Examining students' graduation issues using data mining techniques - The case of TEI of Athens

Manolis Chalaris, Stefanos Gritzalis, Manolis Maragoudakis, Cleo Sgouropoulou, and Katerina Lykeridou

Citation: AIP Conference Proceedings **1644**, 255 (2015); doi: 10.1063/1.4907845
View online: http://dx.doi.org/10.1063/1.4907845
View Table of Contents: http://scitation.aip.org/content/aip/proceeding/aipcp/1644?ver=pdfcov
Published by the AIP Publishing

Examining Students' Graduation Issues Using Data Mining Techniques - The Case of TEI of Athens

Manolis Chalaris[1, a)], Stefanos Gritzalis[2], Manolis Maragoudakis[2], Cleo Sgouropoulou[1] and Katerina Lykeridou[1]

[1]*Technological Educational Institute of Athens, Department of Informatics, Aigaleo 12210, Athens, Greece*
[2]*University of the Aegean, Department of Information and Communication Systems, Samos GR-83200, Greece*

[a)]*Corresponding author: manoshlr@teiath.gr*

Abstract. One of the major issues that Greek Higher Education Institutes face is the delayed completion of studies of their students. For example, in the case of the Technological Educational Institute of Athens, in the academic year 2012-2013, the percentage of graduates with a length of studies of more than 6 years was 53%. This "problem" becomes harder if we consider that according to the new legislation, the Greek Higher Education Institutes (HEI) must cut off access to the students who "linger" too long. This means that many of these graduates wouldn't be able to complete their studies. While many institutes have systems to quantify and report the length of studies of all graduates, far less attention is typically paid to each student's reason(s) for delayed graduation. In this paper, we focus on examining the question of why students delay in the completion of their studies using several data mining techniques. Through the application of data mining techniques new knowledge will be provided to the administration of a HEI that could be used for solving this problem. The data used in our case study come from a questionnaire distributed to graduates of the institute but also from educational data stored in the Institute's student database.

INTRODUCTION

Higher Education Institutes, serving as knowledge centers and human resource developers have mission to generate, accumulate and share knowledge. They need to have a clear mission and attainable objectives that can be interpreted and analysed by using terms of strategic planning in order to solve their educational problems and enhance their quality. One of the major problems HEI face is the length of studies of the students who "linger" that leads in delayed graduation. In Greece, this problem becomes harder if we consider that, according to the new legislation; the Greek HEIs must cut off access to the students who "linger" too long. The administration of the Institutes should find solution to this issue by firstly examining the reasons of the delay completion of studies of their students and then taking the appropriate decisions to eliminate or at least restrict these reasons in order to provide an efficient environment for students to achieve their learning objectives without any problem.

Educational Data Mining has proven to be a relative new research field that helps HEIs to extract significant knowledge from their past and current data using several techniques and methods and thus supporting the decision making process for enhancing the quality of educational activities. A comprehensive study on the development of educational data mining has been conducted by Romero & Ventura in [1] focusing on the application of data mining techniques in educational systems from 1995 until 2005, while Siti Khadijah Mohamada & Zaidatun Tasira [2] presents the latest trends on data mining in educational research. Thoughout recent years, we can find many studies with subject the application of data mining in educational data of HEIs. In [3] Dr Kumar and Chadha presented some experiments of the application of several data mining techniques in a Higher Educational Institute of Sri Lanka and how the extracted information can be used for decision making processes like Organization of Syllabus, Predicting Student Performance etc. Ranjan and Malik [4] presented a holistic model for educational purpose using data mining approach for exploring the effects of probable changes in processes related to admissions, course delivery and recruitments and thus improving the quality of educational processes. Furthermore, Mohammed M.

International Conference on Integrated Information (IC-ININFO 2014)
AIP Conf. Proc. 1644, 255-262 (2015); doi: 10.1063/1.4907845

Abu Tair & Alaa M. El-Halees [5] use educational data mining to discover knowledge that may affect the students' performance while Karel Dejaeger et al. in [6] investigated the construction of data mining models to identify the main attributes of students' satisfaction. Natek and Zwilling [8] explored the possibility to predict the success rate of students enrolled to an academic course using contemporary data mining tools available to HEIs. Finally, Guruler et al [7] developed a knowledge discovery software and tested it on student data of Mugla University. They used a decision tree classification as data mining technique in order to explore which attributes may affect the student success in education.

Our aim is to describe the problem of the Length of Studies more precisely and investigate the question of why students are "lingering" in the Greek Higher Education and thus delay in the completion of their studies. In our case, we applied several data mining techniques like cluster analysis, correlation analysis and decision tree classification in educational data of the Technological Educational Institute of Athens (TEIA) in order to extract useful knowledge for students of all faculties of the Institute concerning the problem of the Length of Studies.

In the following sections the methodology is formulated following four phases: Definition of the problem of the Length of Studies for the TEI of Athens, Collection of the Educational Data, Preparation of the Educational Data and finally Application of the Data Mining Models and Evaluation of the results. In the end of this paper we present our conclusion and some thoughts for future activities.

METHODOLOGY

Technological Educational Institution of Athens (TEI of Athens) is the biggest Technological Institute of Greece and comprises five faculties and 27 Departments. The faculty of Technological Applications (STEF), the faculty of Health and Caring Professions (SEYP), the faculty of Management and Economics (SDO), the faculty of Fine Arts and Design (SKS) and the faculty of Food Technology and Nutrition (STETROD). The faculty with the most Departments is SEYP with twelve, while STETROD has only two. In this section we deal with the application of data mining technology in educational data of TEI of Athens describing first the problem we deal with and then the steps of the methodology we follow for our experiments.

The steps involved in our knowledge discovery process are Data Collection, Data Preparation as well as Modeling and Evaluation. In Data Collection, an initial data set is collected for the selected problem which is obtained by the questionnaire delivered to graduates of academic year 2013-2014. In Data Preparation several pre-processing and transformation activities are conducted in order to construct the final data set. Finally, in the Modeling and Evaluation phase, various modeling techniques are selected, applied in our data and the derived outcomes are evaluated in order to produce discover new knowledge.

The problem of the Length of Studies

A well-known problem in the Greek Higher Education is the late graduation. There are many students that are still "lingering" in the Departments (beyond six years) before graduation. In Technological Educational Institute of Athens in the academic year 2012-2013, the percentage of graduates with a length of studies of more than 6 years was 53%, while in year 2013-2014 it increased in 65%. In the following Table 1 we can see the number of graduates of TEI of Athens the last four academic years. As we can notice, the number of graduates is decreasing. From 4039 graduates in academic year 2009-2010, we had 2997 and 3061 in years 2011-2012 and 2012-2013 respectively.

TABLE 1 Number of Graduates of TEI of Athens the last four academic years

	2009-10		2010-11		2011-12		2012-13	
Faculty	MEN	WOMEN	MEN	WOMEN	MEN	WOMEN	MEN	WOMEN
SEYP	370	1696	335	1423	265	1118	246	1124
SDO	292	678	268	586	220	439	245	454
STEF	436	181	454	186	443	144	420	143
SKS	64	204	65	247	66	197	69	217
STETROD	55	63	34	52	40	65	53	90
TOTAL	**1217**	**2822**	**1156**	**2494**	**1034**	**1963**	**1033**	**2028**
	4039		**3650**		**2997**		**3061**	

In addition to that, the number of students that are still “lingering” in their Departments (beyond six years) is increasing. Figure 1 presents the percentages of students that are still "lingering" in the total of the enrolled students of TEI of Athens in the last four academic years. As we can see, the percentage in academic year 2010-2011 was only 16% and in the next two academic years it increased in 30% and 29%. All these information we collected from the databases of TEIA, highlighted the problem of the delayed graduation of the students of the Institute.

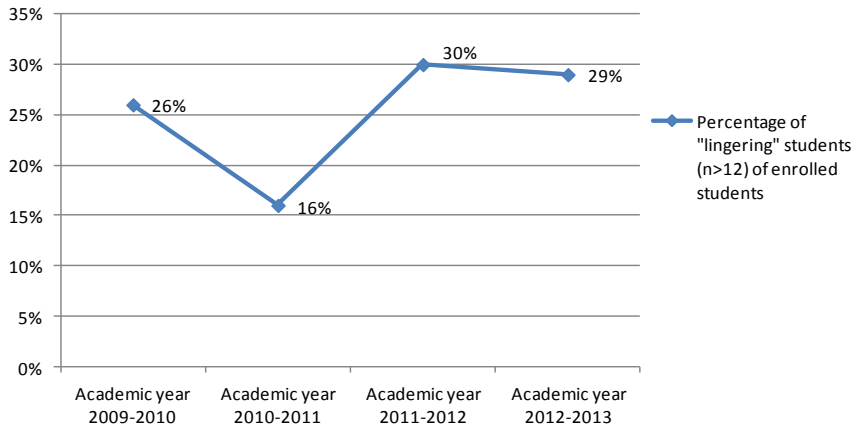

FIGURE 1 Percentage of "lingering" students in the total of the enrolled students of TEIA

Data Collection

The data set used in this study was obtained from a questionnaire delivered to graduates of the TEI of Athens of the academic year 2013-2014. Each Department gathered all filled questionnaires of his graduates and sent them to the Quality Assurance Unit of the Institute where we have digitized and "cleaned" them, in order to offer the possibility of analyzing and visualizing data included in these documents. The structure of the questionnaire included two parts. First part contained personal questions of the graduates like Faculty, Department, Gender, Marital Status, Year of Admission etc. Second part contained ten questions about possible causes that made them delay their graduation. There were three types of questions in the questionnaire, 3 yes-no questions, 5 five-level Likert scale questions and 2 open-ended questions. Figure 2 depicts the number of men, women and the total number of graduates answering the questionnaire per Faculty. As we can see, almost the half of the graduates comes from the Faculty of Health and Caring Professions (SEYP) and the majority is women.

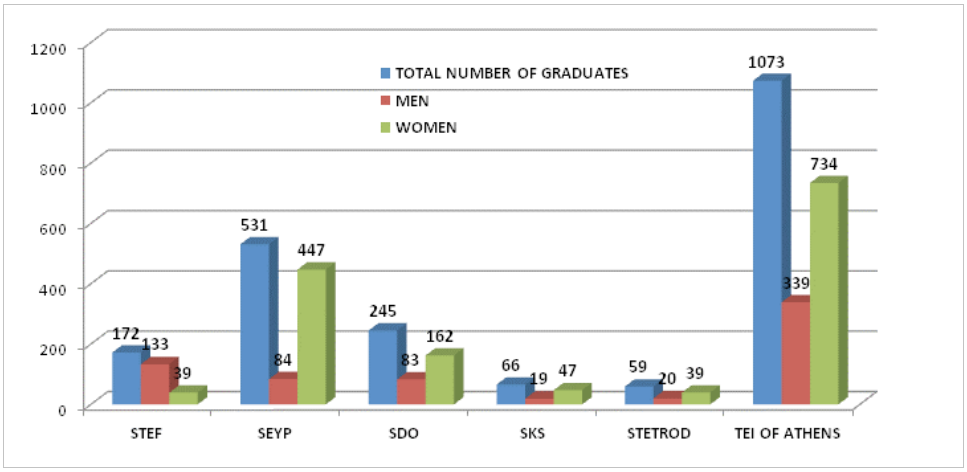

FIGURE 2 Number of graduates' questionnaire per Faculty and per Gender

In Figure 3, we illustrate the average Length of Study of the graduates per Faculty and for the whole Institute, while the red line shows the typical Length of Study for the TEI of Athens which is four years.

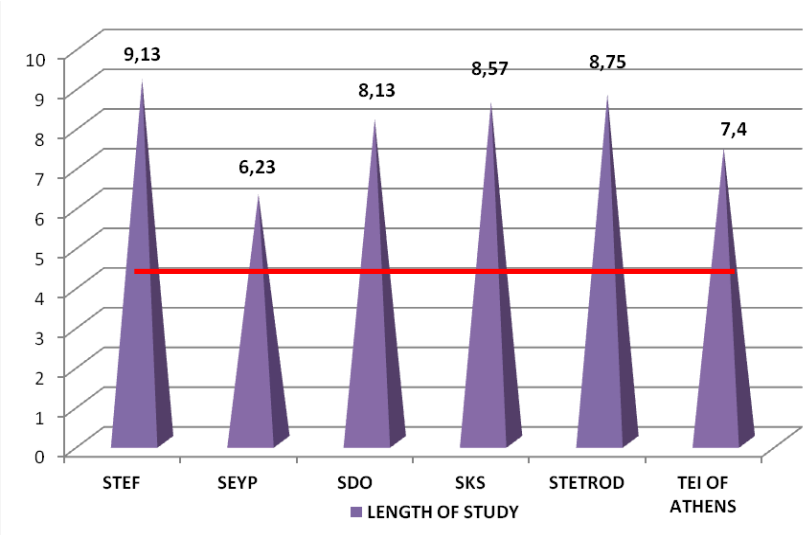

FIGURE 3 Length of Study of the graduates per Faculty and in total

As seen in the Figure, students of the Faculty of Health and Caring Professions needed the fewer years to graduate, 6.23 years, while students from the Faculty of Technological Applications needed 9.13 years which is the highest average. The average Length of Study for the whole Institute is 7.4 years.

Data Preparation

Initially in this step, we gathered our data in one dataset and after data integration we performed data cleansing operations. More specifically, concerning highly incomplete records (a record corresponds to an entire questionnaire) it was deleted from the dataset. In cases of few missing values, the record remained in the data set and its missing values were replaced by the average, if it was a Five-level Likert scale question, or remained missing if it was a yes-no question. The next activity was to select the items that can be proved interesting for conducting the analysis. The selection was made by human experts that participated in our research. Table 2 depicts the items selected for conducting the experiments, the format of each question and its explanation.

TABLE 2 Items for conducting analysis

Item - Attribute	Format of question	Explanation – Comments
Faculty	Personal question	In which Faculty were you studying?
Gender	Personal question	What is your Gender?
Length of Study	Personal question	What was your length of study?
q1: Working	Yes-no question	I was working and I hadn't enough time for studying
q2: Erasmus	Yes-no question	I spent time for Erasmus program and that delayed my study in Greece
q3: Not interesting subject	Five-level Likert scale question	My subject's study was not interesting
q4: Distance	Five-level Likert scale question	I stayed far away from the Institute
q5: Strict teachers	Five-level Likert scale question	The teachers were very strict
q6: unlimited length of study	Five-level Likert scale question	There were no limitation in length of study
q7: Labor market prospects	Five-level Likert scale question	The labor marker prospects discourage me to finish my studies

In some experiments data needed to be transformed in order to apply the model and to achieve more interesting and useful outcomes. For example, we transformed the item "Length of Study" from numerical to ordinal (3 categories) using the following index: to values from 4 (4 are the minimum years of study for a student) till 6 years a "<n+2" value was assigned, to values from 6 till 8 years a ">=n+2 & <2n" value was assigned and to values above 8 years a ">2n" value was assigned.

Modeling and Evaluation

In this section, we describe the experiments conducted on the educational data of TEIA and present the results. For the application of the data mining techniques the Rapid Miner software is used, which is the world-leading open-source system for data mining.

As a first experiment we conducted a correlation analysis in order to examine if the attribute "Length of Study" interacts with other attributes of our dataset. Figure 4 presents the correlation matrix, where we have the correlation coefficients between the attributes used in this experiment. As we can see the attribute "Length of Study" has correlation with the attributes q1_working and some correlation with the attributes q6_unlimited_length_of_study.

Attributes	Length_of_...	q1_working	q2_Erasmus	q3_not_inte...	q4_distance	q5_strict_te...	q6_unlimite...	q7_Labor_...
Length_of_study	1	-0.508	-0.055	-0.009	-0.022	0.021	0.184	0.079
q1_working	-0.508	1	-0.061	0.056	0.069	-0.035	-0.094	-0.028
q2_Erasmus	-0.055	-0.061	1	-0.012	0.009	-0.028	0.006	-0.027
q3_not_interesting_subject	-0.009	0.056	-0.012	1	0.245	0.258	0.297	0.353
q4_distance	-0.022	0.069	0.009	0.245	1	0.217	0.232	0.217
q5_strict_teachers	0.021	-0.035	-0.028	0.258	0.217	1	0.281	0.322
q6_unlimited_length_of_study	0.184	-0.094	0.006	0.297	0.232	0.281	1	0.371
q7_Labor_market_prospects	0.079	-0.028	-0.027	0.353	0.217	0.322	0.371	1

FIGURE 4 Correlation Matrix

Our next experiment was to conduct a cluster analysis using the k-means algorithm. As number of clusters we chose k=3. For determining the optimal number of clusters we used Ward's algorithm, which is also used and proposed in [9]. Concerning the distance measure, we used the Squared Euclidean Distance. As we can see in Table 3, the cluster_0, which is the cluster with the lowest average in years of study (5.625 years), consists in 78.3% of women, the majority (60%) belongs to the Faculty of Health and Caring Professions (SEYP) and 57.9% of these graduates were not working during their studies. In the other two clusters, cluster_1 and cluster_2, which include graduates with an average of 9.085 and 14.163 years of study respectively, the percentage of women is almost equal with this of men, they belong mostly in STEF and SDO and 87.8% in cluster_1 and 93.5% in cluster_2 declared that hadn't enough time for studying because they were working. In a previous work [10], in which we conducted a cluster analysis in student data derived from the theoretical courses questionnaire, we found out that students of the

Faculty of Health and Caring Professions (SEYP) are more consistent in their studies (attendance, studying, and understanding) and evaluated better their teachers and the organisation of the courses they attend than those of the other Faculties. In addition, in [11] some association rules were presented, which showed that students of SEYP participated in the class and used library services more than the other students of the TEI of Athens and they had better percentage of graduation and better Grades, by far. These outcomes fit with the outcome of this study concerning the graduates of SEYP and confirm that students of this Faculty are "lingering" less and graduate earlier than those of the other Faculties. Concerning the other attributes participating in this analysis, there is no useful outcome except that the averages of attribute "q6: unlimited length of study" are higher in the clusters that have higher averages in the attribute "Length of Study".

TABLE 3 Centroid table of cluster analysis

Attribute	cluster_0	cluster_1	cluster_2
%	(456)	(295)	(92)
Faculty= SDO	0.221	0.275	0.326
Faculty= SEYP	0.603	0.237	0.228
Faculty= SKS	0.042	0.098	0.087
Faculty= STEF	0.090	0.298	0.304
Faculty= STETROD	0.044	0.092	0.054
Gender=F	0.783	0.485	0.543
Gender=M	0.217	0.515	0.457
Length of Study	5.625	9.085	14.163
q1: Working=no	0.579	0.122	0.065
q1: Working =yes	0.421	0.878	0.935
q3: Not interesting subject	1.522	1.586	1.522
q4: Distance	1.616	1.675	1.804
q5: Strict teachers	2.147	2.329	2.065
q6: unlimited length of study	1.908	2.386	2.641
q7: Labor market prospects	1.816	2.176	2.109

Our next objective was to apply a classification task, specifically a decision tree classification method, in order to examine the characteristics of the students that graduate early (<n+2 years), late (>=n+2 & <2n years) and very late (>2n years), based of course on the attributes collected through the questionnaire. For this reason, the attribute Years of study is set to have the role of the label (target) attribute. In addition, another data transformation task was conducted. Specifically, the attributes q3 till q7, which had a range of value from 1-5 based on the Five-level Likert scale, transformed using the following index: to values 1 and 2 a "little or no significant reason" value was assigned and to values from 3 to 5 a "quite or very significant reason" value was assigned.

Figure 5 depicts the decision tree that is generated. As splitting criterion we used the measure called gini index and as seen in the Figure the attribute that is more affecting the target attribute and is first selected for splitting the data is q1_working.

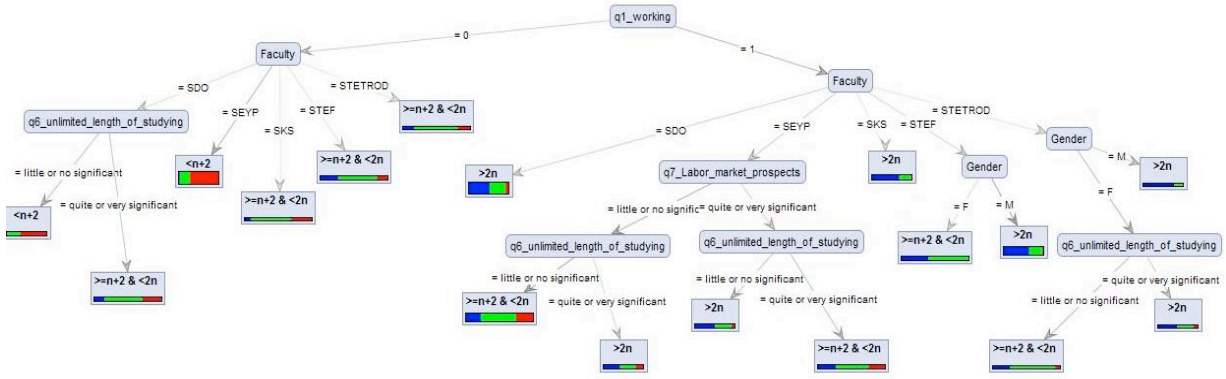

FIGURE 5 Graphical display of the Decision Tree

A significant outcome of this classification tree is that graduate of all Faculties that were working during their studies are classified whether in the category “late” or in the category “very late”. Concerning the graduates that were not working during their studies, there are other attributes that define the category they are classified, such as the Faculty they belong to.

Following are the results in a text view where we can see also the size of the class frequencies of each leaf of the tree.

```
q1_working = 0
|  Faculty = SDO
|  |  q6_unlimited_length_of_studying = little or no significant: <n+2 {>2n=1, >=n+2 & <2n =9, <n+2=18}
|  |  q6_unlimited_length_of_studying = quite or very significant: >=n+2 & <2n {>2n=4, >=n+2 & <2n =14, <n+2=7}
|  Faculty = SEYP: <n+2 {>2n=8, >=n+2 & <2n =48, <n+2=131}
|  Faculty = SKS: >=n+2 & <2n {>2n=1, >=n+2 & <2n =6, <n+2=3}
|  Faculty = STEF: >=n+2 & <2n {>2n=9, >=n+2 & <2n =19, <n+2=5}
|  Faculty = STETROD: >=n+2 & <2n {>2n=2, >=n+2 & <2n =7, <n+2=2}
q1_working = 1
|  Faculty = SDO: >2n {>2n=88, >=n+2 & <2n =66, <n+2=11}
|  Faculty = SEYP
|  |  q7_Labor_market_prospects = little or no significant
|  |  |  q6_unlimited_length_of_studying = little or no significant: >=n+2 & <2n {>2n=23, >=n+2 & <2n =52, <n+2=25}
|  |  |  q6_unlimited_length_of_studying = quite or very significant: >2n {>2n=12, >=n+2 & <2n =11, <n+2=5}
|  |  q7_Labor_market_prospects = quite or very significant
|  |  |  q6_unlimited_length_of_studying = little or no significant: >2n {>2n=15, >=n+2 & <2n =12, <n+2=2}
|  |  |  q6_unlimited_length_of_studying = quite or very significant: >=n+2 & <2n {>2n=8, >=n+2 & <2n =14, <n+2=7}
|  Faculty = SKS: >2n {>2n=28, >=n+2 & <2n =12, <n+2=0}
|  Faculty = STEF
|  |  Gender = F: >=n+2 & <2n {>2n=9, >=n+2 & <2n =13, <n+2=0}
|  |  Gender = M: >2n {>2n=67, >=n+2 & <2n =34, <n+2=2}
|  Faculty = STETROD
|  |  Gender = F
|  |  |  q6_unlimited_length_of_studying = little or no significant: >=n+2 & <2n {>2n=4, >=n+2 & <2n =10, <n+2=1}
|  |  |  q6_unlimited_length_of_studying = quite or very significant: >2n {>2n=5, >=n+2 & <2n =4, <n+2=1}
|  |  Gender = M: >2n {>2n=11, >=n+2 & <2n =3, <n+2=0}
```

In order to evaluate our model, we split our dataset in two datasets, a training one and a testing one in which we have removed the values of the Length of Study attribute. Then we developed a decision tree data mining model using the training dataset and we applied it to the testing dataset in order to predict the length of study of the graduates belonging to this. The accuracy of this model is 55%, which is acceptable but not very high. In Table 4 we see the accuracy in each Faculty of TEIA.

TABLE 4 Accuracy of Classification Model per Faculty

Faculty	Accuracy percentage
Faculty of Technological Applications (STEF)	50%
Faculty of Health and Caring Professions (SEYP)	56%
Faculty of Management and Economics (SDO)	44%
Faculty of Fine Arts and Design (SKS)	80%
Faculty of Food Technology and Nutrition (STETROD)	54%

As we can see in the table, the accuracy of SKS is very high (80%), while in SDO is very low (44%). For improving the accuracy of the model, we assigned to values “>=n+2 & <2n” and “>2n” of the attribute “Length of Study” the “>=n+2” value, since a student is considered “lingering” after he studies for n+2 years and we conducted the experiment again. The accuracy of the new model was 71%, which is a good prediction percentage.

CONCLUSIONS AND FUTURE ACTIVITIES

In this paper, we discussed about students' graduation issues in Higher Education Institutes and especially about the problem of the Length of Studies ("lingering students") which is a major and hot issue in Greek educational area. Our aim is investigate this issue using data mining methods and thus extract useful knowledge for supporting the decision making procedure. Several data mining techniques like cluster analysis, correlation analysis and decision tree classification were applied in educational data of the Technological Educational Institute of Athens (TEIA) and interesting and useful knowledge is extracted.

As main conclusion for solving the problem of the long Length of Study of the students, based on the analysis we conducted, we claim that the focus of the administration of the Institute should be given to the students that are working during their studies especially for the Faculties that have the biggest problem, like STEF. A good strategy could be to allow to students who work, to study part time.

In our future plans we aspire to present a holistic and strategic approach of applying data mining methods in HEIs that will support the decision making of solving more educational problems and enhancing the quality of educational processes. These plans include the identification of indicators that determine the quality of educational processes and the collection of educational data of many stakeholders of TEIA, such as Faculty and Administrative members. Of course, new attribute should be defined also concerning the students and graduates of the Institute that will allow us to have more integrate, accurate and eventually useful outcomes when applying data mining methods in our educational data.

ACKNOWLEDGEMENTS

Research underlying this article has been supported by the national project "MODIP of TEI of Athens" – NRSF 2007-2013

REFERENCES

1. Romero, C. and Ventura, S. (2007) 'Educational data mining: A Survey from 1995 to 2005', Expert Systems with Applications (33), pp. 135-146.
2. Siti Khadijah Mohamada, Zaidatun Tasira (2013) 'Educational data mining: A review', Procedia - Social and Behavioral Sciences 97 (2013) pp.320 – 324
3. Dr V. Kumar, Anupama Chadha (2011), "An Empirical Study of the Applications of Data Mining Techniques in Higher Education", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No.3 March 2011
4. J. Ranjan and K. Malik, "Effective educational process: a data-mining approach", The journal of information and knowledge management systems Vol. 37, No 4, 2007, pp. 502-515
5. M. M. Abu Tair, Alaa M. El-Halees, (2012). Mining Educational Data to Improve Students' Performance: A Case Study, International Journal of Information and Communication Technology Research, Vol.2, No. 2 February 2012
6. K. Dejaeger, F. Goethals, A. Giangreco, L. Mola, B. Baesens, "Gaining insight into student satisfaction using comprehensible data mining techniques", (2011) European Journal of Operational Research, Vol. 218 Iss: 2 pp. 548 - 562
7. H. Guruler, A. Instanbullu and M. Karahasan, "A new student performance analysing system using knowledge discovery in higher educational databases", Computers & Education 55 (2010) pp 247-254
8. S. Natek and M. Zwilling, "Student data mining solution–knowledge management system related to higher education institutions", Expert Systems with Applications 41 (2014) pp 6400-6407
9. P. Belsis, A. Koutoumanos, C. Sgouropoulou (2013). "PBURC: a patterns-based, unsupervised requirements clustering framework for distributed agile software development", Requirements Engineering © Springer-Verlag London 2013
10. M. Chalaris, S. Gritzalis, M. Maragoudakis, C. Sgouropoulou and A. Tsolakidis, (2013). "Improving Quality of Educational Processes Providing New Knowledge using Data Mining Techniques", Procedia - Social and Behavioral Sciences, 3rd International Conference on Integrated Information (IC-ININFO 2013)
11. M. Chalaris, I. Chalaris, Ch. Skourlas, An. Tsolakidis, (2012). "Extraction of rules based on students' questionnaires", Procedia - Social and Behavioral Sciences, Volume 73