International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 6 ǁ 2022 ǁ PP. 1804-1809

Students' Performance Analysis Using Machine Learning Algorithms

Rosemary Vargheese[1], Adlene Peraira[2], Aswathy Ashok[3] and Bassant Johnson[4]
Dept of CSE, Adi Shankara Institute of Engineering and Technology, Kalady, Kerala
rosemaryv.cs@adishankara.ac.in[1]adleneperaira@gmail.com[2]
aswathyashok135@gmail.com[3]bassantjohnson761@gmail.com[4]

Abstract— *an outsized amount of digital data from social media, research, agriculture, medical records and other University and technical organizations face high competition and their challenge is in analyzing their students' performance. The foremost important challenges are in admission, student placement and within the curriculum. The two most important processes during which data are collected and analyzed are admission and placement. The university's rank in the market solely depends on academic performance and placement of the student. Aside from academic performance there are various other factors which help in understanding the final performance of the student. During this project, processing techniques are utilized to understand the performance of students and group the students under various categories as a student must consistently improve to compete in today's world. Almost every university has their own management system to manage the students' records. Currently, there is a student management system that manages the students' records in University of Malaysia Sarawak (UNIMAS), but no permission is provided for lecturers to access the system. This is often actually because the access permission is solely to top management like Deans and Deputy Deans of Undergraduate and Student Development due to its privacy setting. Thus, this project proposes a system named Student Performance Analysis System (SPAS) to remain track of students' results. The proposed system offers a predictive system that's able to predict the students' performance which in turn assists the lecturers to identify students that are predicted to possess bad performance in their courses. The proposed system offers student performance prediction through the principles generated via processing technique. The knowledge mining technique utilized during this project is classification.*

Keywords— Data Analysis, Classification Techniques, Machine learning

---

Date of Submission: 14-06-2022 Date of acceptance: 29-06-2022

I. INTRODUCTION

Machine learning could be a specialization under the vast AI. Machine learning works towards comprehending the complexity of assorted sorts of collected data and identifying the right model for the data by trying several models. This can be effectively systemized with easier interpretation and use by people. Machine learning lies within the engineering science field but is different from basic computing algorithms which are used for problem solving. Within the process of machine learning, the algorithms are designed in a way which allows the system or computer to process the input information, create training sets and produce the desired range specified output using statistical estimation. Students are the greatest asset for various universities. Universities and students play a very important role in producing graduates of high qualities with their academic performance achievement. Academic performance achievement is the level of accomplishment of the students' educational goal that may be measured and tested through examination, assessments and other types of measurements. However, the educational performance achievement varies for different reasons students may have at different levels of performance achievement. In all the elemental parts of a student's personal and professional development is performance evaluation. Performance evaluations emphasize students' strong suits and their forte. This acts as a vital tool in augmenting their strengths and distinguishing areas that require improvement as goals. By having the ability to research the performance of their students, teachers can divert their attention to the mandatory areas, advice and guide the scholars along the proper path and acknowledge and reward their achievements.

II. RELATED WORKS

There are many studies within the literature associated with the students' performance using machine learning. Most research in this area seeks to spot the foremost appropriate algorithms by which predictions will be made and to spot the features which will be used for forecasting. Scientific articles are known, the aim of which is to form an in-depth review of the literature associated with the prediction of scholars' results. In certain

researches, different techniques and differing types of knowledge used are analyzed. "Based on the information collected during this review, the foremost widely used technique for predicting students' behaviour was supervised learning, because it provides accurate and reliable results. Specifically, the SVM algorithm was the foremost utilized by the authors and provided the foremost accurate predictions. Additional to SVM, DT, NB and RF have also been well-studied algorithmic proposals that generated good results. As for the neural networks, they're a less used technique, but they obtain great precision in predicting the students' performance. The characteristics used are Demographic characteristics of scholars and their grades, grades obtained in other courses, highschool exams, behavioural data, students' interaction with certain Moodle modules, students' characteristics and their academic performance. According to another research, there has been a transparent increase in the number of published studies within the field in recent years. "The methodologies that are used may be split into Classification (supervised learning, e.g., Naive Bayes, Decision Trees), Clustering (unsupervised learning, e.g., partitioning data), Statistical (e.g., correlation, regression), data processing (identifying features and trends) and other methods. The features that would predict student performance is broadly split into five categories: demographic (e.g., age, gender), personality (e.g., self-efficacy, self-regulation), academic (e.g., high-school performance, course performance), behavioural (e.g., log data) and institutional (e.g., high-school quality, teaching approach). They used academic data for prediction (e.g., predicting course performance supported high-school performance). The utilization of information describing student behaviour in an exceedingly course (log data), while becoming more popular within the computing education research domain, continues to be relatively rare." In Bulgarian universities, the subject of predicting student performance is poorly developed. There are few studies of Bulgarian authors associated with this issue. A study was presented, which aimed to verify whether there are models available for the students' performance at the university with the data on their personal and pre-university characteristics. The characteristics in the available data that the majority strongly influence the success of the scholars also are sought. Machine learning was used to achieve the goal. After comparing the results of the classifiers Naïve Bayes, Decision Tree, SVM, k-NN, it absolutely was found that SVM performs best, then k-NN. The good performance of the models depends on the sort of information used for forecasting. This explains the very fact that different classifiers offer the most appropriate results in different scenarios. It was discovered that the potential for predicting the performance of scholars in small student groups with very limited attributes like attendance at lectures, access to a virtual learning environment and intermediate grades. From the compared machine learning techniques, K-Nearest Neighbors (KNN) and Random Forest (RF) give the foremost accurate forecasting results. A different approach was implemented to predict student performance in England, where the models where developed for predicting student grades using internal institutional databases and external open data sources (the results of the National Student Survey). It was proved that models supporting internal and external data sources provided better efficiency and are more accurate than models based only on internal institutional data sources.

III. EXISTING METHOD

A background study is completed on reviewing similar existing systems accustomed to perform student performance analysis. Three existing systems for such a system include as follows.

Faculty network (FSS) Shana and Venkatacalam has proposed a framework named Faculty web (FSS) which is low in cost (because it uses cost effective open source analysis software), WEKA to analyse the students' performance in a very course offered by Coimbatore Institute of Technology of Anna University. FSS is in a position to analyse the students' data dynamically because it is in a position to update students' data dynamically with the flow of time to make or add a brand new rule. The update of a latest rule is feasible with the assistance from domain experts and also the rule is set by data processing techniques like classification. Classification technique is employed to predict the students' performance. Besides, FSS specialise in the identification of things that contribute to performance of scholars during a particular course.

Next is the Student Performance Analyser (SPA). SPA is an existing secure online web-based software that allows educators to look at the students' performance and keep track of the school's data. The SPA could be a tool designed for analysing, displaying, storing, and getting feedback of student assessment data. It is a robust analyser tool utilized by schools worldwide to perform analysis and displays the analysis data once raw student data is uploaded to the system. The analysis is completed by tracking the scholar or class to urge the general performance of a student or class. It helps to spot the students' performance which is below the expected level, at expected level or above the expected level. This might allow the educators or staff to spot these students' performance easily.

The third existing system is the Intelligent Mining and Decision Support System(InMinds). InMinds helps University of Malaysia, Sarawak (UNIMAS) to observe the performance of varied areas in every UNIMAS's departments. The system enables top and mid-management in UNIMAS to own a transparent look on the areas that needed attention by viewing the figures, revenues and risks. The features, simple use and adaptability provided by the system makes the performance analysis in UNIMAS to be performed in a perfect

solution. Charts are provided by the system for simple student performance's interpretation. From the reviews on these existing systems, useful techniques and features might be applied into the proposed system for a more robust system's performance. The WEKA is chosen as a tool for data processing because it's open source software.

IV. PROPOSED METHOD

In the proposed system, we are able to produce accurate predictions on student's performance in future which successively does the following :

- Able to help lecturers to automatically predict student's performance in specific courses.
- Able to keep track and retrieve student's performance during a particular course.
- Able to seek attention to poorly performing students on the basis of the students' prediction result.

There are some features from the present systems that are employed during the planning and implementation phase of the proposed system. These features and functionalities include the program, students' performance prediction, displaying them etc. an honest program provides an user-friendly interface because it is simple to navigate and not complicated. Meanwhile, the students' performance prediction is included into the proposed system to ensure the objectives are achieved. Furthermore, the generation of reports in Portable Document Format (PDF) and illustration display like charts in PDF makes student performance analysis easier.

V. METHODOLOGY

There are several phases throughout the project development, which is as follows:

1. Problem and data understanding

The problem and data understanding is critical in determining the success of the scholar Performance Analysis system. Before developing the system, problems and data understanding is important to build the system and achieve the project goal and objectives. The problems of the system are identified and analyzed for its effectiveness and efficiency in terms of functionality. After the problems are identified, the solutions to resolve each problem are identified and picked up through more reading and studying on the related research papers.

Moreover, an interview shall be conducted with the stakeholders and experts in the Machine Learning field to get a clearer look on the system, its working, constraints etc. Besides, other similar systems are studied and analysed for clearer understanding of its features, strengths and weaknesses. This helps to identify the requirements and opportunities for the proposed system.

Other than that, student data is collected during this phase. The students' data like student's results from the past semesters, percentage of marks obtained in secondary and senior secondary classes etc is collected. Table 1 shows the attributes of the dataset collected for processing classification.

TABLE I. ATTRIBUTES OF DATASET

Attributes	Values
Gender	Categorical (0=Female, 1=Male)
10th Marks	Discrete
12th Marks	Discrete
S1 CGPA	Discrete
S2 CGPA	Discrete
S3 CGPA	Discrete
S4 CGPA	Discrete
S5 CGPA	Discrete
Overall CGPA	Discrete

2. System analysis and design

The overall flow of the system is planned, analysed and designed in this phase. Firstly, based on the problem understanding, we analysed the system and user requirements and listed them in table format. Data flow diagram (DFG) was used to chart the input, processes and output of the system. Data flow diagram analyses and draws an overall picture from the context diagram up to the first level.

Besides, the logical design of the proposed system is drawn to ensure the developed system is functioning as expected. The logical design is designed by drawing entity-relationship diagrams (ERD). The

ERD illustrates the data objects, attributes and relation between tables in the database as it is a graphical representation of the entity-relationship data model. The design of the proposed system also provides a detailed idea on the design of the backend consisting of databases and the frontend user interface.

The hardware requirement in this phase is a computer with sufficient tools for analysis and design. The ERD and data flow diagrams are drawn using Microsoft Office Visio 2007.

C. Implementation and testing

During the implementation phase, a dataset of 114 students' records is collected and analysed by using data mining techniques to generate rules for prediction of students' performance in the future semesters. The generation of rules is performed by using an open software platform, Pycharm and Jupyter Notebook. The dataset is divided into training-set and test-set. 80% of the dataset is used for the training set and the remaining 20% is for the test set. The training set is used to train the classification model while the test set is used to test the classification model built for its prediction's accuracy. A comparison of accuracy between different classification techniques are tested to ensure the highest prediction of accuracy could be achieved. The following table shows the accuracy comparison between five different decision trees' classification techniques.

TABLE II. COMPARISON BETWEEN CLASSIFICATION TECHNIQUES

Technique used	Correctly Classified Instances
Decision Tree	79.87%
Random Forest	80.52%
Naive Bayes	79.22%
SVM	81.82%

From the table shown above, the SVM is chosen to be implemented in the proposed system due to its highest accuracy (81.82%) among other classification techniques.A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both regression and classification purposes. It is based on the idea of finding the hyper-plane that best divides the dataset into two classes. Hyper-planes are decision boundaries that help the machine learning model classify the data or data points.

Predicting the performance of the student has been done in two phases, namely (i) training phase, and (ii) testing phase. At the initial stage of the work, certain preliminary tasks have been performed related to the work. The dataset was collected from the students and the pre-existing faculty databases. This collected information about the students was uploaded. Since, there always exists a possibility of samples that sometimes don't satisfy the maximum number of criteria of values with their respective variables or sometimes they may have irrelevant values that are not suitable to present conditions of the work. Those data samples are considered as irrelevant and should be omitted always for the prediction of student performance for a whole class. So pre-processing of data samples was done on them.

The programming skills such as HTML, CSS, PHP and database queries will be applied to build the proposed system once the system has been designed. To write the programming coding, tools for writing codes and a local web server is needed. Pycharm happens to be an apt tool for it.

In order to predict a student's result, lecturer is required to import nine important components listed below for analysis:

i. Gender
ii. 10th Marks
iii. 12th Marks
iv. S1 CGPA
v. S2 CGPA
vi. S3 CGPA
vii. S4 CGPA
viii. S5 CGPA
ix. Overall CGPA

After these data are imported to the database, an analysis is carried out inside the system to predict a student's upcoming performance. Figure 3 illustrates an interface that enables the prediction of a student.

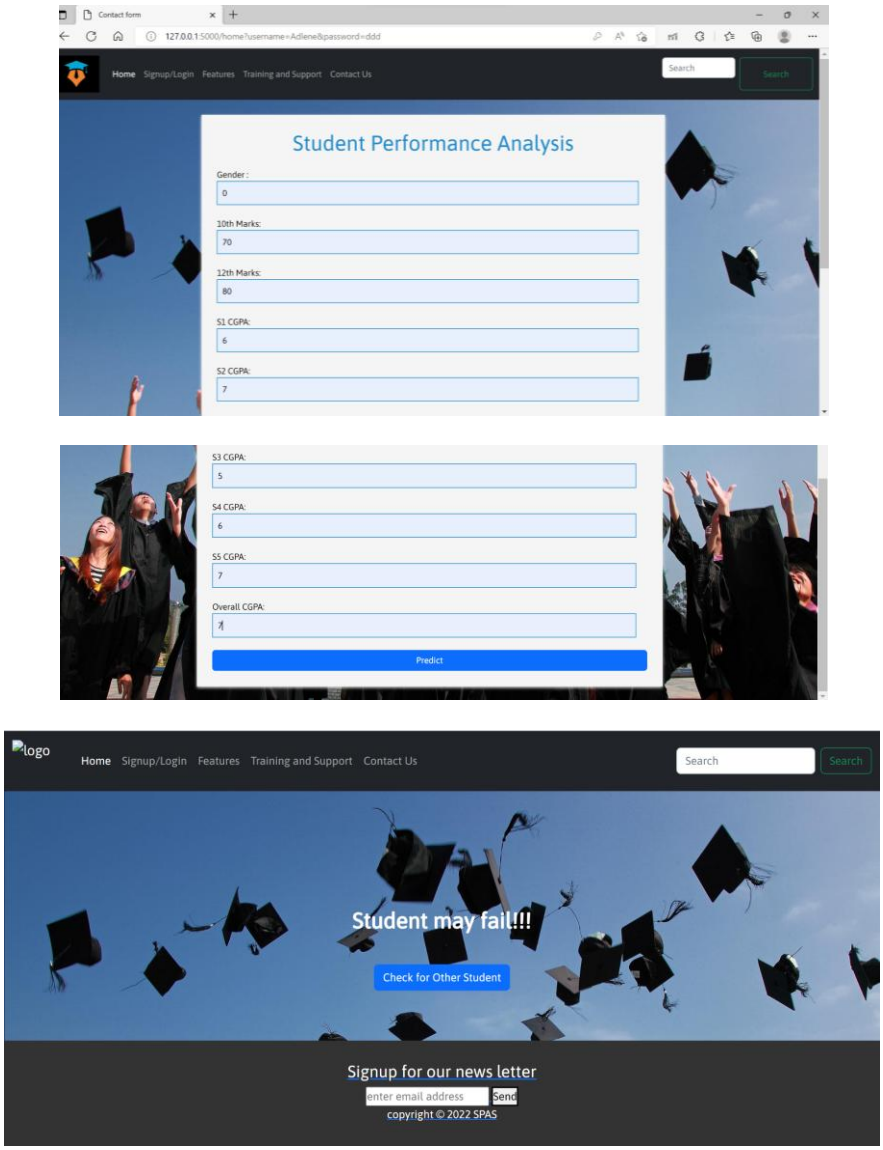

Figure 3. Students' Results Prediction Interface

After the system is built, the unit testing, system testing and user acceptance testing are needed for errors detection before the system is distributed and used by lecturers. This is to ensure the performance of the system is in its optimal state. Besides, the errors and bugs that are detected during testing of the proposed system can be fixed. The unit and system testing will be tested by the developer while the user acceptance testing will be tested by a few end-users to ensure the functionalities of the system are working as expected.

D. Evaluation of System

For the evaluation of the system, five end-users are requested to evaluate the usability of the Students' Performance Analysis System. This ensures the objectives of the proposed system are achieved as well as the ease of navigation across the interfaces of the proposed system. Moreover, the evaluation confirms high effectiveness of the proposed system is achieved. From the evaluation performed, a list of users' recommendations and suggestions are stated as shown below:

i. Apply the students' results prediction to other courses.
ii. Enable the viewing of all semesters' performance when searching for a student's performance.
iii. Provide a class wise report.

E. System Limitation

Some system limitations were identified considering the users' evaluations on the system, which are listed below:

i. Resources and time constraint

ii. Inflexible rules implemented in the system

VI . FUTURE WORK

In this project, the prediction is not updated dynamically within the system's source codes. Thus, in future, a dynamic prediction model could be implemented by training the prediction model itself whenever a new training set is fed into the system. Moreover, the prediction can be offered to the other courses in future as well.

VII . CONCLUSION

In conclusion, the project concentrates on the development of a system for student performance analysis. A data mining technique, Support Vector Machine algorithm is applied in this project to ensure the prediction of the student performance is possible. The main contribution of the SPAS is that it assists the lecturers in conducting student performance analysis. The system assists lecturers in identifying the students' that are predicted to fail in a course.

REFERENCES

[1]. J. Shana, and T. Venkatacalam, "A framework for dynamic Faculty Support System to analyse student course data", International Journal of Emerging Technology and Advanced Engineering, Vol. 2, No. 7, 2012, pp.478-482.
[2]. Quality Assurance Division (2010). InMinds: Intelligent Mining and Decision Support System. *University of Malaysia Sarawak, UNIMAS* [Online]. Available:
[3]. http://www.qad.unimas.my/Function/ICTCompliance/inminds.html
[4]. SPA (2013). What is SPA Standard? *SPA Student Performance Analyser* [Online]. Available:
[5]. https://www.studentperformanceanalyser.com.au/spa/about.html
[6]. V. Kumar, and A. Chadha, "An empirical study of the applications of data mining techniques in higher education", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 2, No..3, 2011, pp. 80-84.