Geo-Based Prediction of Academic Performance

Literature Review

Star Chen (sc7828) - *Computer Science*
Scott Ye (my2152) - *Data Science (CS Concentration)*

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 [2], 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 [3]. 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 [6]. As seen in [15], 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 [13].

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 [16]. 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 [1]. 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, this project aims to investigate the significance of students' geolocation, a representative feature of the physical educational environment, and to predict students' academic performance based on students' geolocation and study behaviors. We will conduct EDM/LA on data from one popular LMS, Moodle for prediction, and multiple machine learning approaches such as regression, decision tree, and random forest classifier will be tested. The prediction will allow instructors to assess students' academic performance accurately and derive 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 based on students' geolocations and study behaviors?

Related Works

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 [1] 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 [14].

In [17], 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. [8] 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. [5] 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. [11] 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. [6] 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. [7] 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 IP geolocation as the main factor in predicting academic performance. Nevertheless, IP geolocation is studied for other purposes in EDM/LA. Komosny et al. [9] have 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. [12] 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 [16] and active geolocation measurement services [10]. Research such as [4] also presents methods to improve the accuracy of IP geolocation.

Conclusion

As is reflected in the previous section, IP geolocation, showing its importance in many fields as well as EDM/LA, has not been studied as the main factor in precedent works on the prediction of students' academic performance. The lack of previous research shows both the challenge and significance of our study. The main tasks of our study are the followings:

T1. Derive geographical location from students' IP addresses.

T2. Preprocess the IP geolocation data.

T3. Build a predictive model with appropriate machine learning techniques for the data.

For T1, approaches in [16] and [10] will be the two possible ways to map IP addresses and geolocation. Other online tools and services are also available for this purpose. We will consider the features of our dataset and apply the best method.

As for T2, Methods in [9] set a good example of the sufficient usage of IP geolocation. IP geolocation can be calculated into multiple attributes, which might improve the performance of our model. Techniques presented in [4] might also be helpful to improve the accuracy of IP geolocation.

In the case of T3, since several machine learning algorithms such as regression, decision tree, and random forest are proved to be performing well in different contexts, we will try applying different algorithms to our model, from the basic ones to the more advanced ones, depending on the availability of our time. Besides the regular predictive models with a result of prediction accuracy, [8] also provides another kind of predictive model by grouping similar students and setting predictive rules, which we may also try and combine into our model, based on the results of data preprocessing and our attempts with other machine learning methods. The final results of our study can be visualized with Google Analytics tools [12].

References

[1] Amjed Abu Saa, Mostafa Al-Emran, and Khaled Shaalan. 2019. Factors affecting students' performance in Higher Education: A systematic review of predictive data mining techniques. *Technology, Knowledge and Learning* 24, 4 (2019), 567–598. https://doi.org/10.1007/s10758-019-09408-7

[2] Olasile Babatunde Adedoyin and Emrah Soykan. 2020. Covid-19 pandemic and online learning: the challenges and opportunities. *Interactive Learning Environments* (2020), 1–13. https://doi.org/10.1080/10494820.2020.1813180 arXiv:https://doi.org/10.1080/10494820.2020.1813180

[3] Ram Gopal, Varsha Singh, and Arun Aggarwal. 2021. Impact of online classes on the satisfaction and performance of students during the pandemic period of COVID 19. *Education and Information Technologies* 26, 6 (2021), 6923–6947. https://doi.org/10.1007/s10639-021-10523-1

[4] C. Guo, Y. Liu, W. Shen, H. J. Wang, Q. Yu, and Y. Zhang. 2009. Mining the Web and the Internet for Accurate IP Address Geolocations. In *IEEE INFOCOM 2009*. 2841–2845. https://doi.org/10.1109/INFCOM.2009.5062243

[5] Raza Hasan, Sellappan Palaniappan, Abdul Rafiez Abdul Raziff, Salman Mahmood, and Kamal Uddin Sarker. 2018. Student Academic Performance Prediction by using Decision Tree Algorithm. In *2018 4th International Conference on Computer and Information Sciences (ICCOINS)*. 1–5. https://doi.org/10.1109/ICCOINS.2018.8510600

[6] Muhammad Imran, Shahzad Latif, Danish Mehmood, and Muhammad Saqlain Shah. 2019. Student Academic Performance Prediction using Supervised Learning Techniques. *International Journal of Emerging Technologies in Learning (iJET)* 14, 14 (Jul. 2019), pp. 92–104. https://doi.org/10.3991/ijet.v14i14.10310

[7] Abhinav Jain and Shano Solanki. 2019. An Efficient Approach for Multiclass Student Performance Prediction based upon Machine Learning. In *2019 International Conference on Communication and Electronics Systems (ICCES)*. 1457–1462. https://doi.org/10.1109/ICCES45898.2019.9002038

[8] Mehmet Kokoç, Gökhan Akçapınar, and Mohammad Nehal Hasnine. [n. d.]. Unfolding Students' Online Assignment Submission Behavioral Patterns using Temporal Learning Analytics. *Educational Technology amp; Society* 24, 1 ([n. d.]), 223–235. https://www.jstor.org/stable/26977869

[9] Dan Komosny and Saeed Ur Rehman. 2022. A Method for Cheating Indication in Unproctored On-Line Exams. *Sensors* 22, 2 (2022). https://doi.org/10.3390/s22020654

[10] Sándor Laki, Péter Mátray, Péter Hága, Tamás Sebők, István Csabai, and Gábor Vattay. 2011. Spotter: A model based active geolocation service. In *2011 Proceedings IEEE INFOCOM*. 3173–3181. https://doi.org/10.1109/INFCOM.2011.5935165

[11] Owen H. T. Lu, Anna Y. Q. Huang, Jeff C.H. Huang, Albert J. Q. Lin, Hiroaki Ogata, and Stephen J. H. Yang. 2018. Applying Learning Analytics for the Early Prediction of Students' Academic Performance in Blended Learning. *Journal of Educational Technology Society* 21, 2 (2018), 220–232. http://www.jstor.org/stable/26388400

[12] Heng Luo, Stevie Rocco, and Carl Schaad. 2015. 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)*. 264–268. https://doi.org/10.1109/EITT.2015.62

[13] Kew Si Na and Zaidatun Tasir. 2017. 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)*. 62–68. https://doi.org/10.1109/LaTiCE.2017.18

[14] Alejandro Peña-Ayala. 2014. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications* 41, 4, Part 1 (2014), 1432–1462. https://doi.org/10.1016/j.eswa.2013.08.042

[15] Cristobal Romero and Sebastian Ventura. 2020. Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery* 10, 3 (2020), e1355. https://doi.org/10.1002/widm.1355 arXiv:https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1355

[16] Yuval Shavitt and Noa Zilberman. 2011. A Geolocation Databases Study. *IEEE Journal on Selected Areas in Communications* 29, 10 (2011), 2044–2056. https://doi.org/10.1109/JSAC.2011.111214

[17] Taeho Yu and Il-Hyun Jo. 2014. 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* (Indianapolis, Indiana, USA) *(LAK '14)*. Association for Computing Machinery, New York, NY, USA, 269–270. https://doi.org/10.1145/2567574.2567594