A Learning Analytics Methodology for Student Performance Assessment in a Distance and Open Education Environment

Vasileios Kagklis

Hellenic Open University, Educational Content, Methodology & Technology Laboratory, 278 Patron-Claus Str., GR-263 35 Patras, Greece
kagklis@eap.gr

Antonis Lionarakis

Hellenic Open University, School of Humanities, 18 Parodos Aristotelous Str., GR-263 35 Patras, Greece
alionar@eap.gr

Elias C. Stavropoulos

Hellenic Open University, Educational Content, Methodology & Technology Laboratory, 278 Patron-Claus Str., GR-263 35 Patras, Greece
estavrop@eap.gr

Vassilios S. Verykios

Hellenic Open University, School of Science & Technology, 18 Parodos Aristotelous Str., GR-263 35 Patras, Greece
verykios@eap.gr

Abstract

Advances in data storage devices and data collection techniques have enabled the capturing and persistent maintenance of all the information related to the interaction between students and instructors in the context of the learning processes, supporting a 360-degree view of the student profile. Data Mining, Data Analytics, and Exploratory Data Analysis techniques can be used to make sense out of these multidimensional and endlessly generated data in order to offer added value to the higher educational academic institutes and universities. In this paper, we present a learning analytics approach that has as its goal to improve both the learning experience of students and the instructional experience of tutors, as well as the institutional strategic view of the university. By exploring and analyzing all the collected data, we build models that explain and assess the effectiveness of the learning environment, so as to address the needs of the aforementioned members of these organizations. In this underlying framework, we also present and evaluate a case study that focuses on analyzing educational data for assessment purposes from a big data infrastructure, which is under development in the Hellenic Open University in Greece.

Keywords: Big Data, Learning Analytics, Student Performance Assessment, Distance Education

1. Introduction

The Hellenic Open University (HOU) was officially established in 1997 and is the only University in Greece that exclusively offers distance education courses. Since its establishment, the HOU has evolved and has been attracting more and more students of all ages, from different cities, and of a variety of profession and financial status, that have shown great interest in studying through a distance learning program. HOU consists of four Schools, offering undergraduate and graduate courses to adult learners. Each course consists of modules and students have to submit 4-6 written assignments throughout the 10-month academic year period and participate in a compulsory sit exam at the end of it. Furthermore, each course module includes five not compulsory face-to-face counselling group sessions that take place in 9 cities all over the country. Refer to http://www.eap.gr for a presentation of the undergraduate and the graduate studies and courses' structure in HOU.

Communication and interaction between tutors and students is mainly held via e-mail and telephone as well as through the Student Information System http://open.eap.gr and the Learning Management System (LMS) https://study.eap.gr/. Students at HOU are provided with a variety of learning materials (printed course material, audio and video material, CD-ROMs/software, etc.) specially prepared for distance learning, most of them located in the digital material repository https://apothesis.eap.gr/. Through the digital repository, students have access to digital educational material as a supplement or alternative material for their study. The Student Information System of HOU contains information supplied by the Register's Office of the HOU, concerning the students' record, the organization of modules into groups, the time, date and place of the counselling group sessions for each module, written assignments and final exams grades, etc. The LMS of HOU is based on the Moodle platform (https://moodle.org) and it has been offering services to students and tutors since the academic year 2013-14. Through this platform, students have the ability to submit their written assignments, work reports and answered questionnaires within their academic studies, while tutors are able to give feedback, annotate and grade their assignments or work reports. In addition, work spaces for asynchronous discussions at module level, discussion groups and online fora of students and tutors are available which are managed via an automated process. The service is configured properly to keep pace with the academic calendar of modules and to provide students and tutors with direct access to the activities of the current week. Moreover, the web conferencing platform http://centra.eap.gr/ is used for synchronous teleconferencing between students and tutors, and offers the possibility for an interactive classroom experience, in addition to the face-to-face counselling group sessions.

The above platforms provide a vast amount of valuable educational and administrative data. Log files and activity reports, forum posts, participation in counselling group sessions, written assignments and final exams' grades, can give feedback about students' activity, attention, active participation and engagement. Along with information from student admissions, all these data can be combined and the exploitation and the analysis of them can provide knowledge to empower both student learning and tutors teaching experience, enrich educational experience and elevate the level of academic excellence.

In this study, we provide a proof of concept of a sample of such data, concerning the "PLS60 - Specialization in Software Engineering", a second year module of the graduate course Master's in Information Systems of the School of Science and Technology. The aim of this programme is to offer its students the opportunity to

acquire specialized knowledge in Information and Communication Technologies, and to prepare them for professional work in the design, development and management of integrated information systems. To fulfil the program requirements, a student has to succeed in four modules and submit and successfully defence a thesis of his/her choice. Module PLS60, in particular, covers subjects from Database Theory and Data Management (ER, MySQL etc.), Data Mining techniques (like classification, clustering, regression, and association rules), Operating Systems (concurrency and paging) and Modern programming paradigms (the Java programming language). To complete the module, students have to successfully submit six written assignments during the academic year, and succeed in the final written examination.

The rest of this paper is organized as follows: Section 2 gives a brief background insight of some previous related works. Section 3 presents the methodology followed for the analysis of the data, describes the data set and the software tools utilized for their processing. Section 4 demonstrates and discusses the results of the analysis and, the final Section concludes this work and presents some future ideas.

2. Related Work

Distance education (White, 1982; Byrne, 1989) is an alternative way of learning from a distance, without the need of physical presence in a classroom. In the last decade, it has gained a lot of attention and popularity. In distance education, online discussion fora are used as a means of communication, as they provide many benefits to students and teachers, countermeasuring the lack of face-to-face communication. According to Tiene (2000), students have been found to be in favor of the self-paced, self-regulated feature of asynchronous discussions compared to their face-to-face counterparts. Duffy, et al., (2002) report that students who have undertaken a degree from a distance obtained significantly higher average marks than those who have undertaken it on campus.

The application of data mining in education is an emerging research field known as Educational Data Mining (EDM) (Baker, 2010). It studies the development of methods for exploring data that come from educational environments. The aim of EDM is to provide a better understanding of students' behavior, to identify how students learn, and how the learning process can be augmented so as to improve the performance of the students (Berland et al, 2014). Each one of these issues asks for a specific solution and its own unique characteristics require a different way of treatment. Therefore, the knowledge discovery process has to be adapted in the needs of the specific problem each time. In their research, Baker and Yacef (2009) suggest four goals of EDM: predicting students' future learning behavior, discovering or improving domain models, studying the effects of educational support that can be achieved through learning systems, advancing scientific knowledge about learning and learners by building and incorporating student models.

The problem of predicting low performance or even the possible drop-out of students (Pierrakeas et al, 2004; Romero & Ventura, 2010; Pal, 2012) has long been recognized and is one of the hottest topics in EDM. A lot of research has been devoted on how to apply EDM techniques effectively, in order to create models that can predict dropout rates and school failure (Romero & Ventura, 2010). More specifically, statistical techniques, such as correlation analysis and regression, and data mining techniques, such as classification and decision trees, have been used to evaluate the prediction results for students' academic success (Hämäläinen & Vinni, 2010).

Data from the online discussion fora can be used to apply text mining and natural language processing (Manning & Schütze, 1999), along with social network analysis and sentiment analysis techniques (Pang et al, 2002; Turney, 2002), in order to extract useful knowledge about the behavior of the students, their mood during the course, their collaboration or communication patterns, or even try to predict their final performance based on these findings. (Turney, 2002) and (Pang et al, 2002) are among the first to use sentiment analysis, combined with machine learning algorithms. Currently, the existing approaches of sentiment analysis can be grouped into four main categories: keyword spotting, lexical affinity, statistical methods, and concept-level techniques (Cambria et al., 2013). Keyword spotting classifies text by affect categories based on the presence of unambiguous affect words (Ortony et al, 1988). Lexical affinity not only detects obvious affect words, but also assigns arbitrary words a probable "affinity" to particular emotions (Stevenson et al, 2007). Statistical methods leverage on elements from machine learning such as latent semantic analysis, support vector machines, "bag of words", etc. More advanced methods try to detect the holder of a sentiment and the target (Kim & Hovy, 2006). Lastly, concept-level approaches leverage on elements from knowledge representation such as ontologies and semantic networks and, thus, are also able to detect semantics that are expressed in a subtle manner (Cambria & Hussain, 2012).

In (Lotsari et al., 2014), social network analysis techniques were applied on educational data originated from the online forums of the Hellenic Open University, to obtain networks of students and instructors, according to their interaction. The analysis of the data has been accomplished by using the R and the Weka tools, in order to analyze the structure and the content of the exchanged messages in these fora as well as to model the interaction of the students in the discussion threads. Lately, in (Kagklis et al, 2015) sentiment analysis and opinion mining was applied on educational data obtained by the online forum of a graduate module, to analyze students' attitude towards the course, model their behavior and detect how this affected their overall performance. Moreover, a recent study on Massive Open Online Courses (MOOCs) is that of (Wen et al, 2014). The authors apply sentiment analysis on students' posts, in order to identify students' opinion for specific features of the course, and to evaluate if there is a connection between the sentiments and the students drop-out rate.

To the best of our knowledge, most of the existing studies focus on the analysis of the data being collected during and/or after the distance learning courses. However, important knowledge can emerge from data obtained by student applications for acceptance in distance learning programs. In (Kagklis et al, 2016) demographical data related to student admission for acceptance in programs offered by the Hellenic Open University, were studied. The authors analyze data to discover patterns and knowledge that can be used to help the strategic placement of the University, and to improve students' learning experience. Moreover, they attempt to correlate the discovered findings with the social and financial status of the applicants' environment.

3. Methodology

As we explained in the previous section, EDM is a research area that investigates the automatic analysis of data coming from an educational context, spanning all levels and types of educational systems. EDM can be considered as a specific area of Data Science, that is a broader term used for encompassing all aspects of management and analysis of data. Learning Analytics (LA) is another term used in parallel with EDM to signify the analysis of data originating from a learning environment. We consider EDM a broader term than

LA, but there is definitely a lot of overlap between the two terms, as well as maybe some issues that are not common to these two concepts.

In this paper, we present a holistic LA methodology for analyzing educational data from a distance learning university. As we explained before, a distance learning university presents a number of singularities that stem from the specific nature of education offering as well as from the subtle characteristics of the student population. As far as the model of education offering that is used in HOU is concerned, it is considered a blended leaning model, where distance offering of materials and tutor guidance and teleconferencing is offered along with a selected number of face to face counseling group sessions that are uniformly distributed in the timeframe of a module offering. Student population is biased towards adults or otherwise older generations, especially in the undergraduate programs, and this bias by itself introduces a lot of peculiarities that are dealt with both methodological and technological solutions that facilitate the learning process in various ways.

The LA methodology that we propose relies on assimilating and analyzing detailed data from all possible phases of student virtual or real presence in the academic environment and even beyond that. Specifically, we are focusing on integrating data from the whole students' lifecycle both static and dynamic as well as real time and historical, before, during and after their fulfillment of the requirements for the degrees they are seeking. The uniqueness of our proposal, is the multimodality, large dimension and size, as well as the complementarity of the sources all these different data reside. Even though we present only a partial view of the results coming from running a large number of experiments for data analysis purposes as a proof of concept, we strongly believe that a 360 degrees view of the students' presence in the university for accelerating and facilitating their learning is still a major step that needs a lot more of experimentation and testing.

The starting point in our methodology is the time that the student is applying for a position in one of the programs the university is offering. We should say here that HOU is one of the Hellenic public universities where the entrance of the students is by ballot, even though by the time of this writing, some of the programs are accepting all of the applicants. A lot of information about the people who applies can be found even from this subset of student admission data, and a lot of knowledge can be produced with respect to the patterns followed by applicants, so that the university can appropriately adjust well in advance the offering of new or the termination of not promising programs of study. A geographical distribution of the applicants can also shed some light to the urban or suburban areas the majority of the applicants originate from, so that new annexes of the university can be planned for future development to accommodate the increasing needs of the tentative student population among other strategic decisions that can be drawn about the placement of the programs.

A major component of the data that are fed into our methodology come from the student information system that contains information related to the student population from a university administration point of view, such as demographical data, previous student education related data, address data, family related information, courses enrolled, counseling group meetings schedule and student participation, grades and transcripts, payment information, and so on and so forth. The largest part of the data that support our LA methodology are coming from the learning management system used for the offering of the online courses.

A great amount of data spanning from logging data, posts to forums, assignments, grades, completion of various learning activities and quizzes have a central role in the analysis techniques used. In particular, we have so far applied in the past various social network analysis techniques for analyzing the group of students interacting with each other and with their instructors through the forums and chat rooms, as well as with fellow students in performing various collaborative assignments and tasks. A number of text mining techniques have been applied for analyzing the content of the messages exchanged through the forums in order to find out the main concepts in the studying materials that caused difficulties to the students, as well as the procedures in the course offering that challenged or puzzled the students most.

A number of classification and clustering techniques, like decision trees and classification rules or hierarchical agglomerative clustering can be used to analyze performance data from assignments and quizzes or various other activities and build a model that indicates what is the relationship between those and the final grade in the course, or even predicting in an early stage the students that have a high degree of probability of failing to complete a course or even that eventually will drop out of the program altogether. We should add to those techniques a number of sequential mining techniques that are important for analyzing logging data with a time dimension, in order to build a profile of the student engagement with the course site and the effect that such an engagement plays in the final performance of the student. A sample of experimental results related to such techniques can be found in the next section. Sequential mining techniques can also be applied to analyzing the participation in various video conferencing sessions offered throughout the academic year in the context of a course offering as an extra guidance for students in performing specific goals such as their software assignments or introducing a software in the form of a tutorial.

Finally, a number of sentiment analysis techniques have been applied to forum data to analyze the sentiment of the students from their postings in the forums they participate in. Even in a preliminary stage, this seems to be a promising direction of investigation, especially in case we start analyzing the student behavior from data posted in various social media sites the students may engage in participating, for collaborating with other fellow students for completing their assignments or preparing themselves for the final exams.

The last portion of the data we envision to use for building global profiles by analyzing the student behavior and engagement is the questionnaires answered during the end of the academic year for evaluating the course and the instructor by the student. HOU was a pioneer among the Greek universities to build an electronic system that supports anonymous evaluation of courses, teaching materials and instructors from students, in addition to allowing instructors to evaluate course materials and course coordinators and the other way around.

As a proof of concept we have used a sample of the data available in our databases related to the information systems program, and especially from the system engineering course, which is a graduate level course. The experimental results from applying some of the aforementioned techniques are presented in the next section, while the rest of the techniques have been applied to a similar set of data and the results can be found in a number of publications from our group.

For our first analysis we utilized a dataset consisting of 794,800 student applications, which were submitted towards 83 different distance learning modules, the equivalent of a course, during a 10-year period (2003-2013). The applications were made by a total of 362,311 different applicants, who could apply once a year in multiple modules. These data were used for geographical analysis, analysis of the applications that the university received the last decade, and analysis of the number of applications based on the age and the sex of the applicants.

We selected a small dataset of totally 61 students, who participated in the computer engineering module to perform a thorough analysis upon their learning activities. We combined the data from their applications, the log files from the online forum, their graduation grade and their performance in this module, and lastly the messages they posted in the corresponding online discussion forum.

The data preprocessing was implemented in Python (Rossum, 1995). The data were stored into an SQL database. The visualizations were created by using Tableau v. 9.3 (http://www.tableau.com/).

4. Results and Discussion

In this section we present the results of our experiments by visually presenting our findings in a number of graphs.

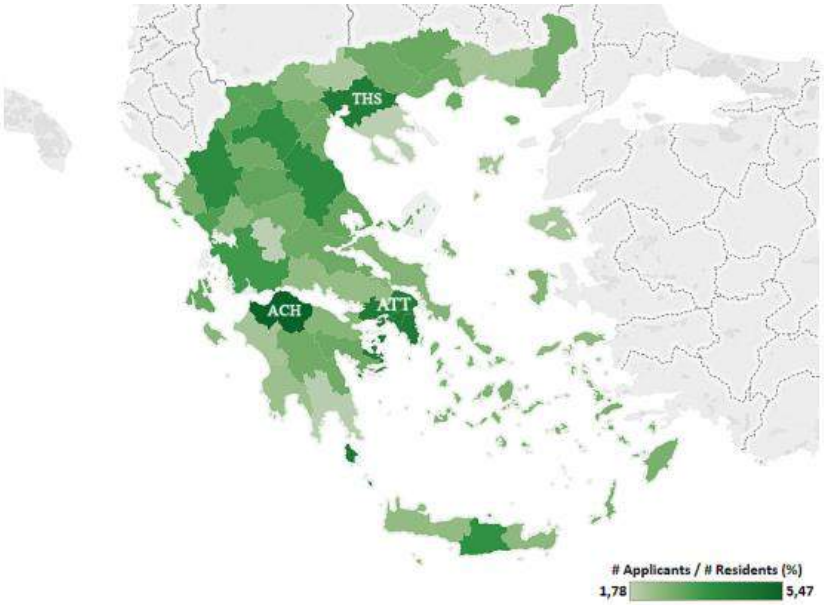

Figure 29 : Number of applicants per 100 people for each geographical region.

Figure 29 demonstrates the distribution of the applicants per region. Results are scaled per 100 residents. The data for the number of residents per region were retrieved from the (Eurostat Census Hub, 2016)[214]. The three regions with the largest percentages are the regions of Achaia (ACH, 5.47%), Attiki (ATT, 4.49%), and Thessaloniki (THS, 4.45%), which are marked with their region code. In these three regions, the three largest cities of Greece are located, with their size and number of residents having the following descending order; ATT, THS, ACH. The region with the largest percentage (almost 5.5%) is the region of Achaia (ACH), in

[214] The data are from the last census, which took place in 2011, and covers the period 2001–2011.

which the base of the HOU is located. Despite the fact that regions with large cities do have public universities or other private institutes, in which people can study, these regions have large percentages as well. On the contrary, peripheral regions usually have no access to higher education and as a consequence, we would expect larger percentages of residents applying to HOU from such regions.

Figure 30 presents the total number of applications per gender during that period. The number of applications has its maximum peak in 2004, while its minimum peak occurs in 2012. The decreasing trend during 2010¬–2012 is associated with the financial crisis that stormed in Greece during that period. Moreover, each year the university received more applications from female than male applicants. The percentage of female applicants was higher by approximately 11 units compared to the percentage of male applicants.

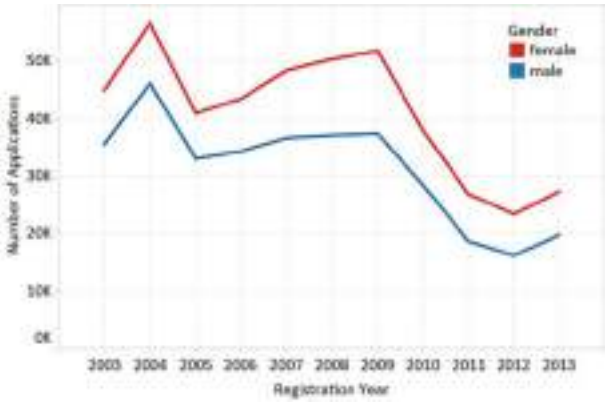

Figure 30: Number of applications per gender during 2003-2013.

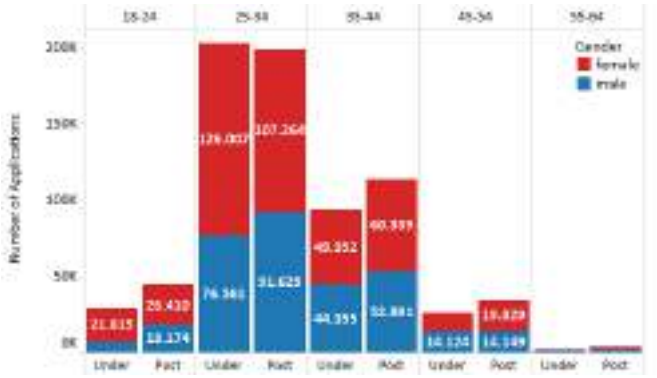

Figure 31: Number of applications per age group and educational level.

Figure 31 displays the number of applications made per age group, in combination with the educational level. We separated the applicants into five groups; 18–24, 25–34, 35–44, 45–54, and 55–64 years old. We chose to divide ages in these ranges because this setup has been used in demographics by numerous newspapers and advertising executives. The 25–34 age group held the majority of the applications. This can be explained by the fact that many students enter a public university to study a subject that quite often is not their first option. Then, after graduating or even after finding a job, they decide to either study or make postgraduate studies related to the subject that was their first option.

Figure 32 presents the number of events in the forum as recorded in the log files, separated by month. The activity of users reaches its peak early in the academic year, during November. Then, in December and January there is a descending trend, probably due to the Christmas holidays. During that period, students have free time to rest or study the material given to them. Therefore, less students participate in the forum. The forum activity then rises and has a small deviation from February until May. Exams take place during June and July, which is the beginning of summertime. Again there is a lower activity in the forum compared to the previous four months, which is almost equal to the one during January. This is probably because students focus on studying and have less time for participating in the online forum discussions.

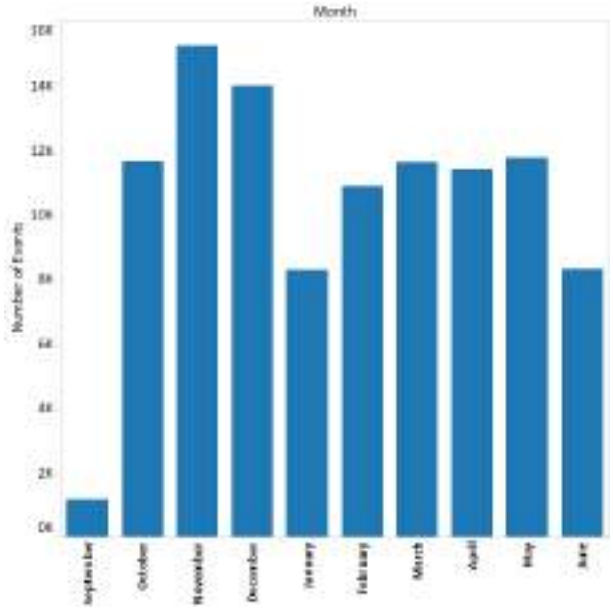

Figure 32: Number of events in the log file separated by month.

Figure 33 and *Figure 34* demonstrate the distribution of the number of events in the forum based on the hour of the day. In *Figure 33*, the colour indicates the average of the average grade of the students' assignments, which they have to submit during the academic year in order to pass the module. Firstly, we observe that outlier hours such as 03:30 and 05:30 have not only the fewest number of events, but also the lowest averages. This is probably because these students work a lot of hours during the day and do not have a lot of free time to study, let alone participate in the online discussion forum.

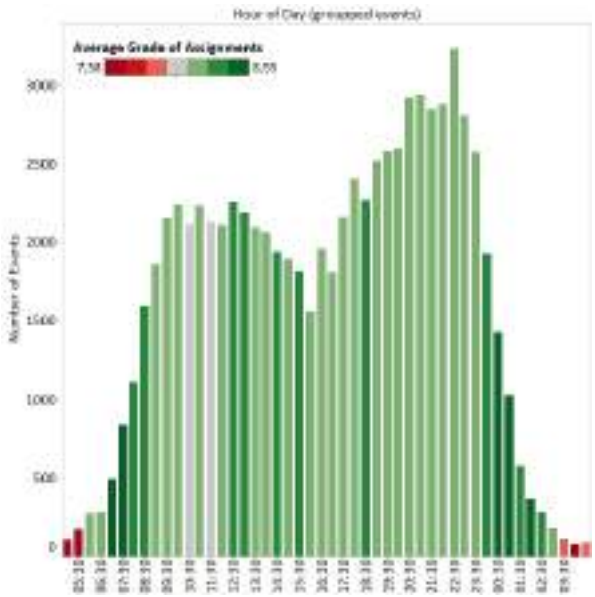

Figure 33: Distribution of the number of events in the log file based on the hour of day. Colour shows the average of the average grade of assignments of the students.

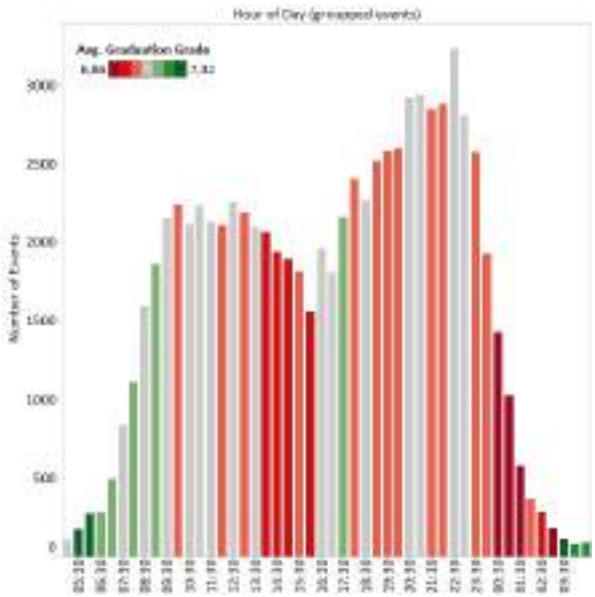

Figure 34: Distribution of the number of events in the log file based on the hour of day. Colour shows the average graduation grade of the students (from their previous education).

In *Figure 34*, the colour indicates the average graduation grade of the students from their previous educational organization, and follows the opposite trend as the one depicted in *Figure 33*. The very same hours of the day that appear with higher averages in *Figure 33*, appear with low averages of graduation grades in *Figure 34*. There is no clear explanation about this pattern.

Additionally, we observe that more events occur in the afternoon or later in the evening. This is not surprising considering that this is an online forum for a distance education module and that the majority of people work either from 9am to 5pm, or 9:30am-2pm and 5:30pm-9pm. Therefore, more students have free time to participate in the forum during afternoon or in the evening, depending on their working hours.

Figure 35 shows the distribution of the number of events in the forum based on the average grade in the students' assignments. The colour indicates the sex of the students. We observe that the bigger the average grade is, the more number of students and events there are. This may indicate that students who participate actively in the discussion forum achieve higher scores in their assignments. Figure 36 shows the distribution of the number of events in the forum based on the students' graduation grade from their previous educational organization. We note a normal distribution, with female students having a lower deviation. We observe that the highest number of events occur around the grade of 8. Students with a good graduation grade tend to participate more in the discussion forum compared to students with lower graduation grades.

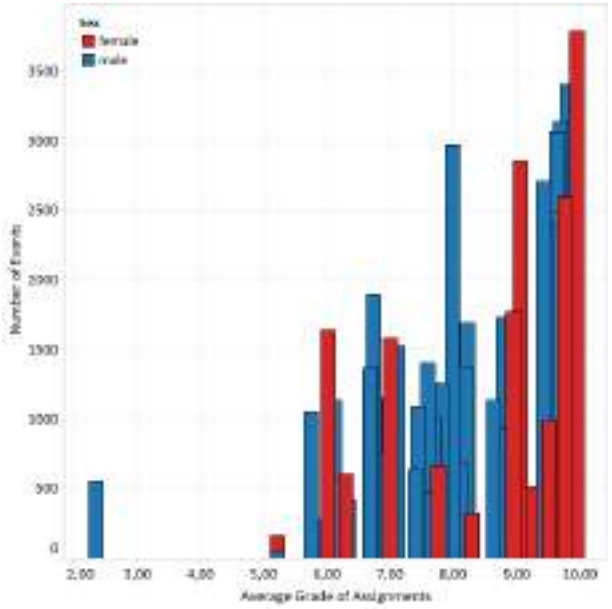

Figure 35: Distribution of the number of events in the log file based on the average grade in the assignments.

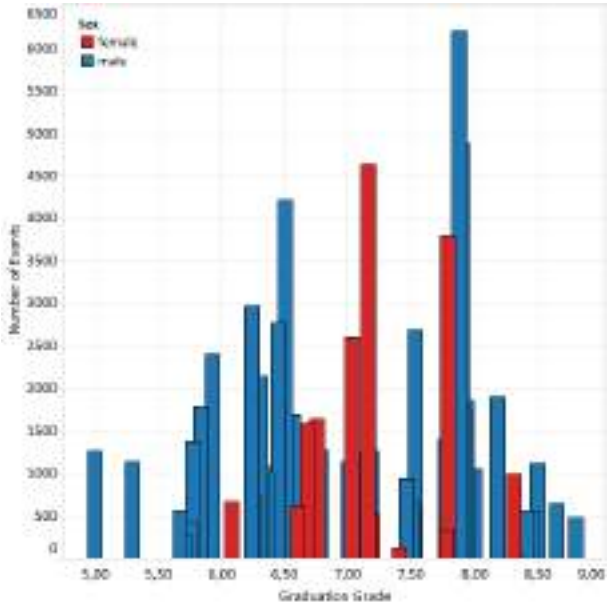

Figure 36: Distribution of the number of events in the log file based on the graduation grade of the students.

In *Figure 37* we depict the distribution of the undergraduate studies of the students. The colour shows details about the sex while the count shows the number of students for each Study.

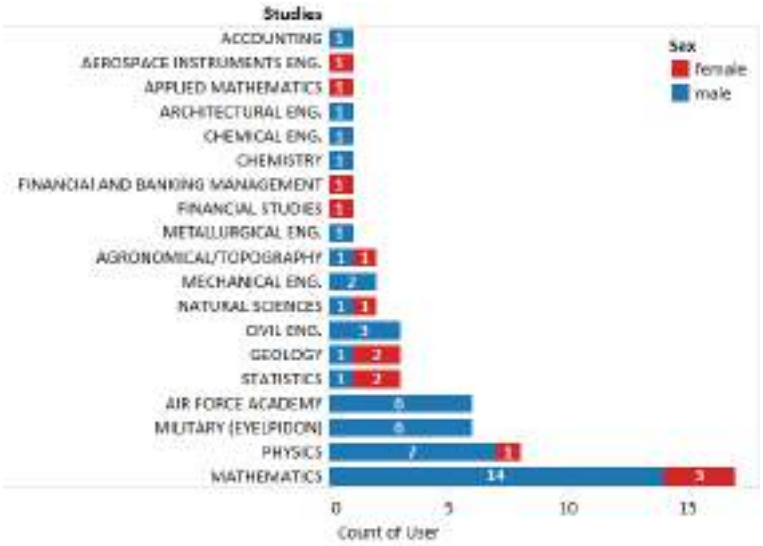

Figure 37: The distribution of students' undergraduate Studies.

Complementary data concerning students' behaviour in the forum are derived from the students' network that is built and presented in *Figure 38*. The students' network is based on their co-occurrence in the same thread and illustrates the interactions among them. Each node presents a student and each edge presents a correlation between two students. Students with higher levels of participation in the discussion forum are at the centre of the network. For instance, students with IDs 83117, 61122 and 83172 are located close to the centre as they have the highest frequency of participation in the forum being the most active participants. *Figure 38* demonstrates the active as well as the peripheral students allowing the tutor to have a visual description of the interactions among his/her students. Furthermore, labels id1, id2, and id3 refer to the course tutors, who seem to be an important factor of the interaction in the forum and they have a central role in the network.

Figure 38: Students' network illustrating interaction among them and their instructors.

5. Conclusions

In this paper we used data from the HOU so as to provide useful information not only for the HOU itself, but also for the tutors who teach the offered distance learning modules. For our first analysis we utilized a large dataset originating from the student applications of the decade 2003-2013. We expected larger percentages of residents applying to the HOU from remote regions but this was not really the case. From our analysis, we concluded that despite the fact that regions with large cities do have public universities or other private institutes, in which people can study, they have larger percentages of residents applying in the HOU.

External factors (such as societal and financial) affected to certain degree the number of applications received. The results indicated that more females applied to the HOU than males, which means that more females wanted to study the subject of their choice in a second-chance organization like the HOU or even to counteract against the gender wage gap. Additionally, people that are between 25 and 34 years old are more likely to apply in distance learning programs, such as the ones offered by the HOU, looking to acquire a better resume for future work placement.

We then utilized a smaller dataset, from students participating in the system software module, and combined the data from their applications, the log files from the online forum, their graduation grade from the previous educational institute, their performance in the module, and lastly the messages they posted in the corresponding online discussion forum.

The amount of spare time that students have is an important factor that can affect their progress even when participating in distance education programs, which are far more flexible than the conventional ones that require their physical presence. The activity in the online discussion forum reaches its maximum peak during the first four months. Tutors should bare this in mind and should at least place a small part of the most important subjects in the curricula during that period, as students appear to have more time and to be more eager to participate and discuss in the forum. This in turn could lead in understanding difficult subjects and concepts more easily.

The majority of the applicants, and therefore the majority of students, were between 25 and 34 years old. At this age, people have a job or at least a part-time job. Thus, students participate in the forum in later hours of the day. Tutors can utilize this information and answer the questions in a shorter time after they have been posted.

Lastly, participating in the forum can have a positive impact on the overall performance of a student. Questions are posted, discussions are made, and knowledge is shared, not only between students and tutors, but also among students. This process helps every participant learn something new or, through repetition, the maintenance of knowledge that one already owns.

References

Baker, R. S. (2010). Data mining for education. International Encyclopedia of Education (3rd edition), 7, 112-118.

Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1 (1), 3-17.

Berland, M., Baker, R. S., & Blikstein, P. (2014). Educational data mining and learning analytics: Applications to constructionist research. Technology, Knowledge and Learning, 19 (1-2), 205-220.

Byrne, T. C. (1989). The evolution of distance education. Calgary, Alberta: University of Calgary Press.

Cambria, E., & Hussain, A. (2012). "Sentic Computing: Techniques, Tools, and Applications". Springer.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. IEEE Intelligent Systems, 28 (2), 15-21. doi:10.1109/MIS.2013.30.

Duffy, T., Gilbert, I., Kennedy, D., & Kwong, P. W. (2002). Comparing distance education and conventional education: Observations from a comparative study of post-registration nurses. Association for Learning Technology Journal, 10 (1), 70-82.

European Statistical System, Census Hub, https://ec.europa.eu/CensusHub2/query.do?step=selectHyper Cube &qhc=false, 2001–2011, [Online; accessed 15-February-2016].

Hämäläinen, W., & Vinni, M. (2010). Classifiers for educational data mining. Handbook of Educational Data Mining Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 57-74.

Kagklis, V., Karatrantou, A., Tantoula, M., Panagiotakopoulos, C.T., & Verykios, V.S. (2015). A learning analytics methodology for detecting sentiment in student fora: A case study in distance education. European Journal of Open, Distance and e-Learning, 18 (2), 75-94.

Kagklis, V., Lionarakis, A., Panagiotakopoulos, C.T., & Verykios, V.S. (2016). Student Admission Data Analytics for Open and Distance Education in Greece. Submitted.

Kim, S.M., & Hovy, E.H. (2006). Identifying and Analyzing Judgment Opinions. Proceedings of the Human Language Technology / North American Association of Computational Linguistics conference (HLT-NAACL 2006). New York, NY.

Lotsari, E., Verykios, V., Panagiotakopoulos, C., Kalles, D. (2014). A Learning Analytics Methodology for Student Profiling Artificial Intelligence: Methods and Applications. Lecture Notes in Computer Science, Volume 8445, 300-312.

Manning, C., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT Press.

Ortony, A., Clore, G., & Collins, A. (1988). The Cognitive Structure of Emotions. Cambridge University Press.

Pal S. (2012). Educational data mining and learning analytics: Applications to constructionist research. International Journal of Information Engineering and Electronic Business (IJIEEB), 4 (2), 1-7.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

Pierrakeas, C., Xenos, M., Panagiotakopoulos, C., & Vergidis, D. (2004). A comparative study of dropout rates and causes for two different distance education courses. International Review of Research in Open and Distance Learning (IRRODL), 5 (2), 117-131.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, 40 (6), 601-618.

Stevenson, R., Mikels, J., & James, T. (2007). Characterization of the Affective Norms for English Words by Discrete Emotional Categories. Behavior Research Methods, 39 (4), 1020-1024.

Tiene D. (2000). Online discussion: A survey of advantages and disadvantages compared to face-to-face discussions. Journal of Educational Multimedia and Hypermedia, 9 (4), 371-384.

Turney, P. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. Proceedings of the Association for Computational Linguistics, 417-424.

Wen, M., Yang, D., & Rosé, C. P. (2014). Sentiment analysis in MOOC discussion forums: What does it tell us? Educational Data Mining.

White, M. (1982). Distance education in Australian higher education a history. Distance Education, 3 (2), 255-278.