Implementation of Big Data Analytics Tool in a Higher Education Institution

Tiago Franco[1(✉)], P. Alves[1], T. Pedrosa[1], M. J. Varanda Pereira[1], and J. Canão[2]

[1] Research Centre in Digitalization and Intelligent Robotics, Polytechnic Institute of Bragança, Bragança, Portugal
{tiagofranco,palves,pedrosa,mjoao}@ipb.pt
[2] JCanão, Viana do Castelo, Portugal
jose.canao@jcanao.pt

Abstract. In search of intelligent solutions that could help improve teaching in higher education, we discovered a set of analyzes that had already been discussed and just needed to be implemented. We believe that this reality can be found in several educational institutions, with paper or mini-projects that deal with educational data and can have positive impacts on teaching. Because of this, we designed an architecture that could extract from multiple sources of educational data and support the implementation of some of these projects found. The results show an important tool that can contribute positively to the teaching institution. Effectively, we can highlight that the implementation of a predictive model of students at risk of dropping out will bring a new analytical vision. Also, the system's practicality will save managers a lot of time in creating analyzes of the state of the institutions, respecting privacy concerns of the manipulated data, supported by a secure development methodology.

Keywords: Big data analytics · Web-based tool · Higher education · Data extraction · Machine learning

1 Introduction

The impact of data-based decision making is known as a management revolution. Today's managers benefit from a range of powerful indicators for making crucial decisions, relevant information that was not feasible to obtain years ago [12].

For the field of education, the researchers point to great potential in the use of educational data [3,8,15]. Through the use of new environments that enable learning, such as online communities, discussion forums, chats, Learning Management Systems, among others, a large amount of data is produced inside the educational institutions. This volume is so large that traditional processing techniques cannot be used to process them, forcing educational institutions that want to take advantage of data to explore big data technologies [17].

Á. Rocha et al. (Eds.): WorldCIST 2021, AISC 1365, pp. 207–216, 2021.
https://doi.org/10.1007/978-3-030-72657-7_20

Concerning the use of educational data, there is a clear difference between institutions that offer fully online training and institutions that have a more traditional education (they usually use online environments, but most of the education remains in person). For online education institutions, data studies are so refined and evolved that these institutions intend to personalize the process working at the individual level, seeking maximum effectiveness by adapting to the difficulties of each student.

For traditional educational institutions, the effective use of data for decision support is still uncommon, generally existing applications are limited to traditional statistical analyzes [6,16]. This shortcoming is related to several technical challenges that are necessary to deal with the multiple sources of educational data that these institutions store for years [7]. Besides, there is also the difficulty of changing the culture of managers, who are sometimes used to old technologies and are not always convinced that they will produce valuable results by investing in the area [5,14].

The US Department of Education [2] suggests and other studies agree [1,5], that implementation should be done progressively, through collaborative projects across departments, and throughout development, it will incorporate other sectors until they get the whole data management ecosystem [14].

Following this line of thought, this article describes a possible software solution that was implemented on the Polytechnic Institute of Bragança (IPB) in Portugal, with the aim of exploring educational data without the need to modify the systems already in operation at the institution.

The purpose of the system is to provide a set of analytical and predictive information about the institution's academic situation for management improvement. As a requirement, the software should have the ability to expand to other educational institutions and has an architecture that aims to guarantee the protection of sensitive data.

Seeking to take advantage of studies already started on improving higher education and the willingness to solve problems already known by IPB, we designed the software in modules. In all, three modules were implemented: the first consists of a machine learning model for predicting students at risk of dropping out; the second consists of a set of graphs and analytical tables that seek to translate the current teaching situation of all courses at the institution; the third module deals with the creation of dynamic reports from the extracted data.

This document is divided into four more chapters that will continue the explanation of the study. Section 2 refers to the developed architecture, detailing its components and connections; The third chapter addresses the reasons why we implemented the 3 modules and their individual characteristics; In Sect. 4, the results of implementing the web tool in the educational institution are presented and discussed; Sect. 5 reports the main conclusions and scientific contributions that we have drawn from this study

2 Architecture

The task of implementing a big data communication structure involving several systems already in operation is not a trivial task. Initially, we seek to deal with the initial phases of the big data analysis process, which are in order of precedence: acquisition, extraction, and integration [3]. For that, we defined that it would be necessary an ETL system (Extraction, Transformation, and Loading) coupled to a server with access to the institution's active systems. Although its name is already very clear about its purpose, an ETL system can also be in charge of other activities that are necessary for the exchange of information between the data sources and the requesting system [18].

The next task in defining the architecture was to create a data collection group component, and its connection with all the other components in order to execute the modules defined. From this, we created the diagram in Fig. 1 to represent the architecture with its components and connections.

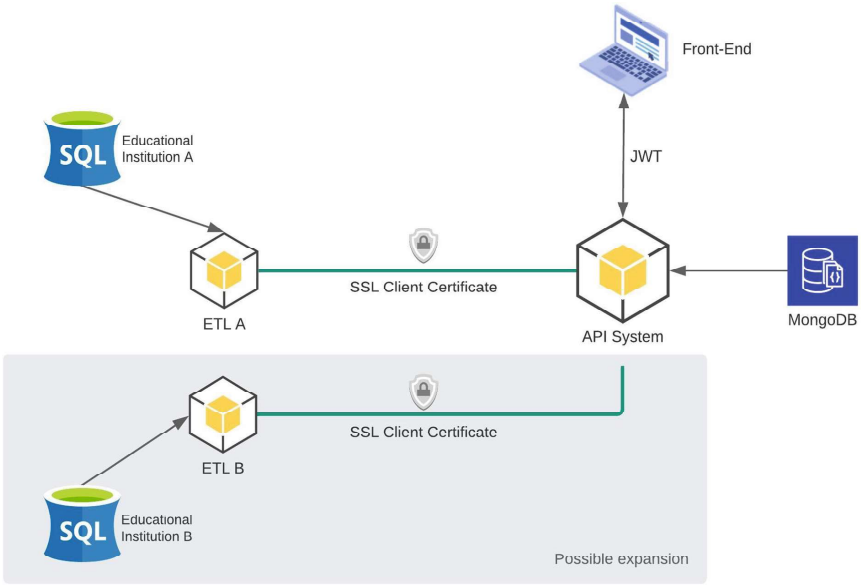

Fig. 1. Structure of the proposed architecture

The API System is a service modeled with the REST architectural style to provide the solution's main feature set. In this component, we manage the users, the triggers for the extraction tasks, process the extracted data, and execute the proposed modules.

This component depends on a secure direct connection to the MongoDB database, where the stored and processed data is located. With this approach,

it is possible to perform complex analyzes to respond to requests from the front end in a few seconds.

The ETL, as already mentioned, is the component that is installed next to the active systems that are intended to extract the data. This component is the only one in the architecture that accesses the educational institution's databases. This strategy makes it possible, when necessary, to encrypt sensitive data, ensuring that they are never stored in MongoDB and protecting them from malicious attacks. The component authentication to the API uses TLS/SSL client certificates to perform secure authentication and prevent dictionary attacks and/or brute force.

During the development of the architecture, we seek to incorporate the current concepts of microservices, big data analytics, and web development. To ensure cybersecurity, we follow the OWASP Secure Coding Practices manual [9]. Also, due to the fact that many users can process personal data, it was decided to consider the mandatory technical requirements defined in the Resolution of the Council of Ministers n° 41/2018 [13].

Thinking about a more commercial aspect of the software, it is possible to expand the application by adding several ETL components in other institutions. As data protection is already guaranteed by ETL on the institution side, the scope of the API component is even greater, as it can group multiple data sources to produce more powerful reports and machine learning models.

3 Implemented Modules

Bearing in mind that the software would not be just a prototype, after the defined architecture, we started to implement the system core. This core included multiple access levels, logging, the establishment of connections between components, the administrative panel, and automatic scheduling of extraction tasks.

To facilitate the audit of extractions, we have created a page to manage them. The main features of this page are: view the ETLs logs sent daily about their operation; check if the extraction tasks were successfully performed and their processing time; manage the frequency of automatic tasks and create an extra sporadic extraction.

With the software core assembled, we compiled a set of analyzes that had passed an initial phase of discussion and could be really implemented, resulting in the following modules.

3.1 School Dropout - Machine Learning Model

One of the main metrics to measure the quality of education in a higher education institution is the student dropout rate. Constantly government entities that deal with education, update their goals on recovery from dropout and seek new solutions to combat it [4].

Likewise, the educational institution that developed this study is also looking for new ways to improve this index. Among the proposed alternatives, the article

[10] proved to be promising for implementation, by validating the capacity of the institution's educational data in predicting possible dropouts through a machine learning model.

For its operation, the model developed takes into account data from three active systems in the institution. The first refers to the student's enrollment information and their status in the course, such as grades, number of approved subjects, academic years. The second system is the attendance record in the classroom. The last refers to the logs generated by the institution's Learning Management System (LMS).

The model requires a weekly snapshot of the student's current situation. With this data, the Random Forest algorithm is used to classify between dropout and non-dropout, using data from the last 4 years of the same week. Afterward, the number of times the student has been classified as dropout during the year is counted as a percentage. This percentage is called the critical rate, indicating students with the highest values as likely dropouts.

In order to implement the visualization of the study results, three pages were implemented. The purpose of the first page is to provide a set of comparisons on the critical rate between schools and courses at the institution. The second page developed, provides a report of the information collected and the results of the model of a specific student.

Seeking to take preventive measures, the third page was designed to support a call system for students at risk of dropping out. In this way, a specialized agent can get in touch with the student and discuss what problems he is facing and possible ways to deal with them. Afterward, the agent can record the reasons for the abandonment (if confirmed) and a report of the conversation.

3.2 Numerical Analysis

Analyzing the difficulties of the educational institution, we observed that an analytical report was produced every year. It's a report used by course coordinators to manage their courses and disciplines. The main analytical components consisted of the sum of the number of students enrolled, evaluated, and approved in each discipline of each course. Our initial idea was to transform this report into dynamic elements on a new page in the application.

In this way, we created a task that could extract all the data related to the material records of a requested year. This extraction resulted in the records being stored in MongoDB, already with the sum of each discipline, that is, it did not bring the students' personal information, only the number of how many students had that record.

With the data available to the front-end, the next step was to develop and increment the elements of the base report. Finally, we personalize the pages by access levels, covering more managers who may use the software.

3.3 Dynamic Reports

Unlike the other modules, the dynamic report was designed in the second stage of implementation. The idea arose from the need to always contact a developer to provide a new analytical element, even if the necessary data has already been extracted.

As the data stored in MongoDB has already been processed and did not contain sensitive data, we developed a tool that allows users to create their own queries to build graphs and tables. In this way, we build a sequence of steps necessary to build one of these elements.

After the user chooses the dataset they want to work with, the first step is to create the filters. Filters make it possible to simplify SQL operators to extract a subset of the original data. The second step is to select the fields that should be returned after the query. The third step is optional, its function is to create fields from calculations in other fields. As an example, it is possible for the teacher to experiment with different formulas to define the composition of the final grade of a discipline.

The fourth step depends on the user's objective, being possible to follow two paths. The first path aims to export a table, having the only functionality of ordering it. The second way is to create a chart. Thus, the fourth stage aims to choose which groupers will compose the chart. Finally, the fifth stage refers to the aesthetic settings of the chart, such as the names of the axes and the title.

With the definition of these elements, we developed a page for the construction of the report effectively. Similar to a text editor, the page has the advantage of being able to import the dynamic elements developed and resize them. With that, different agents of the institution can work with the educational data and create their own reports.

4 Results and Discussion

Following the completion of the software implementation, our work was directed to diagnose the system's impacts on the institution. In this session, a compilation of the impressions found by some managers who used the platform will be presented, comparisons between the results of the machine learning model of the original article [10] and the software developed and particularities found that are worth discussing.

4.1 Analytical Components

A significant improvement was noted in the practicality of creating a numerical analysis of the institution, mainly due to the fact that previously it was necessary to open multiple PDFs to view the same data that is now on a single dashboard. In addition, the amount of information that could be found was expanded, offering a new set of analyzes on teaching.

Similarly, the dynamic reports complement the numerical analyzes, offering the possibility to build an analysis from scratch. In a simplified way, the result

of these components is similar to typical business intelligence software, without the huge set of tools that we usually find, but with the difference of not having the need to import the data into the system, since the data is already available in the software for easy handling.

Figure 2 shows some of the components that can be found on the analysis page. The data displayed is emulated and does not reflect the reality of the institution.

Fig. 2. Screenshot of the analysis page

The first four boxes contain information on the averages of the number of students enrolled, evaluated and approved in the last 4 years of a chosen course. The following two graphs show the decomposition of these averages in the historical form of recent years. At the end of the figure, two graphs are presented referring to enrolled students and dropouts, the first informing the real values and the second in percentage.

4.2 Machine Learning Model

The first tests of the developed module sought to identify whether the extracted and processed data obtain the same result as the article that suggests the model [10]. The results indicate that the behavior of the model was similar to that expected, with a small increase in the number of students selected as dropouts,

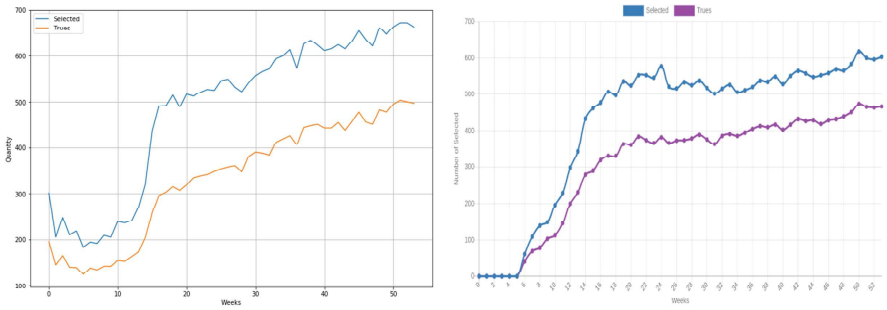

Fig. 3. Performance comparison of machine learning models

but justified by the increase in the number of students enrolled in that year. With that, we can say that the model was successful in its initial purpose.

The two graphs in Fig. 3 illustrate the performance of the predictive model of students at risk of dropping out of the institution in the 2016–2017 academic year. The graphic on the right was taken from the original article and on the left is the product of the implemented software. In both graphs, the upper blue lines represent the number of students that the algorithm selected as dropouts, the bottom lines represent the number of students who actually dropout. Thus, the difference between the lines is the error that the algorithm presents.

Comparing the two graphs, we can see a great divergence between weeks 1 and 20 and a decrease in the distance of the lines in the second graph. The difference in the first few weeks was caused by two factors. The first factor was the adjustments to the semester start dates and the frequency calculation. We found that the old algorithm could incorrectly calculate attendance if the student enrolled late.

The second factor was the addition of two new attributes that improved the model's performance. The first attribute added refers to the number of days it took the student to start the course and the second attribute refers to whether the student works or not during the course. This improvement is reflected throughout the year and can be seen in the decrease in the difference between the two lines of the second graph compared to the first graph.

4.3 Data Extraction

For the predictive model, we initially processed the data for the academic years from 2009 to 2018 in a development environment, leaving us to extract only the data from 2019. After correcting small bugs found, we noticed that the tool took an average of 30 to 40 min to extract all data for a requested week and 2 to 5 min for processing the model.

The differences between the times basically depended on two factors. The first is related to the week in which the data is to be extracted, since at the beginning of the year the amount of information generated by the students was

smaller, resulting in less processing time. The second refers to the time we do the extraction since the server could be overloaded.

In mid-day tests, the most critical moment of use for students, it could take twice as long. With an automatic task scheduler developed, we set up so that the extractions only occur at dawn, a time that had less use of the systems, minimizing the negative impacts of the software on the institution.

As the data required for the analytical module did not require much processing and are not bulky, the extractions took less than 1 min to be successful.

5 Conclusion

In this study, the following steps were described for the implementation of big data analytics tool in a higher education institution. Analyzing the results presented, we can conclude that our system can positively contribute to help the academic managers to improve the success of their students. We can highlight the time savings that the system's practicality has brought to the institution's managers.

Generally, most managers at an educational institution are also teachers, needing to divide their time between multiple tasks. The system presented allows managers to pay more attention only to data analysis since they no longer need to build a series of analytical components. In addition, if necessary, the system allows more complex analyzes to be made, exporting the data, streamlining the process with initial treatment, and covering more external tools.

As described in Sect. 3, two of the three modules implemented were already developed or at least quite discussed. This fact simplified several hours of our work and allowed us to refine the collected studies. This reality can be found in many educational institutions, with papers or mini-projects already developed internally but with several constraints to put it into practice. Once the proposed architecture is developed, it does not require complex refactoring to create a new data extraction in a safe way. Therefore, the inclusion of new modules is encouraged without prejudice to existing ones.

Another positive point observed was in the contact pages of students at risk of dropping out of school. In addition to fulfilling its purpose of enabling preventive contact with students at risk of dropping out, through the record of contacts a history of reasons that led students to dropout was created. In this way, it is possible to analyze the most worrying reasons in order to create a plan of preventive actions by the institution.

A similar strategy that supports the development of the architecture and the whole software can be seen at [11]. The HESA project aims to repeatedly collect a series of data from more than 250 educational institutions in the United Kingdom, creating a government ecosystem of educational data management. Undoubtedly it is an inspiring project for other countries, but this reality depends on a series of facts for its success, besides the monetary, the difficulty of its cooperation of the gigantic team involved. The study presented here, suggests an alternative for institutions that do not intend to wait years for this to happen and already reap the rewards of using data for decision making.

Acknowledgment. This work was supported by FCT - Fundação para a Ciência e a Tecnologia under Project UIDB/05757/2020 and Cognita Project (project number NORTE-01-0247-FEDER-038336), funded by the Norte 2020 - Norte's Regional Operational Programme, Portugal 2020 and the European Union, through the European Regional Development Fund.

References

1. Ali, L., Asadi, M., Gašević, D., Jovanović, J., Hatala, M.: Factors influencing beliefs for adoption of a learning analytics tool: an empirical study. Comput. Educ. **62**, 130–148 (2013). https://doi.org/10.1016/j.compedu.2012.10.023
2. Bienkowski, M., Feng, M., Means, B.: Enhancing teaching and learning through educational data mining and learning analytics: an issue brief, pp. 1–60 (2014)
3. Bomatpalli, T.: Significance of big data and analytics in higher education. Int. J. Comput. Appl. **68**, 21–23 (2013). https://doi.org/10.5120/11648-7142
4. European Commission: Education and training monitor 2019 - Portugal (2019)
5. Daniel, B.: Big data and analytics in higher education: opportunities and challenges. Br. J. Edu. Technol. **46**(5), 904–920 (2015). https://doi.org/10.1111/bjet.12230
6. Daniel, B.: Big data in higher education: the big picture, pp. 19–28 (2017). https://doi.org/10.1007/978-3-319-06520-5_3
7. Daniel, B., Butson, R.: Foundations of big data and analytics in higher education. In: International Conference on Analytics Driven Solutions: ICAS2014, pp. 39–47 (2014)
8. Dutt, A., Ismail, M.A., Herawan, T.: A systematic review on educational data mining. IEEE Access **5**, 15991–16005 (2017). https://doi.org/10.1109/ACCESS.2017.2654247
9. T.O. Foundation: OWASP secure coding practices quick reference guide (2010)
10. Franco, T., Alves, P.: Model for the identification of students at risk of dropout using big data analytics. In: INTED2019 Proceedings, 13th International Technology, Education and Development Conference, IATED, pp. 4611–4620, 11–13 March 2019. https://doi.org/10.21125/inted.2019.1140
11. HESA: About hesa. https://www.hesa.ac.uk/about. Accessed 01 Nov 2020
12. Mcafee, A., Brynjolfsson, E.: Big data: the management revolution. Harvard Bus. Rev. **90**, 60–68 (2012)
13. de Ministros, C.: Resolução do conselho de ministros n° 41/2018. Diário da República n° 62/2018, Série I—28 de março de 2018, pp. 1424 – 1430 (2018). https://data.dre.pt/eli/resolconsmin/41/2018/03/28/p/dre/pt/html
14. Murumba, J., Micheni, E.: Big data analytics in higher education: a review. Int. J. Eng. Sci. **06**, 14–21 (2017). https://doi.org/10.9790/1813-0606021421
15. Romero, C., Ventura, S.: Educational data mining: a survey from 1995 to 2005. Expert Syst. Appl. **33**, 135–146 (2007). https://doi.org/10.1016/j.eswa.2006.04.005
16. Shacklock, X.: The potential of data and analytics in higher education commission (2016)
17. Sin, K., Muthu, L.: Application of big data in education data mining and learning analytics-a literature review. ICTACT J. Soft Comput.: Special Issue Soft Comput. Models Big Data, 4 (2015)
18. Trujillo, J., Luján-Mora, S.: A UML based approach for modeling ETL processes in data warehouses. In: Song, I.Y., Liddle, S.W., Ling, T.W., Scheuermann, P. (eds.) Conceptual Modeling - ER 2003, pp. 307–320. Springer, Heidelberg (2003)