International Journal of Engineering Innovations in Advanced Technology

ISSN: 2582-1431 (Online), Volume-4 Issue-4, December 2022

A Generic Model To Analysis And Predict The Students' Academic Performance

Devula Sri Sai Divya

Research scholar (CSE) in International School of Technology and Sciences for Womens (ISTS women's college) in NH-16, East Gonagudem, Rajanagaram, Andhra Pradesh, 533294

E-mail: devuladivya@gmail.com

D D D Suribabu

Associate professor and hod (CSE) in International School of Technology and Sciences for Womens (ISTS women's college) in NH-16, East Gonagudem, Rajanagaram, Andhra Pradesh, 533294

E-MAIL: suribabu.ddd@gmail.com

***Abstract*: Developing tools to support students and learning in a traditional or online setting is a significant task in today's educational environment. The initial steps towards enabling such technologies using machine learning techniques focused on predicting the student's performance in terms of the achieved grades. The disadvantage of these approaches is that they do not perform as well in predicting poor-performing students. The objective of our work is two-fold. First, in order to overcome this limitation, we explore if poorly per- forming students can be more accurately predicted by formulating the problem as binary classification. Second, in order to gain insights as to which are the factors that can lead to poor performance, we engineered a number of human-interpretable features that quantify these factors. These features were derived from the students' grades from the University of Minnesota, an undergraduate public institution. Based on these features, we perform a study to identify different student groups of interest, while at the same time, identify their importance.**

1. INTRODUCTION

Educational data mining (EDM) is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in. There are several data regarding the students which stay unused with untapped potential of data mining which could revolutionize the field of education. Since the ultimate aim of an educational institution is to create a pool of skilled professionals to take on the society to a next upgraded level, they need to create an environment for their students to grow in every vertical by giving them right exposure and training. Most of the educational institutions, maintain huge databases of students and the information keeps on increasing with time, but there is no action taken to gain knowledge from it. DM has the suitable techniques in mining the data to discover new information and knowledge about students. DM provides various methods for analysis which include classification, clustering, and association rules. Classification, one of the prediction algorithms, classifies the data (constructs a pattern) based on the training set and uses the pattern to classify a new data (testing set). In this paper, we consider the students' academic performance (SAP) system in University Sultan Zainal Abidin (UniSZA), Kuala Terengganu, Malaysia as our existing system. IHL faces a major challenge in order to
improve and manage the organization to be more efficient in managing students' activities. To achieve this target, DM is considered as the one of most suitable technique in giving additional insights to the IHL community to help them make better decisions in educational activities. The IHL make use of WEKA tool in order to build a model and predict the SAP in order for the professors to provide the students with individual attention. In SAP system, the classification method is selected to be applied on the students' data. This system makes use of
one among the three selected classification algorithms; decision tree (DT), Naïve Bayes' (NB), and rule-based (RB). The best technique is used to develop a predictive model for SAP. The patterns

ISSN: 2582-1431 (Online), Volume-4 Issue-4, December 2022

obtained is used to predict the first semester of the first year in two Bachelor of Computer Science (BCS) courses; Bachelor of Computer Science with specialization in Software Development (BCSSD) and Science with specialization in Network Security (BCSNS) at the Faculty of Informatics and Computing (FIC), university Sultan Zainal Abidin (UniSZA), Terengganu, Malaysia. This pattern will be used to improve the SAP and to
overcome the issues of low grades obtained by students. In our proposed system, student performance analysis system (SPAS) provides means for students to get an idea regarding their profile is suitable for getting placements, for which we make a clear-cut analysis of all the data parameters required for the classification and the students' prediction is made based on the collected data. For this, the placement details of the CSE passed out students of the batch 2016 and 2017 of our institution is considered. The data is collected from the concerned and it is normalized. The unnecessary attributes for prediction are removed and the required parameter data is cleaned and converted into Arff format for the analysis in WEKA tool. Various existing algorithms are applied in the WEKA tool and the most efficient algorithm is considered for improvement to be used in SPAS. The cumulative predictor algorithm builds a prediction model on the previous years' student data which can be applied on future data sets. This algorithm is found to be more efficient than the existing systems and this model built is used in SPAS. SPAS provides an interface in a web platform which enables both teachers as well as students to predict the outcome of placements based on the model's results. Data provided to SPAS can also be designed to predict even more information like possible arrear students, product based potential students and much more. Since there is support from the institution, a data analysis to determine the necessary parameters and the collection of information for the parameters can be done from the students and the institution. The existing model focus mainly on probable low graders, so that the professors can provide individual attention, whereas the proposed system allows both students and staffs who can work mutually in order to increase the placements in the college. Also, the existing system, involves generating a prediction model based on common classification algorithm whereas our proposed system makes use of voting classifier which is an ensemble classification technique with more accurate results. SPAS is an application of EDM, which makes use of existing and new collected data, for prediction of students' chances of getting placements. SPAS modules have the following phases:

- data analysis
- data collection
- applying existing algorithm
- development of new algorithm
- interface implementation.

III. LITERATURE SURVEY

This paper proposes a framework for predicting SAP of first year bachelor students in computer science course. The data were collected from eight-year period intakes from July 2006/2007 until July 2013/2014 that contains the students'
demographics, previous academic records, and family background information. DT, NB, and RB classification techniques are applied to the students' data in order to produce the best SAP prediction model. The experiment result shows the RB is a best model among the other techniques by receiving the highest accuracy value of 71.3%. The extracted knowledge from prediction model will be used to identify and profile the student to determine the students' level of success in the first semester. This project acts as the basis of SPAS and gives a clear idea regarding the parameters involved in predicting students' performance (Ahmad et al., 2015). This paper discusses on split cover type dataset into two equal parts randomly. After that, they compared some different algorithms on the split datasets, they proposed that back propagation has an accuracy of 70.0%, J48 having an accuracy of 82.3% and the best performance is C5.0, whose accuracy is 83.7%. From this work, the algorithm J48 is considered as the algorithm for improvisation in SPAS (Jain and Minz, 2008). The main objective of the paper is to provide an overview on the data mining techniques that have been used to predict student's performance. This paper also focuses on how the prediction algorithm can be used to identify the most important attributes in a student's data. Therefore, we arrive at clear cut essential parameters for prediction of performance of students for our model (Shahiria et al., 2015). We could actually improve students' achievement and success more effectively in an efficient way using EDM techniques. It could bring the benefits and impacts to students, educators and academic institutions (Venkatramaphanikumar et al., 2015). Even though many papers in this

ISSN: 2582-1431 (Online), Volume-4 Issue-4, December 2022

educational research focuses on learning methodologies, Pedagogic methods, and factors influencing student performance, only limited work is done towards the prediction of student performance is in current digitally enriched society. The authors have presented a model in which the academic performance of the newly admitted students into the engineering stream was predicted by considering socio-demographic and academic variables. This paper helps to understand the various algorithms which can be used regarding the prediction of SAP and also the parameters influencing the students' performance (Garcia and Mora, 2011). NB classifier was used to predict the academic performance which yields an accuracy of 50%. It used different classification techniques to build performance

prediction model based on students' social integration, academic integration, and various emotional skills. Two algorithms (J48 and random tree) were applied on MCA students for prediction. The algorithms were applied on only

limited data samples and achieved an accuracy of 94% with random tree algorithm and 88% with J48 algorithm (Mishra et al., 2014). The paper involves development of an approach to

predict the performance of a student, who got admission into the Nigerian university. Factors like scores such as current courses, matriculation scores, age, gender, parent history and location have affected the performance of the student in that model (Oladokun et al., 2008). The main aim of the paper is to describe the methodology for the implementation of the initiated data mining project at the University of National and World Economy (UNWE), and to present the results of a study aimed at analyzing the performance of different data mining classification algorithms on the provided dataset in order to evaluate their potential usefulness for the fulfilment of the project goal and objectives. To analyses the data, we use well known data mining algorithms, including two rule learners, a DT classifier, two popular Bayes' classifiers and a nearest neighbor classifier. The WEKA software is used for the

study implementation since it is freely available to the public and is widely used for research purposes in the data mining field (Kabakchieva, 2013). DT is the most widely applied supervised classification data mining technique. The learning and classification steps of DT induction are simple and fast and it can be applied to any domain. For this research work student qualitative data has been taken from EDM and the performance analysis of the DT algorithm C4.5 and proposed algorithm are compared. The classification accuracy of proposed algorithm is higher when compared to C4.5. However, the difference in classification accuracy between the DT algorithms is not considerably higher. In this study C4.5 classifier and proposed algorithm with ensemble techniques such as boosting and bagging have been considered for the comparison of performance of both the algorithms according to parameters accuracy, build time, error rate, memory used and search time for the classification of datasets (Patidar et al., 2015).

3 SYSTEM ARCHITECTURE

SPAS is developed as incremental modules. There is several research studies carried out by experts which form the base in determining the modules in the software development process of SPAS. The development process of SPAS involves the following modules:

• Data analysis: it involves analysis of the parameters necessary to make the prediction by doing research on existing systems and literature surveys by renowned experts in the field.

• Data collection: this phase mainly involves collection of the required data from various sources and normalizing them so that it can be used in data mining process.

• Applying existing algorithm: this phase is carried out using already implemented and available algorithms in WEKA tool.

• Development of new algorithm: after analysis of existing algorithms, their methods and evaluating accuracy, a new and more efficient algorithm is designed in this phase.

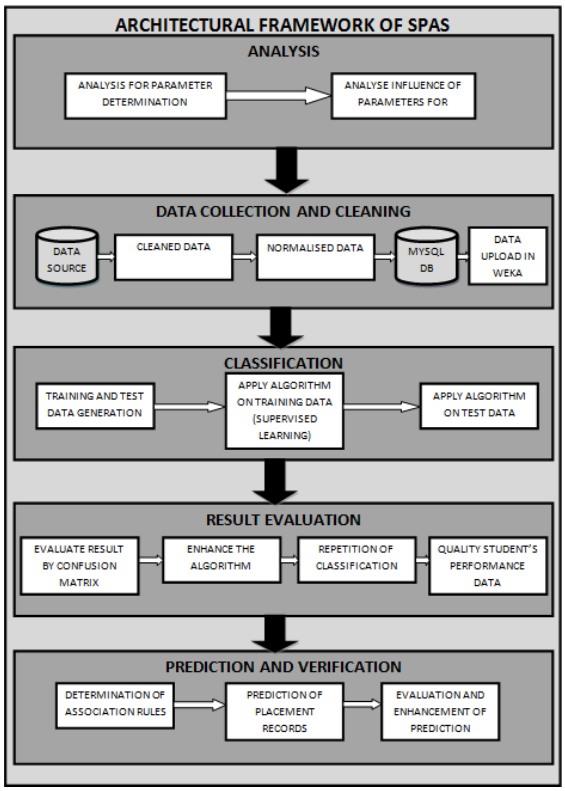
• Interface implementation: the project is made available to users (mainly faculties and students) by implementing the algorithm in a simple interface as a web platform service.

Figure 1 provides an overall idea on the processes involved in each of the modules in the development process of SPAS. Finally, once the framework for SPAS is developed, the SPAS is made available for real-time use to students and

staffs as a web application which enables them to view the prediction.

Figure 1 Architectural framework of students' performance analysis system

ISSN: 2582-1431 (Online), Volume-4 Issue-4, December 2022



4 DATA ANALYSIS

The data analysis phase of SPAS involves a complete analysis of all the data which will be necessary for the prediction of students' placements. The systematic literature review is used to identify the important attributes in predicting students' performance. The attributes that have been frequently used is cumulative grade point average (CGPA) and internal assessment. The main idea of why most of the researchers are using CGPA is because it has a tangible value for future educational and career mobility. It can also be considered as an indication of realized academic potential. Through the coefficient correlation analysis, the result shows that CGPA is the most significant input variable by 0.87 compared to other variables CGPA is the most influence attributes in determining the survival of students in their study, whether they can complete their study or not. In this study, internal assessment was classified as assignment mark, quizzes, lab work, class test and attendance. All attributes will be grouped in one attribute called internal assessment. The attributes are mostly used among the researchers to predict student's performance. Next, the most often attribute being used is students' demographic and external assessments. Students' demographic includes gender, age, family background, and disability. While external assessment is identified as a mark obtained in final exam for a particular subject, the reason of why most of the researchers used student's demographic such as gender is because they have different styles of female and male students in their learning process. It is also found that most of female students have various positive learning styles and behaviors compared to male students. Female students are more discipline and dutiful in their studies, self-directed, always preserved and focused. In other side, female students have effective learning strategies in their study. They have self-motivation, organization and rehearsal that were effectively used by them. Thus, it is proven that gender is one of important attributes influencing students' performance.

The three other attributes mostly used in predicting students' performance are extra-curricular activities, high school background and social interaction network. There are five out of thirty studies that used each one of these

attributes. There are also several researchers in another study who have used psychometric factor to predict students' performance. A psychometric factor is identified as student interest, study behavior, engage time, and family support. They have used this attribute to make a system to look very clear, simple and user friendly. It helps the lecturer to evaluate students' achievement based on their personal interest and behavior. However, these attributes are rarely to apply in predicting students' performance by several researchers because it focuses more on qualitative data and it is also hard to get a valid data from respondents.

5. BUILD MODEL USING CUMULATIVE PREDICTOR ALGORITHM

Cumulative predictor algorithm is a data mining technique which builds a prediction model on existing data and applies the model on new data. Cumulative predictor algorithm is an extension of the classification algorithms in WEKA standard library based on open-source python code, i.e., CP algorithm is built on WEKA jar file. Briefly the algorithm for building the model involves the following processes:

5.1 Data retrieval

• The normalized data is stored and made available in a MySQL database and the connection is established.

• The database with all the attributes as mentioned in the data collection phase is retrieved using MySQL

ISSN: 2582-1431 (Online), Volume-4 Issue-4, December 2022

query by using DataSource.read() command of the WEKA jar
file.
• The data records are stored as instances by creating an instances data type by the name 'train' and we set the class index for our instances by using the command
train.setClassIndex(train.numAttributes()-1);

5.2 Re-sampling

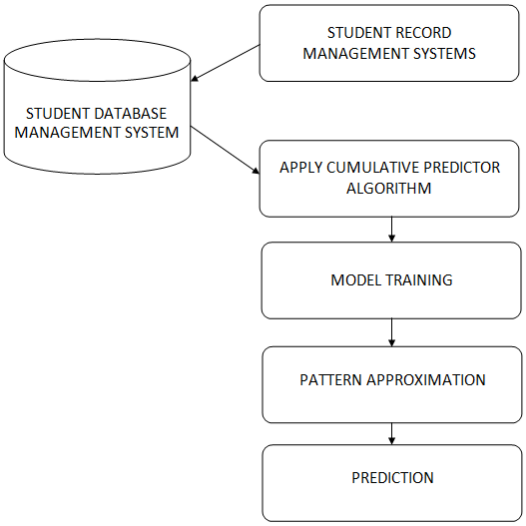
• The data is re-sampled into several subsets for the training of the classifier.
• Re-sampling enables the data to be distributed over the entire training data in order to increase the efficiency of our model built.
• In this phase the data is randomised by the following command:
Random rand new Random((int)
= System.currentTimeMillis());
data.randomize(rand);
• After the data is randomised, re-sampling of the data is carried out using resampleData() function of the Weka library, which is a supervised filter used for resampling the data.
An instance for resample is created as follows and then re-sampling is done:
weka.filters.supervised.instance.Resample sr new
weka.filters.supervised.instance.Resample();

• After re-sampling, several subsets of data are created and stored as separate arff files to be used in the model build for training our prediction model

5.3 PREDICTION MODEL GENERATION

• Bagging is an ensemble classifier technique.
Cumulative Predictor Bagging method is considered for improvement in our system.
• Bagging is based on bootstrap re-sampling. Resample data sets are created and are used for voting to generate the final classifier.
• The model generated by the final classifier is utilized for prediction of students' placement details.
The procedure is as follows:
For each cumulative predictor classifier:
{
• Use training dataset to train the classifier.
• Analyse the data for information entropy and entropy gain ratio to determine the root node.
• Split the node to form the branches based on split criteria.
• Use pair test dataset to test the classifier and get the accuracy.
}
• The new instance will be classified by every cumulative predictor classifier and the instance will be classified as the class Which gets most votes. • Thus for each and every training data subset stored as a separate Arff file, a prediction model is generated by the cumulative predictor algorithm. • The cumulative result of all the models generated above are combined to make the final prediction model as SPAS.model file.
• This model is used for the prediction of students' performance and is explained in Figure 2. On research regarding various factors and formulae, the main influencing parameters which are changed to improve accuracy for cumulative predictor algorithm is found to be:
• standard deviation
• information entropy
• entropy gain ratio.

Figure 2 Prediction model generation



A Standard deviation

Standard deviation reflects the class distribution of the dataset. If an attribute has a low standard deviation, it indicates the data values tend to be very close to its mean value. The distribution is simpler. If an attribute has a large standard deviation, it indicates the data values spread out over a large range of values and it is highly randomized.

B Information entropy

Information entropy describes the expected value of the information in an attribute. It is a measure of the randomness of the values of an attribute. And in forest cover type scenario, with a larger standard

ISSN: 2582-1431 (Online), Volume-4 Issue-4, December 2022

deviation, the information entropy will be larger as well.

C Entropy gain ratio

Entropy gain ratio optimization method in J48 is the java implementation of C4.5 in WEKA. It builds DTs from training dataset like ID3 but uses information gain ratio. J48 builds DT by choosing an attribute with the largest information entropy gain ratio as current split
node.

D Modifying the three parameters

The brief process involving modification of the above parameters is as below:

• If a numeric attribute has a large standard deviation, the information entropy will be multiplied by a larger balancing coefficient α, its split information will multiply by a smaller
coefficient β, α and β are determined by standard deviation.

• Else an attribute with a small standard deviation will multiply by a smaller α and a larger β.

6. CONCLUSION

The purpose of this paper is too accurate identify students that are at risk. These students might fail the class, drop it, or perform worse than they usually do. We extracted features from historical grading data, in order to test different simple and sophisticated classification methods based on big data approaches. The best performing methods are the Gradient Boosting and Random Forest classifiers, based on AUC and F1 score metrics. We also got interesting findings that can explain the student performance.

7.REFERENCES

[1] L. Breiman. Random forests. Machine learning, 45(1):5{32, 2001.

[2] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. Classification and regression trees. CRC press, 1984.

[3] C. Cortes and V. Vapnik. Support-vector networks. Machine learning, 20(3):273{297, 1995.

[4] J. H. Friedman. Greedy function approximation: a gradient boosting machine. Annals of statistics, pages 1189{1232, 2001.

[5] Z. Iqbal, J. Qadir, A. N. Mian, and F. Kamiran. Machine learning based student grade prediction: A case study. arXiv preprint arXiv:1708.08744, 2017.

[6] J. E. Knowles. Of needles and haystacks: Building an accurate statewide dropout early warning system in wisconsin. Journal of Educational Data Mining, 7(3):18{67, 2015.

[7] S. Kotsiantis, C. Pierrakeas, and P. Pintelas. Predicting students'performance in distance learning using machine learning techniques. Applied Articial Intelligence, 18(5):411{426, 2004.

[8] J. McFarland, B. Hussar, C. de Brey, T. Snyder, X. Wang, S. Wilkinson-Flicker, S. Gebrekristos, J. Zhang, A. Rathbun, A. Barmer, et al. Undergraduate retention and graduation rates. In The Condition of Education 2017. NCES 2017-144. ERIC, 2017.

[9] B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch. Predicting student performance: an application of data mining methods with an educational web-based system. In Frontiers in education, 2003. FIE 2003 33rd annual, volume 1, pages T2A{13. IEEE, 2003.

[10] S. Morsy and G. Karypis. Cumulative knowledge-based regression models for next-term grade prediction. In Proceedings of the 2017 SIAM International Conference on Data Mining, pages 552{560. SIAM, 2017.

AUTHORS PROFILE

Devula sri sai divya completed her B.Tech in 2018 in Computer Science and Engineering and has interest in Software Testing, Machine Learning, DBMS, Techniques For Student's performance analysis system using cumulative predictor algorithm as a part of research of her M.Tech Project.

D D D suribabu (M-.tech,Phd) he working as an associate professor and hod in ISTS women's engineering college.