Research Article

Development of MyCGPA for Early Predicting Students' Academic Performance

Muhammad Yazid Abdul Mutalib[1], Zuraini Zainol[2,*], Puteri Nor Ellyza Nohuddin[3], Ummul Fahri Abdul Rauf[4]

[1] Department of Computer Science, Universiti Pertahanan Nasional Malaysia; muhammadyazid250@gmail.com;

[2] Department of Computer Science, Universiti Pertahanan Nasional Malaysia; zuraini@upnm.edu.my; 0000-0002-6881-7039

[3] Institut Visual Informatik, Universiti Kebangsaan Malaysia; puteri@ukm.edu.my; 0000-0003-0627-5630

[4] Pusat Asasi Pertahanan, Universiti Pertahanan Nasional Malaysia; ummul@upnm.edu.my; 0000-0001-6080-528X

* Correspondence: zuraini@upnm.edu.my

Abstract: *One of the primary concerns in higher education is the early identification of underperforming students. To address this issue, the current study proposes the development of a system that would assist academic advisers and faculty management to identify students at risk of low academic performance at an early stage. This system utilises a prediction model based on a dataset of academic and demographic data from the UPNM's Computer Science students. The dataset contains information from 97 students and 21 characteristics. We developed a prediction model for Cumulative Grade Point Average (CGPA) using the regression technique, focusing on three variables: 'activity', 'absence', and 'GPA'. The prototype model was used in the system development process. The findings of this study are valuable for the institution (university), since they enable for the early identification of those who may struggle academically. Future enhancements include increasing the dataset and using more powerful algorithms to predict students' academic achievement.*

Keywords: predictive model; early prediction; machine learning; academic performance; regression.

DOI: 10.5281/zenodo.13370554

1. INTRODUCTION

The evaluation of educational systems and the improvement of teaching and learning methods depend significantly on academic performance (Alhazmi & Sheneamer, 2023). Given that student outcomes are directly influenced by the quality of education, it is imperative for educational institutions to implement strategies that can efficiently enhance and evaluate academic performance. These strategies may encompass individualized counselling sessions, focused scientific activities, and participation in developmental programs (Radwan et al., 2020). Precise predicting of student performance is crucial for the proactive management of these strategies.

Academic achievement can be predicted by educational institutions using data mining (DM) techniques. These methods are designed to reveal hidden patterns within massive datasets (Dunham, 2006; Han et al., 2011). In the field of education, Data Mining (DM) has been widely applied to predict student outcomes. There are several techniques to achieve this, including classification, clustering, regression, correlation analysis and association rule mining. These techniques facilitate predictive modeling, which utilises machine learning (ML) algorithms to recognize patterns and identify potential performance issues at an early stage. While numerous studies focussing on developing predictive models for academic performance (Abdo et al., 2021; Albahli, 2024; Hussain & Khan, 2023; Kukkar et al., 2024; Nohuddin et al., 2022; Nohuddin et al., 2021; Rahayu et al., 2018; Shou et al., 2024), fewer researchers (Alboaneen et al., 2022; Moharekar & Pol, 2021; Sokkhey & Okazaki, 2020) who have delved into the practical implementation or system development of these models.

Despite the progress made in the education domain, the availability of comprehensive systems available for accurately predicting student performance is still inadequate. Thus, the objective of this study is to overcome this limitation by developing MyCGPA, an online system that utilises a Multiple Linear Regression (MLR) model to predict students' CGPA scores. The development of this model will be based on data acquired from a Google Form survey, as well as academic records obtained from UPNM. MyCGPA offers a wide range of features designed specifically for faculty and academic advisors. These features will make it easier to analyse academic data efficiently and provide convenient access to CGPA predictions. In summary, the development of MyCGPA represents a step forward in incorporating predictive analytics into educational practices.

The paper is organised as follows: Section 2 discusses on the research methodology and system structure. Section 3 elaborates the research findings. Followed by Section 4 presents the discussion of research findings and contributions. Finally, Section 5 concludes with a summary and future research directions.

2. METHODOLOGY

The prototype model is a software development model in where prototypes are built, tested and refined until a satisfactory prototype is achieved. It also creates the basis for producing the final system or software. It works best in scenarios where project requirements are not known in detail. The Prototype models includes 4 main phases: planning, analysis, design, and implementation (see Figure 1).

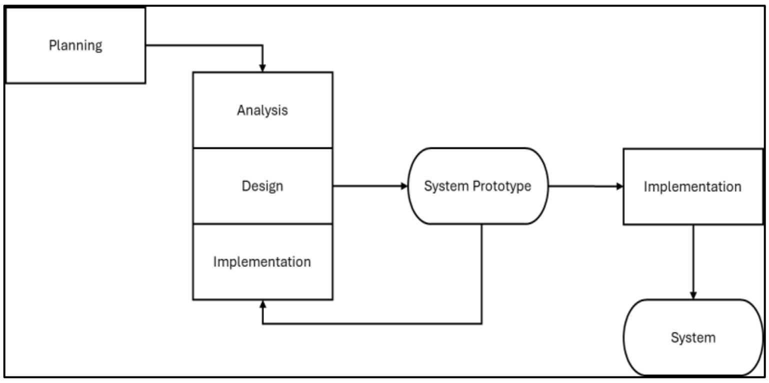


Figure 1. Prototype model adapted from (Dennis et al., 2021).

Phase 1 involved understanding the objectives of study and planning the research requirements. The aim of this study was to create a web application system using the MLR prediction model to assist academic advisors in predicting students' academic results. By identifying students who may be at risk of poor academic performance early, the system facilitates timely interventions, helping to prevent student dropouts.

Phase 2 focused on the pre-processing the collected dataset. This study used 2 main datasets: a survey question dataset and a student academic dataset. The survey questions were developed based on the previous studies by Alhazmi & Sheneamer (2023) and Fadilah et al. (2021). These studies had identified various factors (attributes) that can predict student academic performance such as GPA, CGPA, demographic data like family status, parents' occupations and many more. Data were collected from 97 respondents, who were semester 5 students majoring in Computer Science. The list of attributes can be referred in our previous study (Zainol et al., 2024).

The raw dataset collected in Phase 2 was subjected to data preprocessing. Preprocessing is a crucial step as it ensures high-quality data and error-free analysis. This process also ensures that the dataset selected is clean and ready for use in the modelling phase. Of the original 21 attributes, 20 were selected after removing the 'Matric' attribute. Next, the label encoder technique was applied to convert categorical attributes into numerical values. Therefore, before proceeding the model development, it is essential to numerically encoded the 18 attributes, excluding 'GPA' and 'CGPA'. Label encoder is a preprocessing technique for handling categorical data, was employed to convert all categorical data into numerical form using Python command codes.

The subsequent step involved selecting a set of independent attributes that are most related to the dependent variable. Predictive model might take longer to execute when they are applied to large datasets that contain a large number of attributes. To address this issue, the Ordinary Least Squares (OLS) regression technique was used to identify the most predictive attributes for student performance. OLS is a statistical method used to estimate a linear regression equation, describing the relationship between one or more independent variables. The OLS test results indicated that the variables Activity, Absence, and GPA had an R-squared value of 0.999, indicating a strong relationship. In this study, 'Activity', 'Absence', and 'GPA' were selected as the independent variables, while 'CGPA' was chosen as the dependent variable. These variables are now ready for use in the model development phase.

Phase 3 is a system design phase which involves structuring the system, designing input and output, interface display and system procedures. This phase is carried out by observing the requirements that have been obtained from the previous phase. This phase is divided into the design part which is physical and logical. The physical design consists of the database design and the software specifications required to develop a system. Logical design consists of the design of modules and interfaces. In addition, logical design also describes what the system does such as input, output, process and database design. Several types of diagrams are used such as Use-Case Diagrams, Sequence Diagrams and flow chart diagrams as a medium to explain the logical design of the system.

All these phases aim to facilitate the understanding of the design of the system to be built. This study uses the MLR algorithm to build a prediction model. As mentioned in Phase 2, the selected data set will be divided into 2 main parts: training set and test set according to the ratio of 60:40. At this stage, the MLR algorithm will train the training data set to produce a predictive model. This trained MLR model will be the classifier, and the test data will be used to evaluate the model.

Phase 4 is an important phase to evaluate the model that has been built. During this phase, we revisit the analysis phase to identify any necessary improvements or additional details. The purpose of this phase is to ensure that both the data and the built model are well prepared to produce good results. With the right selection of variables, an efficient model for predicting students' academic performance can be built. In order to ensure that the predictive model works efficiently, a predicting accuracy test was carried out. To ensure the predicting model's efficiency, an accuracy test was conducted, showing a prediction accuracy of 85.12% for the MLR model. The results are compiled into report and presented to end users. The MLR model developed in Phase 3 will be implemented in the application MyCGPA application. This system this designed to predict a student's CGPA score based on three inputs such as 'Activity' , ' Absence' and 'GPA'.

3. SYSTEM ANALYSIS

According to Tilley and Rosenblatt (2016), system analysis phase consists of 4 main activities such as requirements modelling, data and process modelling, object modelling, and consideration of development strategies (see Figure 2). The details of each activity will be explained in the following sub sections.

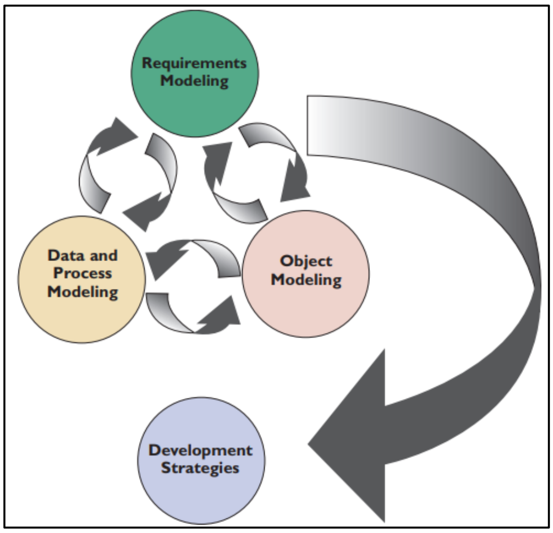


Figure 2. System analysis phase tasks adopted from (Tilley & Rosenblatt, 2016).

4.1 Requirement Modelling

Requirements modelling involves utilising various fact-finding techniques to identify the necessary research requirements. This approach seeks to achieve a thorough comprehension of user needs. The interview is a widely used technique. An interview session was conducted between the researchers and the lecturers. A set of enquiries were posed to collect valuable insights for this investigation. Throughout the interview, several important points were highlighted. Therefore, it is essential to grasp the importance of this system for users. All the interviewees agreed on the analysis of student’s academic performance prediction. By applying a regression technique can be highly advantageous for lecturers, faculties, and other stakeholders as it empowers them to anticipate student performance in advance, facilitating early detection of any potential concerns. Furthermore, the interviewer has proposed several enhancements for the system, such as the option to download data, present results through different graphs, and design a user-friendly interface.

4.2 Data and Process Modelling

Figure 3 illustrates the MyCGPA system use case which portrays user interacts with 6 use cases in the system. Firstly, the student authenticates using their username and password, navigates to their dashboard to access up-to-date academic and personal information, and enters data such as involvement in organizations, attendance records, absences, and their current CGPA. The system utilises this data to predict the student's GPA for the forthcoming semester and exhibits the prediction. Ultimately, the student concludes their session by logging out to assure the security of their session. This use case outlines the procedure for handling academic data, making predictions, and ensuring the security of the user's session.

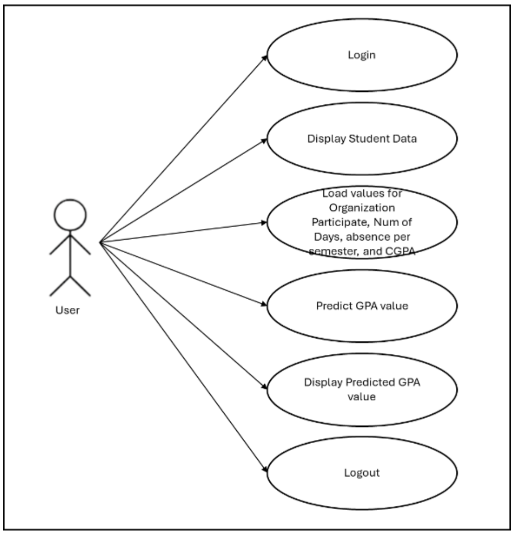

Login
Display Student Data
Load values for Organization Participate, Num of Days, absence per semester, and CGPA
Predict GPA value
Display Predicted GPA value
Logout
User

Figure 3. Use case diagram for predicting student academic performance.

4.3 Object Modelling

Object modelling is a technique used to seamlessly incorporate data and processes into entities known as objects, which serve as representations of individuals, objects, and occurrences. The aim of this modelling is to explain the interaction between the object and the developed system.

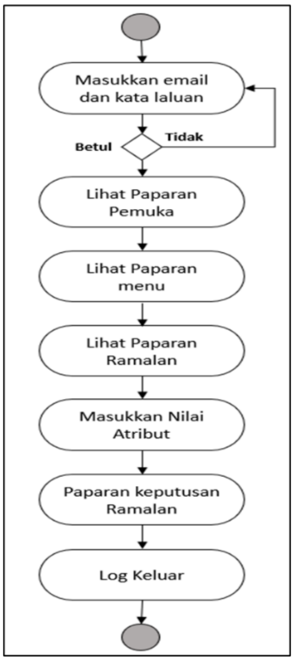


Figure 4. Activity diagram for predicting student academic performance.

Figure 4 shows an activity diagrams which is commonly used to illustrate the sequence of operations within a system. Initially, users are required to enter a username and password to continue logging into the system. If the login information is valid, the user will be redirected to the system dashboard. Otherwise, they will be redirected to the login page to retry the login. Once inside the front panel, users can choose between accessing the user account page to modify their password or delete their account or proceed directly to the menu display to initiate the forecasting process. Once the prediction attribute values are entered, the results will be generated and displayed. After completing the task, the user can safely log out of the system.

4.4 Development strategies

Through the analysis of requirements modelling, data and process modelling, and object modelling, a comprehensive understanding of the functioning of the web system has been achieved. During the design transition stage, the concepts observed will be translated into the system and database design. The techniques employed to represent the design include flow chart, and entity relationship diagram (ERD).

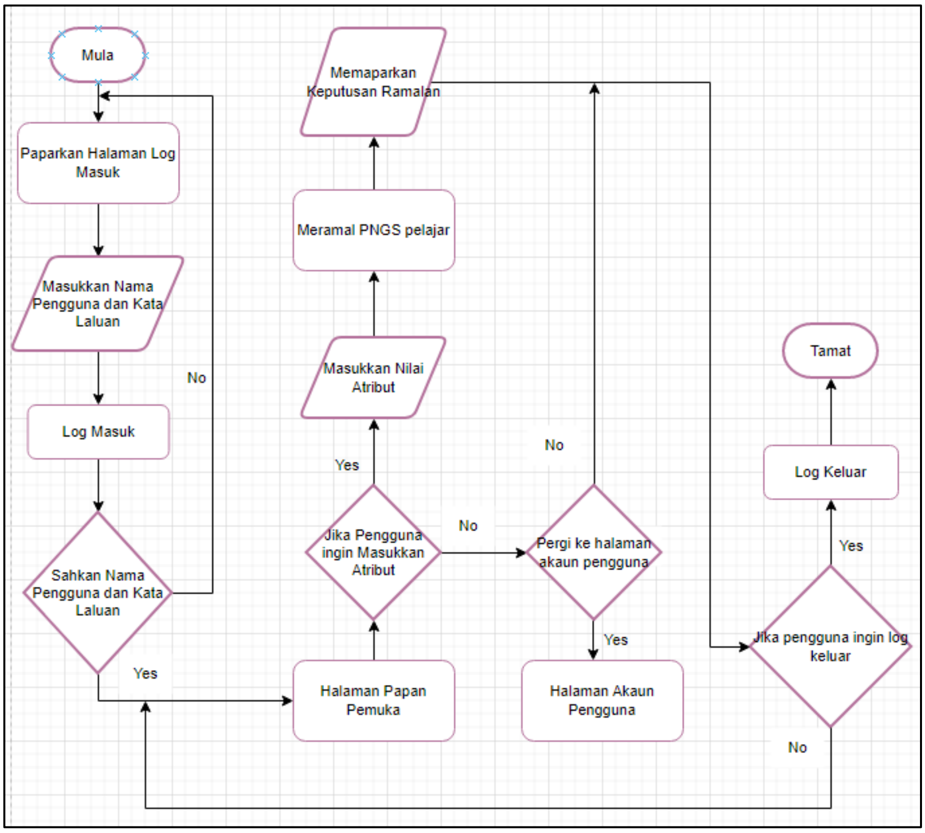

Mula
Paparkan Halaman Log Masuk
Masukkan Nama Pengguna dan Kata Laluan
Log Masuk
Sahkan Nama Pengguna dan Kata Laluan
No
Yes
Halaman Papan Pemuka
Jika Pengguna ingin Masukkan Atribut
Yes
No
Masukkan Nilai Atribut
Meramal PNGS pelajar
Memaparkan Keputusan Ramalan
Pergi ke halaman akaun pengguna
No
Yes
Halaman Akaun Pengguna
Jika pengguna ingin log keluar
Yes
No
Log Keluar
Tamat

Figure 5. Flow chart for predicting student academic performance.

Figure 5 depicts the process of deploying the system using a flow chart. It begins with the user logging into the system using the registered username and password. After the username and password have been validated, the user will be directed to the dashboard page; otherwise, the user will be required to log in again. Next, the user can input the value of the prediction attribute, and the prediction result will be displayed. ERD is used to describe the database design as well as the relationships between the same database tables. As shown in Figure 6, a user has more than one factor.

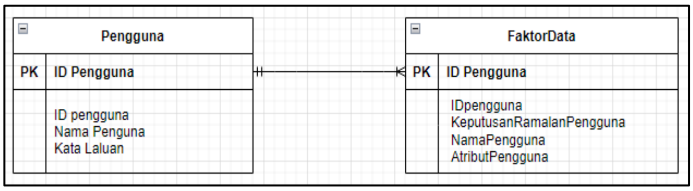
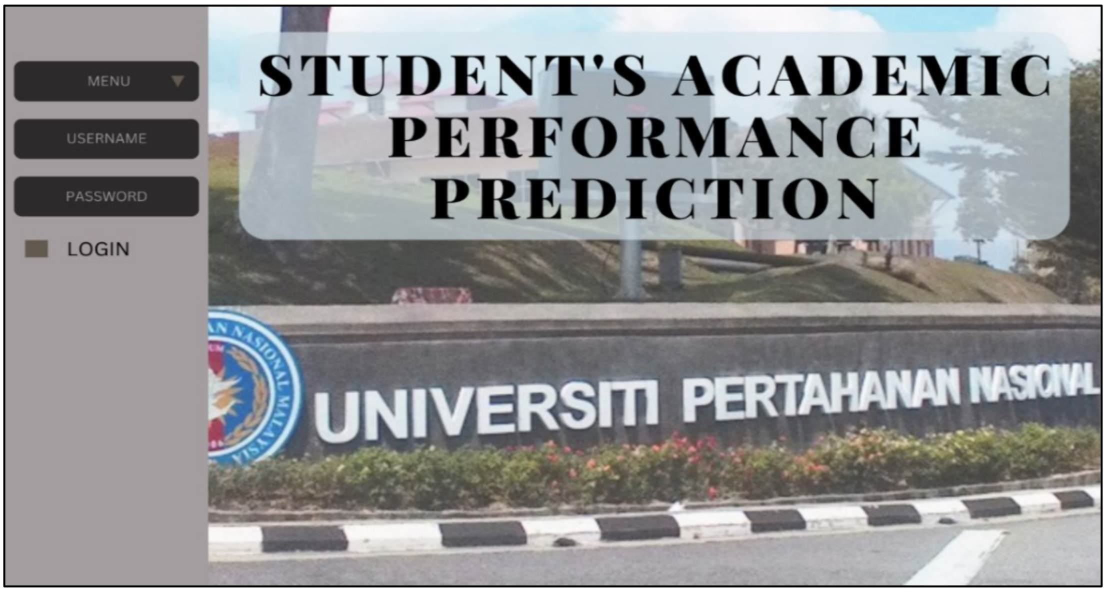


Figure 6. ERD for predicting student academic performance.

4. FINDINGS

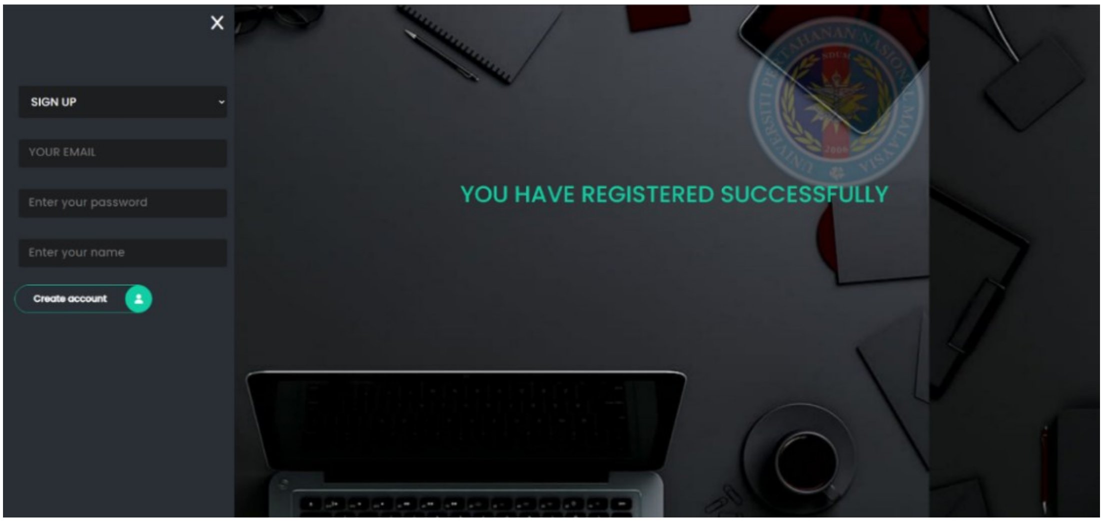
This system is developed using the Flask micro web framework for the front-end and Python programming language for the back end. MyCGPA contains 2 levels of access: administrators and users. Administrators are responsible for managing access and controlling the system during the registration process of membership.



STUDENT'S ACADEMIC PERFORMANCE PREDICTION

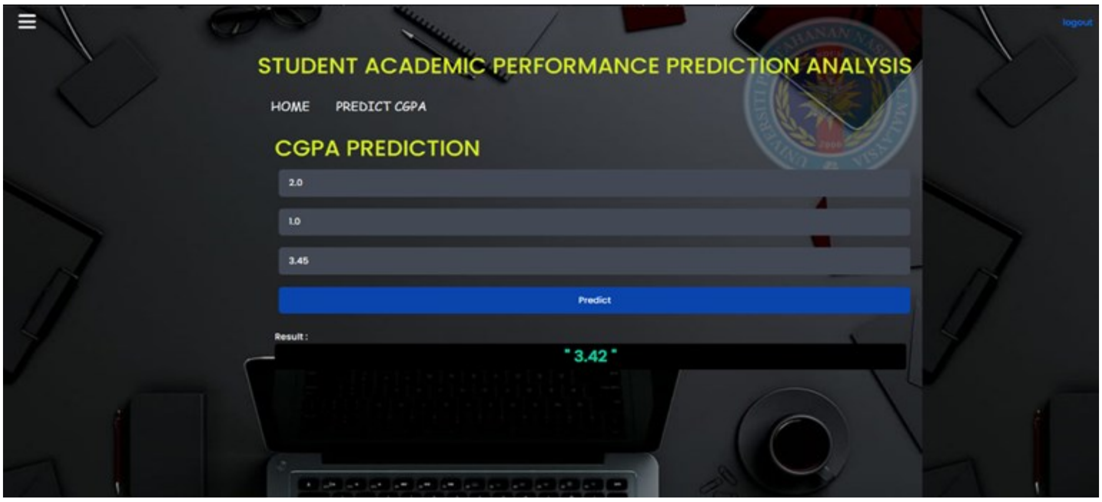
Figure 7. MyCGPA Login Page.

Figures 7 shows the interface of the MyCGPA system login page. First time user need to register and receive approval from the system administration before entering the system. All registered users can log into the system by entering their correct email and password. If the information entered is incorrect, users are unable to log in and an error message error will be displayed. All saved passwords will be hashed and securely stored in the database. The user will be directed to MyCGPA system main page once the information input is correct (See Figure 8).



YOU HAVE REGISTERED SUCCESSFULLY

Figure 8. MyCGPA Main Page.



STUDENT ACADEMIC PERFORMANCE PREDICTION ANALYSIS

CGPA PREDICTION

2.0

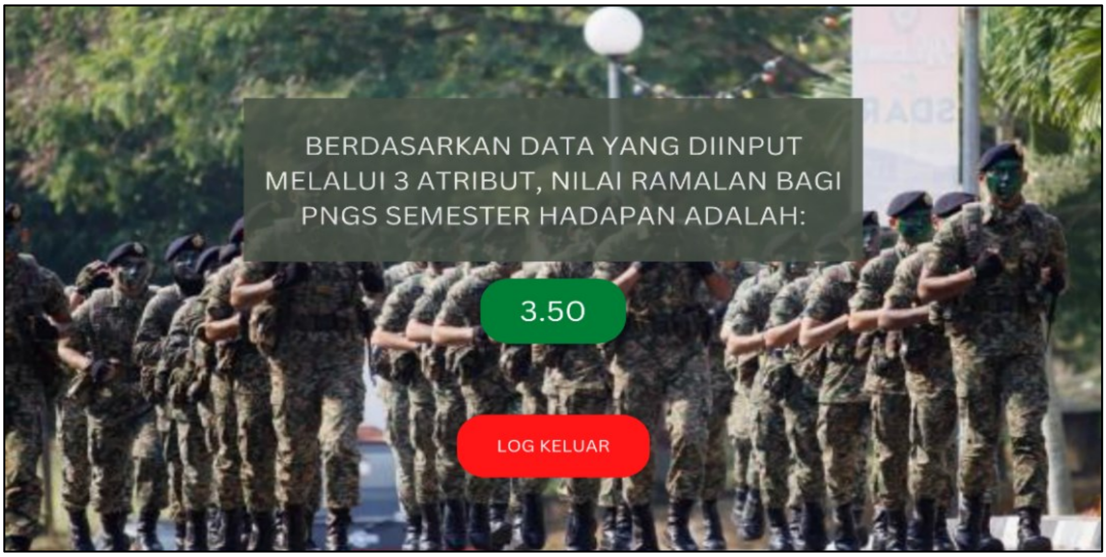
1.0

3.45

Result :

" 3.42 "

Figure 9. Prediction dashboard.



BERDASARKAN DATA YANG DIINPUT MELALUI 3 ATRIBUT, NILAI RAMALAN BAGI PNGS SEMESTER HADAPAN ADALAH:

3.50

Figure 10. Prediction result.

Figure 9 shows the prediction dashboard page which includes two navigation menu buttons e.g. 'home' and 'predict cgpa'. On this page, user must enter three inputs such as 'activity' , 'absence' and 'gpa'. The system will automatically calculate the result once the user clicks the 'predict' button. Figures 10 illustrates the prediction result. Users can easily view their academic prediction results on this page.

5. IMPLEMENTATION AND TESTING

User Acceptance Testing (UAT) involves the evaluation and validation of the developed system by end users. It is crucial in ensuring that the system meets the expectations of the end user (Fang & Darman, 2022). It allows for a thorough evaluation of whether the system has successfully passed all necessary requirements.

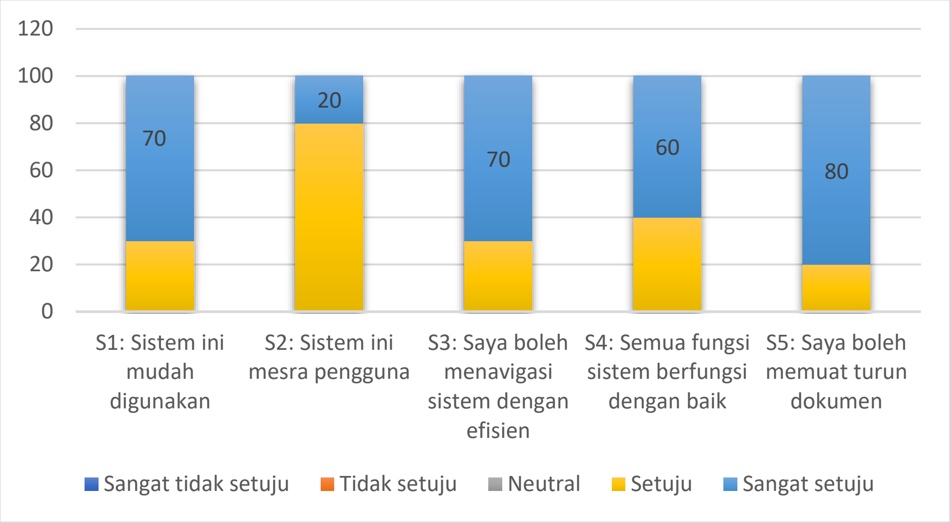


Figure 10. UAT result.

Figure 10 displays the results of the system's user acceptance test. The questionnaire was completed by 10 respondents, including lecturers and students from the Department of Computer Science. The majority of respondents, 70% found the system easy to use and reported overall satisfaction. 30% of the respondents were satisfied with their experience.

The system's user-friendliness received positive feedback from the majority (80%) of the 10 respondents, with an additional 20% strongly agreeing. The response to statement "S3: Saya boleh menavigasi sistem dengan efisien" shows that the majority of respondents believe that users can traverse the system swiftly and effectively, with 70% strongly agreeing and 30% agreeing.

The majority of users have provided good comments on the system's features, with 60% strongly agreeing. Overall, the majority of respondents expressed satisfaction with the MyCGPA system's usefulness. According to the findings, the majority of respondents were very satisfied with the system's capacity to download documents, with 80% strongly agreeing and 20% agreeing. These findings clearly show that most users value the system's capacity to download documents

6. DISCUSSION

The development of this system has successfully met the objectives such as (a) Collecting datasets with attributes required for predicting student academic performance; (b) Identifying the important attributes (variables) for developing the prediction model; and (c) Developing a web system that can predict student academic performance based on the input attributes (variables).

The advantages of the MyCGPA system are: (a) Simplifying the process of analyzing data among students; (b) Efficiently produce CGPA prediction results for the upcoming semester; (c) Having a secure and systematic database to store data and information, and (d) Providing a user-friendly graphic display of CGPA prediction results.

There are several weaknesses identified in the MyCGPA system including (a) The system is unable to produce accurate results due to lack of data; (b) Data errors may occur from manual entry by users due to user negligence, and (c) The system requires internet access to be used, which may limit accessibility.

As a result of the research done, several suggestions that can be made to improve MyCGPA system in the future. Among the improvements that can be made are: (a) Enhance the accuracy of predicting results by using a larger data set to train into a predictive model; (b) Improve the accuracy of predicting results by using more appropriate prediction algorithms, and (c) Develop an online predicting system so that access to the system is easier to use.

7. CONCLUSION

This study proposes a system that can predict the performance of students' academic. In this study, the MyCGPA successfully predicted the CGPA score using three main variables: 'Activity', 'Absence', and 'GPA'. The use of the MLR algorithm has the potential to assist academic advisors and universities in predicting CGPA scores, particularly for low-performing students. In conclusion, the MyCGPA is user-friendly and can be effectively utilized by various parties within the university, including students. This system also provides multiple functionalities to users, especially faculty members or academic assessors, enabling them to analyse student data more efficiently and produce the CGPA prediction results in an easily understandable format.

Acknowledgement

The researchers would like to thank Universiti Pertahanan Nasional Malaysia for funding this study (grant number SF0139-UPNM/2023/SF/ICT/4).

References

Abdo, A. M., Rasid, N. M. A., Badli, N. A. H. M., Sulaiman, S. N. A., Wani, S., & Zainol, Z. (2021). Student's Performance Based on E-Learning Platform Behaviour using Clustering Techniques. *International Journal on Perceptive and Cognitive Computing, 7*(1), 72-78.

Albahli, S. (2024). Efficient hyperparameter tuning for predicting student performance with Bayesian optimization. *Multimedia tools and applications, 83*(17), 52711-52735.

Alboaneen, D., Almelihi, M., Alsubaie, R., Alghamdi, R., Alshehri, L., & Alharthi, R. (2022). Development of a web-based prediction system for students' academic performance. *Data, 7*(2), 21.

Alhazmi, E., & Sheneamer, A. (2023). Early predicting of students performance in higher education. *IEEE Access, 11*, 27579-27589.

Dennis, A., Wixom, B. H., & Roth, R. M. (2021). *Systems Analysis and Design* (8th ed.): John Wiley & Sons.

Dunham, M. H. (2006). *Data mining: Introductory and advanced topics*: Pearson Education India.

Fadilah, K. I. M., Zainol, Z., Ebrahim, M., & Lee, A. S. H. (2021). *Covid-19 Effect On Undergraduate Computing Students' Performance At Higher Education: Pilot Study.* Paper presented at the 2021 6th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE).

Fang, S. Y., & Darman, R. (2022). The Development of e-PSM System for Undergraduate Students of Faculty Food Science and Nutrition (FSMP), University Malaysia Sabah (UMS). *Applied Information Technology And Computer Science, 3*(2), 388-410.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.

Hussain, S., & Khan, M. Q. (2023). Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning. *Annals of data science, 10*(3), 637-655.

Kukkar, A., Mohana, R., Sharma, A., & Nayyar, A. (2024). A novel methodology using RNN+ LSTM+ ML for predicting student's academic performance. *Education and information technologies*, 1-37.

Moharekar, T. T., & Pol, U. R. (2021). Academic Performance Prediction Application (APPA). *YMER International Open Access Journal, 20*(12), 179-196.

Nohuddin, P. N., Zainol, Z., Omar, M. A., Al Hijazi, H., & Noordin, N. A. (2022). Understanding Malaysian B40 Schoolchildren's Lifestyle and Educational Patterns Using Data Analytics. In *Sustainable Development*

Through Data Analytics and Innovation: Techniques, Processes, Models, Tools, and Practices (pp. 171-189): Cham: Springer International Publishing.

Nohuddin, P. N. E., Zainol, Z., & Hijazi, M. H. A. (2021). Study of B40 Schoolchildren Lifestyles and Academic Performance using Association Rule Mining. *Annals of Emerging Technologies in Computing (AETiC), 5*(5), 60-68.

Radwan, R. M., Mustapha, A., & Abdullah, B. (2020). Ramalan Prestasi Akhir Pelajar Melalui Kaedah Perlombongan Data [Student Final Performance Forecasting Through Data Mining Methods]. *J Asian Journal of Civilizational Studies, 2*(4), 1-16.

Rahayu, S. B., Kamarudin, N. D., & Zainol, Z. (2018). Case Study of UPNM Students Performance Classification Algorithms. *Journal of Engineering & Technology, 7*(4.31), 285-289.

Shou, Z., Xie, M., Mo, J., & Zhang, H. (2024). Predicting Student Performance in Online Learning: A Multidimensional Time-Series Data Analysis Approach. *Applied sciences, 14*(6), 2522.

Sokkhey, P., & Okazaki, T. (2020). Hybrid machine learning algorithms for predicting academic performance. *International Journal of Advanced Computer Science and Applications, 11*(1), 32-41.

Tilley, S., & Rosenblatt, H. J. (2016). *System Analysis and Design (Shelly Cashman Series)* (11th ed.): Cengage Learning.

Zainol, Z., Nohuddin, P. N. E., Husin, H. S., Rauf, U. F. A., & Mutalib, M. Y. A. (2024). A Regression Analysis for Predicting Student Academic Performance. In *Tech Horizons: Unveiling Future Technologies* (pp. 59-66): Springer.