Computer Science
&
Data Science

Capstone Report - Fall 2022

Geo-Based Prediction of Academic Performance

Sida (Star) Chen,
Mengshuo (Scott) Ye

supervised by
Ratan Dey & Promethee Spathis

Preface

This study was carried out by a team of two senior students majoring in Computer Science and Data Science. Acknowledging the prevailing concerns about the quality of online learning since the COVID-19 and desperate need for effective assessment of student's academic performance, we aim to derive prediction models for academic outcomes based on geolocations. The prediction models constructed in our study will aid educators in predicting the academic performance of students with single rules, which enables pre-intervention for preventing academic failures.

Acknowledgements

We would like to express our very great appreciation to Prof. Promethee Spathis for his valuable guidance and constructive suggestions during the planning and development of this study as well as the dataset he collected that we relied on for our study. We would also like to thank Prof. Ratan Dey for being willing to collaborate with us and provide a solid start for this research project. Finally, We wish to thank Prof. Li Guo and Prof. Olivier Marin for mentoring our capstone course and for all the effective communications throughout the semester that enabled us to proceed smoothly with our project.

Abstract

This study aims to investigate the significance of students' geolocation and to predict students' academic performance based on changes in geolocation and study behaviors. Although the online Learning Management System (LMS) guarantees the continuation of education during uncertain times, such as global pandemics, war, and local emergencies, it is difficult for instructors to ensure teaching outcomes without accurate assessments of students' performance in online settings. Meanwhile, predicting students' performance is challenging due to the ever-increasing data on LMS, while some crucial factors, such as geolocation, have been largely ignored. In our study, we attempt to use two different approaches, one of which is Decision Tree, and the other is Clustering Analysis, Markov Chain, and Association Rule Mining, to discover the significance of geolocation. The latter approach not only shows how students' academic performance is affected by their geolocation and study behaviors but also generates many heuristic rules that can be used directly by instructors to conduct necessary pre-intervention on possible academic failures of the students.

Keywords

Geolocation, Learning Analytics, Clustering, Decision Tree, Markov Chain, Association Rule Mining

Contents

1 Introduction

The outbreak of COVID-19 has had an impact on every aspect of life, including education. Global educational activities have been severely disrupted, as the widespread closures of schools have limited students' access to receive in-person instruction. In response to the stay-at-home orders, instructors worldwide have shifted the teaching mode to online. With the help of Learning Management Systems (LMS) such as Canvas, Moodle, and Blackboard, students are able to complete coursework regardless of location or time availability.

While the LMS guarantees the continuation of education during the pandemic, feedback from educators shows concerns about this new learning format. Compared to traditional face-to-face instruction, online learning is susceptible to drawbacks and challenges, such as heavy workload, human and pet intrusion, course incompatibility, and absence of assessment [1], which negatively impact students' learning outcomes. Moreover, the instructor's quality is the most prominent factor that determines students' satisfaction with online classes, which directly contributes to their academic performance [2]. This requires instructors to have an accurate assessment of students' performance in order to deliver the course materials properly and make meaningful interventions to ensure the teaching outcomes.

Every time a student utilizes the LMS, his or her actions will be recorded in log files. With the accumulation of educational data available on the LMS platforms, Educational Data Mining (EDM) and Learning Analytics (LA) can be applied to exploit educational data and generate crucial assessments of student performance [3]. As seen in [4], EDM refers to the application of data mining techniques on educational datasets to address important education questions, and LA stands for the collection, analysis, measurement, and visualization of educational data to optimize learning and teaching environments. EDM and LA are interdisciplinary areas with a common interest in enhancing educational practice using data-intensive approaches, including causal mining, prediction, clustering, knowledge tracing, etc. The Effective deployment of interventions, based on the result from EDM/LA analysis, has a noticeable impact on students' success and academic performance outcomes [5].

Many factors have been identified as central to students' academic performance in educational studies. The student's geographical location (geolocation) is one important area for educational research [6]. However, there are few studies that investigate geolocation as a factor of academic outcomes, especially in the context of students' performance prediction for online learning [7].

Under the mode of online teaching, there are significant disparities in geolocation between each individual student, and the datasets available on LMS have made further explorations of geolocation on students' performance possible.

To address the research need, we investigated the significance of students' geolocation, a representative feature of the physical educational environment, and predicted students' academic performance based on students' geolocation and study behaviors. We conducted EDM/LA on data from one popular LMS, Moodle for prediction. We used students' IP addresses recorded on Moodle to acquire their geolocation (IP geolocation). Multiple machine learning approaches such as K-Means Clustering and Decision Tree classifiers were tested. Moreover, we derived a set of prediction rules from Markov Chain (MC) analysis and Association Rule Mining (ARM). These prediction rules will allow instructors to assess students' academic performance accurately and implement a list of possible pre-intervention measures that are beneficial for the students. In particular, the following research questions guide our analysis of student performance prediction:

RQ1. What is the impact of stay-at-home orders on students' academic performance?

RQ2. Can academic performance be predicted from students' geolocations and study behaviors?

2 Related Work

Many studies have been done in the past on the prediction of students' academic performance, which is a significant subject in EDM/LA. A systematic review in [7] has shown the most popular factors studied in recent works. Students' previous grades and class performance, e-learning activity, demographics, and social information are the most common and widely used factors for prediction, while other factors are less frequently studied. Various machine learning methods and techniques can be applied for different factors to improve the accuracy of the prediction [8].

In [9], Yu et al. adopt a straightforward approach for their learning analytics model. Multiple linear regression analysis was conducted for predicting students' learning outcomes (i.e. final grades) based on their behavior datasets on the Moodle-based LMS. The derived model accounts for 33.5% of the variance in the final grade, and four factors are confirmed to be significantly correlated with academic performance.

Kokoç et al. [10] study the temporal aspect of online assignment submission behavior and its relationship with students' academic performance. Cluster analysis and Markov Chains are conducted to observe the transition in behaviors of online assignment submission among several

groups of students with similar behaviors. Then, association rule mining is used to model predictive rules between the students' behaviors and academic performance. The study has built several predictive rules with high confidence that can prevent students from possible academic failures.

Hasan et al. [11] combine both students' academic information and their activities in LMS into a classification model to predict their academic performance. During data preprocessing, integer data are masked into four scales to improve prediction accuracy. Several classifiers are used to predict academic performance. The result shows that Random Forest, Naïve Bayes, and SMO have the best prediction accuracy.

Lu et al. [12] concentrate on students' online behaviors in video-viewing, out-of-class practice, and scores. Principal components are selected for different datasets and are trained in a linear regression model. The research shows that students' final academic performance can be predicted by the sixth week of the semester, and the dataset provided by blended learning contributes to better prediction results than the traditional learning dataset.

Imran et al. [3] focus on a variety of factors including student grades, demographic, social, and school-related features. A supervised learning decision tree model is built with three classifiers, namely J48, NNge, and MLP, for prediction purposes. The study reveals that J48 has the best performance with an accuracy of 95.78%.

Jain et al. [13] work on a dataset consisting of factors such as students' parent education and social information. Several different machine learning algorithms, including Decision Tree, Random Forest, Gradient Boosting, and Extreme Gradient Boosting, are adopted in the proposed models. Parameter tuning and attribute selection are conducted to improve prediction accuracy. The research reaches a high accuracy of 95% with the Random Forest classifier.

The works mentioned above have studied many factors and built proper machine learning models to predict students' academic performance. However, no previous research is found to use geolocation as the main factor in predicting academic performance. Nevertheless, geolocation is studied for other purposes in EDM/LA. Komosny et al. [14] have derived students' IP geolocation from IP addresses and applied IP geolocation information to calculate a score that reflects students cheating risk. IP geolocation, with an inevitable error margin, is filtrated and processed with its confidence area to improve its accuracy. The study also shows information that can be calculated by students' IP geolocation, such as distance and travel speed. Luo et al. [15] use Google Analytics to visualize multiple factors, including geolocation. Meanwhile, fundamental

studies are conducted with respect to IP geolocation. The two common approaches to acquiring geographical locations from IP addresses are IP geolocation databases [6] and active geolocation measurement services [16]. Research such as [17] also presents methods to improve the accuracy of IP geolocation.

3 Solution and Results

3.1 Dataset Observation

The dataset was collected by Professor Promethee Spathis from his course at Sorbonne Université in Paris in the Fall 2020 semester with a total of 249 students. All tasks, including assignments, labs, and exams, are taken through Moodle, and the grade and time of every task are recorded for each student. Every time a student accesses Moodle, his/her activities will be recorded in a log file that contains the task name, event name, time, and IP address. There are over ten event names, including "Quiz attempt viewed", "Course module viewed", "Quiz attempt started", "Quiz attempt abandoned", etc., that reflect the students' specific actions on the task.

During the semester, due to the stay-at-home orders, two lockdowns occurred respectively on October 5, when the number of in-person students was limited to half of the campuses' capacity, and on October 29, when the university campuses decided to be fully closed. The two lockdowns made this semester a special case since students' studying patterns, schoolwork, and geographical locations were greatly affected. Students had to travel more frequently between campuses and their hometowns than in other semesters. The frequent change in geolocation could influence students' behaviors and academic performance. Also, the grade for assignments and labs became less reliable as it was in a non-lockdown semester, especially for assignment and lab grades around the two lockdown periods. Hence, we believe studying geolocation is necessary since it can fix the gap caused by the unreliability of those common factors, such as assignment grades, and will be helpful to predict students' academic performance.

3.2 Data Processing

3.2.1 IP Address to Geolocation

IP geolocation databases [6] and active geolocation measurement services [16] are two common methods for acquiring geographical locations from IP addresses. For simplicity, we adopted geolocation databases for the geolocation conversion. Geolocation DB is a free-to-use database.

However, it does not provide much transparency regarding the accuracy of its data. As a result, using this database introduces an unknown level of inaccuracy to geolocation data. As an improvement, using Abstract API, we created a data frame that contains the corresponding information of each IP address, including longitude and latitude, city and country, and the use of VPN. There are 6646 unique IP addresses in total.

From the geolocation data that we retrieved from students' IP addresses recorded by their online activities, we observed that geolocation can reflect much information on students' learning status.

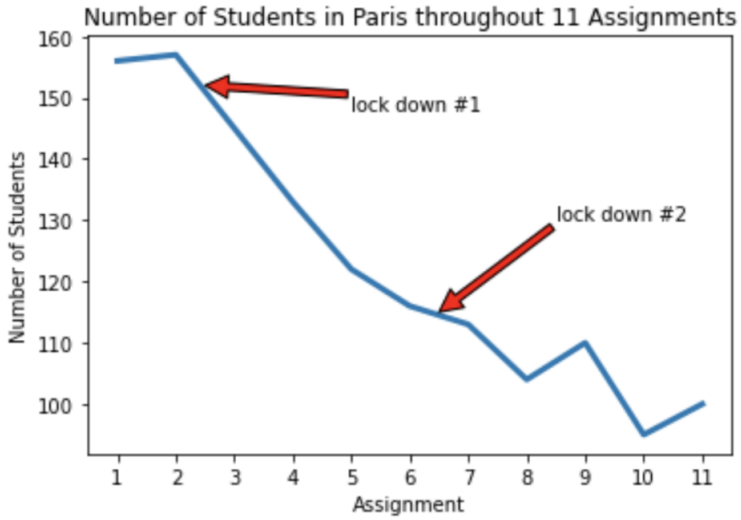

Figure 1: Number of Students Remained in Paris

Figure 1 shows a decline in the number of students who submitted their assignments in Paris, where the campus is located. The red arrows indicate two lockdowns on the campus. Classes were converted to online mode because of the lockdown. The decline reflects the influence of the lockdown on the students' decision of whether to stay around the campus or return home. Figure 2 also illustrates the changes in students' geolocations before and after the lockdowns. By comparing the students' geolocations between the time of HW02 and HW11, we can observe a clear trend, in which students moved away from Paris to other areas in France, out of France, and even out of Europe.

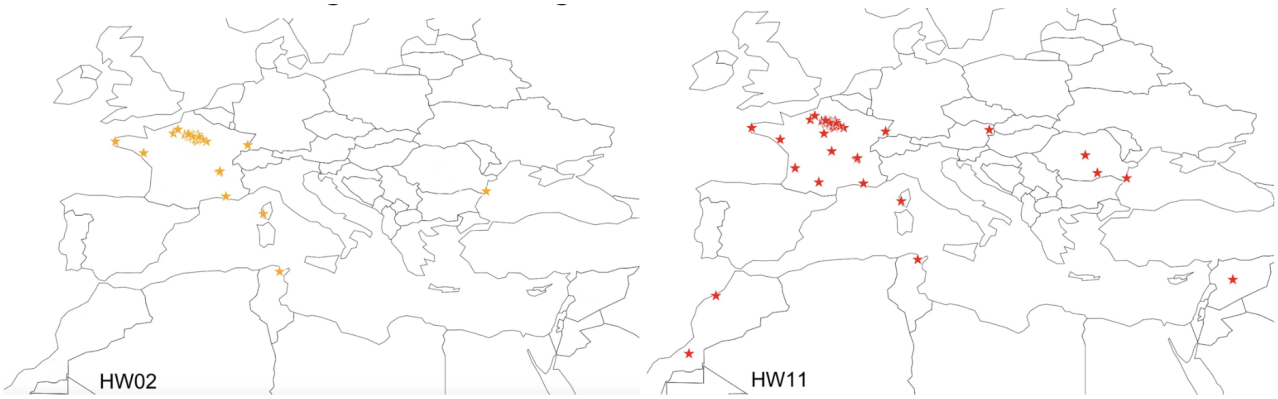

Figure 2: Students' Geolocations Before and After Lockdowns

3.2.2 Distance and Travel Frequency

A single log record can only reflect students' momentary geographical information. However, considering the nature of our dataset, we believe that what matters is the change in locations both within one task and among different tasks. There are several ways to reflect the changes in geolocations. [14] have shown that Distance and Travel Speed can be calculated by geolocations. Even though Travel Speed does not work in our case since we focus on the whole semester rather than merely one exam taken in a short period of time, Distance is a good variable that reflects students' travel behaviors.

We arranged all the students' geolocation log information by DateTime, calculated the Distance between every two continuous logs using the haversine formula, and summed the total traveled distances for each student throughout the semester using the following formula:

$$D_{total} = \sum_{n=0}^{k-1} haversine\{(lat_n, lon_n), (lat_{n+1}, lon_{n+1})\}$$

The haversine function implemented in Python is defined as below:

```
def haversine(lat1, lon1, lat2, lon2):
    lat1, lon1 = np.radians([lat1, lon1])
    lat2, lon2 = np.radians([lat2, lon2])
    a = np.sin((lat2-lat1)/2)**2 + \
        np.cos(lat1) * np.cos(lat2) * np.sin((lon2-lon1)/2)**2
    return 6371 * 2 * np.arcsin(np.sqrt(a))
```

Besides Distance, we also calculate another variable, Travel Frequency, to directly reflect students'

changes in geolocation. We observed that most students traveled around Paris or France, and only a few traveled to countries outside Europe, indicating that there were more French students than international students. Hence, we divide the locations into five categories, Paris, Île-de-France excluding Paris (IdF), France excluding Île-de-France (France), Europe excluding France (Europe), and out of Europe. Every time a student travels from one category to another, the Travel Frequency is incremented by one.

3.2.3 Grade

Since our dataset does not include final grades that indicate students' academic performance, we use the following formula to calculate the Final Grade for every student.

$$\text{Final Grade} = 50\% \times \text{Midterm Exam Score} + 50\% \times \text{Final Exam Score}$$

Both Midterm Exam Score and Final Exam Score were converted to a scale of 100. The assignment and lab grades are not considered, and their unreliability is explained in the previous section. Based on the numerical Final Grade, we mapped students' grades on a pass-and-fail basis. Students with a Final Grade greater or equal to 50 will be marked as Passed. Otherwise, they will be marked as Failed.

3.3 Decision Tree

As the previous work suggests, the use of Decision Tree Models is well suited for the prediction tasks in educational scenarios. The decision tree has multiple advantages in this case. First, the prediction result from the decision tree is easy to interpret (for instructors) and similar to human decision-making. Secondly, it can easily handle the mixed type of features in our Moodle datasets. More importantly, according to [18], feature importance can be derived during training, which will enable us to concentrate on a few crucial factors for the prediction.

We implemented a decision tree classifier based on the total distance and travel frequency that we processed. Utilizing the scikit-learn package in Python, we split the data into training and testing, and we constructed a decision tree to predict the final performance (Passed/Failed). For our first experiment, we obtained 0.9875 for the Training set score and 0.5072 for the Test set score.

To control the potential overfitting issue, we pruned the decision tree with max_leaf_nodes =

10, min_samples_leaf = 5, max_depth = 5 and received an improved Test set score of 0.58.

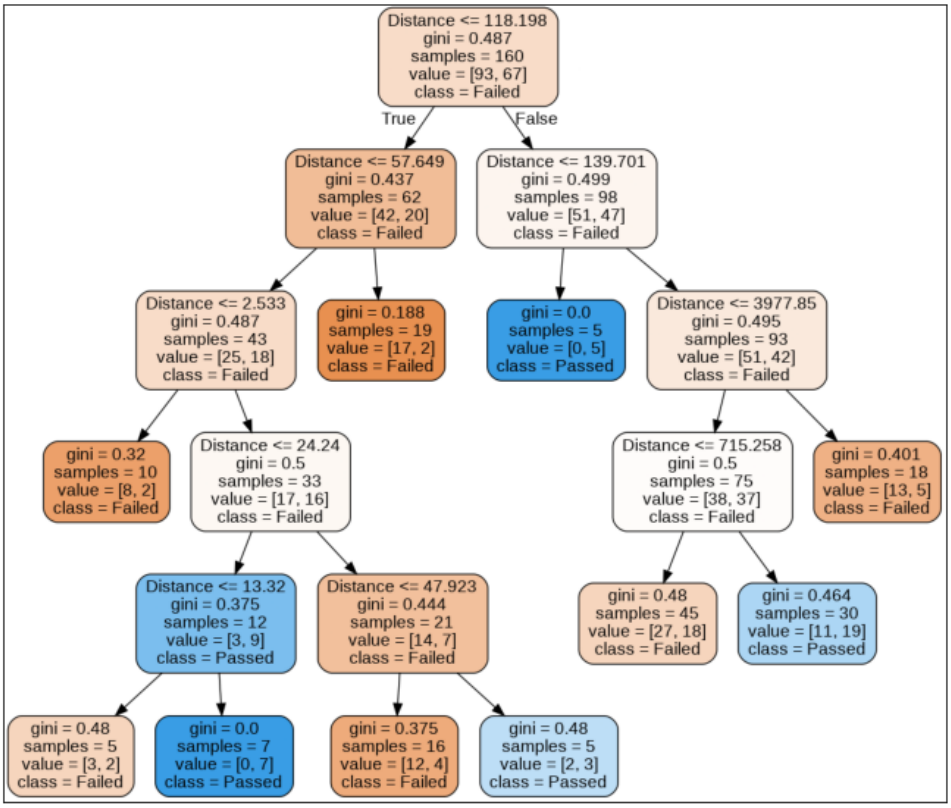

Figure 3: Visualization of the Decision Tree Classifier

In addition, by analyzing the numerical feature importance, we observed that Distance plays a dominant role in students' final performance.

Variable: Distance Importance: 0.88

Variable: Travel Frequency Importance: 0.12

The preliminary results from our decision tree model exhibit the ranking of two geolocation features in their importance. As shown in Figure 3, all the decisions are made based on Distance in this pruned tree, which confirms the importance of the Distance factor. The results also confirm the correlation between students' geolocation features and their studying outcomes. However, the relatively low accuracy score for classification suggests that aggregating all the geolocation features as a whole might omit crucial information regarding the unique change patterns in geolocation. Instead, since a student's geolocation feature is observed repeatedly throughout the semester, a longitudinal approach may yield better prediction results.

3.4 Markov Chain and Association Rule Mining

In order to study how students' geolocation features change throughout the semester with the influence of the lockdowns, we divided the semester into three periods and focused on assignments as indicators. Period-I contains Assignment 1-4, Period-II contains Assignment 5-8, and Period-III contains Assignment 9-11. Period-I reflects the influence of the first lockdown which occurred between Assignment 2 and 3. Period-II reflects the influence of the second lockdown between Assignment 6 and 7. Period-III shows how the end of the semester affects students' changes in geolocation.

Based on the feature importance from our decision tree model, we abandoned the variable Travel Frequency. We calculated the Distance of each student during every period. Then for each period, based on the Distance and its statistical distribution, we clustered the students into four groups: Most (Distance $>$ 500 km), Moderate (50 km $<$ Distance $\leq$ 500 km), Least (0 km $<$ Distance $\leq$ 50 km), and None (Distance = 0km).

Then we applied Markov Chain (MC) and Association Rule Mining (ARM) to the student groups. [10] has shown the effectiveness of these two methods in analyzing the changes in students' assignment behaviors throughout the semester. MC is conducted with R package *klaR* and ARM is performed with the *arules* package. Figure 4 shows the result of MC.

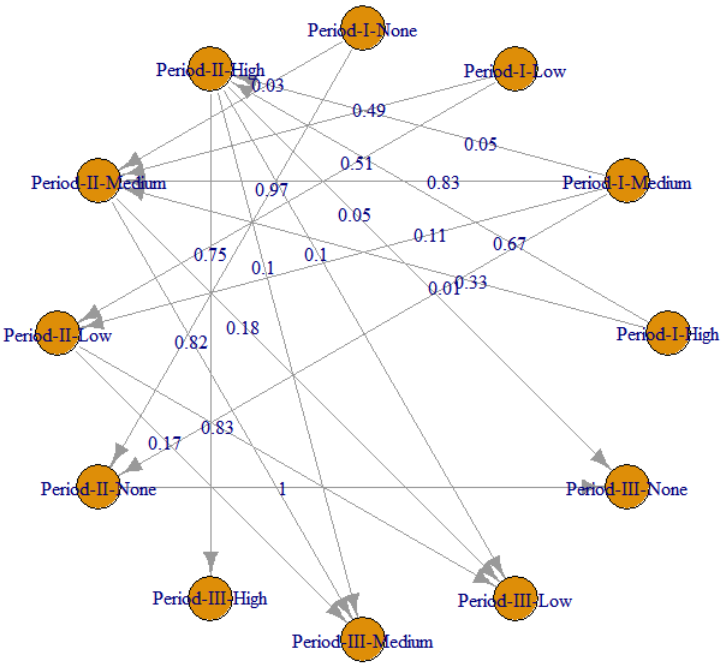

Figure 4: Transition Probabilities among Different Groups

Each node represents a group of students for a period, and the arrows between the nodes show the probability of students transit from one group to another between different periods. Our result of MC reveals how students' geolocation features change among periods. For example, for students belonging to the Most group in Period-I, though the transition probability from Period-I-Most to Period-II-Most is 0.53, which is the highest among all the probabilities of transition from Period-I-Most, it also indicates that Period-I-Most has a probability of 0.47 transiting to other groups in Period-II. As another example, for the None group in Period-I, the transition probabilities to Period-II-None and Period-II-Least are both 0.37. These data show that students' geolocation features did change over periods, which was influenced by the lockdowns.

	lhs		rhs	support	confidence	coverage	lift
[1]	{Period.I.Most, Period.III.Least}	=>	{Passed}	0.01357466	0.7500000	0.01809955	1.726562
[2]	{Period.II.None, Period.III.Least}	=>	{Failed}	0.03167421	0.7000000	0.04524887	1.237600
[3]	{Period.II.None, Period.III.Moderate}	=>	{Failed}	0.01809955	1.0000000	0.01809955	1.768000
[4]	{Period.I.Moderate, Period.II.None}	=>	{Failed}	0.03167421	1.0000000	0.03167421	1.768000
[5]	{Period.I.Least, Period.II.None}	=>	{Failed}	0.04524887	0.7142857	0.06334842	1.262857
[6]	{Period.II.Most, Period.III.Moderate}	=>	{Passed}	0.01357466	0.7500000	0.01809955	1.726562
[7]	{Period.II.Most, Period.III.None}	=>	{Passed}	0.01357466	0.7500000	0.01809955	1.726562
[8]	{Period.I.None, Period.II.Moderate}	=>	{Failed}	0.02262443	0.8333333	0.02714932	1.473333
[9]	{Period.I.Least, Period.III.Moderate}	=>	{Failed}	0.06334842	0.8235294	0.07692308	1.456000
[10]	{Period.I.Moderate, Period.III.None}	=>	{Passed}	0.02262443	0.8333333	0.02714932	1.918403
[11]	{Period.I.Moderate, Period.II.Least}	=>	{Passed}	0.03619910	0.7272727	0.04977376	1.674242
[12]	{Period.I.Most, Period.II.Least, Period.III.None}	=>	{Passed}	0.01357466	0.7500000	0.01809955	1.726562
[13]	{Period.I.Least, Period.II.None, Period.III.Most}	=>	{Failed}	0.01357466	1.0000000	0.01357466	1.768000
[14]	{Period.I.Moderate, Period.II.None, Period.III.Least}	=>	{Failed}	0.01357466	1.0000000	0.01357466	1.768000
[15]	{Period.I.Least, Period.II.None, Period.III.Least}	=>	{Failed}	0.01357466	0.7500000	0.01809955	1.326000
[16]	{Period.I.Moderate, Period.II.None, Period.III.Moderate}	=>	{Failed}	0.01357466	1.0000000	0.01357466	1.768000
[17]	{Period.I.None, Period.II.Moderate, Period.III.Moderate}	=>	{Failed}	0.01357466	1.0000000	0.01357466	1.768000
[18]	{Period.I.Least, Period.II.Moderate, Period.III.Moderate}	=>	{Failed}	0.03167421	0.7777778	0.04072398	1.375111
[19]	{Period.I.Least, Period.II.Least, Period.III.Moderate}	=>	{Failed}	0.03167421	0.8750000	0.03619910	1.547000
[20]	{Period.I.Moderate, Period.II.Moderate, Period.III.None}	=>	{Passed}	0.01809955	0.8000000	0.02262443	1.841667

Figure 5: ARM Results

We then conduct the ARM on the groups. Figure 5 shows 20 prediction rules along with the support, confidence, coverage, and lift values. The "lhs" column of each rule describes a pattern of changes in students' geolocation features. The "rhs" column shows the prediction based on the "lhs". Most of the rules for Failed follow the pattern, in that students with an increasing trend of changes in geolocation features will fail, e.g. Rule 2, 3, 8, 9, 13, 17, 18, 19. Similarly, students with a decreasing trend of changes in geolocation features usually pass, e.g. Rule 1, 6, 7, 10, 11, 12, 20.

There are also rules that cannot be explained by the pattern above, e.g. Rule 4 and 5. These two rules indicate that students in Period-I-Moderate or Period-I-Least, and later transferred to Period-II-None will fail. The possible explanation is that, when we calculated the Distance for every period, we kept all the students as long as they have a log record in the dataset, even if some of them might have missed assignments or labs. A detailed discussion of the indication will be delivered in section 4.

3.5 Clustering Analysis

Building on the previous model, we decided to implement a clustering analysis with the geolocation features together with other behavioral features in completing the assignment before conducting MC and ARM. This adaptation will address our RQ2 by deriving feasible prediction rules based on both students' geolocations and study behaviors. Besides the Distance for each period, we derived additional features from the assignment log files with event names including 'Quiz attempt submitted', 'Quiz attempt summary viewed', and 'Quiz attempt viewed'. For each assignment, we derived the following features:

- Duration - Time difference between 'Quiz attempt submitted' and the first 'Quiz attempt viewed'.
- Attempts - Total number of 'Quiz attempt summary viewed' and 'Quiz attempt viewed'.
- Start - Time difference between the first 'Quiz attempt viewed' and the assignment open time. i.e. How soon they started working after an assignment was posted.
- Complete - Time difference between 'Quiz attempt submitted' and the assignment deadline. i.e. How promptly they completed each assignment before the deadline.

According to the Period-I, Period-II, and Period-III defined above, we averaged the above statistics and combined it with the Distance to produce the refined dataset for clustering. To begin with, we adopted K-Means Clustering in Python, and we set the number of clusters to be 3 (with names to be defined later). We normalized the data to ensure that our five features are in the same scale and conducted the clustering. After we attained the clusters, we visualized the clustering results using Principal Component Analysis (PCA) and *plotly*. The visualizations of three clusters in Period-I in 2D and 3D are displayed in Figure 6 and Figure 7.

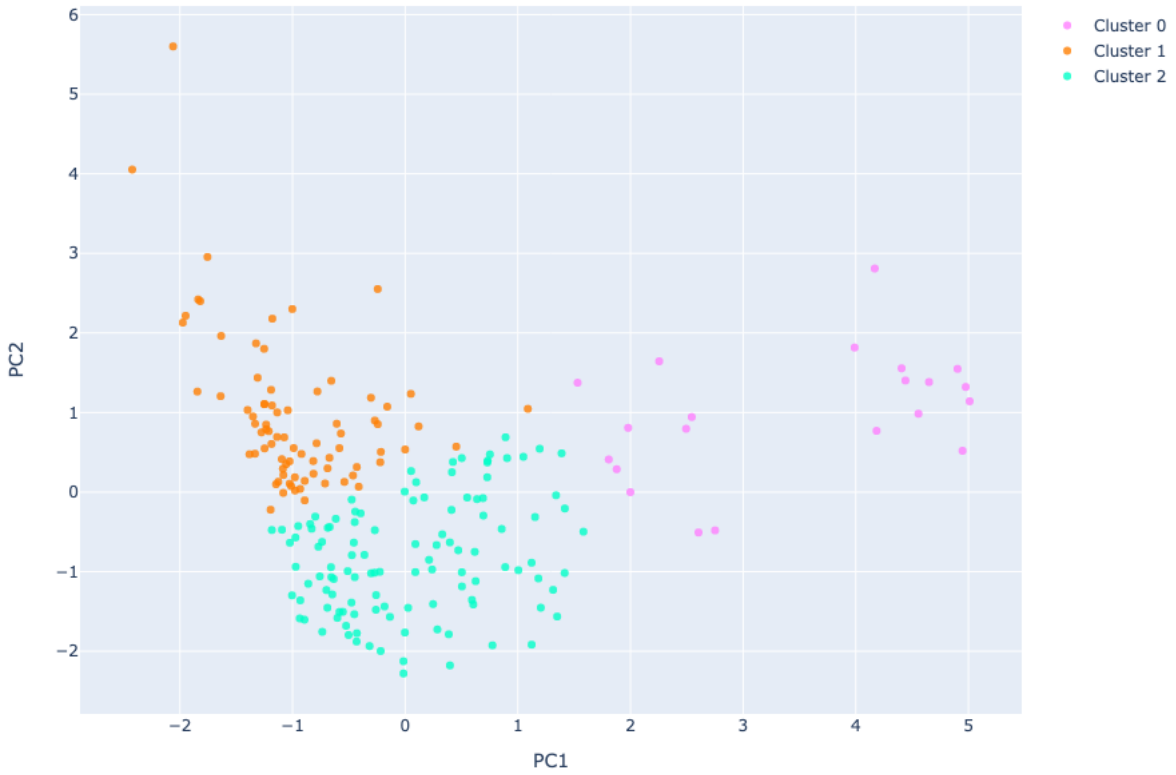

Figure 6: Visualizing Clusters in 2D Using PCA

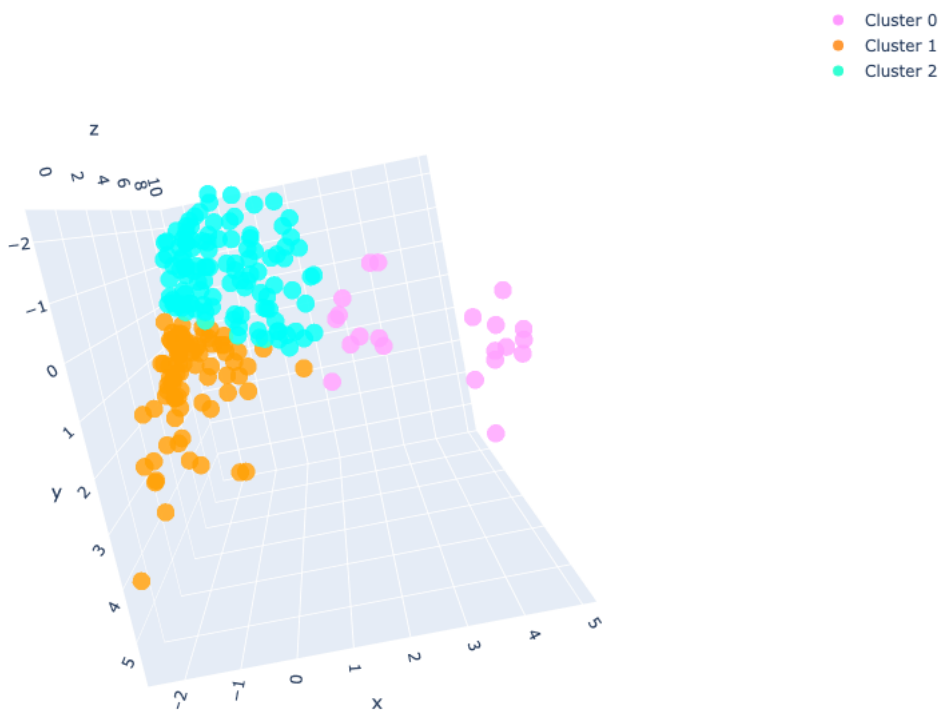

Figure 7: Visualizing Clusters in 3D Using PCA

The visualization verifies the existence of the clusters. To endow each cluster with its practical meaning, we proceeded to visualize the clustering results for each period along with their feature characteristics using the *seaborn* package so that we can compare the clusters in parallel. The visualization results for the three periods are given in Figure 8, Figure 9, and Figure 10.

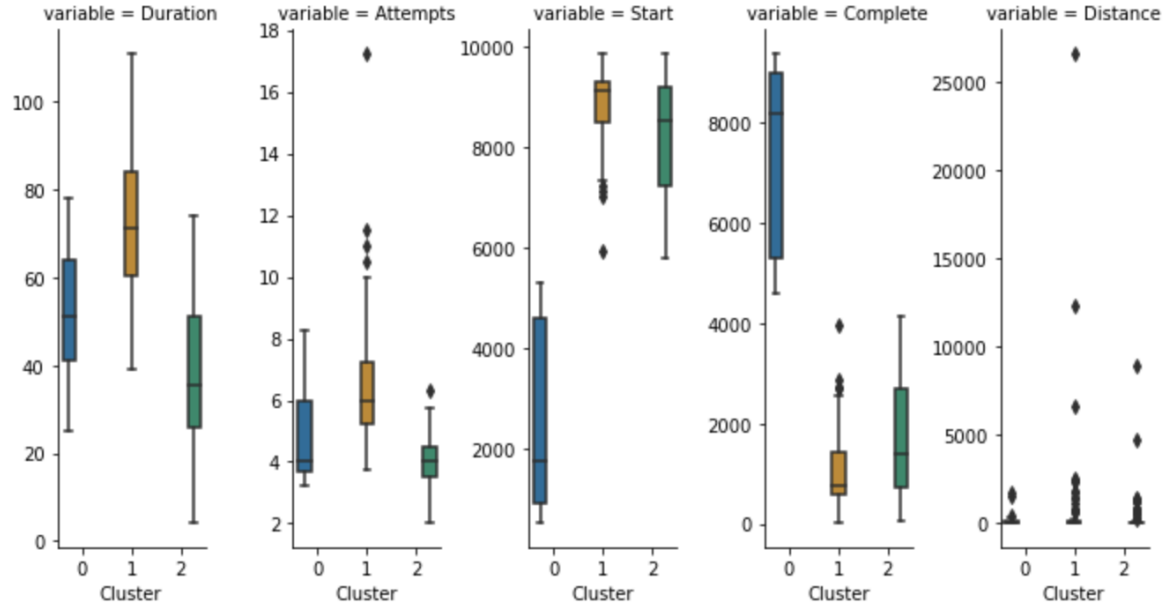

Figure 8: Period-I

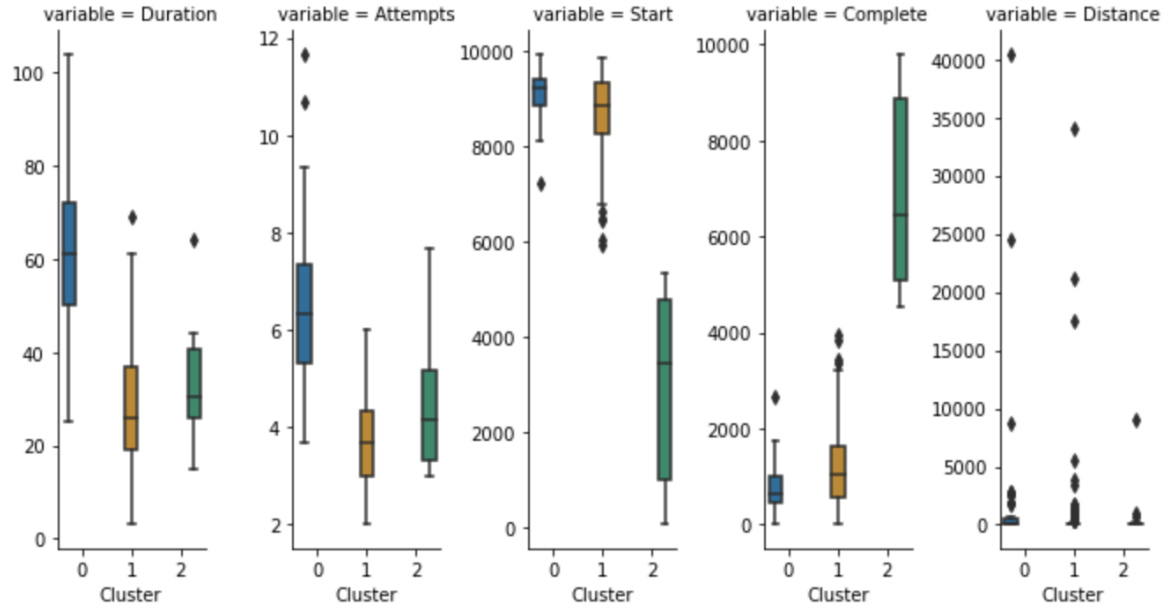

Figure 9: Period-II

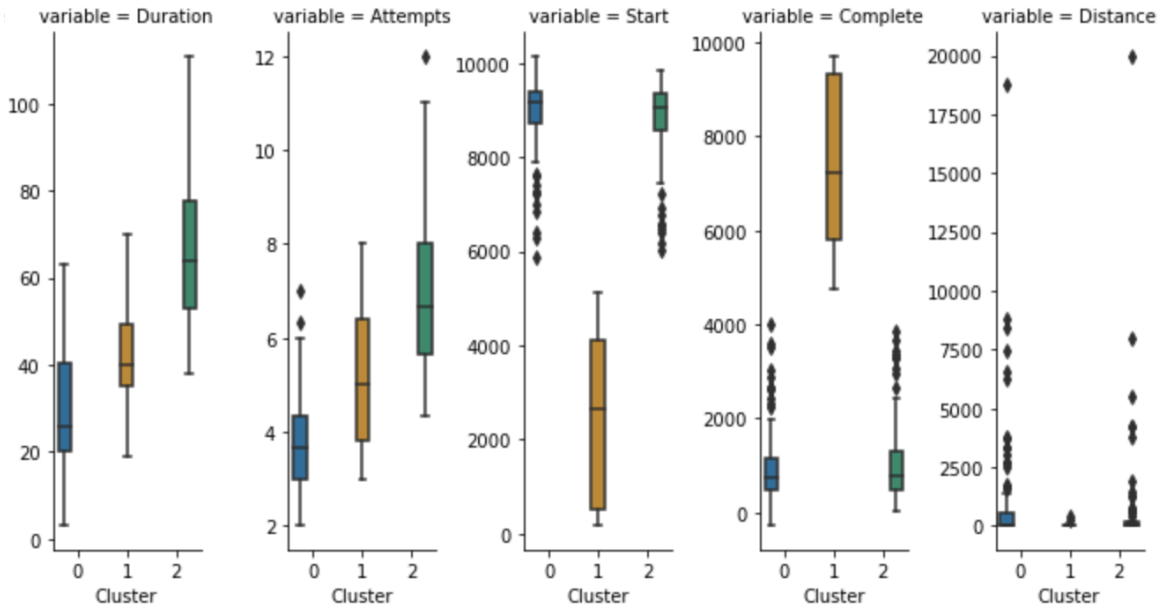

Figure 10: Period-III

The clustering showed that students displayed similar patterns in all three periods. We proposed the clustering name to be Dedication Level (High/Medium/Low), and we manually categorize the clusters into one of three levels. The students excluded from the clustering analysis are manually assigned to the None group.

1. **High Dedication** - The students in Cluster 0 in Period-I, Cluster 2 in Period-II, and Cluster 1 in Period-III.

   They start the assignment much earlier than the given deadline (small Start, large Complete), spend more time completing the assignment (Duration), and the number of views (Attempts) is relatively higher. Across all three periods, students with High Dedication Level experienced the least changes in geolocations as their Distance remains to be significantly smaller than the individuals in other clusters.

2. **Medium Dedication** - The students in Cluster 2 in Period-I, Cluster 1 in Period-II, and the students in Cluster 0 in Period-III.

   The prominent features of these students are they start and complete the assignment at the last moment (large Start, small Complete), and spend the least time and number of attempts to complete the assignment (Duration and Attempts). This means that even though students in these clusters submitted the assignment right before the deadlines, they gave a minimum effort for the assignment.

3. **Low Dedication** - The students in Cluster 1 in Period I, Cluster 0 in Period II, and Cluster 2 in Period III exhibit similar assignment submission patterns.

   These students start their assignment submission near the deadline (small Complete), and they spend the most time (Duration) completing the assignment than the other two groups. These signs revealed that although these students were not as familiar with the course materials as the Medium Class, they still did not dedicate enough effort in studying. Note that there are no significant disparities between the Distance for Medium Dedication and Low Dedication groups, but the figures for both of these groups are substantially higher than the High Dedication group.

4. **None Group** - Students who did not submit their assignments frequently were manually assigned to this group.

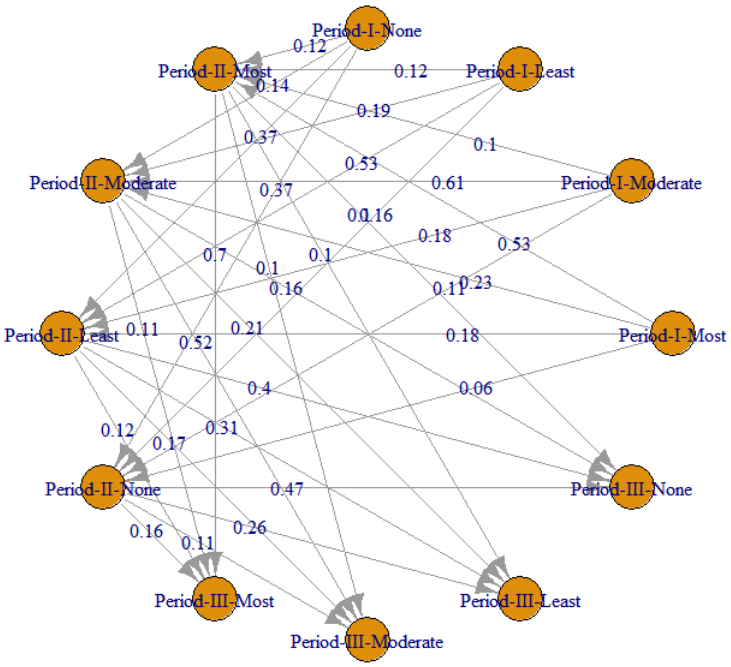

Figure 11: Transition Probabilities among Different Dedication Levels

Then we apply MC and ARM to the clusters. Figure 11shows the result of MC with clustering. Compared to MC without clustering, MC with clustering shows lower probabilities in transition among different Dedication Levels, implying that usually, students' behavior would not vary too much throughout the semester. Using ARM with clusters, 26 prediction rules are generated along with the support, confidence, coverage, and lift values, as shown in Figure 12.

	lhs		rhs	support	confidence	coverage	lift
[1]	{Period.III.High}	=>	{Passed}	0.04435484	0.7333333	0.06048387	1.818667
[2]	{Period.II.High}	=>	{Passed}	0.05645161	0.7000000	0.08064516	1.736000
[3]	{Period.II.None}	=>	{Failed}	0.11290323	0.8484848	0.13306452	1.421785
[4]	{Period.I.None}	=>	{Failed}	0.11290323	0.8235294	0.13709677	1.379968
[5]	{Period.III.None}	=>	{Failed}	0.11693548	0.8529412	0.13709677	1.429253
[6]	{Period.I.Low}	=>	{Failed}	0.19354839	0.6075949	0.31854839	1.018132
[7]	{Period.II.High, Period.III.High}	=>	{Passed}	0.04435484	0.7333333	0.06048387	1.818667
[8]	{Period.I.High, Period.III.High}	=>	{Passed}	0.03225806	0.6666667	0.04838710	1.653333
[9]	{Period.I.Medium, Period.III.High}	=>	{Passed}	0.01209677	1.0000000	0.01209677	2.480000
[10]	{Period.I.Medium, Period.II.High}	=>	{Passed}	0.02419355	1.0000000	0.02419355	2.480000
[11]	{Period.I.None, Period.II.None}	=>	{Failed}	0.11290323	0.8484848	0.13306452	1.421785
[12]	{Period.II.None, Period.III.None}	=>	{Failed}	0.11290323	0.8484848	0.13306452	1.421785
[13]	{Period.I.None, Period.III.None}	=>	{Failed}	0.11290323	0.8484848	0.13306452	1.421785
[14]	{Period.I.Medium, Period.II.Low}	=>	{Failed}	0.03225806	0.6153846	0.05241935	1.031185
[15]	{Period.I.Low, Period.III.Low}	=>	{Failed}	0.12500000	0.6326531	0.19758065	1.060121
[16]	{Period.II.Medium, Period.III.Low}	=>	{Failed}	0.06451613	0.6400000	0.10080645	1.072432
[17]	{Period.I.Low, Period.II.Medium}	=>	{Failed}	0.10080645	0.6410256	0.15725806	1.074151
[18]	{Period.I.High, Period.II.High, Period.III.High}	=>	{Passed}	0.03225806	0.6666667	0.04838710	1.653333
[19]	{Period.I.Medium, Period.II.High, Period.III.High}	=>	{Passed}	0.01209677	1.0000000	0.01209677	2.480000
[20]	{Period.I.None, Period.II.None, Period.III.None}	=>	{Failed}	0.11290323	0.8484848	0.13306452	1.421785
[21]	{Period.I.Low, Period.II.Low, Period.III.Low}	=>	{Failed}	0.08870968	0.6285714	0.14112903	1.053282
[22]	{Period.I.Low, Period.II.Low, Period.III.Medium}	=>	{Passed}	0.01612903	0.8000000	0.02016129	1.984000
[23]	{Period.I.Medium, Period.II.Low, Period.III.Medium}	=>	{Failed}	0.01612903	1.0000000	0.01612903	1.675676
[24]	{Period.I.Low, Period.II.Medium, Period.III.Low}	=>	{Failed}	0.03629032	0.6428571	0.05645161	1.077220
[25]	{Period.I.Medium, Period.II.Medium, Period.III.Low}	=>	{Failed}	0.02822581	0.7000000	0.04032258	1.172973
[26]	{Period.I.Low, Period.II.Medium, Period.III.Medium}	=>	{Failed}	0.06451613	0.6400000	0.10080645	1.072432

Figure 12: Results of ARM with Clustering

The ARM results with clustering exhibit similar trends as the previous without clustering predicts. In this case, students' behavior and their geolocation feature jointly influenced their Dedication Levels and ARM outcomes. From observation, students who have a High Dedication Level in any period enjoy an overwhelmingly positive learning outcome as predicted. In addition, Similar to what we obtained in Section 3.4, that "students with a decreasing trend of changes in geolocation features usually pass", we have rule number 9, 10, and 19 where students moved from Medium Dedication Level (large Distance) to High Dedication Level (small Distance) throughout the semester. Furthermore, we witness improved support and coverage for ARM with clustering than the previous one without clustering.

4 Discussion

4.1 What is the impact of stay-at-home orders on students' academic performance? (RQ1)

Due to the two lockdowns and the consequent stay-at-home orders, students travel much more than in a normal in-person or online semester. We first attempted to use two variables, Distance and Travel Frequency, to indicate the travel behavior of students. According to our decision tree model in section 3.3, Distance is proven to be more important than Travel Frequency. This is because Distance can also be understood as a weighted sum of each time of travel. If a student travels from category Paris to category out of Europe, this time of travel should be given more

weight than traveling from category Paris to category IdF. Hence, Distance can reflect Travel Frequency to a large extent, and is consequently a better variable to directly indicate the influence of stay-of-home orders on the students.

In section 3.4, by dividing the semester into three periods and grouping students with the values of Distance, we further refine the influence of stay-of-home orders on different students. With MC, the transition probabilities among groups indicate that stay-at-home orders can affect students to different extents during different periods. It proves that, in our case, ARM is more appropriate than a traditional machine learning approach, such as Regression and Decision Tree, which focus on the variables themselves rather than the changes of a certain variable over time. The rules produced by ARM show how the trends of changes in geolocation features can influence students' academic performance. In general, students with an increasing trend of changes in geolocation features are more likely to fail, while students with a decreasing trend are more promising in passing the course. Though many of the rules have high confidence, the support and coverage values are usually not very high. From this, we can know that the patterns of changes in students' geolocation features vary for individuals.

There are also two rather abnormal rules, which can be illustrated by the None group of students. We expect the None group contains students staying in the same region and not traveling throughout the whole period. If so, the None group would include those who are the least affected by the stay-at-home orders. However, students who constantly miss their assignments and labs, and thus have limited log records, are also classified into the None group. While the travel behavior of these students cannot be derived due to the lack of log records, the fact that they miss many assignments and labs might have a greater influence on their academic performance compared to the influence of stay-at-home orders.

The results from both MC and ARM indicate that, though stay-at-home orders have significant impacts on students' academic performance, considering it together with other factors will help acquire a better result. The next section will explain how we use geolocation jointly with study behaviors to improve prediction accuracy.

4.2 Can academic performance be predicted from students' geolocations and study behaviors? (RQ2)

From the results of ARM that we derived with clustering, the answer to this question is definite. [7] has shown that students' study behaviors are correlated with their performance, but studies

on students' physical environment as a predictive factor are lacking. In our study, by using the Decision Tree Classifier, we have proved the feasibility of predicting students' academic performance directly from geolocation factors (with Distance as a dominant factor). With the aim of improved classification accuracy, we modified our methods to exploit the predictive power of geolocation features on academic performance.

Considering the nature of Moodle dataset that reflects a trend with continuous observations throughout the semester, we adopted a longitudinal approach and divided the data into three periods for further analysis. Through MC and ARM in Section 3.4, we derived a set of prediction rules that reflect students' performance based on the change in their traveling distance across each period. These prediction rules not only confirm the predictive power of geolocation features (Distance, specifically) but also yield a satisfying confidence rate of approximately greater than 75% for each rule.

To achieve a more complete model with higher accuracy, we relied on clustering analysis that factors in both study behavior and geolocations, and we recognized there are common patterns for each cluster that link the features together. Then, we interpreted the clusters as groups with different Dedication Levels (High/Medium/Low). Within each Dedication group, Distance is directly correlated with other behavioral features. For instance, having a High Dedication Level over a long Distance is impossible in any period. The clustering result again verified the intertwined relationship between geolocation factors and study behaviors, as they complement each other to influence the Dedication Level of each individual.

Finally, we repeated the MC and ARM with a focus on the changes in Dedication Levels. This time, we obtained prediction rules with increased support and coverage, since the students with similar characteristics were clustered into the same Dedication Level group. Additionally, we discovered the overwhelming trend that students who achieve a High Dedication Level in any period are likely to have a positive learning outcome. Therefore, based on results from the Decision Tree classification, Clustering analysis, MC, and ARM, we are assured that either geolocations or study behaviors can be used for predicting students' academic performance, and we are able to obtain more accurate predictions when the two factors are analyzed in a joint approach.

4.3 Challenges and improvements

The main challenge is the lack of previous studies on the prediction of academic performance using geolocation as a factor. Though geolocation is studied in other fields of EDM/LA, the variables derived by geolocation might not fit the context of prediction. For example, Travel Speed is a good variable for cheat detection in [14], since the study only concentrates on a short period of time. However, its meaning is not well adapted to our cases since our focus is rather the whole semester. Besides Distance, we attempted to create our own variable, Travel Frequency, which is proven to be not as effective as Distance.

Hence, our project can be improved by digging deeper into geolocation. We believe that geolocation is a significant factor in the prediction of academic performance. There should be not merely Distance but also other information that we can derive from geolocation. In the case of Travel Frequency, a redefinition of this variable or an alternative way to calculate it might endow greater importance on the variable. The more variables we have that are derived from geolocation, the more thorough understanding we would gain regarding the relationship between geolocation and academic performance.

In addition, the dataset that we used to derive our prediction models is from the Fall 2020 semester when every student was online amid the pandemic. On one hand, our study provides a straightforward solution for effective predictions for academic performance. On the other hand, questions may arise regarding the applicability of our prediction rules, since the Fall 2020 semester was a special case that included direct impacts from the pandemic and may not reflect the general situations for online learning. Thus, it is necessary to validate and refine our prediction models on datasets from other time periods so that the prediction results will become independent of the extraneous factors.

Moreover, our study has revealed that academic performance can be jointly predicted by students' geolocations and study behaviors. The results from clustering analysis and ARM have also shown their individual and combined contribution to each student's learning outcomes. However, one complication is that the causal relationship between geolocations and study behaviors is still unclear. This question is worth considering because depending on their causal relation, i.e. whether it is one's change in geolocations that lead to his/her change in study behaviors, the opposite direction, or the two factors are merely correlated through external factors, educators may resort to different approaches for timely and effective intervention. Therefore, future work is needed to understand the causal relationship within these factors and their respective mechanisms

in determining students' academic performance.

5 Personal Contributions

In the data preprocessing stage, I worked with my teammate together to process data and converted files into the required format for later use. Specifically, among the data mentioned in section 3.2, I acquired Travel Frequency and Grade from the raw dataset. I defined the variable Travel Frequency by observing the characteristic of the IP geolocation data, though it turned out to be not as effective as the other variable Distance. I conducted Markov Chain and Association Rule Mining in both section 3.4 and 3.5. In section 3.4, I defined the three periods (whose definition is later refined with the co-effort of me and my teammate) and grouped the students accordingly to the Distance. In section 3.5, I conducted Markov Chain and Association Rule Mining on the clustering analysis result obtained by my teammate. In section 4, I focus more on our RQ1 in section 4.1 and the first two paragraphs in section 4.3. But in fact, I believe my teammate and I fairly contribute to section 4, since it is a product of our discussion and reflection on this project.

6 Conclusion

In this study, we identified the significance of geolocation in predicting students' academic performance and we constructed efficient prediction rules that incorporate both changes in geolocation and study behaviors as factors for the prediction. After conducting a systematic literature review, we derived geographical location from students' IP addresses, and we preprocessed the dataset to get Distance, Traveling Frequency, and Grades for preliminary analysis. Specifically, by using the Decision Tree Classifier, we have proved the feasibility of predicting students' academic performance directly from geolocation factors, and we obtained their relative importance with Distance as a dominant factor. Considering the nature of the longitudinal data, we then divided our dataset into three periods and carried out MC and ARM based on changes in Distance. The resulting prediction rules yield a satisfying confidence rate of approximately greater than 75%. To improve the accuracy of our prediction, we relied on clustering analysis that factors in both study behavior and geolocations to construct clustering groups with different Dedication Levels (High/Medium/Low/None). Finally, we repeated the MC and ARM with a focus on the changes in Dedication Levels across three periods and obtained prediction rules with increased support

and coverage.

More importantly, the results of this study address the two research questions, that stay-at-home orders have significant impacts on students' academic performance, and that we can obtain more accurate prediction results when students' geolocations and study behaviors are analyzed jointly. As for the practical outcomes, the finalized prediction rules we developed through comprehensive clustering analysis, MC, and ARM guarantee an improved prediction accuracy of above 0.85 on average. These prediction rules provide straightforward criteria that enable instructors to assess students' academic performance accurately. Meanwhile, we can derive simple conclusions to assist instructors' evaluation and implementation of possible pre-intervention measures. Examples of students who require extra attention:

1. Students who observe a decreasing trend in Dedication Level throughout the semester have a higher chance of failure.

2. Students who have never achieved a High Dedication Level in any period are prone to fail the course.

3. Students who traveled more than 5,000 km can never receive a High Dedication Level in that period.

In the meantime, our study also admits a few limitations. Firstly, the change in geolocation is a complex factor. Currently, we have only examined the predictive power of geolocation by converting it into Distance and Travel Frequency. However, whether this Distance is the pivotal factor that carries the geolocation information and whether our conversion preserves the crucial information remains unknown. Secondly, our prediction model is established on datasets from a special period with the severe impact of the global pandemic. Therefore, the model might not generate the most accurate results for general situations. Thirdly, the causal relationship between geolocations and study behaviors is still unclear to us, which may lead to potential issues regarding educators' decisions on the optimal interventions that prevent academic failure.

As for future work, researchers may exploit the geolocation data for other valuable derived features that not only preserve vital information but also generate the most predictive power for students' academic performance. To ensure the applicability of the prediction rules, it is of great interest to validate and refine our models on other datasets so that the prediction results will become independent of the extraneous factors. Lastly, future work is needed to understand the

causal relationship between geolocations and study behaviors as well as their respective mechanisms in determining students' academic performance.

References

[1] O. B. Adedoyin and E. Soykan, "Covid-19 pandemic and online learning: the challenges and opportunities," *Interactive Learning Environments*, pp. 1–13, 2020. [Online]. Available: https://doi.org/10.1080/10494820.2020.1813180

[2] R. Gopal, V. Singh, and A. Aggarwal, "Impact of online classes on the satisfaction and performance of students during the pandemic period of covid 19," *Education and Information Technologies*, vol. 26, no. 6, p. 6923–6947, 2021.

[3] M. Imran, S. Latif, D. Mehmood, and M. S. Shah, "Student academic performance prediction using supervised learning techniques," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 14, no. 14, p. pp. 92–104, Jul. 2019. [Online]. Available: https://online-journals.org/index.php/i-jet/article/view/10310

[4] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1355, 2020. [Online]. Available: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1355

[5] K. S. Na and Z. Tasir, "A systematic review of learning analytics intervention contributing to student success in online learning," in *2017 International Conference on Learning and Teaching in Computing and Engineering (LaTICE)*, 2017, pp. 62–68.

[6] Y. Shavitt and N. Zilberman, "A geolocation databases study," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 10, pp. 2044–2056, 2011.

[7] A. Abu Saa, M. Al-Emran, and K. Shaalan, "Factors affecting students' performance in higher education: A systematic review of predictive data mining techniques," *Technology, Knowledge and Learning*, vol. 24, no. 4, p. 567–598, 2019.

[8] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Systems with Applications*, vol. 41, no. 4, Part 1, pp. 1432–1462, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417413006635

[9] T. Yu and I.-H. Jo, "Educational technology approach toward learning analytics: Relationship between student online behavior and learning performance in higher education," in *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, ser. LAK '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 269–270. [Online]. Available: https://doi.org/10.1145/2567574.2567594

[10] M. Kokoç, G. Akçapınar, and M. N. Hasnine, "Unfolding students' online assignment submission behavioral patterns using temporal learning analytics," *Educational Technology amp; Society*, vol. 24, no. 1, p. 223–235. [Online]. Available: https://www.jstor.org/stable/26977869

[11] R. Hasan, S. Palaniappan, A. R. A. Raziff, S. Mahmood, and K. U. Sarker, "Student academic performance prediction by using decision tree algorithm," in *2018 4th International Conference on Computer and Information Sciences (ICCOINS)*, 2018, pp. 1–5.

[12] O. H. T. Lu, A. Y. Q. Huang, J. C. Huang, A. J. Q. Lin, H. Ogata, and S. J. H. Yang, "Applying learning analytics for the early prediction of students' academic performance in blended learning," *Journal of Educational Technology Society*, vol. 21, no. 2, pp. 220–232, 2018. [Online]. Available: http://www.jstor.org/stable/26388400

[13] A. Jain and S. Solanki, "An efficient approach for multiclass student performance prediction based upon machine learning," in *2019 International Conference on Communication and Electronics Systems (ICCES)*, 2019, pp. 1457–1462.

[14] D. Komosny and S. U. Rehman, "A method for cheating indication in unproctored on-line exams," *Sensors*, vol. 22, no. 2, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/2/654

[15] H. Luo, S. Rocco, and C. Schaad, "Using google analytics to understand online learning: A case study of a graduate-level online course," in *2015 International Conference of Educational Innovation through Technology (EITT)*, 2015, pp. 264–268.

[16] S. Laki, P. Mátray, P. Hága, T. Sebők, I. Csabai, and G. Vattay, "Spotter: A model based active geolocation service," in *2011 Proceedings IEEE INFOCOM*, 2011, pp. 3173–3181.

[17] C. Guo, Y. Liu, W. Shen, H. J. Wang, Q. Yu, and Y. Zhang, "Mining the web and the internet for accurate ip address geolocations," in *IEEE INFOCOM 2009*, 2009, pp. 2841–2845.

[18] S. J. Hong, "Use of contextual information for feature ranking and discretization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 9, no. 5, pp. 718–730, 1997.