Cluster and Logistic Regression Distribution of Students' Performance by Classification

Nareena Soomro[1(✉)], Fahad Razaque[2], Safeeullah Soomro[2], Shoaib Shaikh[3], Natesh Kumar[4], Ghulam e Mustafa Abro[3], and Ghulam Abid[3]

[1] Department of Computing, Indus University, Karachi, Sindh, Pakistan
nainee_soom@yahoo.com
[2] College of Computer Studies, AMA International University, Salmabad, Kingdom of Bahrain
fahad.indus1337@gmail.com, s.soomro@amaiu.edu.bh
[3] Hamdard University, Karachi, Sindh, Pakistan
skshaikh@outlook.com, mustafa.abro@hamdard.edu.pk, engrabid1246@gmail.com
[4] Usman Institute of Technology, Karachi, Sindh, Pakistan
nateshsolanki1992@yahoo.com

Abstract. In the research cluster based logistic regression model on student result, performance at computing department, and other demographics to predict whether or not student will annually enroll if admitted that help the campus administrators to manage registrations. In this study, deals with performance and analysis of examination results' performance of students from computing department by also establishing general assessment. However, it cannot be stand-alone and only serves to compliment campus administrator of decision making procedure to manage registrations effectually. Predict students of educational performance are critical for scholastic departments because planned program can be scheduled in maintaining performance of students during their period of studies in departments. The demographic profile of students and fourth year of academic are used as predictor variable for performance of students educational in academic program.

Keywords: Logistic regression · Performance · Cluster · Probability

1 Introduction

Research purpose is data mining method's efficiency apply in education and benefit educational department better use this method to notify student graduation policies. Advance university standing, improved student retention foremost to graduation recovers admission executive and reduce on recruiting costs. From side of student, retention chief to qualification was social, own and financial insinuations. This study's purpose was a variable's prediction that needs influence on educational student's performance was significant as auxiliary programs could be applied to avoid failures. It perceived relationship between the possibility of educational failure and level of

M. H. Miraz et al. (Eds.): iCETiC 2018, LNICST 200, pp. 209–219, 2018.
https://doi.org/10.1007/978-3-319-95450-9_18

knowledge in software engineering and computer science [9]. In this research, logistic regression model to predict computing students' performance in their four years using educational, mental and professional learning and motivational policies as variables. It was examined interactions between student and personal contextual characteristics, educational preparation and performance traits, quantitative education by logistic regression was conducted to scrutinize association between variables and ability to predict student persistence in academic. It should do a better job of summit students' wants and sighted them down to degree achievement. It is significant for the future of students, higher education, and society as a whole.

Educational organizations are progressively interested in intensive care act of their students, which contributes increase to necessity to investigation, collate, scrutinize and interpret data, in order to have proof to notify an educational strategy that was formulated to progress student's performance, excellence training and support resources; producing involvement policies to mitigate factors that will definitely influence student performance.

The core research of department of Computing and Technology is to deliver quality learning to students and to advance quality of decision-making. The approach to accomplish quality level in academic is by learning information from instructive data to study key attributes that might affect performance of students. It could be used to suggestion supportive and positive references to educational organizers in department to improve decision making process, educational performance of students, teaching and reduce failure ratio, to well recognize behavior of students, to assist lecturers, and various other profits [13].

Study on predicts feature causative to students of educational success would be helpful to educational area, communal and others who is concerned with improving performance of students throughout universities times.

Academic Data Mining was one of emerging field which comprise procedure of examined students' details by different elements such as earlier semester marks, attendance, assignment, discussion, lab work were of used to improved bachelor academic performance of students, and overcome difficulties of low ranks of students [14]. It was extracted useful knowledge from academic students data collected from department of Computing. Subsequently preprocessing data, which was applied data mining techniques to discover classification and clustering and outlier detection. In this study, classification method was described which based on K mean algorithm and Cluster based logistic regression model for predicting the students' efficiency of academic.

2 Methodology

Figure 1 demonstrates of data preparation and data pre-processing contain data set that taken student's data from department of computing. The data preparation determination was examined and transformed raw data in order to create them mean more and improved data quality. Without data preparation, hidden information was not easily accessed using data mining models [15]. Data Pre-processing step was executed to develop excellence of data set through removed incomplete values. Data set considered, 61 records were removed from 660 entire data and simply 591 records were prepared for data mining method later. When pre-processing method applied on data set, 91 records were eliminated from data set of 591 records which left only 500 clean records. The total, data of student comprised 160 missing values in numerous parameters from 660 records was ignored from data set. The total numbers of records was reduced to 500 [14].

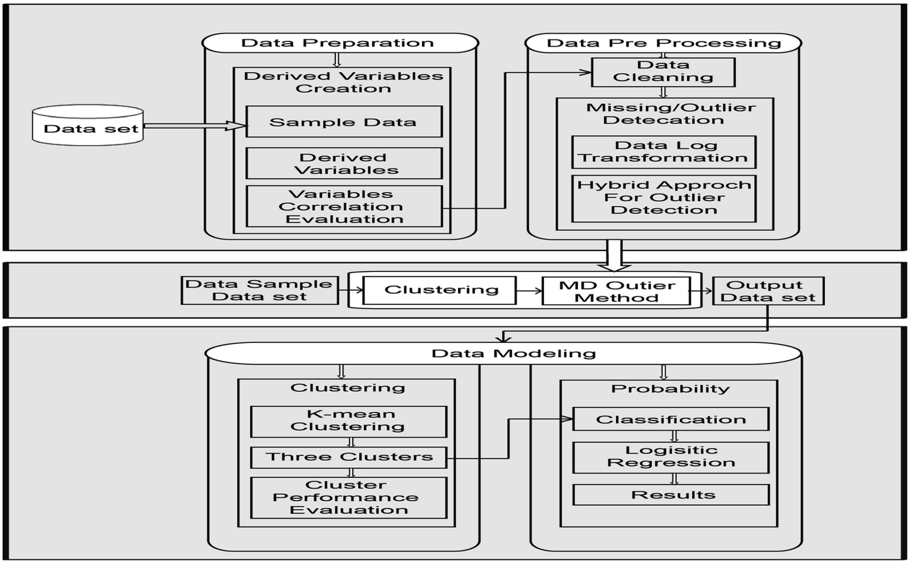

Fig. 1. Research schema of data analysis

It shows that data preparation to expose further concealed info in data by making variables, and to clean raw data set by using the method of hybrid outlier detection. Data set is organized for modeling; adapt clustering and predictive models to execute the analysis. K-means clustering method is modified to divider sessions into three player groups, at same the time, the target variable used in predictive models is also generated. Model of logistic regression is practical to recognize which behavioral pointers are highly related with gambling addiction because of its highest total accuracy.

Clustering was recognized descriptive model, which dividers data into clusters set, such as remarks with parallel appearances were assembled. Hence, the cluster was collected of data items which were parallel to each one but disparate to those facts in further clusters [2]. A good cluster model was hypothetical to ensure that intro cluster resemblance was high, however intercluster resemble ought to be low [3]. Numerous diverse cluster procedures were developed, but furthermost extensively used was k-means algorithm that was also used in the study.

K-means attempted to divider n remarks in a data set into k number of clusters in which every remark belongs to cluster with nearest centroid [2]. Additionally, clustering was frequently significant beginning point to other forms of data modeling [6].

In this study, outlier detection and deletion is most significant task in phase of data preparation as researchers identify numerous irrational items in the data set. The presence of those irrational data points will introduce complexity into data models, and finally reach specious deductions. Because of conduct review and compare different methods in order to discover a most appropriate technique. Outlier referred to those data points that were substantial unrelated in data set. While many outlier detection approaches was proposed, most of them can be classified into four kinds, which are distributed, density, distance and clustering based. Distribution method, for example Standard Deviation is mainly applied to deal with univariate data set [5], but the data set is multivariate with several variables.

MD analysis follows Chi-Square distributions, which have critical values table is used as a means to determine threshold that is decided by significance level (p) and degrees of freedom (df). The Significance level is usually set at 0.05 ($p = 0.05$), which is most commonly used number and has already been accepted as a standard by researchers [12]. Different Chi-Square test, MD is evaluated with degrees of freedom equal to the number of independent variables involved in the calculation ($df = n$) [8]. Though the MD approach has been commonly used, some researchers pointed out that it is not appropriate to deal with outliers in a large data set, since the distance between observation and center of the whole data set needs to be calculated which increase the computation time but decrease the accuracy [10].

Probabilistic classifier that is able to predict, input observation, a probability distribution over classes set, instead of output most likely class that observation has to belong. It delivered classification that could be beneficial in own right [4].

Regression models were typically adapted to discover which predictors were highly linked to variable, and how modifications of predicators affect aim variable. It was most operative when it was used to predict data set taking big quantity of observations but trivial variables number. Moreover, regression models effort fine to predict data set when predicators and variable have underlying association and modifications between them was estimated to be predictable [1].

A regression model was linear regression and logistic regression model usually used. In investigation, logistic regression instead of linear regression as latter was primarily applied to predict association between single input variable and aim variable with category [6].

3 The Graduate Students Data Set and Preprocessing

The data set comprises graduate students' information composed of Department of computing and Technology. The student's data set as sample data contains about 500 records and 13 attribute. Table 1 shows the attributes, description and the possible values that exist in the data set.

The department of Computing and Technology awarded their graduate bachelor degree included two areas for bachelor degree in Computer Science and Software

Table 1. The graduate students data set description

Variable	Description	Possible values
Faculty	The name of the faculty	Computing
Computing program	The name of the program	BS (Computer Science), BS (Software Engineering)
Bachelor academic year	The year of academic	1st Year, 2nd Year, 3rd Year, 4th Year
H.S.C or equivalent study medium	The type of medium	Urdu, English
1st year semester 1	Semester 1 (GPA)	GPA (1.00 to 4.00)
1st year semester 2	Semester 2 (GPA)	GPA (1.00 to 4.00)
2nd year semester 1	Semester 1 (GPA)	GPA (1.00 to 4.00)
2nd year semester 2	Semester 2 (GPA)	GPA (1.00 to 4.00)
3rd year semester 1	Semester 1 (GPA)	GPA (1.00 to 4.00)
3rd year semester 2	Semester 2 (GPA)	GPA (1.00 to 4.00)
4th year semester 1	Semester 1 (GPA)	GPA (1.00 to 4.00)
4th year semester 2	Semester 2 (GPA)	GPA (1.00 to 4.00)
Classes are mostly	The procedure of class	Lecture & discussion, lecture & lab, lecture, discussion & practical lab, lecture based

Engineering. The data preparation and preprocessing of data set and to get better input data for data mining techniques, It was some preprocessing for composing data earlier loaded data set of software of data mining, immaterial attributes was removed. The elements as selected as shows in Table 1 were treated by the rapid miner software to apply the data mining approaches on them.

Figure 2 shows that two axis X and Y which was X include academic year and y contains that distinguished actions of the centers for all three clusters (C1, C2, C3) using K-mean algorithm in same dataset during the academic years. It is showing that movements of the three clusters' centers (shown as symbols circle, cross and triangle) are volatile and they heavily depend on the random choice of academic year (shown as colored circles, blue, red, green and orange). C1 Student ratio highest in first year, C2 Student ratio highest in second year and fourth year, and C3 Student ratio highest in fourth year from academic year.

Predictor of educational failure or success such as statistically important ($p \leq 0$) Fig. 3 indicate probability of achievement according to different values of continuous variables where probability of achievement is directly proportional to score obtained in

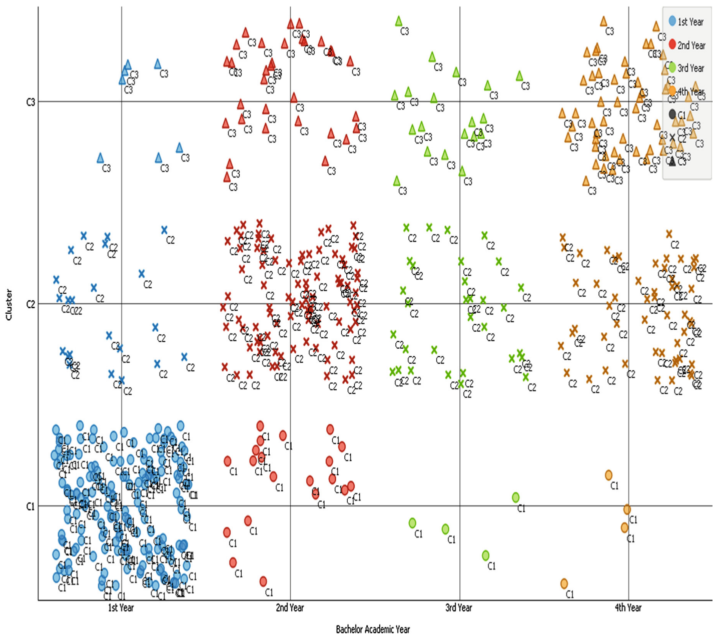

Fig. 2. Strength of students clusters wise using k mean. (Color figure online)

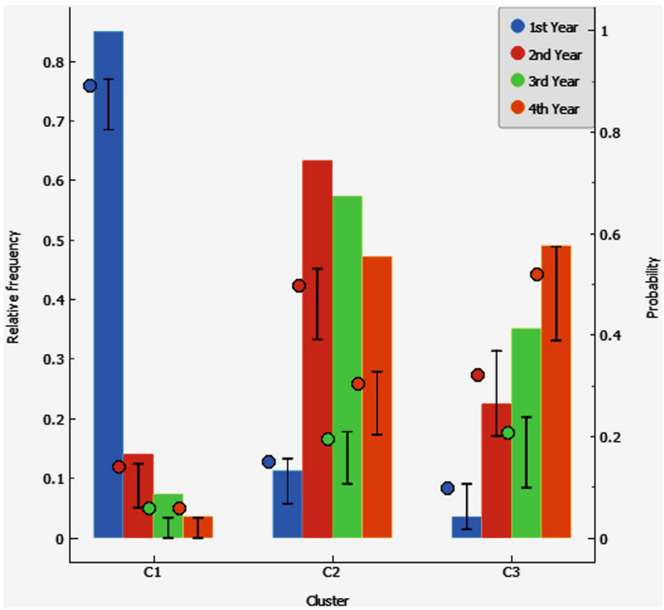

Fig. 3. Distribution of 'Cluster' grouped by 'Bachelor Academic Year' (relative probability)

academic. Examining three clusters is different importance in cluster analysis that distinguishable performing successfully [7]. Essential assumption in relation to investigation is that data must be approximately distributed in year wise using K mean cluster as C1 contain first year probability was 0.8, C2 have second year 0.7 probability and third year 0.64 probability, and C3 in fourth year 0.59 probability.

Figure 4 illustrate that number of students was 500 in clusters as logistic regression classifiers predicted in percentage of student as C1 96.9%, C2 98.5% and C3 97.3%. C2 have highest percentage that was best cluster. It helps to find out which behavior are highly related to dependence by identifying which predictor variables contribute more to target variable. Figure 5 shows that two axes X and Y which was Y include academic year and X contains cluster (C1, C2, C3) that was proved as C2 best cluster because student ratio is high.

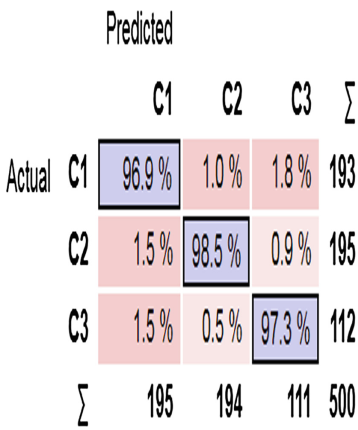

Fig. 4. Logistic regression (showing proportion of predicted)

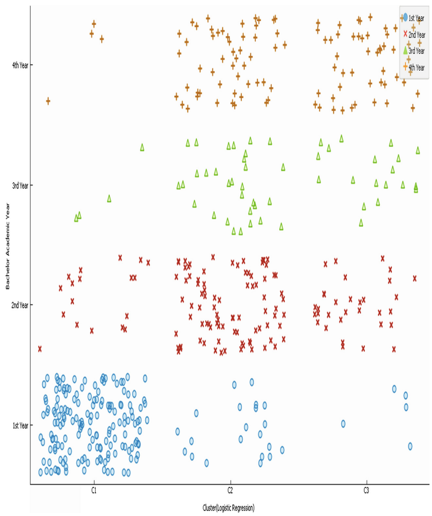

Fig. 5. Classify clusters (c1 c2 c3) using logistic regression

However, it needs to check whether the logistic regression is statistically significant by using the relative frequency ratio. If the probability in the graph is less than 0.05, the Cluster is statistically significant and the predicator variables have an impact on the target variable. Based on the result, the regression is finally determined to be applied for cluster analysis that was C2 cluster probability ratio high (Fig. 6).

The comparison of clusters by performance indicators assists to find out unique behavioral characteristics of each cluster and therefore to differentiate them. Typically, researchers compare the cluster center of each variable within each cluster. But we consider that comparison of cluster center is not sufficient since cluster center only indicates the behavior of academic years in program and is highly affected by extreme

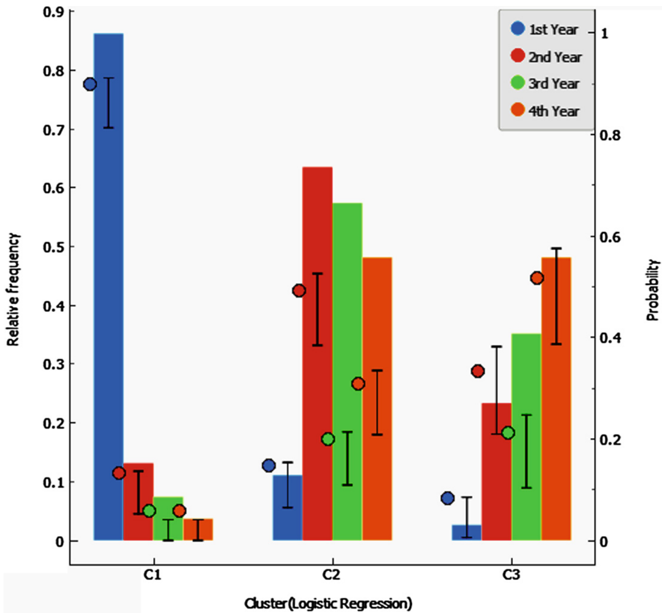

Fig. 6. Clarify the clusters (c1 c2 c3) of probability through logistic regression

values within each cluster. Figure 7 shows that Number of clusters as logistic regression classifiers reaches almost the same outcomes in terms of C1, C2 and C3. By further looking at the logistical regression C1 is slightly better than C3. However, it is hard to determine which cluster is better on basis of slight difference of cluster between each logistic regression classifier.

		Predicted			
		C1	C2	C3	Σ
Actual	C1	97.2 %	0.9 %	2.3 %	212
	C2	0.9 %	96.5 %	0.6 %	112
	C3	1.9 %	2.7 %	97.1 %	176
	Σ	213	113	174	500

Fig. 7. Probability measured cluster wise (C1, C2, C3).

In Fig. 8 logistic regression C1 (blue circle) at height zero and C2 (red cross) at height one. Every cluster pushes on distribution, though not with equal force. The C2 push dividing line towards the C1 and C1 push it back towards the C2, that logistic

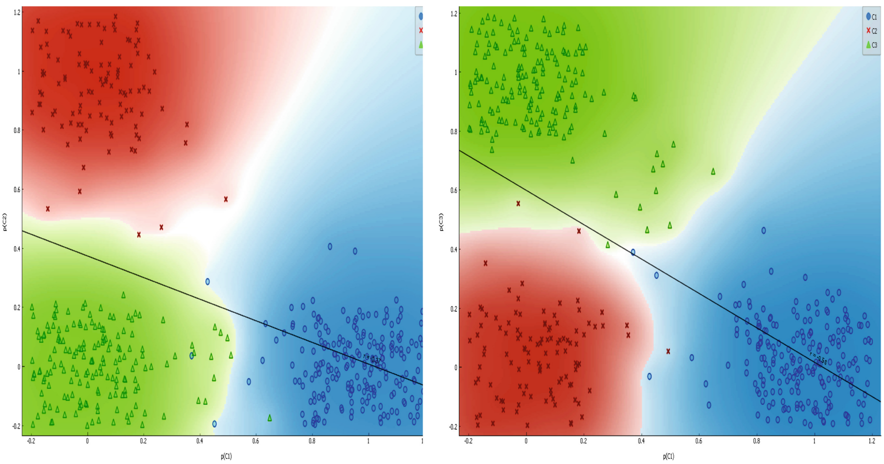

Fig. 8. Logistic regression clusters difference of P(C1) and P(C2), P(C1) and P(C3). (Color figure online)

regression algorithm selects could be thought of as equilibrium clusters of all these forces. The distribution on lift includes blue circle in green part of distribution and vice versa. The distribution on the right does a much better job of matching blue to blue and green to green, so this would be closer to probability of cluster 1 (P(C1)) and probability of cluster 3 (P(C3)) chosen by logistic regression.

Distribution is close to plane on one side of line, but below plane on other side, cross section of distribution is curve as logistic. Logistic regression is blue dots line at height one, green and red dots lines points at height zero, distribution that minimizes distances from distribution function based on logistic curve to dots lines and in plane.

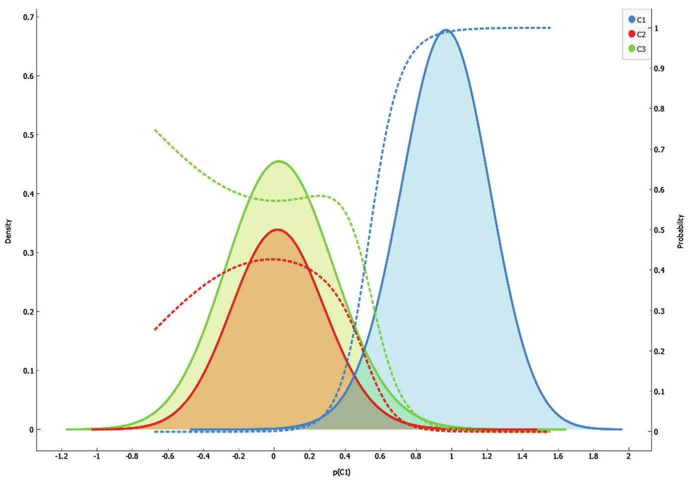

Fig. 9. Logistic regression clusters difference of P(C1), P(C2) and P(C3). (Color figure online)

It uses logistic function to fit output values between zero and one, just like a probability [11]. Figure 9 shows that were three cluster difference in three clusters include P(C1) logistic regression was R = 0.95 approximately 1.0 best score by regression line.

4 Conclusion

In this paper, students' performance is known problem which has its importance in found education and research strategies in order to raise quality level. The paper of such phenomena is benefit to society; sought to analyze association between student variables related with student determination and their capability to predict constant enrollment. It requires fundamental measures to be taken to increase not only financial situations but standards for research environment in university. Incessant development of commercial situations of students and creation of non-stressful situations are important aspects in growing performance of student. Consuming historic enrollment information predictive model was made to estimate enrollment probability of future student. Logistic regression classifier, related four year bachelor academic and demographic data on students to relative probability, was estimated. Subsequently enrollment pattern may modification such as in Campus policies, model needs to be always modified and validated year after year to improve its predictive power. This study cannot be used as stand-alone but helps to admissions administrators in decision making process to competently succeed enrollments.

References

1. Armstrong, J.S.: Illusions in regression analysis. Int. J. Forecast. **28**, 689–694 (2012)
2. Correa, A., González, A., Nieto, C., Amezquita, D.: Constructing a credit risk scorecard using predictive clusters. In: SAS Global Forum, p. 128. SAS Institute Inc. (2012)
3. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2006)
4. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning, p. 348. [I]n: Data Mining Applications the Interest is Often More in the Class Probabilities Themselves, Rather Than in Performing a Class Assignment (2009)
5. Jayakumar, D.S., Thomas, B.J.: A new procedure of clustering based on multivariate outlier detection. J. Data Sci. **11**, 69–84 (2013)
6. Linoff, G.S., Berry, M.A.: Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. Wiley Publishing Inc., Indianapolis (2011)
7. Mooi, E., Sarstedt, M.: A Concise Guide to Market Research. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-12541-6
8. Northern Arizona University: Multiple Regression (2002). Northern Arizona University. http://oak.ucc.nau.edu/rh232/courses/EPS625/Handouts/Regression/Multiple%20Regression%20-%20Handout.pdf. Accessed 8 Dec 2013
9. Oliver, T.A.L., Smith, C., Winston, S.J., Geranmayeh, F., Behjati, S., Kingston, O., Pollara, G.: Impact of UK academic foundation programmes on aspirations to pursue a career in academia. Med. Educ. **44**, 996–1005 (2010)
10. Pachgade, S.D., Dhande, S.S.: Outlier detection over data set using cluster-based and distance-based approach. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **2**(6), 12–16 (2012)

11. Allison, P.D.: Logistic Regression Using the SAS System: Theory and Application. SAS Institute and Wiley, Cary (2003)
12. Taylor, C.: What Is the Difference Between Alpha and P-Values? (2013). About.com. http://statistics.about.com/od/Inferential-Statistics/a/What-Is-The-Difference-Between-Alpha-And-P-Values.htm. Accessed 20 Nov 2013
13. Kumar, V., Chadha, A.: An empirical study of the applications of data mining techniques in higher education. Int. J. Adv. Comput. Sci. Appl. **2**(3), 80–84 (2011)
14. Razaque, F., Soomro, N., Shaikh, S.A., Soomro, S., Samo, J.A., Kumar, N., Dharejo, H.: Using Naïve Bayes algorithm to students' bachelor academic performances analysis. In: 4th IEEE International Conference of Applied Science and Technology, ICETAS (2017)
15. Pyle, D.: Data Preparation for Data Mining. Academic Press, Norwell (1999)