NOVA IMS Information Management School

# MDSAA

Master Degree Program in

**Data Science and Advanced Analytics**

## Predicting Academic Success

A Comprehensive Analysis using Moodle Log Data

and Learning Analytics

Carolina Pratas Ferreira Mira

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

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

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

# Predicting Academic Success

A Comprehensive Analysis using Moodle Log Data

and Learning Analytics

by

Carolina Pratas Ferreira Mira

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Business Analytics

## Supervised by

Roberto Henriques, PhD, NOVA Information Management School

July, 2024

# 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 acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

*[Lisbon, 15/07/2024]*

# ABSTRACT

There has been an advance in computing with the rise of large volumes of data. Education was one of the fields that has advanced to this point. As new technologies and information are continually integrated, datasets are now available from students' interactions with educational software and online learning platforms. Educational software like Moodle exemplifies e-learning solutions that combine traditional teaching methods with information technology resources.

In this line of thought, this study investigates the potential use of Learning Analytics and Moodle log data to predict academic success in higher education, considering the first part and the whole course. The research uses a variety of machine learning algorithms, including Random Forest, Logistic Regression, Gradient Boost, Support Vector Machine, and Neural Networks, to uncover patterns in student behaviour and academic performance. The data used in the study is related to Moodle log data and sociodemographic information that comes from NOVA Information Management School's Master of Data Science and Advanced Analytics program, of three academic years (2021-2023). To achieve the objective a CRISP-DM methodology was implemented to serve as the base model for Machine Learning models.

The Random Forest model had the highest prediction values; however, overfitting was detected in all models. Key findings show that student involvement, as measured by interactions with course materials and demographic parameters such as age and nationality, has a significant impact on academic success.

The study underlines the importance of early and persistent student involvement with Moodle and offers instructors techniques to improve student engagement and performance. Future work will include the development of early warning systems and complete dashboards to offer instructors real-time knowledge.



# KEYWORDS

Learning Analytics; Academic Success; Machine Learning; Moodle.


## Sustainable Development Goals (SDG):

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| **LA** | Learning Analytics |
| **LMS** | Learning Management System |
| **ML** | Machine Learning |
| **SVM** | Support Vector Machine |

# 1. INTRODUCTION

## 1.1 CONTEXT AND RELEVANCE

In recent years, we observed a sweeping transformation across various domains, mainly driven by incredible technological progress, transforming the world. Hussaini (2022) emphasizes these crucial changes, which have had a profound impact on multiple sectors of the world, notably in the field of education. With the constant integration of new technologies and information, the evolution of the student learning processes has become imperative. The significance of this evolution was especially evident during the COVID-19 pandemic when the necessity to adapt various forms of education became apparent (Abu Talib et al., 2021).

Recent advancements in Learning Analytics have created a way to develop innovative methodologies designed to harness educational datasets to enhance and support the learning process (Chatti et al., 2012). Increasingly, substantial datasets are becoming available from students' interactions with educational software and online learning (Siemens & Baker, 2012).

Educational software platforms like Moodle represent E-Learning solutions that combine traditional teaching methods with information technology resources. These platforms aim to improve students' performance through distance learning. (Gogan et al., 2015) as students need not only to gain subject-specific knowledge in their classes but also to learn how to explore and critically evaluate the vast array of information available online.

Understanding and studying Moodle Log Data is crucial in predicting academic success. However, while a significant body of research has been developed on predictive modelling using Moodle log data and Learning Analytics (LA) to identify students at risk of academic, there is a substantial gap in understanding how to interpret the results of predictive modelling for the practical implementation of intervention strategies by educators and administrators. Most existing studies have focused on the development of predictive models but there has been insufficient emphasis on creating methods like early warning systems capable of assisting educators in providing prompt feedback and developing corrective measures before issues become critical (Conijn et al., 2017; Santos & Henriques et al., 2021).

Furthermore, there is a lack of study on how to effectively communicate predictive modeling results to educators and administrators to promote intervention efforts and improve student outcomes. Without effective communication and practical application of these models, their potential to improve educational practices is underutilized.

Bearing this scenario in mind, the motivation for this thesis is driven by the urgent necessity to explore and understand student behaviour to enhance student grades and overall academic performance. The main goal is to develop a predictive model using Moodle log data to identify students at risk and provide actionable insights. Furthermore, the goal is to create measures to avoid failing students in future classes. Additionally, this research aims to establish efficient

means of communicating predictive modelling results to educators and university administrators, empowering them to implement targeted intervention strategies.

This research provides information and answers that will improve current academic grades and prepare the way for long-term success in future educational courses.

## 1.2 OBJECTIVES

To address this challenge, this study is dedicated to investigating and developing a predictive model that uses Moodle log data, to obtain a profound understanding of students' academic performance. Furthermore, the study aims to create a way to help establish an efficient means of communicating this information to educators and university administrators, empowering them to implement targeted intervention strategies.

So, the guiding research questions of the study are:

- How can Moodle log data and learning analytics be effectively integrated to predict students' academic success in the first part and the entire course?
- How can predictive modelling results be effectively communicated to educators and administrators to support intervention strategies and improve student outcomes?

## 1.3 STUDY OUTLINE

The first chapter of this thesis presents the introduction with context and relevance, followed by the objectives and a brief description of the structure of the work.

The second chapter introduces the literature review of the main essential aspects of the study: Learning Analytics, Moodle, and Academic Success. Furthermore, we analyzed the most relevant previous research in this field in the literature review.

The following section explains all the steps used to consider a Crisp-DM methodology. In this chapter, all data understanding, data preparation, and modelling are done.

The last two chapters present the conclusion, limitations and future work. The conclusion provides an overview of the work done and the most important results and insights from the study. The limitations and future work provide the main obstacles faced and suggest possible future research directions.

# 2. LITERATURE REVIEW

The following literature review synthesizes some critical domains of study to provide a comprehensive framework and a context for the current research. It also references an overview of essential related work that answers the question: "What is the state of the art of predicting the academic performance of students using *Learning Analytics*, *Moodle* log data and *Multimodal Analysis*."

Some data repositories, such as Scopus, Scispace and Mendeley, were used to select the papers. The papers were selected based on their relevance to the research questions considering the keywords and trying to cover different perspectives and methodologies. The chosen research studies were published between 2017 and 2023 and covered various topics to comprehensively understand the study's current advancements, challenges, and potential future directions.

These studies are beneficial for finding out which procedures have been used and which have given better results. These papers discuss many techniques, including choosing which variables to use, which algorithms work best, and which pre-processing activities get the most outstanding results. By considering the existing body of research, this study aims to contribute to expanding and improving the knowledge in the field.

## 2.1 LEARNING ANALYTICS

*Learning analytics* is an emerging field that utilizes computational analytics to understand and enhance learning. *Learning Analytics* is a new field of study, considered an interdisciplinary field between *Statistics*, pedagogy, *Machine Learning*, *Artificial Intelligence*, *Business Intelligence, and Learning Technology*. LA aims to use educational data to improve learning, teaching and learning environments (Banihashem et al., 2018).

Various techniques are used in LA, including monitoring student's online activities, assessing their level of involvement, and forecasting the students' outcomes and academic results.

This field is essential to teachers, and institutions can use *Learning Analytics* to inform their teaching strategies and student assistance systems by utilizing various data sources and analytical methods.

Research shows that *Learning Analytics* could bring significant benefits by providing educators with insights into students' habits and progress in learning, identifying students at risk of failure, and improving some learning outcomes (Banihashem et al., 2018).

## 2.2 MOODLE

*Moodle* is a widely adopted open-source *Learning Management System* (LMS) that offers extensive applications to empower educators and administrators with a robust, secure, and integrated system to develop personalized learning environments. This platform gives a vast array of log data capturing students' interactions, activities, and progress. It can capture data from forum discussions, resource access, quiz attempts, downloads, etc.

*Moodle* is used as a tool to support traditional learning, and it can be used by universities, schools, and individual instructors, considering its easy usability since it is straightforward to learn how to use the system (Matijašević-Obradović et al., 2017).

## 2.3 ACADEMIC SUCCESS

*Academic success* in university is a multidisciplinary concept that includes academic performance, intellectual and social abilities, professional and personal development, and participation in learning experiences. It can be influenced by student's interactions with people and university software and by the individual characteristics of each student (Alyahyan & Düştegör, 2020).

Umbach & Wawrzynski (2005) in their study found that student-faculty interaction positively promotes cognitive development, satisfaction with their university experience, and overall academic success.

In another way, Robbins et al. (2004) showed that psychosocial and study skill characteristics significantly predict university students' performance and perseverance, suggesting that academic success is more than cognitive abilities but also includes personal and behavioural factors.

## 2.4 RELATED WORK

Mwalumbwe & Mtebe (2017) developed a study to forecast student performance using Moodle LMS at Mbeya University of Science and Technology (MUST) to understand how learners use the system and find some strategies to improve the system's performance. The research focused on two courses, Applied Biology I and Services and Installation II, with 171 students. This research mainly aimed to find specific factors significantly influencing a student's academic success. A Linear Regression model was used to predict students' grades, and Adjusted R Square was used to measure the model's performance.

To conclude, the study finds that Forum Posts (beta value = 77.1%) are a critical factor that influences the learning performance of students in both courses, which means that participation in discussion posts is positively correlated with academic performance. Other factors that positively impacted the student's grades were the interaction with peer students (with beta value = 19.6%) and completing the proposed exercises (beta value = 51.5%). Surprisingly, it was found that some factors, such as downloads, login frequency, and time

spent in the LMS, had no significant effect on student's learning performance in both courses. Finally, for the first course, the Linear Regression predicted an R squared of 0.954, and for the second course, it was 0.943. It means that in Applied Biology I, the model explains 95.4% of the variance, and Services and Installation II explains 94.3%.

Kadoic & Oreski (2018) Developed research using a different approach to analyze the data. Instead of using predictive models, they used visualization methods. This study examines the behaviour and performance in a Moodle-based course at the University of Zagreb's Faculty of Organization and Informatics. The main goal is to analyze the log files of 73 students produced by the LMS to interpret the relations between students' actions and their academic performance and understand the data patterns.

So, the guide questions of the investigation are:

- RQ1: To what extent are individual variables derived from log data a reliable predictor of academic success?
- RQ2: What is the level of similarity in student LMS usage between genders?

Surprisingly, students with the highest grades appeared to be the most active the day before lectures, seminars, and exams. This result challenges conventional assumptions about study habits, suggesting a possible correlation between academic achievement and last-minute involvement.

When analyzing the second research question, the authors found some patterns in the data. Female students were more engaged and accomplished, showed a higher frequency of log entries, and consequently had a higher average grade when compared to male students. Another pattern found was a peak in log entries on the day before exam days, indicating a possible effort in preparation for the exam.

Quinn & Gray (2019) investigated the possibility of predicting students' academic success in a blended learning environment for higher education using data from the LMS Moodle. This was accomplished by creating student activity metrics from the Moodle course logs for higher education.

The study aims to forecast a student's alphabetic grade as well as whether they would pass or fail the course. It is important to notice that classrooms in further education are often smaller than those at universities, so it was necessary to combine data from several classes, which could have reduced predicted accuracy because there were more sources of variance.

In this study, there were two research questions: Using measures created from Moodle activity from the entire course duration, is it possible to predict student academic performance on further education courses? Using the same data but only for the early weeks of a course, the first six weeks and the first ten weeks, is it possible to predict whether a student will pass or fail? They used data from 29 classes in 9 different modules. The variables were created from Moodle data for the entire course duration and the first six and ten weeks

of a course. The different grade levels that were used were Early Exit (0% or no grade), Pass (50-64%), Merit (65-79%), Distinction (80%+), and Fail (1-49%).

The results showed that it was reasonably easy to predict a student's alphabetic grade (accuracy= 60.5%), and Random Forest was the algorithm with the highest accuracy values. Using all the course data, it was possible to predict with high accuracy whether students would pass or fail the Pass/Fail classification task (accuracy= 92.2%). However, using data from the first six weeks, it performed poorly in predicting failing students.

A paper by Anagnostopoulos et al. (2020) describes an innovative approach to detect high-risk students in Moodle, i.e., students at risk of failure in online courses, by comprehending a wide range of factors aligned with the course design. The researchers used an approach implemented in Moodle LMS, with data from two cohorts of an online course. The first cohort contains detailed data about students' behaviour and interactions with the learning objects that support the course's activities to comprehend the relation with the final test performance.

To achieve the best performance of the model, a model-enhancing technique was used using a Majority Voting ensemble classifier with six models: Naïve Bayes, Support Vector Machine, K-NN, RIPPER, and C4.5. The study involves data from 183 students attending the cohort.

The result of the Majority Voting gave the best prediction, with an accuracy of 73.31%, demonstrating that predicting the model with two distinct cohorts gives statistically significant accuracy values.

The main results of this research highlight the importance of course design in predicting student success. More precisely, carrying out a set of course activities such as watching a video, interacting with multimedia content, doing a laboratory exercise or a self-assessment quiz, and passing formative assessment tests demonstrate a real effort from the student that is crucial for the successful completion of the entire course.

For future work, the researchers are developing a Moodle plugin for the application of an Early Warning intelligence system capable of providing pertinent feedback for improving student achievement.

Gaftandzhieva et al. (2022) provided a case study predicting the students' final grades based on their attendance at Zoom online classes and their actions in the Moodle learning management system. To address the three study topics outlined, a few machine learning and statistical techniques were used to answer the three research questions defined:

- RQ1: Do the features of learning resources, activities, and attendance of the students demonstrate any correlation with the final academic grade of the learner?
- RQ2: Is attendance significantly associated with academic performance?

- RQ3: Can machine learning algorithms be utilized to predict the learner's final academic grade?

The authors used a data frame with the final grades and the activities in the online course of 105 students from an Object-Oriented Programming class at the University of Plovdiv during the 2021-2022 year.

A Chi-square test was utilized to answer the first two questions and determine the relationships between the students' final grades and the event context, including the lectures, source code, exercise, and assignment. To predict the final academic grades of learners, four Machine Learning algorithms were used - Random Forest, XGBoost, KNN and SVM - performing experiments on 4-week, 8-week and the entire dataset.

After some visualizations, it was possible to conclude that 32.4% of students had "Good" academic achievement, and 21% had a "Satisfactory" group. However, the students with the lowest percentage (9.5%) are those whose academic performance is rated as "Excellent." The percentages for the academic performance categories labelled as "Very Good" and "Fail" are almost identical; academic performance is categorized as "Fail" for 18.1% of students and "Very Good" for 19% of pupils.

The results showed a substantial correlation between academic grades and all event settings, meaning that students' engagement and attendance in the lessons and activities impacted their academic performance. Considering the Machine Learning Algorithms applied, they did not perform well in a four-week data frame. However, in the eight-week dataset, all the algorithms performed moderately well (with more than 60% accuracy). Random Forest was the model with the highest accuracy, recall and F1-score values, with 71%, 78% and 77%, respectively. It concludes that in this study, the Random Forest may be used to predict which students will fail after eight weeks.

Kaensar & Wongnin (2023) conducted research with the primary objective of addressing the challenges in education faced during the COVID-19 pandemic when everything was forced to switch to entirely online instruction. The researchers want to determine whether it was possible to identify students' performance by examining the students' interactions with log data from LMS at four distinct course completion levels: 25%, 50%, 75% and 100%. This research wants to study if it is possible to identify student performance at different course completion levels and which Machine Learning algorithms predict the best results at each level and in the overall of course.

They used a dataset from Database Systems at Ubon Ratchathani University about students during the academic year 2021–2022. This dataset contained 54,803 events, considering students' interactions in courses, assignments, forums, files, quizzes, labels, videos, and attendance.

To address these objectives, six ML algorithms were used to predict student performance: Neural Network, Random Forest, Decision Tree, Logistic Regression, Linear Regression, and Support Vector Machine. The main conclusions indicated that the Decision Tree classifier perform better than the other algorithms in the course completion, with an accuracy of 81.10%. Considering the different levels of completion, SVM was the most accurate algorithm at the first level (25%) with an accuracy of 86.90%, while Linear Regression was the best algorithm in the other stages, with accuracy values of 70.80% at the 50% and 80.20% at the 75% levels, respectively.

According to the results of Decision Trees, proactive students, particularly those who pay more attention and submit assignments and quizzes, had a positive impact on their performance. On the other way, students who fail to attend and have minimal interaction with course content may also fail a class.

A different approach has been made by Reimondo Tamba et al. (2023), where their study deviates from the traditional predictive analysis, focusing on a cluster analysis and association analysis using Moodle activity log data. The main goal of the study is to organize students into distinct groups according to their engagement within the learning course, using the K-Means Algorithm. Simultaneously, it is essential to find relationships between different students' actions using association analysis methods like the Apriori method.

The clustering implementation divides students into three groups, i.e., clusters, where each cluster indicates the student's level of activity – active, less active, and very inactive. The clustering was done considering 55 students, which reveals 21 active students, 9 less active and 25 very inactive students. Association rules provide valuable insights into understanding the students' behavior in LMS, especially in the strong rules generation events that are course viewed and course module viewed.

As a result, using two distinct data mining functionalities, clustering to create possible groups and understanding information about student activity and association rules to find relationships between online activities, could be helpful for analyzing and visualizing activity log data in an LMS.

## 2.5 SUMMARY TABLE

Table 1 - Literature Review Summary Table

| Author | Title | Techniques | Domain | Best Performance |
|---|---|---|---|---|
| **Imani Mwalumbwe & Joel S. Mtebe** | Using Learning Analytics to predict students' performance in Moodle Learning Management System: A case of MBEYA | Linear Regression | University Education | Adjusted R Square = 0.954: Course 1<br>Adjusted R Square = 0.943: Course 2 |

| | | | | |
|---|---|---|---|---|
| | University of science and technology | | | |
| **Nikola Kadoić & Dijana Oreški** | Analysis of Student Behavior and Success Based on Logs in Moodle | Visualization methods | University Education | |
| **Rory Joseph Quinn & Dr. Geraldine Gray** | Prediction of student academic performance using Moodle data from a Further Education setting | Random Forest, Gradient Boosting, KNN and Linear Discriminant Analysis | Further Education Setting | Random forest – 92.2% - accuracy |
| **Anagnostopoulos & Kytagias & Xanthopoulos & Georgakopoulos** | Intelligent Predictive Analytics for Identifying Students at Risk of Failure in Moodle Courses | Naïve Bayes, Support Vector Machine, K-NN, RIPPER, C4.5, Majority Voting | University Education | 73.31% accuracy in Ensemble Majority Voting |
| **Silvia Gaftandzhieva & Ashis Talukder & Nisha Gohain & Sadiq Hussain & Paraskevi Theodorou & Yass Khudheir Salal & Rositsa Doneva** | Exploring Online Activities to Predict the Final Grade of Student | Random Forest, Extreme Gradient Boosting, K-nearest neighbors, and Support Vector Machine | University Education | Random forest – 71% accuracy, 78% recall, 77% - F1-score |
| **Kaensar & Wongnin** | Analysis and Prediction of Student Performance Based on Moodle Log Data using Machine Learning Techniques | Neural Network, Random Forest, Decision Tree, Logistic Regression, Linear Regression, and Support Vector Machine | University Education | 81.10% accuracy in Decision Trees |
| **Reimondo Tamba & Krista Lumbantoruan & A. Pakpahan & Samuel Situmeang** | A cluster and association analysis visualization using Moodle activity log data | K-Means and Apriori Algorithm | University Education | Apriori Algorithm: 0.73 - Support; 0.91 - Confidence<br>K-Means: 0.54 – silhouette |

This literature review illustrates a comprehensive landscape in the domain of predicting academic performance using Learning Analytics and Moodle log data. The studies highlight how various methodologies, including linear regression, visualization techniques, Machine Learning algorithms, Ensemble Methods, and Clustering approaches, have been used successfully. Moreover, the review provides some practical examples of how predictive results might be communicated to educators to support timely interventions. By understanding the current state of the art, this research contributes to establishing valuable knowledge and to the expansion of educational practices, addressing some existing gaps.

# 3. METHODOLOGY

A scientific framework was used to achieve the work project goals. This methodology refers to CRISP-DM (Cross-Industry Standard Process for Data Mining) which is a process that serves as the base model for applying Machine Learning projects (Hotz, 2023). Figure 1 below visually represents the methodologies and the six phases presented.

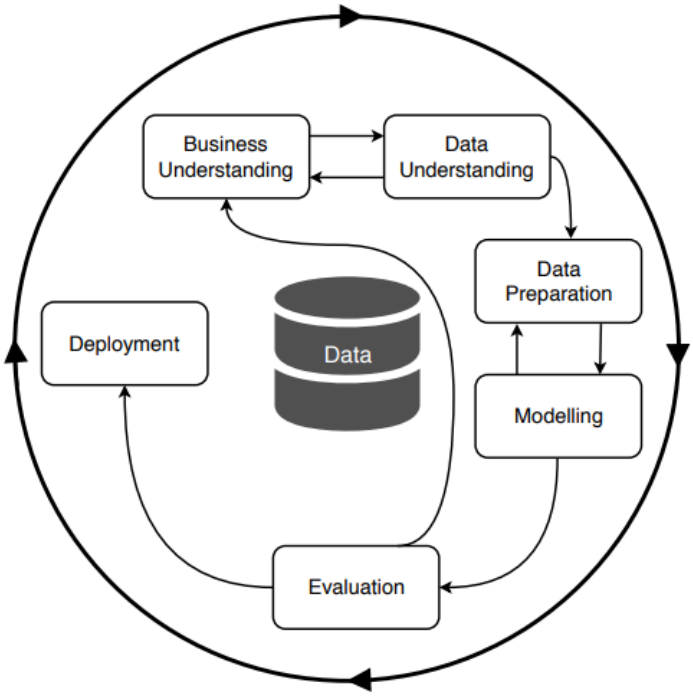



Figure 1 - CRISP-DM diagram from Plumed & Ochando et al. (2019)

As observed, CRISP-DM has six phases: Business understanding, Data understanding, Data preparation, Modelling, Evaluation and Deployment.

**Business Understanding** is the phase where the primary objective is to clarify and understand the research problem and the project goals. A Literature Review and domain analysis are included to determine the data's scope, objectives, and success criteria. Stakeholder interviews, requirement collecting, and problem formulation are some strategies used in this phase.

During the **Data Understanding** phase, the researchers focus on identifying, collecting and analyzing the data to gain insights into its structure, quality and possible issues. Researchers identified data sources, collected initial data, explored the data, and verified its quality. Descriptive statistics, data profiling and visualization techniques are used in this phase to better understand the data and identify initial issues.

The **Data Preparation** phase is essential to selecting, cleaning, and transforming the data for analysis and modelling. Relevant data are selected, cleaned, and transformed through various

processes, including handling missing values, removing outliers, normalization, and scaling. Feature engineering is also performed to construct new features that could be beneficial for modelling.

**Modelling** is where we try various data mining algorithms to build predictive or descriptive models based on the study objectives. The method involved choosing relevant modelling techniques, creating test designs, producing models, and evaluating their performance. Algorithm selection, model training, and hyperparameter tuning are used to create effective models.

The **Evaluation** phase focuses on evaluating which model has the best performance, i.e., which algorithm fits better with the business objectives. This phase included evaluating model performance, examining models against company objectives, and determining the next steps. Cross-validation techniques and evaluation metrics are utilized to compare and choose the best model.

Finally, during the **Deployment** phase, the evaluation models and the results are presented in a real-world context. This step involved planning the deployment, monitoring, and maintaining its performance over time. Model integration, system implementation, and continuous monitoring techniques were used to ensure the model's effectiveness.

By using this methodology, this research aims to improve the accuracy and efficacy of the models and consequently helps with decision-making.

The Business Understanding phase was completed in sessions one and two, and we are now moving on to the Data Understanding phase.

## 3.1 DATA UNDERSTANDING

### 3.1.1 Data

This study used collected log data from a Master of Data Science and Advanced Analytics in Nova IMS Moodle in three years between 2021 and 2023. It was extracted in three .csv files and each one was imported to a Pandas Data frame in a Jupyter Notebook. As the three tables had the same variables a unique data frame was created by concatenating them. As a starting point, this dataset has 2919589 observations and 10 variables and regarding the dataset's content, after verifying the dataset data types, it was concluded that there was an issue: the variable *cd_discip* was an integer not a float and *timecreated* must be a datetime type.

In Table 2, we can observe a general structure and description of the variable's existence in the data frame.

Table 2 – Description Table of Log Variables

| Variable name | Description | Data Type | Value Example |
|---|---|---|---|
| ***ID*** | ID of the log | Int64 | 25908150 |
| ***EVENTNAME*** | Description of the event | Object | \core\event\user_loggedin |
| ***COMPONENT*** | Moodle component | Object | core |
| ***ACTION*** | Moodle action | Object | loggedin |
| ***TARGET*** | Moodle target | Object | user |
| ***ID_STUDENT*** | Student's ID | Int64 | 11524 |
| **COURSEID** | Marter's ID | Int64 | 1943 |
| ***FULLNAME*** | Course's name | Object | Estatística para a Ciência de Dados |
| ***CD_DISCIP*** | Course's code | Int64 | 200178 |
| ***TIMECREATED*** | Data time of log | Object | 2020-11-20 14:25:17 |

It also provided social demographic information about the 717 distinct students who frequent the Master. It was extracted as a .csv file imported to a Pandas Data frame in the same Jupyter Notebook as before. This dataset comprises 12 variables and 7884 observations, and considering the datatypes all variables are consistent. Table 3 gives the features names, description, data type, and a value example for all fields of the data frame.

Table 3 – Description Table of Sociodemographic Variables

| Variable name | Description | Data Type | Value Example |
|---|---|---|---|
| ***CD_LECTIVE*** | School year code | Int64 | 202122 |
| ***CD_COURSE*** | Marter's ID | Int64 | 7512 |
| ***CD_DISCIP*** | Course's code | Int64 | 200142 |
| ***DS_DISCIP_PT*** | Course's name PT | Object | Inteligência Computacional para Otimização |
| ***DS_DISCIP_EN*** | Course's name EN | Object | Computational Intelligence for Optimization |
| ***ID_STUDENT*** | Student's ID | Int64 | 15807 |
| ***GENDER*** | Student's gender | Object | F |
| ***DT_BORN*** | Student's born date | Datetime64[ns] | 2000-11-06 |
| ***DS_NATIONALITY*** | Student's nationality | Object | Portugal |
| ***DS_NATURAL*** | Student's ID naturality | Object | São Domingos de Benfica |

Considering all the provided data and the initial analysis, some fields are not relevant in the context of the study, so the columns *cd_course*, *DS_Natural* and *courseid* were deleted.

### 3.1.2 Exploratory Analysis

For the explanatory analysis, we decided not to merge the data frames and make an analysis of each one, so considering the "Moodle data frame" as the log data and the "students data frame" as the information about students.

A statistical analysis was conducted to a numerical analysis to gain a better understanding of the dataset. The analysis of the mean, median, standard deviation, maximums, minimums, and unique values given an initial behaviour of the data and could reveal some initial issues with the dataset. Tables 4 and 5 show the results of the analysis for both datasets.

Regarding the codes and ID's values, they are not important for the analysis since they represent unique values for each student or course. Considering that, the only measurable numeric variable is *Final Grade*, where we can notice that the scale goes from zero to twenty and the mean is fifteen values. Comparing the *fullname* and *DS_Discip_PT*, which represent the name of the course, after some imputation explained in the following, they have the same number of unique values so there are no inconsistences values. It is important to notice that *Final Grade, ds_nationality* and *fullname* have missing values.

Table 4 - Descriptive Statistics Table for categorical variables

| Variable name | Count | Unique |
|---|---|---|
| **eventname** | 2919589 | 115 |
| **component** | 2919589 | 21 |
| **action** | 2919589 | 30 |
| **target** | 2919589 | 62 |
| **fullname** | 1685689 | 31 |
| **DS_Discip_PT** | 7884 | 31 |
| **DS_Discip_EN** | 7884 | 31 |
| **Gender** | 7884 | 2 |
| **DS_nationality** | 7884 | 55 |
| **DS_Natural** | 7698 | 294 |
| **Status** | 7884 | 7 |

Table 5 - Descriptive Statistics Table for numerical variables

| Variable | Count | Mean | Std | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|---|---|
| **Final Grade** | 2919589 | 15.41 | 3.76 | 0.0 | 14.0 | 16.0 | 18.0 | 20.0 |

Following this inquiry, several of visualizations were created to improve our understanding of the data. Visualization is critical in identifying insights hidden inside datasets, hence, some plots have been developed throughout the following sections to highlight relevant results.

As been seen before, we are studying the logs of 717 distinct students. The gender variable was examined by counting the various occurrences in the dataset. As we can notice in Figure 2, there are more male students (3951 students) than female ones (2300 students) in our dataset.

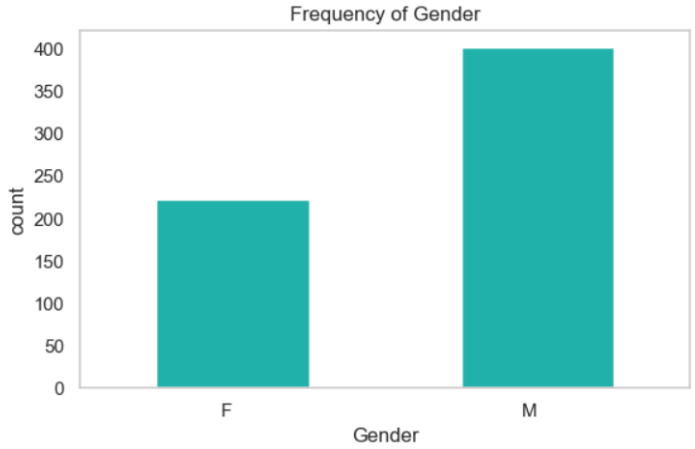



Figure 2 - Distribution of students by gender

Considering the distribution of the final grades (Figure 3), it is possible to notice that the histogram presents a distribution slightly skewed to the left, where the frequency increases as the value of the grades increases until reaching the peak of the grades. The grade peak occurs at 17 values, with a frequency of 918. It is also noted that there was a small number of students with grades between zero and nine values.

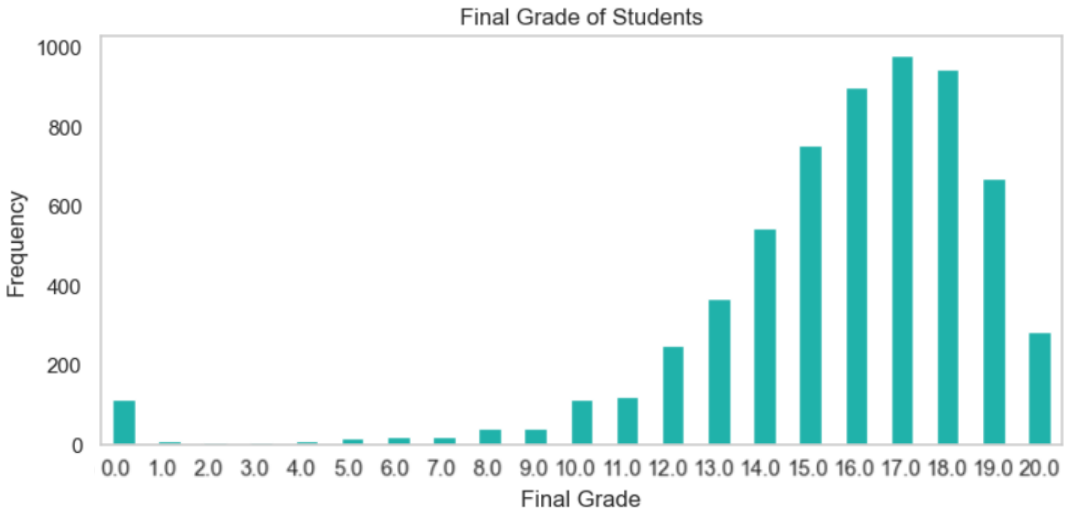



Figure 3 - Distribution of final grades of students

Given this, the most common status (Figure 4) is *Approved*, followed by *Reproved*. Right from the start we can observe that the future target of the dataset is not balanced.

As previously stated, because the other categories have minimum values, they will be replaced entirely with "Approved" or "Not Approved".

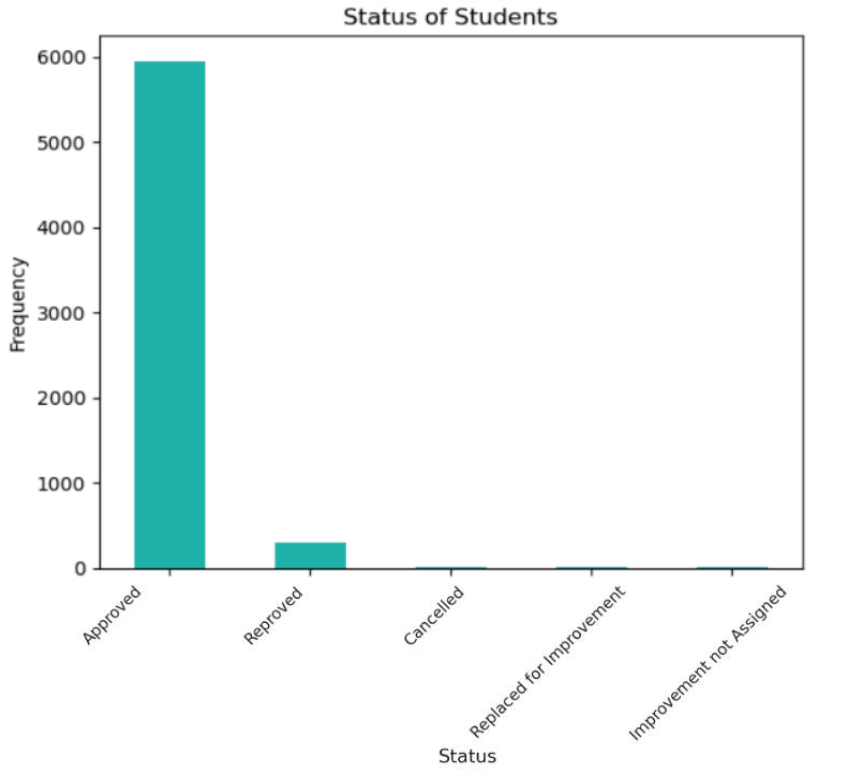



Figure 4 - Distribution of student status

When we analyze the number of logs over time (Figure 5), as expected the month with the highest frequency and with the principal peaks is January (month one), since it corresponds to the exam's month and the middle of the year. In September and October (month eight and night, respectively) represent the second and third months with more logs, coinciding with the start of classes. Contrarily, the summer months are the months with almost zero logs.

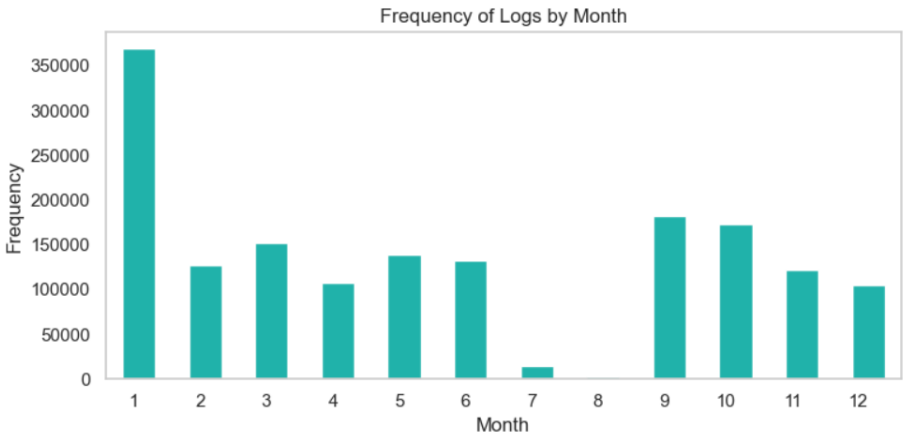



Figure 5 - Monthly distribution of log counts

The following Figure 6 shows us the activity level by weekday and hour of the day. There was generally more activity in the afternoon, emphasizing that the most intense activity occurs on Wednesday at 3 PM. As expected, there are almost no logs between 1 AM and 7 AM, and on the weekends, the activity is also less than on the other weekdays. It's also possible to notice that between 12 PM and 2 PM, there's a noticeable reduction in activity, possibly indicating a common lunch hour.

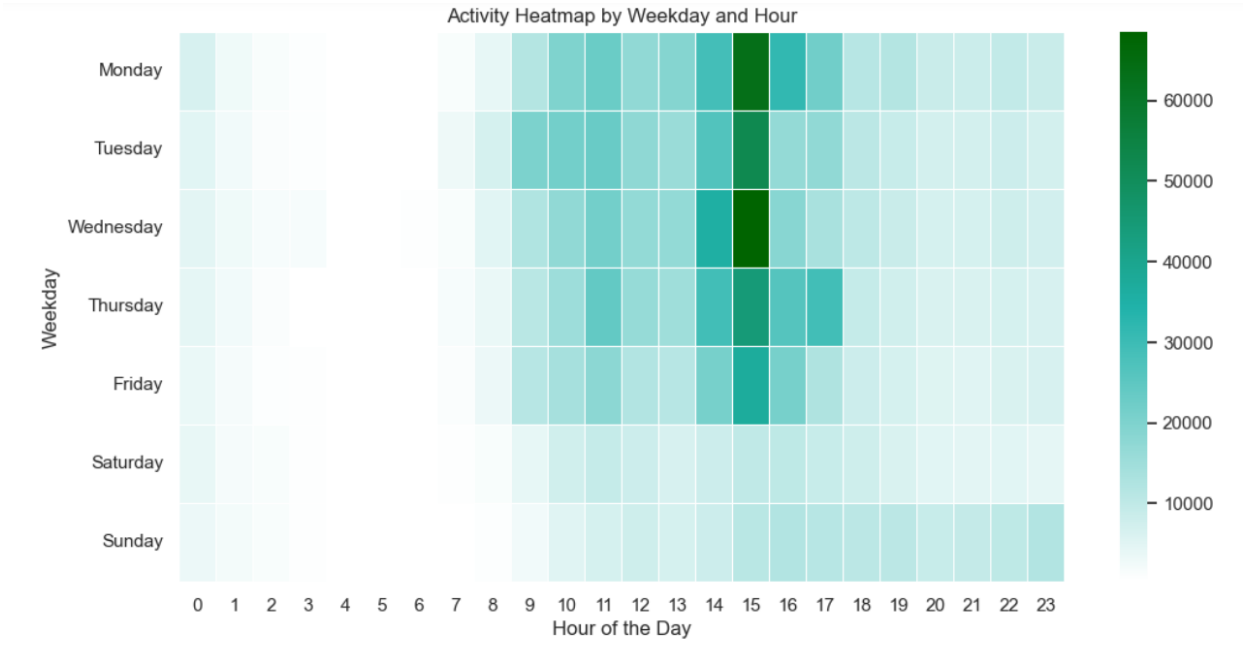



Figure 6 - Heatmap of log counts by Weekday and Hour

## 3.2 Data Preparation

After understanding the data, we discovered several mistakes and noise that must be addressed before modelling. To do so, we first prepare the data, which entails data cleaning procedures such as addressing missing values, smoothing noisy data, and eliminating outliers.

### 3.2.1 Duplicates and Irrelevant columns

Some variables are not interesting in the context of the problem, that were dropped for the analysis, as mentioned before.

Considering duplicates, the *student's data frame* had 42 duplicated values which were dropped. In order to merge the datasets in the future, only one combination of *id_student* and *discp_description_pt* was required, as a student can only have one final grade per course. This should have been seen, so the problem had to be resolved. We identified a solution by selecting only the row with the highest grade in each course for each student.

### 3.2.2 Missing Values

Missing data occurs when you don't have data stored for certain variables or participants. Data can go missing due to incomplete data entry, equipment malfunctions, lost files, and many other reasons (Pritha Bhandari, 2021).

In our data, we found three variables with missing values in both datasets, such as Final Grade, *fullname* and *cd_discip*, with 17, 42 and 44 percentage of missing values, respectively.

Several methods can be used to deal with missing values: accept, delete, and impute. In this case, we used different methods for each variable, considering their future usability. Considering the course code, we should eliminate null values to obtain a valid combination of *id_student* and *discp_description_tp* when merging the datasets. Regarding the Final Grade, we also dropped them since NULL values correspond to the student missing an assessment. Finally, the missing *fullname* values were replaced by a string " " because we didn't know the course's true name and the corresponding code was null.

### 3.2.3 Incoherence Checking

At the beginning of the analysis, incoherence was noticed. Some courses were named with odd characters, creating more than one class for each. It also had to code to define the semester and trimester. One example of that was: "EstatÃ-stica para a CiÃªncia de Dados T1" were not in the same class as "Estatística para a Ciência de Dados". The solution that was found was to impute these values to create coherent classes manually. In the same feature, there are some classes corresponding to the final thesis, and these subjects should not be analyzed with the others since they do not have different assessment methods.

By looking at the possibility of having students with a Final Grade equal to or higher than ten but with Status as "Reproved", none of these cases were seen. However, in this analysis, we might conclude that there are many classes in the *Status* variable, and it should be reduced only to "*Approved*" and "*Not Approved*".

### 3.2.4 Feature Engineering

Feature engineering involves extracting, changing, or creating new features from the original variables to improve prediction performance. It also involves creating new variables, balancing data, scaling and encoding variables.

In this technique, existing variables in the dataset are used to create additional features that improve the performance of ML models. In this case, the creation of the *age* variable derived from birth dates, the classification of a *target* variable into two classes: "*Approved*" for grades equal to or exceeding ten, and "*Not approved*" otherwise, the inclusion of *month* and *weekday* variables derived from student log data and the combination of *action* + *target* variables.

Considering that we also want to consider the first part of the course to the predictive models, a new dataset was created only with the information regarding the initial part of each course. It's important to take into account that the courses are structured differently, some are divided into trimesters, while others are in semesters, so the first part of the course is not the same period in both. From now on, all changes were made equally in both datasets.

After that, it was essential to merge the datasets to predict the students' status, considering the logs they made in Moodle. Considering that we only want one row for each combination of student and course, a different method to merge the datasets was used. The idea was to store only the information about the student (final grade, age, gender and nationality) and join the *action* and *target* variables, so the classes of action-target variables were divided into columns, where the value of each one is the number of actions-target that occurred. For example, the student with ID 5420 made three created actions in a specific target, so the value in the column was three and so on.

### 3.2.5 Train test split

In machine learning problems, it's essential to split the dataset into train and test to predict the models in the train and evaluate them in the test.

In this case, as we had the same ID of the student many times, considering all the courses that he is enrolled in, a *GroupShuffleSplit* method was used. This method ensures that the same group is not represented in both train and test datasets, i.e., the same student ID is only included in the train or test, never in both. It was used to maintain the integrity of individual student's data and ensure that the model was tested on completely unseen data, creating more reliable data.

### 3.2.6 Outliers

Outliers are values within a dataset that vary significantly from the others—they're either much larger or significantly smaller. Outliers may indicate variabilities in a measurement, experimental errors, or a novelty (Kirstie Sequitin, 2023).

Outliers can be addressed using various methods, including quantile-based algorithms, the mean or median, or manual imputing based on study objectives. Alternatively, one might choose to keep or omit the outlier results from the analysis. In this case, all numerical variables about the action and target of students have outliers (Appendix A.1). In this case, the outliers were kept considering the real number of logs of each student. Examining the *Age* variable, it was observed that there are some outliers with age greater than 41. However, given the context of the data related to a Master's, where students often come from a variety of age groups and professional backgrounds, these values were kept.

### 3.2.7 Feature Selection

The most significant features must be selected before applying the Machine Learning models. Some different methods were applied in both datasets and in the end, a combination of the output of all of them was used to select these variables.

The variables presented in Appendix A.2 and Appendix A.3 didn't show any variability in the data, i.e., only one value in all variables. Therefore, these variables were not maintained in the analysis since they did not add anything to it.

The analysis of the correlation matrix was the initial feature selection approach employed, to extract the variables with the highest correlation values with the target variable and find the most correlated independent features.

As we can observe (Appendixes A.4 and A.5), there are no high correlation values with the target. However, there are some independent variables with high correlation values between each other. Some examples of these cases are *created calendar_event* and *added booking* (1.00), *created submission* and *submitted assessable* (0.99), and *deleted calendar_event* and *removed booking* (1.00), which means that these variables were explaining the exactly or almost the same and we only need one. Given that the variables showed in Appendixes A.6 and A.7 were dropped to avoid the multicollinearity problem.

Next, we used SelectKBest analysis to identify the k characteristics that best fit the highest-scoring model. It includes a score function parameter calculated using f_classif (ANOVA), mutual_info_classif, and chi2. Each score function gave a different set of selected variables.

Finally, a Random Forest model aids in the identification of variables by assessing the purity of each node - the lower the purity, the better.

We chose which variables to keep and which to delete for each approach. The conclusion was to keep the features that appear at least in two of the previously mentioned statistical tests. In Appendixes A.8 and A.9, the final chosen variables can be observed for future use in the predictive models.

### 3.2.8 Balanced Data

Ensuing if the data is balanced in the target variable, i.e. each category has a relatively equal representation in the dataset, is a crucial step for ML models. The use of imbalanced data may lead to models that tend to favour the majority class, resulting in poor performance of the model and, consequently, low prediction accuracy. Many techniques deal with these problems, such as random under-sampling, random over-sampling and SMOTE. Under-sampling is the process of randomly eliminating instances from a majority class of a dataset and assigning them to the minority class. Contrarily, random over-sampling involves randomly duplicating the values of the minority class. Finally, SMOTE is an oversampling technique that focuses on reducing the overfitting problem in the data.

For this reason, balanced data was analyzed to ensure no discrepancies in representation in the target classes. In this particular case, we were presented with an unbalanced dataset, in which 96,6% of the values are in target 1. For this reason, a SMOTE method with a sampling strategy equal to 0.3 was used.

### 3.2.9 One-Hot Encoding & Data Scaling

To prepare for modelling, we first scaled and transformed the data. Features vary in magnitude, units, and range. We started by converting category variables into dummy variables, i.e., a new variable that takes a binary value to indicate the category's presence or absence. For example, a new gender variable was created, gender_M, with "1" if it is a male student and "0" if it is a female student. The same happened in ds_nationality_Portugal, with a "1" if the student was from Portugal and a "0" otherwise.

Finally, we scaled the dataset to correct the difference in scales of the data. There are three techniques that usually are used to solve this problem: Min-Max scaler, Standard Scaler and Robust scaler. For every model, every scaler was tested to find the best result.

The Min-Max scaler reduces the data within a range between zero and one. The transformation is given by the formula:

$$X\_scaled = \frac{\mathrm{X} - \mathrm{Xmin}}{\mathrm{Xmax} - \mathrm{Xmin}}$$

Figure 7 - Min-Max Normalization formula

, where *Xmin* and *Xmax* are the minimum and maximum values of the feature, respectively.

The Standard Scaler transform the data by removing the mean and scaling it to unit variance. The formula of this scaler is:

$$X\_standardized = \frac{\mathrm{X}-\mu}{\sigma}$$

Figure 8 - Standard Scaler formula

, where μ is the mean and α is the standard deviation of the corresponding feature.

The main idea is to standardize features with mean zero and standard deviation one.

Finally, the Robust scaler transforms the features using statistics that are robust to outliers. This removes the median and scales the data according to the quantile range. The formula is:

$$X\_robust = \frac{\mathrm{X} - \mathrm{median}}{\mathrm{IQR}}$$

Figure 9 - Robust Scaler formula

## 3.3 MODELLING

Once all pre-processing stages have been tackled, and since we are currently working with predictive modelling, explicitly dealing with classification problems, five *Machine Learning* Algorithms were used to find the model with the best performance. It is essential to note that all algorithms were implemented using the *Scikit-Learn* library from *Python*. The best parameters for each model were obtained through the Hyperparameter Tuning method. The following pages gave us an overview of models and the parameters used for each.

### 3.3.1 Random Forest

*Random Forest* is an ensemble learning algorithm that builds multiple decision trees and merges them to get a more accurate prediction finding of the average. *Decision trees* exhibit a hierarchical structure as a tree containing a root node, branches, internal nodes, and leaf nodes. The following figure 10 shows an example of a Random Forest Model to understand this concept better.

*Random Forest* is a known algorithm for its ability to handle high dimensional data, flexibility, and robustness and to reduce the risk of overfitting. It is widely used for both classification and regression problems.

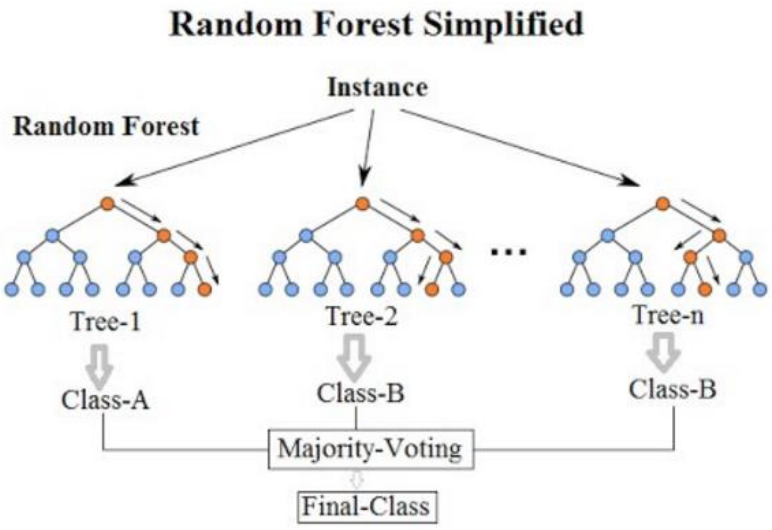



Figure 10 - Random Forest representation from Kannan, A., Kolovich, B., Lawrence, B., & Rafiqi, S. (2018)

### 3.3.2 Logistic Regression

*Logistic Regression* is a statistical method used for binary classification problems, where the probability of a binary event occurring is calculated based on one or more predictor variables. Logistic Regression analyzes relationships between variables and is trained using a method called *Maximum Likelihood Estimation* (MLE), where the objective is to find the values of the coefficients that maximize the probability of observing the given set of outcomes in the data.

Typically, a threshold of 0.5 is applied to a decision boundary, indicating that if the probability is greater or equal to 0.5, the class is one and class 0 otherwise.

### 3.3.3 Gradient Boost

*Gradient Boost* is an ensemble technique that combines multiple weak models to create a robust predictive model, i.e., this method aims to decrease prediction error in the best possible next model after combining previous models. Its main objective is to iteratively train new models, considering and improving the errors made by its predecessors. In each interaction, the algorithm calculates the gradient of the loss function of the current ensemble and then trains a new model to minimize this gradient.

This model has shown to be effective since it can be used for both regression and classification problems, and it can handle complex and non-linear relationships within the data.

### 3.3.4 Support Vector Machine

*Support Vector Machine (SV)* is a powerful algorithm that aims to find an optimal hyperplane in a high dimensional space that separates the dimensional space into categories, classifying new data points. Many possible hyperplanes could be chosen, but the goal is to find the maximum margin between the points of each class, where the margin is the distance between the hyperplane and the nearest data points.

SVM is capable of handle with complex and non-linear relationships in the data using kernel functions. It is functional in high-dimensional data, and it's capable of performing well in diverse datasets.

### 3.3.5 Neural Networks

This ML model draws inspiration from the human brain, i.e., biological neurons and their messages. Neural networks consist of at least three interconnected node layers: input, hidden (one or more), and output. Training data are used to increase the accuracy of this model.

The input variables are weighted to represent the importance of each feature. The total input, including previous values, is converted to output using an activation function (e.g., sigmoid function). Combining multiple factors leads to more accurate forecasts (Figure 11).

Neural networks are versatile and can be applied to various tasks, including image and speech recognition, natural language processing, and game playing. It has been a successful model since it can handle complex and high-dimensional data.

To fit the model to our data, we used the MLPClassifier (Multi-layer Perceptron Classifier), which employs Neural Networks for classification problems.

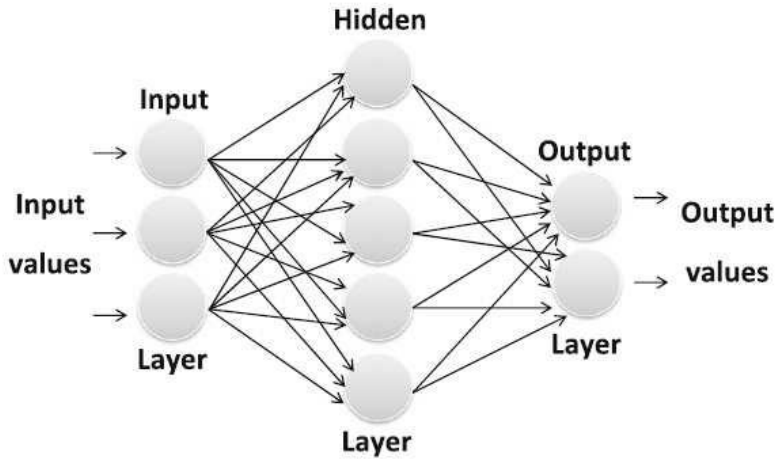



Figure 11 - Neural Networks representation from Bozkus, E. (2021)

### 3.3.6 Hyperparameter Tuning

Hyper-parameters are the parameters used to either configure a ML model or to specify the algorithm used to minimize the loss function (Yang & Shami, 2020). Tuning the hyper-parameters is a crucial key for data scientists since they directly control the behaviours of training algorithms and have a significant effect on the performance of the models (Wu et al., 2019). The process of Hyperparameter tuning is considered iterative since you try out all different combinations of parameters and values until find the model with the highest performance.

With this in mind, a *Grid Search Cross Validation* with 3-Fold was implemented with *Scikit-Learn* Library to find the best performance in all models. The following tables, show the different hyperparameters that were tested in each ML model.

Table 6 - Hyperparameter Tuning Table for Random Forest

| Parameter | Values | Description |
|---|---|---|
| **n_estimators** | 100, 200, 300 | The number of trees in the forest |
| **max_depth** | None, 10, 20, 30 | The maximum depth of the tree. |
| **min_samples_split** | 2, 5, 10 | The minimum number of samples required to split an internal node |
| **criterion** | ‘gini’, ‘entropy’, ‘log_loss’ | The function to measure the quality of a split. |
| **max_features** | ‘sqrt’, ‘log2’, None | The number of features to consider when looking for the best split |
| **Min_samples_leaf** | 1, 2, 4 | The minimum number of samples required to be at a leaf node. |

Table 7 - Hyperparameter Tuning Table for Logistic Regression

| Parameter | Values | Description |
|---|---|---|
| **C** | 0.01, 0.1, 1, 10, 100 | Inverse of regularization strength |
| **penalty** | 'l1', 'l2' | Specify the norm of the penalty |
| **solver** | ''liblinear', 'sag', 'saga' | Algorithm to use in the optimization problem. |

Table 8 - Hyperparameter Tuning Table for Gradient Boost

| Parameter | Values | Description |
|---|---|---|
| **n_estimators** | 100, 200, 300 | The number of boosting stages to perform |
| **learning_rate** | 0.01, 0.1, 0.2 | Learning rate shrinks the contribution of each tree by learning_rate |
| **min_samples_split** | 2, 5, 10 | The minimum number of samples required to split an internal node |
| **max_depth** | 3, 5, 10 | Maximum depth of the individual regression estimators |
| **Min_samples_leaf** | 1, 2, 4 | The minimum number of samples required to be at a leaf node. |

Table 9 - Hyperparameter Tuning Table for SVM

| Parameter | Values | Description |
|---|---|---|
| **C** | 0.1, 1, 10 | Regularization parameter. The strength of the regularization is inversely proportional to C. |
| **gamma** | 'scale', 'auto' | Kernel coefficient |
| **kernel** | 'rbf', 'linear', 'poly' | Specifies the kernel type to be used in the algorithm. |

Table 10 - Hyperparameter Tuning Table for Neural Networks

| Parameter | Values | Description |
|---|---|---|
| **solver** | 'adam', 'sgd' | The solver for weight optimization. |
| **alpha** | 0.0001, 0.001, 0.01 | Strength of the L2 regularization term. |
| **activation** | 'relu', 'tanh' | Activation function for the hidden layer. |

| | | |
|---|---|---|
| **hidden_layer_sizes** | (50,), (100,), (50,50) | The ith element represents the number of neurons in the ith hidden layer. |
| **Learning_rate** | 'constant', 'adaptive' | Learning rate schedule for weight updates. |

### 3.3.7 Metrics

Classification metrics are evaluation metrics used to measure the performance of a model.

When a model is created, it is necessary to compare their performance to the already existing ones. Evaluation serves two purposes: methods that do not perform well can be discarded, and the ones that seem promising can be further optimized (Rainio & Teuho, 2024).

There are several metrics that can be used in a classification problem: Accuracy, Precision, Recall, F1-score, and ROC Curve.

**Accuracy** is a metric that measures the proportion of correct predictions (true positives and true negatives) compared to the total number of predictions made in a Machine Learning model.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Figure 12 – Accuracy metric formula

**Precision** measures how often a model correctly predicts the positive class of the target, i.e. the proportion of true positives in all positive predictions.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Figure 13 - Precision metric formula

In another way, **Recall** is a metric that assesses the correctly identified positive instances (true positives) from all positive instances in the dataset (true positives and false negatives).

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Figure 14 – Recall metric formula

The **F1-score** metric uses both precision and recall. It calculates the harmonic mean of the precision and recall of a model. This metric is quite similar to accuracy, but the F1-score is better in imbalanced classes in the target variable since it doesn't treat all types of correct and incorrect classifications equally.

$$F1 - score = \frac{2 \text{ x Precision x Recall}}{\text{Precision} + \text{Recall}}$$

Figure 15 - F1 score metric formula

The **ROC Curve** is a graph that illustrates the performance of a model using a true positive rate against the false positive rate and **AUC** is the Area under the ROC Curve that provides a single measure of the model's performance across all classification thresholds.

### 3.3.8 Overfitting

Overfitting is a common problem that occurs in ML models. It is a challenge where a model fits nicely on the training set while performing poorly on testing data. This means that the model learns the details and noise of the train, which negatively impacts the performance of new data. The primary indicator of overfitting is a significant gap between the performance of the training data and the performance of the test data (Ying, 2019).

Recognizing overfitting is a crucial skill for any *Data Scientist* creating predictive models to prevent and treat this issue. There are several ways to deal with overfitting, such as Cross-validation, feature selection, L1 / L2 regularization, and Dropout. The problem of overfitting was addressed in all the following developed models.

### 3.3.9 Feature Importance

At the end of the modelling part, a SHAP methodology was used to explain the importance of each feature in the context of the predictive model. SHAP values show how each feature affects the final prediction when compared to the others. This was essential to understanding the variables that influence the results of the model (Lundberg & Lee, 2017).

# 4. RESULTS AND DISCUSSION

Using datasets collected from Moodle log data and student information, this study applies algorithms and techniques to identify patterns and predictors of student performance.

The main objective of the present study is to develop a predictive model to find students at risk of failure in the whole and in the first part of the course.

The results from this analysis contribute to the emerging fields of educational Data Mining and provide practical insights for educators and administrators to improve student's engagement and educational outcomes. Our approach includes different analytical strategies, including an initial analysis of the student's behaviour and five different ML models, to find the most robust one in predicting student success. The following section presents the principal results retrieved from the whole analysis.

In the following Table 11, it is provided the results of the performance of various ML models on both training and testing datasets across the metrics referred before: Accuracy, F-score, and AUC. When developing the predictive models using the entire available dataset, all ML models demonstrated good results across various scaling methods. Considering the presence of outliers in the dataset, the scaling method chosen was Robust Scaler since it is resilient to outliers.

The *Random Forest* model showed the best performance, achieving perfect scores in all considering metrics in training (1.0 in accuracy, F-score and AUC). In testing, accuracy and F-score also showed almost perfect results (0.97, 0.98 respectively), but AUC decreases for values of 0.75, possible indicating a problem of overfitting.

Although with lower values than *Random Forest*, the *Gradient Boost* model also stands out for an exceptional performance in train. The model had all metric values equal to 1.0, indicating a strong predictive capability of the model. As it happened in the previous model, a problem of overfitting was viewed with an AUC value of 0.68 in the test.

While presenting more modest scores than the other models, *Logistic Regression*, *Neural Networks* and SVM maintained consistent performance in all metrics. However, in accuracy and F-score metrics, these models presented worst results in the training. The F-score is slightly higher, meaning there was an outstanding balance between precision and recall. As seen in the other models above, the performance in the test, when considering the AUC metric, showed a problem of overfitting with values of 0.69, 0.65 and 0.71 in *Logistic Regression*, SVM and *Neural Networks*, respectively.

Table 11 - All course Table Results

| Model | Train/Test | Accuracy | F-score | AUC |
|---|---|---|---|---|
| **Random Forest** | Train | 1.0 | 1.0 | 1.0 |
| **Random Forest** | Test | 0.97 | 0.98 | 0.75 |
| **Logistic Regression** | Train | 0.85 | 0.91 | 0.87 |
| **Logistic Regression** | Test | 0.96 | 0.98 | 0.69 |
| **Gradient Boost** | Train | 1.0 | 1.0 | 1.0 |
| **Gradient Boost** | Test | 0.97 | 0.98 | 0.68 |
| **SVM** | Train | 0.92 | 0.95 | 0.95 |
| **SVM** | Test | 0.96 | 0.98 | 0.65 |
| **Neural Networks** | Train | 0.96 | 0.97 | 0.99 |
| **Neural Networks** | Test | 0.94 | 0.97 | 0.71 |

Considering that Random Forest presented the best results, from now on a feature importance analysis will be done considering only the chosen model.

The chart based on SHAP values in Figure 16 visualizes the importance of the different features in the Random Forest model. We can observe that some student interactions feature such as *viewed course*, *downloaded all files* and *viewed attempt* demonstrated high variability in influencing the model across different instances.

Some specific nationalities like Portugal and Brazil, showed a significant negative impact on the results. When observing the *age* variable, we can notice that it had a concentration of points toward lower feature values, indicating that lower values on age contribute positively to the predictive model. Contrarily, *view_submission_status* and *submission add* had a significant positive impact on the results of the students.

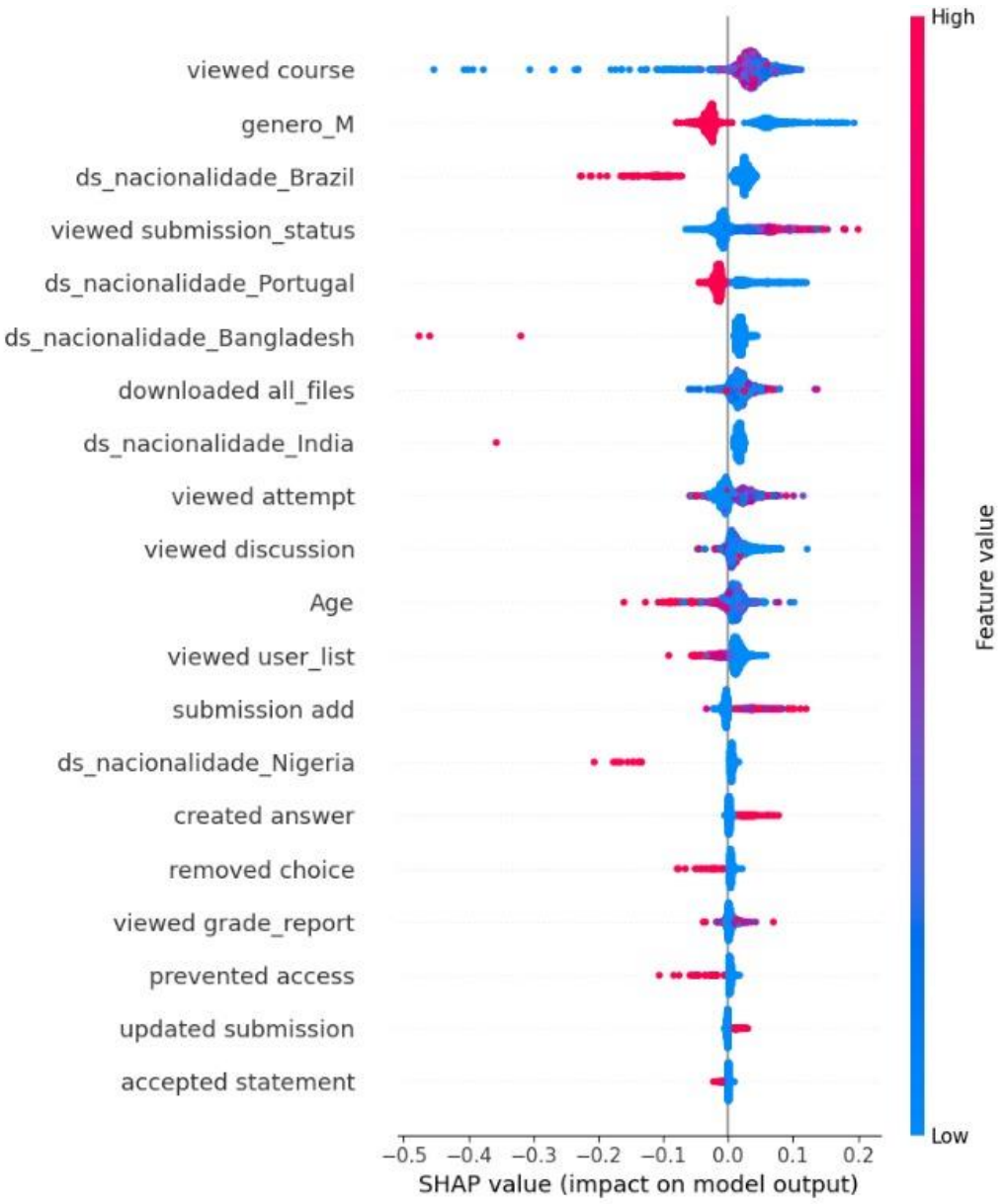



Figure 16 - Feature importance for Approved students in all course

When looking at the waterfall plot presented below (Figure 17), it's possible to take some considerations of how a feature value of a single instance contributes directly to the model prediction. In this case, it is possible to observe that *viewed course* = 81, *submission add* = 2 and *created answer* = 1 contributed positively to the final prediction (+0.04, +0.03 and +0.02, respectively). This could mean that high engagement and participation could be a strong indicator of approval. Contrarily, being a male student has a slight negative contribution to the model, showing a minor impact on the final prediction.

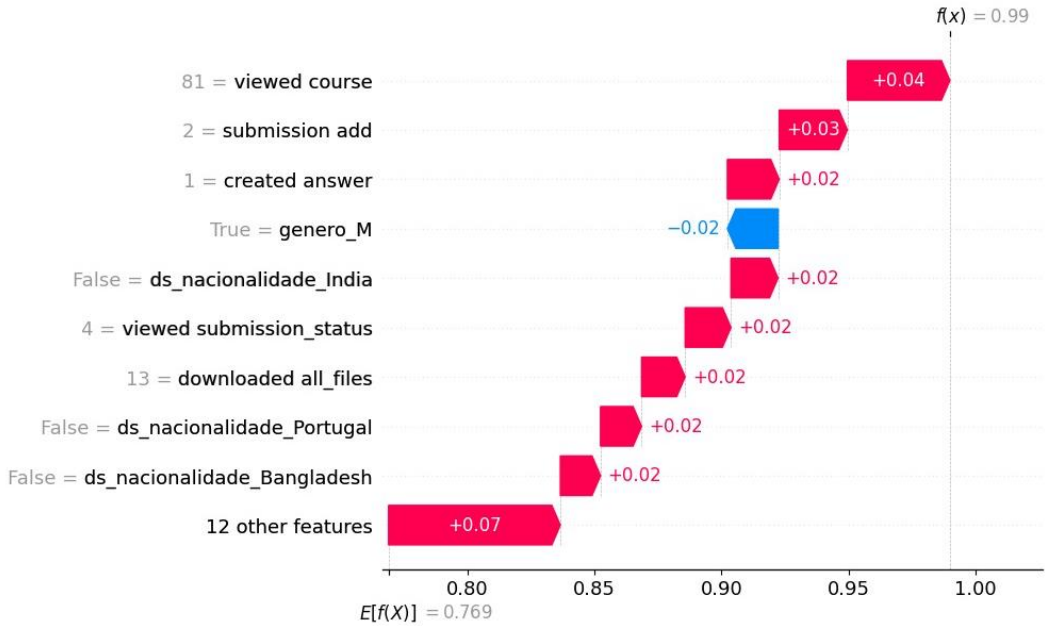



Figure 17 - Waterfall plot of a single instance for Approved students in all course

Considering only the first part of the course, the results are in the following Table 12.

As we can observe, the results were generally lower compared with the log data of the entire course. As it happened in the previously developed models, the *Random Forest* proved to be the model with the best performance with nearly perfect scores in training data. However, it had very low values in AUC in the test (0.62) giving a strong indication of possible overfitting in the model.

The *Logistic Regression* model showed a different behavior. In Accuracy and F-score metrics, they both performed better in the test, showing possible evidence of underfitting. Considering AUC as it happened in the other models, it showed worst results in the test (0.70) possibly indicating overfitting.

When looking at the *Gradient Boost* model, shows coherent values in train and test considering accuracy and F-score metrics. Contrarily, observing the AUC metric, there's evidence of overfitting in data since it has a value of 0.98 in train and 0.53 in the test.

Finally, *SVM* and *Neural Networks* also showed evidence of overfitting and underfitting. Both showed strong test accuracy and F-score values.

Table 12 - First part of the course Table Results

| Model | Train/Test | Accuracy | F-score | AUC |
|---|---|---|---|---|
| **Random Forest** | Train | 0.94 | 0.96 | 0.98 |
| **Random Forest** | Test | 0.94 | 0.97 | 0.62 |
| **Logistic Regression** | Train | 0.83 | 0.90 | 0.78 |
| **Logistic Regression** | Test | 0.94 | 0.97 | 0.70 |
| **Gradient Boost** | Train | 0.94 | 0.96 | 0.98 |
| **Gradient Boost** | Test | 0.94 | 0.96 | 0.53 |
| **SVM** | Train | 0.87 | 0.92 | 0.90 |
| **SVM** | Test | 0.94 | 0.97 | 0.59 |
| **Neural Networks** | Train | 0.91 | 0.94 | 0.95 |
| **Neural Networks** | Test | 0.92 | 0.96 | 0.65 |

In the following graph (Figure 18), it is possible to observe a Summary Plot Analysis from Shap Values that provides a comprehensive visualization of how the different features influence the models' predictions. The *age* feature has a significant impact on the model, with a wide distribution. Lower values of age tend to have a positive impact on the success of the students.

Some features that correspond to engagement with course material such as *added group_member*, *viewed discussion*, *downloaded all_files* and *updated choice* significantly influence the model's prediction. This suggests the importance of promoting active engagement for the students with the course material.

Some nationalities such as Bangladesh and India appear to slightly negative impact on the model. In the same way, being from Portugal slightly decreases the probability of academic success.

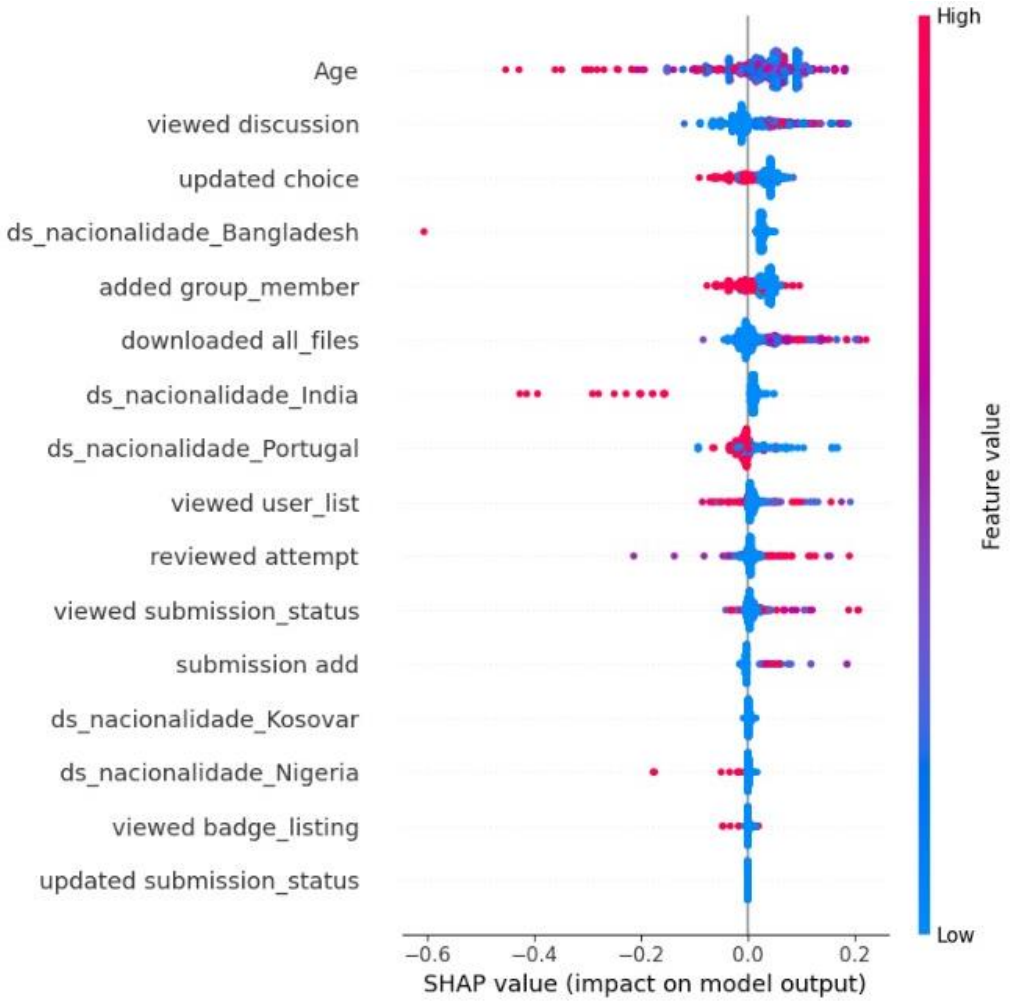



Figure 18 - Feature importance for Approved students in the first part of the course

Taking into account the following Figure 19, we can observe that the most significant impact in the model is *downloaded all_files* = 10 (+0.05), indicating that students who downloaded all files in the first part of the course are more likely to succeed.

Social demographic features such as *age* = 25, *ds_nationality_Bangladesh* and *ds_nationality_India* equal to False slight positive impact the model's performance.

Other actions that contributed positively to the models were *added group_member* = 0 and *view discussion* = 1, meaning that engaging with Moodle is beneficial.

We also can observe that being a Portuguese student slightly decreases the likelihood of passing the course (-0.01).

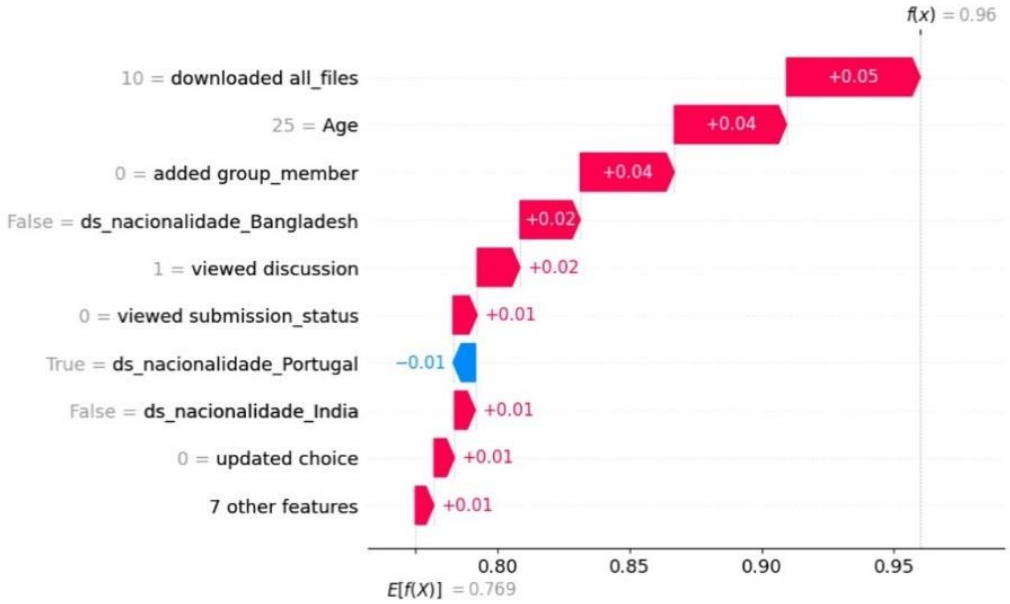



Figure 19 - Waterfall plot of a single instance for Approved students in the first part of the course

The results indicate that models like *Random Forest* and *Gradient Boost* are effectively able to distinguish between the two classes in training data. However, the value in the test indicates that the model performance degraded when applied to new unseen data. It suggests a problem of overfitting. The other models presented more modest results, with a problem of Overfitting and Underfitting, i.e., the model may not have enough complexity to capture the patterns in the train dataset.

The findings suggest that early and consistent interaction with Moodle is a strong indicator of academic success. This implies that educators should focus on increasing student interaction with Moodle to improve outcomes.  It's also important to consider personalized approaches in E-learning environments to be accessible to everyone in every background, considering social demographic factors.

# 5. CONCLUSIONS

The main objective of this thesis was to develop a predictive model capable of classifying whether a student will pass or fail the course in the first and the whole course, based on Moodle log data and social demographic information about the students. Furthermore, the objective was to support educators and administrators by creating a strategy to communicate the results of the model. This study demonstrates the potential of combining Moodle log data and *Learning Analytics* to predict academic achievement in university students.

To address this challenge, a methodology of *CRISP-DM* was used and a combination of five Machine Learning models was developed: *Random Forest, Logistic Regression, Gradient Boost, SVM and Neural Networks.*

The study showed that we developed robust models capable of identifying students at risk of failure, even so with a present of overfitting and underfitting. The principal results in the first part and in the whole course indicate that *Random Forest* and *Gradient Boost* algorithms provided the best accurate predictions, always with train values above 0.9 in all metrics. However, overfitting is a concern that needs to be addressed in future research.

Our results revealed that early and consistent interaction with Moodle is a strong indicator that influences academic success. Students who frequently engage with course materials actively (e.g., *viewed courses, downloaded files, viewed discussion forum and updated choice*) tend to demonstrate better results. Demographic factors, such as *age* and *nationality* also played a crucial role in impacting the predictive model.

Regarding the first research question defined earlier "How can Moodle log data and Learning Analytics be effectively integrated to predict students academic success in the first part and the entire course?" the study demonstrated the models provided a comprehensive view of the student's behaviours and the highlights the importance of the interactions with the Moodle platform. The predictive models had poor performance in the first part of the course, but even so, both are capable of identifying students who are less engaged with the course, allowing for timely interventions and strategies.

Considering the second question "How can predictive modelling results be effectively communicated to educators and administrators to support intervention strategies and improve student outcomes?" the study shows the importance of giving clear insights to educators. By highlighting some metrics, such as engagement to Moodle and the demographic ones, educators can better understand which students are at risk and come up with early strategies to help them.

To conclude, this assumes that teachers should consider everyone's background and find ways to encourage students to have daily contact with Moodle. The educators can adopt several strategies to encourage active participation in the Moodle platform through discussion forums, frequent and mandatory discussion posts and quizzes. They should ensure that the

course content is interactive and engaging, for example by using multimedia elements like videos, photos and quizzes or even create a competition system with rewards for those who regularly engage and complete the activities presented on the platform. They can also develop personalized plans based on student engagement and demographic insights, by providing additional support. It's also important to maintain available lines of communication to provide regular feedback and updates on the student performance or even maintain updated the announcement feature of Moodle.

By explicitly linking the findings to the research questions, this thesis demonstrates how *Moodle log data* and *Learning Analytics* can be effectively used to predict academic success as well as how the results of predictive modeling can be communicated to improve educational outcomes.

# 6. LIMITATIONS AND FUTURE WORK

While this study provided useful information on predicting academic success, it is important to take into account that there were some limitations. The extent of the analysis was limited by the computer resources available. The dimensionality of the dataset and the complexity of the predictive models needed extensive memory and processing resources. As a result, it was not possible to perform an extensive hyperparameter tuning process, which could potentially provide better results. Considering this, future studies could consider using more powerful computing resources in order to try more different combinations of hyperparameters.

Considering the findings of this Master Thesis, several directions for further research and studies were proposed. One interesting strategy is to develop an early warning system capable of identifying students who are at risk of failure early in the semester using predictive models. By providing this approach, this system could support educators and administrators to find and help students before it is too late.

Another possible strategy to implement is to create a comprehensive dashboard for professors. The dashboard could provide real-time data on student engagement, performance and progression. This might lead to a strategy for educators adjust their tactics of teaching and provide personal help to students who may be at risk.

Moreover, another promising direction could be including multimodal analysis to give more valuable insights from the data. Multimodal analysis could use data from various sources such as textual material, video recordings, audio feedback and so on.

Lastly, another direction for future work to improve the performance of the ML models involves the inclusion of more data. Since we just consider one master's from NOVA IMS University, a wide range of additional degrees and master's can be used to provide better insights. This can provide more volume of data with diverse learning contexts.

# BIBLIOGRAPHICAL REFERENCES


Abu Talib, M., Bettayeb, A. M., & Omer, R. I. (2021). Analytical study on the impact of technology in higher education during the age of COVID-19: Systematic literature review. *Education and Information Technologies*, 26(6). https://doi.org/10.1007/s10639-021-10507-1

Accuracy vs. precision vs. recall in machine learning: what's the difference? (n.d.). *Evidentlyai.com.* https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *In International Journal of Educational Technology in Higher Education* (Vol. 17, Issue 1). https://doi.org/10.1186/s41239-020-0177-7

Anagnostopoulos, T., Kytagias, C., Xanthopoulos, T., Georgakopoulos, I., Salmon, I., & Psaromiligkos, Y. (2020). Intelligent predictive analytics for identifying students at risk of failure in moodle courses. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12149 LNCS. https://doi.org/10.1007/978-3-030-49663-0_19

Banihashem, S. K., Aliabadi, K., Pourroostaei Ardakani, S., Delaver, A., & Nili Ahmadabadi, M. (2018). Learning Analytics: A Systematic Literature Review. *Interdisciplinary Journal of Virtual Learning in Medical Sciences*, 9(2). https://doi.org/10.5812/ijvlms.63024

Bhandari, P. (2023). Missing Data | Types, Explanation, & Imputation. *Scribbr*. https://www.scribbr.com/statistics/missing-data/

Bozkus, E. (2021). Evaluation of occupational accidents with artificial neural networks in occupational health and safety management systems. https://www.researchgate.net/publication/357285118_EVALUATION_OF_OCCUPATIONAL_ACCIDENTS_WITH_ARTIFICIAL_NEURAL_NETWORKS_IN_OCCUPATIONAL_HEALTH_AND_SAFETY_MANAGEMENT_SYSTEMS

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, 4(5–6). https://doi.org/10.1504/IJTEL.2012.051815

Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting student performance from LMS data: A comparison of 17 blended courses using moodle LMS. *IEEE Transactions on Learning Technologies*, 10(1). https://doi.org/10.1109/TLT.2016.2616312

Gaftandzhieva, S., Talukder, A., Gohain, N., Hussain, S., Theodorou, P., Salal, Y. K., & Doneva, R. (2022). Exploring Online Activities to Predict the Final Grade of Student. *Mathematics*, 10(20). https://doi.org/10.3390/math10203758

Gogan, M. L., Sirbu, R., & Draghici, A. (2015). Aspects Concerning the Use of the Moodle Platform – Case Study. *Procedia Technology*, 19. https://doi.org/10.1016/j.protcy.2015.02.163

Hotz, N. (2023). What is CRISP DM? Data Science Process Alliance. https://www.datascience-pm.com/crisp-dm-2/

Kadoic, N., & Oreski, D. (2018). Analysis of student behavior and success based on logs in Moodle. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018 - Proceedings*. https://doi.org/10.23919/MIPRO.2018.8400123

Kaensar, C., & Wongnin, W. (2023). Analysis and Prediction of Student Performance Based on Moodle Log Data using Machine Learning Techniques. *International Journal of Emerging Technologies in Learning*, 18(10). https://doi.org/10.3991/ijet.v18i10.35841

Kannan, A., Kolovich, B., Lawrence, B., & Rafiqi, S. (2018). Predicting National Basketball Association success: A machine learning approach. *SMU Data Science Review*, 1(3), Article 7. https://scholar.smu.edu/datasciencereview/vol1/iss3/7

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *In Proceedings*. http://arxiv.org/abs/1705.07874

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., & Flach, P. (2019). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/tkde.2019.2962680

Matijašević-Obradović, J., Dragojlović, J., & Babović, S. (2017). The Importance of Distance Learning and the Use of Moodle Educational Platform in Education. https://doi.org/10.15308/sinteza-2017-236-241

Mwalumbwe, I., & Mtebe, J. S. (2017). Using learning analytics to predict students' performance in moodle learning management system: A case of Mbeya University of science and technology. *Electronic Journal of Information Systems in Developing Countries*, 79(1). https://doi.org/10.1002/j.1681-4835.2017.tb00577.x

Quinn, R. J., & Gray, G. (2019). Prediction of student academic performance using Moodle data from a Further Education setting. *Irish Journal of Technology Enhanced Learning*, 5(1). https://doi.org/10.22554/ijtel.v5i1.57

Reimondo Tamba, A., Lumbantoruan, K., Pakpahan, A., & Situmeang, S. (2023). A cluster and association analysis visualization using Moodle activity log data. *International Journal of Informatics and Communication Technology (IJ-ICT)*, 12(2). https://doi.org/10.11591/ijict.v12i2.pp150-161

Robbins, S. B., Le, H., Davis, D., Lauver, K., Langley, R., & Carlstrom, A. (2004). Do Psychosocial and Study Skill Factors Predict College Outcomes? *A Meta-Analysis. Psychological Bulletin*, 130(2). https://doi.org/10.1037/0033-2909.130.2.261

scikit-learn. (2019). sklearn.preprocessing.MinMaxScaler — scikit-learn 0.22.1 documentation. *Scikit-Learn.org.* https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

Sequitin, K. (2021). Data Analytics Explained: What Is an Outlier? *Careerfoundry.com.* https://careerfoundry.com/en/blog/data-analytics/what-is-an-outlier/

Siemens, G., & Baker, R. S. J. D. (2012). Learning analytics and educational data mining: Towards communication and collaboration. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/2330601.2330661

Umbach, P. D., & Wawrzynski, M. R. (2005). Faculty do matter: The role of college faculty in student learning and engagement. *Research in Higher Education*, 46(2). https://doi.org/10.1007/s11162-004-1598-1

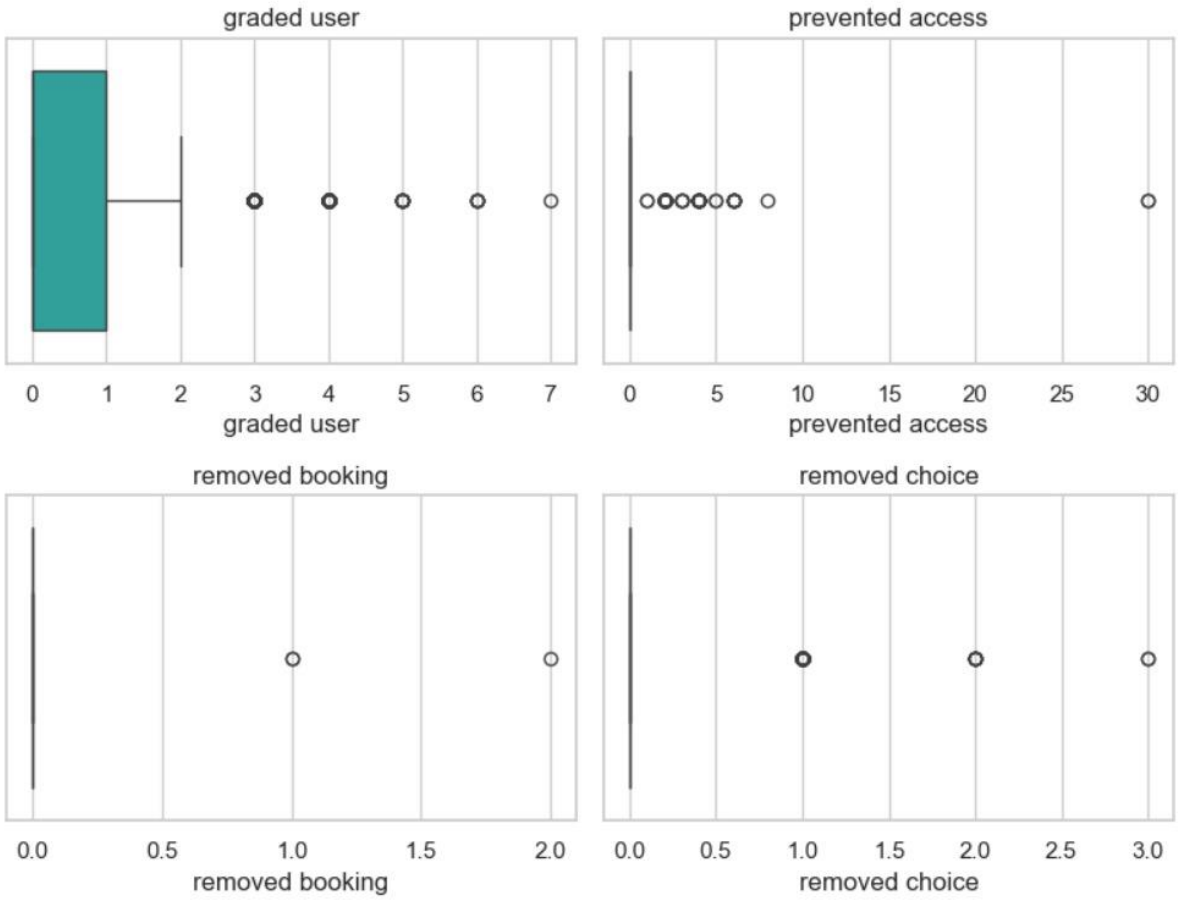
Wu, J., Chen, X. Y., Zhang, H., Xiong, L. D., Lei, H., & Deng, S. H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology*, 17(1). https://doi.org/10.11989/JEST.1674-862X.80904120

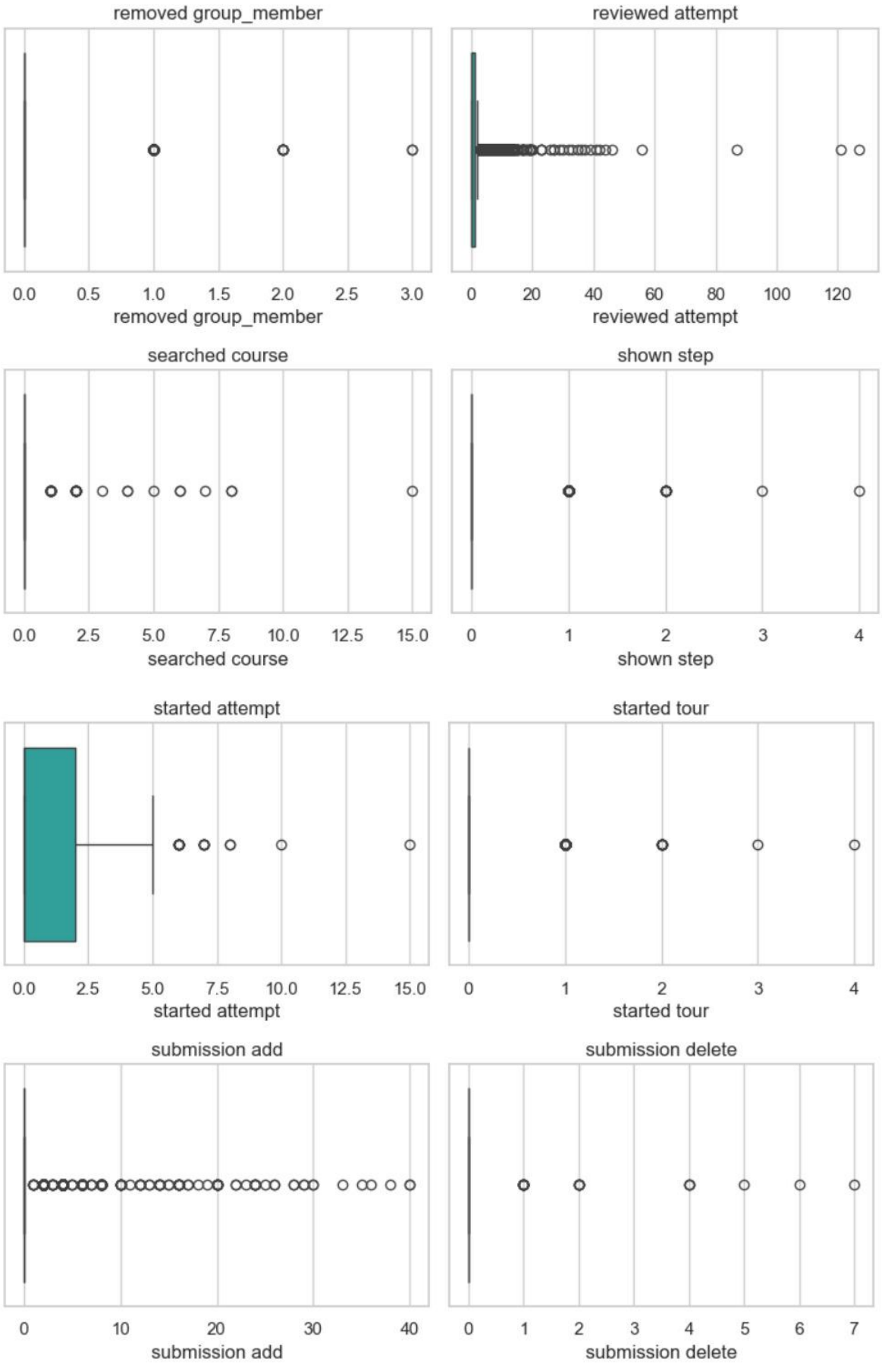
Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415. https://doi.org/10.1016/j.neucom.2020.07.061

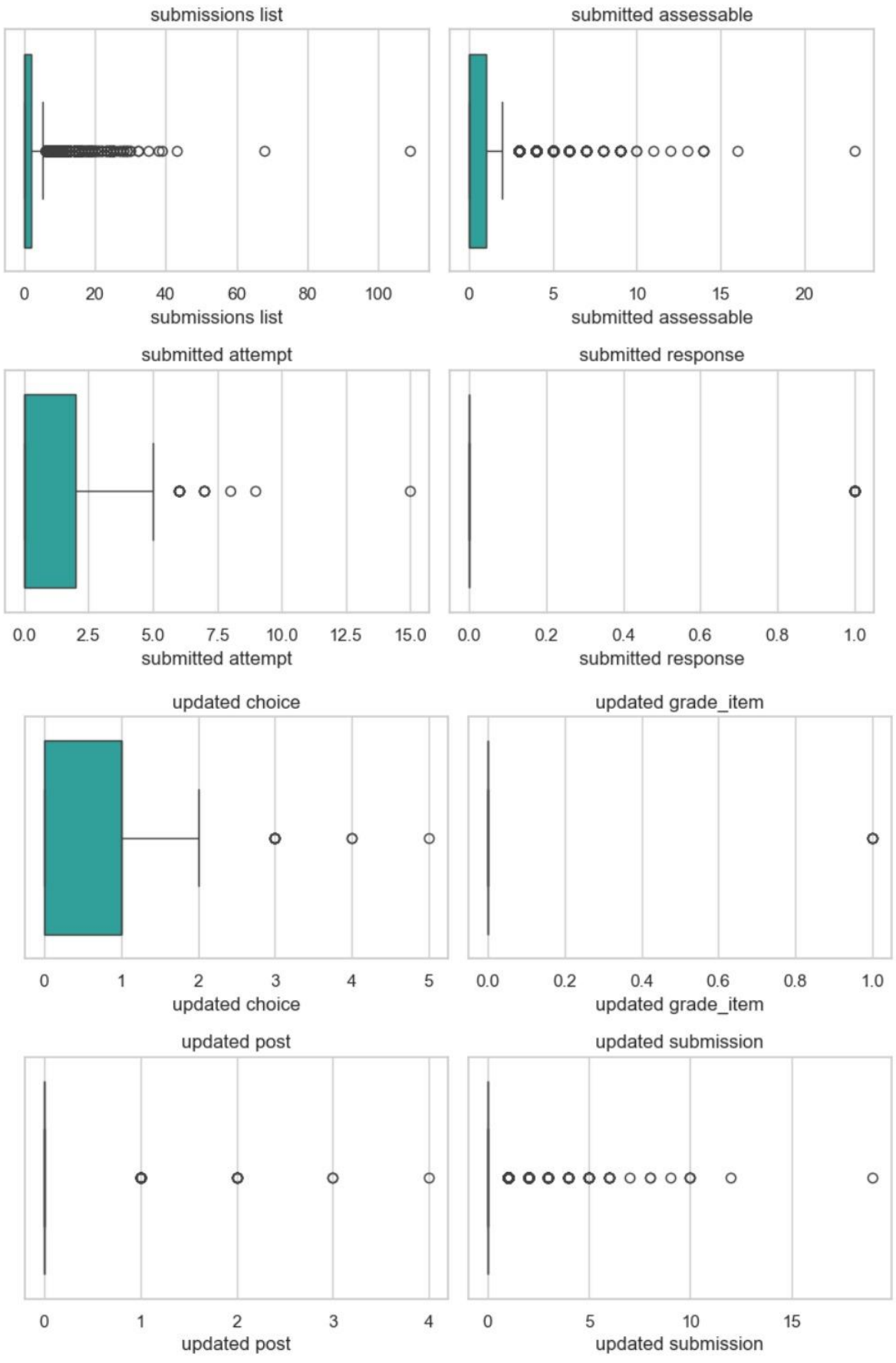
Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, 1168(2). https://doi.org/10.1088/1742-6596/1168/2/022022

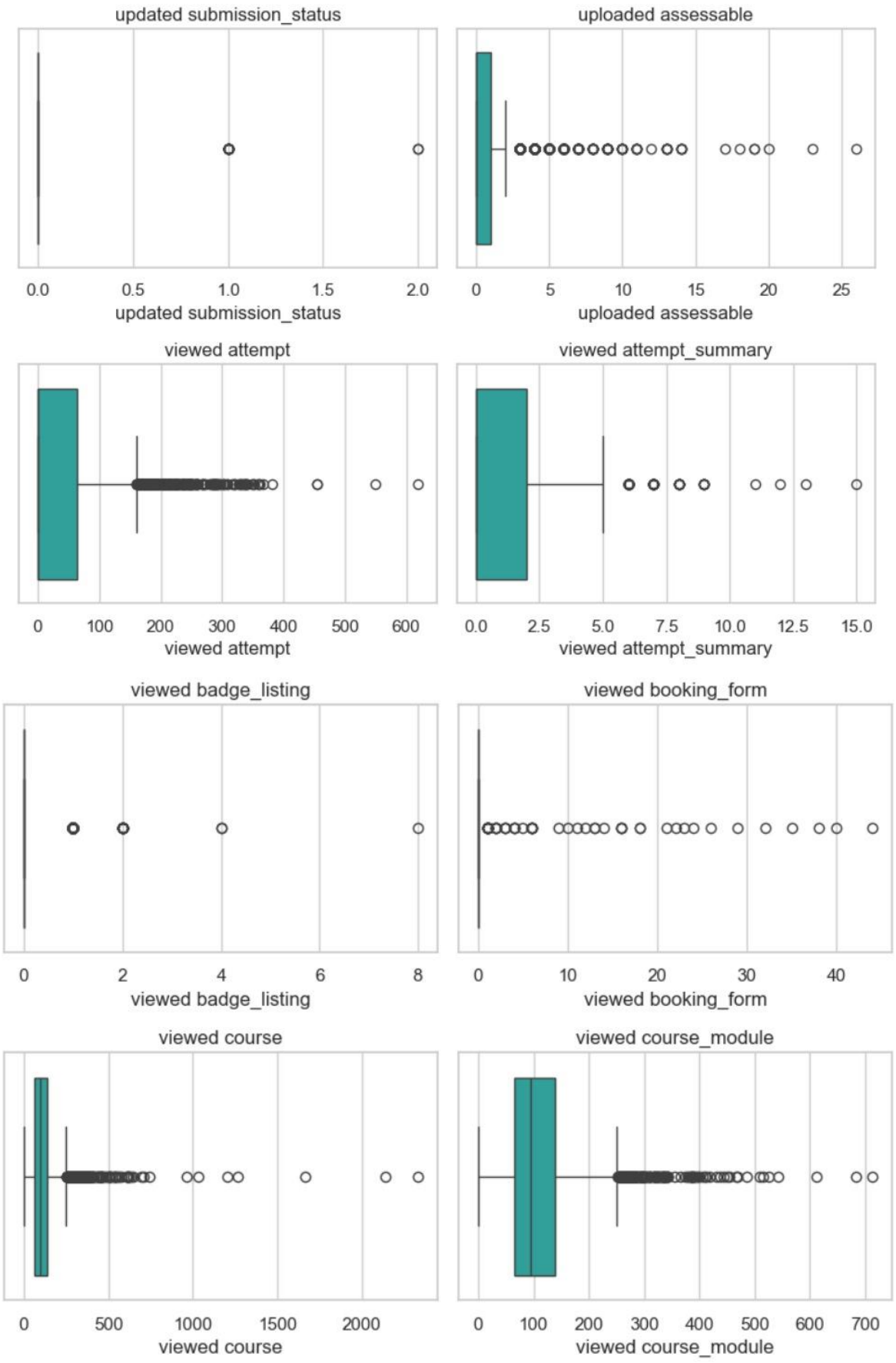
# APPENDIX A

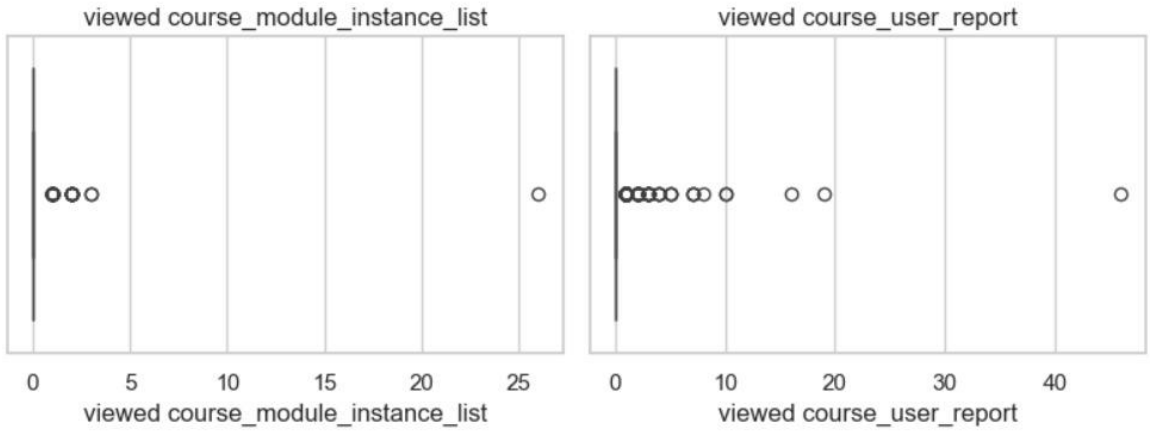
## A.1 OUTLIERS



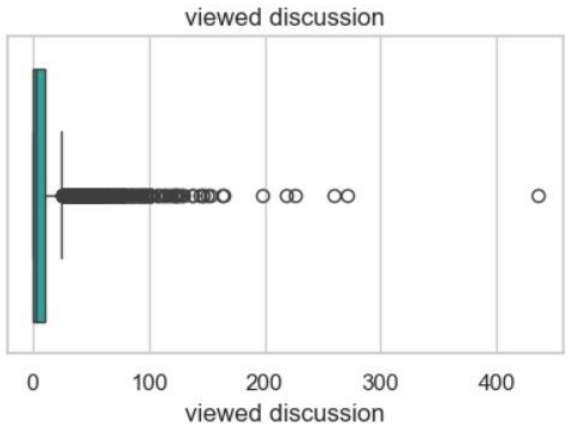

removed group_member
0.0 0.5 1.0 1.5 2.0 2.5 3.0
removed group_member
reviewed attempt
0 20 40 60 80 100 120
reviewed attempt
searched course
0.0 2.5 5.0 7.5 10.0 12.5 15.0
searched course
shown step
0 1 2 3 4
shown step
started attempt
0.0 2.5 5.0 7.5 10.0 12.5 15.0
started attempt
started tour
0 1 2 3 4
started tour
submission add
0 10 20 30 40
submission add
submission delete
0 1 2 3 4 5 6 7
submission delete

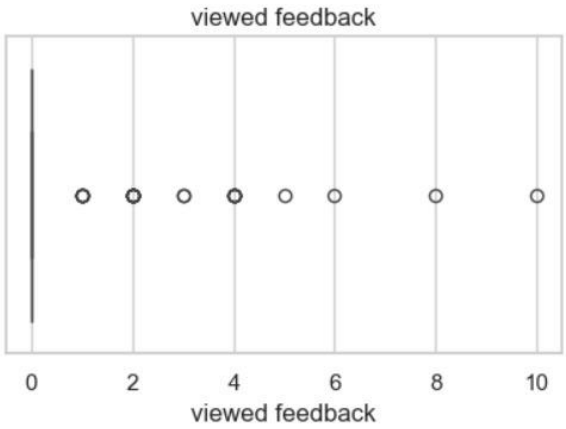

submissions list
0
20
40
60
80
100
submissions list
submitted assessable
0
5
10
15
20
submitted assessable
submitted attempt
0.0
2.5
5.0
7.5
10.0
12.5
15.0
submitted attempt
submitted response
0.0
0.2
0.4
0.6
0.8
1.0
submitted response
updated choice
0
1
2
3
4
5
updated choice
updated grade_item
0.0
0.2
0.4
0.6
0.8
1.0
updated grade_item
updated post
0
1
2
3
4
updated post
updated submission
0
5
10
15
updated submission

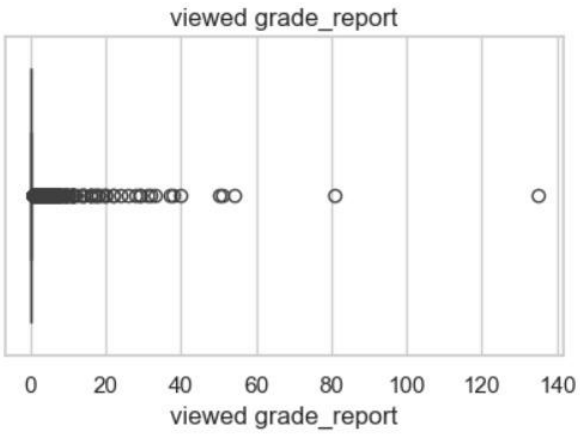

updated submission_status
0.0
0.5
1.0
1.5
2.0
updated submission_status
uploaded assessable
0
5
10
15
20
25
uploaded assessable
viewed attempt
0
100
200
300
400
500
600
viewed attempt
viewed attempt_summary
0.0
2.5
5.0
7.5
10.0
12.5
15.0
viewed attempt_summary
viewed badge_listing
0
2
4
6
8
viewed badge_listing
viewed booking_form
0
10
20
30
40
viewed booking_form
viewed course
0
500
1000
1500
2000
viewed course
viewed course_module
0
100
200
300
400
500
600
700
viewed course_module

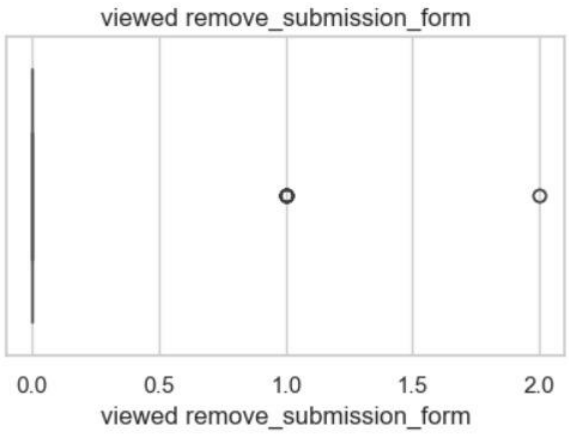

viewed course_module_instance_list
0
5
10
15
20
25
viewed course_module_instance_list

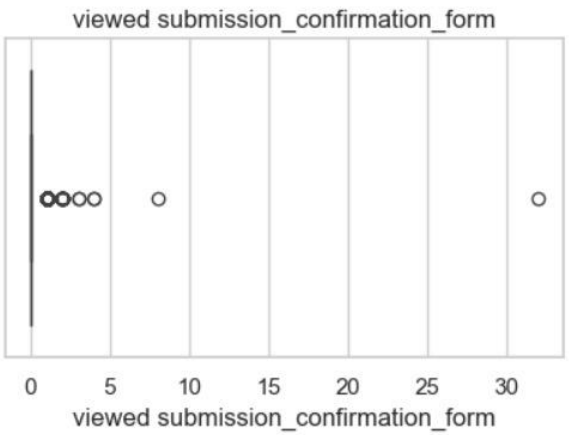

viewed course_user_report
0
10
20
30
40
viewed course_user_report

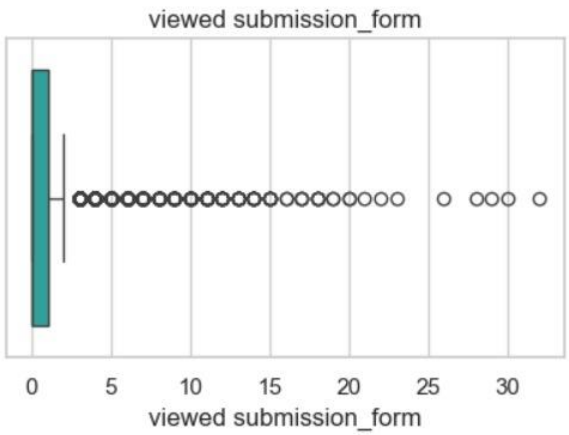

viewed discussion
0
100
200
300
400
viewed discussion

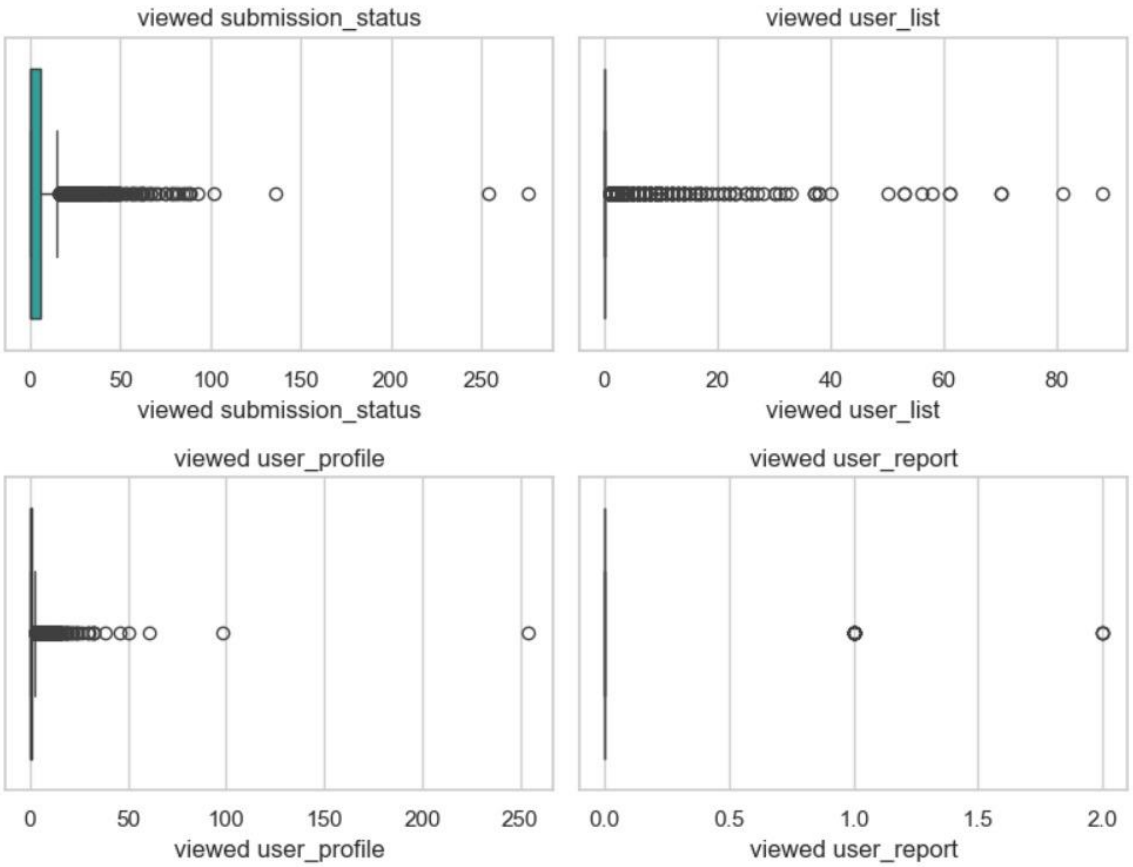

viewed feedback
0
2
4
6
8
10
viewed feedback

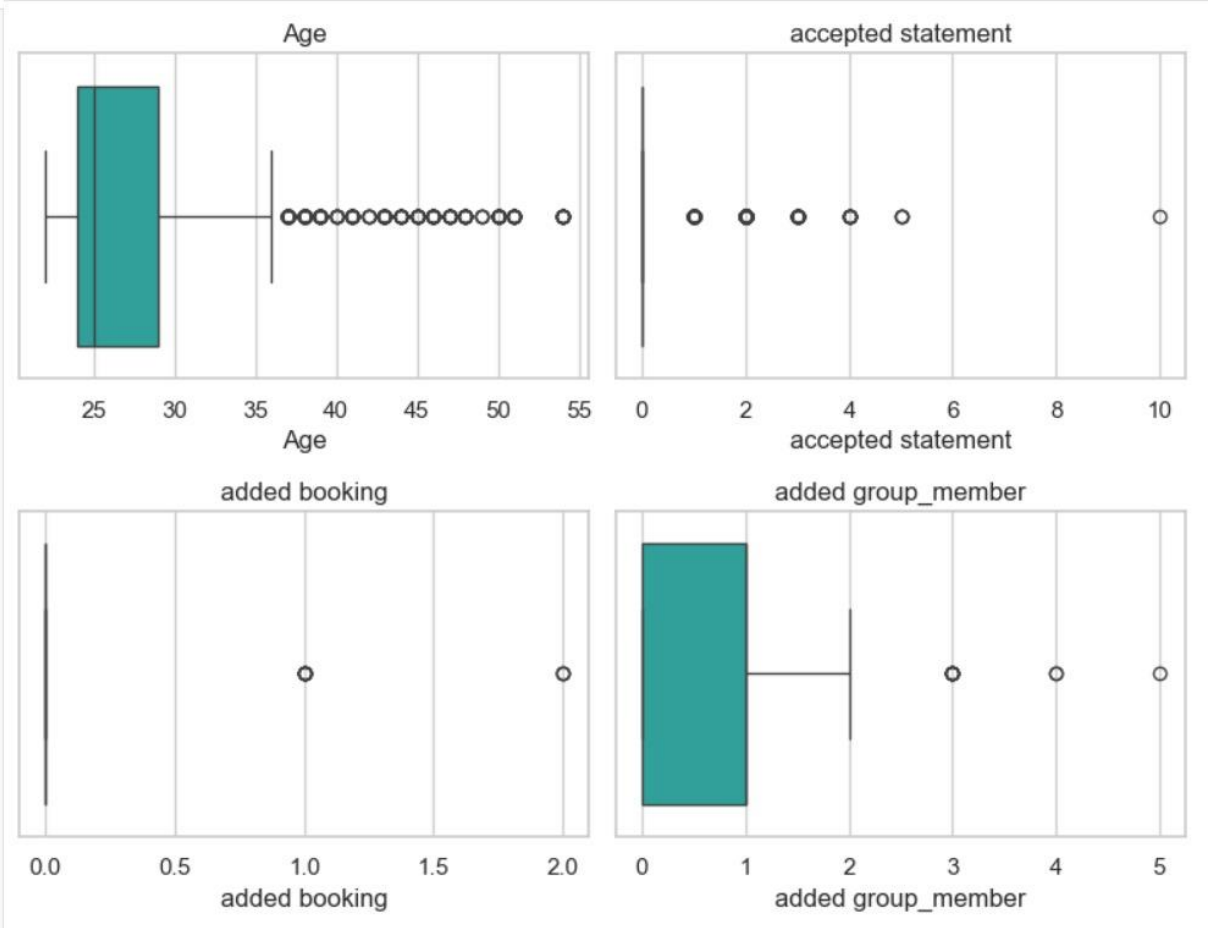

viewed grade_report
0
20
40
60
80
100
120
140
viewed grade_report

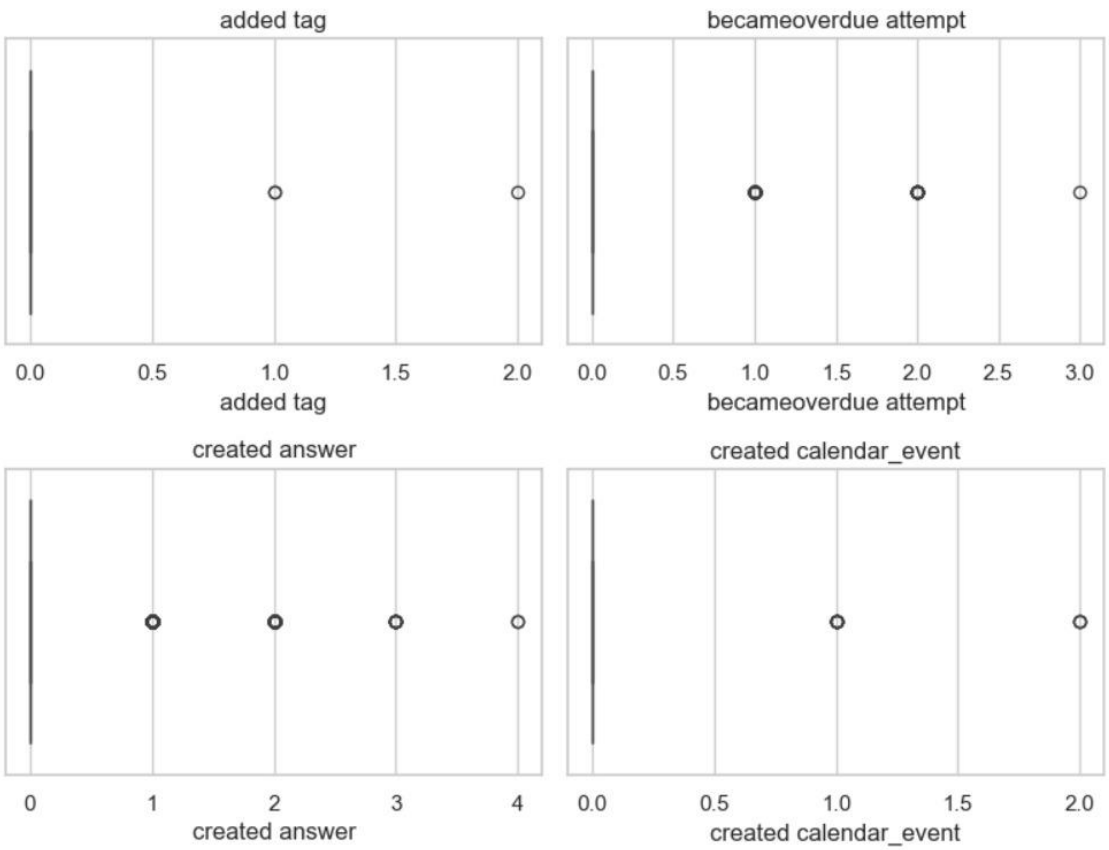

viewed remove_submission_form
0.0
0.5
1.0
1.5
2.0
viewed remove_submission_form

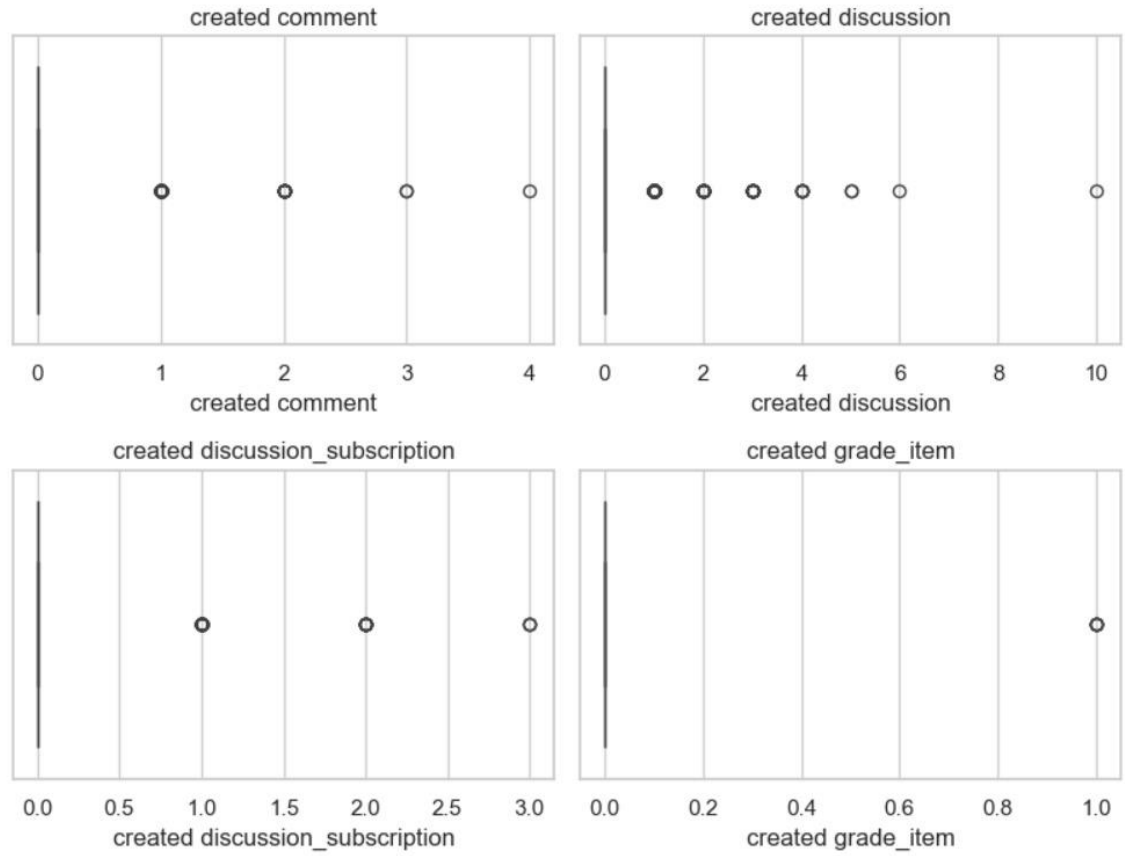

viewed submission_confirmation_form
0
5
10
15
20
25
30
viewed submission_confirmation_form

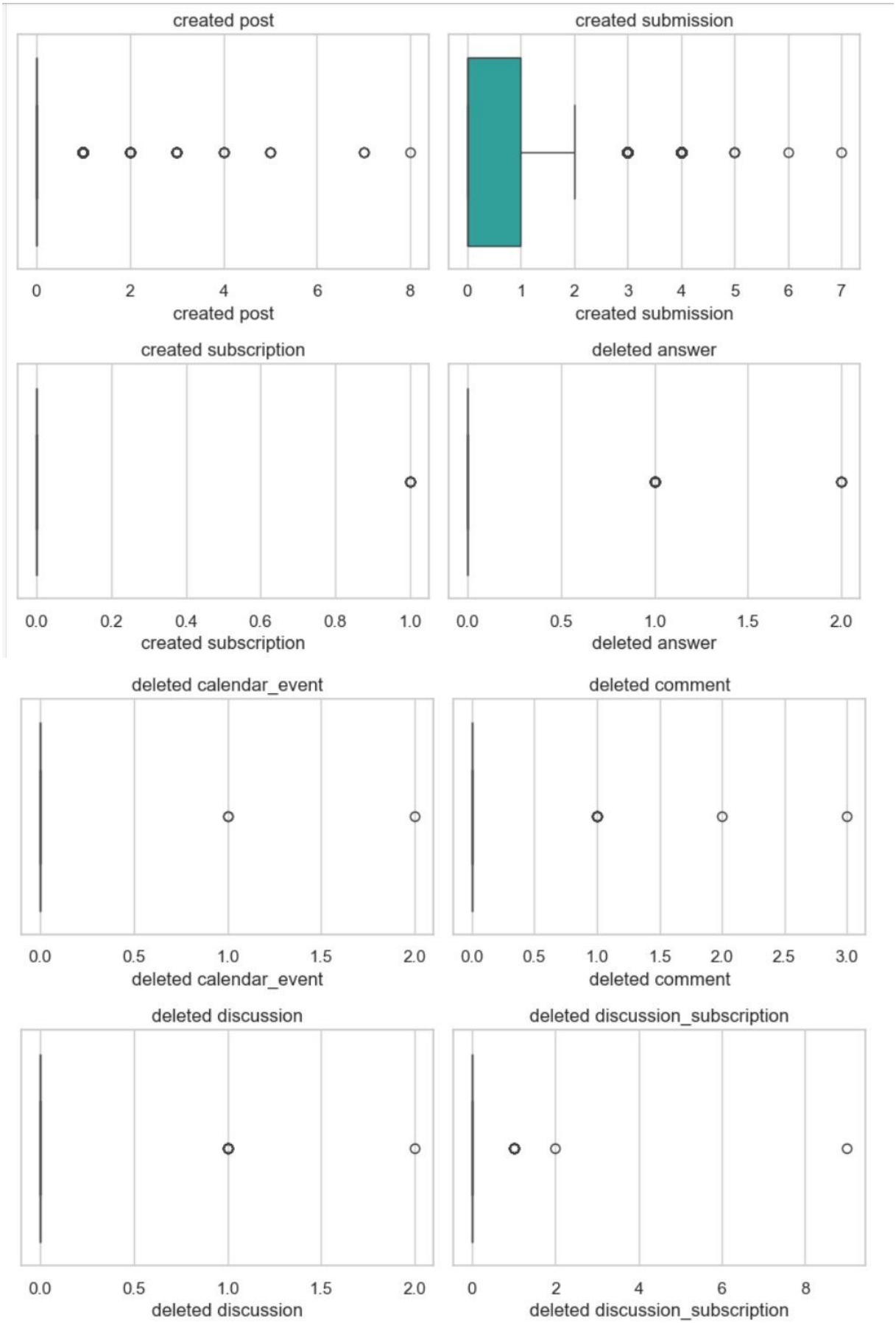

viewed submission_form
0
5
10
15
20
25
30
viewed submission_form


viewed submission_status
0
50
100
150
200
250
viewed submission_status
viewed user_list
0
20
40
60
80
viewed user_list
viewed user_profile
0
50
100
150
200
250
viewed user_profile
viewed user_report
0.0
0.5
1.0
1.5
2.0
viewed user_report


Age
25
30
35
40
45
50
55
Age
accepted statement
0
2
4
6
8
10
accepted statement
added booking
0.0
0.5
1.0
1.5
2.0
added booking
added group_member
0
1
2
3
4
5
added group_member


added tag
0.0
0.5
1.0
1.5
2.0
added tag
becameoverdue attempt
0.0
0.5
1.0
1.5
2.0
2.5
3.0
becameoverdue attempt
created answer
0
1
2
3
4
created answer
created calendar_event
0.0
0.5
1.0
1.5
2.0
created calendar_event


created comment
0
1
2
3
4
created comment
created discussion
0
2
4
6
8
10
created discussion
created discussion_subscription
0.0
0.5
1.0
1.5
2.0
2.5
3.0
created discussion_subscription
created grade_item
0.0
0.2
0.4
0.6
0.8
1.0
created grade_item


created post
0
2
4
6
8
created post
created submission
0
1
2
3
4
5
6
7
created submission
created subscription
0.0
0.2
0.4
0.6
0.8
1.0
created subscription
deleted answer
0.0
0.5
1.0
1.5
2.0
deleted answer
deleted calendar_event
0.0
0.5
1.0
1.5
2.0
deleted calendar_event
deleted comment
0.0
0.5
1.0
1.5
2.0
2.5
3.0
deleted comment
deleted discussion
0.0
0.5
1.0
1.5
2.0
deleted discussion
deleted discussion_subscription
0
2
4
6
8
deleted discussion_subscription

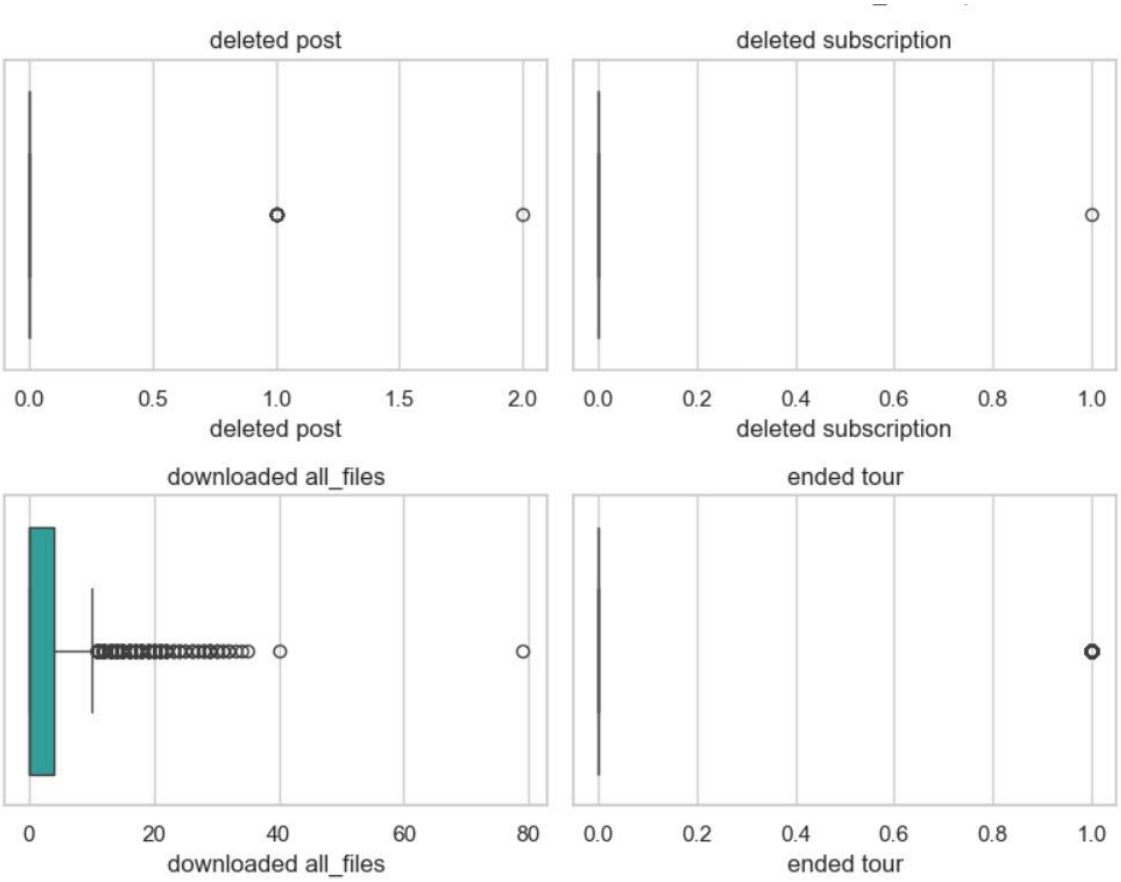



Figure 20 - Outliers of all numeric variables

## A.2 NON-VARIABILITY COLUMNS – ENTIRE COURSE

Table 13 - Columns with high variability in the entire dataset

| **Non-Variability columns – Entire Course** | |
|---|---|
| Assigned role | Created course_module |
| Created couse_section | Created user_enrolment |
| Deleted user_enrolment | Started attempet_preview |
| Unassigned role | Updated course |
| Updated course_module | Updated course_section |
| Updated folder | View report |
| Ds_nationality Bielorrússia | Ds_nationality Sweden |

## A.3 NON-VARIABILITY COLUMNS – INITIAL PART OF THE COURSE

Table 14 - Columns with high variability in the initial part dataset

| Non-Variability columns – Initial part | |
|---|---|
| Assigned role | Created course_module |
| Created couse_section | Created user_enrolment |
| Deleted user_enrolment | Unassigned role |
| View report | Updated course |
| Updated course_module | Updated course_section |
| Updated folder | ds_nacionalidade_Coreia (República da) |
| ds_ nationality_Ireland | ds_ nationality_Lithuania |
| ds_ nationality_Norway | ds_ nationality_Taiwan (Província da China) |

## A.4 HIGH CORRELATED COLUMNS – ENTIRE COURSE

Table 15 - High correlated columns in the entire dataset

| High correlated columns – Entire Course | | |
|---|---|---|
| **Column 1** | **Column 2** | **Correlation** |
| Accepted statement | viewed submission_confirmation_form | 0.90 |
| Added booking | Created calendar_event | 1.0 |
| Added group_member | Updated choice | 0.99 |
| Created submission | Submitted assessable | 0.99 |
| Created submission | Uploaded assessable | 0.95 |
| Created submission | viewed submission_form | 0.97 |
| Created submission | Viewed submission_status | 0.85 |
| Deleted calendar_event | Removed booking | 1.0 |
| Deleted discussion | Deleted post | 0.94 |
| Ended tour | Shown step | 0.97 |
| Ended tour | Started tour | 0.97 |
| Graded user | Reviewed attempt | 0.75 |
| Graded user | Started attempt | 0.86 |
| Graded user | Submitted attempt | 0.86 |
| Graded user | Viewed attempt | 0.87 |
| Graded user | Viewed attempt_summary | 0.82 |
| Removed choice | Removed group_member | 1.0 |

| | | |
|---|---|---|
| Reviewed attempt | Started attemmpt | 0.74 |
| Reviewed attempt | Submitted attempt | 0.74 |
| Reviewed attempt | Viewed attempt | 0.71 |
| Reviewed attempt | Viewed attempt_summary | 0.73 |
| Shown step | Started tour | 0.99 |
| Started attempt | Submitted attempt | 0.98 |
| Started attempt | Viewed attempt | 0.94 |
| Started attempt | Viewed attempt_summary | 0.93 |
| Submission add | Submissions list | 0.70 |
| Submitted assessable | Uploaded assessable | 0.96 |
| Submitted assessable | viewed submission_form | 0.98 |
| Submitted assessable | Viewed submission_status | 0.85 |
| Submitted attempt | Viewed attempt | 0.93 |
| Submitted attempt | Viewed attempt_summary | 0.93 |
| Submitted response | Viewed feedback | 0.82 |
| Updated submission_status | Viewed remove_submission_form | 0.81 |
| Uploaded assessable | Viewed submission_form | 0.95 |
| Uploaded assessable | Viewed submission_status | 0.83 |
| Viewed attempt | Viewed attempt_summary | 0.88 |
| Viewed course | Viewed course_module | 0.82 |
| Viewed submission_form | Viewed submission_status | 0.86 |

## A.5 HIGH CORRELATED COLUMNS – INITIAL PART OF THE COURSE

Table 16 - High correlated columns in the initial part dataset

| **High correlated columns – Initial Part** | | |
|---|---|---|
| **Column 1** | **Column 2** | **Correlation** |
| Accepted statement | viewed submission_confirmation_form | 0.96 |
| added group_member | updated choice | 0.99 |
| added tag | deleted discussion | 0.70 |
| added tag | deleted post | 0.70 |
| created submission | submitted assessable | 0.96 |
| created submission | uploaded assessable | 0.79 |
| created submission | viewed submission_form | 0.95 |
| deleted discussion | deleted post | 1.00 |
| ended tour | shown step | 0.98 |
| ended tour | started tour | 0.98 |
| graded user | reviewed attempt | 0.74 |

| | | |
|---|---|---|
| graded user | started attempt | 0.86 |
| graded user | submitted attempt | 0.91 |
| graded user | viewed attempt | 0.87 |
| graded user | viewed attempt_summary | 0.87 |
| removed choice | removed group_member | 1.00 |
| shown step | started tour | 0.99 |
| started attempt | submitted attempt | 0.95 |
| started attempt | viewed attempt | 0.99 |
| started attempt | viewed attempt_summary | 0.90 |
| submitted assessable | uploaded assessable | 0.77 |
| submitted assessable | viewed submission_form | 0.92 |
| submitted assessable | viewed attempt | 0.95 |
| submitted assessable | viewed attempt_summary | 0.93 |
| updated submission_status | viewed remove_submission_form | 1.00 |
| uploaded assessable | viewed submission_form | 0.78 |
| viewed attempt | viewed attempt_summary | 0.90 |
| viewed course | viewed course_module | 0.91 |

### A.6 HIGH CORRELATED COLUMNS DROPPED – ENTIRE COURSE

Table 17 - High correlated columns dropped in the entire dataset

| **High correlated columns dropped - Entire Course** | |
|---|---|
| viewed submission_confirmation_form | Added booking |
| Added group_member | Created submission |
| Submitted assessable | Removed booking |
| Deleted post | Shown step |
| Ended tour | Graded user |
| Reviewed attempt | Started attempt |
| View attempt_summary | Removed group_member |
| Reviewed attempt | Submitted attempt |
| Submissions list | Submitted assessable |
| Viewed submission_form | Submitted response |
| Viewed remove_submission_form | Uploaded assessable |
| Viewed course_module | |

### A.7 HIGH CORRELATED COLUMNS DROPPED – INITIAL PART OF THE COURSE

Table 18 - High correlated columns in the initial part dataset

| **High correlated columns dropped – Initial Part** | |
|---|---|
| Added tag | Created submission |

| | |
|---|---|
| Deleted discussion | Ended tour |
| Graded user | Started attempt |
| Uploaded assessable | Viewed attempt |
| Viewed submission_form | Started tour |
| View attempt_summary | |

## A.8 FEATURE SELECTION – ENTIRE COURSE

Table 19 - Selected features in the entire course dataset

| **Feature Selection – Entire Course** | |
|---|---|
| Age | Gender_M |
| Ds_nationality_Bangladesh | Ds_nationality_Brazil |
| Ds_nationality_India | Ds_nationality_Nigeria |
| Ds_nationality_Portugal | Created answer |
| Removed choice | Viewed attempt |
| Viewed course | Viewed user_list |
| Downloaded all_files | Submission add |
| Viewed discussion | Viewed submission_status |
| Accepted statement | Created discussion |
| Prevented access | Updated submission |
| Viewed grade report | |

## A.9 FEATURE SELECTION – INITIAL PART OF THE COURSE

Table 20 - Selected features in the initial part of the course dataset

| **Feature Selection – Initial Part** | |
|---|---|
| Age | Ds_nationality_Bangladesh |
| Ds_nationality_India | Ds_nationality_Portugal |
| Ds_nationality_Nigeria | Ds_nationality_Kosovar |
| Reviewed attempt | Submission add |
| Updated submission_status | Viewed badge_listing |
| Viewed submission_status | Added group_member |
| Downloaded all_files | Update choice |
| Viewed discussion | Viewed user_list |


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

Universidade Nova de Lisboa