus on a particular subject (mentioned in the each module outline) and also considering the past results for the similar kind of subjects that the student has obtained.

In order to build the sub-system, we have gathered past marks of the student as input data through the application which helps to analyze position of the student for similar past subjects. The position will be measured by grades and the number of hours will be depended on that. When a student selects a subject here, the relative cluster which the subject belongs to will be sorted and from the grades on that cluster an average number of hours will be suggested. To calculate the above estimated time, we have used the machine learning techniques and linear regression models. The parameters are the number of

So after combing these four components, there comes a full student guidance system which is capable of helping academics in university.



In figure 3.6 shows that the full functioning system along with all the components.

As these components are working fine, then the system will function fully as expected by integrating all the components. Overall, we have used machine learning, natural language processing, and artificial intelligence in order to take the highest accuracy within the components. As all the machine learning models have used the python as the backend we have created APIs to interact with the frontend of our system. The frontend of our application is being built using the react-native mobile application and the ionic as mobile and web application technologies respectively.

IV. RESULTS AND DISCUSSION

Four components in this research have achieved a more than 75% accuracy and the system is fully functioning. The overall performance of the system will directly influence the students to achieve the expected academic results soon as these models are well tested.

In an AI-based mark advisor, we used multiple linear regression models to train the system and make predictions. The model is trained by dividing the dataset which is collected from the student survey and it is divided into 80% for training and 20% for testing. So at the very end, the model is giving 89% accuracy with the capability of predicting the student cumulative grade point average for a certain module based on certain subjects marks. The reason why we choose the multiple linear algorithms to build this module is,

1. Predict the results based on the value of two or more than other variables.
2. The value of the dependent variable at a certain value of the independent variables.

So final outcome will be having the cumulative Grade Point Average while this system is capable of giving the final output results by having the linear regression models which takes the input as the selected subject marks in the particular year and get the results for the final semester year exam.

The AI-Based lecture identifier is having over 90% accuracy over identifying the lectures that the questions belong to. This is highly needed for students as if the system can mention what are the student weaknesses and really good areas, then the students can easily find solutions for them. The system can

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

make reports with the student's performances and on each subject make pie charts showing weaknesses and strengths in these subjects based on the questions and answering.

As the input to the system, we apply question and answers, then apply the natural language processing tasks like tokenization and lemmatization in order to text preprocessing. We have used these methods as these are highly tested and giving more accurate results. Then we put all the words in the questions and also answers to a separate bag of word models and get the highest probability that the words appear in each question to belong to the lecture. So through this method, there is the method which highly predictive as this method is capable of finding the best possible lecture that the questions asked. If we make the algorithm using any other method, as we previously done using the sentence splitting and text wrangling it takes much code lines which is head to take much more time in executing the program. So this by using tokenization and lemmatization we can save lot of code lines and also it executing time.

The main objective of the bag of word model using in here is that the keep probability of having the most used keywords and compare it with the lecture content keywords. The system is capable of making a prediction using the keywords and give the best possible lecture that the question comes by using the bag of word model. If we use any other method like word2vec, there using a very small number of texts and if we want to keep the highest amount of a number of texts, the word2vec model is not suitable. Another straight forward and classical method which gives decent to good results in practice is Latent Semantic Indexing (LSI). This method is based on a low-rank approximation of a matrix of BoW vectors of the corpus and using the approximated singular vectors as document representation. So by comparing these models, the best model which is fit to analyze questions and answers is the bag of word model.

The third component in the student learning guidance system for university students is an AI-based human-like interactive university chat-bot. As this one is using the natural language processing for analyzing the real-time questions of students and answers according to them, there is the high accuracy of giving the exact answer to the student question based on the pool of questions and answers that the chat-bot have trained.

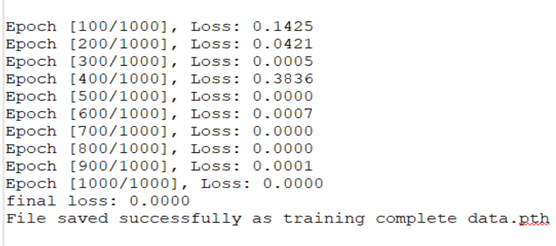
To preprocess the text we have used the method of stemming and tokenization. The stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots. This allows in identifying the words specifically from many number of words. These methods will save time for the execution of the code and execute the training process for the minimum amount of time. Then these processed data are passed to a bag of word model which is highly capable of keeping a large number of texts and then pass to the feed-forward model. As this model is capable of capturing the more complex representations like,

1. Image recognition tasks
2. Natural Language processing
3. Bio- informatics task.

So this model can be used to represent more complex functions easily is the main advantage that we can get from the module. If we used the rules-based chat-bots, then the system can only answer only the pre-designed rules. Main disadvantage of having the rule-based chat-bots are they are using the graphical user interface where a bot builder will design paths using a decision tree. But as we used the artificial intelligence to build this chat-bot, this model is capable of automatically learn after an initial training period by the bot developer.

Other than this module is getting support from the GPU as the training time for the bot will become low and it helps to quickly the deploy the update where as possible.

As figure 1.4 in the above, the system is 90% accuracy and the system can find the most probable answer to the question ask from the chat-bot.



```
Epoch [100/1000], Loss: 0.1425
Epoch [200/1000], Loss: 0.0421
Epoch [300/1000], Loss: 0.0005
Epoch [400/1000], Loss: 0.3836
Epoch [500/1000], Loss: 0.0000
Epoch [600/1000], Loss: 0.0007
Epoch [700/1000], Loss: 0.0000
Epoch [800/1000], Loss: 0.0000
Epoch [900/1000], Loss: 0.0001
Epoch [1000/1000], Loss: 0.0000
final loss: 0.0000
File saved successfully as training complete data.pth
```

Figure 4.1 Chat-bot training process.

As the results showing in the figure 4.1 showing the training process of the chatbot. Bag of word model and the feed forward model is the best way to make the chat-bot for this kind of university chat-bot which is able to giving the real time answers.

The proposed workload calculator sub-system is capable of calculate the estimated workload of each student based on the lecture contents. In there most accurate time duration that the student should be work to achieve the expected results will be calculated by linear regression algorithms. We have used the student marks for several semester subjects which is inter-related to each other and then module outline information to calculate the expected time duration to calculate the time duration that each week that the student might need to work to achieve the expected results. The sub-system is much accurate which is able to give the real-time calculation for each students.

V. LIMITATION AND CHALLENGES

There are many hardships faced during the implementation of a smart student guidance system for university students. Because of a lack of sample data and a lesser number of resources to train the machine learning models we had to survey the faculty of computing, 4th-year students of Sri Lanka Institute of Information Technology to prepare the dataset. The prepared dataset contained around 400 records of student data which included the semester grades of each student. Then we had to convert those grades into marks range. Other than that, making the matching between different courses is the other main problem which we have got. To overcome this challenge we have given points to each subject based on the year, the number of hours allocated for that subject in the course outline, and the marks obtained by the students for that subject in past. After analyzing these we could prepare the dataset which we highly needed to feed the machine learning models.

One other reason that we had to face during the implementation of the system is making an algorithm for the workload calculator for students and test the system with data. To proceed with the above machine learning algorithm, we had

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

to collect the data related to the module outline of each subject and make the algorithm that inter-related with these parameters. In there the accuracy could not calculate as this is one is calculating the time duration for each student to calculate.

Other than that in making the chatbot, making the dataset is hard to find as it is hard to train for all the modules. To overcome this challenge, we only took the Object-Oriented Concepts module questions and answers and trained the system according to them. There were 400 questions selected and trained the algorithm accordingly. The training time takes much time and we could reduce the time by getting the GPU support from the computer.

These are the main limitations and the challenges that we face during the implementation of each module.

VI. CONCLUSION

Smart Student Guidance System is essential for undergraduates in order to find the best possible solutions to student's questions arise in the academic years. The student GPA analyzer is capable of making predictions of the next semesters GPA based on the past data which will help students to choose the modules and also subject streams most of the time. This also helps the academic staff to make future decisions to adopt considering the current student performance in the certain batch. The multi - linear regression in this module make the key role in this part as it gives the most accurate GPA prediction for the students.

In AI-based lectu re identifier makes identify the which questions comes from which lecture easily. The natural language processing techniques like tokenization, lemmatization and bag of word model keep the final output of the system most suitable and accurate to the system. Students can identify the weak points and also from where the questions are arising easily through this module easily.

In the uni-bot implementation we have used the tokenization, stemming and lemmatization to preprocess the data and feed to the bag of the model which is implemented to get the highest probability words that occur in the selected bag of words. Then these processed data is passed to the feed forwarding model and find the most suitable sentence by using the question asked by the student.

So as we discussed in the methodology part, the loss error in this module is zero. The best technology that we have used to build this component is using the above-mentioned techniques and also the feed-forward model. With the help of the estimated workload calculator, the students can easily find how much time he/she should work on the particular subject within that week according to his performance in the previous semesters. To build this model we have used the multilinear regression model as this mode is using the multiple variables to calculate the final output.

As we conclude, all the models doing an excellent performance in order to build the smart student guidance system for undergraduates. The combination of all the components helps students to make use of the system in various ways. But one of the limitations of the system is unable to have included the voice-based chat system also. So this will be implemented in the future release as this will highly useful for the students.

In future work, we will test the module with more real data and expand the system to cover all the modules in the university. So that will help all students make use of this system for more usefully. Other than that, the voice-based chat will also be implemented as mentioned above and each module can be separately set up for every university based on their requirements is the main advantage that we can get from these modules.

REFERENCES

[1] Stefanie Hassel and Nathan Ridoul. An Investigation of First-Year Students' and Lecturers' Expectations of University Education, US National Library of Medicine National Institutes of Health. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5790796/ [Accessed : Feb 15 2020].

[2] Role of GPA, Before, During and After Grad School. https://www.phdstudent.com/surviving-grad-school/transitioning-to-graduate-study/role-of-gpa-before-during-and-after-grad-school

[3] Nirogini Yogendra and Anthony Andrew, "A Study On The Factors Influencing On Grade Point Average (GPA) With Special Reference To Third Year Commerce And Management Students Of Eastern University, Sri Lanka

[4] Yil,Sayi and Aralik. "The Reasons of lack of motivations from the student's and teacher's voices". The Journal of Academic Social Science 2013.

[5] Erkan Er, Identifying At-Risk Students Using Machine Learning Techniques: A Case Study with IS 100. https://www.researchgate.net/publication/271297360_Identifying_At-Risk_Students_Using_Machine_Learning_Techniques_A_Case_Study_with_IS_100

[6] Walid Mohamed Aly, Osama Fathy Hegazy, Heba Mohmed Nagy Rashad, "Automated Student Advisory using Machine Learning" International Journal of Computer Applications.

[7] Neuza Pedro, E-learning & Higher Education: Strengths and Weaknesses from Students' Perspective. Conference: Proceedings of E-learn 2011 World Conference on E-learning in Corporate.

[8] Patrick Bii,Jackson Too, Reuben Langat, "An investigation of student's attitude towards the use of chat-bot technology in instruction: the case of knowie in a selected high school" Educational Research in Octber 2013.

[9] Laura Silva, "Smartphone Ownership Is Growing Rapidly Around the World, but Not Always Equality." Global Attitudes & Trends. February 2019.

[10] E-learning & Higher Education: Strengths and Weaknesses from Students' Perspective. Available from: https://www.researchgate.net/publication/306272403_E-learning_Higher_Education_Strengths_and_Weaknesses_from_Students'_Perspective [accessed Feb 16 2020].

(This work is licensed under a Creative Commons Attribution 4.0 International License.)