Abu Zohair *International Journal of Educational Technology in Higher Education* (2019) 16:27
https://doi.org/10.1186/s41239-019-0160-3

International Journal of Educational Technology in Higher Education

RESEARCH ARTICLE Open Access

Prediction of Student's performance by modelling small dataset size

Lubna Mahmoud Abu Zohair

Correspondence:
Department of Engineering and IT,
The British University in Dubai,
Dubai, United Arab Emirates

Abstract

Prediction of student's performance became an urgent desire in most of educational entities and institutes. That is essential in order to help at-risk students and assure their retention, providing the excellent learning resources and experience, and improving the university's ranking and reputation. However, that might be difficult to be achieved for startup to mid-sized universities, especially those which are specialized in graduate and post graduate programs, and have small students' records for analysis. So, the main aim of this project is to prove the possibility of training and modeling a small dataset size and the feasibility of creating a prediction model with credible accuracy rate. This research explores as well the possibility of identifying the key indicators in the small dataset, which will be utilized in creating the prediction model, using visualization and clustering algorithms. Best indicators were fed into multiple machine learning algorithms to evaluate them for the most accurate model. Among the selected algorithms, the results proved the ability of clustering algorithm in identifying key indicators in small datasets. The main outcomes of this study have proved the efficiency of support vector machine and learning discriminant analysis algorithms in training small dataset size and in producing an acceptable classification's accuracy and reliability test rates.

Keywords: Classification algorithms, Machine learning, Learning analytics, Visualization, Small dataset

Introduction

Extensive efforts have been made in order to predict student performance for different aims, like: detecting at risk students, assurance of student retention, course and resource allocations, and many others. This research aims to predict student performance to engage distinct students in researches and innovative projects that could improve universities reputation and ranking nationally and internationally. However, analyzing students records for startup to medium size institutes or schools, like the XYZ University in Dubai which have small size of students records, have never been explored in educational or learning analytics domain. Yet, that were investigated in other fields, like: health sciences and Chemists (Ingrassia & Morlini, 2005; Pasini, 2015). So, this project aims to explore the utilization possibility of small students' dataset size in educational domains.

Additionally, in most researches that were aimed to classify or predict, researchers used to spend much efforts just to extract the important indicators that could be more useful in constructing reasonable accurate predictive models. They will either use

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

features ranking algorithms or will look at the selected features while training the dataset on different machine learning algorithms, like in (Comendador, Rabago, & Tanguilig, 2016; Mueen, Zafar, & Manzoor, 2016). Instead, and until recently, there have been no research efforts to investigate the ability of visualization or clustering techniques in identifying such indicators for small dataset, especially in the learning analytics domain (Asif, Merceron, Ali, & Haider, 2017). If such studies will be conducted, its outcomes might prove the feasibility of mitigating the hassle that is normally spent on features extraction or selection processes.

So, this research aims to narrow the aforementioned gaps by solving the following research questions:

- What is the best machine learning classification model for classifying student's dissertation project grade, using small dataset size, with a reasonable and significant accuracy rate?
  - What are the main key indicators that could help in creating the classification model for predicting students' dissertation project grades?
- Could students' performance in any course (excluding the Dissertation) be predicted with a reasonable and significant accuracy rate using only students' pre-admission records, course names, and instructors' name attributes?

The overall study is explained in four sections, including this introduction. The following section will talk about the used methodology. And the third section will demonstrate the analysis results. Finally, and in the last section, results will be interpreted and discussed, and the research will be concluded.

Research methodology

To achieve the project's aims, quantitative simulation research methods were conducted as suggested in the framework phases shown in Fig. 1. In these phases the dataset will be prepared to be passed through visualization and clustering techniques, i.e. like heat map and hierarchical clustering, to extract the top correlated indicators. Then, the indicators will be used in different classification algorithms and the most accurate model will be the chosen for predicting student performance in dissertation projects and all courses grades. In between, and before the classification models' evaluation phase, the datasets will pass through a pre-processing (cleansing, missing data

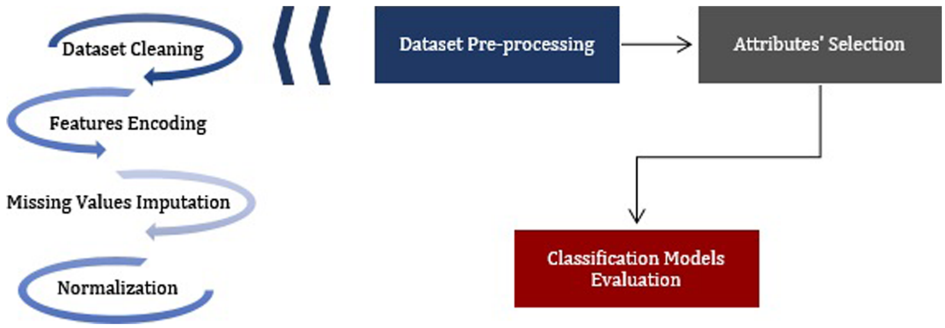

Fig. 1 Dataset analysis framework

imputation, …) stage to make it ready for the analysis phase. That will be more detailed in the following sections.

Participants and datasets

In this study, the records of fifty graduated students in one master's program were collected from the administration department. These records include students' ID, age, bachelor degree name, bachelor degree accumulated grade, courses taken during their master's study with their grades and instructors name of each course. Table 1 shows the list of the main used attributes, their datatypes, and other related details. From that records, 2 datasets were created to answer the research questions and Table 2 illustrates the descriptive statistics of that sets. These records were provided after to comply with the university's data privacy obligations requirements and the replacement of students' IDs and instructors' names with other unique identifiers.

Tools

To utilize from the provisioned dataset, multiple modifications have been created to prepare the dataset for analysis. Microsoft Excel and Python Integrated Development Environment version 3.6.2 were used for that. Additionally, R studio (version 1.1.456) was used to visualize the dataset and select the key attributes. Besides, it has been used for training the dataset with different classification algorithms and evaluate them in order to select the most accurate machine learning classification algorithm.

Data Analysis & Procedures

As illustrated in Fig. 1, three main phases have been followed to answer the research questions. The following sections will explain these phases in more details.

Dataset pre-processing phase

Initially, the datasets contained valueless attributes, missing instances, inadequate attributes' data types and other problems that raise the necessity of preparing it first before feeding it to the analysis phase. Therefore, the datasets were passed through the following preparation stages:

Dataset cleaning Firstly, irrelevant attributes to this study (like: Model code, assessment status, Status, Course description, Academic Year, and Bachelor institution) were eradicated. After that, students with incomplete records, like those who had no grades'

Table 1 Main dataset attributes

Students Attributes	Data Type	Attributes' Details
ID	Ordinal	1,2,3 4, …., 52 (total were 50 students)
Age	Ratio	Between 30 and 52
B.Sc. degree	Nominal	Computer Science, Information system, …
B.Sc. grade	Ratio	3.25, 3.62, …
Course names	Nominal	Introduction to AI, Knowledge management, ….
Course Grades	Ordinal	A, B, C, & F
Instructor Names	Nominal	Instructor 1, instructor 2, …

Table 2 Summery statistics for dataset 1 and 2

Descriptive Statistics	Dataset1	Dataset2
Number of instances	273	38
Dependent Variables (DV)	Grades (All Courses Grades)	Grade (Dissertation Grade)
DV Mean	3.39	3.44
DV Median	4	4
DV Mode	4	4
Accuracy Baseline	P (4) = (136/234) * 100 = **58.1%**	P (4) = (23/38) * 100 = **60.5%**

details in most of their courses or those who didn't have any course records were excluded from the list. Up to that stage, the remaining number of students and their attributes were thirty-eight and seven, in respectively, as illustrated in Table 1. Last, since it's been noted that the number of the courses were decreased since 2010 from nine to seven courses, and to treat all students equally in the analysis phases, the number of courses for all students were decreased from nine to seven by removing the retired courses.

Features encoding In this stage the datatypes of all attributes have been changed to numeric attributes for many reasons. First, some machine learning algorithms, which have proved to be efficient in dealing with small datasets size, such as Linear Discriminant Analysis(LDA) (Sharma & Paliwal, 2015) and Multiple Perceptron Artificial Neural Network (MLP-NN) (Ingrassia & Morlini, 2005; Pasini, 2015) algorithms, requires numeric types of attributes. And the Support Vector Machine algorithm, which was used as well, was designed to work efficiently with numerical attributes. Also, as a best practice in dealing with MLP-NN, in general, attributes have to be in numeric form and be normalized to achieve best classification results. By normalization, attributes' values will be changed and normalized into ranges (either [0,1] or (Mueen et al., 2016)) before feeding them into the classification models. Lastly, since R studio was used for training the classification algorithms, and it executes its operations in RAM, dealing with categorical variables or strings will require more space, runtime, and more processing overhead (since characters are converted to combinations of bytes, especially while dealing with long course names) compared with numerical datatype attributes. This effect on processing performance might not be observed while dealing with the small sample, however, its' always important to comply with best practices to achieve successful analysis results. So, the attributes' conversion to numeric type was done using "ifelse" function in Excel and the following attributes were encoded: B.Sc. Degree, Course Grades and Names, and Instructors' names. The corresponding numbers of each encoded attribute are shown in Tables 3, 4, 5 and 6.

In order to answer research question 1, new arrangements and changes have been made to the dataset, and new attributes have been added. Figures 2 and 3 shows the newly populated datasets with the final arrangements. That arrangements have been programmed to be done automatically using python, and Fig. 4 shows the screenshot of the executed code. So, that new datasets will be used to answer the research questions.

Missing value imputation Visually, and using Amelia library in R, missing values were identified using the missmap function. This function outputs a heat map that marks

Table 3 Encoding course names

Module Description	Courses
Informatics Research Methods	0
Knowledge Representation	1
Learning from Data	2
Introduction to AI	3
Knowledge Management	4
Web Design Project	5
Applied Databases	6
Knowledge Engineering	7
Data Mining and Exploration	8
Introduction to Computational Linguistics	9
Speech Processing	10
Dissertation	11

missing values with different colors. So, both datasets were fed to that function to visually identify the missing values. In the case of Additional file 3, 'dissertation instructor' attributes' were missing most of its values; thus, the variable was deleted. Instead, for the remaining missing values in dissertation grade attribute (i.e. Grade) and course1 grade (i.e. Grade1), they were replaced with the mode value of both attributes. The corresponding code in R is attached in the Additional file 1. Besides, for Additional file 2, the mode of grades attribute's values (i.e. Grades) was the replacement. Compared to mean, median, or regression imputation, and other imputation methods, imputing using mode value will:

- preserve the new encoded numeric (ordinal) attribute datatype from being changed to continuous ones.
- Avoid producing values that will not belong to any of the Grades attribute's classes.

Normalization Normalization is considered one of the recommended pre-processing practices that shall precede training the dataset to some kinds of classification or prediction algorithms, i.e. like the neural network machine learning algorithm. That algorithm recommends making the instances values within specific ranges, either [0,1] or [− 1,1], since scaling to these ranges tend to give better results (Rotich, Backman, Linnanen, & Daniil, 2014). In this project, MinMaxScaler (which scales instances to this range [0,1]) was used as the normalization method and calculated in R using the following equation (assuming the range is [a,b]:

Table 4 Encoding grades

Grade	Grades (new variable name)
A	4
B	3
C	2
Fail	1

Table 5 Encoding instructor names

Instructor Name	Instructors (new variable name)
Instructor 2	2
Instructor 3	3
Instructor 4	4

$$\mathrm{Y{-}Normalized} < -([\mathrm{b}-\mathrm{a}]*(\mathrm{X}-\min(\mathrm{X}))/(\max(\mathrm{X})-\min(\mathrm{X}))) + \mathrm{a}$$

Attributes selection phase

After the pre-processing phase, features selection process was started. The heat map visualization and hierarchical clustering methods were used to help in visualizing the relations between variables and in identifying the main indicators that could help in predicting dissertation and courses' grades. In a nutshell, the heat map is a simple and organized way to display a colorful matrix of data, where its columns represent the dataset attributes and the rows are their corresponding values. The R code for to the used visualization and clustering methods are attached in the Additional file 1. Also, the key indicators - which were identified visually- were compared to those which were selected by the classification algorithms while training the datasets. This comparison is needed to confirm if the key attributes were visually identified correctly, especially, in case if the relationships between attributes cannot be clearly identified.

Classification model evaluation

Multiple machine learning classification algorithms were used to train the datasets, including: MLP-ANN, Naïve Bayes(NB), Support Vector Machines(SVM), K Nearest Neighbor (KNN), and LDA. The idea was to evaluate which one will be better in terms of the ability to produce reasonable accurate prediction rate of students' performance for small size datasets. MLP-ANN and LDA were chosen because some researchers discovered their efficiency dealing with small dataset size and in producing more accurate results, especially, in the fields of face and speech recognition and financial market forecasting (Mustafa, Allen, & Appiah, 2017; Pasini, 2015; Sharma & Paliwal, 2015). MLP is a type of artificial neural network that allows the processing of multiple inputs to produce multi-label output. It accepts nominal or numerical attributes and it can be used as a classification or regression algorithm. Nonetheless, LDA is a dimensionality reduction algorithm that tries to create a linear relationship between different classes, while minimizing the scatter of each class and maximizing the distance between the labels centroids and the central point of all of them (Qiao, Zhou, & Huang, 2009). It predicts the class of a variable using two or multi numeric attributes. On the other hand, NB algorithm computes the probability that a certain class label will appear given that a certain condition has already been occurred. This classifier was fundamentally designed to accept categorical attributes, but also it could support normally distributed numerical inputs. It is an advantageous method since it can utilize from a small size training set to create the classification model (Dey, Chakraborty, Biswas, Bose, & Tiwari, 2016). Also, it is equipped with a kernel density estimator that can handle non-parametric variables. As for KNN, it can be used for classification or prediction problems, where by knowing K value (number of instances) and utilizing numerical

Table 6 Encoding Bachelor Degree Specialization. Noting that specialization that are almost similar to each other were grouped in one category

BSc Degree	BSc Deg. (new variable name)
Mathematics	0
BSc Computing1	1
Information Technology	1
Operations & Information Management	2
Management Information System	2
Business Information Technology	2
Electronic Engineering	3
Electrical Engineering	3
Engineering	3
Engineering	3
Electrical & Electronics Engineering	3
Electrical Engineering	3
Electrical Engineering - Computers & Control Section	3
Electronics Engineering (Computer & Control)	3
Computer Information Systems	4
Information System	4
Computer Information Systems - Circuits and Systems	4
Computer Science/Information System	4
Computer Engineering	5
Computer Engineering	5
Computer Engineering	5
Computer Engineering	5
Computer Engineering	5
Computer Science	6
Computer Science Mathematical Statistics	6
Computer Science	6
Computer Science	6
Computer Science	6
Computer Science	6
Computer Science	6
Computer Science	6
Computer Science	6
Computer Science	6
Computer Systems	7
Computer Systems	7
Software Engineering	8
Software Engineering	8
Software Engineering	8

variables the algorithm can predict the class labels based on the most occurring labels in the k nearest ones. Likewise, SVM works for classification and prediction problems, and the idea behind it is to find a line that best isolates multi group labels. It is

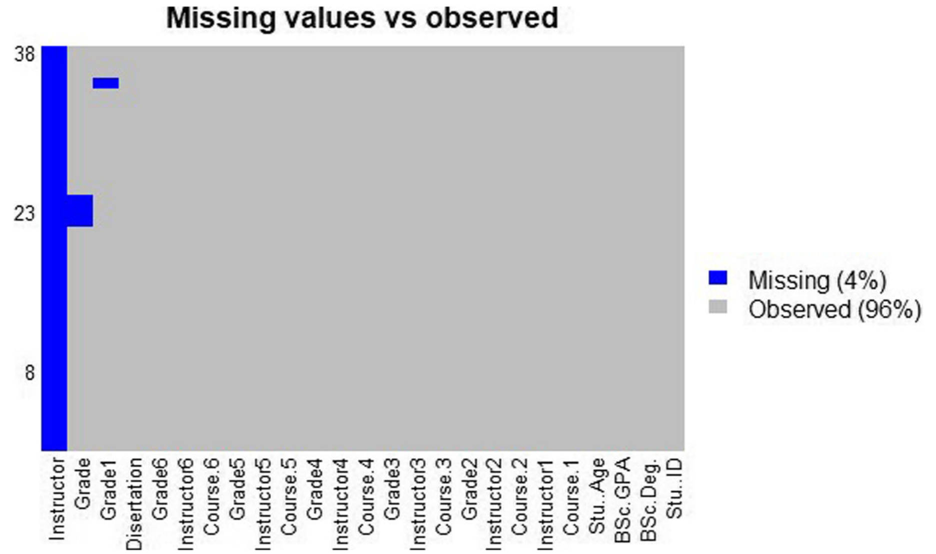

Fig. 2 The dataset1 attribute structure after the feature encoding process

developed to deal with numeric attributes, as it deals with nominal ones after converting them to numeric datatypes.

Abstractly, the aforementioned explanation about the selected machine learning algorithms described why they were selected and, most importantly, helped in knowing the attributes' types that shall be used in each algorithm to allow it to perform efficiently. However, since no machine learning algorithm is considered good in all use case scenarios (like in training small sample size or accurately predicting students' performance (as what literature in (Asif et al., 2017) suggests)), this research will examine all the aforementioned algorithms and will evaluate them in terms of their classification accuracy rates to end up selecting the most accurate algorithm to create students' grades' classification model.

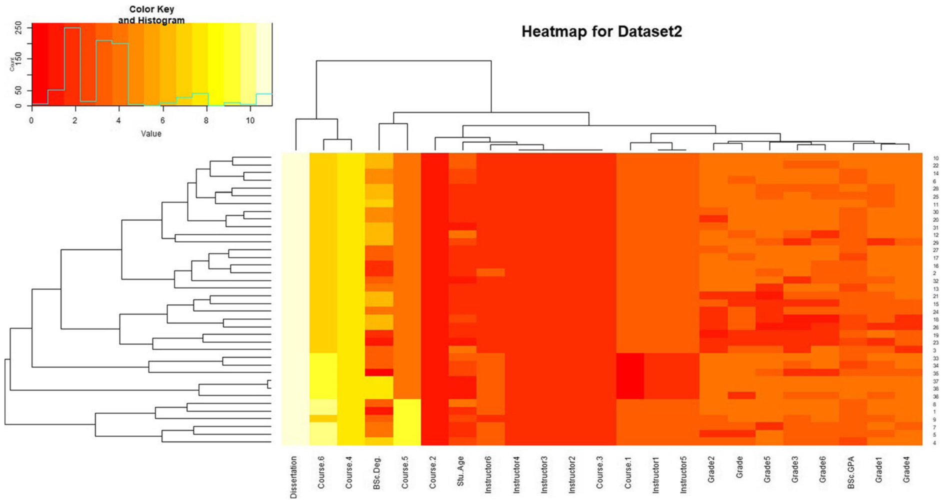

Fig. 3 The dataset2 after the new attributes creation and arrangements

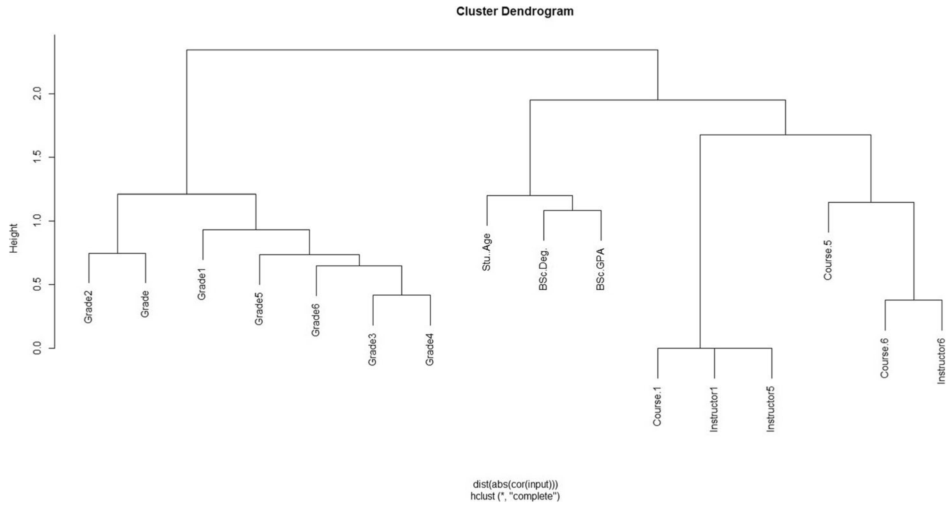

Fig. 4 Python code for the new arrangement for the dataset, required for research question 1 needed analysis

The used evaluation metrics for the best performed classifiers are: the accuracy (the right predictions subdivided by the total predictions) and Cohen's kappa (which is more reliable accuracy metric). Notably, since the datasets are small, Leave-One-Out Cross Validation (LOOCV) technique is used as a validation method since it's considered as the most preferable and advisable validation method for small size sets (Rao, Fung, & Rosales, 2008). Instead of segmenting the dataset into training and testing sets, the efficiency of LOOCV lies in its ability to utilize from all the dataset instances (except one) to train the machine learning models. Besides, this process iterates to test one data point in each iteration, and the average accuracy of all tested points will be the output accuracy rate of each classification model. As a baseline from which the reasonableness of the evaluation accuracy results will be compared with (i.e. the point that should be improved), the probability of the occurrence of the grade value (the mode value, i.e. the most occurred grade) will be used and will be measured using this equation:

$$\text{The probability of Grade``x''occurrence} = (\text{Number of Grade``x''Instances}/\text{Total Grades Instances})*100\%$$

That method is called zeroR classification, and it's a function in Weka tool, which calculates the probability for attribute's values occurrence (Litman & Forbes-Riley, 2004). Also, Cohen's Kappa (K) will measure the rate of models' accuracy in comparison with the accuracy of the random occurrence of attributes values. The kappa baseline starts from zero, which means that the algorithm produces an accuracy rate which is similar to the accuracy of the stochastic prediction. This algorithm considered an efficient and reliable evaluation metric for nominal attributes, also, in dealing with imbalanced (non-parametric) dataset attributes (or if there'll be a skewness in class frequency distribution) (Kuhn, 2008; Mchugh, 2012). Last, the overall accuracy p-value will be used to examine how reasonable or significance are the classifiers' accuracy in predicting the class of interest in relative to the baseline, i.e. no information rate (NIR). The applied alpha is 0.05 and the null hypothesis will be rejected if $p < 0.05$. So, the proposed hypotheses are:

– Null hypothesis (H0): there is no difference between the accuracy predicted by classification algorithms and NIR (accuracy of the random prediction).
– Alternative hypothesis (H1): there is difference between the accuracy predicted by classification algorithms and NIR (accuracy of random prediction).

Results

Datasets summary statistics

Since the accuracy is the main key metric that the evaluation of machine learning models will be relying on, the baseline accuracy is calculated at first (also called 'no information' rate) for both datasets. The calculated baseline and the results obtained from the preliminary descriptive analysis of the datasets of interest are shown in Table 2 and the related code is shown in the Additional file 1. After that, missing values were identified and found (mainly) in Additional file 3, as shown in Fig. 5. Some missing values were treated by eliminating the attribute (like: dissertation instructor name), and the other missing values were replaced with the mode of the corresponding attributes.

Key attributes

To achieve the first aim of this research, Additional file 3 was assessed for its key indicators using the heatmap.2 function, which is imported from gplots Library in R. In that function, the attributes in that dataset were grouped according to their similarity with the help of agglomerative traditional hierarchical clustering algorithm that is embedded within the heatmap function. In other words, since clustering is performed for rows and columns, then, the attributes and values that are similar to each other were grouped close to each other in one cluster. So, after observing Additional file 3 and its relative heat map and the column's dendrogram figures, i.e. Figure 6 and Fig. 7 (in respectively), the top five features that were found close to the dissertation grade attribute (i.e. Grade) were: Grade2 (Grade for Course 2), Grade 1 (Grade for Course 1), Grade 5 (Grade for Course 5), Grade 6 (Grade for Course 6), and Grade 3 (Grade for Course 3).

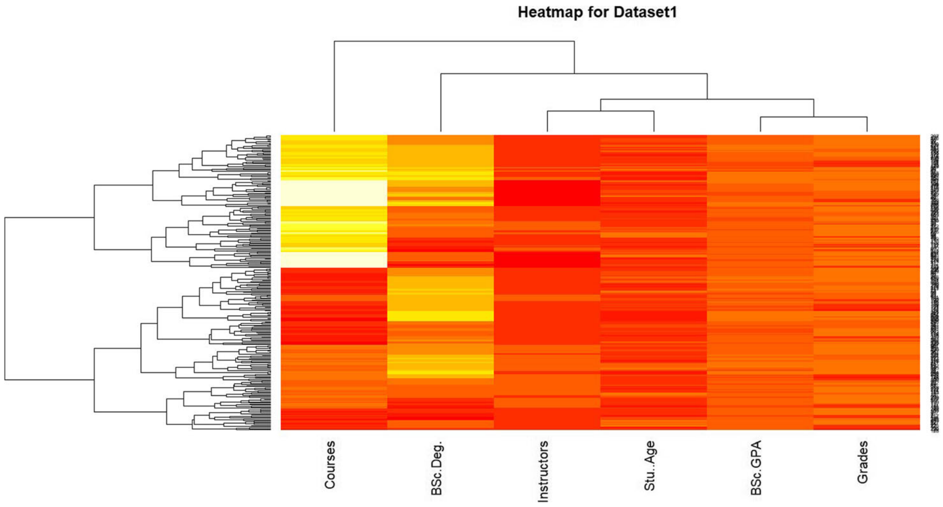

Fig. 5 Dataset 2 attributes and missing instances of each

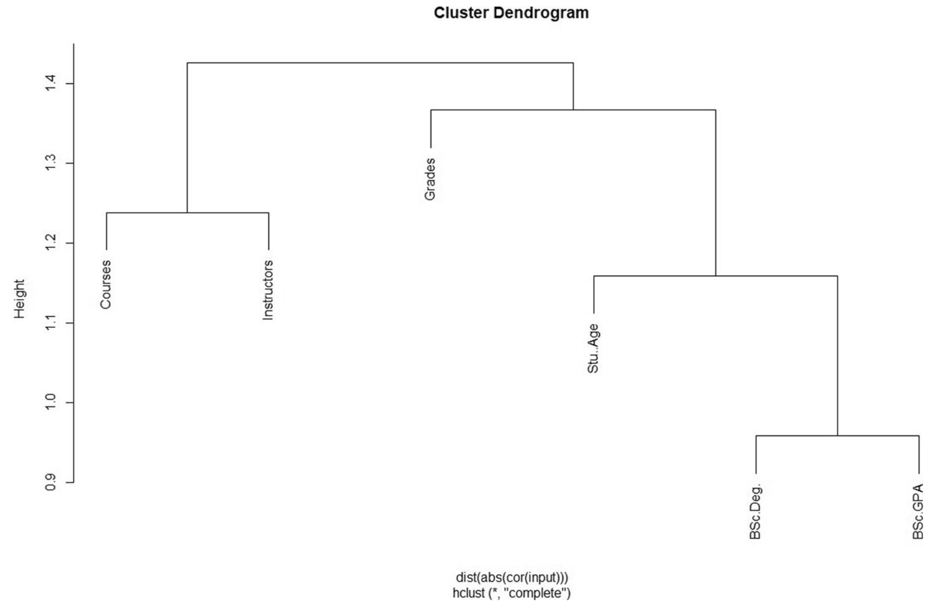

Fig. 6 Visualizing Dataset2 using Heat map

These attributes are considered the main key indicators for predicting student grade in dissertation course, as they all have correlations that allowed them to be in one cluster at dendrogram height 1.5. And that answered the sub-question of the first research question, and proved the efficiency of visualization and clustering in identifying the key attributes. Providing that the success factors of the visualizations analysis lies in the scaling of large values attribute, i.e. Student Age, into a range that is commonly used in other attributes, which is (Mueen et al., 2016; Sharma & Paliwal, 2015), using this equation:

$$(4-1)*(\text{inputStu..Age}-\min(\text{inputStu..Age})))/(\max(\text{inputStu..Age})-\min(\text{inputStu..Age}))+(1)$$

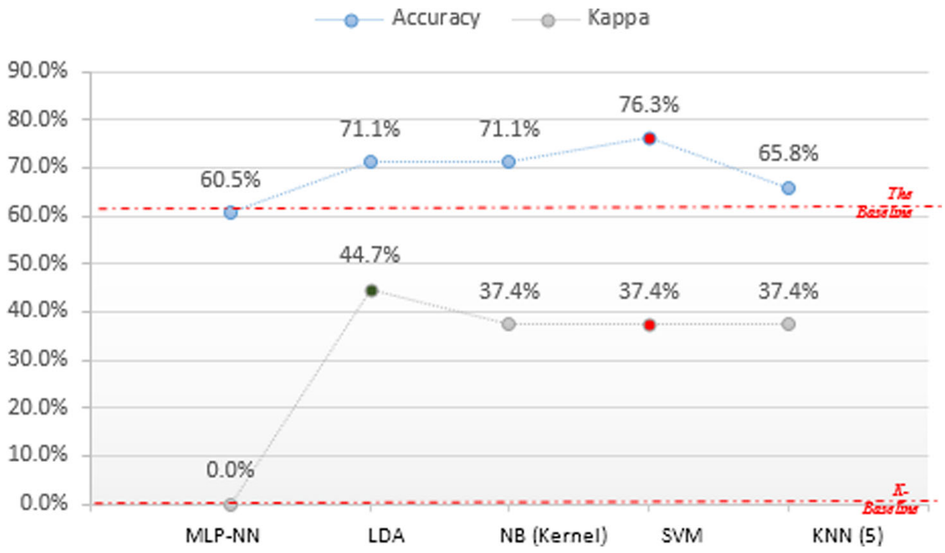

Fig. 7 Visualizing Dataset 2 using Traditional Dendrogram

In addition, before passing the dataset to dendrogram visualization function, the attributes that had zero or very low standard deviation or variance were nullified to avoid invalid correlations and output errors.

The same aforementioned visualizations techniques were repeated to identifying the best indicators in Additional file 2, to help in answering research question 2. So, as a result of its visualization, Fig. 8 and Fig. 9 illustrate the features which were correlated with students' Grades in all courses (i.e. Grades), and they are: students' Age (Stu.Age), bachelor GPA (BSc.GPA) and specialization (BSc.Deg). However, the visualization of their relations barely appeared in the heat map, but were clearly forming one dendrogram cluster at 1.4 height. Therefore, and as a partial answer to the second question, the clustering analysis was obviously showing that pre-admission attributes (i.e. students' age, bachelor degree and GPA) were having significant impacts on student grades compared to other attributes. Another thing, heatmap visualization was perfect in showing the dominant grade label in both datasets, as the color that represents grade 'A' in grade 4 attribute was the widely spread one, but for grade 1, grade 'Fail' was the dominant grade.

Evaluation of classification models

The extracted key indicators, which was extracted from the visualization analysis, were fed in the five chosen classification algorithms. But, it's worth mentioning that since the chosen classification algorithms have the capability to train two different attributes' types, i.e. nominal and numeric, both were tried and trained. Then, the accuracy results were evaluated to see which variable type can work efficiently with each classification algorithm in training the datasets of interest. As a result, Fig. 10 and Fig. 11 clearly show that SVM model (with radial kernel) reported the highest accuracy rate in predicting students grades in all courses (i.e. Grades attribute) and dissertation project (i.e. Grade attribute). Noting that the underlined (U) x-axis names of algorithms were those

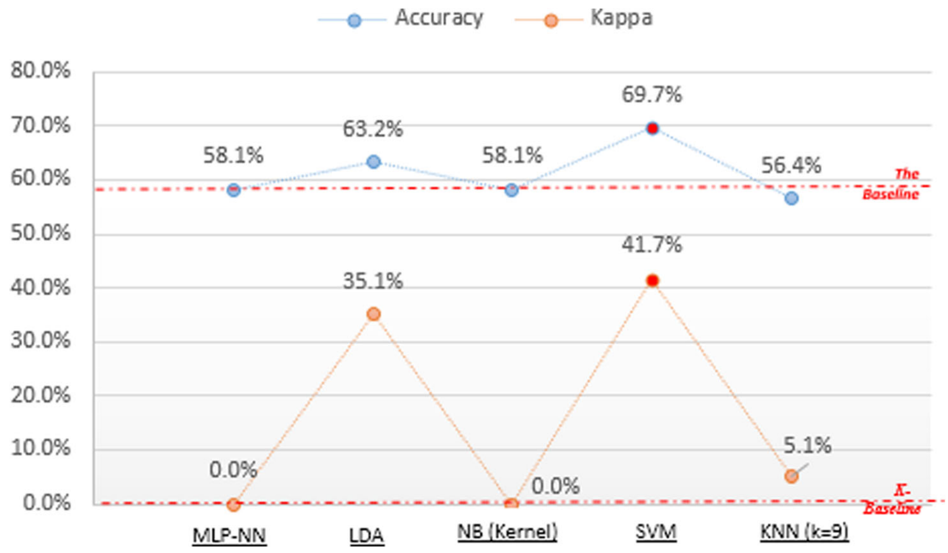

Fig. 8 Visualizing Dataset1 using Heat map

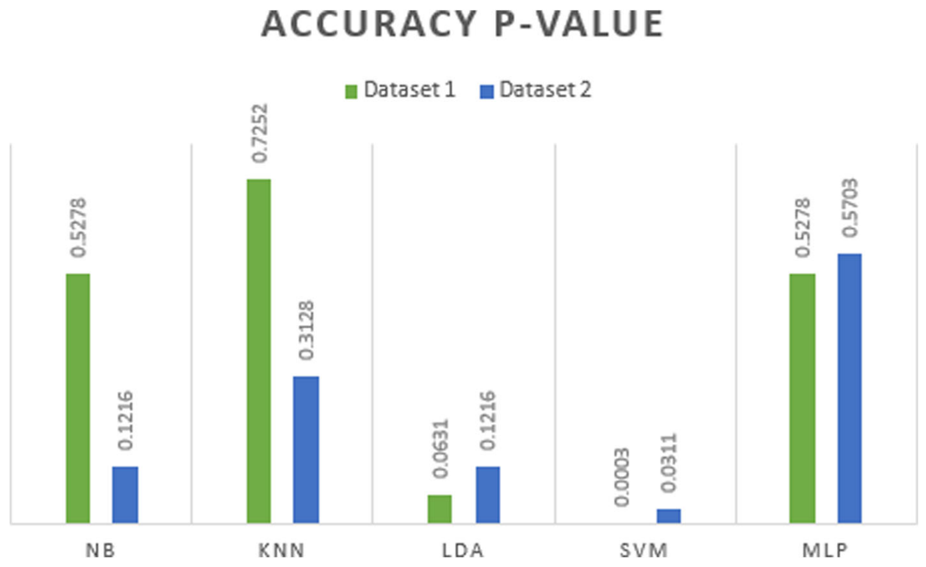

Fig. 9 Visualizing Dataset1 using Traditional Dendrogram

that worked efficiently with nominal attributes. The predictions rates, in comparison with the baseline, are 76.3% and 69.7% for dissertation grade and all courses grade class, in respectively. SVM with radial kernel function was chosen since it has the ability to train and deal with imbalanced datasets. Additionally, kappa results showed that LDA's accuracy in predicting student's dissertation grade is 44.7% and that considered better than predicting the same class labels randomly. However, for all course classification (i.e. Grades attribute), SVM's kappa was the highest and its is better than the baseline recording 41.7% accuracy rate. All the aforementioned related results and the comparison between different attributes types are placed in Tables 7, 8, and 9.

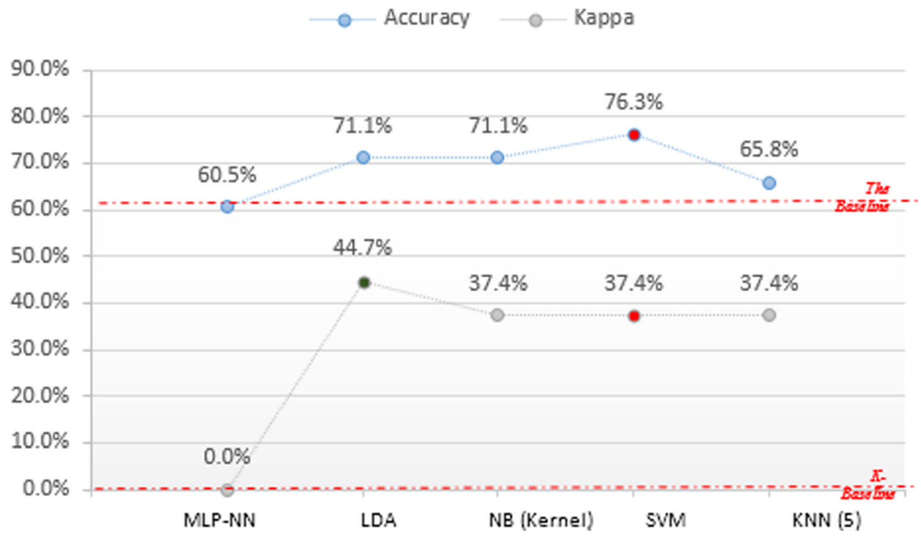

Fig. 10 Accuracy and Kappa evaluation for different machine learning models while training the dataset2 for predicting students' grade in dissertation course

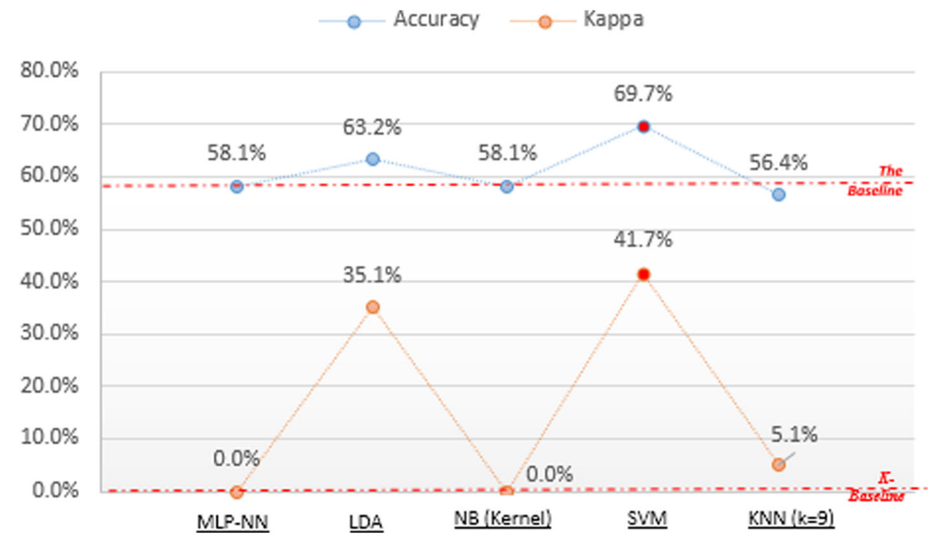

Fig. 11 Accuracy and Kappa evaluation for different machine learning models while training the dataset1 for predicting students' grades in all courses

Now, to evaluate the significance or the credibility of the achieved accuracy rates in contrast with the random prediction ones, the *p*-values were extracted from the confusion matrix function of all trained machine learning algorithms for Additional files 2 and 3. The outcomes, as presented in Fig. 12, indicate the significance of SVM's accuracy results in accurately classifying students grades in all courses and the dissertations one because the recorded p-value for successfully classifying them were 0.0003 and 0.03, respectively. So, since the p-values were less than 0.05, the proposed null hypothesis has been rejected. To rephrase, the accuracy results achieved by the SVM Radial classifiers in both datasets exceeded the baseline and were accepted as significance accuracy rate, making SVM kernel model a perfect classification model among other tested algorithms. Thus, that answers the remaining parts (about the significance of the accuracy rate) of research questions 1 and 2.

Discussion & Conclusion

Predicting students' performance for post graduate study is important for any educational institutions. It is important especially, for those who are aiming to give students

Table 7 Accuracy and Kappa results for nominal & numeric attributes for dataset 2

Dataset2	Numeric Attributes		Nominal Attributes		*Notes*
Classifications Algorithms	*Accuracy*	*Kappa*	*Accuracy*	*Kappa*	
MLP - ANN	60.5%	0.0%	60.5%	0.0%	With three hidden layers, i.e. 3,2,1, however, other values were used aswell, but the accuracy results maintained the same
LDA	71.1%	44.7%	57.9%	19.3%	
NB	65.8%	32.1%	71.1%	37.4%	using Kernel
SVM	68.4%	33.4%	76.3%	49.3%	were sigma = 0.1590384 and C = 1, for nominal The final values used for the model were sigma = 0.0564085 and C = 1
KNN	65.8%	28.8%	65.8%	31.9%	At K = 7, for nominal at k = 5

Table 8 Accuracy and Kappa results for nominal & numeric attributes for dataset 1

Dataset1	Numeric Attributes		Nominal Attributes		*Notes*
Classifications Algorithms	*Accuracy*	*Kappa*	*Accuracy*	*Kappa*	
MLP - ANN	58.1%	0.0%	58.1%	0.0%	
LDA	56.4%	−1.0%	63.2%	35.1%	
NB	57.7%	0.1%	58.1%	0.0%	Using kernel
SVM	58.1%	0.0%	69.7%	41.7%	where c = 1, c = 0.25 ('sigma' was held constant at a value of 0.2410613, Accuracy was used to select the optimal model using the largest value, for numeric. The final values used for the model were sigma = 0.2410613 and C = 0.25.), for nominal
KNN	55.6%	11.4%	56.4%	5.1%	k = 7,k = 9

opportunities in doing something useful in their field of study, and those who are aiming to well manage the needed teaching resources for excellent learning experiences, like XYZ University. XYZ is a start-up research-based institute which aims to improve its reputation and ranking by selecting high performing students to engage them in solving real world issues. So, predicting distinguished students is an urgent desire. Additionally, knowing students' performance in each course beforehand is a main requirement in order to help at risk students by mitigating the challenges that they are facing in their learning journeys and helping them excel in the learning process. Whilst, such predictions, especially, for a new university is a challenge since there are no enough dataset records to be analyzed. Nonetheless, our results prove the possibility of doing so with reasonably significant accuracy rates. The support vector machine classifier with radial kernel was the one which proved its efficiency (among the rest of classifiers) in predicting students' performance in all courses' grades, including their dissertation projects' grade. The main reason that may be attributed to that classifier's success is the model training method that its used, which relies only on a few data points or samples (those which are very close to the hyperplane) to build its classification model. That result did not match the research findings in (Mustafa et al., 2017; Pasini, 2015; Sharma & Paliwal, 2015) which proves the efficiency of LDA and MLP-NN in treating small dataset sizes. But it agrees with (Asif et al., 2017) that there is no perfect classifiers that can work efficiently for similar dataset characteristics in different use case scenarios.

Moreover, since the attributes values in each class (for both datasets) were imbalanced, and for generalizability purpose (i.e. to measure the accuracy while avoiding the bias that may be created by that imbalanced data while training the dataset), another

Table 9 Dataset 1 and 2 with latest chosen attributes type, that records the highest accuracy and kappa results

Dataset2				Dataset1			
Classifications Algorithms	*Accuracy*	*Kappa*	*Attribute Type*	Classifications Algorithms	*Accuracy*	*Kappa*	*Attribute Type*
MLP - ANN	60.5%	0.0%	Numeric	MLP - ANN	58.1%	0.0%	Nominal
LDA	71.1%	44.7%	Numeric	LDA	63.2%	35.1%	Nominal
NB	71.1%	37.4%	Nominal	NB	58.1%	0.0%	Nominal
SVM	76.3%	37.4%	Nominal	SVM	69.7%	41.7%	Nominal
KNN	65.8%	37.4%	Nominal	KNN	56.4%	5.1%	Nominal

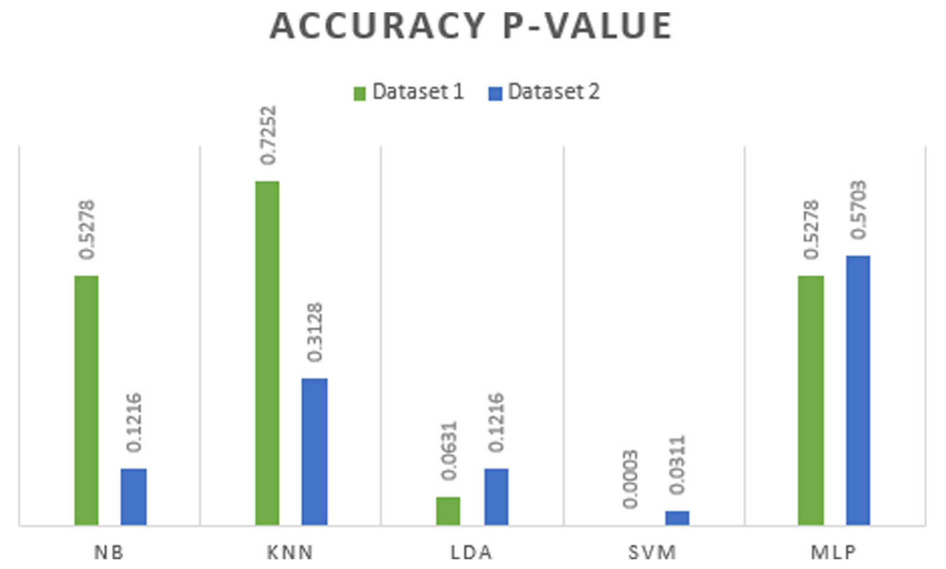

Fig. 12 Accuracy (ACC > NIR) P-Value results for the used classification algorithms

performance measure was used and it's called balanced accuracy. Balanced accuracy is an evaluation metric that takes the average of sensitivity (or recall) and specificity to calculate the accuracy rate of a certain attribute's class (Brodersen, Ong, Stephan, & Buhmann, 2010). So, the balanced accuracy rate was extracted from the confusion matrix of the classification results of all tested classifiers for only class A (4) of grades attribute. Then, they were compared with the calculated accuracy baseline, and the result is shown in Fig. 13. LDA has recorded the highest accuracy rates with values 79% and 77% for the classification of class A for Additional files 2 and 3, in respectively. Noting that, although SVM produced acceptable accuracy results, it is still susceptible more than LDA to be biased with imbalanced dataset observations while training the model. Despite that, both have proved to be reliable since their kappa results not only

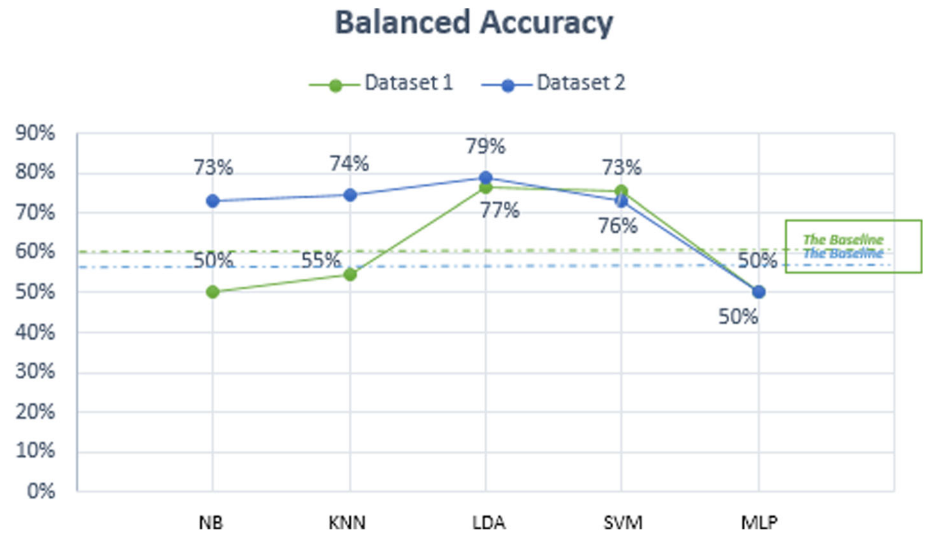

Fig. 13 Class A Balanced Accuracy Rates for Dataset 1 and 2

exceeded their baselines but also they were fluctuating between fair and moderate agreement levels, as suggested by Landis, Kock, and Fleiss's kappa guidelines (Byrt, Bishop, & Carlin, 1990; Fleiss & Paik, 2003). Kappa and balanced accuracy considered reliable metrics, but examining the efficiency of LDA and SVM Radial models with more evaluation and diagnostic metrics, like ROC Curve or F-Scores, will be an interesting topic for further investigation in the future.

Overall, students' grades in most courses were correlated with students' grade is dissertation course (i.e. grade attribute in Additional file 3), however, students grade in course 1 and course 2 (i.e. grade1 and grade 2 attributes) were considered the highest key indicators for predicting students' performance in dissertation project (i.e. in predicting grade attribute). Moreover, the classification accuracy rates of LDA and SVM algorithms were the highest, and considered highly significant in comparison with the stated baseline (i.e. the baselines which were specified in Table 2: 58.1 and 60.5). Another key findings, and unlike predicting grades in Additional file 2, the indication of the key attributes in classifying students grade in dissertation project (i.e. grade attribute) using visualization and clustering techniques were found much easier, less complicated, and more timely efficient for small datasets in comparison with the methods followed in the following reviewed literatures (Asif et al., 2017; Comendador et al., 2016; Mueen et al., 2016) (where they spent time and efforts in running multiple classifiers and extracting key attributes based on their coefficients or by extracting those key indicators that were selected in building the classification/prediction models). Besides, LDA and SVM algorithms proved the importance of students' pre-admission records (like: student age) in predicting their performance in all their courses' grades (attribute grads in Additional file 2). That outcome was identified by evaluating the calculated accuracy rate of LDA and SVM classification algorithms that were significant and exceeded the calculated baseline. Despite that extracting the best indicators for predicting grades attribute in Additional file 2 was difficult using the heat map (since the number of the attributes were very few in comparison to the number of instances), the indicators were easily identified using the dendrogram. And that concluded the main findings of this project.

Last important note, this research studied only students' administration records to form the classification models, ignoring by that other variables that could affect students' learning outcomes, like: attendance, instructor course delivery, and many others. That was because the main focus in this project was to explore the feasibility of utilizing from small dataset size in predicting student performance and to shed the light on the importance of visualization and dendrogram in identifying valuable predictors.

Additional files

Additional file 1: R code related to visualization and classification model evaluations. (DOCX 20 kb)

Additional file 2: Dataset 1. (CSV 6 kb)

Additional file 3: Dataset 2. (CSV 2 kb)

Acknowledgements

This is to express my warm thanks to the administration team at the British University in Dubai for providing the datasets of interest and making this project reaching its current successful state.

Authors' contributions

The author read and approved the final manuscript.

Funding

The paper's publication is not funded by any institute or any other entities.

Availability of data and materials

The used datasets of participants are included in the attached additional files named dataset1.csv and dataset2, However, they are not for publication as they are considered sensitive data for the university and its students as well.

Competing interests

The author declares that she/he has no competing interests.

Received: 7 October 2018 Accepted: 27 June 2019
Published online: 02 August 2019

References

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers in Education*, *113*, 177–194.

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *Proceedings - international conference on pattern recognition*.

Byrt, T., Bishop, J., & Carlin (1990). Prevalence adjusted bias. *Journal of Clinical Epidemiology*.

Comendador, B. E. V., Rabago, L. W., & Tanguilig, B. T. (2016). An educational model based on knowledge discovery in databases (KDD) to predict learner's behavior using classification techniques. In *2016 IEEE Int. Conf. Signal process. Commun. Comput*, (pp. 1–6).

Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. (2016). Sentiment analysis of review datasets using Naïve Bayes' and K-NN classifier. *Int. J. Inf. Eng. Electron. Bus.*

Fleiss, J. L., & Paik, M. C. (2003). The measurement of interrater agreement, in statistical methods for rates and proportions. In *Statistical methods for rates and proportions*.

Ingrassia, S., & Morlini, I. (2005). Neural network modeling for small datasets. *Technometrics*.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*.

Litman, D. J., & Forbes-Riley, K. (2004). Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics - ACL '04*.

Mchugh, M. L. (2012). Interrater reliability: The kappa statistic importance of measuring interrater reliability theoretical issues in measurement of rater reliability. *Biochem Med (Zagreb)*.

Mueen, A., Zafar, B., & Manzoor, U. (2016). Modeling and predicting students' academic performance using data mining techniques. *Int. J. Mod. Educ. Comput. Sci.*, *8*(11), 36–42.

Mustafa, M. K., Allen, T., & Appiah, K. (2017). A comparative review of dynamic neural networks and hidden Markov model methods for mobile on-device speech recognition. *Neural Computing and Applications*.

Pasini, A. (2015). Artificial neural networks for small dataset analysis. *Journal of Thoracic Disease*.

Qiao, Z., Zhou, L., & Huang, J. Z. (2009). Sparse linear discriminant analysis with applications to high dimensional low sample size data. *International Journal of Applied Mathematics*.

Rao, R. B., Fung, G., & Rosales, R. (2008). On the dangers of cross-validation. An experimental evaluation. In *Proceedings of the 2008 SIAM international conference on data mining*.

Rotich, N. K., Backman, J., Linnanen, L., & Daniil, P. (2014). Wind resource assessment and forecast planning with neural networks. *Journal of Sustainable Development of Energy, Water and Environment Systems*.

Sharma, A., & Paliwal, K. K. (2015). Linear discriminant analysis for the small sample size problem: An overview. *International Journal of Machine Learning and Cybernetics*.Sharma, A. and Paliwal, K. K., "Linear discriminant analysis for the small sample size problem: An overview," *International Journal of Machine Learning and Cybernetics*, 2015.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com