NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Outlier detection and characterization for students

Gabriel Avezum

Dissertation

presented as partial requirement for obtaining the Master Degree Program in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Outlier detection and characterization for students

by

Gabriel Avezum

Dissertation presented as partial requirement for obtaining the Master's degree in Advanced Analytics, with a Specialization in Data Science

Supervisor: Prof. Doutor Roberto André Pereira Henriques

January/2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading
to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon,10/07/2023

ABSTRACT

Nowadays, web-based educational systems are being used as an essential tool to the learning process and with them the universities are capable to collect data from the students and build studies to understand their behavior. In these studies, the outliers' students are not the main topic, or they are removed because of their extreme behavior, so this thesis focuses on detecting the outliers and understanding their behavior. Focusing on these students a methodology is proposed to find two clusters using outliers' techniques on their grades and Moodle usage, the main goal was to detect the subjects with a high volume of usage in Moodle and with a bad performance on the final grade, this analysis was done for each subject that the student did. In the end we detect a total of five students that require an intervention by the university to understand how their performance could improve or if they are having some troubles in the use of the online tool. The methodology proposed shows an efficient and easy way to find the students and could be replicated easily for other universities and be implemented as an active tool to assist the university.

KEYWORDS

Outlier 1; Educational data mining 2; Moodle 3

Sustainable Development Goals (SGD):

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

EDM	Education data mining
LA	Learning Analytics
KNN	k-nearest neighbour
CBLOF	Cluster-based local outlier score
LOF	Local Outlier Factor
DBSCAN	Density Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise
ECOD	Empirical Cumulative distribution-based Outlier detection

1. INTRODUCTION

Data science impacts many aspects of our life. It has been involved in many areas, and education is not an exception (Novoseltseva, 2022). Nowadays, web-based educational systems have been rising exponentially and they led us to store a huge amount of potential data from multiple sources with different formats and with different granularity levels (Romero & Ventura, 2014). Educational data can be gathered from various sources with different formats.

When using any these types of tolls, the educational data is usually gathered from three sources, Log files (records all students-system interaction), Quiz/test (information about quiz/test utilization) or Portfolio (information about students) (Romero & Ventura, 2014). This study is focus on extracting knowledge and insights from a data extract from the first two sources that are recorded using Moodle.

The most studies on education data are focused on studying the whole data of learners and extract information from them, but only a few focus on the most extreme events and usually these students are removed from the general study. This study proposes a method and an approach to analyze this public.

This study is divided in six parts, where the first is a brief introduction about the topic of analysis on student data, the second is the literature review that was done on outlier's techniques and the most recent publications on the use of outliers detection on student data, the third is described the methodology used to extract the desired results from the data, the fourth is the presentation of the results, the fifth is the conclusion of the study and finally the last chapter is the recommendation for future works and limitations in this work.

1.1. EDUCATIONAL DATA MINING AND LEARNING ANALYTICS

With this amount of data been gathered a challenge arise that is the exponential growth of education data and to transform this data into insights that will help the university to take a decision to improve the learners study process and the educators instructing (Baker, 2015).

A long the years of research in educational field two different communities emerged in with a joint interest in how educational data can be exploited to benefit education and the science of learning (Baker & Inventado, 2014), these areas are learning analytics (LA) and Educational Data Mining (EDM).

EDM was created to extract knowledge from these data using data mining techniques (Romero & Ventura, 2013). So, the researchers in EDM started to use all the available techniques in data mining to extract information from the educational data, such as visualization, clustering, outlier detection, relationship mining, causal mining, etc. (Romero & Ventura, 2020). So, the results are obtained by extracting insights from data that are already collected and the main techniques used are derived from the machine learning or data mining area.

LA can be defined as the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs (Lang, Siemens, Wise, & Gasevic, 2017. In LA to obtain some of the results the researcher

needs to apply some test or different treatments to the students/teachers and compare the results before and after. In this case most techniques used are from the statistical area.

EDM and LA involve the implementation of techniques from various disciplines, such as visual data analytics, recommender systems, information retrieval, etc. EDM and LA can be described as a combination of three fields: education, statistics, and computer science (Romero & Ventura, 2020).

In the view of Baker and Iventado (2014) there is considerable thematic overlap between EDM and LA. The overlap and differences between the communities is organic, developing from the interests and values of specific researchers rather than reflecting a deeper philosophical split or antagonism (Baker & Inventado, 2014).

Data mining has already been successfully applied to other areas or domains such as business, bioinformatics, genetics, medicine, and so forth (Romero and Ventura, 2013). Although algorithms used in both areas are similar, the final goal are different (Romero and Ventura, 2010). For example, an insurance company can have an objective to predict the churn of a customer while an e-commerce is to increase profit. So, for education we need to set objectives and stakeholders.

Since EDM/LA has a wide range of possible objectives Romero and Ventura (2013) proposed to classify these objectives according to the stakeholder. These classifications can be one of the following categories:

- Learners: to support the learners, to provide adaptive feedback or recommendations to learners, to improve learning performance, and so on

- Educators: To find a pattern in their students' learning processes and use this to modify their own teaching methods, and so on

- Researchers: To develop and compare techniques to be able to recommend the most useful one for each specific educational task or problem, and so on

- Administrators: To understand how to optimize the institutional resources (human and material) and their educational offer, and so on (Romero & Ventura, 2013).

1.2. KNOWLEDGE PROCESS

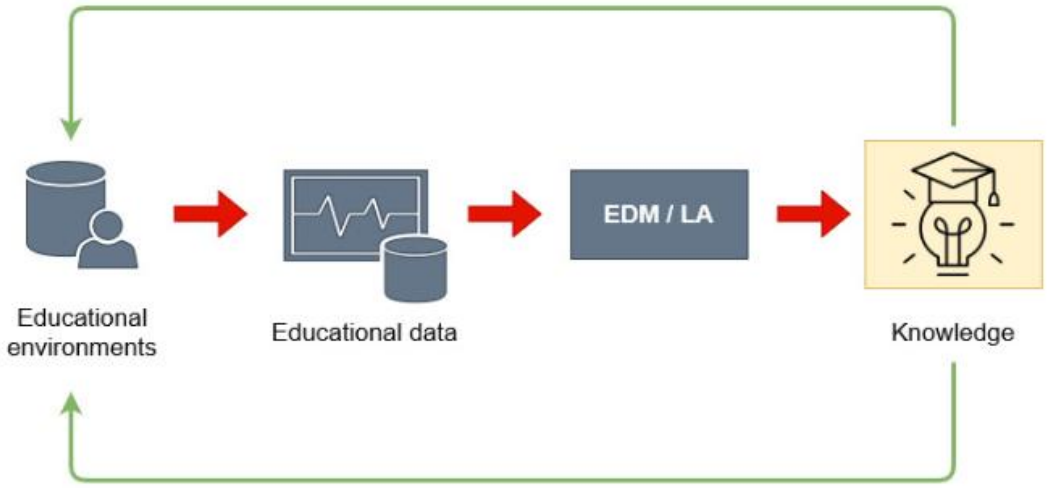

Figure 1.2.1 The Educational data mining/Learning analytics knowledge extraction process (Novoseltseva, 2022).

Figure 1.2.1 demonstrate how the knowledge extraction process works, first in the educational environments, Moodle in our case, while the students interact their data is recorded and compelling all the student's information the educational data is created. Finally, to extract he knowledge from the data the EDM/LA techniques are used. Once this ultimate step is done this can be an input for news studies. This process is designed to reach a specific objective.

In this process, the goal is not just to turn data into knowledge, but also to filter mined knowledge for decision-making about how to modify the educational environment to improve student's learning (Romero & Ventura, 2020).

1.3. APPLICATIONS

After being able to define our objectives and the process to extract knowledge it is needed to apply the necessary methods. Most traditional data mining techniques including but not limited to classification, clustering, and association analysis techniques have been already applied successfully in the educational domain (Baker, 2010).

Some EDM/LA techniques are as follows:

•Prediction: A supervised learning method that predicts value, class, or probability of the desired outcome through a combination of other variables (Novoseltseva, 2022): For example, dropout prediction, score prediction.

•Casual mining: the goal is to find whether one event was the cause of another event (Spirtes et al. 2000) distinguished from prediction in its attempts to find not just predictors but actual. It is different from prediction because it tries to find a causal relationship (Baker & Inventado, 2014). For example, finding features of students' behavior evoke dropout.

•Clustering: Identify groups of instances that are similar in some respects (Romero & Ventura, 2013). For example, merging students to clusters based on similar performance or learning strategies.

•Association rule mining: the goal is to create if-then rules that will return a value if a set of values is found(Baker & Inventado, 2014). For example, IF student is frustrated THEN the student frequently asks for help.

• Social network analysis: to study and evaluate the relationships between students, teachers or universities using a network information (Romero & Ventura, 2013).

• Text mining: to extract high-quality information from the text (Novoseltseva, 2022). For example, analyzing the contents of a discussion forum.

• Outlier detection: to discover data points that are significantly different than the rest of data (Romero & Ventura, 2013).

1.4. RESEARCH QUESTIONS AND OBJECTIVE

Even though the data mining techniques are well used in the educational field, not all have the same focus (Aldowah et al., 2019). One of these techniques is the outliers that are barely explored in education literature (Novoseltseva, 2022). In the topic about anomaly detection the main research is focused in using supervised learning and in detecting only the negative behaviors of the students (Guo et al., 2022). The main algorithms used on the research is necessary to remove the outliers to continue the study, so these students are treated as uncommon events and excluded from the study, and we can't address specific insights. When we find who are the outliers and why the university could help the students to address some gap on their learning, one example is the student who uses a lot of the resources on Moodle but has the minimum grade to pass the final exam.

Unexpected phenomena occur from time to time, leading to undesired consequences. For example, students who drop out of university can face social stigma or less job opportunities and a higher chance of involvement with the criminal justice system (Amos, 2008).

This dissertation is focused in using the learner's data to explore the anomaly detection using outlier techniques to identify who are the students that have difficulties on the learning. In this study is important to remember that the outlier can have two directions, can be an extreme high or low value, because the direction will impact the further analysis.

To accomplish this, we need to find two types of students, the first are the ones who is an outlier on the grade and is a low value and at the same time is an outlier on the Moodle usage, on the latest can be on both directions. The second type of students is the ones that are not an outlier on the grades but is an outlier on the Moodle on both directions. With this the following research question was elaborated:

•Who are the students with difficulties on learning?

My research objective will be:

•To characterize the outliers after the end of a semester and the secondary objectives are:

a. SRO1: To identify the outliers by grades,

b. SRO2: To identify the outliers by Moodle usage

c. SRO3: Compare both groups and find the match

d. SRO4: Find the students that are outliers on Moodle, but they are not on the grade

e. SRO5: Count how many times a student has the previous behavior.

f. SRO6: Assign the students with high appearance as the students with difficult on learning

2. LITERATURE REVIEW

In this chapter, the theorical bases and related research will be presented. Initially the most common and up to date techniques of outlier detection are going to be presented, then it will be discussed the usage of outlier detection in another fields of study and finally it will be introduced the studies in education that uses outliers and anomaly detection.

2.1. OUTLIER DETECTION TECHNIQUES

In the recent years the technologies are more developed and with that the amount and complexity of data collected increases. Resulting in appearance of patterns that do not has the expected behavior or that differ significantly from most data objects in certain ways (Novoselteva, 2022). This conduct is known as outliers, identifying, and analyzing these unexpected patterns is important in vairous fields as they can provide valuable information. For example, unexpected geological activity could be an indication of an earthquake or tsunami. The process of finding outliers in data is called outlier detection (Novoselteva, 2022).

Grubbs (1996) gave the first definition of outlier: an outlying observation, or outlier, is one that appears to deviate significant from other members of the sample in which it occurs. Hawkins, 1980 defined an outlier as an object that is so different from the others in the same dataset that will raise a concern that it was generated by a different mechanism. Generalizing outlier is a data point that lies significantly outside the range of the rest of the data in a dataset.

Outlier detection is a widely studied topic with important applications in data analysis. Many systems, such as credit card transaction monitoring, use outlier detection to identify patterns in data that do not conform to the expected distribution or behavior. This can help detect malicious activity, for example. In finance, traders may use outlier detection to identify unusual market movements and make trading decisions. There are many different methods for detecting outliers, including statistical methods, machine learning algorithms, and visualization techniques (Novoselteva, 2022).

So, techniques were created to identify these observations, because it will influence the analysis of the dataset if they remain on the dataset, or it is desired to study these observations. In the literature, a diverse set of outlier detection techniques and their classifications have been discussed (Novoselteva, 2022), but the most common are statistic-based, distance-based, density-based, cluster-based techniques and tree-based techniques (Hodge, et al, 2004).

2.1.1. Statistic-based techniques

Statistical outlier detection approaches are usually classified into two different groups – parametric and non-parametric methods. The parametric techniques assume that the data has a known distribution. The nonparametric techniques do not have any assumption about distribution (Novoselteva, 2022).

One of the most used parametric techniques of a statistical method for outlier detection is the Z-score method (Grubbs, 1969), which calculates the number of standard deviations a data point is from the mean. Data points that are more than a certain number of standard deviations away from the mean are outliers. This method assumes that the data is normally distributed.

Beside the Z-score method there are others that are used like the Rosner test (Rosner,1983), Dixon test (Gibbons, Bhaumik, Aryal, 2009) and Gaussian mixture model (Tang, Yuan, & Chen, 2015)

The most used non-parametric outlier factor is the Interquartile range (IQR), which is calculated by the difference between the 75^{th} and 25^{th} percentiles of the data and then a constant value is multiplied, and the result is added in these percentiles, all the data outside these bounds is considered an outlier.

The techniques describe above only works for univariate data, so in the case of a multivariate outlier these methods will not provide a sound outlier identification.

Novel models are being designed in the non-parametric space to fill the gap in the outlier detection in the multivariate domain, like the Empirical Cumulative distribution-based Outlier detection (ECOD) (Li, et al, 2022).

These techniques are easy to implement and mathematically sound, however the parametric require that the data follows a known distribution, and, in many cases, this is not possible.

2.1.2. Distance-based techniques

Distance-based methods for outlier detection utilizes the distance between each pair of objects in the dataset. These methods define a neighborhood of each data point within a certain distance threshold. Depending on the distance threshold definition, outliers can be identified as:

- the object *O* is an outlier, if at least *p* of objects lies greater than distance *D* from *O* (Knorr, e.t al, 2000)
- An object *O* is considered an outlier if it is among the top n objects with the highest average distance to their k-th nearest neighbor. (Hautamaki, Karkkainen, Franti, 2004)

2.1.3. Cluster-based techniques

Cluster-based methods for outlier detection operate by applying a clustering algorithm to the data and then identifying data points that either belong to a small or sparse cluster, or do not belong to any cluster at all. These data points are outliers (Novoselteva, 2022).

In the step of the clustering the objects any clustering algorithm can be used, but the most used is the k-means, which is usually faster than hierarchical approaches.

One of the algorithms is the Cluster-based local outlier score (CBLOF) (He, Deng, 2003) and another famous one is the Density Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise (DBSCAN) (Ester, Kriegel, Sander, Xu, 1996).

Cluster-based local outlier factor labels an observation as outlier according to two factors, the first the size of the cluster the observation belongs to and the second is the distance between the observation and its closest cluster (He, Deng, 2003).

2.1.4. Density-based techniques

Density-based outlier detection is a method of identifying points in a dataset that are significantly different from the others in terms of the local density of points around them. This can be done by calculating the density of points in the vicinity of each point, and flagging points that have a much lower density than the others (Novoselteva, 2022). This method is useful for identifying anomalous observations that may be due to errors in data collection or measurement or may represent rare or unusual events.

The goal is to compare the density around an object with the density around its local neighbors. To accomplish this goal a division between these two densities is done, so when an object is inline the value will be close to one, because the density around this object will be like the density around its neighbors, while the density around an outlier object is significantly different from the density around its neighbors, so the value would be very different from one (Novoselteva, 2022). The method most used is the Local Outlier Factor (LOF) (Breunig, Kriegel, et al, 2000)

2.1.5. Tree-based techniques

Tree-based outlier detection is based in the idea that an anomaly is so different from other observation that a single split of a tree is possible to isolate this point to the other points. The method starts building a tree where each leaf will have exactly one observation, after each observation has a score based how deep is on the tree, so observations that appear in the top of the tree will present a higher score (Liu, et al , 2008).

The most common technique in the class is the IForest, where the user just needs to select the percentage that will be considered as an outlier then the algorithm will give a score for each observation on the dataset and the top percentage will be defined as outliers.

2.2. OVERVIEW OF OUTLIER DETECTION APPLICATIONS

Outlier detection is a well-studied area with many applications in various domains. Resulting in a large and diverse literature on outlier detection algorithms. It is important to consider domain

knowledge when using these techniques, particularly when dealing with contextual anomalies and specific outlier scenarios. Outliers may also be referred to as intrusions, fraud, exceptions, misuses, novelties, or irregularities depending on the specific application domain.

In the area of Fraud detection, the outlier technique is a tool used to detect malicious activity, which usually appears in the bank systems, credit card systems or insurance agencies. Abdallah, Maarof and Zainal (2016) described a fraud detection system that use behavioral profiling methods, where it models everyone's behavioral pattern, monitoring it for any deviation from the norm and pointing the outliers when the behavior is very different from the pattern detected.

Salem, Liu, and Mehaoua (2013) proposed an anomaly detection in medical wireless sensor networks, a type of outlier detection, has been utilized for various purposes in the medical field, such as remotely monitoring patient vital signs.

Kuna, García-Martinez, and Villatoro (2014) created a process to combine different outlier detection algorithms in alphanumeric data within information systems' audit logs to help the auditor in the outlier detection within the data. This was achieved by combining outlier detection algorithms and classification algorithms to validate the results.

2.3. OUTLIER DETECTION IN EDUCATION

Ueno (2004) developed an e-learning platform system including an online outlier detection using Bayesian predictive distribution to identify students with irregular learning process in an online test. The main idea is to use a new response time to calculate the Bayesian predictive test statistic using the students' previous data on the same test. This test statistic follows a t distribution, so when the value from the Bayesian predictive test is greater than the value of t in t distribution with α an irregular process is detected.

Oeda and Hashimoto (2017) used log-data from a programming lesson, learner's history of UNIX commands and source-code editing, to identify the students who cannot keep up with programming lessons. K-means techniques was used to compare the timing of active behavior and clustering using dynamic time warping. The final clusters show three main trends of students' outlying behavior:

1. a high number of inputs
2. a low number of inputs
3. suddenly increasing number of inputs.

Carneiro et al. (2019) tried to identify the students who make minimal use of the resources offered by the e-learning platforms but still achieve passing grades because of the face-to-face exams weight. The dataset is the Moodle log of the approved students and detect the outliers by applying the isolation forest technique. After this an analysis using all the variables was made between the anomalous group and the normal group, in the end is concluded that the student in the anomalous group has an adequate use of the platform and had a higher performance on the test.

Dobashi et al. (2021) classified learning patterns and outliers using the learning material's clickstream and the result of 13 quizzes. The clickstream only contains the interactions in the Moodle platform

and are included in-class and out-of-class activity. Then the proposed method is to calculate the deviation from the mean for both variables and classified them into four different learning patterns. They classified the outliers being the students with a high deviation. In this study the authors didn't use any cluster technique to find the four groups in the learning process and the outlier's identification also was done without any unsupervised technique.

Rotelli and Monreale (2022) presented an analytical methodology for estimating the amount of time spent by students on quality learning (i.e., time-on-task) for students working in online learning environments, their experiments and comparisons on real-world data reveal that using a local outlier detection strategy allows for a better estimation of time-on-task than existing global approaches, some outlier techniques were used but none was the best for all datasets.

Novoselteva (2022) use four unsupervised algorithms to detect outliers: distance-based k-nearest neighbors (kNN), cluster-based local outlier factor (CBLOF), histogram-based outlier score (HBOS), and density-based outlier factor (LOF). To decide the hyperparameters for each algorithm the author build a graph to show the impact in the outlier score when the hyperparameters are changed and the values was decided when the output was robust. The percentage of outliers was set as 10%.

These techniques were used in two different datasets with the same purpose, first to identify the outliers in the datasets containing the performance of the students and then create clusters for these outliers and characterize each cluster. For each one of the outlier techniques was created a set of clusters using K-means and the characterization of each cluster was done, finally the author underlined the interaction between all these different set of clusters.

The most papers on the outlier detection are based on the time-on-task problem, where the challenge is to identify an outlier based on the student behavior and not between all the students, so this type of study is not relevant for our method, because to classify a student as an outlier on the grades or moodle usage is necessary to compare to all the other students. The most recent study that is possible to compare the behavior of all the students is from Novoselteva (2022), that after the outlier detection the goal was to build clusters only on the outliers, but this final step has some clusters with only three or less students this could return a miss interpretation because number of observations are not representative to drawn any conclusion. Also, building a k-means cluster using only outliers is not the a recommend approach, because this algorithm uses the distance as the metric and on outliers the distance is biased.

3. METHODOLOGY

Before beginning the description of the methodology is presented a diagram showing the process that will take to deliver the results:

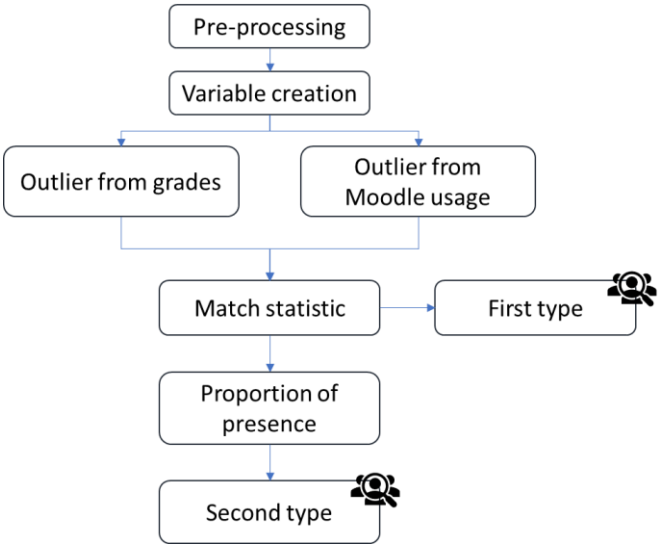

Figure 3.1 – Methodology diagram to be performed

It starts by applying the pre-processing step on the full dataset when the dataset is ready the variables are created and with this variables is possible to use the outliers techniques in the Moodle usage and grades, after the match statistic is built to locate the first type of student and from the students that are not present on the match statistic the proportion of presence will be calculated to locate the last group.

3.1. PRE-PROCESSING

The database containing the grades and the Moodle data are separated, so to do the analysis it is necessary to confirm if the students with Moodle data in a subject has a grade in this same subject, to accomplish this the datasets were merged using the user id and the subject id.

After the join was identified some subjects with less than five students, so these cases were dropped in both datasets, also the students with a grade equal to 0 was dropped because this could happen if the students dropped out and the Moodle will contaminate the analysis.

After this one more step was done in the grades database, because a student can have more than one grade in the same subject since the student can try a test more than one time, so was decided to keep the highest score.

3.2. VARIABLES USED

The variables are presented and described on the following list, the group userid and subject id will be as key:

- **Num_appe**: number of times that the userid and subject appeared together on the dataset
- **num_failed_log**: number of times that the key appeared with a failed login action
- **num_course_mod**: number of times that the student entered on the class module
- **num_mod_resource**: number of times that the student entered on a resource of a class module
- **num_mod_page**: Number of times that the student entered in the class page
- **num_mod_forum**: number of times that the student entered on the forum of a class
- **num_mod_folder**: number of times that the student entered on a folder of a class
- **num_discussion**: number of times that the student joined on a discussion
- **num_download:** number of times that the student did a download of a class material
- **num_created**: number of times that the student created a discussion, forum question or any other creation action
- **num_num_sent**: number of times that the student sent a message
- **num_view_grade**: number of times that the student viewed the grade
- **num_mod_url**: number of times that the student entered an URL on the class module
- **num_viewd_submission_form**: number of times that the student viewed the submission form after a submission is done
- **num_viewd_submission_status**: number of times that the student viewed the submission status after a submission is done
- **num_logged_time**: number of times that the student entered
- **diff_data_acesso_quiz**: difference in days between the first day that the student entered in the course module and the first day of exam or quiz

Before starting the analysis all the features above were scaled using the standard scaler approach.

3.3. TECHNIQUES FOR GRADES OUTLIERS

Since the grades is a univariate variable the technique to be used needs to be performed well in this type of variables, so a good approach will be using the statistic-based type because it is a set of techniques that performs well in univariate variables.

It was decided to test the Z-score and IQR test to find the outliers for the grades.

3.4. TECHNIQUES FOR MOODLE OUTLIERS

The selected algorithms are: Iforest, LOF and DBSCAN. For Iforest and LOF we need to set a contamination value or check the dispersion of the scores, since we are going to select outlier for each subject it is not possible to check the scores and identify a cut point because it will produce a

large volume of graphs, so was decided to use a contamination of five percent that returns an average of ten percent of outlier in all the dataset.

For the LOF it is needed to set the numbers of neighbors' parameter that an observation will be defined as outlier, so following Breunig, Kriegel, et al (2000) the values will vary between 20 and 70 and for each number of neighbors the resulting score will be stored, and the heights score for each observation will be kept. In the end each observation can have a score from a different number of neighbors, but it is guaranteed that the score will be heights possible.

After the treatments discussed on 3.2 is done it is possibly to build variables on the Moodle data, these variables were grouped by userid and subject code, so the outliers' techniques could be used in this same groups.

3.5. MATCH STATISTIC

The outlier detection will be done on two separate occasions, the first to identify the outliers in the Moodle data and the second in the final grades that were results by the same subjects that Moodle data were recorded. This analysis will be done for each subject presented in both databases, so for each subject the two types of outliers will be found. Is important to find for each subject for two points, the first is that in each subject the professor can give a different number of contents that the student can use on the system; second is that the average grade in each subject is very different, a 15 in one could be the highest one and in another could be the mean.

To identify the outliers in the first dataset, where the variables describing their interaction with the platform like the time in the platform, how many times the student enter, number of accesses the link to the material and so on. This information will be for all the subjects that the students took in the two years of analysis. In the second dataset will be done a univariate outlier technique.

With the outliers from the Moodle and the grades it is possible to perform a match statistic, where it is verified the percentage of the Moodle usage outlier is present in the grade's outliers, first it is necessary to find the observations that have a match in the outliers, this can be achieve be applying the following formula:

$$flag\_match_i = \begin{cases} 1, & flag\_outlier\_moodle_i = 1 \; and \; flag\_outlier\_grade_i = 1 \\ 0, & flag\_outlier\_moodle_i \neq 1 \; or \; flag\_outlier\_grade_i \neq 1 \end{cases}$$

Where i varies between 1 and the total length of the database and $flag\_outlier\_moodle_i$ and $flag\_outlier\_grade_i$ are equal to 1 when the observation is an outlier in the Moodle usage and the grades respectively. Now it is possible to calculate the match statistic with the following formula:

$$match\_statistic = \frac{\sum_{i=1}^{n} flag\_match_i}{\sum_{i=1}^{n} flag\_outlier\_grade_i}$$

This formula will return the percentage of the outliers in grade that are also an outlier in the Moodle variables.

In the same moment a distribution of the grades will be presented with a hue on the outliers, so we can check where the outliers of Moodle usage are on the grades graph. So, it is possible to check if an

outlier of the Moodle corresponds to an outlier of the grades, or the outliers from Moodle are in the grade's distribution.

3.6. SELECTING THE ALGORITHMS

Now that we have two outlier's methods for grades and three for the Moodle usage, we need to select a combination to perform an analysis in the Moodle usage and find some patterns, so a table is proposed where the rows will be the techniques used in the 3.3 section and the columns the section 3.4 and the values in the table will be the match statistic for each method.

3.7. PROPORTION OF PRESENCE

To identify the second type of students the proportion of presence is defined, where first is needed to divide the total number of times that the students is an outlier based only on the Moodle variables by the total number of subjects that the student did. Latter this proportion is sorted on the descending order and is possible to check the students with high proportion on this statistic, where the top students from the table will have an outlier in most of the subjects that they frequented.

With this final list we can confirm if the first type can be present and enforce that this student needs some help to understand better how to use Moodle. The students that are not the first type but is on the second type top list shows that they don't are very far from their colleague's base on the grade but have some trouble to use the Moodle or they need to study more to achieve the same that their colleagues.

4. RESULTS AND DISCUSSION

4.1. MATCH STATISTIC

With the variables created is possible to start analysing the data and applying the outliers techniques necessary in both datasets and build the match table that was describe before. With all done the following table is genereted:

Table 4.1.1 – Match statistics for each algorithm

	Moodle Outlier		
Grade outlier	**DBSCAN**	**LOF**	**IForest**
Z-score	9.09%	4.55%	4.55%
IQR	12.12%	9.09%	9.09%

In this table we can check that the best match is the IQR on the grades and the DBSCAN in the Moodle data, so the next analysis will be done using this couple of methods. Also, the IQR presents a higher match in all the Moodle outliers than the Z-score. IForest and LOF present the same performance using the match statistic as an evaluation metric.

Table 4.1.2– Match number and mean grade for each class

		DBSCAN	
		0	**1**
IQR outlier	0	(539 ; 14.04)	(48 ; 13.86)
	1	(29 ; 11.48)	(4 ; 10.25)

The above table shows the number of pair students, subjects are in each combination, where 0 means no outlier and 1 is outlier, so we can conclude that the match found only 4 students and the grade mean in this group is equal to 10.25. So, with this we locate the first group of students that would need some assistance, because they have barely the minimum grade do pass and in the same time their behavior on the Moodle is not in the same pattern than the other students.

Table 4.1.3– userid with match static equals to 1

userid
394
964

1687
3179

The students above were the ones with the match statistic equal to one, so they are the first type of students.

Next is a histogram plot on the grades with a division on the outliers and non-outliers found using both methods, so is possible to check the dispersion of these outliers and check the motive that the match statistic is low.

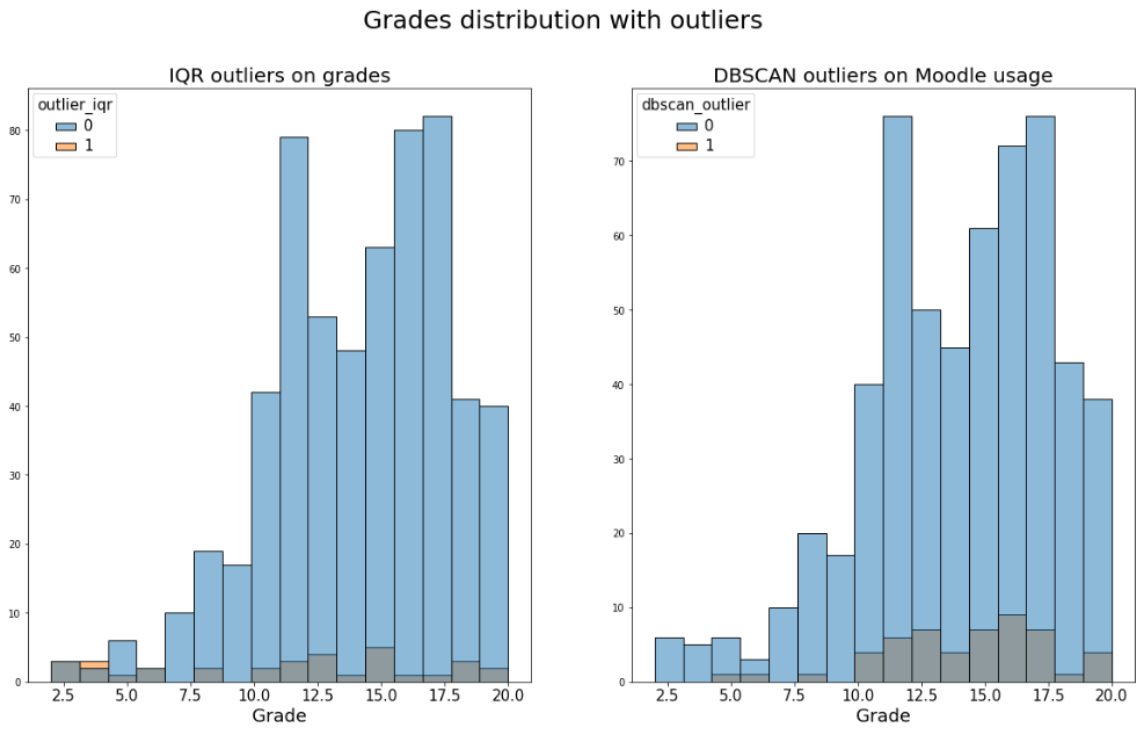

Figure 4.1.1– Histogram plot by outlier metric

The orange bars in the both graphs are the outliers, and it is easy to understand the low number of the match statistic, because 50% of the Z-score outliers has a grade lower than 10, but for the Moodle outliers this area represents only 5.7% of the outliers. The outliers based on grades has a mean equal to 11.6 while the Moodle outliers has a mean of 14.17. This shows that most outliers of Moodle usage are approved, and high grades normally is not an outlier when only the grades are considered.

One point of interest is the number of outliers from Moodle data with a grade between 10 and 15, this could show some students with difficult to learn since their usage is high enough to be an outlier, but their grades is not between the best ones.

One way to do this analysis is to check the distribution of some Moodle variables on the students with grades between 10 and 15 and verify the difference between the outliers and non-outliers, so with this is possible to identify the students with some difficult in learning.

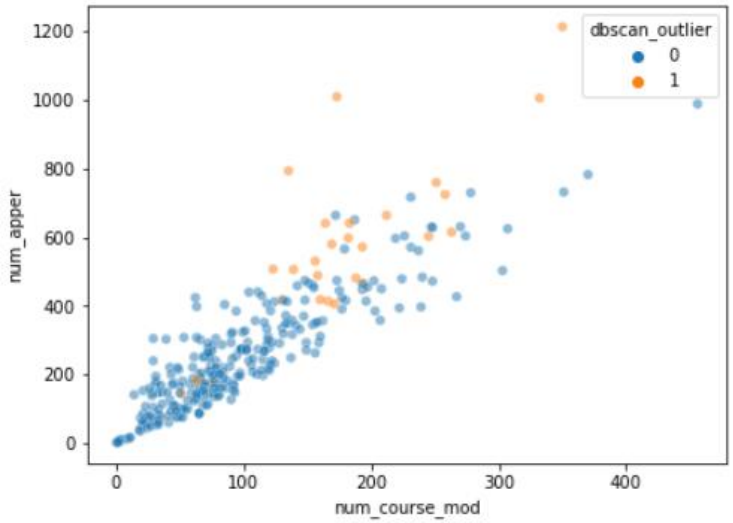

Figure 4.1.2– Scatterplot between num_apper and num_course_mod

The graph 4.2 was built with only the students with grades between 10 and 15 and the variables were without the scaling treatment, on this graph the orange dots are the outliers on the Moodle data and the blue ones are the inlines. In the image the anomaly points are more present when the number of times that the student entered on the class module is around 200 and number of times that the userid and subject appeared together on the dataset is above 400, so students with this type of behavior shows that access numerous times the page but the scores weren't between the heights.

4.2. PROPORTION OF PRESENCE

On the previous chapter was presented that the data have some students with a grade between 10 and 15 and are outliers on Moodle data, since the minimum grade to be approved is 10 some students are using a lot the online tool but have a low grade, so the proportion of presence will be calculated focus on this student to achieve this will be selected the students with grades between 10 and 13.

Table 4.2.1– userid with proportion of presence and other metrics

userid	Number of outliers	Number of subjects	Proportion of presence	Mean grade	Mean grade of all subjects
394	6	6	100%	12.00	13.69
6929	3	5	60%	10.40	11.33
3126	1	2	50%	13.00	15.60
964	1	3	33%	11.66	9.75
1068	1	4	25%	11.25	12.20
1488	1	4	25%	12.25	13.27
3165	1	4	25%	12.00	14.35
1544	1	4	25%	11.50	12.42
927	1	4	25%	10.75	12.45
1111	1	5	20%	11.60	10.57

The table 4.2.1 presents the ten students with the highest proportion of presence, the first column is the userid of the student, the second is the number of times that the student is an outliers on the Moodle data, the third is the total subjects, the fourth the proportion of presence statistic, the sixth is the mean grade and the sixth is the mean grade of all the subjects of the students and not only the ones between 10 and 13.

The userid 964 and 394 appears in the table 4.1.3 and on the 4.2.1, they are the ones with most problems on the learning, because they have a high proportion of subjects with low grades and with a high use of the online tool, also they have grades that are very far from their colleagues, so it is proposed to the university to invite these students to a talk to understand the reason of these poor performance and to understand how the school can help the student.

At the same time the students 1687 and 3179 don't appear on table 4.2.1, so they don't need to have an action like the description above, but they require some action or follow-up by the university.

The other students are necessary to do an analysis one by one to understand if is necessary to contact them. The students 3126, 3165 don't need to be contact because they have a high grade of the overall subjects, and their presence is cased only because they had a low number of subjects between 10 and 13. The student 6929 is desired to have some contact by the university, because it has a high value on the proportion, have a significant number of subjects on these group and the overall grade is low. The rest of the students is only with a presence between 20% and 25%, so is indicated to keep a close attention to understand if they start to drift to the high values.

A check on the variables for the students 964, 394, 6929, 1687 and 3179 was done to confirm that they are outlier on Moodle usage is not because of low values on the dataset.

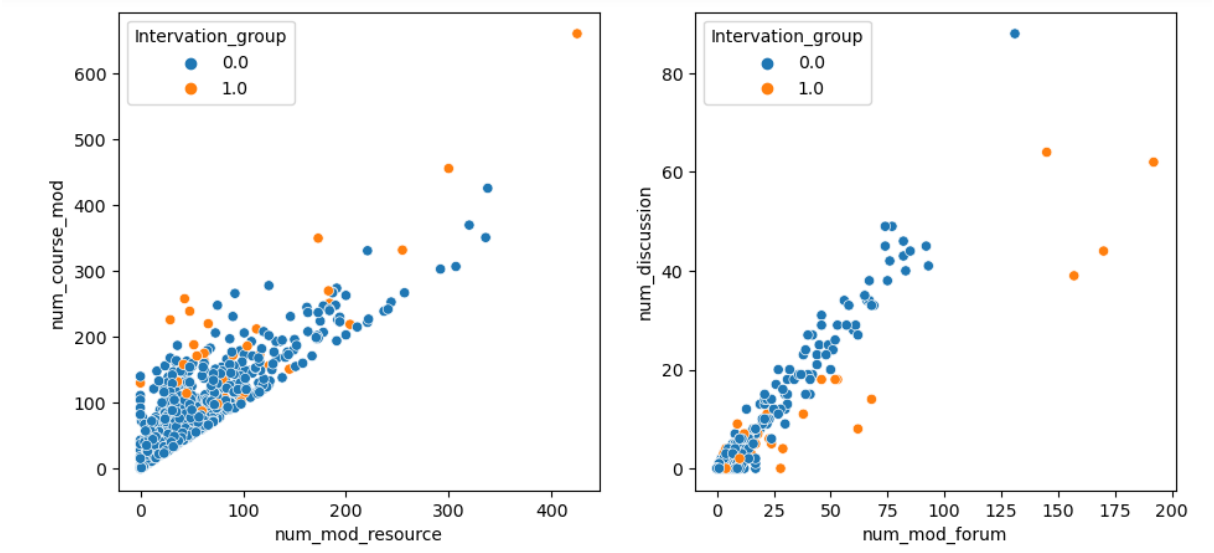

Figure 4.2.1– Two scatterplots with a hue on the students that was proposed an intervention

The figure above the orange dots are the students that were proposed to have an intervention. Is easy to see that their numbers are not low in some cases, they appear more than once because each value is for the userid and subject, is possible to understand why they become an outlier on the Moodle usage. In some cases, the students are in the middle of the normal points, but the technique

that we used to identify the outliers are based on multivariate analysis and in the figure is only shown a bivariate approach to be easy to understand.

5. CONCLUSION

Since the universities nowadays have a high number of students is impossible to track each one of them and provide customized feedback, so is necessary to build automatic methods to select the students with more difficulties for the universities to intervene and help them, resulting in a more accurate action.

For the lack of studies focused on detecting the deficient performance of the students, this study presented a methodology to identify two types of situations that the student can be found using outliers' techniques on the Quiz/tests results and Log files from Moodle. These groups were designed to find the students with a bad performance on grades and a high use of web-based educational system. Also, these students usually are discarded from traditional analysis and this study focus only on them.

In the Results chapter was presented the statistics responsible to locate the groups described above and in total was found three students to have a more active action by the university and two that require more attention by the university to understand their behavior. With this the university can focus in only five students and can perform a customized action for each one and can help them achieve their academical goal with more assurance.

Furthermore, the results presented that the extreme values of Moodle usage are not really correlated with the final grades, so when the student has a high interaction with the platform is not necessary that it will result in a high final grade.

To conclude is evident that the outlier techniques that are not well explored on the EDM is a good tool to identify students with a bad performance. These techniques didn't work by themselves, two auxiliary statistics were proposed to help find the desired groups. This indicates that simple formulas can enhance the results from the traditional machine learning or data mining approaches and should be part of further studies.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

Since multivariate techniques were used to identify the anomalies based on the Moodle data it is difficult to locate the variables that can separate the students that are outliers and had very different grades and since thy all are outliers is not adequate to apply a cluster technique to identify their differences.

Also, the Moodle outliers were identify based on the variables presented in this project, so the results are really linked with these variables, so one could change these variables or create new ones and the conclusions could be different. Since, in further analysis is recommended to spend more time to create new variables to extract information of how the students use the Moodle.

In the results none direct action were describe because is out of the scope of this study and that can be accomplish by different ways and should be designed by a specialist in educational mentoring.

The study was considering two different bachelor's degree and a data from two years, another analysis that could be done in further works is to check if the conclusions on this study are kept in all the subgroups.

7. REFERENCES

Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. Journal of Network and Computer Applications, 68, 90-113.

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. Telematics and Informatics, 37, 13–49.

Amos, J. (2008). Dropouts, Diplomas, and Dollars: US High Schools and the Nation's Economy. Alliance for Excellent Education.

Baker, R.S.J.d. (2010) Data Mining for Education. In McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition), vol. 7, pp.112-118. Oxford, UK: Elsevier.

Baker, R. S., & Inventado, P. S. (2014). Educational Data Mining and Learning Analytics. Em J. A. Larusson & B. White (Orgs.), Learning Analytics (p. 61–75). Springer New York.

Baker, R. S. (2015). Big data and education (2nd ed.). New York, NY: Teachers College, Columbia University

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (pp. 93-104).

Carneiro, R. E., Drapal, P., Fagundes, R. A. A., Maciel, A. M. A., & Rodrigues, R. L. (2019). Anomaly Detection on Student Assessment in E-Learning Environments. 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), 168–169.

Dobashi, K., Ho, C. P., Fulford, C. P., Lin, M.-F. G., & Higa, C. (2021). Classification of Learning Patterns and Outliers Using Moodle Course Material Clickstreams and Quiz Scores. 6.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd (Vol. 96, No. 34, pp. 226-231).

Gibbons, R. D., Bhaumik, D. K., & Aryal, S. (2009). Statistical methods for groundwater monitoring. John Wiley & Sons.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. Technometrics, 11(1), 1-21.

Hautamaki, V., Karkkainen, I., & Franti, P. (2004). Outlier detection using k-nearest neighbour graph. In Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. (Vol. 3, pp. 430-433). IEEE.

Hawkins, D. M. (1980). Identification of outliers (Vol. 11). London: Chapman and Hall.

He, Z., Xu, X., & Deng, S. (2003). Discovering cluster-based local outliers. Pattern recognition letters, 24(9-10), 1641-1650.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. Artificial intelligence review, 22(2), 85-126.

Knorr, E. M., Ng, R. T., & Tucakov, V. (2000). Distance-based outliers: algorithms and applications. The VLDB Journal, 8(3), 237-253.

Konijn, R. M., & Kowalczyk, W. (2011). Finding fraud in health insurance data with two-layer outlier detection approach. In International conference on data warehousing and knowledge discovery (pp. 394-405). Springer, Berlin, Heidelberg.

Kuna, H. D., García-Martinez, R., & Villatoro, F. R. (2014). Outlier detection in audit logs for application systems. Information Systems, 44, 22-33.

Lang, C., Siemens, G., Wise, A., & Gasevic, D. (2017). Handbook of learning analytics. SOLAR, Society for Learning Analytics and Research. New York, NY: SOLAR

Li, Z., Zhao, Y., Hu, X., Botta, N., Ionescu, C., & Chen, G. (2022). Ecod: Unsupervised outlier detection using empirical cumulative distribution functions. IEEE Transactions on Knowledge and Data Engineering.

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In 2008 eighth ieee international conference on data mining (pp. 413-422). IEEE.

Novoseltseva, D. (2022). Outlier Detection as a Tool for Reinforcing Data Analysis and Prediction in Education. [Doctoral dissertation, Université Paul Sabatier - Toulouse III]. HAL.

Oeda, S., & Hashimoto, G. (2017). Log-Data Clustering Analysis for Dropout Prediction in Beginner Programming Classes. Procedia Computer Science, 112, 614–621.

Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (pp. 427-438).

Romero, C., Romero, J. R., & Ventura, S. (2014). A survey on pre-processing educational data. In Educational data mining (pp. 29-64). Springer, Cham.

Romero, C., & Ventura, S. (2013). Data mining in education: Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12–27.

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. WIREs Data Mining and Knowledge Discovery, 10(3).

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.

Rosner, B. (1983). Percentage points for a generalized ESD many-outlier procedure. Technometrics, 25(2), 165-172.

Rotelli, D., & Monreale, A. (2022). Time-on-Task Estimation by data-driven Outlier Detection based on Learning Activities. LAK22: 12th International Learning Analytics and Knowledge Conference, 336–346.

Salem, O., Liu, Y., & Mehaoua, A. (2013). Anomaly detection in medical wireless sensor networks. Journal of Computing Science and Engineering, 7(4), 272-284.

Spirtes, P., Glymour, C., & Scheines, R. (2000). Causation, prediction, and search. New York: MIT Press.

Tang, X. M., Yuan, R. X., & Chen, J. (2015). Outlier detection in energy disaggregation using subspace learning and Gaussian mixture model. International Journal of Control and Automation, 8(8), 161-170.

Ueno, M. (2004). ONLINE OUTLIER DETECTION SYSTEM FOR LEARNING TIME DATA IN E-LEARNING AND IT'S EVALUATION. 7.

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa