PREDICTING MANIPULATION ATTEMPTS BY STUDENTS ON LEARNING MANAGEMENT SYSTEMS: AN APPROACH USING MACHINE LEARNING MODEL

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

MEHMET MELİH GÖRMEZOĞLU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

DECEMBER 2023

Approval of the thesis:

PREDICTING MANIPULATION ATTEMPTS BY STUDENTS ON LEARNING MANAGEMENT SYSTEMS: AN APPROACH USING MACHINE LEARNING MODEL

Submitted by MEHMET MELİH GÖRMEZOĞLU in partial fulfillment of the requirements for the degree of **Master of Science in Information Systems Department, Middle East Technical University** by,

Prof. Dr. Banu Günel Kılıç
Dean, **Graduate School of Informatics**

Prof. Dr. Altan Koçyiğit
Head of Department, **Information Systems, METU**

Prof. Dr. Soner Yıldırım
Supervisor, **Computer Education and Instructional Technology, METU**

Prof. Dr. Sevgi Özkan Yıldırım
Co-Supervisor, **Information Systems, METU**

Examining Committee Members:

Prof. Dr. İhsan Tolga Medeni
Management Information Systems, Ankara Yıldırım Beyazıt University

Prof. Dr. Soner Yıldırım
Computer Education and Instructional Technology, METU

Asst. Prof. Dr. Özden Özcan Top
Information Systems, METU

Date: _*18.12.2023*

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Mehmet Melih Görmezoğlu

Signature : ____________________

ABSTRACT

PREDICTING MANIPULATION ATTEMPTS BY STUDENTS ON LEARNING MANAGEMENT SYSTEMS: AN APPROACH USING MACHINE LEARNING MODEL

Görmezoğlu, Mehmet Melih

MSc., Department of Information Systems

Supervisor: Prof. Dr. Soner Yıldırım

Co-Supervisor: Prof. Dr. Sevgi Özkan Yıldırım

December 2023, 62 pages

This study focuses on the identification of students' behavior, spanning from 1st grade to 8th grade, within a designated Learning Management System, specifically aiming to detect potential instances of attempting to "game the system." The analysis employs a two-step approach: firstly, utilizing K-means clustering to reveal patterns in students' behavior based on the log data from the Learning Management System, and subsequently applying the XGBoost classification method to predict whether a student is engaged in attempts to manipulate the system. The selection of relevant features is informed by domain knowledge, providing an insight of the key indicators. The study concludes by offering improvement suggestions for Learning Management Systems, aimed at enhancing predictive outcomes of “gaming the system” behaviors and fostering a more robust educational environment to mitigate such behaviors.

Keywords: Data Mining, Data Science, Clustering Method, Classification Method, Gaming the System

ÖZ

ÖĞRENME YÖNETİM SİSTEMLERİNDE ÖĞRENCİLERİN MANİPÜLASYON GİRİŞİMLERİNİN TAHMİN EDİCİLERİ: BİR MAKİNE ÖĞRENİMİ MODELİ YAKLAŞIMI

Görmezoğlu, Mehmet Melih
Yüksek Lisans, Bilişim Sistemleri Bölümü
Tez Yöneticisi: Prof. Dr. Soner Yıldırım
Tez Eş Danışmanı: Prof. Dr. Sevgi Özkan Yıldırım

Aralık 2023, 62 sayfa

Bu çalışma, belirlenmiş bir Öğrenme Yönetim Sistemi içinde 1. sınıftan 8. sınıfa kadar olan öğrencilerin davranışlarının tanımlanmasına odaklanmaktadır. Özellikle "sistemi manipüle etme" girişimlerini tespit etmeyi amaçlayarak gerçekleştirilen bu analiz, iki aşamalı bir yaklaşımı benimsemektedir. Bu yaklaşım ilk olarak, Öğrenme Yönetim Sistemi'nden gelen öğrenme aktiviteleri verilerine dayanarak öğrenci davranışlarındaki desenleri ortaya çıkarmak için K-Means kümeleme yöntemini kullanma ve ardından öğrencilerin sistemi manipüle etme girişimlerinde bulunup bulunmadığını tahmin etmek için XGBoost sınıflandırma yöntemini uygulama şeklindedir. Tahmine ilişkin değişkenler, alan bilgisi temel alınarak seçilmiş olup, temel göstergelere ilişkin bir anlayışını sunar. Çalışma, Öğrenme Yönetim Sistemleri için iyileştirme önerileri sunarak sona erer ve bu öneriler, sistemin manipüle edilmesine yönelik aktivitelere ilişkin tahmin sonuçlarını artırmayı ve benzer davranışları hafifletmeyi amaçlayan daha sağlam bir eğitim ortamı oluşturmayı hedefler.

Anahtar Kelimeler: Veri Madenciliği, Veri Bilimi, Kümeleme Metodu, Sınıflandırma Metodu, Sistemi Manipüle Etme

To my beloved family…

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Prof. Dr. Soner Yıldırım for his invaluable guidance, encouragement, and support throughout this research. Without his guidance, I would not have known about the research opportunities in this domain. I am grateful for our enlightening, fruitful, and heartfelt discussions in developing this study.

I would like to thank my co-advisor, Prof. Dr. Sevgi Özkan Yıldırım, for sharing her support and encouragement on this study.

I am thankful to Assoc. Prof. Dr. Erkan Er for their help, comments, and valuable feedback from the very beginning of this study. Without his mentorship and encouragement, many of this study's accomplishments would not have been possible.

I also would like to thank Asst. Prof. Dr. Özden Özcan Top and Prof. Dr. İhsan Tolga Medeni for their valuable participation as jury members of my thesis.

I would like to express my gratitude to Ahmet Bodur and Ayşe Nur Özdere Yüksel for their support, encouragement, and positive energy that helps me stand up when I feel down.

I am grateful for each of my family members for their support, encouragement, and understanding in any condition.

Finally, I extend a very special thanks to my love, my dear wife, and partner in crime, Damla Görmezoğlu for supporting me with infinite love in every moment of my life especially during the tough times of this study.

TABLE OF CONTENTS

CHAPTERS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AUCPR	Area Under Curve Precision-Recall
CART	Classification and Regression Trees
DM	Data Mining
EDM	Educational Data Mining
GBDT	Gradient-Boosted Decision Tree
GTS	Gaming The System
ICT	Information and Communication Technology
ITS	Intelligent Tutoring System
k-NN	k-Nearest Neighbors
LA	Learning Analytics
LDA	Linear Discriminant Analysis
LMS	Learning Management System
LRM	Latent Response Model
LV-GD	Latent Variable-Based Gaming Detection
ML	Machine Learning
OBT	One Big Table
RF	Random Forest
SVM	Support Vector Machine
XGBoost	Extreme Gradient Boosting

CHAPTER 1

INTRODUCTION

In the contemporary landscape, data has emerged as an invaluable asset, often heralded as the "new oil" of the digital era. We find ourselves in an epoch where nearly every action, transaction, and interaction contribute to an ever-expanding digital footprint, thus generating a wealth of data. This ubiquity of data has made a paradigm shift in the way we comprehend and navigate the world around us. Data has evolved into a cornerstone that underpins informed decision-making and strategy formulation, impacting a myriad of domains.

The utility of data extends across a multitude of sectors, fundamentally shaping our lives and propelling progress in unprecedented ways. In the field of healthcare, data-driven insights are revolutionizing patient care, diagnosis, and treatment options. In the financial sector, data analytics serves as a guidance for investment decisions and risk assessment. Governments, in their pursuit of addressing societal challenges, are leveraging data for policy formulation. Educational institutions are harnessing the power of data to personalize learning experiences and optimize student outcomes. In this data-driven era, the ability to harness, analyze, and derive actionable insights from data stands as the bedrock of success and innovation. It is through the responsible and ethical management of data that solutions to complex problems are unveiled, processes are optimized, and societal progress is achieved.

In the educational domain, data has evolved into an indispensable tool for informed decision-making, fundamentally altering the way of education and structure learning experiences. Its significance lies in its capacity to equip educators, administrators, and policymakers with invaluable insights into the dynamics of the learning process. By methodically collecting and analyzing data on student performance, engagement, and behavior, educational institutions can tailor their teaching methods, curriculum design, and support services to ensure that each learner is equipped to excel. Data-driven education facilitates early intervention strategies, aiding in the identification of struggling students and providing them with targeted assistance to bridge learning gaps.

The role of data in education extends beyond the classroom. It facilitates institutional planning and resource allocation, enabling schools and universities to optimize their budgets and infrastructure. By scrutinizing data on student enrollment, demographic trends, and learning outcomes, educational institutions can make evidence-based decisions on staffing, facility management, and the development of new academic programs. Data also allows for the measurement of the effectiveness of educational policies and the assessment of long-term educational goals, fostering accountability and transparency in the education sector. Thus, data has become an essential tool for enhancing the quality of education, fostering student success, and ensuring that educational systems remain adaptable and responsive to the evolving needs of learners in the 21st century.

The fusion of education and technology has ushered in transformative approaches to learning and teaching, with Learning Management Systems at the forefront of this revolution. Learning Management System (LMS) platforms have assumed an indispensable role within educational institutions and organizations, providing educators and learners with powerful tools for accessing, delivering, and managing online education. However, as the proliferation of these platforms has advanced, so has the challenge of safeguarding academic integrity, which has been further compounded by the phenomenon known as "gaming the system."

"Gaming the system" within the context of LMS usage refers to the deliberate manipulation of the platform, its regulations, or its assessment mechanisms by students, with the aim of gaining an unfair advantage in terms of grades, progress, or coursework completion. This infringement encompasses a wide spectrum of illicit activities, including plagiarism, cheating on quizzes and exams, and the exploitation of technical vulnerabilities within the LMS. The proliferation of such behaviors imperils the core principles of equitable education and presents a substantial obstacle for educators and institutions committed to preserving the integrity of online learning environments.

Amid this sophisticated landscape, the role of data, Information and Communication Technology (ICT), machine learning, and the informed utilization of data in decision-making has assumed paramount importance. The detection and prevention of gaming the system within LMS platforms necessitates an interdisciplinary and technology-driven approach. Educational institutions are endowed with extensive repositories of data, capturing a wealth of information regarding student behaviors, patterns of interaction, and academic performance. Leveraging this data through ICT infrastructure and advanced machine learning (ML) techniques is essential, not only for early detection but also for proactive decision-making.

The repercussions of gaming the LMS are many-sided, jeopardizing academic integrity, compromising the quality of education, and undermining the credibility of awarded qualifications. As students continually devise new strategies to game the system, educators find themselves in a perpetual battle to uphold fairness and reliability. Thus, there is an imperative need for robust and intelligent tools to detect, prevent, and deter such activities, along with the use of data-driven insights for more effective real-time decision-making.

This study serves as a comprehensive exploration of the pressing issue of gaming the system within LMS platforms, with a central focus on the vital role of data, ICT, and machine learning in addressing this challenge. It delves into the motivations driving students to manipulate the system and proposes innovative, data-driven approaches to safeguard academic integrity in online education. By shedding light on this intricate issue, this research aims to equip educators and administrators with the knowledge, tools, and insights necessary to create secure, equitable, and data-informed learning environments.

Subsequent chapters will delve into the motivations, methods, and data sources relevant to gaming the LMS, provide an insight in detection strategies, and present data-driven solutions to overcome this challenge. Ultimately, this work seeks not only to contribute to the ongoing discourse on academic integrity but also to empower stakeholders with the capacity to harness data and technology for more effective decision-making in the field of LMS.

1.1. Purpose of the Study

The E-learning environment has garnered increasing significance, with schools, educators, and students aspiring to access educational content conveniently through such platforms. Learning Management Systems (LMS), which are crucial in the field of E-learning, and other learning tracking systems, have evolved beyond their role as mere content presentation platforms. They have transitioned into sophisticated tools capable of amassing substantial volumes of data concerning student behavior and performance.

The existence of platforms capable of collecting such extensive data underscores the importance of enabling educators to proficiently analyze this wealth of information. In this context, the field of learning analytics assumes predominant significance, as it contributes significantly to the coherent interpretation of this data, thereby providing valuable insights for educational stakeholders.

This study seeks to address a concern within the domain of Learning Management Systems: the identification and prediction of attempts to "game the system" by users of a specified LMS. Furthermore, the research aims to explore potential enhancements to the data and methodologies employed by the LMS to fortify its effectiveness in the context of E-learning.

1.2. Research Questions

- To what extent do the independent variables of exam duration, class enrollment, subject of the exam, the count of active study materials, and total study time to identify and predict instances of users attempting to manipulate the system within an E-learning environment?

- What recommendations can be proposed to enhance the analytical capabilities of Learning Management Systems (LMS) for improved outcomes effects on prediction accuracy and reduce the gaming the system?

1.3. Significance of the Study

The significance of the E-learning environment has surged, capturing the attention of schools, educators, and students who seek the ease of accessing educational content through these digital platforms. Within this evolving landscape, Learning Management Systems, which serve as pivotal components of E-learning, have transcended their conventional roles as mere content delivery systems. They have transformed into sophisticated instruments adept at accumulating substantial datasets pertaining to student behavior and performance.

The presence of platforms capable of amassing such extensive data underscores the pressing need to empower educators with the ability to proficiently dissect and derive insights from this trove of information. It is in this context that the field of learning analytics emerges as critically significant, facilitating the coherent interpretation of these datasets and, in turn, delivering invaluable perspectives to educational stakeholders.

This study undertakes the crucial mission of addressing a paramount concern within the field of Learning Management Systems: the detection and anticipation of efforts to "game the system" by users of a specified LMS. Furthermore, the research endeavors to explore potential enhancements to the data and methodologies employed by the LMS, thereby reinforcing its effectiveness within the broader context of E-learning.

1.4. Limitations of the Study

- It is assumed that the LMS system used in the study was used by students as desired.
- It is assummed that each student only have one account in the system.
- The data is limited to as provided from LMS provider.
- The research is limited to 1999 unique students using a certain LMS system in Turkiye.

CHAPTER 2

LITERATURE REVIEW

The reviewed literature in this chapter investigates the insightful approach of Big Data in education, Learning Analytics (LA), Learning Management Systems, and Educational Data Mining (EDM) techniques, all with a common goal of comprehending and enhancing student behavior and educational outcomes. These studies collectively highlight the focal role played by data-driven methodologies in the domain of education, focusing on diverse aspects such as student engagement, off-task behavior, gaming the system, and predicting academic performance. They emphasize the need to develop more precise LMS platforms, recommend systems for student evaluation, and employ machine learning techniques for personalized learning experiences. Through the thoughtful application of data analytics, these authors aim to empower educators and institutions to make informed decisions, cultivate effective teaching methods, mitigate detrimental student behaviors, and ultimately elevate the overall quality of education in various educational settings.

These studies rely on some fundamental concepts while emphasizing the importance of data and its effects on education domain. The following concepts are explained in further detail.

2.1. Big Data

Along with technological advancements and integration of information and communication technologies (ICT) in our daily life, there has been a significant increase in the utilization of technological instruments in various sectors. This utilization provides increasing data collection and where the Big Data notion comes to life. Big data refers to large and complex data sets that require special processing and analytics approach to extract insights. It comprises structured, unstructured, and semi-structured data, and its analysis can lead to better decision-making and efficient policy making capability for organizations. The term is used across various disciplines and is associated with challenges such as data storage, analysis, visualization, and privacy concerns (Sagiroglu & Sinanc, 2013). Big data's significance lies in its potential to reveal hidden patterns and correlations, offering advantages to businesses and organizations in gaining a competitive edge (Emetere, 2019; Crawford, 2013). Big Data often defines with 4 V, the “3 Vs” that characterize it (i.e., volume, velocity, and variety) or the “4 Vs” (adding veracity to the previous three) are responsible for the fact that it exceeds an organization’s own data as well as its storage or compute capacity for accurate and timely decision-making (Vossen, 2014).

In the literature, aim of using big data to enhance student academic performance through personalized learning experiences, improve grading practices, increase student engagement and motivation, reduce dropout rates, and optimize e-learning in higher education. It can also help in analyzing the causes of problems in the e-learning system and provide real-time data about academic performances of students.

Big data analysis in education has the potential to enhance student academic performance by providing personalized learning experiences. Additionally, big data can be utilized to reduce dropout rates by providing early warning indicators to identify students who are at risk of dropping out (Sui & Sui, 2023). Furthermore, big data can help higher education optimize e-learning by determining and analyzing the causes of problems in the system. The digital platforms contributed to the collection of more data about students' academic performance, their learning styles, preferences, and tendencies, which can be utilized for different purposes such as improving digital learning platforms and individualization of learning processes (Duykuluoğlu et al, 2023).

2.2. Learning Analytics

Learning analytics is a field of educational technology and data analysis that focuses on the collection, interpretation, and application of data to enhance the learning and teaching processes within the education domain. It leverages the power of technology and data science to gather and analyze information about students' interactions with learning resources, such as digital tools, courses, and assessments.

Learning analytics is a developing field that focuses on analyzing data from learners to enhance educational practices. It involves the use of analytical tools to collect, process, and interpret data from virtual learning environments, aiming to optimize learning outcomes and teaching methods (Štrukelj, 2015). Educational data mining, academic analytics, teaching analytics, and assessment analytics are closely related concepts that contribute to the understanding and improvement of learning processes and educational decision-making (Yin-Kim, Yau et al., 2020). Learning analytics provides teachers with new insights into their students' learning processes, enabling them to make more effective use of their resources and influence the metrics agenda towards richer conceptions of learning (Clow, 2013). This multidisciplinary approach integrates studies of learning with technological capabilities, emphasizing the importance of diverse stakeholders and productive multivocality in the field of learning analytics (Suthers & Verbert, 2013).

2.3. Educational Data Mining

In line with the development of ICT, digital footprint is increasing for every user of these systems. Over the last two years alone 90 percent of the data in the world was generated (Marr, 2018). Being able to analyze this exponentially increasing data provides great advantage for decision-makers in their policy making activities. In accordance with this purpose, Data Mining (DM) concept has emerged.

DM is a process of discovering meaningful patterns, trends, and insights from large and complex datasets. It is a subset of the broader field of data analysis and plays a crucial role in various domains, including finance, education, healthcare, and more (Kaya Keleş, 2017). DM provides us with a series of new technologies to assist us in revealing previously hidden patterns, which have the potential to help us innovate and develop new theories, thus promoting revolutionary influences on new theory development in many disciplines (Shu & Ye, 2023).

As observed in various domains, the application of DM has extended to the field of education. DM encompasses a diverse array of techniques designed to facilitate the comprehension of hidden relationships within extensive datasets. In parallel, the discipline of Educational Data Mining (EDM) not only serves the purpose of uncovering these relationships but also provides invaluable insights in learning and teaching. Consequently, this concept, which departs from the conventional definition of Data Mining by specifically focusing on the analysis of educational data, has contributed to the emergence of a distinct and evolving research area over time.

EDM is an emerging discipline that focuses on exploring and analyzing the vast amount of data generated in educational settings. It aims to understand students' learning behaviors, improve educational processes, and make informed decisions (Harikumar, 2014). EDM applies data mining techniques such as prediction, classification, relationship mining, clustering, and social network analysis to educational data (Patham et al., 2014; Khare et al., 2018). It has been successfully used in various educational systems, including traditional educational systems, web-based educational systems, intelligent tutoring systems, and e-learning platforms (Khare et al., 2018). The main objectives of EDM include measuring student performance, assessing students, studying student behavior, predicting student outcomes, detecting undesirable behaviors, grouping students, and developing student models. Bayesian Network and Random Forest are effective techniques for predicting student performance, while Social Network Analysis is useful for detecting undesirable student behaviors. Clustering and Social Network Analysis are effective techniques for grouping students and student modeling (Romero & Ventura, 2010).

2.4. Learning Management Systems

Learning management systems (LMS) are online applications that organizations and educational institutions use to manage and offer online courses and programs. These systems provide resources like video lectures, assignments, games, quizzes, and progress tracking to support learning and training. LMS can be software systems that are customized, products that are commercial, or products that are free and open source.

LMS technologies revolutionize the education system by understanding the needs and expectations of each user group and developing tools and resources that correspond to their specific needs (Simanjuntak et al., 2022). The use of LMS during the COVID-19 pandemic has shown improvement in student learning outcomes, especially in the cognitive domain (Awad et al, 2019). LMSs offer various benefits and features that facilitate and enhance student learning, such as testing, training, bookkeeping, tracking, and plagiarism prevention (Ahmed & Mesanovic, 2019).

On the other hand, Intelligent tutoring systems (ITS) are computer learning environments that adapt to students at a fine-grained level and implement complex principles of learning. ITS have been developed for various subjects and have been shown to improve learning compared to traditional teaching methods. ITS are computer programs that use artificial intelligence techniques to interact with students and provide tailored experiences. Originally, these systems focused on modeling the student's developing knowledge, but now they also adapt to the student's activity patterns and estimates of their knowledge (Lesgold, 1992). ITS are tutor behavior systems that support student learning and retention based on their characteristics and needs (Amastini, 2014). An intelligent tutoring system can be implemented using filters, predictive modeling, and a knowledge warehouse. This system dynamically selects content for individualized presentation to learners (Bergeron, 2008).

The difference between LMS and ITS is depicted in the table below:

Table 1: Difference Between LMS and ITS

	LMS	ITS
Purpose	Focus on managing and delivering educational content and resources. Provide a platform for organizing courses, delivering materials, tracking student progress.	Provide personalized and adaptive instruction. Acts like a tutor, offering individualized content, feedback, and guidance to students on improving their learning outcomes.
Instruction	Provide a framework for instructors to create and deliver content, including courses, assignments, and assessments.	Deliver tailored instruction that adapts to the learner's specific needs, pacing, and style.
Adaptability	Generally static and course oriented. Focus on delivering predefined content.	Highly adaptive. Use AI and data-driven approaches to understand each learner's strengths and weaknesses.
Use Cases	Suitable for managing and delivering a broad range of courses and resources, making it widely used in schools,	Beneficial for personalized instruction and improving student competence

There is a substantial amount of research available regarding the utilization of LMS or ITS to anticipate students' academic performance and comprehend their behavioral patterns in relation to usage of these systems, which is derived from log data of these systems. In this chapter, an extensive examination is conducted on the existing literature surrounding the implementation of machine learning models for the purpose of predicting students' academic accomplishments and their behaviors while using these systems as well as for identifying instances of cheating attempts within online examination platforms and detecting instances of misuse within learning platforms.

Cantabella, et al. (2019) present a case study analyzing student behavior in an LMS at the Catholic University of Murcia over four academic years, considering learning modality, number of accesses to the LMS, tools used by students, and associated events. Statistical and association rule techniques were applied using a Big Data framework to manage the large volume of data generated by users in the LMS, and the obtained results were evaluated to detect trends and deficiencies in the use of the LMS by students.

Kondo, et al. (2017) highlight the importance of utilizing educational big data for this purpose and discusses the potential of using online log data to detect off-task behavior and its impact on learning outcomes. The authors also emphasize the significance of identifying the explanatory variables that have a comparatively greater impact on learning outcomes, which can be used in institutional research.

Han Hu, et al. (2014) emphasize that analyzing learning portfolios and using data mining techniques, educators can gain insights into students' learning performance and identify those who are at risk of course failure. It is also indicated that the development of a precise LMS that can assess student learning performances using web-based learning portfolios is a challenging task, but data mining can help in extracting valuable knowledge from large repositories of learner data.

Dutt and Ismail (2019) discuss the significance of LMS in educational institutions and their role in facilitating and tracking course content for learners. Authors introduce the concept of Educational Data Mining (EDM), which utilizes data from LMS and other educational settings to understand student academic performance. In this study, use of machine learning techniques, including Classification and Regression Trees (CART), Linear Discriminant Analysis (LDA), k-Nearest Neighbors (k-NN), Support Vector Machine (SVM), and Random Forest (RF) is mentioned. The authors use kappa statistics as performance metric and RF classifier perform better among the other techniques.

Duhaim, et al. (2022) propose a recommendation system for evaluating students' answers and detecting cheating during online exams using statistical methods, similarity measures, and clustering algorithms. In this study K-means, Hierarchical, and Expectation Maximization clustering algorithms are employed to develop a well-designed clustering strategy for detecting cheating. The authors are implementing detecting cheating mechanisms in the light of IP addresses identification, by time (time taken or time late) and identifying clusters with K-means clustering method.

Romero and Ventura (2013) present an overview of EDM, discussing its growth, objectives, techniques, tools, and applications in education. Authors highlight the iterative cycle of hypothesis formation, testing, and refinement in EDM, and the various tasks and applications it can be used for, such as predicting student performance and personalizing learning.

Huang, et al. (2023) present a new approach called latent variable-based gaming detection (LV-GD) that controls for contextual factors and provides more robust estimates of student-level gaming tendencies. LV-GD applies a statistical model on top of an existing action-level gaming detector developed based on a typical human labeling process, without additional labeling effort. It controls for contextual factors and provides more robust estimates of student-level gaming tendencies. This study is based on the use of human coders to label gaming on student attempts or actions, inter-rater reliability analysis, and the association between gaming estimates and learning as a measure of validity.

Baker, et al. (2004) focus on a specific behavior called "gaming the system," which refers to behavior aimed at performing well in an educational task by taking advantage of properties and regularities in the system, rather than thinking about the material. The study found that students who frequently engage in gaming the system have lower learning outcomes. The authors aim to develop a machine-learned Latent Response Model (LRM) that can identify if a student is gaming the system in a way that leads to poor learning. The LRM was trained on multiple sources of data, including log files of student actions, human-coded observations of student behavior, and student learning outcomes.

Beal and Cohen (2008) discuss the importance of tracking student learning and the need for accessible and meaningful information about student performance in educational settings. Intelligent tutoring systems (ITSs) are highlighted as a potential solution to assess and track students' progress in real-time, providing up-to-the-minute assessments of their performance. The paper emphasizes the use of data mining techniques to analyze students' behavior and interactions with ITSs, including methods for handling hidden state variables and testing hypotheses.

Fernando Raguro, et al. (2022) focus on the extraction of student engagement and behavioral patterns in online education using decision tree and K-means algorithm. In this study, educational data mining techniques are used to extract hidden knowledge from educational data, which helps improve teaching methods and learning processes. The researchers utilized machine learning algorithms, specifically the Decision Tree Algorithm and the K-means Algorithm, to extract predictive rules sets and profile students' behaviors in the LMS environment.

Vasic, et al. (2015) aim to predict students' knowledge levels using data gathered from student activity logs in LMS systems. The goal is to classify students based on their knowledge levels, which can then be used to implement teaching models and improve student capabilities. In this study, the system logs from Moodle LMS were used for prediction of learning outcomes based on Bloom's taxonomy which is commonly used for classifying learning outcomes. According to Bloom’s taxonomy there are 6 mayor outcomes - Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation. The classification process in this study is based on the Naive Bayes classifier, which consists of a training phase and a classification phase. The Naive Bayes classifier implementation from the Mahout library on Hadoop server was used for classification.

de Sande, et al. (2010) analyze data from online tests and written exams in a Signals and Systems course for undergraduate students that online tests are delivered through a learning management system called Moodle. The LMS generates quizzes by randomly selecting questions from an item bank, with 1 or 2 questions from each category to create a 10-item quiz. The final exam, which was the same for all students, was taken in an examination classroom under the supervision of the teachers. Correlations and analysis of variance were used to compare the marks obtained in the online tests and the final exam, and to determine if different groups of marks were statistically different.

Paquette and Baker (2019) discuss the use of knowledge engineering and machine learning methods in modeling student behaviors in digital learning environments. The study highlights the lack of direct comparison between these two approaches and aims to compare their relative advantages in the context of modeling "gaming the system" behavior. Knowledge engineering involves developing models based on experts' knowledge, while machine learning uses data-driven algorithms to discover relationships between student behaviors. The authors introduce a hybrid approach that combines elements of both knowledge engineering and machine learning.

Zhao, et al. (2023) examine academic cheating among Chinese second to sixth graders using a machine learning approach to demonstrates that machine learning can be effectively used to analyze developmental data. The study uses Random Forest machine learning model to predict cheating behavior with a mean accuracy of 81.43%. Categorical variables are dummy-coded, and two dummy variables are created for school type (with school A as the reference) and three dummy variables for information about siblings (with the only child as the reference). The study included questions about school, participants' age, gender, information about siblings, and achievement level to identify important predictors of cheating. In this study, the logistic regression algorithm is also used to analyze the data, but it did not consider higher order interactions like the other algorithms.

Paquette, et al. (2014) focus on understanding how experts code student disengagement behaviors, specifically gaming the system, in online learning environments. The study aims to use cognitive task analysis to elicit expert knowledge about how they code gaming behaviors in Cognitive Tutor Algebra, and to build a cognitive model based on this knowledge to gain insights into the behaviors that constitute gaming the system. In this study, cognitive task analysis was employed to elicit expert knowledge on how they code gaming behaviors in Cognitive Tutor Algebra. The knowledge elicitor, who acted as the interviewer, coded a few clips while thinking aloud, with the expert providing feedback and corrections to improve the understanding of the coding process. The sessions were recorded and used by the elicitor to develop an initial version of the cognitive model. This model was then executed on a subset of the data, which was divided into a training set and a test set. The training set contained randomly selected clips coded as gaming and non-gaming, while the test set remained unseen. The findings of this study can inform the design and development of interventions to prevent and mitigate gaming behaviors in online learning environments. By identifying specific behaviors associated with gaming, educators and system designers can implement targeted strategies to engage students and discourage exploitative behaviors.

Muldner, et al. (2011) provide insights into the gaming behaviors of students in ITS and their impact on learning outcomes. According to the authors, understanding how students' game the system can help in designing interventions to prevent or mitigate gaming behavior. The use of data mining techniques can provide valuable insights into the impact of instructional interventions, such as hints, on student learning. In this study, as data mining model, a simple knowledge-tracing model, a dynamic Bayesian network, was used to infer student learning from problem-solving actions.

These literature reviews collectively point out the significance of data-driven approaches, data mining techniques, and the utilization of learning management systems to gain insights into student behavior and improve educational outcomes. The authors emphasize the need to analyze student engagement, identify off-task behaviors, and detect gaming the system, a behavior that negatively impacts learning. They stress the importance of developing precise LMS platforms and implementing recommendation systems for student evaluation. Additionally, these studies highlight the role of machine learning techniques in predicting student performance and personalizing learning experiences. By using data to inform their research, the authors aim to provide educators and institutions with tools and strategies to enhance teaching methods, mitigate problematic behaviors, and ultimately foster improved learning outcomes for students across various educational settings.

CHAPTER 3

METHOD

In this study, an examination of LMS server log data is undertaken with the aim of predicting instances where students may have attempted to manipulate or exploit the system, commonly referred to as "gaming the system." Due to the absence of pre-existing labels in the dataset, a labeling process becomes imperative before employing classification algorithms, a category falling under supervised learning. Consequently, the K-means clustering algorithm is employed to disclose inherent patterns in student behaviors derived from log data. Afterwards, an XGBoost machine learning algorithm is applied, leveraging an ensemble of decision trees and gradient boosting techniques to facilitate predictive modeling. The data is gathered from a private LMS called “Morpa Kampüs” which aims to support the students from the 1st grade to 8th grade within Turkish Education System. This LMS, while serving as a supplementary tool for the curriculum, does not impact the actual grading system with the exams conducted within the LMS.

The given LMS is tailored to support both primary and middle school students and teachers in their lessons with curriculum-aligned contents. The online educational materials categorized by the grade levels afford students the ability to monitor their individual progress, enable teachers to oversee their students' advancements, and empower parents to closely observe the developmental trajectory of their children. Furthermore, school administrators can leverage the system to evaluate teachers' work and assess the academic performance of students within their respective institutions.

The primary objective of this study is to analyze LMS log data to predict instances of “gaming the system” behavior among students during exams. Subsequently, the research aims to determine the factors and variables in the predictive analysis, interpret their significance and impact on the accuracy of predictions.

The raw data has no labels about the students who are attempting to “gaming the system”. The conventional methodology for predicting student behavior typically involves the utilization of classification algorithms when labeled data is available. In our dataset, the absence of labels directs the implementation of a clustering algorithm to assign categorizations to the data points. K-means clustering algorithm is used to determine the patterns in student behaviors based on log data of the LMS. Then XGBoost classification algorithm is employed to predict suspicious behavior on gaming the system.

The study is conducted with Python (v.3.9) programming language with pandas, numpy, sklearn, xgboost, matplotlib, seaborn, openpxl libraries. The system has 16GB RAM, 4GB graphic memory, i7 quad CPU processor, and Windows 10 as operating system.

3.1. Data Collection

The log data for the LMS has been made available by the system administrator. The dataset spans from September 1, 2021, to September 30, 2021, and is presented in a partial form. The raw data is structured in Microsoft Excel format with a ".xlsx" extension, comprising seven distinct sheets, outlined as follows:

- Member list: In the raw data there are 1999 rows and 5 features for 1999 unique student id.

Table 2: Data Types of Member List Dataset

Categorical	Student unique ID City unique ID District unique ID School unique ID Student’s enrolled class

- Login logs: In the raw data there are 21569 rows and 5 features for 21569 unique login id for 1999 unique student id.

Table 3: Data Types of Login Logs Dataset

Categorical	Login unique ID Student unique ID
Continuous	Login duration (in minutes)
Timestamp	Login time Logout time

- Games logs: In the raw data there are 6535 rows and 7 features for 26 unique game id for 677 unique student id.

Table 4: Data Types of Games Logs Dataset

Categorical	Student unique ID Game unique ID Login unique ID
Continuous	Score, Duration (in seconds)
Timestamp	Start time End time

- Lectures logs: In the raw data there are 19061 rows and 11 features for 1334 unique study material id for 1554 unique student id.

Table 5: Data Types of Lectures Logs Dataset

Categorical	Student unique ID Study material unique ID Study material type Lecture type Subject Unique ID Login unique ID
Continuous	Participation rate Performance

Table 5 cont.

Timestamp	Start time End time

- Studies logs: In the raw data there are 10965 rows and 11 features for 994 unique study material id for 1311 unique student id.

Table 6: Data Types of Studies Logs Dataset

Categorical	Student unique ID Study material unique ID Study material type Lecture type Subject Unique ID Login unique ID
Continuous	Participation rate Performance
Timestamp	Start time End time

- Exams logs: In the raw data there are 8305 rows and 11 features for 894 unique exam id for 1079 unique student id.

Table 7: Data Types of Exams Logs Dataset

Categorical	Student unique ID, Exam unique ID Subject unique ID

Table 7 cont.

	Exam type Login unique ID
Continuous	Correct answers count Wrong answers count Blank answers count Score Duration (in seconds)
Timestamp	Start time End time

- Subject logs: In the raw data there are 5643 rows and 5 features for 716 unique subject id.

Table 8: Data Types of Subjects Logs Dataset

Categorical	Student’s enrolled class Lecture type Subject Unique ID
Continuous	Active material count Total material count

3.2. Data Preprocessing

Due to the study aims to predict the students who are attempting to “gaming the system”, the main datapoints to conduct this study are exams log data. To enrich the information on exams data before conducting our analysis, it is aimed at creating “one big table (OBT)” which consists with exams log data and relevant data from other sheets. Primary and foreign keys of each table are used to join related columns to create OBT. Exams logs, subject logs are joined using subject unique id. According to unique subject id, cumulative study and lecture time are calculated for each student according to their starting time of the exam regarding the subject. After joining operations, the populated data under OBT has 1079 unique students' data to conduct this study.

The created OBT has 7878 rows and 19 features. The details of the features are given below:

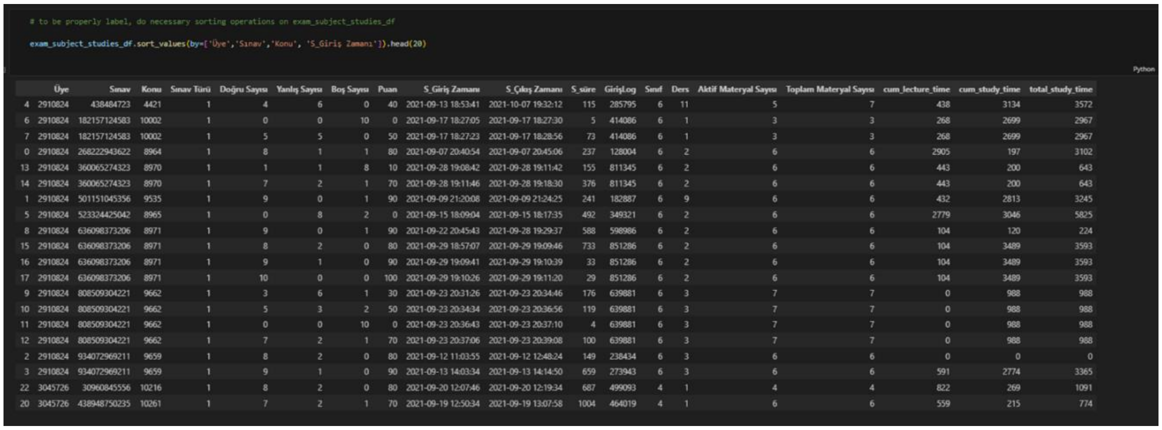
- Student unique ID: It shows the unique id number of students registered to the system.
- Exam unique ID: It shows the unique id number of exam in the system.
- Subject unique ID: It shows the unique id number of subject that related to the exam in the system.
- Exam type: It shows the exam type designed in the system.
- Correct answers count: It shows the count of correct answers for the questions of an exam.
- Wrong answers count: It shows the count of wrong answers for the questions of an exam.
- Blank answers count: It shows the count of left blank for the questions of an exam.
- Score: It shows the score of an exam according to the correct answers. Wrong and blanks answers do not have effect the score.
- Start time: It shows the exam starting time as timestamp in day.month.year hour:minute:second format.
- End time: It shows the exam compete time as timestamp in day.month.year hour:minute:second format.
- Duration: It shows how long students spend time in the exam.
- Login unique ID: It shows the unique id number of each login in the system.
- Student's enrolled class: It shows the students' enrolled class registered in the system.
- Lecture type: It shows the lecture type designed in the system.
- Active material count: It shows the count of active material according to subjects in the system.

- Total material count: It shows the count of total material according to subjects in the system.

- Cumulative lecture time: It shows how log students spend time in lectures until starting the exam in related the subject.

- Cumulative study time: It shows how log students spend time in study until starting the exam in related the subject.

- Total study time: It shows how log students spend time in study and lecture until starting the exam in related the subject.



```python
# to be properly label, do necessary sorting operations on exam_subject_studies_df

exam_subject_studies_df.sort_values(by=['Üye','Sınav','Konu', 'S_Giriş Zamanı']).head(20)
```

	Üye	Sınav	Konu	Sınav Türü	Doğru Sayısı	Yanlış Sayısı	Boş Sayısı	Puan	S_Giriş Zamanı	S_Çıkış Zamanı	S_süre	GirişLog	Sınıf	Ders	Aktif Materyal Sayısı	Toplam Materyal Sayısı	cum_lecture_time	cum_study_time	total_study_time
4	2910824	438484723	4421	1	4	6	0	40	2021-09-13 18:53:41	2021-10-07 19:32:12	115	285795	6	11	5	7	438	3134	3572
6	2910824	182157124583	10002	1	0	0	10	0	2021-09-17 18:27:05	2021-09-17 18:27:30	5	414086	6	1	3	3	268	2699	2967
7	2910824	182157124583	10002	1	5	5	0	50	2021-09-17 18:27:23	2021-09-17 18:28:56	73	414086	6	1	3	3	268	2699	2967
0	2910824	268222943622	8964	1	8	1	1	80	2021-09-07 20:40:54	2021-09-07 20:45:06	237	128004	6	2	6	6	2905	197	3102
13	2910824	360065274323	8970	1	1	1	8	10	2021-09-28 19:08:42	2021-09-28 19:11:42	155	811345	6	2	6	6	443	200	643
14	2910824	360065274323	8970	1	7	2	1	70	2021-09-28 19:11:46	2021-09-28 19:18:30	376	811345	6	2	6	6	443	200	643
1	2910824	501151045356	9535	1	9	0	1	90	2021-09-09 21:20:08	2021-09-09 21:24:25	241	182887	6	9	6	6	432	2813	3245
5	2910824	523324425042	8965	1	0	8	2	0	2021-09-15 18:09:04	2021-09-15 18:17:35	492	349321	6	2	6	6	2779	3046	5825
8	2910824	636098373206	8971	1	9	0	1	90	2021-09-22 20:45:43	2021-09-28 19:29:37	588	598986	6	2	6	6	104	120	224
15	2910824	636098373206	8971	1	8	2	0	80	2021-09-29 18:57:07	2021-09-29 19:09:46	733	851286	6	2	6	6	104	3489	3593
16	2910824	636098373206	8971	1	9	1	0	90	2021-09-29 19:09:41	2021-09-29 19:10:39	33	851286	6	2	6	6	104	3489	3593
17	2910824	636098373206	8971	1	10	0	0	100	2021-09-29 19:10:26	2021-09-29 19:11:20	29	851286	6	2	6	6	104	3489	3593
9	2910824	808509304221	9662	1	3	6	1	30	2021-09-23 20:31:26	2021-09-23 20:34:46	176	639881	6	3	7	7	0	988	988
10	2910824	808509304221	9662	1	5	3	2	50	2021-09-23 20:34:34	2021-09-23 20:36:56	119	639881	6	3	7	7	0	988	988
11	2910824	808509304221	9662	1	0	0	10	0	2021-09-23 20:36:43	2021-09-23 20:37:10	4	639881	6	3	7	7	0	988	988
12	2910824	808509304221	9662	1	7	2	1	70	2021-09-23 20:37:06	2021-09-23 20:39:08	100	639881	6	3	7	7	0	988	988
2	2910824	934072969211	9659	1	8	2	0	80	2021-09-12 11:03:55	2021-09-12 12:48:24	149	238434	6	3	6	6	0	0	0
3	2910824	934072969211	9659	1	9	1	0	90	2021-09-13 14:03:34	2021-09-13 14:14:50	659	273943	6	3	6	6	591	2774	3365
22	3045726	30960845556	10216	1	8	2	0	80	2021-09-20 12:07:46	2021-09-20 12:19:34	687	499093	4	1	4	4	822	269	1091
20	3045726	438948750235	10261	1	7	2	1	70	2021-09-19 12:50:34	2021-09-19 13:07:58	1004	464019	4	1	6	6	559	215	774

Figure 1 : Snapshot of One-Big-Table

The distribution of the enrolled class of students are given below:

Table 9: Number of Students by Enrolled Class

Enrolled Class	Total Students	Percentage
1st Grade	150	7,50%
2nd Grade	291	14,56%
3rd Grade	351	17,56%
4th Grade	392	19,61%
5th Grade	293	14,66%
6th Grade	193	9,65%
7th Grade	137	6,85%
8th Grade	192	9,60%
Total	**1999**	**100,00%**

3.3. Data Analysis

In the absence of pre-existing labels, manual labeling for a classification algorithm proves to be a time-consuming endeavor, particularly when dealing with large datasets. Additionally, such a process is susceptible to biases from the evaluator and demands substantial domain knowledge. To avoid these challenges, we employed the K-means clustering algorithm with Euclidian distance on the log data to identify inherent data patterns. Afterwards, we utilized a rule-based decision-making process to pinpoint the cluster encapsulating students attempting to "game the system." This approach facilitated the creation of labels through clustering, which are then applied to conduct our classification study.

K-means clustering is an unsupervised machine learning algorithm used for partitioning a dataset into K distinct subsets. The aim is to group similar data points together and separate dissimilar ones. The algorithm follows these steps:

- Initialization: Randomly select K data points as initial cluster centers (centroids).

- Assignment: Assign each data point to the cluster whose centroid is the closest based on a Euclidean distance.

- Update Centroids: Recalculate the centroids by taking the mean of all data points assigned to each cluster.

- Repeat Assignment and Update: Iterate steps 2 and 3 until convergence, where convergence occurs when the assignment of data points to clusters and the centroids stabilize.

The algorithm aims to minimize the within-cluster sum of squares (WCSS), which represents the sum of squared distances between each data point and its assigned cluster centroid. At the last iteration, K clusters with centroids that represent the "center" of each cluster.

The application of the Elbow method facilitates the determination of the number of clusters in our analysis. To determine the optimal cluster number, we utilize seven features as input for the Elbow method. These features encompass; correct answers count, wrong answers count, blank answers count, score, duration, start time, and end time. To gauge the influence of each feature, progressively incorporating them into the model one at a time.

According to this method, our data shows an optimal separation into two clusters.

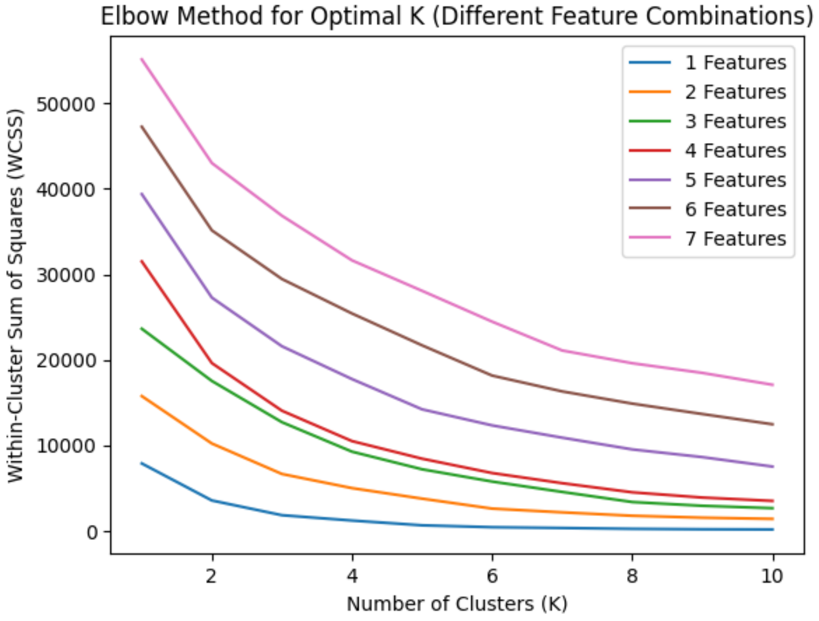


Figure 2 : Elbow Method to Determine Number of Clusters

The Elbow method applied to our data reveals a division into two clusters. The clustering algorithm separates data for cluster 0 as 76% and cluster 1 for 24%. This outcome aligns with a binary labeling approach, distinguishing between instances of "gaming the system" and those not engaging in such behavior. Given the lack of clarity regarding the interpretation in clustering algorithms, supplementing the analysis with domain knowledge backed rule-based labeling assists in clarification which cluster corresponds to students attempting to "game the system".

Rule-based labeling is applied according to the specified criterion: a student is designated as "gaming the system" if they undertake the exam again on the same subject within the same login session, while the previous exam is still ongoing, and if their subsequent score demonstrates an increase. The outcome of the rule-based labeling process identifies 120 attempts as instances of "gaming the system," whereas 7758 attempts are categorized as not involving such behavior.

The labels are used to identify which clusters consist of the student who attempts to “gaming the system”. Notably, 78% of students engaging in this behavior are associated with cluster 1. Consequently, cluster 1 is designated as the label indicative of suspicion regarding "gaming the system" for our classification analysis.

After labeling our data, the analysis reveals that a predominant number of attempts to "game the system" are attributed to 4th grade students.

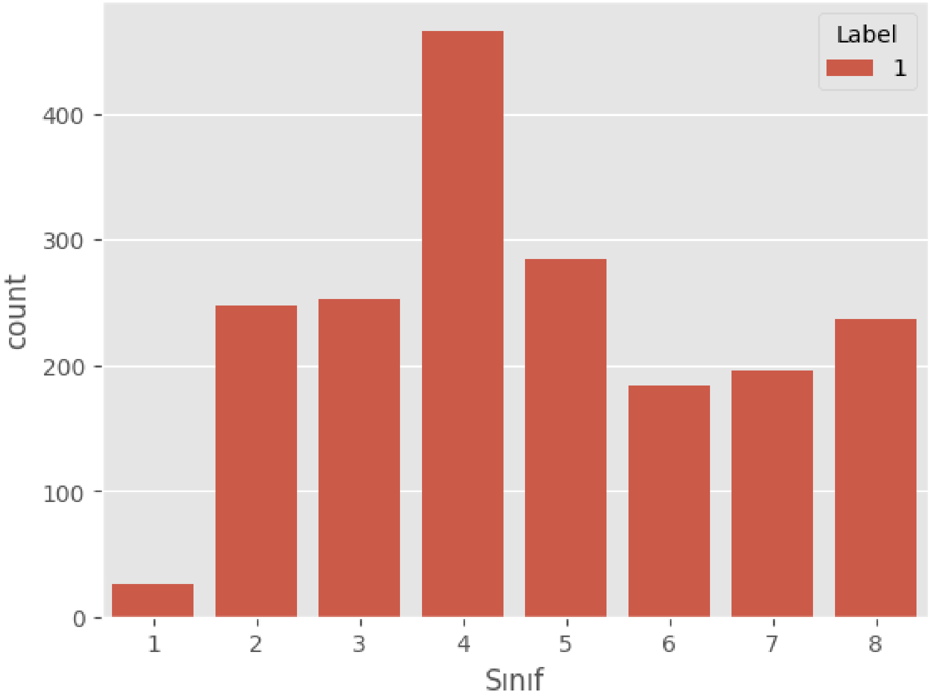


Figure 3: Distribution of Gaming the System by Students' Enrolled Class

Furthermore, the analysis indicates disparity in the duration of exams between students who attempt to "game the system".

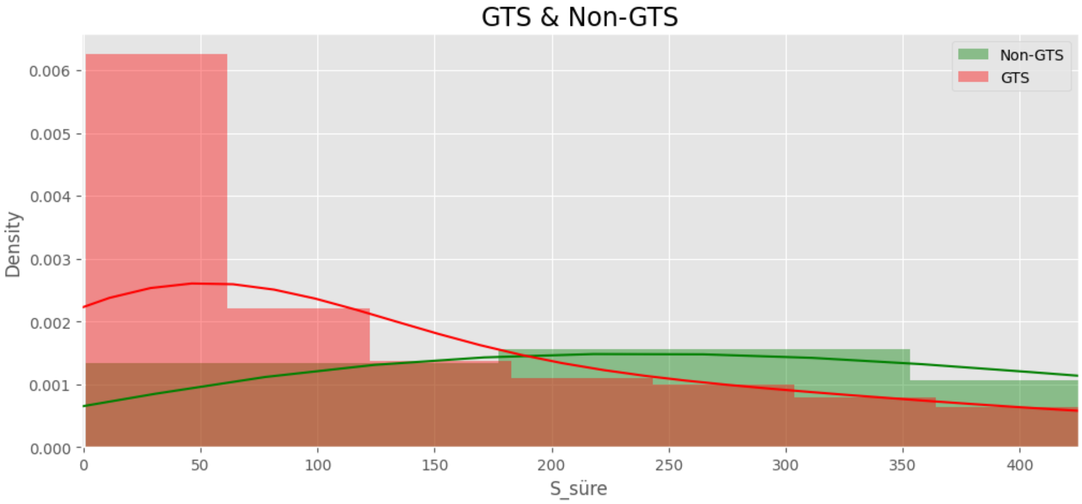


Figure 4 : Distribution of Gaming the System by Exam Duration

With the creation of data labels, our dataset is now prepared for utilization in classification operations.

Classification algorithms play a crucial role in supervised machine learning, where the goal is to categorize instances into predefined classes or labels based on their features. Numerous classification algorithms are well-suited for handling this type of data. Popular choices include decision tree-based methods like Random Forest and Gradient Boosting, with XGBoost being particularly effective. Support Vector Machines (SVMs), k-Nearest Neighbors (k-NN), and Naive Bayes are also commonly employed. These algorithms leverage patterns and relationships within the tabular structure to learn decision boundaries that separate different classes.

Given the tabular structure of our data and the imbalance in target variables, it is decided using XGBoost as classification algorithm for this study. XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems (Nvidia, 2023)

There are some key features and concepts associated with XGBoost:

- Gradient Boosting Algorithm: XGBoost is an implementation of the gradient boosting framework, which builds a predictive model in the form of an ensemble of weak learners (usually decision trees). It sequentially adds trees to correct the errors of the previous ones.

- Regularization: XGBoost incorporates L1 (Lasso) and L2 (Ridge) regularization techniques to prevent overfitting and improve model generalization. Regularization penalizes overly complex models by adding a term to the loss function based on the complexity of the model.

- Tree Pruning: XGBoost uses a depth-first approach for growing trees and applies pruning to control their depth. Pruning helps prevent overfitting and contributes to the algorithm's efficiency.

- Continuous Variables: XGBoost utilizes a process similar to traditional gradient boosting algorithms. It recursively partitions the data based on the continuous feature values, creating decision trees that split the data into subsets.

- Categorical Variables: XGBoost employs a technique called the “tree method" to handle categorical variables directly. It internally encodes categorical variables into numerical values, assigning unique integers to each category. This encoding allows XGBoost to incorporate categorical features into the tree-building process.

- Weighted Instances: XGBoost allows for the assignment of different weights to instances in the training data. This is particularly useful for imbalanced datasets, where you can assign higher weights to minority class instances, ensuring that the algorithm pays more attention to learning patterns in the underrepresented class.

- Feature Importance: XGBoost provides a feature importance score, allowing users to understand the relative importance of different features in making predictions.

In the classification algorithm, the identification of features involves an accurate analysis utilizing the correlation matrix, specifically focusing on features exhibiting low correlations. Consequently, features such as subject, exam duration, enrolled class, active study material count, and total study time have been selected as significant determinants.

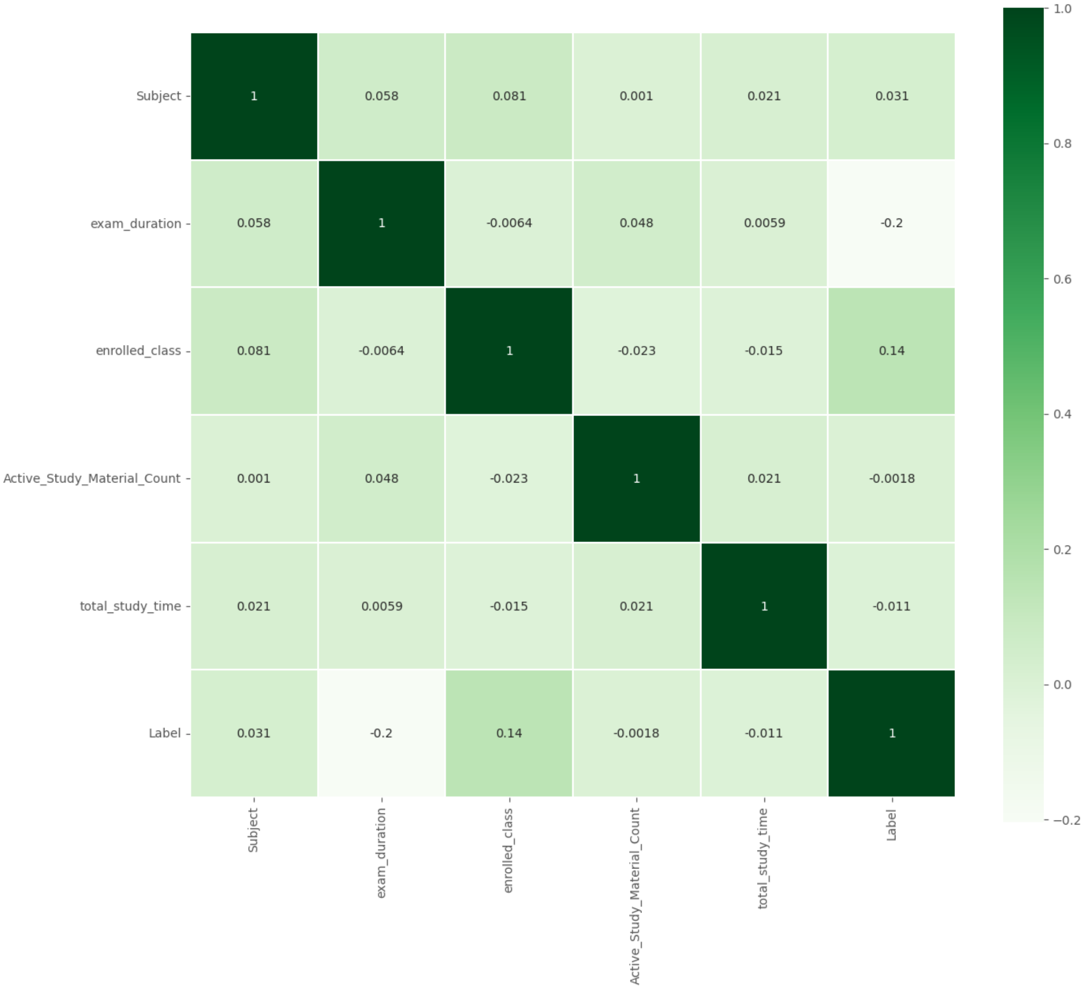


Figure 5: Correlation Matrix for Classification

Since our data imbalanced, the evaluation metric selected as “precision-recall”. Precision and recall are preferred as evaluation metrics on imbalanced datasets due to their sensitivity to the challenges posed by a significant class imbalance. In scenarios where one class is underrepresented, accuracy can be misleading, making it crucial to assess a model's performance specifically with respect to the minority class. Precision measures the accuracy of positive predictions, highlighting how well the model correctly identifies instances of the minority class. Recall, on the other hand, gauges the model's ability to capture a substantial proportion of actual positive instances. This focus on both false positives and false negatives allows for understanding of the model's behavior whether prioritizing precision, recall.

Our initial trained model produces an evaluation metric score of 0.59. While this metric shows promise, there remains scope for improvement. The feature importance of this model is outlined below:

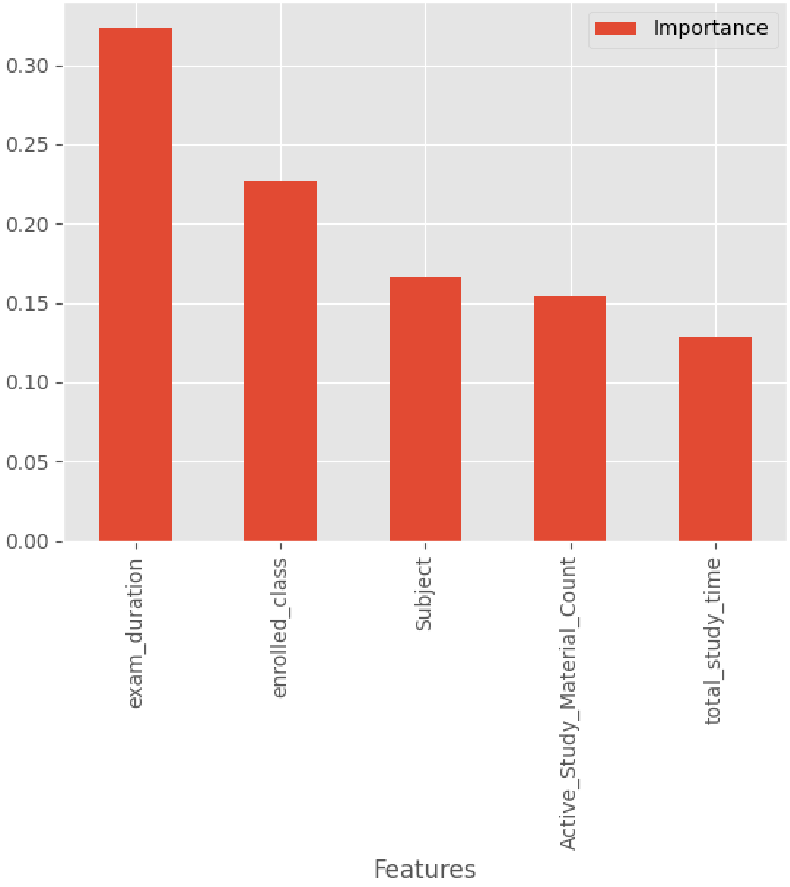


Figure 6: Feature Importance for Trained Classification Model

The trained model shows that the five most important determinants of the classification method are exam duration, enrolled class, subject, active study material count, and total study time.

The confusion matrix generated by the trained model is presented below:

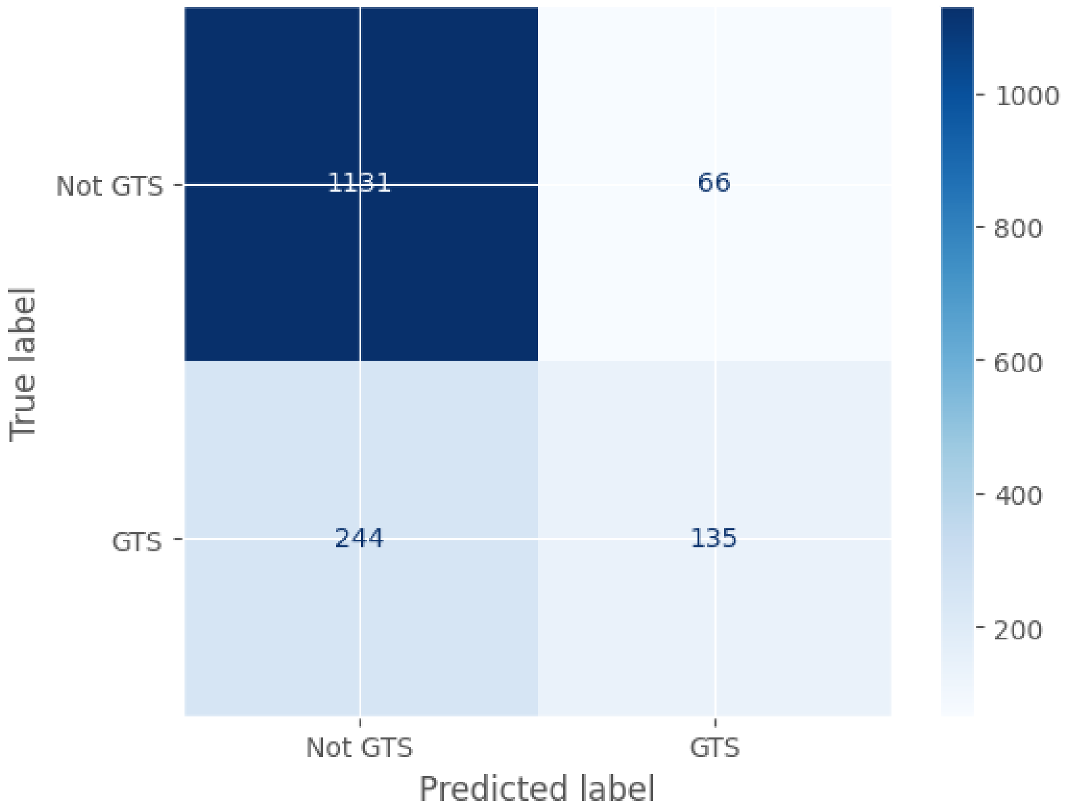


Figure 7: Confusion Matrix for Trained Classification Model

The confusion matrix shows that in our test subset, 66 students fall into false positive as Type I error. On the other hand, 244 students fall into false negative as Type II error.

Utilizing the GridSearchCV technique, an extensive search is conducted to identify the optimal parameters for the classification model. The most effective parameters determined through GridSearchCV with 3-fold cross validation are as follows:

- gamma: 0.25
- learning_rate: 0.05
- max_depth: 5
- reg_lambda: 10.0
- scale_pos_weight: 1

After training our classification model with given parameters, it produces an evaluation metric score of 0.60. Feature importance of the model with optimal parameters is given below:

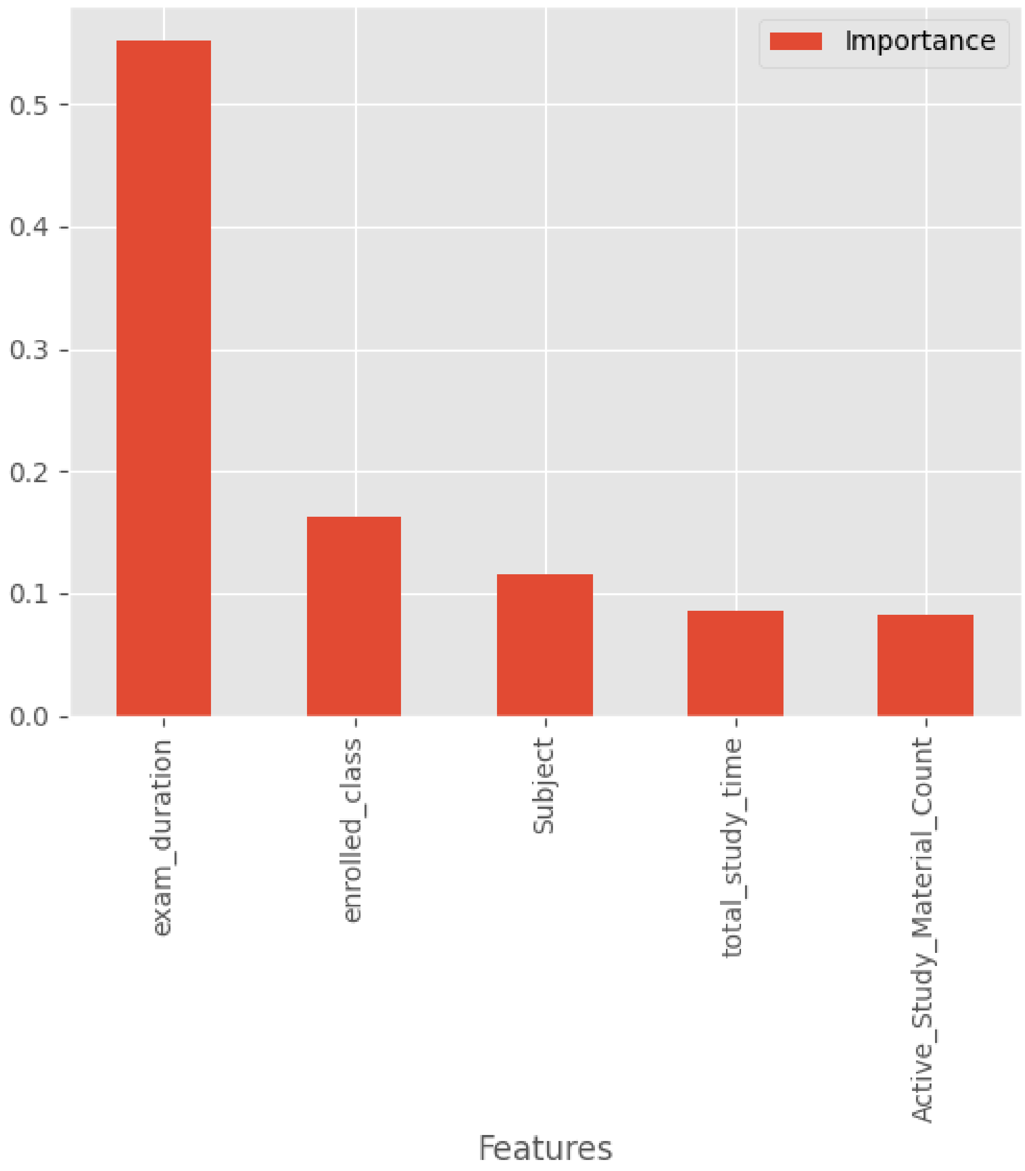


Figure 8: Feature Importance for Trained Classification Model with Optimal Parameters

The confusion matrix generated by the trained model with optimal parameters is presented below:

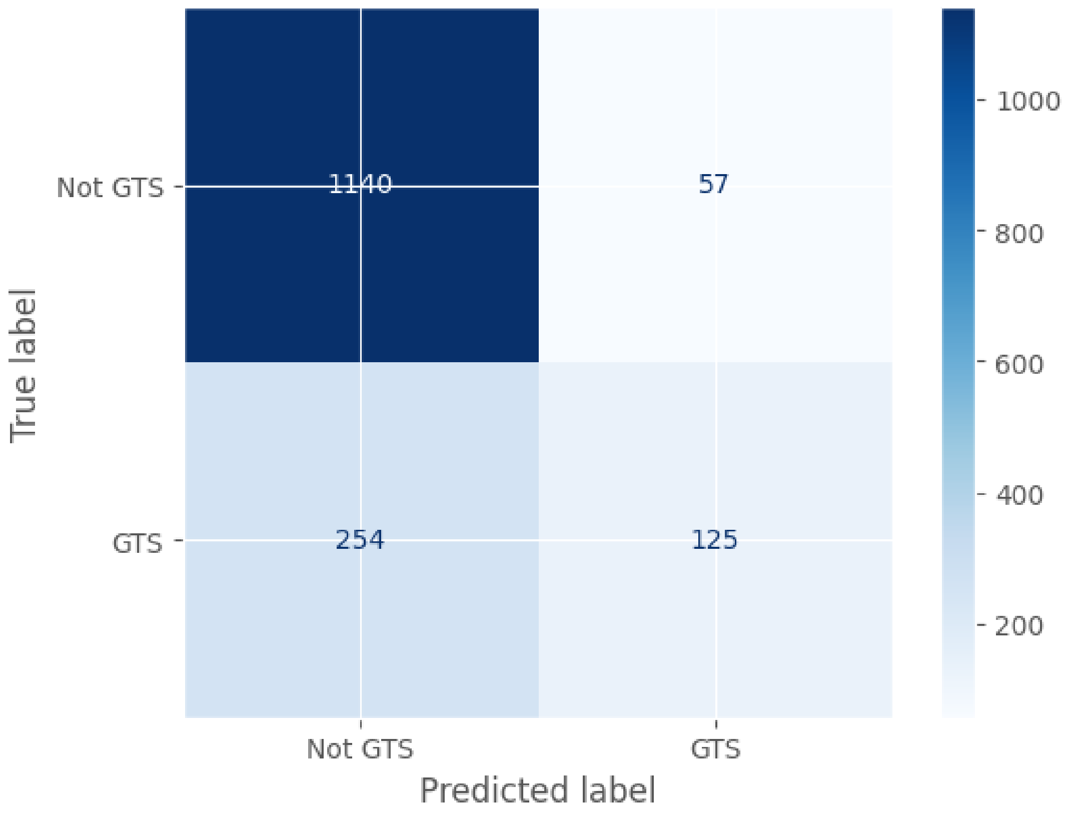


Figure 9: Confusion Matrix for Trained Classification Model with Optimal Parameters

In this section, we provide a comprehensive overview of the methods employed during the analysis phase of our study. The results derived from these methodologies will be presented in the subsequent section.

CHAPTER 4

RESULTS

The results of the outcomes of a comprehensive investigation into LMS server log data are presented. The primary objective is to determine patterns in student behaviors, particularly instances indicative of attempts to manipulate the system, commonly known as "gaming the system." The initial phase involved the application of the K-means clustering algorithm to unlabeled data, enabling the identification of inherent behavior patterns. Subsequently, the predictive capabilities of the XGBoost machine learning algorithm were implemented in order to analyze what extend to the independent variables of exam duration, subject of the exam, class enrollment, the count of active study materials, and total study time to identify and predict instances of users attempting to manipulate the system within an LMS that defined as research question of this study. This section delves into the findings obtained through these methodologies, providing a detailed examination of the identified patterns and the effectiveness of the predictive model in detecting and categorizing instances of interest within the student log data.

4.1. K-Means Clustering Algoritm Results

In accordance with the methodology outlined in the previous chapter, the investigation of LMS data proceeded in the absence of pre-existing labels. To distinguish meaningful patterns in students' behavior within this unlabeled dataset, the K-means clustering algorithm is employed. Drawing on domain knowledge, a selection of relevant features including correct answers count, wrong answers count, blank answers count, score, duration, start time, and end time of the exam were identified. These features were regarded crucial for characterizing students' behavior during exams. Through the exploration of within-cluster sum of squares values, it is observed that the most substantial decrease occurred when the data is clustered into two distinct groups. This finding indicates the optimal partitioning of the data into two separate clusters, revealing distinct patterns in students' exam behavior.

The outcomes of the clustering algorithm reveal a distinct distribution of students within the identified clusters. Specifically, a majority, comprising 75.92% of the student population, is assigned to cluster 0, indicating a common pattern of behavior among this subgroup. In contrast, 24.08% of students are categorized into cluster 1, highlighting a distinguishably different behavioral profile within this subset of the student population. These cluster assignments provide insights into the prevalence and diversity of behaviors exhibited by students present in the LMS data under investigation.

The students count of clusters are 5981 students fall into cluster 0 and 1897 students are assigned as cluster 1.

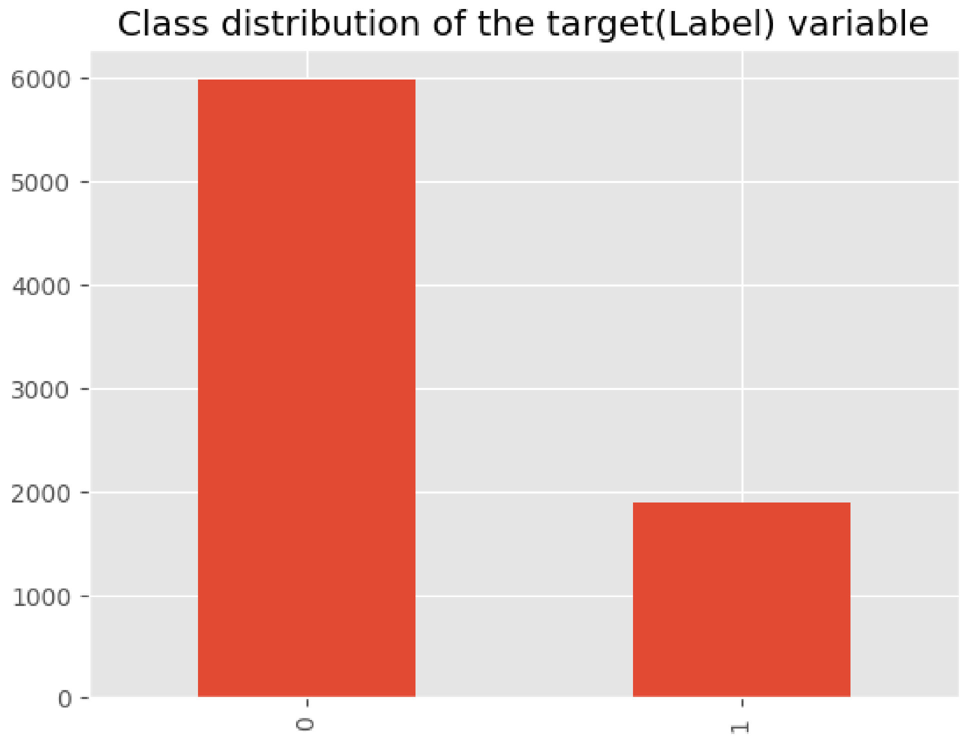


Figure 10: Distribution of Clusters of LMS Log Data

Students affiliated with cluster 0 exhibit higher average exam scores, standing at 90.14. In contrast, students aligned with cluster 1 demonstrate lower average exam score registering at 30.47.

The incorporation of domain knowledge into a rule-based labeling approach used to figure out the cluster that encapsulates students engaging in the behavior of "gaming the system." Following this rule-based labeling methodology, a total of 120 students are identified as attempting to "game the system," while 7758 students do not exhibit such behavior. The result shows that 78% of the students labeled as engaging in "gaming the system" through the rule-based approach are found within cluster 1. Cluster 1 is designated as the cluster of students suspected of "gaming the system."

4.2. XGBoost Classification Algoritm Results

The clustering algorithm is used in assigning labels to identify students suspected of "gaming the system," then facilitating the construction of a classification model for predicting whether a student is engaged in such behavior. The dataset is partitioned into training and test sets, with an 80% allocation to training data and 20% to test data. To maintain a balanced representation of the imbalanced data in both training and test datasets, a stratified sampling method is employed. This strategic sampling approach ensures that the distribution of suspected "gaming the system" instances remains proportional across both training and test datasets, enhancing the robustness and generalizability of the subsequent classification model.

Initially, the XGBoost classification model is trained with default parameter values, incorporating an early stopping criterion set at 10 rounds. Given the imbalanced nature of the data, the area under the precision-recall curve (AUCPR) is employed as the evaluation metric, yielding a value of 0.59 According to this model, the feature importance is outlined below:

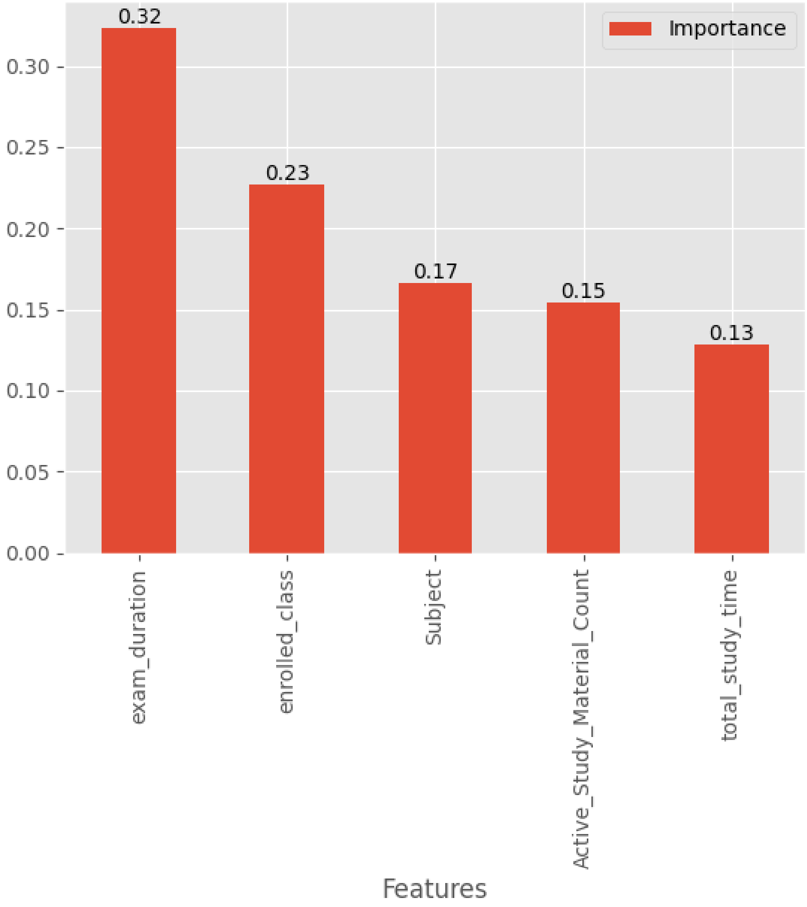


Figure 11: Feature Importance of Classification Model with Data Labels

In the classification model, the exam duration feature emerges as the most influential among other features, contributing with a weight of 32%. Following, the enrolled class of students constituting 23% of the model's predictive power. The subject of the exams and its relevance to active study materials contribute with weights of 17% and 15%, respectively. Finally, the total study time until the exam contributes with a weight of 13% of the predictive capacity of the model.

The confusion matrix of classification model outlined below:

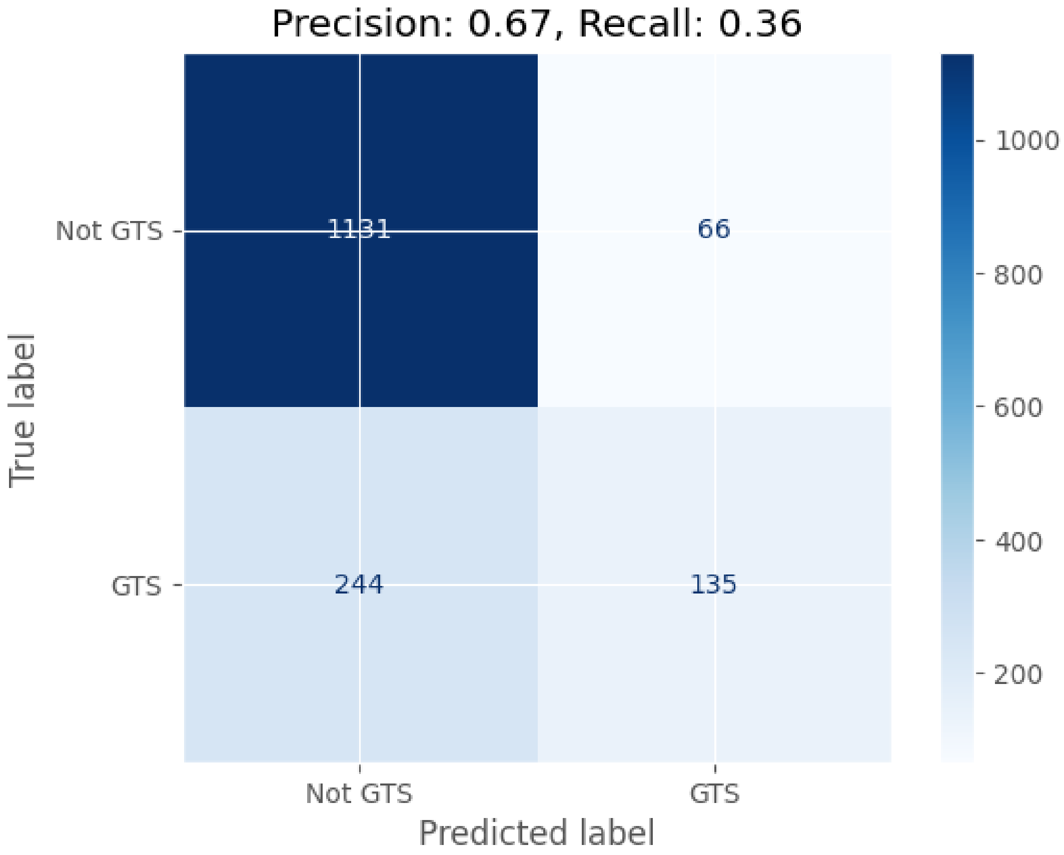


Figure 12: Confusion Matrix for Trained Classification Model with Precision, Recall Scores

The confusion matrix shows that the classification model has a 0.67 precision score. Precision score shows the ratio of correct positive predictions to total predicted positives. The classification model produced a true positive outcome for 135 students, indicating that these individuals were identified as exhibiting manipulation attempts within the Learning Management System. On the other hand, the model generated a Type I error for 66 students, signifying instances where students were incorrectly classified as engaging in manipulation attempts.

4.3. Classification Algoritm Results With Optimal Parameters

Following an initial assessment of the classification model's general performance, a more in-depth investigation is undertaken to determine if the model outperforms a baseline counterpart. Employing GridSearchCV, optimal parameters are identified to enhance the model's predictive capabilities. With the application of these optimized parameters, the evaluation metric score becomes 60% that shows a slight improvement of nearly 1%. The updated feature importance analysis result is outlined below:

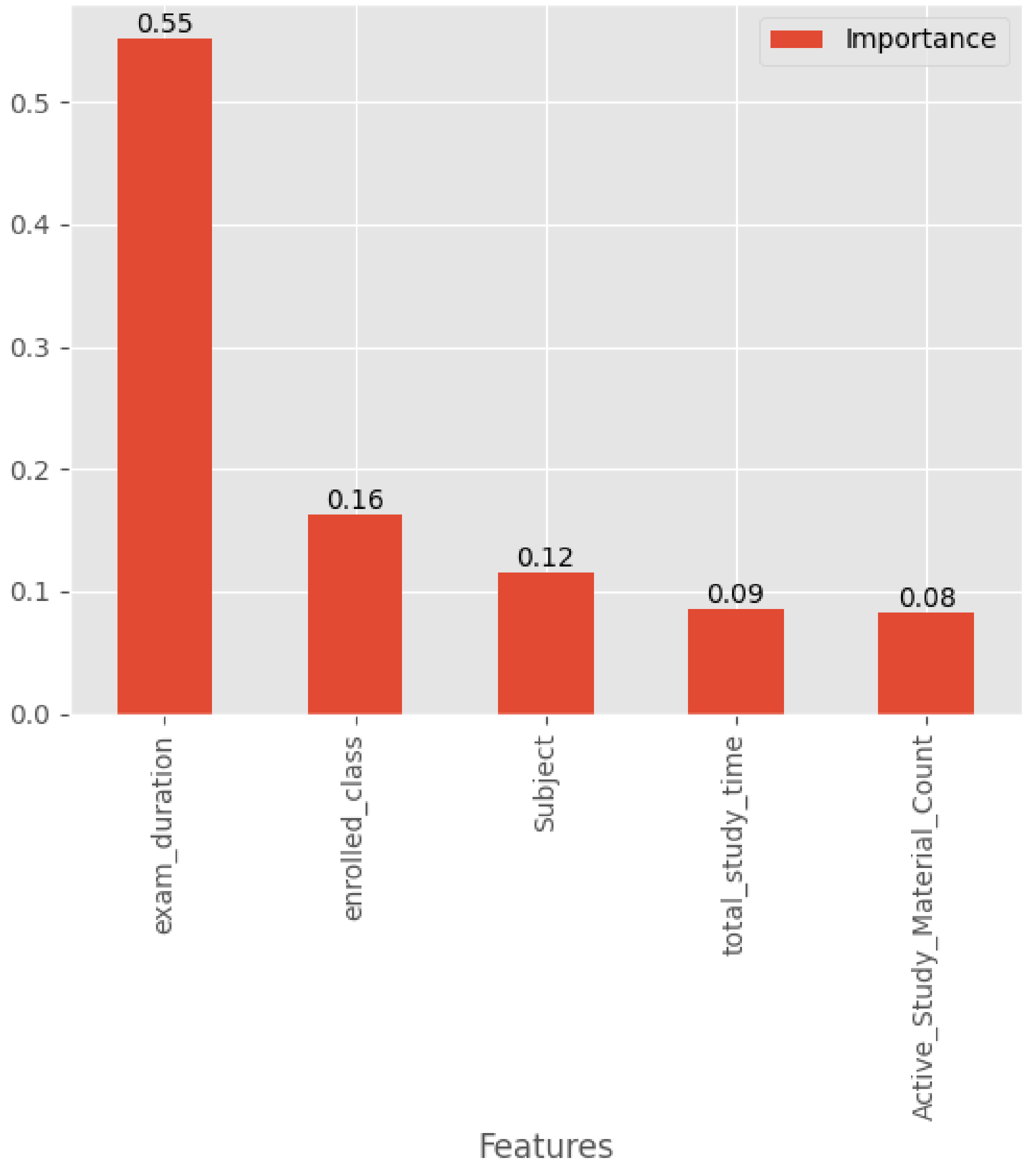


Figure 13: Feature Importance of Classification Model with Optimal Parameters and Data Labels

In the classification model with optimal parameters, the exam duration feature improves its model contribution from 32% to 55% and remains the most influential among other features. Following, the enrolled class of students constituting 16% of the model's predictive power. The subject of the exams and its relevance to active study materials contribute with weights of 12% and 9%, respectively. Finally, the total study time until the exam contributes with a weight of 8% of the predictive capacity of the model.

The confusion matrix of classification model with optimal parameters is outlined below:

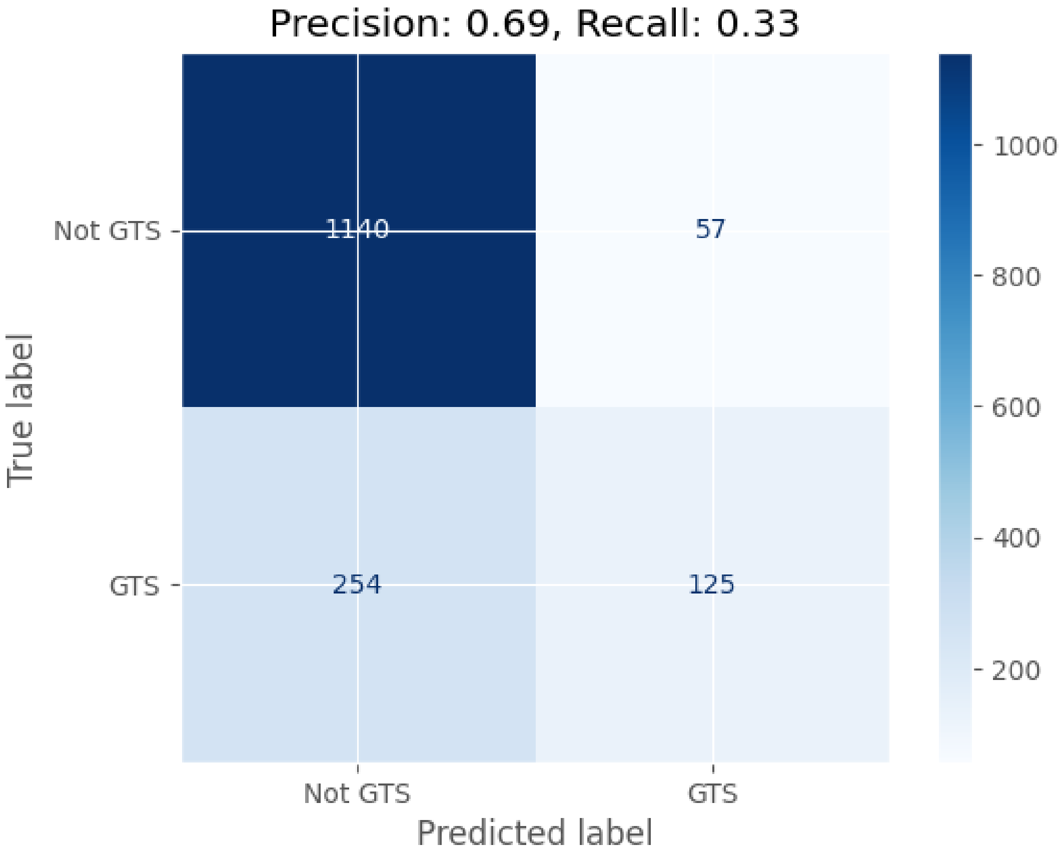


Figure 14: Confusion Matrix for Trained Classification Model with Optimal Parameters and Precision, Recall Scores

The confusion matrix shows that the classification model has a 0.67 precision score. Precision score shows the ratio of correct positive predictions to total predicted positives. The classification model produced a true positive outcome for 135 students, indicating that these individuals were identified as exhibiting manipulation attempts within the Learning Management System. On the other hand, the model generated a Type I error for 66 students, signifying instances where students were incorrectly classified as engaging in manipulation attempts.

Aligned with our research question, this study employs, as suggested with domain knowledge, exam duration, class enrollment, the count of active study materials, and total study time as key features for predicting instances of users attempting to "game the system" within an LMS. In both the baseline model and the model with optimal parameters, the rank order of features according to feature importance scores remains consistent. The classification models consistently highlight the role of "exam duration" as the most influential feature in predicting manipulation attempts. Following hyperparameter optimization, the classification model with optimal parameters reinforces the significance of exam duration by assigning it an increased weight in determining the predictive capability of the model. The robustness of this feature is further emphasized by the evaluation metric, which demonstrates a substantial drop from 60% to 37% when exam duration is excluded from the model.

CHAPTER 5

DISCUSSION AND CONCLUSION

In this study, by using the data obtained from the learning activities on an LMS used by from 1st grade students to 8th grade students, it was investigated to predict the instances of users attempting to manipulate the system within an LMS and which of the variables used during this study were effective on these results.

In existing literature, numerous studies have investigated the detection of "gaming the system" behaviors within educational settings. However, a common trend among these studies is the reliance on labeled data, typically annotated by domain experts, through observational methods, or based on reports.

Baker, et al. (2004) aim to develop a machine-learned Latent Response Model (LRM) that can identify if a student is gaming the system in a way that leads to poor learning. The LRM was trained on multiple sources of data, including log files of student actions, human-coded observations of student behavior, and student learning outcomes. Paquette and Baker (2019) point out the use of knowledge engineering and machine learning methods in modeling student behaviors in digital learning environments. The study highlights the lack of direct comparison between these two approaches and aims to compare their relative advantages in the context of modeling "gaming the system" behavior. Knowledge engineering involves developing models based on experts' knowledge, while machine learning uses data-driven algorithms to discover relationships between student behaviors.

The use of labeled data allows researchers to train and evaluate machine learning models in a supervised manner, enhancing the accuracy of identifying instances of system manipulation. While this approach has proven valuable, it is essential to acknowledge potential biases in the labeling process and explore alternative methodologies that address challenges associated with obtaining labeled data, especially in instances where domain expertise, observation, or reporting may be subjective or incomplete. This study contributes to the existing literature by exploring predictive models in an environment where labeled data is not readily available, aiming to enhance the robustness and applicability of detection methodologies.

In the absence of labeled data in the log records of the Learning Management System (LMS), the initial step involves uncovering patterns within the data to discern different types of data points. A prevalent approach in the literature for such unsupervised learning tasks is the utilization of K-means clustering algorithms. By applying K-means clustering, researchers aim to reveal latent structures within the unlabeled data, facilitating a deeper understanding of student behaviors or system interactions.

Duhaim, et al. (2022) use K-means, Hierarchical, and Expectation Maximization clustering algorithms are employed to develop a well-designed clustering strategy for detecting cheating. The authors are implementing detecting cheating mechanisms in the light of IP addresses identification, by time (time taken or time late) and identifying clusters with K-means clustering method. Fernando Raguro, et al. (2022) focuses on the extraction of student engagement and behavioral patterns in online education using decision tree and K-means algorithm.

In this study, a selection of features was chosen to employ in the K-means clustering algorithm to observe students' behavior during exams within an LMS. The identified features include correct answers count, wrong answers count, blank answers count, score, duration, start time, and end time of the exam. The rationale behind the selection of these specific features lies in their representation to capturing variations in student behavior during exams such as instances were observed where students left a significant number of questions blank in the initial trial then to exhibit an increased score in subsequent trials within the same session, raising suspicions of potential manipulation. Given the complexity of student interactions with the exam content and the dynamic nature of these features, their incorporation into the K-means algorithm enables the identification of distinct behavioral patterns at detecting anomalies or irregularities within the LMS.

The clustering algorithm, employing the Elbow method to determine the optimal number of clusters, has yielded a division into two distinct clusters. To interpret these clusters, a domain knowledge-backed rule-based identification process was implemented, revealing that cluster 1 comprises nearly 76% of students exhibiting suspicious activity, possibly attempting to "game the system." While this cluster does not directly identify students engaged in such behavior, it serves as an indicator of students with potentially irregular activities.

The observed disparity in average exam scores between the two clusters aligns with existing literature on the relationship between gaming system behaviors and academic performance. In a study conducted by Baker et al. (2004), findings indicate that students who frequently engage in gaming the system tend to demonstrate lower learning outcomes. The results of the current study, where students in cluster 1, associated with suspicious activities indicative of potential attempts to "game the system," exhibit a substantially lower average exam score of 30.47 compared to their counterparts in cluster 0 with a score of 90.14, substantiate and echo the earlier research.

These parallel findings reinforce the notion that behaviors suggestive of gaming the system may indeed be associated with diminished academic achievement. The integration of empirical evidence from this study, coupled with insights from existing literature, contributes to a more comprehensive understanding of the impact of such behaviors on student learning outcomes within the context of LMS.

In the literature, if labels have been created or obtained from the data, classification algorithms become a valuable tool for predicting and identifying students who exhibit suspicious behaviors indicative of gaming the system. Various studies have employed classification algorithms to categorize and predict such behaviors based on labeled data. Commonly used algorithms include logistic regression, decision trees, support vector machines, and ensemble methods like Random Forest and Gradient Boosting.

Fernando Raguro, et al. (2022) utilized machine learning algorithms, specifically the Decision Tree Algorithm and the K-means Algorithm, to extract predictive rules sets and profile students' behaviors in the Learning Management System (LMS) environment. Zhao, et al. (2023) examines academic cheating among Chinese second to sixth graders using a machine learning approach to demonstrates that machine learning can be effectively used to analyze developmental data. The study uses Random Forest machine learning model to predict cheating behavior.

In this study, the XGBoost algorithm is selected as the classification model to predict instances of suspicious behaviors, particularly attempts to game the system within the Learning Management System. The features used for this classification models are subject, exam duration, class enrollment, the count of active study materials, and total study time. Initial results reveal a precision score of 67% without any hyperparameter optimization. Additionally, the Random Forest algorithm was explored as an alternative for the analysis; however, it did not outperform XGBoost, yielding a precision score of 61%. The preference for XGBoost over Random Forest aligns with the observed precision scores, suggesting that XGBoost is better suited for the specific characteristics of the dataset and task at hand. The utilization of XGBoost indicates its capability to handle complex relationships within the data, contributing to improved predictive performance. Further fine-tuning of hyperparameters in XGBoost enhances its precision score by avoiding overfitting, providing an opportunity for optimizing the model's performance in identifying instances of gaming the system within the LMS.

The results of the classification algorithm without hyperparameter optimization reveal a confusion matrix where 135 students are correctly predicted as True Positive, while 66 students are incorrectly predicted as False Positive (Type I error). Afterwards, training the classification algorithm with optimal hyperparameters, the updated confusion matrix shows improvement. Specifically, there are now 125 students correctly predicted as True Positive, and the number of students predicted as False Positive (Type I error) decreases to 57. This modest improvement in precision metrics is gained, but the notable enhancement is the reduction in Type I errors.

Minimizing Type I errors is crucial to prevent the mislabeling of students as engaging in gaming the system incorrectly. This improvement not only contributes to the accuracy of predictions but also holds significance in avoiding potential discouragement among students who may be erroneously flagged for suspicious behavior.

5.1. Feature Importance of Classification Model

In the feature importance aspect, exam duration is the most important feature in the classification model aligns with a logical interpretation of student behavior during the testing process. The insight that students may engage in a trial-and-error strategy, learning from initial attempts to improve subsequent results, is plausible. When students are given the opportunity to review their initial responses and learn from their mistakes, this behavior is reflected in the extended exam duration. The observed pattern suggests an adaptive learning approach, where students leverage insights gained during the initial attempt to enhance their performance in subsequent trials.

Class enrollment is the second most important feature of detecting gaming system behavior. Students can be overwhelmed with the increased courseload compared to their previous class workload. If the previous course load is lower than the current one, the student is more likely to justify academically dishonest practices (Chow et al., 2021).

Subject difficulty as a crucial factor in detecting potential cheating behavior is insightful. The difficulty level of a subject can influence student behavior, potentially steering some students toward engaging in dishonest practices. Several factors contribute to this phenomenon. Students might feel compelled to cheat when faced with a subject perceived as challenging, either due to a lack of preparation or a desire to achieve a more favorable outcome. The perceived difficulty could create pressure, leading some students to resort to dishonest tactics to cope with the academic challenges presented by the subject.

The count of active study materials and total study time are identified as less important features in detecting potential gaming of the system aligns with an interesting behavioral insight. It suggests that, in the context of gaming the system, students might be more focused on specific strategies during the exam (e.g., trial-and-error, revisiting questions) rather than extensive preparation or engagement with study materials. This finding implies that, when students are attempting to manipulate the system, the emphasis may shift away from comprehensive study efforts or prolonged engagement with learning materials. Instead, their behaviors during the exam, such as the number of attempts or the duration of the exam, become more prominent indicators.

In this study, exam score was also tried in the classification task which could be seen as meaningful predictor when predicting exam related data. However, the issue of data leakage arises when the exam score, a potentially significant feature for the classification task, is utilized both in the clustering algorithm and subsequently in the classification model. This dual usage can lead to unintended information transfer between the two processes, influencing the results in an undesirable manner. To address this concern, an approach involves excluding the exam score from the clustering algorithm, recognizing that it is a key feature for the classification task. However, this adjustment may result in a lack of clear and significant cluster detection when relying solely on the remaining features.

5.2. Limitations

Identifying and addressing data quality issues is crucial for ensuring the reliability and validity of study outcomes. In the context of your study on the Learning Management System (LMS), the following suggestions aim to enhance data quality and contribute to more robust data analysis in future studies. In the studied raw data, exam duration of some data points exceeds 200 days where the longest session duration in the data is around 12 hours. In some data points in the raw data has also some discrepancies between exam duration and the difference between exam start time and exam ending time. These occurrences point out the data quality issues that the data generated by the system.

The study's dataset, encompassing student behaviors on exams, is confined to the period between September 1, 2021, and September 30, 2021. This period falls into fall semester during ongoing Covid-19 precautions. This constraint encapsulates only 1999 unique instances of student behavior, and after data preprocessing phase there are 1079 unique instances of student behavior remaining for OBT, presenting a limitation to the study's generalizability. A more expansive dataset, covering a broader temporal scope and incorporating a larger sample of diverse student behaviors, would likely contribute to a more robust and comprehensive predictive model, thereby addressing the potential limitations stemming from the restricted dataset employed in this investigation.

The raw LMS data provides games logs dataset. However, this dataset does not provide subject_id or any other foreign key. The lack of these identifiers hinders the ability to join this dataset to our one big table which is created with combined features with the aid of feature engineering to gain more information, limiting the extent of information available for analysis. The inclusion of "games logs" data that incorporates related subject_id information could be used for enriching our combined dataset. Integrating these game logs presumably enhances the capacity to analyze user behavior in a subject-specific context and has the potential to uncover patterns and correlations between gaming behaviors and academic subjects, affording a more comprehensive understanding of how users interact with the Learning Management System across different domains of study.

5.3. Assumptions & Suggestions

In the context of this study, several recommendations are proposed to enhance the effectiveness of the Learning Management System (LMS) in preventing gaming the system behaviors. These suggestions are as follows:

- Delaying the presenting of the exam answers, such as implementing a time lag of 5 minutes post-completion could mitigate potential gaming behaviors within the LMS. By introducing a time delay before presenting the answers, students are less inclined to employ immediate post-exam feedback for subsequent attempts, thus reducing the effectiveness of gaming strategies. This approach could discourage the reliance on immediate answer disclosure as a means of gaming the system.

- Multiple-choice questions can be combined with open-ended questions. This can discourage trial-and-error behavior. This in turn should lead to more productive thought by the student, as evidenced in their answers such as thinking about concepts rather than test-taking strategies such as process of elimination or guessing (Meir et al., 2019).

- The inclusion of time spent on each question in the exams dataset could provide deeper insights into student behavior, particularly those attempting to game the system. This enriched information allows for a more detailed profiling of student interactions, enabling the identification of patterns associated with gaming behaviors. Students employing strategies such as prolonged consideration of specific questions or rapid responses to others may exhibit distinct temporal patterns, aiding in the early detection of irregularities.

- The potential inclusion of subject_id in the games logs dataset as a foreign key, facilitated by the Learning Management System (LMS) data architecture allows for a more unified and consolidated dataset. This utilization facilitates efficient data linking and cross-referencing between different datasets, enabling researchers to draw correlations between gaming activities and subject-specific games log.

- Further exploration may involve finding alternative features or additional data sources to compensate for the exclusion of exam scores in the clustering process, ensuring the robustness and effectiveness of both the clustering and subsequent classification tasks.

- Improving data quality is an essential consideration for robust data analytics within the LMS. The presence of inconsistencies among related features poses a challenge to sound data analysis, potentially impeding the accuracy and reliability of findings.

- It is assumed that the LMS system used in the study was used by students as desired.

- It is assummed that each student only have one account in the system.

5.4. Implication for Further Research

The findings of this study interpret various aspects of student behavior within the LMS and offer insights into the detection and prevention of gaming the system activities. However, the complexity and dynamic nature of educational technologies justify further investigation and exploration. As such, the implications for future research are as follows:

- Expanding the scope of this analysis by incorporating a larger dataset holds the potential to provide a more robust evaluation of the proposed methodology's accuracy and effectiveness. With increased data volume, researchers can assess the scalability and generalizability of the methodology across a more diverse set of scenarios and student populations. This expanded analysis could uncover additional patterns and potential challenges that might not be evident in a smaller dataset, enhancing the methodology's reliability and applicability in varied educational contexts.

- Extending the analysis to include other Learning Management Systems can serve as a valuable step in assessing the generalizability and applicability of the proposed methodology beyond the specific system under investigation. Understanding how the methodology performs in different settings contributes to establishing its generalizability and robustness. It also enables the identification of potential system-specific factors that may influence the methodology's outcomes.

REFERENCES

Ahmed, K., & Mesanovic, M. (2019). Learning Management Systems and Student Performance. *International Journal for E-Learning Security, 8*, 582–591. https://doi.org/10.20533/ijels.2046.4568.2019.0073

Alshareef, F. (2020a). Educational Data Mining Applications and Techniques. *International Journal of Advanced Computer Science and Applications*, *11*(4).

Alshareef, F. (2020b). Educational Data Mining Applications and Techniques. *International Journal of Advanced Computer Science and Applications*, *11*(4).

Amastini, F. (2014). Intelligent Tutoring System. *Ultima InfoSys : Jurnal Ilmu Sistem Informasi*, *5*(1), 1–7. https://doi.org/10.31937/si.v5i1.212

Awad, M., Salameh, K., & Leiss, E. L. (2019). Evaluating Learning Management System Usage at a Small University. *Proceedings of the 2019 3rd International Conference on Information System and Data Mining*, 98–102. https://doi.org/10.1145/3325917.3325929

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004). Detecting Student Misuse of Intelligent Tutoring Systems. In J. C. Lester, R. M. Vicari, & F. Paraguaçu (Eds.), *Intelligent Tutoring Systems* (Vol. 3220, pp. 531–540). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-30139-4_50

Beal, C. R., & Cohen, P. R. (2008). Temporal Data Mining for Educational Applications. In T.-B. Ho & Z.-H. Zhou (Eds.), *PRICAI 2008: Trends in Artificial Intelligence* (Vol. 5351, pp. 66–77). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-89197-0_10

Bergeron, B. P. (2008). Learning & retention in adaptive serious games. *Studies in Health Technology and Informatics*, *132*, 26–30.

Cantabella, M., Martínez-España, R., Ayuso, B., Yáñez, J. A., & Muñoz, A. (2019). Analysis of student behavior in learning management systems through a Big Data framework. *Future Generation Computer Systems*, *90*, 262–272. https://doi.org/10.1016/j.future.2018.08.003

Chow, H., Jurdi-Hage, R., & Hage, H. S. (2021). Justifying academic dishonesty: A survey of Canadian university students. *International Journal of Academic Research in Education*, *7*(1), 16–28. https://doi.org/10.17985/ijare.951714

Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education*, *18*(6), Article 6.

Crawford, K. (2013). Think Again: Big Data. *Foreign Policy*. https://www.microsoft.com/en-us/research/publication/think-big-data/

de Sande, J. C., Fraile, R., Arriero, L., Osma, V., Oses, D., & Godino, J. I. (2010). *CHEATING AND LEARNING THROUGH WEB BASED TESTS*.

Duhaim, A. M., Al-mamory, S. O., & Mahdi, M. S. (2022). Cheating Detection in Online Exams during Covid-19 Pandemic Using Data Mining Techniques. *Webology*, *19*(1), 341–366. https://doi.org/10.14704/WEB/V19I1/WEB19026

Dutt, A., & Ismail, M. A. (2019). Can We Predict Student Learning Performance from LMS Data? A Classification Approach. *Proceedings of the 3rd International Conference on Current Issues in Education (ICCIE 2018)*. Proceedings of the 3rd International Conference on Current Issues in Education (ICCIE 2018), Yogyakarta, Indonesia. https://doi.org/10.2991/iccie-18.2019.5

Duykuluoğlu, A., Dumitrascu, A., Martinez, M. N., & Simarro, J. F. B. (2023). Educational Big Data and Its Functions. *Technium Education and Humanities*, *4*, 30–39. https://doi.org/10.47577/teh.v4i.8424

Emetere, M. E. (2019). Big Data and Further Analysis. In M. E. Emetere, *Environmental Modeling Using Satellite Imaging and Dataset Re-processing* (Vol. 54, pp. 141–170). Springer International Publishing. https://doi.org/10.1007/978-3-030-13405-1_5

Fernando Raguro, Ma. C., Carpio Lagman, A., P. Abad, L., & S. Ong, P. L. (2022). Extraction of LMS Student Engagement and Behavioral Patterns in Online Education Using Decision Tree and K-Means Algorithm. *2022 4th Asia Pacific Information Technology Conference*, 138–143. https://doi.org/10.1145/3512353.3512373

Harikumar, S. (2014). A Study on Educational Data Mining. *International Journal of Computer Trends and Technology*, *8*, 90–95. https://doi.org/10.14445/22312803/IJCTT-V8P117

Huang, Y., Dang, S., Elizabeth Richey, J., Chhabra, P., Thomas, D. R., Asher, M. W., Lobczowski, N. G., McLaughlin, E. A., Harackiewicz, J. M., Aleven, V., & Koedinger, K. R. (2023). Using latent variable models to make gaming-the-system detection robust to context variations. *User Modeling and User-Adapted Interaction*, *33*(5), 1211–1257. https://doi.org/10.1007/s11257-023-09362-1

Keleş, M. K. (2017). *AN OVERVIEW: THE IMPACT OF DATA MINING APPLICATIONS ON VARIOUS SECTORS*.

Khare, K., Lam, H., & Khare, A. (2018). Educational Data Mining (EDM): Researching Impact on Online Business Education. In *On the Line: Business Education in the Digital Age* (pp. 37–53). https://doi.org/10.1007/978-3-319-62776-2_3

Kondo, N., Okubo, M., & Hatanaka, T. (2017). Early Detection of At-Risk Students Using Machine Learning Based on LMS Log Data. *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, 198–201. https://doi.org/10.1109/IIAI-AAI.2017.51

Lesgold, A. (1992). Going form Intelligent Tutors to Tools for Learning. *Proceedings of the Second International Conference on Intelligent Tutoring Systems*, 39.

Marr, B. (2018, May 21). *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read*. Forbes. https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/

Meir, E., Wendel, D., Pope, D. S., Hsiao, L., Chen, D., & Kim, K. J. (2019). Are intermediate constraint question formats useful for evaluating student thinking and promoting learning in formative assessments? *Computers & Education*, *141*, 103606. https://doi.org/10.1016/j.compedu.2019.103606

Muldner, K., Burleson, W., Van De Sande, B., & VanLehn, K. (2011). An analysis of students' gaming behaviors in an intelligent tutoring system: Predictors and impacts. *User Modeling and User-Adapted Interaction*, *21*(1–2), 99–135. https://doi.org/10.1007/s11257-010-9086-0

Paquette, L. (2014). *Towards Understanding Expert Coding of Student Disengagement in Online Learning*.

Paquette, L., & Baker, R. S. (2019). Comparing machine learning to knowledge engineering for student behavior modeling: A case study in gaming the system. *Interactive Learning Environments*, *27*(5–6), 585–597. https://doi.org/10.1080/10494820.2019.1610450

Pathan, A. A., Hasan, M., Ahmed, Md. F., & Farid, D. Md. (2014). Educational data mining: A mining model for developing students' programming skills. *The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014)*, 1–5. https://doi.org/10.1109/SKIMA.2014.7083552

Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Romero, C., & Ventura, S. (2013). Data mining in education. *WIREs Data Mining and Knowledge Discovery*, *3*(1), 12–27. https://doi.org/10.1002/widm.1075

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. *2013 International Conference on Collaboration Technologies and Systems (CTS),* 42–47. https://doi.org/10.1109/CTS.2013.6567202

Shen, Y., Yin, X., Jiang, Y., Kong, L., Li, S., & Zeng, H. (2023). "Intellectual Companion": A Whole-Process Educational Big Data that Helps Improve Regional Education Quality. In *Case Studies of Information Technology Application in Education: Utilising the Internet, Big Data, Artificial Intelligence, and Cloud in Challenging Times* (pp. 137–145). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-9650-4_24

Shu, X., & Ye, Y. (2023). Knowledge Discovery: Methods from data mining and machine learning. *Social Science Research, 110,* 102817. https://doi.org/10.1016/j.ssresearch.2022.102817

Simanjuntak, M. P., Marpaung, N., Sinaga, L., & Siagian, E. (2022). *The use of moodle as a learning management system to improve student learning outcomes*. 140004. https://doi.org/10.1063/5.0114301

Sui, X., & Sui, Y. (2023). Research on the Application of Educational Big Data Analysis in Online Learning Behavior of Computer Basic Teaching. In S. Patnaik & F. Paas (Eds.), *Recent Trends in Educational Technology and Administration* (pp. 35–42). Springer International Publishing.

Vasic, D., Kundid, M., Pinjuh, A., & Seric, L. (2015). Predicting student's learning outcome from Learning management system logs. *2015 23rd International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, 210–214. https://doi.org/10.1109/SOFTCOM.2015.7314114

Vossen, G. (2014). Big data as the new enabler in business and other intelligence. *Vietnam Journal of Computer Science, 1*(1), 3–14. https://doi.org/10.1007/s40595-013-0001-6

What is XGBoost? (2023, November 25). NVIDIA Data Science Glossary. https://www.nvidia.com/en-us/glossary/data-science/xgboost/

Yau, J. Y.-K., & Ifenthaler, D. (2019). Learning Analytics: International Perspectives, Policies, and Contributions. In M. A. Peters & R. Heraud (Eds.), *Encyclopedia of Educational Innovation* (pp. 1–6). Springer. https://doi.org/10.1007/978-981-13-2262-4_123-1

Zhao, L., Zheng, Y., Zhao, J., Li, G., Compton, B. J., Zhang, R., Fang, F., Heyman, G. D., & Lee, K. (2023). Cheating among elementary school children: A machine learning approach. *Child Development, 94*(4), 922–940. https://doi.org/10.1111/cdev.13910