UNIVERSITY ADMISSION SYSTEMS USING DATA MINING TECHNIQUES TO PREDICT STUDENT PERFORMANCE TO SUPPORT DECISION MAKING

PRANEETH REDDY PENUGONDA[1], MEENALOCHANI GANDHAM [2], VENKATA SATYA KOPPULA [3], RADHA KRISHNA BALUSU [4].

[1, 2, 3, 4.] *SCHOOL OF COMPUTER SCIENCE AND ENGINEERING VELLORE INSTITUTE OF TECHNOLOGY VELLORE (TN.), INDIA*.

***

ABSTRACT – The admissions process is a struggle for universities nowadays, particularly those that focus on STEM fields like computer science and engineering. In order to identify students who would be successful in their programmers, universities should use objective criteria for admissions. The suggested technique was tested using a dataset consisting of 2,039 students from 2016 through 2019 who were enrolled in the Information and Computer Science College of the a Saudi state institution. The findings show that early academic success at university may be predicted prior to admission using certain criteria. The findings also suggest that a student's score on the Scholastic Proficiency Admission Test is the best predictive factor for admission. This score should thus be given greater weight in selection procedures.

***Index terms* – Data Mining, SVM, ANN, Decision tree.**

I. INTRODUCTION

In order to identify students who would be successful in their programmes, universities should use objective criteria for admissions. Additionally, each school should use the most advanced methods for estimating an individual student's potential in the classroom before enrolling them. This would help policymakers at universities establish effective admissions standards. Nonetheless, most universities have difficulty analysing their massive educational datasets to foretell students' success [1]. This is because they rely only on traditional statistical approaches, as opposed to more modern and effective prediction methods like Educational Data Mining (EDM), the most widely used method for assessing and predicting students' performance [2–6]. In order to anticipate students' performance, EDM first involves gathering meaningful information and trends from a massive educational database [2]. Better data allows for more planned approaches to boosting students' academic standing.

This research aims to aid colleges in their admissions processes by providing more accurate predictions of applicants' future academic achievement using data mining methods.

In various areas, this research adds to the existing body of knowledge. At first, we use four data mining classification models to foretell candidates' early academic success based on their profiles before they enrol. Quiz and final exam scores, extracurricular involvement, student demographics, cumulative grade point average, and social network contacts are among the profile data most often utilised for predicting students' success in higher education (e.g. [7]–[10]). However, factors that may be used to predict student achievement, like as pre-admission test results, are seldom taken into account in the admissions process (e.g. [11]–[13]). This research focuses on these underappreciated indicators. We also compare four categorization methods for making predictions about students' performance and determine which one is most effective in terms of accuracy, precision, recall, and F1-Measure.

Second, we use a correlation coefficient analysis to find out how the selection criteria for incoming freshmen affect their GPA after the first semester. We also determine the best predictive admissions criteria for student achievement so that decision-makers may give that factor greater weight.

Third, the institution where this research was done adopted a new admissions policy that gave different weight to the qualities that were shown to be most important. By comparing the cumulative grade point averages of freshmen accepted under the old and new systems, this research demonstrated the wisdom of the latter. The number of freshmen with outstanding or very good cumulative GPAs rose by 31%, while the number with acceptable or bad GPAs fell by 18%.

The huge sample size of 2,039 students from the Faculty of Information and Computer Sciences (CCIS) of Prince Nourah bint Abdulrahman University (PNU) in Riyadh, Kingdom of Saudi Arabia, sets this research apart from others in the area of forecasting student performance (KSA). It has more female students than any other institution in the world. Most previous research in this area validates the efficacy of their models using significantly smaller samples.

II. BACKGROUND WORK

Attributes and prediction techniques are the two most important aspects in predicting student achievement. The cumulative grade point average (CGPA) of college students is the single most important factor in determining their academic success. Many studies have benefited from its use (e.g. [7]–[10]). Assessments, quiz scores, lab work, & final exam grades are additional indicators employed by academic performance studies (e.g. [8], [9],). Only a small number of studies have accounted for demographics, student activities, and social networks as independent variables.

However, in the admissions process, input factors such as pre-admission exams are seldom employed (e.g. [11]-[13]) to predict student success in university. This is what we'll be looking at in detail.

Many different data mining categorization methods have been used to attempt to foretell how well students would do in their courses. In one research, for instance, ANN is used to predict how well 505 students would do in their eighth-semester classes. Using Decision Trees, a model was presented to predict student success in specific courses with little data (32 and 42 students). An analysis of 1,600 students' grades for a single class using Naive Bayes. The research uses SVM on a dataset of 1,074 individuals to forecast the academic success of at-risk freshmen.

Predicting first-year cumulative grade point averages in the computer engineering departments of Saudi public institutions was investigated here using a variety of admissions factors as input qualities. Few research conducted in KSA and published on the topic have focused outside of medical schools (e.g. [11]–[13], [15]). While EDM may help uncover hidden patterns in huge datasets, it has not been applied in these investigations. Study, which used one of the EDM approaches (i.e., J48 decision tree) to usually considered' final GPA based on grades in all classes, is one of the few on the issue that has been limited to a computer science institution. Authors gathered information from the transcripts of 236 Computer Science College students at King Saud University (KSU) from Saudi Arabia. They determined which classes had the most influence on the cumulative grade point average. However, they only used a single EDM methodology on a tiny dataset to make predictions about student performance & did not double-check their work with any other EDM methods.

III. PROPOSED SYSTEM

This study used the Linear Regression method, a common method for determining the connection between independent variables (i.e., predictors) or a predictor variables (i.e., future academic performance), to answer the first question (Can the admission criterion that best predicts future academic performance be identified?) (i.e., response). We used this model to analyse the correlation between the three entrance factors (HSGA, SAAT, and GAT) and the cumulative grade point average (CGPA) after the first two semesters of study. We utilised the correlation coefficient, a standard statistical measure of the strength and direction of linear correlations between two variables, to characterise the linear link between each admission criterion and the CGPA. In addition, we employed the determination coefficient to characterise the relative impact of each admission criterion on the students' first-year CGPA.

Using four well-known data mining classification techniques—Artificial Neural Network (ANN), Decision Tree (DT), Support Vector Machine (SVM), or Naive Bayes—we created four prediction models to address the second issue in this research.

Although there are many other methods for classifying DMs, the following four are the ones we settled on for this article due to the salient qualities they share.

A. *ARTIFICIAL NEURAL NETWORK (ANN)*

The artificial neural network (ANN) is a common tool in evolutionary computational modelling (EDM) because it attempts to replicate the way the human brain works to address difficult issues. It is made up of a collection of modules that take in a set of weighted inputs and produce a corresponding output. Predicting student performance using ANN has been the subject of several published publications (e.g., [8]). We utilised it as well since it can learn from a small sample and can discover all conceivable relationships between variables. The accuracy with which ANN models predicted which applicants would be approved and which would be rejected was also superior to that of traditional classification methods. Due to the limited size of the datasets, the ANN model in this investigation used a simple topology called Multilayer Perception (MLP).

B. DECISION TREE

In a decision tree, the nodes are ranked from most important to least important. Each node in the graph represents a property of an instance, or the edges between them stand for the range of values the property may take. Since this method is so popular among scientists, we choose to employ it (e.g., [1], [6], [7], [9]). It makes uncomplicated and easy-to-understand value predictions. In addition, it performs well with both category and numeric features, requires nothing in the way of complicated data preparation, and expresses rules that can be simply understood and comprehended by users [6].

C. SUPPORT VECTOR MACHINE (SVM)

In order to divide items into their respective categories, this method constructs a hyperplane. The generalisation error of a SVM method decreases with increasing distance from of the hyper - plane to the closest object. SVM is utilised in this work because it is well-suited for tiny datasets, and it has been used in only a limited number of previous studies (e.g., [7]). Additionally, it is quicker than the other methods.

D. NAIVE BAYES

Naive Bayes is a straightforward probabilistic method that uses independent assumptions across variables to apply Bayes' theorem. It calculates the chances of each item belonging to each category. Due to its computational efficiency, widespread usage in relevant literature, and general ease of implementation, this method was selected for this investigation.

3) We used accuracy, recall, precision, or F1-Measure metrics to analyse and compare the performances of the four models in order to answer the third research question in this work (Which datamining prediction approach performs best in this study?). (For a detailed breakdown of evaluation metrics, please refer to Section A, Experimental Design below.)

4) We developed this same second stage of the research to make a comparison the first-year CGPAs of incoming freshmen admitted in 2018-2019 using the new admission weights of criteria to the inaugural CGPAs of former students admitted in 2016-2 using the old admission weight of each criterion, in order to answer the 4th question in this research.

IV. RESULTS

Data from PNU's CCIS student database was used for this analysis. However, the approaches used are universal and may be implemented in any university setting. These figures are from the Admitted & Registration Deanship's computerised academic database. The Institution Review Panel at PNU provided the necessary ethical clearance (Number 19-0152). In the first phase of our research, we gathered data on 1,569 students across all three departments and two cohorts: 902 students from the 2016–2017 school year and 667 students from the 2017–2018 school year.

Second, we gathered 470 student records across the three departments from the 2018-2019 school year, when admissions were made using the revised weightings. Similar situations like those described in the first section of the research occurred in other student bodies. We utilised this information to evaluate how newly accepted students' first-year GPAs stacked up against those of incoming students under the previous weighting system.

Based on the initial numeric parameter CGPA, we built a category target variable (class). The PNU grading system uses a five-point scale, broken down as follows: exceptional (4.5), extremely good (3.75 to 4.5), great (2.75 to 3.75), ordinary (2.0 to 2.75), and bad (2.0). An illustration of the grading scale used by students in the 2016–17 and 2017–18 academic years.

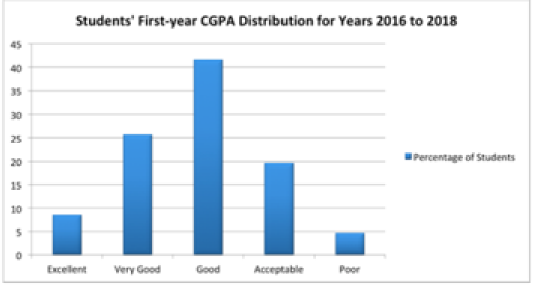

Fig. 1: Students' first-year CGPA distribution for academic year 2016-2107 and academic year 2017-2018.

Just like we did with the data sets spanning 2016–2018 academic years, we performed which was before to the set of data spanning 2018–2019. There were 437 student files left over from the 2018-2019 school year.

We used Artificial Neural Networks (ANNs), Decision Trees, Support Vector Machines (SVMs), and Naive Bayes to build four different prediction models. Each model was developed using a ten-fold cross validation, with 9 data sets used during training and 1 set employed while testing. This procedure was carried out a total of ten times, once for every one of the distinct sets.

The following ideas may be used to evaluate the efficacy of data mining models:

Number of occasions for which a positive prediction was made (True Positive Rate, or TP).

Number of times a negative result was wrongly labelled as positive; sometimes called the "False Positive Rate" (FP).

Number of occurrences when a negative prediction was made accurately; sometimes called the "True Negative Rate" (TN).

One measure of accuracy is the False Negative Rate (FN), which is the percentage of positive events that were wrongly labelled negative.

Predictive accuracy is quantified as the proportion of times an outcome is accurately anticipated.

Preciseness equals (Total Probability) divided by (Total Probability) plus (First Probability) plus (Final Probability) (1)

Measured by (2), recall is the proportion of true positives that were accurately anticipated.

Calculating Recall: TP / (TP + FN) (2)

Accuracy is quantified as the proportion of true positives, and it may be calculated as (3):

TP / (TP + FP) = Accuracy (3)

The F1-Measure highlights a classifier's performance on both common and unusual categories, and it represents the trade-off between recall and accuracy. There are four units of measurement for it:

Formula for F1 Measurement = 2 * Recall * Precision * (Recall + Precision) (4)

To improve first-year student performance, the PNU Admitted & Registration Deanship has chosen to implement changes to the present admissions procedure based on the study's findings and suggestions. For the 2018-2019 school year, the dean's office decided to increase the weight of the SAAT criterion and adjust the weights of the other two criteria (HSGA and GAT) to 30% and 40%, respectively. Prior to this change, the percentages were 60%, 20%, and 20%.

Figure 2 displays a comparison of first-year CGPAs from new students to those from prior years.

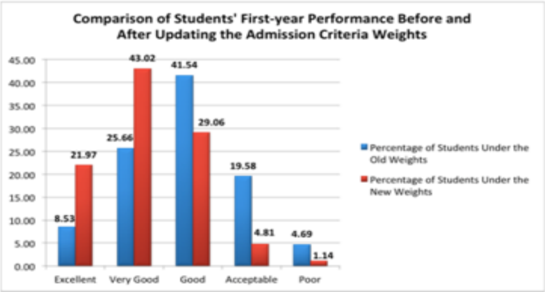

Fig. 2: Comparison of students' first-year performance before and after updating admission criteria weights.

V. CONCLUSION

This research was conducted with the aim of assisting universities in making informed admissions choices based on accurate predictions of prospective students' academic success after they are admitted. Artificial Neural Networks (ANNs), Decision Trees, Support Vector Machines (SVMs), and Naive Bayes were used to propose and create four different models for making predictions. The research used a database including 2,039 student records from PNU, the biggest university in KSA. However, the approaches used are universal and may be implemented in any university setting.

The results of the research lend credence to the employment of prediction models in higher education, where they may be put to good use in the allocation of scarce resources. Moreover, the findings demonstrate that, with sufficient pre-admission data, a high-performance model to predict students' early performance may be created. In this research, for instance, the ANN model achieved an accuracy rate of roughly 79.22% in terms of its performance. The research also found that the ANN method achieved the highest levels of accuracy and precision, while the Decision Tree method achieved the highest levels of recall and F1-Measure. Out of all the methods, Naive Bayes performed the poorest.

VI. FUTURE ENHANCEMENT

Students' personalities, demographics, families, and communication abilities are just a few of the pre-admission characteristics that have been shown to influence future academic success; further research is required. It's possible to employ other data mining methods, such as clustering, in subsequent research.

REFRENCES

[1] H. Guruler, A. Istanbullu, and M. Karahasan, ''A new student performance analysing system using knowledge discovery in higher educational databases,'' Comput. Edu., vol. 55, no. 1, pp. 247–254, Aug. 2010.

[2] S. K. Mohamad and Z. Tasir, ''Educational data mining: A review,'' Procedia Social Behav. Sci., vol. 97, pp. 320–324, Nov. 2013.

[3] A. Peña-Ayala, ''Educational data mining: A survey and a data miningbased analysis of recent works,'' Expert Syst. Appl., vol. 41, no. 4, pp. 1432–1462, Mar. 2014.

[4] C. Romero and S. Ventura, ''Educational data mining: A review of the state of the art,'' IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 40, no. 6, pp. 601–618, Nov. 2010.

[5] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, ''Educational data mining and learning analytics for 21st century higher education: A review and synthesis,'' Telematics Informat., vol. 37, pp. 13–49, Apr. 2019.

[6] C. Anuradha and T. Velmurugan, ''A comparative analysis on the evaluation of classification algorithms in the prediction of students performance,'' Indian J. Sci. Technol., vol. 8, no. 15, pp. 974–6846, Jan. 2015.

[7] V. L. Miguéis, A. Freitas, P. J. V. Garcia, and A. Silva, ''Early segmentation of students according to their academic performance: A predictive modelling approach,'' Decis. Support Syst., vol. 115, pp. 36–51, Nov. 2018.

[8] M. Mayilvaganan and D. Kalpanadevi, ''Comparison of classification techniques for predicting the performance of students academic environment,'' in Proc. Int. Conf. Commun. Netw. Technol., Sivakasi, India, Dec. 2014, pp. 113–118.

[9] S. Natek and M. Zwilling, ''Student data mining solution–knowledge management system related to higher education institutions,'' Expert Syst. Appl., vol. 41, no. 14, pp. 6400–6407, Oct. 2014.

[10] T. M. Christian and M. Ayub, ''Exploration of classification using NB tree for predicting students' performance,'' in Proc. Int. Conf. Data Softw. Eng. (ICODSE), Bandung, ID, USA, Nov. 2014, pp. 1–6.

[11] J. Albishri, S. Aly, and Y. Alnemary, ''Admission criteria to Saudi medical schools. Which is the best predictor for successful achievement?'' Saudi Med. J., vol. 33, pp. 1222–1228, 2012.

[12] M. O. Al-Rukban, F. M. Munshi, H. Abdulghani, and I. Al-Hoqail, ''The ability of the pre-admission criteria to predict performance in a Saudi medical school,'' Saudi Med. J., vol. 31, pp. 560–564, 2010.

[13] A. M. Alhadlaq, O. F. Alshammari, S. M. Alsager, K. A. F. Neel, and A. G. Mohamed, ''Ability of admissions criteria to predict early academic performance among students of health science colleges at King Saud University, Saudi Arabia,'' J. Dental Educ., vol. 79, pp. 665–670, Jan. 2015.

[14] S. M. Hassan and M. S. Al-Razgan, ''Pre-university exams effect on students GPA: A case study in IT department,'' Procedia Comput. Sci., vol. 82, pp. 127–131, 2016.

[15] M. F. Al-Qahtani and T. M. Alanzi, ''Comparisons of the predictive values of admission criteria for academic achievement among undergraduate students of health and non-health science professions: A longitudinal cohort study,'' Psychol. Res. Behav. Manage., vol. 12, pp. 1–6, Dec. 2018.