Influence Factors on Academic Integrity revealed by Machine Learning Methods

Jennifer Landes[1], Sonja Köppl[1] and Meike Klettke[2]

[1]Hochschule Neu-Ulm, Faculty of Business and Economics
[2]Regensburg University, Faculty of Computer Science and Data Science

Abstract
Academic integrity in higher education can be influenced by individual or by institutional factors. Cheating behavior undermines the academic integrity of the learning environment and can have negative consequences for both the individual student and the academic community. To understand the factors that influence the cheating behavior of students, a quantitative study was conducted, specifically focusing on the types of exams and assignments that are most susceptible to cheating. The collected data has been analysed with Machine Learning methods and the results have been visualised. This survey is a part of a dissertation project and the survey results will be used for an eye-tracking experiment to measure cheating behavior of students. Long-term aim is to develop online exam methods which are not susceptible to certain cheating methods.

Keywords
Academic Cheating, Online Exam, Machine Learning, Clustering, Empirical Evaluation

1. Introduction

Academic integrity is a fundamental value in higher education, and it is essential for students to maintain ethical behavior and honesty in their academic pursuits. It can be influenced by individual student characteristics or by institutional factors [1]. To underpin the importance of academic integrity, Mccabe et al. [2] discuss several findings: integrity; cheating is prevalent and increasing; college is a critical time for ethical development; students face significant pressures to cheat; students are being taught that cheating is acceptable; and the fact, that today's college students will become tomorrow's leaders. However, there has been a growing concern regarding academic dishonesty among students, especially during the Covid-19 semesters. During these courses, which were mainly teached online, the suspicion grew, that many students took advantage to cheat. Therefore, there is a high necessary to look deeper in the factors, which influence cheating and in the cheating behavior in online exams. Academic misconduct among students has been a persistent concern for educational institutions. Cheating behavior undermines the academic integrity of the learning environment and can have as well as negative consequences for the individual student and for the academic community.

Structure of the article. This paper presents in the chapter 2 a short insight in two related works, which dealt with academic cheating. This will be followed by chapter 3, where the collected influence factors for academic cheating will be presented and the data collection process in the quantitative study will be presented. The following chapter 4, includes the data analysis, first the descriptive values of the study and the two clustering methods K-menas and DBscan. In the last chapter present a discussion and interpreation of both clustering results, a study outlook and the study limitations.

Aim of the work. In this paper, the issue of academic misconduct will be analysed. To understand the different factors that influence cheating behavior of students, a quantitative study at Hochschule Neu-Ulm was conducted, specifically focusing on the types of exams and assignments most susceptible to cheating. The collected data was first visualised and in a second step analysed with Machine Learning methods. The analysis was conducted by these steps: A descriptive analysis to reveal statistical information of the dataset, a selection of the dataset focusing on used cheating methods, a clustering of selection with k-Means and DBSCAN, a matching of clustering results to the complete dataset, a comparison of both clustering results and finally the interpretation of both results.

2. Related Work

A study by Janke et al. [3] examined factors regarding cheating behavior among students. The sudden shift to online teaching and exams during the COVID-19 pandemic led to a rise in cheating rates. The study proposed three hypotheses: the unproblematic digitization, the selective behavioral change, and the strong threat to integrity hypothesis. They conducted a national online survey in Germany in November/December 2020, reaching 3,005 students from all federal states and various types of academic institutions. After reducing, the survey included 1,608 students with diverse characteristics, including gender, age, and academic background. The results indicate that the majority of students had no prior experience with online exams, and most of them perceived online exams as less controllable and more prone to cheating than traditional exams. However, the study found no evidence of a general increase in academic dishonesty, although the use of unauthorized aids during online exams was more common than in traditional exams. Overall, the study suggests that the shift to online exams is not necessarily associated with a higher risk of academic dishonesty, but it requires careful monitoring and preventive measures to maintain academic integrity.

Mccabe et al. [2] conducted a large-scale study on cheating in academic institutions over a fifty-year period. They found that most college-bound students are exposed to cheating cultures during their high school years and that more than two-thirds of college students engaged in academic dishonesty in the previous year. Cheating is prevalent in graduate

and professional schools, with varying levels in different fields. The authors also found that there has been a shift in cheating-related attitudes and definitions among students, and both individual and contextual factors influence academic integrity and cheating behavior. They suggest that a strong ethical environment, fostered by factors such as peer disapproval and a well-run honor code, can play a key role in reducing cheating.

3. Data Collection

3.1. Methodology

Prior work shows, that the focus layed on measuring the cheating amount in online exams. So, therefore it is needed to examine the influence factors in detail. On the one hand, the used cheating method as well as the task method can be explored, which has a higher risk for cheating. As literature reveals, academic misconduct can be influenced by a variety of factors, which can be classified as extrinsic and intrinsic motivation. Intrinsic motivation refers to subjective and individual factors stemming from the student's personality, including self-motivation, self-efficacy, job opportunities, and adaptive comparative behavior. Extrinsic motivation refers to situational and organizational factors that affect the student from outside, such as living conditions, family circumstances, friends or classmates, learning mechanisms, examination form, course structure, instructor, and technical issues. Sanctions can also have an impact on academic misconduct. Figure 1 depicts the main influence factors, which are the basic concept for the survey design [4, 5].

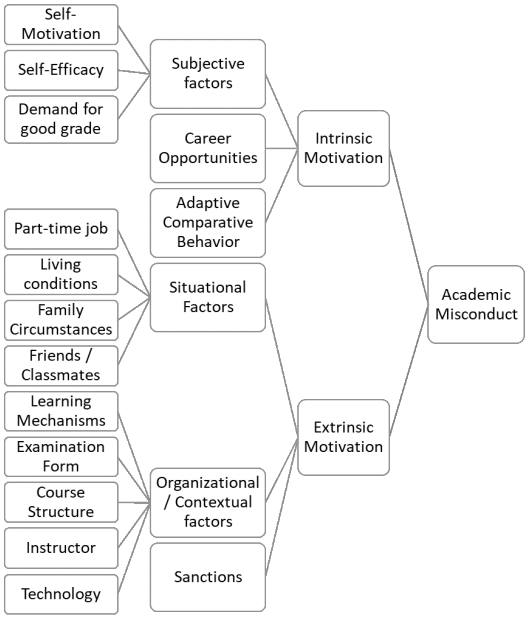


Figure 1: Collected Influence Factors based on [4]

3.2. Survey design

To examine individual as well as contextual factors, which influence the cheating behavior, a student survey was designed and conducted. The study involved the creation of an online survey with Lime Survey that captured information on both personal and academic activities of the students. The survey also captured the different cheating methods that students were aware of and when they would apply them. The survey was distributed to all students of the Hochschule Neu-Ulm through an email distribution list during the time of 06.12.2022 to 02.01.2023. Additionally, the survey was also presented in four lectures of industrial engineering by Professor Dr. Sonja Köppl to students from the first to fifth semester of their bachelor. The survey consisted of 42 questions divided into 5 groups, and it took approximately 12 minutes to complete. The groups were divided as follows:

- Part A: General questions about the course of study
- Part B: General questions about personal life
- Part C: Questions about exams
- Part D: Questions about cheating
- Part E: Demographic questions

Part A included questions about the course of study, semester, and grade point average. The next section examined student satisfaction with their studies and the university, personal motivation, and academic pressure. Part B comprised questions on lecture preparation, leisure activities, interests, part-time jobs, volunteer work, social media behavior, family obligations, and religiosity. Section C focused on online exam participation, equipment requirements, and comparisons between face-to-face and online exams in terms of comfort, fairness, and performance. Part D of the questionnaire dealt with questions about attitudes towards cheating, consequences of cheating, known cheating methods, the influence of the lecturer on cheating behavior, and the application of cheating methods in exams and task types. The final section E of the questionnaire collected demographic data such as age, gender, and living arrangements.

4. Data Analysis

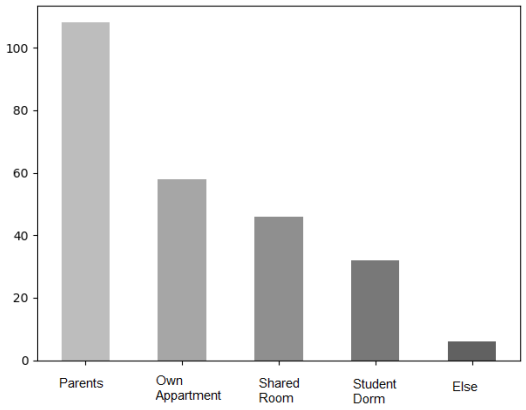
4.1. Descriptive Analysis

The analysed demographic data of the students included their course of study, semester, age, gender, and place of residence. Most participants came from the course Business Administration (19.21%), followed by Industrial Engineering (15.68%), Business Psychology (13.72%), Healthcare Management (13.33%), and Information Management and Corporate Communication (11.37%).

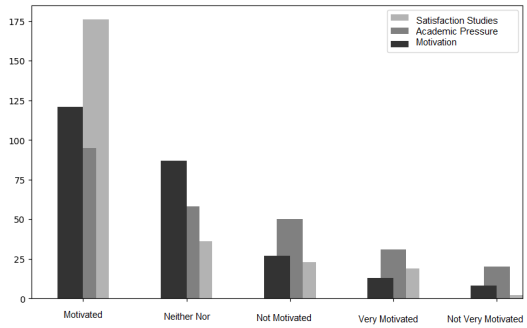
The students' average age was 23.32 years, with a range from 18 to 56 years old. Most participants were in their 4th semester and the average grade was 2.14. More females (57.58%) than males (33.46%) completed the survey, and most lived with their parents (43.2%).

Regarding satisfaction (compare figure 2), 77.73% were satisfied with their studies, and 49.4% with the university. 22.83% felt high pressure to perform, 37.4% felt some pressure, and 20.07% felt no pressure. 47.26% reported feeling motivated in their studies.

On average, participants spent 10.26 hours on hobbies and sports. Meeting friends was the most popular hobby (172 participants), followed by going to the gym (105), reading (90), and going to a bar or club (89). Playing poker (5), handball (4), and martial arts (4) were the least popular hobbies.



(a) Living Habits



(b) Satisfaction with Studies, Academic Pressure and Motivation

Figure 2: Descriptive Values

Tasks and Methods

An analysis of cheating methods was made, the results reveal, that the five most commonly used methods are cheating sheets, communication with others, preparation of material, use of multiple devices and translation programs. An additional analysis shows the occurrence of cheating per task type and per exam type for each cheating method (compare figure 3 and 4).

The digital exam forms are:

- Oral: An exam conducted through spoken communication between the examiner and the student on a video conference.
- Written: Students write their answers in a digital format and upload it to a portal or send it to the examiner.
- IT Pool: Students are all examined on computers in an IT pool and have limited access to programs and internet.
- Take Home Moodle Test: An exam administered through the Moodle learning management system, completed by students outside the classroom. The test has to be completed in a limited time like a real exam.
- Take Home Moodle Assignment: An assignment given to students through Moodle to be completed outside the classroom. The time space is not limited to an exam time duration.

And the task types are:

- Definition Task: A task that requires students to provide the meaning or definition of a concept or term.
- Transfer Task: A task that assesses the ability of students to apply knowledge or skills learned in one context to solve problems in a different context.
- Open Task: A task that allows students to explore different approaches and solutions without strict guidelines.
- Single Choice: A task where students choose one correct answer from a list of options.
- Multiple Choice: A type of task where students choose multiple correct answers from a list of options.
- Maths/Coding: A task that involves mathematical calculations or coding skills.

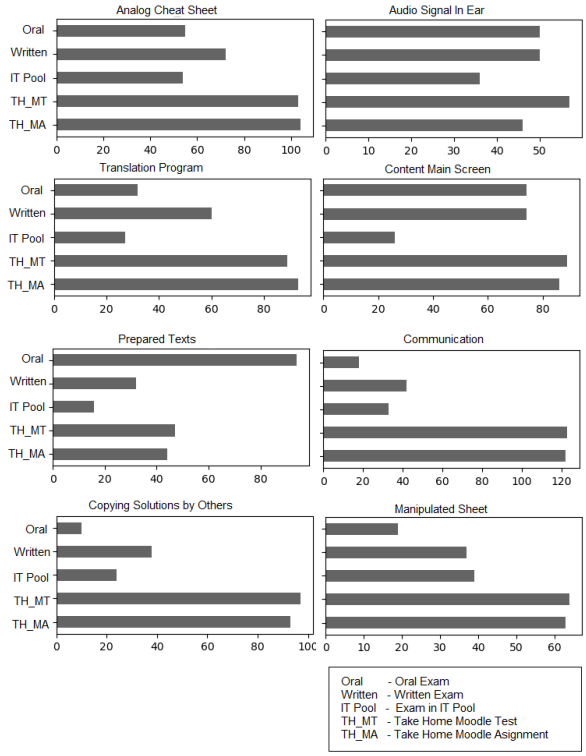


Figure 3: Occurrence of Cheating Methods per Exam Form

The first method is the use of analog cheat sheets. This method involves writing down definitions, math equations, or short answer responses on a piece of paper and referring to it during the exam. Students typically use this method in take-home Moodle tests where they are allowed to use materials during the exam. The second method is communication with other students during the exam. This type of cheating usually occurs in multiple-choice questions in take-home Moodle tests. Students collaborate with one another to share answers, which may give them an unfair advantage. The third method is the use of pre-written text materials. Students may read from prepared texts during oral exams or refer to notes during open-ended questions. The fourth method is the use of multiple devices during the exam. Students may use a second screen or another device to display

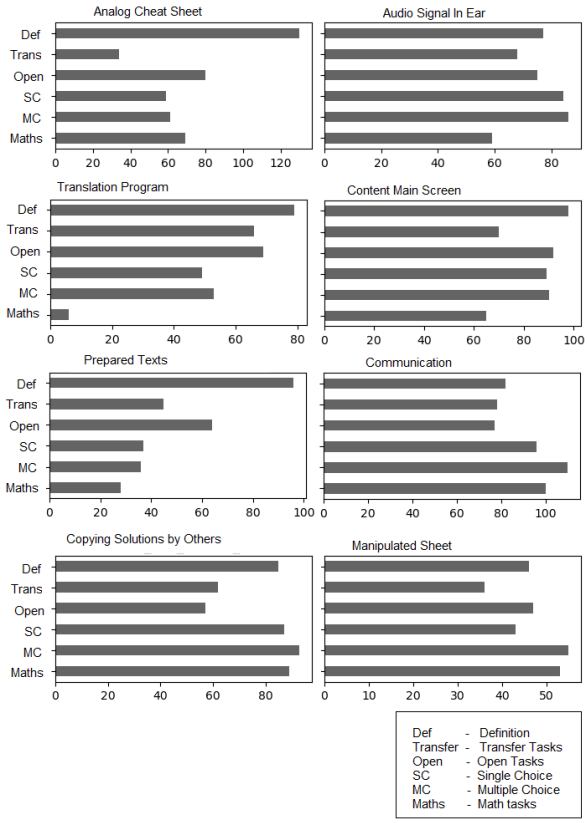


Figure 4: Occurrence of Cheating Methods per Task Type

notes, definitions, or other materials during the exam. This method is commonly used in take-home Moodle exams for short answer, multiple-choice, and open-ended questions. The fifth and final method is the use of translation programs during the exam. This type of cheating occurs in take-home exams, where students may use online translation programs to translate questions and provide answers in a different language. This method is commonly used in open-ended questions. The results strongly indicate, that digital exam formats have much higher rates in cheating potential.

4.2. Clustering

To gain insights in the collected data, two clustering methods were chosen to combine data and to identify similar groups of patterns in student behavior. Clustering is a method of unsupervised learning and involves the use of an unlabeled dataset consisting of a collection of examples $\{x_ß\}_{ß=1}^{N}$. Here, each $\{x_ß\}$ represents a feature vector, and the objective of an unsupervised learning algorithm is to develop a model that can process a feature vector x and transform it into either another vector or a value that can be employed to address a practical problem. The developed model assigns each feature vector in the dataset an identification number for its respective cluster [6]. K-means was chosen due to its widespread usage and reputation as a simple and efficient clustering algorithm. Its popularity makes it an ideal choice for establishing a benchmark and facilitating comparisons with other clustering methods. As a second method, DBSCAN was selected as a density-based algorithm, offering an alternative approach to centroid-based techniques like K-means. The aim was to investigate whether this density-based approach would yield notable distinctions in results and capture clusters that may be overlooked by K-means. In k-means, the clusters are named in numerical order, starting from 0. This naming convention is used to distinguish and identify individual clusters in the algorithm's results. In DBSCAN, the clusters are named based on the significance of cluster assignments. Outlier points, which do not belong to any cluster, are often labeled as -1. The first cluster is labeled as 0, and the second cluster is labeled as 1. This naming convention allows clear differentiation of outliers from actual clusters and provides a unique identification for each cluster.

4.2.1. k-Means

The well-known k-Means clustering algorithm [7] forms k clusters around centroids in a feature space whereby k is a predefined input parameter. In each step the distance of each data point to each centroid is calculated and the function

$$J = \sum_{i=1}^{k} \sum_{\mathbf{x_j} \in S_i} |\mathbf{x_j} - \boldsymbol{\mu_i}|^2$$

is optimized whereby $\mathbf{x_j}$ represents a data point and $\boldsymbol{\mu_i}$ represents a centroid of the cluster S_i. After each step, the cluster centers are updated until there are no further changes (convergence of the algorithm). With that, the algorithm forms k non-overlapping clusters. [8, 7, 9]

Data Preparation

The first step in the process involved importing an Excel spreadsheet using Pandas. The variables were converted from binary responses (Yes/No) to numerical values (0/1). The missing values for age, semester, and great point average (GPA) were replaced with their mean. Next, the missing values were filled using the "StandardScaler" method for data normalization. Then, one hot encoding was performed on the categorical variables (cheating attitude, major, gender, residence, motivation, performance pressure, technical equipment, preferred exam format, consequences of cheating, satisfaction with studies, and interest in technology) to convert them into numerical data.

Implementation

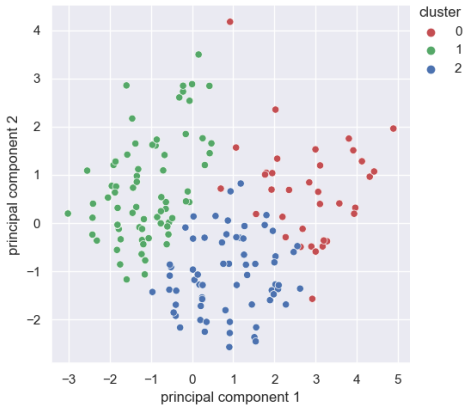
A dataframe object was created containing only the cheating methods: analog cheat sheet, manipulated exam materials, displaying content on main or second screen, displaying content on other devices, virtual camera, audio signals in ear, faking technical problems, reading prepared texts, translation programs, communicating with other students, and copying solutions from others.

Then, a principal component analysis was conducted to reduce the dimensionality of this dataset. The number of principal components was determined using the calculation of the "explained variance ratio". The analysis revealed that 6 principal components were needed to obtain sufficient information for clustering. In this case, there are six principal components: the first principal component explains 32.07% of the total variance, the second principal component 11.27%, the third principal component 8.46%, the fourth principal

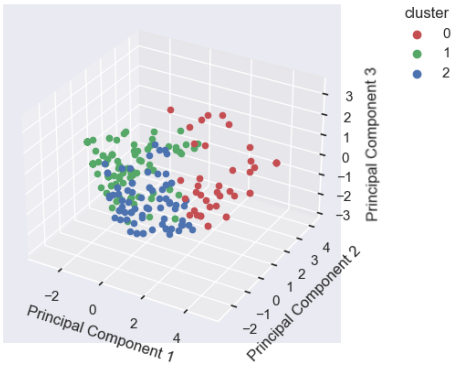
component 7.96%, the fifth principal component 6.53%, and the sixth principal component 5.75%.

Based on this data, clustering with k-Means was performed. The visualizations in 2D and 3D in figure 5 show three distinct clusters. During the clustering process, the value of the k parameter was manipulated to explore its effect on the resulting clusters. Various visualizations were explored using different numbers of clusters. Through an evaluation of the results, it was observed that the grouping exhibited the highest efficacy and meaningfulness when employing n=3 clusters. This decision was made by considering both the interpretability and distinctiveness of the resulting clusters. By opting for three clusters, the visual representation depicted clear boundaries and discernible patterns, facilitating a comprehensive understanding of the underlying structure present in the data.

```
# k-Means-Modelling
km_model = KMeans(n_clusters=3,
    random_state=42).fit(principalDf1)
sns.relplot(x = "principal component 1",
     y = "principal component 2", hue =
    "cluster", data = cluster_y)
```



(a) 2D Cluster Visualisation



(b) 3D Cluster Visualisation

Figure 5: Clustering Visualisations

Interpretation

To analyse the clusters, a column with the respective cluster was appended to the original table. After that, each cluster was filtered, and an individual evaluation was made based on the mean values for each category in each cluster.

Cluster 0 shows an increased tendency to cheat. In this cluster, almost all means of the cheating methods used are the highest. The cluster can be categorized as follows: The average GPA is the highest at 2.08 compared to the other two clusters, and the age of 22.98 indicates that this group is the youngest compared to the other clusters. The participants are predominantly male, have a high technical interest, live in their own apartment or a shared flat, are on average between the 3rd and 4th semester, and study Digital Enterprise Management, Game-Production Management, Information Management in Healthcare, Business Informatics, or Industrial Engineering. Furthermore, the evaluation shows that the average values for extensive and predominantly very time-consuming and active hobbies as well as time for voluntary and social media activities are the highest. The participants are also motivated and have high performance pressure, which would increase the tendency to cheat. Regarding the exam format, the participants perceive online exams as fairer and more pleasant than, for example, Cluster 1, which tends towards presence formats. Measures such as failing the exam or being excluded from the exam would deter cheating.

In a stark contrast to Cluster 0, it is evident that participants in Cluster 2 view cheating as unethical, are religious, and prefer presence formats. The participants predominantly study Business Administration, Digital Medicine and Care Management, Physician Assistant, or Business Psychology. The participants do not have a high technical interest, which could lead to a decrease in the incentive to use and experiment with technical cheating methods.

4.2.2. DBSCAN

DBSCAN is a density-based clustering algorithm. This algorithm requires the definition of two hyperparameters, E and n. E defines the radius of the neighborhood around each data point and is used to associate the data points to a cluster, n defines the minimum number of data points of each cluster. The clustering process can be defined as follows:

- Let X be the set of n data points, and let x_i be the i-th data point.
- The neighbourhood of x_i within the radius E is defined as: $N_E(x_i) = \{x_j | dist(x_i, x_j) \leq E\}$, where dist($x_i$, x_j) is the distance between x_i and x_j.
- A core point is defined as a data point that has at least n data points within its neighbourhood: core point : $x_i \in X \mid |N_E(x_i)| \geq n$.
- A border point is a data point that is not a core point but is within the neighbourhood of a core point: border point : $x_i \in X \mid \exists x_j \in X, x_j$ is a core point and $x_i \in N_E(x_j)$.
- A noise point is a data point that is neither a core point nor a border point [10, 11].

Implementation

The algorithms does the same preprocessing steps as the k-Means method. Then, the DBSCAN model is initialized with an value for E of 2.2 and minimum number of samples of 15 to form a dense region (in the source code the variable

eps is used for E and min_samples presents the value of n data points). The model is then applied to a standardized dataset, X-stand1, and the resulting cluster labels are printed.

Next, the DBSCAN algorithm is applied to a dataset, principalDf1, and the resulting clusters are visualized by a scatter plot with the principal components on the x and y axes, and the clusters indicated by different colors (see Figure 7).

The resulting cluster labels are converted into a Pandas Series and added to the original one-hot-encoded dataset. The observations are then grouped by their cluster numbers and the mean values of each column in each cluster are calculated and printed to the console. Determining the means of the points in DBSCAN allows for the representation of a cluster by providing a central point that can describe or visually represent the cluster.

```
# Init Model
dbscan = DBSCAN(eps=2.2, min_samples=15)
dbscan.fit(principalDf1)
# Visualize the clusters
plt.figure(figsize=(5, 5))
principalDf1 = principalDf1.rename(
    columns={"principal component 1":"
    PC1","principal component 2":"PC2"})
sns.scatterplot(x="PC1", y="PC2", data=
    principalDf1, hue=dbscan.labels_,
    palette="Set1")
```

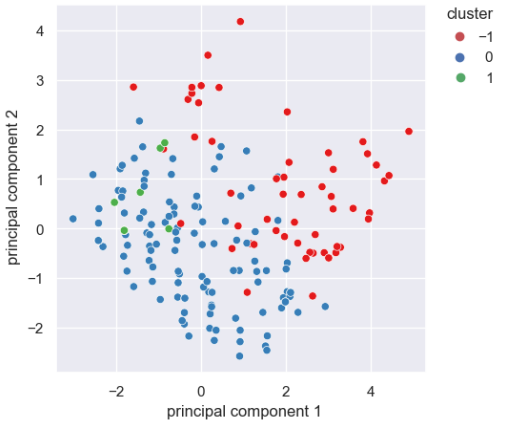


Figure 6: DBSCAN clustering

Interpretation

The majority of male participants has a high interest in technology and an increased likelihood of using cheating methods found in Cluster -1. The cheating methods employed by these participants include displaying content on the main screen, second screen, or other devices, using a virtual camera, receiving audio signals in the ear, pretending technical problems, reading prepared texts, using translation programs, communicating with other students, completely copying solutions, having someone else take the exam, cheating on take-home exams and submissions, cheating on pool exams, cheating on written Zoom exams.

Cluster -1 is characterized by a higher frequency of social media activities and hobbies such as football, tennis, dancing, yoga, fitness, martial arts, horse riding, jogging, chess, painting, cinema, and bars/clubs. Participants in this cluster report the highest number of volunteer hours, which is almost double the number reported by participants in other clusters. The attitude towards cheating in this cluster is generally permissive, with a tendency to cheat. Participants are primarily enrolled in Digital Enterprise Management, Information Management and Corporate Communications, Business Information Systems, and Industrial Engineering programs.

Compared to Cluster 0, significant differences are observed in the following categories: there is no academic pressure in Cluster -1, and there is no preference for any particular type of examination format. However, online examinations are perceived as fairer. The consequence of being expelled from university is a significant deterrent against cheating.

Cluster 0, on the other hand, comprises participants with a negative attitude towards cheating, and the mean values for cheating methods are not as high as those in Cluster -1. The hobbies of these participants include rock climbing, basketball, and poker playing. The represented study programs are Healthcare Business Administration and Data Science Management.

Cluster 1 consists of participants, mostly female and religious, with a low interest in technology and the lowest likelihood of employing cheating methods. The preferred cheating method in this cluster is using an analogue cheat sheet. The study programs represented in this cluster are Business Administration, Digital Medicine and Care Management, and Game Production and Management. Hobbies include yoga, pilates, handball, and socializing with friends. Participants in this cluster are aware of the consequences of cheating, such as being excluded from the exam, failing the exam, and having to give an oral explanation before the exam, which acts as a deterrent against cheating. They perceive in-person exams as fairer. These participants are highly motivated and feel significant academic pressure.

5. Discussion and Results

This paper aimed to identify and reveal factors that influence academic misconduct based on relevant literature by using clustering algorithms. A survey was developed to obtain necessary information through a quantitative study in multiple categories. 460 students participated in the survey, of which 263 completed the survey in its entirety. The results revealed that cheating behavior among students is influenced by various factors, including personal factors such as working time, family situation, academic pressure or organisational factors like the exam format. The analysis was carried out through a descriptive analysis and two different clustering methods k-Means and DBSCAN. The clustering process involved clustering a dataset that contained only cheating methods, and then assigning the resulting groups to all categories. The clusters generated by both methods exhibit significant similarities, but there are also some differences.

Similarities

- Both analyses identify clusters with participants who have a negative attitude towards cheating (Cluster 1 in DBSCAN, Cluster 2 in k-Means).

- Both methods identify a group with a tendency to cheat: Cluster 0 in k-Means is characterized by predominantly male participants with a high technical interest and a tendency to cheat, whereas Cluster -1 in DBSCAN is characterized by male participants with a high interest in technology and a permissive attitude towards cheating.
- Online exams are perceived as fairer compared to in-person exams by certain clusters (Cluster 0 in k-Means, and in both clusters in DBSCAN).
- Both analyses identify participants with high academic pressure and motivation in their studies.

Differences

- The two clustering analyses identify different numbers of clusters and their characteristics.
- The cheating methods used by participants in the different clusters vary across the two analyses.
- Cluster 2 in k-Means is characterized by younger male participants with a high technical interest who have a tendency to cheat, whereas there is no corresponding cluster in DBSCAN.
- The hobbies and study programs of participants in the different clusters differ between the two analyses.

The study and analysis provided insights into the factors that influence cheating behavior among students. The descriptive analysis revealed the prefered cheating method or exam format, time spent for their private hobbies, interests, living habits, working or volunteering hours, motivation or academic pressure of the students. The results suggest that the exam format, academic pressure and the perceived fairness are significant predictors of cheating behavior. Students who reported high levels of motivation and academic pressure were more likely to engage in cheating behavior.

For the cluster analysis in k-Means as well as DBSCAN the information was selected based on the cheating methods and then mapped to the complete data set. Both clustering results reveal tendencies, that a high technical interest and the online format influence a higher rate in cheating. Furthermore, the clustering identified in both methods a group of younger male students with a large number of hobbies and social media hours which use several cheating methods. It also showed, that the participants with a lower cheating tendency have a ethical attitude, prefer presence formats, are at a higher age and are aware of the consequences, when they get caught in exam cheating.

This study has implications for educators and academic institutions, highlighting the need to address academic integrity issues and to create a culture of academic honesty. Further research is necessary to explore how academic institutions can effectively address academic integrity issues and promote ethical behavior among students.

Future Work

This analysis is part of a PhD project, which identifies and evaluates the attitude and habits of students in regard to academic cheating. Further analysis with other machine learning algorithms are planned. In a next step, a second data collection at Regensburg University is planned to compare the data sets between boths institutions. Furthermore an eye tracking study will be conducted, to reveal patterns in eye moving while students are cheating.

Study Limitations

In this data analysis missing values got replaced by their mean values. In the further PhD thesis there will be used alternative strategies to deal with missing values, like to use zero values or use different case scenarios.

References

[1] I. Krumpal, R. Berger (Eds.), Devianz und Subkulturen: Theorien, Methoden und empirische Befunde, Kriminalität und Gesellschaft, Springer Fachmedien Wiesbaden, Wiesbaden, 2020. URL: http://link.springer.com/10.1007/978-3-658-27228-9.

[2] D. L. McCabe, K. D. Butterfield, L. K. Treviño, Cheating in college: why students do it and what educators can do about it, The Johns Hopkins University Press, Baltimore, 2012.

[3] S. Janke, S. C. Rudert, Petersen, T. M. Fritz, M. Daumiller, Cheating in the wake of COVID-19: How dangerous is ad-hoc online testing for academic integrity?, Computers and Education Open 2 (2021) 100055. URL: https://www.sciencedirect.com/science/article/pii/S2666557321000264.

[4] L. Hillebrecht, Einflussfaktoren des Studienerfolgs im Vollzeit-Studium, in: Studienerfolg von berufsbegleitend Studierenden, Springer Fachmedien Wiesbaden, Wiesbaden, 2019, pp. 77–124. URL: http://link.springer.com/10.1007/978-3-658-26164-1_3.

[5] C. Konegen-Grenier, Studierfähigkeit und Hochschulzugang., Kölner Texte &[und] Thesen. 61, Deutscher Instituts-Verl., Köln, 2002.

[6] P. Gupta, N. K. Sehgal, Machine Learning Concepts, in: P. Gupta, N. K. Sehgal (Eds.), Introduction to Machine Learning in the Cloud with Python: Concepts and Practices, Springer International Publishing, Cham, 2021, pp. 3–22. URL: https://doi.org/10.1007/978-3-030-71270-9_1.

[7] A. K. Jain, R. C. Dubes, Algorithms for clustering data, Prentice-Hall, Inc., 1988.

[8] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, A. Wu, An efficient k-means clustering algorithm: analysis and implementation, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 881–892. doi:10.1109/TPAMI.2002.1017616, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[9] K. P. Sinaga, M.-S. Yang, Unsupervised K-Means Clustering Algorithm, IEEE Access 8 (2020) 80716–80727. doi:10.1109/ACCESS.2020.2988796, conference Name: IEEE Access.

[10] J. Sander, M. Ester, H.-P. Kriegel, X. Xu, Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications, Data Mining and Knowledge Discovery 2 (1998) 169–194. URL: https://doi.org/10.1023/A:1009745219419.

[11] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise (1996).

Appendices

Full descriptive analysis: https://www.dropbox.com/s/nevkfbxueqakcwn/Deskriptive%20Auswertung.pdf?dl=0
https://www.dropbox.com/s/bnksjy0r1x6upt0/Umfrage_484229_Untersuchung_von_Einflussfaktoren_auf_die_Studienleistung_bei_online_Prfun
Full set of survey data: https://www.dropbox.com/s/8p6sz7gfg4rnykv/Umfragedaten.xlsx?dl=0

Category	Cluster 0	Cluster 1	Cluster 2
Academic semester	3,73529412	4,652173913	3,635658915
Grade point average	2,08294118	2,119021739	2,270229008
Alter	22,98	23,49173913	23,2940458
Cheating methods	Manipulated exam materials, Display content main, second screen, other devices, Virtual camera, Audio signal in ear, Pretending technical difficulties, Reading prepared texts, Translation programs, Communication with others, Copying complete solutions, Someone else takes exam	Analoger Spickzettel	
Exam format where cheating occurred	Pool Exam	Take Home Assignment, Take Home Test, Written Zoom Exam, Oral Zoom Exam	
Hobbies / Interest	Tennis, Dancing, Yoga, Gym, Pilates, Climbing, Martial arts, Meeting Friends, Painting, Bar/Club, Politics, Crafting, Riding	Swimming, Jogging, Chess, Poker, Reading, Cinema, Art, Music	Soccer, Baskeball, Handball, Religious, Fashion
Interest in Technology	very interested	Very interested	Somewhat interested
Not interested at all Working Hours	6,94117647	7,179347826	6,908396947
Volunteering	0,40625	0,350649351	0,309090909
Volunteering hours	3,08823529	1,043478261	1,374045802
Time Social Media	12	9,125	9,473282443
Family situation	Taking care of siblings	Childcare	Family member requiring care
Fields of study	Digital Enterprise Management, Game-Production Management, Information Management in Healthcare, Business Informatics, Industrial Engineering	Healthcare Business Administration, Data Science, Information Management Automotive, Information Management and Communication	Business Administration, Digital Medicine and Care Management, Physican Assistant, Business Psychology
Gender	Male	Female	Not specified
Living situation	Own apartment, shared flat	Other	with parents, Student dormitory
Motivation / Pressure / Satisfaction	Motivated, high pressure, high satisfaction on studies	very motivated, neither nor motivation, medium pressure, not satisfied on studies	very unmotivated, neutral pressure, no satisfaction on studies
Exam Format / Technical Equipment	Technical equipment missing or limited, participation in online exams possible, comfortable with online exam fairer with online exam	All necessary technical equipment available better with online exams, comfortable with in-person exam, fairer with in-person exam	Better with in-person exam
Consequences to deter cheating	Failing the exam, exclusion from exam	Expulsion, notation by proctor presented to exam committee	Verbal explanation of possible consequences before the exam, personal verbal warning during exam
Attitude	Attitude towards cheating (cheating = yes)	OK if friends cheat, neutral attitude, generally think cheating is ok	Believe cheating is wrong, if others do.

Table 1
Selected k-Means mean values

Category	-1	0	1	Category	-1	0	1
Semester	3,8	4,182352941	3,3	Business Administration	0,184615385	0,191860465	0,2
Grade Point Average	2,126769231	2,212093023	2,2225	Healthcare Business Administration	0,061538462	0,162790698	0,1
Age	22,97907692	23,58023256	22,232	Data Science Management	0	0,01744186	0
Analog cheat sheet	0,907692308	0,761627907	1	Digital Enterprise Management	0,046153846	0	0
Manipulated exam materials	0,507692308	0,093023256	0,6	Digital Medicine and Care Management	0	0,01744186	0,05
Display of content on main screen	0,646153846	0,238372093	0	Game Production and Management	0,030769231	0,046511628	0,05
Display of content on second screen	0,784615385	0,331395349	0	Information Management Automotive	0,015384615	0,023255814	0
Display of content on other devices	0,723076923	0,418604651	0,1	Healthcare Information Management	0,030769231	0,005813953	0,05
Virtual camera	0,415384615	0,023255814	0	Information Mgmt & Crp. Communications	0,169230769	0,098837209	0,05
Audio signal in ear	0,6	0,093023256	0	Physician Assistant	0,015384615	0,058139535	0,2
Pretending technical problems	0,6	0,23255814	0	Business Information Systems	0,046153846	0,023255814	0
Reading prepared texts	0,584615385	0,470930233	0	Industrial Engineering	0,184615385	0,151162791	0,1
Translation programs	0,630769231	0,313953488	0	Business Psychology	0,138461538	0,13372093	0,15
Communication with other students	0,923076923	0,529069767	0,75	Female gender	0,6	0,558139535	0,65
Completely copying solutions	0,769230769	0,191860465	0,05	Male gender	0,369230769	0,325581395	0,3
Having someone else take the exam	0,6	0,093023256	0,1	Not specified gender	0,030769231	0,11627907	0,05
Cheating on Take Home Exam	0,338461538	0,203488372	0,15	Living with parents	0,384615385	0,430232558	0,45
Cheating on Take Home Submission	0,338461538	0,203488372	0,25	Own apartment	0,215384615	0,23255814	0,2
Cheating on a pool exam	0,061538462	0,040697674	0	Other living situation	0,030769231	0,023255814	0
Cheating on a written Zoom exam	0,092307692	0,069767442	0	Living in a student dorm	0,123076923	0,127906977	0,1
Cheating on an oral Zoom exam	0,061538462	0,098837209	0,05	Living in a shared apartment	0,230769231	0,151162791	0,25
Football	0,138461538	0,122093023	0,1	No information on living situation	0,015384615	0,034883721	0
Basketball	0,030769231	0,046511628	0	Motivation: Motivated	0,461538462	0,459302326	0,6
Tennis	0,076923077	0,052325581	0	Motivation: Highly motivated	0,061538462	0,040697674	0,1
Dancing	0,184615385	0,087209302	0,05	Motivation: Highly unmotivated	0,015384615	0,040697674	0
Yoga	0,123076923	0,058139535	0	Motivation: Unmotivated	0,076923077	0,110465116	0,15
Swimming	0,046153846	0,093023256	0,15	Motivation: Neither	0,384615385	0,343023256	0,15
Fitness studio	0,446153846	0,395348837	0,4	Disagree with academic pressure	0,123076923	0,122093023	0,1
Pilates	0,092307692	0,046511628	0,15	Fully agree with academic pressure	0,246153846	0,209302326	0,3
Climbing	0,061538462	0,063953488	0,05	Agree with academic pressure	0,307692308	0,401162791	0,3
Martial arts	0,046153846	0,005813953	0	Completely disagree with pressure	0,153846154	0,052325581	0,05
Horse riding	0,061538462	0,034883721	0,05	Agreement: Neither	0,169230769	0,215116279	0,25
Handball	0	0,011627907	0,1	All necessary technical equipment	0,661538462	0,709302326	0,8
Jogging/Running	0,215384615	0,203488372	0,15	Technical equipment: Lack, can participate	0,246153846	0,255813953	0,15
Chess	0,061538462	0,046511628	0	Technical equipment: lack, partially participate	0,030769231	0,011627907	0
Poker	0	0,029069767	0	Exam form: Both are equally good	0,4	0,319767442	0,35
Meeting friends	0,723076923	0,63372093	0,8	Exam form: No comparison yet	0,215384615	0,238372093	0,3
Reading	0,430769231	0,308139535	0,45	Exam form: Online is better	0,169230769	0,180232558	0,15
Painting	0,169230769	0,098837209	0,1	Exam form: In-person is better	0,153846154	0,209302326	0,1
Cinema	0,230769231	0,156976744	0,15	Exam comfort: Both equally comfortable	0,338461538	0,168604651	0,25
Bar/Club	0,476923077	0,284883721	0,45	Exam comfort: No comparison yet	0,215384615	0,156976744	0,15
Working hours	6,969230769	7,218023256	5,35	Exam comfort: Online is more comfortable	0,215384615	0,244186047	0,05
Volunteering	0,413793103	0,317241379	0,25	Exam comfort: In-person more comfortable	0,215384615	0,395348837	0,5
Volunteering hours	2,076923077	1,325581395	0,9	Fairness of exam: Both are equally fair	0,261538462	0,25	0,3
Volunteering related to studies	0,153846154	0,037974684	0	Fairness of exam: No comparison yet	0,184615385	0,174418605	0,15
Religious activities	0,410714286	0,414285714	0,55556	Fairness of exam: Online is fairer	0,092307692	0,063953488	0
Social media time	11,16923077	9,14244186	9,5	Fairness of exam: In-person is fairer	0,430769231	0,470930233	0,5
Politics	0,384615385	0,215116279	0,1	Consequences: Failing the exam	0,323076923	0,220930233	0,45
Crafting	0,215384615	0,098837209	0,05	Consequences: excluded from the exam	0,153846154	0,13372093	0,2
Handiwork	0,2	0,215116279	0,2	Consequences: Being expelled from university	0,338461538	0,308139535	0,15
Art	0,230769231	0,186046512	0,2	Consequences: Oral explanation before exam	0,046153846	0,087209302	0,1
Fashion	0,353846154	0,325581395	0,35	Consequences: Oral warning during exam	0,092307692	0,122093023	0
Music	0,538461538	0,534883721	0,75	Consequences held: Note to exam committee	0,015384615	0,034883721	0,05
Taking care of siblings	0,061538462	0,034883721	0,05	Satisfaction with studies: Do not agree	0,061538462	0,069767442	0,15
Caring for a family member	0,030769231	0,058139535	0,05	Satisfaction with studies: Fully agree	0,123076923	0,081395349	0,05
Childcare	0,046153846	0,040697674	0	Satisfaction with studies: Agree	0,661538462	0,691860465	0,7
Attitude: Cheating is okay with friend	0,015384615	0,011627907	0	Satisfaction with studies: Do not agree at all	0	0,011627907	0
Attitude: Neutral towards cheating	0,6	0,593023256	0,75	Satisfaction with studies: Neither agree/disagree	0,153846154	0,145348837	0,1
Attitude: Other opinion	0,107692308	0,046511628	0,05	Interest in technology: Slightly interested	0,569230769	0,563953488	0,5
Attitude: Cheating is morally wrong	0,138461538	0,23255814	0,15	Interest in technology: Not interested at all	0,092307692	0,162790698	0,45
Attitude: Cheating is okay	0,138461538	0,11627907	0,05	Interest in technology: Very interested	0,307692308	0,238372093	0

Figure 7: Selected DBSCAN mean values