


A Machine Learning Approach for Quantifying Academic Misconduct

Almasi S. Maguya¹ 

¹Mzumbe University, Faculty of Science and Technology, Morogoro, Tanzania

Corresponding author : Almasi S. Maguya

E-mail : asmaguya@mzumbe.ac.tz

ABSTRACT

Evidence from the literature continues to reveal the problem of academic misconduct, particularly cheating, among university students. To deal with this problem effectively, a clear understanding of its magnitude is necessary for planning and resource allocation. This paper proposes a machine learning algorithm to quantify the magnitude of academic misconduct among undergraduate students. In this study, cluster analysis was employed with outlier detection and removal. The algorithm was trained on a dataset comprising 678 short texts. Results indicated that over 80% of students engage in the practice of academic misconduct. This shows that academic misconduct among undergraduate students poses a serious risk to the quality of graduates. This paper proposes a machine learning algorithm to quantify academic misconduct. The proposed algorithm is based on a modified k-means clustering algorithm that automatically detects and removes outliers. Universities can adopt the proposed method to combat the growing problem of academic misconduct among undergraduate students. The proposed approach for quantifying the magnitude of academic misconduct is more reliable and cost-effective than traditional (survey-based) methods.

Keywords: Unsupervised machine learning, machine learning, academic dishonesty, clustering, outlier removal

Submitted : 29.09.2024
Revision Requested : 06.11.2024
Last Revision Received : 11.11.2024
Accepted : 11.11.2024
Published Online : 10.12.2024



This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

1. INTRODUCTION

Academic dishonesty among students is a growing concern in academia. This is particularly true in higher education. While there are many forms of academic dishonesty, such as plagiarism, cheating—in its many forms—is considered one of the most widespread forms of academic dishonesty (Rettinger & Kramer, 2009; Anitha & Sundaram, 2021). Cheating, in the academic context, can be considered the act of gaining an unfair advantage by unauthorized copying or stealing someones' ideas. This unfair advantage can be gained in a number of ways, including copying answers from someone, taking an examination on behalf of another person, and bringing unauthorized materials into the examination room (Pino & Smith, 2003).

The ramifications of academic dishonesty are far-reaching and affect not only students but also educational institutions. Zhao et al. (2023), for example, found that cheating has a negative effect on learning among middle school children, whereby children who performed well as a result of being allowed to cheat in one test could not perform as well in a similar test when they were denied the opportunity to cheat. Arguably, this finding applies to students in higher learning institutions as well. A similar study by Malik et al. (2023) showed that students tend to perform better in online examinations than in physical ones simply because there is a good opportunity for cheating in online examinations.

Due to its importance in academia, the problem of cheating has received considerable attention. Specifically, many researchers have investigated the factors that motivate students to engage in cheating behavior. For example, in their study on students' perceptions and motivations for cheating, Waltzer and Dahl (2023) found that while most students did not realize their acts as cheating, those who did considered it acceptable, citing task feasibility and assignment goals as excuses.

Peer effects and procrastination from activities such as excessive television time were also found to contribute to cheating behavior (Pino & Smith, 2003; Fontaine, Frenette, & Hébert, 2020). In the former case, if a student embeds themselves in a company that engages in cheating behavior, he or she will become more susceptible to engaging in such behavior. Regarding the latter, if a student spends too much time on leisure activities, such as watching television and spending time on social media, the temptation to engage in cheating behavior to compensate for the wasted time will be high. Moreover, excessive pressure due to the desire to perform better in examinations and deliver homework and assignments on time has also been cited as motivating factors (Grenness, 2023).

Modern technologies are also an important motivating factor for cheating. For example, the proliferation of generative artificial intelligence (AI) tools such as ChatGPT and others has made it even simpler and tempting for students to engage in cheating behavior (Lancaster & Co tarlan, 2021; Lund et al., 2023; Uzun, 2023; Peres, Schreier, Schweidel, & Sorescu, 2023). This proliferation, coupled with easy access to mobile computing devices such as smartphones and tablets, raises new concerns to the problem. With the help of these tools, for example, students can generate essays in a matter of seconds. Similarly, the adoption of online learning by many academic institutions across the globe has contributed to this problem (Jenkins, Golding, Le Grand, Levi, & Pals, 2023; Newton & Essex, 2023). Malik et al. (2023) highlighted that cheating is generally more tempting and easier on online examinations than physical examinations. A comprehensive study by Noorbehbahani et al. (2022) provides a systematic review of the many facets of this problem.

The consequences of cheating are well documented in the literature. For example, cheating behavior practiced during academic life tends to propagate to professional life; that is, students who engage in cheating behavior during their studies tend to exhibit the same behavior in their professional life (Nonis & Swift, 2001; Carpenter, Harding, Finelli, & Passow, 2004). This observation means that the integrity of academic institutions is at stake because of students who are cheating. It is not surprising, therefore, that considerable research effort is being invested in combating this problem.

Over the years, different approaches have been used to combat the problem of academic dishonesty, and cheating in particular. Traditionally, the problem of cheating has been combated through the establishment of a number measures. These measures include a mix of punitive and non-punitive measures. There is no evidence on the suitability of these measures, but mixed opinions among researchers on their effectiveness indicate that neither is universally accepted (Simon et al., 2003; Gallant & Drinan, 2006).

Other approaches used to address this problem include determining the magnitude of the problem (quantifying it) and detecting its presence. The former has been mainly conducted through survey-based studies, whereas the use of modern technology, especially machine learning (ML), has been reported in the latter.

In this paper, we have proposed an ML method to quantify cheating behavior. Like other ML-based methods, the proposed method uses historical data to identify patterns. In this case, the patterns of interest are those related to cheating behavior. The data can be initially structured or unstructured. Specifically, this study addresses the question

“How can machine learning be used to quantify cheating behavior among university students?”. The proposed ML method to quantify cheating behavior does not suffer from the bias and sampling errors inherent in surveys.

2. TRADITIONAL METHODS FOR COMBATING CHEATING

To combat cheating effectively, it is important to determine the magnitude of the problem, that is, to estimate the proportion of students engaging in the behavior. This is an important problem that has been studied extensively for many years. The method of choice for many past studies has been surveys. Studies like these have been conducted at many universities across the world where a representative sample of the target student population is interrogated (see, for example, Awdry (2021); DiPaulo (2022a); Newton and Essex (2023)).

One limitation of these methods is the possibility of respondents understating their cheating behavior, which could lead to bias. This phenomenon can be attributed to the lack of awareness and understanding among students about the concept of academic dishonesty and its consequences (Locquiao & Ives, 2020; Chala, 2021). For this reason, most of these studies report the possibility of understating cheating behavior as one of their limitations (Chala, 2021; Orok et al., 2023). Another limitation of the survey methods, which has also been reported in previous studies, is sampling error. This error occurs when a study fails to select a sample that is representative of the population under study, making it impossible to generalize the results.

3. MACHINE LEARNING METHODS FOR COMBATING CHEATING

The past few years have witnessed a notable increase in the use of technology in combating cheating behaviors among students. In particular, the application of machine learning (ML) techniques in this area has received significant attention. This increasing adoption of ML techniques to combat cheating behavior can be attributed to the success of the technology in other domains as well as the abundance of academic data that has been accumulating for many years (Clare, Walker, & Hobson, 2017). These data are available in a variety of formats: unstructured (such as student essays and other works of writing) and structured (such as examination scores). Moreover, deficiencies in traditional assessment practices have also prompted the use of ML techniques in this area (Swiecki et al., 2022).

Consequently, many scholars have approached the cheating problem using ML techniques. Renzella, Cain, and Schneider (2022), for example, used ML for student identity in oral examinations. While oral assessment offers several advantages, such as fostering student interpersonal and communication skills, it is prone to impersonation. This observation is particularly true in large classes where no instructor may know the identity of many students. The student identity validation approach proposed in this study uses ML to determine whether two audio samples resemble speech from the same person. This approach can be used, for example, to facilitate oral discussions between students and tutors while providing the ability to detect contract cheating cases.

Some situations—such as conducting examinations in large classes—call for real-time detection of cheating behavior. However, detecting cheating behavior in real-time is a significant challenge. In such situations, setting, copying answers from fellow students, or unauthorized sources can easily go unnoticed. While combating this problem using traditional methods has proven challenging, it has been demonstrated that post-examination analysis of examination data using ML techniques can be used to detect cheating cases. Meng and Ma (2023), for example, proposed a statistically defensible ML approach for labeling true test cheaters using examination data. The proposed method identifies irregular statistical patterns in examination data. Similarly, many other scholars have employed ML in their studies that deal with detecting cheating behavior (for example, Khabbachi, Zouhair, Mahboub, and El-ghouch (2023); Akiful, Roy, Abdullah, Priota, and Onim (2022); Kaddoura and Gumaiei (2022); Bernius, Krusche, and Bruegge (2022); Kamalov, Sulieman, and Santandreu Calonge (2021)).

However, most existing studies focus on detecting cheating as part of a wider effort to address this problem. While detecting cheating behavior is important in addressing the problem, understanding the *magnitude* of the problem is a necessary step when planning for appropriate mitigation measures, such as resource allocation. For example, while the use of proctoring systems has shown some success, the decision to adopt such systems should be taken only when the magnitude of the problem is known in advance and justifies their adoption, as these systems tend to be costly (Nigam, Pasricha, Singh, & Churi, 2021).

For example, at Mzumbe University, where this study was conducted, final-year students in information technology (IT) programs are required to complete a substantial software development project in their final year of studies. Over the years, it has been learned that some students engage in cheating behavior while undertaking these projects. Cheating mainly occurs in the form of repeated work performed by another person in previous years (either as is or with minor modifications). Despite this understanding, it has been challenging to determine the magnitude of the problem due to

the sheer number of projects completed each year and the fact that these projects are supervised by different supervisors. Given this challenge, choosing an appropriate control mechanism to address the problem was challenging.

4. MATERIALS AND METHODS

The aim of this study was to estimate the extent of cheating (hereafter referred to as the cheating index or C-index) among final-year university students using unsupervised ML. Specifically, the study employed cluster analysis with outlier detection and removal. The underlying assumption is that similar titles (those repeated as is or with modifications) tend to cluster together, while new titles tend to segregate themselves from the rest as outliers. With this assumption in mind, the extent of cheating (as a proportion of the original number of titles) can be obtained by detecting and removing outliers from each cluster and then finding a percentage.

In this section, we describe the test data used in the study as well as details on how the data were processed and analyzed to answer the main research question posed above.

4.1. Test data

This study was conducted at Mzumbe University in Tanzania. The data used in this study pertain to final-year student project titles. These project titles are typically short texts with an average length of 35 characters. In total, 678 titles spanning a period of 5 years (2018–2022) were used. This number represents an average of 135 titles per year.

The titles were collected from final-year students' project reports. The students were enrolled in three ICT-based bachelor's programs: Information Technology and Systems (ITS), Information and Communication Technology with Management (ICT-M), and Information and Communication Technology with Business (ICT-B).

Students enrolled in these programs are required to complete a software development project during their final year of studies as a partial fulfillment of their program. This project is usually a web or mobile application completed in one semester. Examples of the project titles used in this study are as follows:

Tanzania online business license application system
Computerisation of transfer for public school teachers
Accident detection and notification mobile application

After extraction from the students' final year project reports, the titles were stored in a UTF-8 plain text file (one title per line) for subsequent processing and analysis.

4.2. Data Analysis and C-index Computation

To analyze the test data and compute the C-index, we employed a workflow that consisted of a sequence of steps: data preprocessing, feature extraction, determination of a suitable number of clusters to use in a clustering algorithm, cluster analysis with outlier removal, and finally computation of the C-index (see Fig. 1). Each step produces an output that serves as input to the next step. The workflow was implemented in Python.

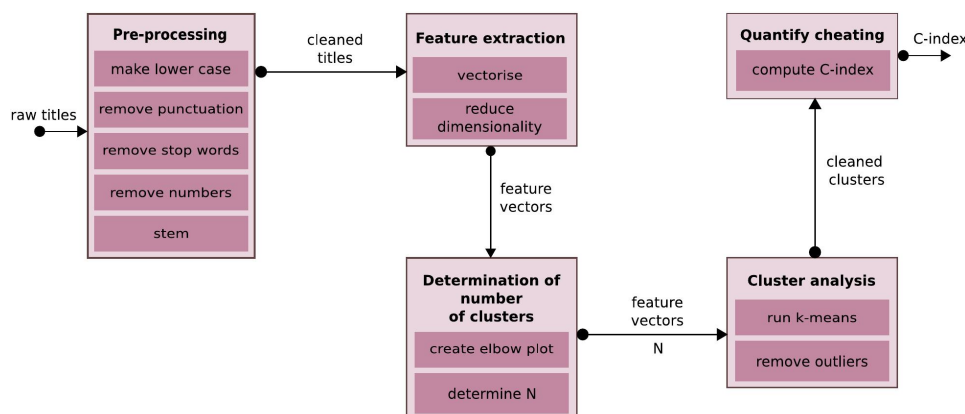


Figure 1. Proposed workflow for estimating C-index. The arrows indicate the sequencing of steps, and N represents the appropriate number of clusters for use in cluster analysis.

4.2.1. Data Pre-processing

The aim of the data pre-processing step was to prepare the data for the subsequent steps. First, each title in the data file was converted to lowercase letters to make the titles uniform and simplify other preprocessing steps. Second, all punctuation symbols were removed because they carry little or no useful information. Third, common terms, such as *system* and *app*, were removed for the same reason. The fourth and fifth steps removed numbers (digits) and normalized the words in each title by *stemming*, respectively. In addition, the titles were checked for misspellings and typos using a spellchecker. This step generated cleaned titles (documents) for the next step.

4.2.2. Feature Extraction

In this step, the titles are transformed into numerical feature vectors using the term frequency inverse document frequency (TF-IDF) algorithm. This text representation algorithm was selected because it is simple, considers the relative importance of individual terms, and takes care of stop words automatically. A drawback of the proposed model is that it typically produces extremely sparse feature vectors. The frequency of term t in document d is expressed as follows:

$$tf(t, d) = \log(f_{t,d})$$

where $f_{t,d}$ is the frequency of t in d . The inverse document frequency of term t in a set D of N documents is given by

$$idf(t, D) = \log\left(\frac{N}{n_t}\right)$$

where n_t is the number of documents in which t appears in. TF-IDF is computed as follows:

$$tf-idf(t, d, D) = tf(t, d) \cdot idf(t, D).$$

Both the term frequency and the inverse document frequency are usually scaled logarithmically to prevent bias in longer documents and in terms that appear more frequently relative to others (Aggarwal, 2022).

After the documents were vectorized, the dimensions of the resulting feature vectors were reduced by projecting them to a lower-dimensional feature space namely \mathbb{R}^3 . This projection was performed to simplify handling and facilitate data visualization. The resulting TF-IDF feature vectors were extremely sparse; thus, this projection was performed using the singular value decomposition (SVD) technique, which, unlike regular principal component analysis (PCA), can handle sparse matrices. This step produces feature vectors that are used to determine the suitable number of clusters to use in the cluster analysis step.

4.2.3. Determining the Number of Clusters

The main input to the k-means clustering algorithm is a predetermined optimal number of clusters to be used. In practice, this number can be determined using an elbow plot. Therefore, the number of clusters to be used was determined using the elbow plot shown in Fig. 2. Although the elbow appeared at four clusters in the figure, five clusters provided more plausible results. Thus, the number of clusters was five ($N = 5$).

4.2.4. Cluster Analysis with Outlier Removal

In this step, two actions were performed. First, the feature vectors were clustered using the k-means algorithm. Second, each cluster was subjected to an outlier removal process. The k-means algorithm is widely used in machine learning and pattern recognition tasks to partition (cluster) data due to its simplicity and ease of implementation. Mathematically, the algorithm aims to partition given data $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ into N groups (clusters) C_1, C_2, \dots, C_N such that the sum of the within-cluster squared distance between each point and the centroid of the cluster is minimized, that is

$$\arg \min_C \sum_{n=1}^N \sum_{\mathbf{x} \in C_n} \|\mathbf{x} - \mathbf{c}_n\|^2$$

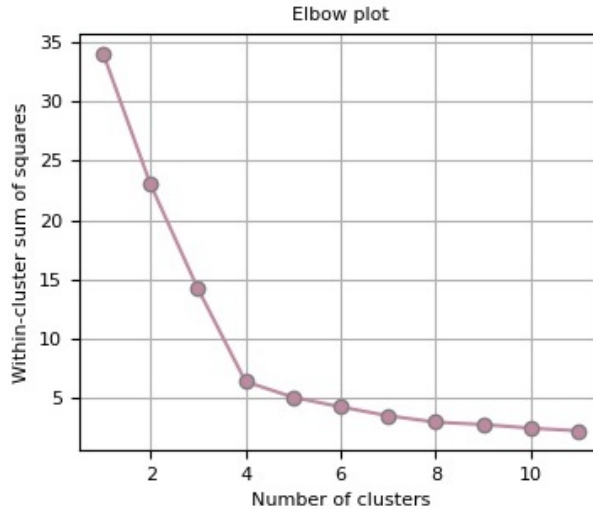


Figure 2. Elbow plot showing the optimal number of clusters for the k-means algorithm.

where \mathbf{c}_n is the centroid of the n^{th} cluster. We ran k-means using the value of N determined above. The clusters obtained were expected to contain some outliers; thus, each cluster was subjected to an outlier removal process. The steps involved in the outlier removal process are summarized in Fig. 3.

The algorithm examines each cluster independently. For each point in a cluster, its k nearest neighbors are determined. Next, the distance between the point and its furthest neighbor (d_{max}) and the median interpoint distance (d_{med}) between the point and all its neighbors are computed (see Fig. 4). If the magnitude of the difference between d_{max} and d_{med} (σ) exceeds a predetermined threshold θ , the point is designated an outlier and marked for removal by adding it to a set of outlier points. Finally, all points in the outlier set are removed from the cluster.

The points are not removed immediately because doing this would distort the structure of the cluster under consideration and would confuse the algorithm. The values of the parameters k and t were determined experimentally; we found that the values $k = 10$ and $\theta = 0.018$ served their purpose reasonably well.

```

input : Feature vectors  $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^3$ 
output:  $N$  clusters  $\{P_1, \dots, P_N\}$  with codebook  $C = \{\mathbf{c}_1, \dots, \mathbf{c}_N\}$  and outliers
          removed, where  $P_i = \{\mathbf{x} : \|\mathbf{x} - \mathbf{c}_i\| \leq \|\mathbf{x} - \mathbf{c}_j\| \forall j \neq i\}$ 

1 Initialise:  $N, \theta, k$ ;
2  $Outliers \leftarrow \emptyset$ ;
3  $\{P_1, \dots, P_N\} \leftarrow \text{k-means}(X, N)$ ;
4 for  $P \in \{P_1, \dots, P_N\}$  do
5   for  $\mathbf{x} \in P$  do
6      $kNN(\mathbf{x}, k) \leftarrow S_x \subseteq X$  s.t.  $|S_x| = k$  and  $\forall \mathbf{x}' \in X \setminus S_x$ ,
        $\|\mathbf{x} - \mathbf{x}'\| \geq \max_{\mathbf{x}'' \in S_x} \|\mathbf{x} - \mathbf{x}''\|$ ;
7      $d_{max} \leftarrow \max\{\|\mathbf{x} - \mathbf{q}_i\|\}, i = 1, \dots, k$  and  $\mathbf{q}_i \in S_x$ ;
8      $d_{med} \leftarrow \text{median}\{\|\mathbf{q}_i - \mathbf{q}_j\|\}, i = j = 1, \dots, k, i \neq j$  and  $\mathbf{q}_i, \mathbf{q}_j \in S_x$ ;
9      $\sigma = |d_{max} - d_{med}|$ ;
10    if  $\sigma > \theta$  then
11       $Outliers \leftarrow Outliers \cup \{\mathbf{x}\}$ ;
12    end
13  end
14   $P \leftarrow P \setminus Outliers$ ;
15 end

```

Figure 3. Proposed k-means clustering with outlier removal.

4.2.5. Computing C-index

After removing the outliers from each cluster, the cheating index is computed as the proportion of points remaining after removing the outliers:

$$C\text{-index} = \frac{\text{Number of remaining feature vectors}}{\text{Original number of feature vectors}} \times 100\%$$

5. RESULTS

The aim of this study was to use ML to estimate the extent of cheating among final-year university students. In this section we present the results of the analysis, both before and after the removal of the outliers. To realize this, we use visualization and various common clustering performance metrics.

The clustering results of the project titles prior to the removal of outliers are shown in Fig. 5. Similarly, Fig. 6 shows the clustering results of the titles after removing the outliers.

The silhouette score (a measure of cohesion among members of a cluster) of the clusters before the removal of outliers was 0.0082. After removing the outliers, the silhouette score improved to 0.7475, which reflects cluster homogeneity. Similarly, the within-cluster sum of squared errors (WCSSE) decreased significantly after the removal of outliers (Table 1). Using these results, the computed value of the cheating index was 81.3

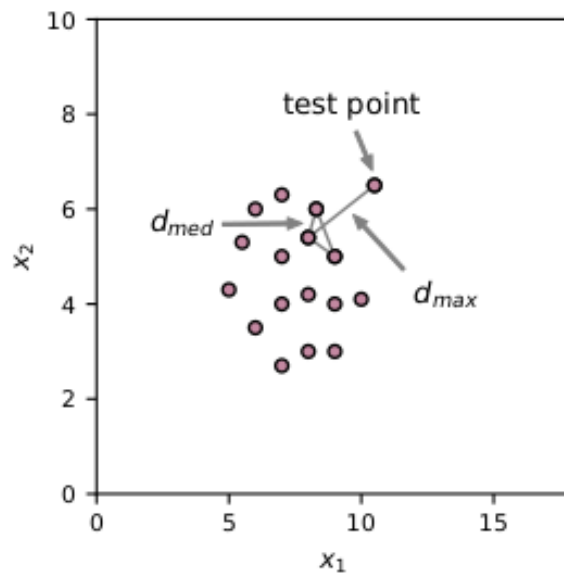


Figure 4. Detecting an outlier using distances between a test point and its nearest neighbors. Here, three neighboring points ($k = 3$) are used.

Table 1. Within-cluster sum of squared errors for five clusters before and after outlier removal

	C1	C2	C3	C4	C5
Before	0.7459	0.6669	1.3891	1.9139	0.3428
After	0.5838	0.2019	0.8301	0.0310	0.0000

6. DISCUSSION

Fig. 5 shows three clusters which are highly cohesive (C1, C2, and C3) in which most feature vectors are contained. In contrast, the remaining clusters (C4 and C5) were highly incoherent. The dispersed vectors in C4 and C5 indicate that project titles corresponding to those vectors have a higher degree of uniqueness than those in C1–C3. Points in these clusters represent titles that differ from the rest, which translates to the absence of cheating behavior.

When the points in C4 and C5 were removed (Fig. 6), the remaining clusters consisted of vectors corresponding to repeated titles (presence of high degree of cheating behavior). The dramatic drop in WCSSE (Table I) and the increase in the silhouette score from 0.0082 to 0.7475 indicate tight coherence among the clusters, which translates to a high degree of resemblance between the project titles represented by the vectors in those clusters. Here, the number of remaining clusters (3) indicates that over the five years that the test data span, students have been repeating project titles from three pools of titles represented by points in those clusters. The high silhouette score and small WCSSE values indicate that these titles were either reused as is or with only minor alterations (such as replacing words with synonyms or rearranging words in the title).

The high value for the cheating index (81.3%) was due to the fact that only a few vectors were identified as outliers and thus were removed (c.f. Fig. 5 (a)–(b) and Fig. 6 (a)–(b)). Because there is a one-to-one correspondence between the titles and students who worked on those titles,

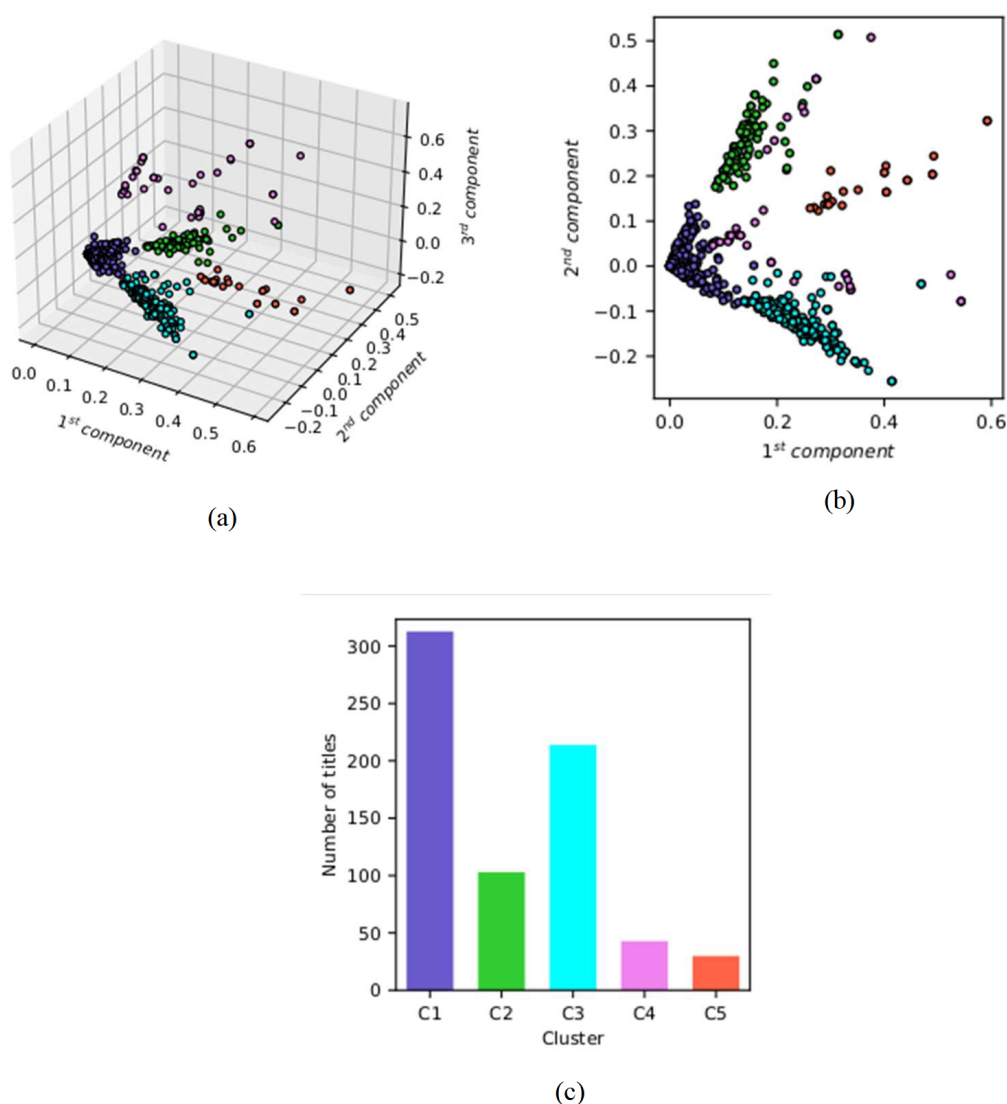


Figure 5. (a)–(b): Visualization of k-means clusters projected onto 3D and 2D feature spaces before removal of outliers; (c): Bar chart showing the color-coded number of points in each cluster.

this value can be regarded as the percentage of students who engaged in cheating behavior during the study period (2018–2022). This finding on the magnitude of cheating behavior is highly consistent with other studies that used traditional methods to determine the prevalence of academic misconduct among university students (Salehi & Gholampour, 2021; Wang & Xu, 2021; Awdry, 2021; DiPaulo, 2022b). Although both methods yield similar results, methods based on ML offer greater flexibility and are less laborious.

Although the present study does not address the reasons for this high rate of cheating, other studies may explain this observation (Anitha & Sundaram, 2021; Waltzer & Dahl, 2023). In addition, the temptation to reuse source code from previous years and not engage in the taxing endeavor of writing code from scratch can also be attributed to this finding.

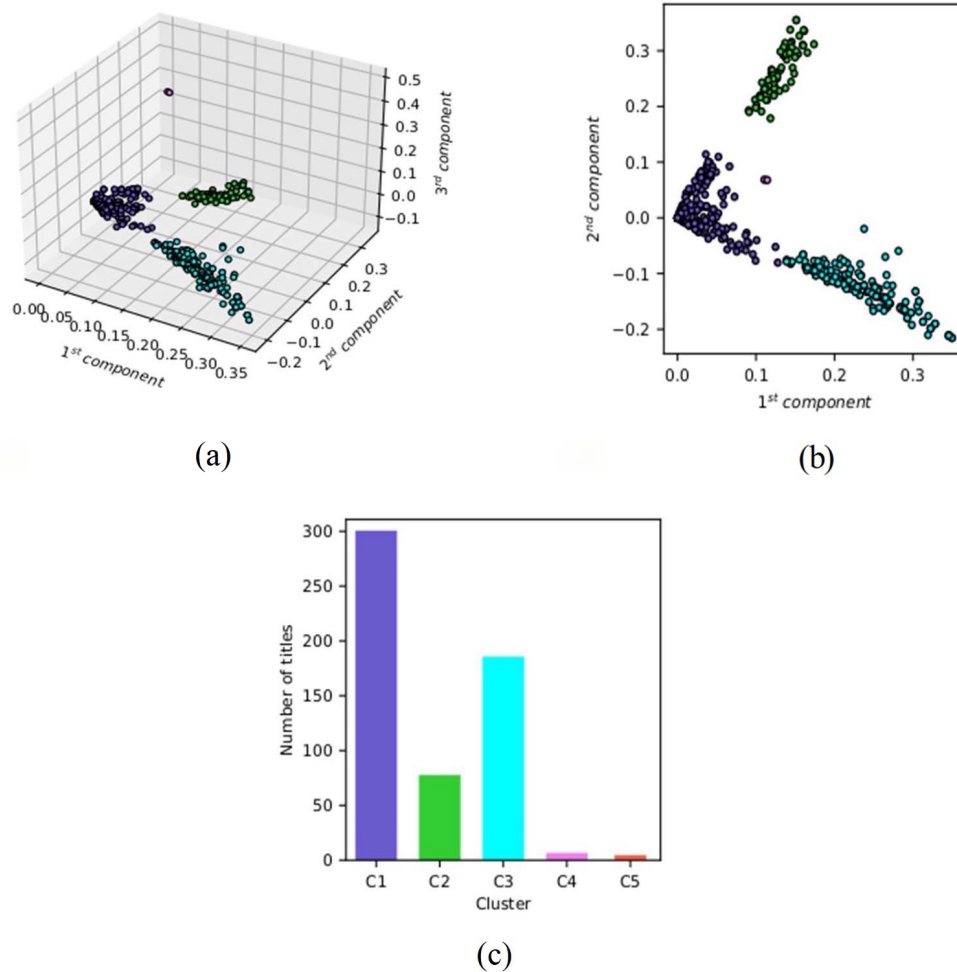


Figure 6. (a)–(b): Visualization of k-means clusters projected onto 3D and 2D feature spaces after removal of outliers; (c): Bar chart showing the color-coded number of vectors in each cluster.

The present study focused on quantifying the extent of cheating as opposed to most existing studies (which focus of detection and prevention of cheating). The results of this study can be used as a foundation for future studies. For example, some of these methods (e.g., Ranger, Schmidt, and Wolgast (2022); Chang and Chang (2023)) rely on supervised learning—particularly classification—which requires labeled data to work. The proposed method can be used to automatically create reliable training datasets for training classification models.

7. CONCLUSION

Understanding the extent of academic dishonesty is a key step in combating this problem; however, the literature is sparse on the use of machine learning (ML) techniques to quantify the magnitude of the problem. In this paper, an ML approach to quantify the magnitude of cheating behavior among final-year university students is proposed. The proposed method is based on cluster analysis coupled with outlier detection and removal. Results show that over 80% of the students engaged in cheating behavior during the period that the test data spans. This finding is consistent with existing studies that use traditional methods (surveys) to determine the extent of cheating behavior among university students. Because large amounts of academic data are readily available, the proposed method offers more flexibility and is more cost-effective than traditional methods. In addition, the proposed method can be used to create reliable datasets for studies that use ML to detect and prevent academic dishonesty.

Peer Review: Externally peer-reviewed.

Conflict of Interest: The author have no conflict of interest to declare.

Grant Support: The author declared that this study has received no financial support.

ORCID ID of the author

Almasi S. Maguya 0000-0002-1345-121X

REFERENCES

- Aggarwal, C. (2022). *Machine learning for text*. Springer International Publishing.
- Akiful, H. A., Roy, K., Abdullah, N., Priota, N. Z., & Onim, S. H. (2022). Performance Analysis of Machine Learning Models for Cheating Detection in Online Examinations. In *2022 25th international conference on computer and information technology (ICIT)* (pp. 342–347). doi: 10.1109/ICIT57492.2022.10055801
- Anitha, P., & Sundaram, S. (2021). Prevalence, types and reasons for academic dishonesty among college students. *Journal of Studies in Social Sciences and Humanities*, 7(1), 1–14.
- Awdry, R. (2021). Assignment outsourcing: Moving beyond contract cheating. *Assessment & Evaluation in Higher Education*, 46 (2), 220–235. doi: 10.1080/02602938.2020.1765311
- Bernius, J. P., Krusche, S., & Bruegge, B. (2022). Machine learning based feedback on textual student answers in large courses. *Computers and Education: Artificial Intelligence*, 3, 100081. doi: <https://doi.org/10.1016/j.caeai.2022.100081>
- Carpenter, D. D., Harding, T. S., Finelli, C. J., & Passow, H. J. (2004). Does academic dishonesty relate to unethical behavior in professional practice? An exploratory study. *Science and engineering ethics*, 10, 311–324.
- Chala, W. D. (2021). Perceived seriousness of academic cheating behaviors among undergraduate students: an Ethiopian experience. *International Journal for Educational Integrity*, 17(1), 2.
- Chang, S.-C., & Chang, K. L. (2023). Cheating Detection of Test Collusion: A Study on Machine Learning Techniques and Feature Representation. *Educational Measurement: Issues and Practice*, 42 (2), 62–73. doi: <https://doi.org/10.1111/emip.12538>
- Clare, J., Walker, S., & Hobson, J. (2017). Can we detect contract cheating using existing assessment data? Applying crime prevention theory to an academic integrity issue. *International Journal for Educational Integrity*, 13(1), 1–15.
- DiPaulo, D. (2022a). Do preservice teachers cheat in college, too? A quantitative study of academic integrity among preservice teachers. *International Journal for Educational Integrity*, 18(1), 2.
- DiPaulo, D. (2022b). Do preservice teachers cheat in college, too? A quantitative study of academic integrity among preservice teachers. *International Journal for Educational Integrity*, 18(1), 2.
- Fontaine, S., Frenette, E., & Hébert, M.-H. (2020). Exam cheating among Quebec’s preservice teachers: the influencing factors. *International Journal for Educational Integrity*, 16(1), 1–18.
- Gallant, T. B., & Drinan, P. (2006). Organizational theory and student cheating: Explanation, responses, and strategies. *The Journal of Higher Education*, 77 (5), 839–860.
- Grenness, T. (2023). "If You Don't Cheat, You Lose": An Explorative Study of Business Students' Perceptions of Cheating Behavior. *Scandinavian Journal of Educational Research*, 67(7), 1122–1136. doi: 10.1080/00313831.2022.2116479
- Jenkins, B. D., Golding, J. M., Le Grand, A. M., Levi, M. M., & Pals, A. M. (2023). When opportunity knocks: College students' cheating amid the COVID-19 pandemic. *Teaching of Psychology*, 50(4), 407–419.
- Kaddoura, S., & Gumaï, A. (2022). Towards effective and efficient online exam systems using deep learning-based cheating detection approach. *Intelligent Systems with Applications*, 16, 200153. doi: <https://doi.org/10.1016/j.iswa.2022.200153>
- Kamalov, F., Sulieman, H., & Santandreu Calonge, D. (2021). Machine learning based approach to exam cheating detection. *Plos one*, 16(8), e0254340.
- Khabbachi, I., Zouhair, A., Mahboub, A., & Elghouch, N. (2023). Reduce Cheating in e-Exams Using Machine Learning: State of the Art. In M. Lazaar, E. M. En-Naimi, A. Zouhair, M. Al Achhab, & O. Mahboub (Eds.), *Proceedings of the 6th international conference on big data and internet of things* (pp. 225–238). Springer International Publishing.
- Lancaster, T., & Cotarlan, C. (2021). Contract cheating among STEM students through file sharing websites: A COVID-19 pandemic perspective. *International Journal for Educational Integrity*, 17(1), 1–16.
- Locquiao, J., & Ives, B. (2020). First-year university students' knowledge of academic misconduct and the association between goals for attending university and receptiveness to intervention. *International Journal for Educational Integrity*, 16(1), 5.
- Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, 74(5), 570–581.
- Malik, A. A., Hassan, M., Rizwan, M., Mushtaque, I., Lak, T. A., & Hussain, M. (2023). Impact of academic cheating and perceived online learning effectiveness on academic performance during the COVID-19 pandemic among Pakistani students. *Frontiers in Psychology*, 14(2), 1124095.
- Meng, H., & Ma, Y. (2023). Machine Learning-Based Profiling in Test Cheating Detection. *Educational Measurement: Issues and Practice*, 42(1), 59–75. doi: <https://doi.org/10.1111/emip.12541>

- Newton, P. M., & Essex, K. (2023). How common is cheating in online exams and did it increase during the COVID-19 pandemic? A Systematic Review. *Journal of Academic Ethics*, 1–21.
- Nigam, A., Pasricha, R., Singh, T., & Churi, P. (2021). A systematic review on AI-based proctoring systems: Past, present and future. *Education and Information Technologies*, 26 (5), 6421–6445.
- Nonis, S., & Swift, C. O. (2001). An examination of the relationship between academic dishonesty and workplace dishonesty: A multicampus investigation. *Journal of Education for business*, 77(2), 69–77.
- Noorbehbahani, F., Mohammadi, A., & Aminazadeh, M. (2022). A systematic review of research on cheating in online exams from 2010 to 2021. *Education and Information Technologies*, 27(6), 8413–8460.
- Orok, E., Adeniyi, F., Williams, T., Dosunmu, O., Ikpe, F., Orakwe, C., & Kukoyi, O. (2023). Causes and mitigation of academic dishonesty among healthcare students in a Nigerian university. *International Journal for Educational Integrity*, 19(1), 13.
- Peres, R., Schreier, M., Schweidel, D., & Sorescu, A. (2023). On ChatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice. *International Journal of Research in Marketing*. doi: <https://doi.org/10.1016/j.ijresmar.2023.03.001>
- Pino, N. W., & Smith, W. L. (2003). College students and academic dishonesty. *College Student Journal*, 37(4), 490–500.
- Ranger, J., Schmidt, N., & Wolgast, A. (2022). Detecting Cheating in Large-Scale Assessment: The Transfer of Detectors to New Tests. *Educational and Psychological Measurement*, 0(0), 00131644221132723. doi: 10.1177/00131644221132723
- Renzella, J., Cain, A., & Schneider, J.-G. (2022). Verifying student identity in oral assessments with deep speaker. *Computers and Education: Artificial Intelligence*, 3 , 100044. doi:10.1016/j.caeai.2021.100044
- Rettinger, D., & Kramer, Y. (2009). Situational and Personal Causes of Student Cheating. *Research in Higher Education*, 50 , 293–313. doi: <https://doi.org/10.1007/s11162-008-9116-5>
- Salehi, M., & Gholampour, S. (2021). Cheating on exams: Investigating Reasons, Attitudes, and the Role of Demographic Variables. *SAGE Open*, 11(2), 21582440211004156. doi: 10.1177/21582440211004156
- Simon, C. A., Carr, J. R., McCullough, S. M., Morgan, S. J., Oleson, T., & Ressel, M. (2003). The other side of academic dishonesty: The relationship between faculty scepticism, gender and strategies for managing student academic dishonesty cases. *Assessment & Evaluation in Higher Education*, 28(2), 193–207.
- Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan S. . . Gañsević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, 3 , 100075. doi:10.1016/j.caeai.2022.100075
- Uzun, L. (2023). ChatGPT and academic integrity concerns: Detecting artificial intelligence generated content. *Language Education and Technology*, 3(1), 45–54.
- Waltzer, T., & Dahl, A. (2023). Why do students cheat? Perceptions, evaluations, and motivations. *Ethics & Behavior* , 33 (2), 130–150. doi: 10.1080/10508422.2022.2026775
- Wang, Y., & Xu, Z. (2021). Statistical Analysis for Contract Cheating in Chinese Universities. *Mathematics*, 9 (14). doi: 10.3390/math9141684
- Zhao, L., Peng, J., Dong, L. D., Compton, B. J., Zhong, Z., Li Y. . . Lee, K. (2023). Academic cheating interferes with learning among middle school students. *Journal of Experimental Child Psychology*, 226(2), 10556.

How cite this article

Maguya, A. S. (2024). A Machine Learning Approach for Quantifying Academic Misconduct. *Acta Infologica*, 8(2), 188-198. <https://doi.org/10.26650/acin.1557985>