Online-Journals.org

iJET International Journal of Emerging Technologies in Learning

iJET | eISSN: 1863-0383 | Vol. 18 No. 23 (2023) | OPEN ACCESS

https://doi.org/10.3991/ijet.v18i23.45499

PAPER

Selecting the Best K Features for Predicting Student Participation in Generic Competency Development Activities in Higher Education

Adam Ka Lok Wong[1], Joseph Chi Ho So[1](✉), Kia Ho Yin Tsang[1], Ran Wei[2]

[1]School of Professional Education and Executive Development, The Hong Kong Polytechnic University, Hong Kong, China

[2]Department of Science and Environmental Studies, The Education University of Hong Kong, Hong Kong, China

sochhome@yahoo.com

ABSTRACT

Generic competency (GC) is an essential but often overlooked aspect of developing students in higher education. While there is much research about using technologies to develop discipline-specific skills for students, the use of technologies in GC development is insufficient. In particular, more research is needed on using technologies to predict student participation in GC development activities (GCDAs). Machine learning (ML) can use student characteristics, known as features, to predict their involvement in GCDAs. However, too many features will slow down the prediction process and reduce the ability to pinpoint the best features for prediction. This study explored an effective way to identify the minimal number of features essential for predicting student participation in GCDAs. The findings help educators develop recommendation systems to help students select the most beneficial GCDA for their holistic development. We collected 98 features from 9570 students from a community college. Then, we applied the Principal Component Analysis and SelectKBest algorithms to reduce the number of features from 98 to 8. Finally, we compared the accuracy of predictions using KNN and ANN based on the all-feature dataset with those based on the reduced-feature dataset. The results showed that the reduced-feature dataset maintained good prediction accuracy and enabled the educator to recommend the GCDAs to students. The findings could drive further research and development in applying machine learning technologies to enhance the recommendations for GCDAs for higher-education students.

KEYWORDS

machine learning, education, feature engineering, SelectKBest, KNN, ANN

Lok Wong, A.K., Ho So, J.C., Yin Tsang, K.H., Wei, R. (2023). Selecting the Best K Features for Predicting Student Participation in Generic Competency Development Activities in Higher Education. *International Journal of Emerging Technologies in Learning (iJET)*, 18(23), pp. 197–213. https://doi.org/10.3991/ijet.v18i23.45499

Article submitted 2023-08-07. Revision uploaded 2023-10-03. Final acceptance 2023-10-07.

1 INTRODUCTION

1.1 Generic competence and generic competence development activities

Generic competence (GC) is widely accepted as a critical element in developing higher education (HED) students. The term GC is sometimes called 'generic skills' or 'generic attributes'. While definitions of GC vary, this research adopted the definition that Young & Chapman [1] used. Their definition stated that GC referred to "competencies that can be applied across different job and life contexts". These generic skills usually include skills and abilities in several dimensions, including leadership, teamwork, problem-solving, communication, critical and creative thinking, and social responsibility [23]. These generic competencies are often transferable skills that can be applied to different disciplines. Due to their broader scope of application than discipline-specific skills, there is a growing recognition of the significance of GC in HED [2] [5] [23] [26].

The need for appropriate GC development is not only actively discussed in academia; employers are also very concerned about the competencies of their potential employees. A report by HKEDB [10] noted that the need for GC, including work attitudes, interpersonal skills, and analytical and problem-solving abilities, has continuously ranked more highly than the technical knowledge required for the job over the past 12 years of the survey.

Generic Competence Development Activities (GCDAs) refer to the activities that can help students develop generic competencies. Student activities, including co-curricular and extra-curricular activities (collectively called GC development activities (GCDAs), are essential elements of GC development [4] [17]. Unlike the formal curriculum, engagement in activities supporting whole-person development is usually much less structured. The diversity of activities and students' freedom of choice is much greater than in the formal curriculum. Moreover, participation in GCDAs is also dependent on the students' personalities. Shiah [22] showed that students' personalities can affect their involvement in extra-curricular activities and the development of career skills.

Students in HED have to plan for their campus life and engage in activities to enable them realize their lifetime goals. Otherwise, students may miss promising opportunities or spend time and effort participating in activities that do not match their developmental needs [27]. Students can choose to participate in various GCDAs, but how do they know what activities suit them? Not only do the students themselves need a more precise idea, but their advisors also need better criteria. The advisors need to consider the characteristics of the students, their career goals, and the characteristics of the activities the school offers. As there are so many characteristics to consider, the advice given to a student relies heavily on the advisor's personal experience, so subjectivity is inevitable. We found that there is a lack of research on using ML to help identify the GC development needs of students. Hence, there is a need to develop a systematic and evidence-based approach to facilitate the whole operation.

In machine learning, the term feature refers to a characteristic applied in building a prediction model. Therefore, a feature is a characteristic of an entity such as a student or a GCDA, but not all characteristics of the student or GCDA will be chosen as features. The objective of this research is to find an optimal set of features that can be used to build a model to predict student participation in GCDAs.

1.2 The students and the higher education institution

This study was conducted in a self-financing community college in Hong Kong that enrolls more than 4000 students yearly. The students join the institution after taking their DSE examinations. The institution offers two-year associate degrees and higher diploma programs of various specialized disciplines for local students to choose from, including but not limited to business, science, arts, and social science, to continue their study path after graduating from secondary school. The associate degree (AD) programs prepare students to continue their studies upon graduation to pursue their bachelor's degree in a related discipline. The higher diploma (HD) programs prepare students to become employees in an industry related to the discipline of the HD. To enroll in the institution, students must have achieved Level 2 or above in five subjects in the Hong Kong Diploma of Secondary Education Examination (HKDSE). The five subjects must include English Language, Chinese Language, Mathematics, Liberal Studies, and one elective. The HKDSE is a public exam in Hong Kong organized by Hong Kong Examinations and Assessment Authority (HKEAA). The HKDSE is a significant requirement for entering and applying for higher education in Hong Kong. In Table 2, the subjects are shown with a *HKDSE_* prefix. For example, *HKDSE_ENG* stands for the subject of English Language.

When students complete their studies at the institution, over 90% of them will continue their studies at other local higher educational institutions called articulation partners to get a bachelor's degree. The feedback from these articulation partners indicates that the graduates have sound discipline-specific skills & knowledge. However, they need to improve soft skills such as interpersonal skills, social skills, leadership, critical thinking, and global awareness. In other words, when students graduate from the institution, many still have much room for improvement in their GCs.

To address the above problem, the institution offers many GCDAs through its student affairs office (SAO) to help students to improve their GCs. These GCDAs are promoted to students at orientation time, published at the SAO website, and announced to students as each event is offered. However, student participation in these GCDAs is very low. This is because both the students and their academic advisors need effective ways to match the characteristics of the GCDAs to the developmental needs of each student. The following section describes the characteristics of the GCDAs and the characteristics of the students.

1.3 GCDAs, Holland code and institutional intended learning outcomes

As stated above, in addition to the formal curriculum, the institution also offers various GCDAs. The characteristics of GCDAs refer to the theme, subtheme or projects a specific GCDA belongs to and the Holland Code intended learning outcomes (ILOs) it covers. These GCDAs are grouped under themes and subthemes, also known as projects. Some projects, such as language enhancement workshops, are free of charge. Other projects, such as overseas service trips, are significantly subsidized by the institution. However, despite the promotions by the institution, these free or subsidized GCDAs often had low student participation rates. The GCDAs are designed according to the Holland Code career theory and to cover the ILOS of the institution. Holland Code is a system that helps people to align their personalities with their careers.

A match in personality and career can help a person thrive in his or her career and boost satisfaction. When personalities and careers lack consilience, it can negatively impact career path, performance, and satisfaction [8]. The Holland Code, an occupation-based classification system, classifies people's different interests in work according to six different human personalities. These include Realistic, Investigative, Artistic, Social, Enterprising, and Conventional [7]. The letter and the meaning of the codes are listed below.

- R – Realistic
- I – Investigative
- A – Artistic
- S – Social
- E – Enterprising
- C – Conventional

Tracey and Rounds [28] suggested an exclusionary relationship between these six personalities. They suggested that these types are mutually exclusive because each occupational interest type represents a group with similar interests and personality traits. In other words, a person is most likely to belong to one of these types rather than multiple types. This is because each type represents a particular occupational interest and personality trait that does not usually exist in the same person simultaneously [9]. For this reason, each program of study has only 3 to 4 related Holland Codes out of the possible 6. For example, the program of Associate in Information Technology focuses on the R (Realistic), I (Investigative), and E (Enterprising) of the Holland Codes. The mapping of the GCDAs to Holland Code and ILOs is shown in the following table.

The students reported their self-rated values for each of the Holland Code twice while studying at the institution. When they started studying at the institution, they entered their self-rated values for each Holland Code during orientation workshops. When they graduated from the institution two years later, they entered self-rated values of each Holland Code again in an exit survey. Table 2 shows these values with the prefix of *HC_Entry_* and *HC_Exit_*, respectively. Two examples are shown below:

- *HC_Entry_R*: self-report value for the Realistic characteristic in Holland Code at *orientation (entry)*
- *HC_Exit_R*: self-report value for the Realistic characteristic in Holland Code at *graduation (exit)*

The students reported their self-rated values for each of ILOs twice while studying at the institution. When they started studying at the institution, they entered their self-rated values for each ILOs during orientation workshops. When they graduated from the institution two years later, they entered self-rated values of each ILO again in an exit survey. Table 2 shows these values with the prefix of *Entry_* and *Exit_*, respectively. Two examples are shown below:

- *Entry_Lifelong Learner*: self-report value for the ILOs of Lifelong learner at *orientation*
- *Exit_Lifelong Learner*: self-report value for the ILOs of Lifelong learner at *graduation*

The self-rated values for the Holland Code and ILOs are what we see as the characteristics of each student. They describe the people the students are.

Table 1a. Mapping of the themes and projects of the GCDAs to Holland codes

Themes	Projects	Holland Code
Sustainability and Knowledge Enrichment	Art Appreciation	A
Contribution and Services/Global Exposure	Volunteer and Community Services	S, R, C
Global Exposure	Global Exploration and Cultural Exchange	R, S
Leadership and Communication	Art Creation	A
	Ceremony Presenter Training	S
	Student Ambassador Programme	S
	Chinese Enhancement Programme	S, A
	English Enhancement Programme	S, A
Sustainability and Knowledge Enrichment	Complementary Studies Programme	R, I, A, S, E, C (depends on the courses they applied)
Physical and Psychological Wellness	Sports and Fitness	S
	Psychological Wellness	S
Career Development	Career Projects	
	Mentorship Programme	E, C, S, R
	Self-initiated Project Scheme	
	Young Entrepreneur Scheme	E, S
	Future and New Skills Training	I
	Work Ethics Awareness	R
Recognition	Challenge and Explore Series	R, I, S
Unknown	Overseas Service Trips	R, I, S
	Self-learning Language Centre	A

Table 1b. Mapping of projects in GCDAs to the ILOs (Leftmost six columns)

Projects	Competent (Associate) Professional	Effective Communicator	Innovative/ Practical Problem Solver	Lifelong Learner	Ethical Leader/Citizen	Global Awareness
Art Appreciation		✓		✓		
Volunteer and Community Services	✓		✓		✓	✓
Global Exploration and Cultural Exchange		✓	✓			✓
Art Creation			✓			
Ceremony Presenter Training		✓				
Student Ambassador Programme		✓				
Chinese Enhancement Programme		✓				
English Enhancement Programme		✓				
Complementary Studies Programme				✓		
Sports and Fitness				✓		
Psychological Wellness				✓		
Career Projects	✓		✓	✓		
Mentorship Programme	✓			✓		
Self-initiated Project Scheme			✓	✓		
Young Entrepreneur Scheme	✓		✓			
Future and New Skills Training	✓		✓			
Work Ethics Awareness					✓	
Challenge and Explore Series			✓			
Overseas Service Trips					✓	✓
Self-learning Language Centre		✓		✓		

2 MACHINE LEARNING & FEATURE SELECTION

2.1 Role of machine learning in education

Machine learning (ML) is getting increasing attention in the education field, and recent development of machine learning has been applied to enhance education quality [31]. ML is often considered an essential part of education technology because it can solve technical problems such as identifying unknown patterns or predicting student performance for timely interventions [12]. ML can be applied to analyze the data gathered through learning management systems (LMS), student management systems (SMS), and student surveys. The application of machine learning, such as educational chatbots, has great potential to complement human educators and educational administrators. For example, it can be a 24/7 counsellor, answering and clarifying any questions from absent students in higher education [19]. Some studies have investigated how ML is to be used in institutional service provision in higher education [13] [18]. A deep learning approach has even been proposed for analyzing people's sentiments (positive, negative, and neutral) towards higher education distance learning [14].

Providing personalized recommendations to students addressing individuals' unique developmental needs is highly desirable. In recent years, educational recommender systems (ERS) have attracted significant attention as a solution for learners. ERS is crucial in helping learners find educational resources relevant to their material and context, and it can also help students to select some adequate courses [15]. Some studies have also shown that recommender systems greatly help and support students' eLearning [3] [25]. However, implementing ERS is always challenging because of the considerable student population and the difficulty in analyzing quantitative measurement across different data sources, particularly concerning GCDAs among students. In particular, we cannot find relevant studies about recommending appropriate GCDAs for students using machine learning. Research on this part is still in the developmental stage.

Universities and educational institutions are eager to predict the enrolment of various student activities accurately, as the provision involves precious college resources, and the enrolment also reflects the success of the activities. However, many factors may affect students' involvement in the GCDAs. Our study investigates the feasibility of applying ML to predict the students' participation in the GCDAs. In deciding the features to be included in the dataset, it is crucial to determine which attributes are significant for the machine learning training to reduce unnecessary performance deterioration due to irrelevant features [21]. Features are similar to characteristics. They both describe the GCDAs and students, but feature is the term used in machine learning. The present study attempts to reduce the features required for ML models yet maintain the same level of performance as the ML models trained with the whole dataset. Artificial Neural Network (ANN) is used to examine the performance. ANN resembles the human brain by comprising input, hidden, and output layers linked by nodes that analyze the correlation between input and output variables [16] [29].

2.2 The challenge of feature selection

The processing time and accuracy of an ML algorithm depend on number and quality of the input variables, which are called features. If only a few features are

fed into an ML algorithm, the processing time is very short, but the accuracy may not be good because the most important features may not be included. If all the possible features are fed into an ML algorithm, the processing time will be excessive, and the accuracy may only improve marginally. Therefore, one of the challenges in ML problems is to find the optimal number of features that can provide a reasonably good accuracy [24].

This study contributes to the literature in ML by identifying the most effective predictors of GCDAs in higher education. This will help educators develop recommendation systems to help students select the most beneficial GCDA for their holistic development. This study collected 98 features from 9570 students during their study at a community college from 2019 to 2021. The features in the dataset are shown in Table 2. This study aims to investigate the effect on prediction accuracy when the SelectKBest algorithm is applied to reduce the number of features. Firstly, the all-feature data were preprocessed and then fed to the SelectKBest algorithm to reduce the number of candidate predictors. Finally, the reduced set of features was used in two prediction algorithms to compare the accuracy of the reduced feature set. We suggest using the findings to drive further research and development in machine learning technologies to enhance the recommendations for GCDAs to students in higher education.

Table 2. All 98 possible features for training the models for predicting GCDA participation

Demographics						
Programme of Study	Gender					
Academic Profile—HKDSE Subject Results						
HKDSE P Score	HKDSE ENG	HKDSE CHI	HKDSE MATH	HKDSE MATH_ EXT1	HKDSE MATH_ EXT2	HKDSE LIB_STDY
Academic Profile—Performance at the institution						
CumlGPA Semester 1	CumlGPA Semester 2	CumlGPA Semester 3	CumlGPA Semester 4	LCH Avg		
Activity Participation—GCDA participation by Project (P_), Theme (T_) and Subtheme (ST_)						
P_Acapella Ensemble	P_Artist-in-Campus	P_Volunteer Network	P_Campus TV	P_Challenge & Explore Series	P_Chinese Mainl& Services	P_Complementary Studies Programme
P_Counselling Services	P_Executive Leadership	P_Green Lifestyle Series	P_Hydroponic Gardening	P_Local Community Services	P_Mentorship Programme	P_University Admission Talks
P_Overseas Service Trips	P_Photography Production	P_Physical Education	P_Physical Wellness	P_Wofoo Leaders' Network	P_Self-learning Language Centre	P_Work Ethics Awareness
P_Career Projects	P_Psychological					
T_Leadership & Communication	T_Career Development	T_Contribution & Services	T_Counselling Services	T_Further Studies	T_Global Exposure	T_Physical & Psychological Wellness
T_Sustainability & Knowledge Enrichment	ST_Aesthetics Appreciation	ST_Communication Enhancement	ST_Community Service Opportunities	ST_Outbound Experience	ST_Physical Wellness	ST_On-Campus Services
ST_Complementary Stduies	ST_Complementary Studies	ST_Leadership	ST_Psychological Wellness	ST_Sustainability	ST_Career Exploration & Planning	

(Continued)

Table 2. All 98 possible features for training the models for predicting GCDA participation *(Continued)*

Activity Participation—GCDA participation by Intended Learning Outcomes (ILOs)						
ILO_Lifelong Learner	ILO_Effective Communicator	ILO_Competent (Associate) Professional	ILO_Ethical Leader/Citizen	ILO_Innovative/ Practical Problem Solver		
GC_Language & Communication	GC_Physical & Psychological Wellness	GC_Career & Personal Development	GC_Creativity & Innovation	GC_Critical Thinking & Problem Solving	GC_Global Citizenship	GC_Social Responsibility
GC_Teamwork & Collaboration	GC_Chinese Language Proficiency	GC_English Language Proficiency				
Self-perception & Characteristic—Self-reported Holland Code (HC) R,I,A,S,E,C						
HC_Entry_R	HC_Entry_I	HC_Entry_A	HC_Entry_S	HC_Entry_E	HC_Entry_C	
HC_Exit_R	HC_Exit_I	HC_Exit_A	HC_Exit_S	HC_Exit_E	HC_Exit_C	
Self-perception & Characteristic—Self-reported Achieved Levels of Learning Outcomes (ILOs)						
Entry_Lifelong Learner	Entry_Competent Professional	Entry_Critical Thinker	Entry_Effective Communicator	Entry_Problem Solver	Entry_Ethical Citizen	Entry_Global Awareness
Exit_Ethical Citizen	Exit_Lifelong Learner	Exit_Competent Professional	Exit_Critical Thinker	Exit_Effective Communicator	Exit_Problem Solver	Exit_Global Awareness

3 PROPOSED METHODOLOGY

The proposed methodology of this study is shown in Figure 1. Firstly, the all-feature dataset was preprocessed using one-hot encoding and scaling. The features were divided into 4 major categories: demographics, academic profile, activity participation, and self-perception & characteristics. Then the data was divided into training and testing sets to train two models, KNN and ANN, using all the features. The principles of these two models will be explained in another section below. The accuracy of these two all-feature models on the testing set provided a baseline (control) for evaluating the performance of the reduced-feature models, which also used KNN and ANN. The accuracies of these two all-feature models are named KNN-control and ANN-control. Then we used the SelectKBest and PCA algorithms to reduce the number of features. After that, the reduced-feature dataset was used to train another set of KNN and ANN models. Finally, the accuracy of the reduced-feature models on the testing set were compared with the baseline accuracies (i.e., all-feature dataset), which are the KNN control and ANN control. The steps in Figure 1 correspond to the following steps.

1. Calculate all-feature accuracies: Use KNN and ANN to calculate the accuracy of prediction using all features (KNN-control and ANN-control).
2. Reduce features: Use PCA to find the number of features required to explain 80% of variance in the dataset.
3. Reduce features: Use SelectKBest to find and rank the features based on findings from the PCA step.
4. Calculate reduced-feature accuracies: Use KNN and ANN to calculate the accuracy of prediction using the reduced-feature dataset and compare with KNN-control and ANN-control.

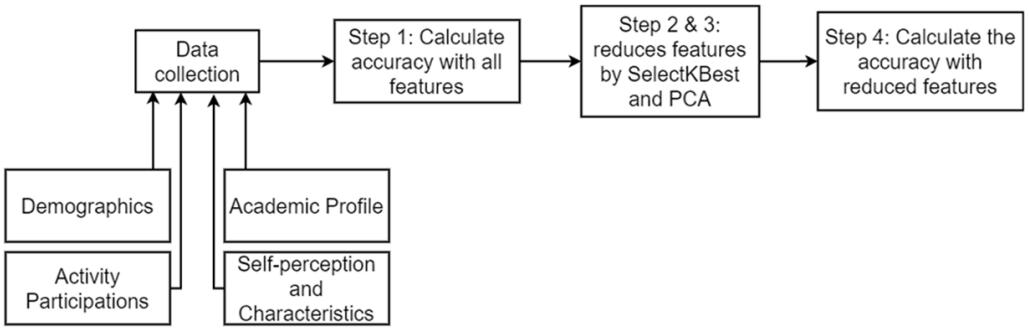

Fig. 1. Proposed methodology to reduce the number of features

3.1 Feature reduction

Due to the sheer volume of features and records, pre-processing and feature engineering are required before feeding the dataset to a machine learning (ML) model. It is commonly used to produce a meaningful feature set for the ML algorithm [11]. The participation hours in each project of activities are calculated, and the naming of CGPA is unified to semester one and semester 4. The dataset featured in this study contains 98 features and 9570 records after pre-processing and feature engineering. The features include student demographics, academic profile, Holland Code, Student Development Assessment survey and activity participation. The demographics and academic profile include the gender, program of studies, public examination results and the CGPA of each student. The Holland Code and Student Development Assessment survey record the students' self-perception about their characteristics and potentially desired career paths. However, 98 features are too many for the ML model as some features are redundant or insignificant to the prediction of the ML model. To facilitate the recommendation system in making a more reasonable prediction on student activity participation in our case, we proposed using SelectKBest and Principal Component Analysis (PCA) from the Python Sckit-Learn library to reduce features. After the features are reduced to a much smaller set, the data was fed into two ML algorithms to compare their accuracies. For this study, the two algorithms were K-nearest neighbor (KNN) and artificial neural network (ANN). This study uses these two algorithms to determine the difference between the effects of the two feature-reduction methods. kNN and ANN were chosen as they are proven ML algorithms that can provide reasonably good performance. They can be used alongside PCA and SelectKBest to observe the changes in the performance of ML algorithms when different feature reduction methods are applied.

3.2 Baseline, K-nearest neighbor (KNN) & Artificial Neural Network (ANN)

This study is a classification machine learning problem. In a classification problem, the goal is to find best features as independent variables to predict the target dependent variable (also called a label or target) using a chosen model. In this case, a label, called "Max Project" was created as a new column in the dataset. The "Max Project" indicates the project that a student spent the most hours in and is used as the label in this study. Those who have 0 hours across all projects are filtered as no activity participation yields meaningless results. After filtering, a total of 2269 records remain.

KNN is a data mining algorithm utilized for classification purposes. For example, it has been used as an ML algorithm to classify autism spectrum disorder among people belonging to different age groups [24]. It is suitable for this study because the target of the algorithm is to find out, that is, to classify, which project in the GCDAs is the Max Project for a student. The KNN algorithm involves the following steps [24].

1. Obtain the unclassified data
2. Evaluate the distance from new data to all other already categorized data
3. Calculate k value
4. Review the list of classes at the minimum distance, counting the number of every appearing class
5. Selection of the class that occurs most often as the right one
6. Classify actual data with the class obtained in (5)

The distance D between two points a and b can be calculated using the following formulas, where a_i, b_i stands the values of the i-th feature for two data points in the dataset.

$$D = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2} \quad (1)$$

In the above equation, if the values of a_i and b_i are numbers representing categorical values, they must be changed into ones and zeroes using *one-hot encoding*. Furthermore, these values must be normalized so that some features do not dominate the final calculated values. For example, the values of self-reported ILOs range from 1 to 5, while the values of self-reported Holland Code range from 1 to 3 only. The self-reported ILOs will always dominate over the self-reported Holland Codes and become the more important feature among the two features. *Normalization* solves this problem by converting these values into 0 to 1, regardless of the actual values before normalization.

ANN resembles the human brain by comprising input, hidden, and output layers linked by nodes that analyze the correlation between input and output variables [16] [29]. It has been used as a classifier to predict student academic performance based on the interactivity with the e-learning management [6]. When using ANN, the same target label, Max Project, was used to calculate the accuracy of the model. It is worth noting that due to the presence of hidden layer(s) in an ANN model, the model can provide predictions, but it cannot rank the input features in terms of their relative importance.

3.3 SelectKBest and Principal Component Analysis (PCA)

In this study, we used the modules of SelectKBest and Principal Component Analysis (PCA) from the Python Sckit-Learn library to reduce features. PCA and SelectKBest can be found in the Python Sckit-Learn library [20]. PCA is used to find the prominent features. SelectKBest selects the best features determined by Chi square. Dominant patterns in a data matrix are extracted using PCA with a complementary set of score and loading plots [30]. Our study uses PCA to find the prominent features that could explain most of the variances in the dataset. In theory, the more number of features chosen, the higher the variance can be explained by the features. We studied the number of features required to achieve the level of variance

explained from 0.80, 0.90, 0.95 and 0.99. Figure 2 shows the relationship between the number of features and the variance in the dataset explained by the features. The figure shows that when there are only 10 features to be chosen as principal components, the variance explained was already at 80%. After that the accuracy marginally increased by only 1% or less for each feature added. While PCA provides a good estimate of the number of features that can be used to be transformed into components that can explain to a specific degree of variance in the dataset, it does not directly indicate the relative importance of these features. This implies that the PCA does not provide actionable estimates for the educator to form the basis of recommendations to student about GCDAs. This is where SelectKBest can complement the limitations of PCA.

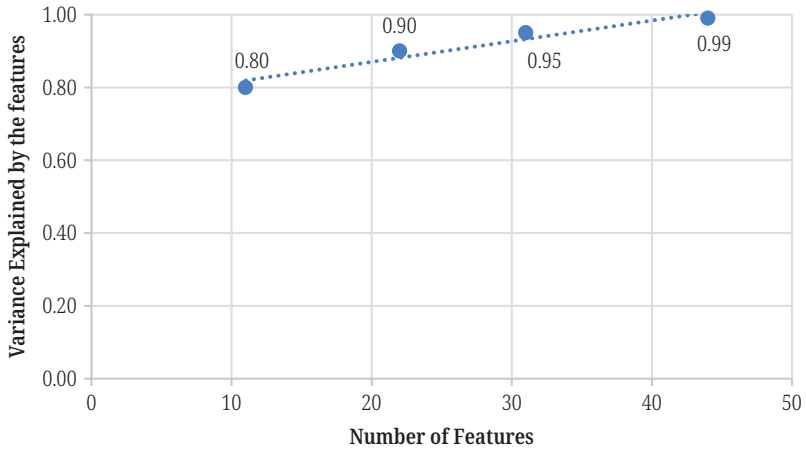

Fig. 2. Number of features versus variance in the dataset explained by the features

SelectKBest can complement the limits of PCA in terms of feature reduction. Unlike PCA, SelectKBest allows the user to directly specify a number called Kf, the number of most important features, and returns the names of the features and their relative importance. This enables the educator to take action and make recommendations based on a reasonably small number of features. Based on the findings from the PCA, the number of features in SelectKBest should be less than 10. In this study, the values of Kf were chosen to be 3, 5 and 8. Using the respective reduced-feature datasets, we applied KNN and ANN to calculate the prediction accuracies.

3.4 Evaluate metrics

Accuracy is a metric for researchers to measure the correctness of a machine learning model. The equation for accuracy is given below. A prediction is regarded as correct if the predicted *Max Project* is the same as the actual *Max Project* participated by a student. If the accuracy is 0.00, it means all the model's predictions about the testing dataset are incorrect. If the accuracy reaches 1.00, it means the model can perfectly predict or classify all given records in the testing dataset.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (2)$$

4 RESULTS & DISCUSSIONS

4.1 Accuracies

The accuracy of an ML model indicates the fraction or count of correct predictions made by the respective ML model [20]. Our study indicates the percentage of correct predictions of "Max Project" made by respective ML models compared to the test data fed to the ML models. As stated in section 3, we applied KNN and ANN on the all-feature dataset and calculated the accuracy of predictions of these two algorithms. The resulting accuracies are called KNN-control and ANN-control. Similarly, we applied KNN and ANN on the reduced-feature dataset and calculated the accuracy of predictions of these two algorithms. The resulting accuracies are called KNN and ANN. Figure 3 compares these accuracies when the number of features in the reduced-feature dataset is 3, 5, and 8, respectively. Note that KNN-control and ANN-control are calculated from the all-feature dataset and, therefore, remain the same.

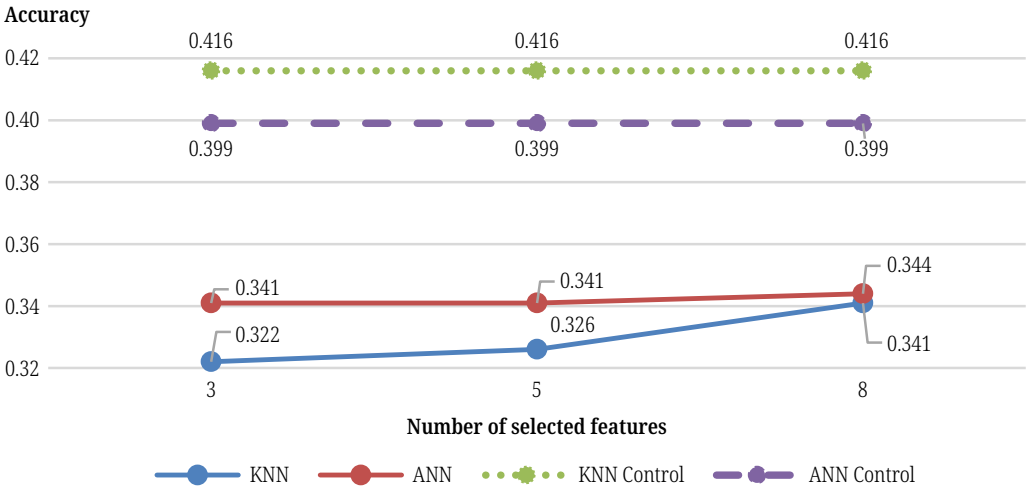

Fig. 3. The accuracies of using all-feature dataset and reduced-feature dataset

The results show that the accuracy computed by KNN and ANN based on the reduced-feature dataset is just marginally less than those computed based on the all-feature dataset. Furthermore, the accuracies are much better than a random guess. It is because the number of possible candidates for the "Max Project" feature is 25. If a project is chosen at random as the prediction, the possibility of the randomly chosen project being the correct Max Project is only 1 in 25, or 0.04. Since both the accuracies of both KNN and ANN models using the reduced-feature dataset are close to 0.34, the two ML models using the reduced-feature dataset are still promising.

It is also observed that the accuracies improve as the number of features increases from 3 to 8. The accuracy of KNN increased significantly from 0.322 to 0.341, while ANN's accuracy increased marginally from 0.341 to 0.344. When the number of features is 8, both KNN and ANN achieved similar accuracies. Furthermore, reducing features from 98 to 8 means the institution's advisors can make recommendations based on only a few known features.

4.2 Selected features

The main objective of this study is to explore an effective method to identify the minimal number of features essential for predicting student participation in GCDAs. We applied the method to create a reduced-feature dataset that can provide good prediction accuracy for the target variable Max Project. Table 3 shows the most important K features when the SelectKBest algorithm is applied and when K changes from 3 to 8. The SelectKBest algorithm can calculate a score for each of the K features. A feature with a higher score value means that the feature is a more important predictor of the target than a feature with a lower score [32]. It is worth noting that the different programs of study are listed as different features. This is because of using one-hot encoding in the data preprocessing stage. The program of study is a feature that contains categorical values such as Associate of Science, Associate In Chinese Language And Literature, and CumlGPA Semester 3. The values of this feature are nominal and do not have any order. In one-hot encoding, for each level of a categorical feature, we create a new variable, and each category is mapped with a binary variable containing either 0 or 1.

Table 3 shows the relative importance of the features in the reduced-feature dataset. The top three features with the highest scores are Associate In Design (Visual Communication), Associate In Chinese Language And Literature, and CumlGPA Semester 3. These features consistently demonstrate the top three most decisive influences on the target variable, Max Project. The fourth and fifth features have very similar scores to the third feature. That means the features Associate In Health Studies and CumlGPA Semester 4 contain essentially the same valuable information for predicting the GCDA as the third feature. However, as the number of features increases to 8, the scores decrease significantly. This indicates that the last three features, Associate Of Science, Associate In Design (Advertising Design), and Associate In Business (International Business) are much less effective predictors of the target variable, Max Project.

Table 3. The most important features in the reduced-feature dataset

Features in the Reduced-Feature Dataset	Scores in Descending Order
1. Associate in Design (Visual Communication)	173
2. Associate in Chinese Language and Literature	100
3. CumlGPA Semester 3	99
1. Associate in Design (Visual Communication)	173
2. Associate in Chinese Language and Literature	100
3. CumlGPA Semester 3	99
4. Associate in Health Studies	99
5. CumlGPA Semester 4	98
1. Associate in Design (Visual Communication)	173
2. Associate in Chinese Language and Literature	100
3. CumlGPA Semester 3	99
4. Associate in Health Studies	99
5. CumlGPA Semester 4	98
6. Associate of Science	90
7. Associate in Design (Advertising Design)	73
8. Associate in Business (International Business)	70

4.3 Interpreting & using the results

Although the main objective of this study is to explore an effective way to identify the minimal number of features essential for predicting student participation in GCDAs, it would be helpful to briefly discuss how to use the selected features. There are some interesting points that can inform computer scientists and student advisors in higher education about using machine learning in making recommendations.

Firstly, the table results are consistent with advisors' intuitions based on their experience. The CumlGPA Semester 3 and CumlGPA Semester 4, rank very high in the table. It is because all students in the institution aim to pursue a bachelor's degree upon graduation. The students usually will receive conditional offers from the articulation partners in Semester 3. The condition of the offer is usually a specific graduation GPA, say 3.0. As long as the student achieves the minimum condition, the actual GPA at the end of Semester 4 does not matter. By Semester 3, the students will have a very good idea of whether they will fulfill that condition. Therefore, many students are willing to spend less time on their academic studies to develop their generic competencies. Secondly, the program of study should not be treated as one single characteristic for recommending GCDAs to students. The table shows that if a student belongs to Associate In Design (Visual Communication) or Associate In Chinese Language And Literature, the advisor should consider the program's characteristics more than the student's academic results in Semester 3. As explained in section 1.3, each program of study has only 3 or 4 related Holland Codes. For example, suppose the student belongs to the Associate of Information Technology. In that case, the advisor can focus on the projects that are relevant to the Holland Codes R, I and E. Then, the recommendation can be made based on the mapping between Holland Code and GCDA projects, as shown in Table 1a. Finally, the specialty programs, such as design and Chinese Language, rank higher than the generic programs, such as science and business.

5 CONCLUSIONS, LIMITATIONS AND FURTHER RESEARCH

Feature selection is crucial in machine learning tasks as it helps identify the most informative and discriminative features for model training. This study aimed to find an effective way to identify the minimal number of features essential for predicting student participation in GCDAs. For that aim, we investigated the effect on prediction accuracy when the Principal Component Analysis and SelectKBest algorithms were combined to reduce the number of features. We applied these two algorithms to reduce the number of features from 98 to 8. We found that using the proper data pre-processing, such as one-hot encoding and scaling, the accuracies of both KNN and ANN using the reduced-feature dataset remain comparable with those of the all-feature dataset. We found that the reduced-feature dataset maintained good prediction accuracy and enabled the educator to recommend to students about the GCDAs.

There are certain limitations this study faced. Firstly, other than their program of study, academic results, personalities, and other factors may affect student's actual participation in GCDAs. For example, students' decisions to join a GCDA may depend on whether their friends joined the activity, if the activity costs too much, or if it clashes with their lessons. The current study is limited by the historical dataset, which has not included these factors. Secondly, the study is also limited by the demographic data collected by the institution. It would be more useful if more demographic data could be collected. In the institution, there is a significant number of non-Chinese students. It would be interesting to include race as a demographic feature. Finally, the data is

restricted by the format of students' responses in the historical dataset. For example, for the self-reported ratings on Holland Code, a student can only respond using "Yes" or "No" to a question such as "I like to work alone." If a finer scale, such as a four-point scale, is used, the student can choose a response from "Strongly Agree", "Agree", "Disagree", and "Strongly Disagree". This may improve the Holland Code as a predictor of GCDAs.

For future research, the reduced-feature dataset can be used to develop a recommendation system for students to get the most suitable GCDAs for them. In the long run, data will be tracked longitudinally, which will eventually be a part of the learning portfolio to showcase the whole-person development of individual students.

6 ACKNOWLEDGMENT

This study was supported by the Faculty Development Scheme (No. UGC/FDS24/E09/20) of the University Grant Committee of Hong Kong.

7 REFERENCES

[1] J. Young and E. Chapman, "Generic competency frameworks: A brief historical overview," *Education Research and Perspectives*, vol. 37, no. 1, pp. 1–24, 2010.

[2] S. Barrie, "A conceptual framework for the teaching and learning of generic graduate attributes," *Studies in Higher Education*, vol. 32, no. 4, pp. 439–458, 2007. https://doi.org/10.1080/03075070701476100

[3] O. Bourkoukou and E. Bachari, "A big-data oriented recommendation method in E-learning environment," *International Journal of Emerging Technologies in Learning*, no. 10, 2022. https://doi.org/10.3991/ijet.v17i10.27861

[4] W. Chan, "Students' understanding of generic skills development in a university in Hong Kong," *Procedia—Social and Behavioral Sciences*, vol. 2, no. 2, pp. 4815–4819, 2010. https://doi.org/10.1016/j.sbspro.2010.03.776

[5] L. Duggan, "A quantitative analysis of students' perception of generic skills within an undergraduate electronics/mechanical engineering curriculum," *Online Submission*, 2014.

[6] T. Hamim, F. Benabbou, and N. Sael, "Survey of machine learning techniques for student profile modeling," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 4, pp. 136–151, 2021. https://doi.org/10.3991/ijet.v16i04.18643

[7] J. L. Holland, "A theory of vocational choice," *Journal of Counseling Psychology*, vol. 6, no. 1, pp. 35–45, 1959. https://doi.org/10.1037/h0040767

[8] J. L. Holland, "Exploring careers with a typology: What we have learned and some new directions," *Am. Psychol.*, vol. 51, no. 4, pp. 397–406, 1996. https://doi.org/10.1037/0003-066X.51.4.397

[9] J. L. Holland, "Making vocational choices: A theory of vocational personalities and work environments," *Psychological Assessment Resources*, 1997.

[10] C. S. Hong Kong, "Survey on opinions of employers on major aspects of performance of sub-degree graduates" *HK Education Bureau*. https://www.cspe.edu.hk/

[11] Y. Jiang, "Deep neural networks: Which is better for sensor-free affect detection?," *Artificial Intelligence in Education*, pp. 198–211, 2018. https://doi.org/10.1007/978-3-319-93843-1_15

[12] C. Korkmaz and A. P. Correia, "A review of research on machine learning in educational technology," *Educational Media International*, vol. 56, no. 3, pp. 250–267, 2019. https://doi.org/10.1080/09523987.2019.1669875

[13] V. Kuleto, "Exploring opportunities and challenges of artificial intelligence and machine learning in higher education institutions," *Sustainability*, no. 18, 2021. https://doi.org/10.3390/su131810424

[14] I. Lasri, A. Riadsolh, and M. Elbelkacemi, "Self-attention-based Bi-LSTM model for sentiment analysis on tweets about distance learning in higher education," *International Journal of Emerging Technologies in Learning*, vol. 18, no. 12, 2023. https://doi.org/10.3991/ijet.v18i12.38071

[15] O. Mazhoud, A. Kalboussi, and A. H. Kacem, "Educational recommender system based on learner's annotative activity," *International Journal of Emerging Technologies in Learning*, no. 10, 2021. https://doi.org/10.3991/ijet.v16i10.19955

[16] M. Mohseni-Dargah, Z. Falahati, B. Dabirmanesh, P. Nasrollahi, and K. Khajeh, "Chapter 12—Machine learning in surface plasmon resonance for environmental monitoring," in *Artificial Intelligence and Data Science in Environmental Sensing*, M. Asadnia, A. Razmjou, and A. Be-Heshti Eds, Eds. Academic Press, 2022, pp. 269–298. https://doi.org/10.1016/B978-0-323-90508-4.00012-5

[17] T. Nghia, "Developing generic skills for students via extra-curricular activities in Vietnamese universities: Practices and influential factors," *J. of Teaching and Learning for Graduate Employability*, vol. 8, no. 1, 2017. https://doi.org/10.21153/jtlge2017vol8no1art624

[18] Y. Nieto, V. Gacia-Diaz, C. Montenegro, C. C. Gonzalez, and R. Gonzalez Crespo, "Usage of machine learning for strategic decision making at higher educational institutions," *IEEE Access*, vol. 7, pp. 75007–75017, 2019. https://doi.org/10.1109/ACCESS.2019.2919343

[19] K. Palasundram, N. M. Sharef, N. Nasharuddin, K. Kasmiran, and A. Azman, "Sequence to sequence model performance for education chatbot," *Interternational Journal of Emerging Technologies in Learning (iJET)*, vol. 14, pp. 56–68, 2019. https://doi.org/10.3991/ijet.v14i24.12187

[20] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *arXiv [cs.LG]*, 2012.

[21] G. T. Reddy *et al.*, "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020. https://doi.org/10.1109/ACCESS.2020.2980942

[22] Y. J. Shiah, Y. Huang, F. Chang, C. F. Chang, and L. C. Yeh, "School-based extracurricular activities, personality, self-concept, and college career development skills in Chinese society," *Educational Psychology*, vol. 33, no. 2, pp. 135–154, 2013. https://doi.org/10.1080/01443410.2012.747240

[23] J. C. So, S. Lam, and Y. So, *A Case Study Of Generic Competencies Among Science And Technology Tertiary Graduates in Hong Kong. Teaching, Asses. and Learning for Engg*. Bali, IEEE, 2013.

[24] R. Sujatha, S. L. Aarthy, J. Chatterjee, A. Alaboudi, and N. Z. Jhanjhi, "A machine learning way to classify autism spectrum disorder," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 6, pp. 182–200, 2021. https://doi.org/10.3991/ijet.v16i06.19559

[25] J. Talaghzi, M. Bellafkih, A. Bennane, M. M. Himmi, and M. Amraouy, "A combined E-learning course recommender system," *International Journal of Emerging Technologies in Learning (Online)*, vol. 18, no. 6, 2023. https://doi.org/10.3991/ijet.v18i06.36987

[26] H. Tait and H. Godfrey, "Defining and Assessing Competence in Generic Skills," *Quality in Higher Education*, vol. 5, no. 3, pp. 245–253, 1999. https://doi.org/10.1080/1353832990050306

[27] M. Torenbeek, E. Jansen, and A. Hofman, "The effect of the fit between secondary and university education on first-year student achievement," *Studies in Higher Edu*, vol. 35, pp. 659–675, 2010. https://doi.org/10.1080/03075070903222625

[28] T. J. Tracey and J. B. Rounds, "Evaluating Holland's and Gati's vocational-interest models: A structural meta-analysis," *Psychological Bulletin*, vol. 113, 1993. https://doi.org/10.1037/0033-2909.113.2.229

[29] N.-T. Vu and K.-U. Do, "Chapter 27—prediction of ammonium removal by biochar produced from agricultural wastes using artificial neural net-works: Prospects and bottlenecks," in *Soft Computing Techniques in Solid Waste and Wastewater Management*, R. R. Karri, G. Ravindran, and M. H. Dehghani Eds, Eds. Elsevier, 2021, pp. 455–467. https://doi.org/10.1016/B978-0-12-824463-0.00012-4

[30] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometr. Intell. Lab. Syst.*, vol. 2, no. 1–3, pp. 37–52, 1987. https://doi.org/10.1016/0169-7439(87)80084-9

[31] H. Zeineddine, U. Braendle, and A. Farah, "Enhancing prediction of student success: Automated machine learning approach," *Comput. Electr. Eng.*, vol. 89, no. 106903, p. 106903, 2021. https://doi.org/10.1016/j.compeleceng.2020.106903

[32] J. Gómez-Ramírez, M. Ávila-Villanueva, and M. Á. Fernández-Blázquez, "Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods," *Sci. Rep.*, vol. 10, no. 1, p. 20630, 2020. https://doi.org/10.1038/s41598-020-77296-4

8 AUTHORS

Adam Ka Lok Wong is a senior lecturer at the School of Professional Education and Executive Development of the Hong Kong Polytechnic University. Dr. Wong had over 20 years of experience in the Information Technology industry before becoming a lecturer in 2020. He has implemented government-funded research and education enhancement projects. His teaching and research areas are pedagogy, machine learning and artificial intelligence (E-mail: adam.wong@cpce-polyu.edu.hk; spklwong@speed-polyu.edu.hk).

Joseph Chi Ho So graduated from The University of Hong Kong with BEng (EEE). He then obtained MSc with Distinction from The Hong Kong Polytechnic University and a PhD in Information Engineering from The Chinese University of Hong Kong. Dr So had been working in at Pacific Century CyberWorks (PCCW) and a multinational internet infrastructure service provider. He currently is a principal lecturer and the Head of CPCE Student Affairs Office (CSAO) in The Hong Kong Polytechnic University. He has published over 20 articles in various journals and international conferences. Dr So has been the project leader of three projects sponsored by the Education Bureau. He also currently acted as a reviewer in ELSEVIER and IEEE journals and other international conferences. He was recognized as Peer Reviewer Extraordinaire from MERLOT in 2019 to 2022. He is currently a Senior Member in IEEE, a Chartered Engineer and a member of The Hong Kong Institution of Engineers (HKIE), the Institution of Engineering and Technology (IET), and a Certified Financial Technologist (E-mail: sochhome@yahoo.com).

Kia Ho Yin Tsang is a Research Assistant of the School of Professional Education and Executive Development, The Hong Kong Polytechnic University. He obtained his Bachelor of Science (Honors) in Applied Sciences (Information Systems and Web Technologies) from the same institution in 2021 (E-mail: kia.tsang@speed-polyu.edu.hk).

Ran Wei received the Doctorate Degree in Education from the Department of Science and Environmental Studies, Faculty of Arts and Sciences, the Education University of Hong Kong. Now, she is working as a research fellow at The Hang Seng University of Hong Kong. Ran Wei's current research mainly focuses on education, environmental education, the development of the green school project in mainland China, and students' and teachers' environmental literacy in Hong Kong and mainland China (E-mail: rwei@hsu.edu.hk).