Xinlei Hong, Shaowei He,Yangzhi Xu, Sicheng Xu

Professor Zhichun Liu

HUDK 4050 Core Methods in Educational Data Mining

10 December 2021

HUDK 4050 Creative EDM Assignment

Our team chose to improve performance on ACA1 for this creative EDM assignment. The purpose of ACA1 was to help Alex choose the best school for her degree in education. The improvements include refining the quality measures and implementing the social network analysis. Data mining and analytics present a huge potential for extracting useful information from large datasets. Such findings can be used to inform certain actions. Data mining and analytics can be applied in education through specific methodologies known as educational data mining and learning analytics. In this research study, a student known as Alex seeks to choose a school based on the degrees offered close to her family's and the level of safety. The goal of the report is to map Alex's school choices by degree, safety, and distance close to her family. The mapping uses the network analysis technique to compare the number of rape cases reported and the city where the university is close to her family. The learning need, in this case, is to ensure that the school that the student chooses to study for a degree is safe from rape incidents. This problem is meaningful because Alex desires to avoid being exposed to a university with high cases of rape and stay close with her family. Even though she prioritizes certain cities close to her family over others, she is also determined to study the degree she will choose in a university that is free from rape incidents. The educational data mining project will find this actionable information from two datasets, namely the university data sets and the crime data. Both data sets

have specific information on the location of specific universities. The researcher will use the network analysis technique to compare the number of rape cases reported with the cities where the universities are located.

<div align="center">A Brief Literature Review</div>

Baker and Yacef (2009) review the historical trends in Educational Data Mining. Educational data mining insists on prediction methods. They apply existing models to create useful and actionable scientific discoveries from educational data sets. The scientific literature on machine learning and data mining has informed the development of educational data mining methods. The works in educational data mining can be categorized into web mining, and statistics, and visualization. Further, the web mining category can be divided into specific techniques for classification, clustering, association rule mining, and outlier detection, text mining, and sequential pattern mining.From a second viewpoint, Baker categorizes educational data mining techniques under prediction, relationship mining, clustering, data discovery, and data distillation classes. Further, the prediction methods include density estimation techniques, regression techniques, and classification techniques. Relationship mining methods include sequential pattern mining, correlation mining, and association rule mining. The most prominent methods for educational data mining are those applied in relationship mining.In the past, research on Educational Data Mining has been applied to determine the association between different causative factors for student failure and the non-retention of students in specific courses. Such student models may also be used to find new data mining rules on student motivation, student attitudes, their current knowledge, and meta-cognition abilities. School instructors can then respond to these individual differences among students to achieve significant

improvement in their learning. The student attributes can be captured in large data sets and modeled in real-time to create higher-level constructs that were not immediately clear at the beginning. Educational data mining methods have also been used in recent years to perform student modeling and predict student failure, student likelihood to drop out of college, and student efficacy.

Educational data mining methods have been used to find empirical evidence needed to refine the well-known educational phenomena and educational theories. In this regard, the information yielded from educational data mining helps educators to design better learning systems. As an example, Heffernan (2009) used educational data mining methods to examine the effect of self-discipline on learning. The study findings showed that self-discipline causes a marginal impact on learning. This finding is important, especially after the same study established that self-discipline correlates to fewer mistakes and higher incoming knowledge. In another study, Madhyastha and Tanimoto (2009) used educational data mining techniques to examine whether there is a relationship between consistency and student performance.

Clear Analysis Plan

The aim of this educational data mining exercise is to map Alex's choices of school by degree, location near to her family, and safety in order to determine a safe university for her. In this study, the researcher will use the network analysis technique to address Alex's problem. The researcher applied network analysis to compare the rates of rape cases to the specific cities where the university is located close to her family. First, the researcher will attach R Software packages from the tidyverse library needed to solve Alex's problem. The rear library will be imported for reading the data contained in the datasets. The researcher will also declare the potential conflicts

within the tidyverse library to avoid discrepancies in the final results. For refining the quality measures part, the team narrowed down Alex's ideal school lists by the following conditions: Safety, urban, diversity, and quality. The researcher will use three data sets to perform data mining. The first data set is the school data set that contains the details of specific universities and their locations. The second data set is the crime data set that contains details of rape cases reported in each city. The third data set is the distance of each main city in the United States. The researcher will take an initial step of data preparation. During dare preparation, the researcher will merge the data from different sources, clean the data to remove any missing values, and arrange the data in descending order. In this specific case, the data column for rape cases will be ordered by starting with the entries with the least number of rape cases to those with the highest number. Under the quality section, not only the highest degree offered was considered this time, ACT scores and SAT scores were both taken into consideration because both scores reflected the overall academic levels were high. On the other hand, both scores also showed the academic atmosphere of the university was very strong. For implementing the social network analysis part, social network graphs were created to visualize the interconnections among the choices of schools by the distance of each city. The researcher applied the Social Network Analysis technique to examine the relationship between the university choices made by Alex due to distance from her family and the choices of the university made on the basis of high safety. The researcher wanted to visualize how the separate data entities are interconnected. In this case, the research used cities that host the universities which are close to her family's city as the nodes. The nodes were conceived as the representation of connected objects to be explored through

network analysis. Equally, the nodes are presented as a larger list that contains additional lists of other cities which show Alex's preferences.

Conclusion

This creative EDM assignment was an overall revision of the team's ACA1 project. The team became more thoughtful about this assignment. The purpose of the assignment was to help Alex narrow the list of schools for her Education degree in a more comprehensive way. More variables were taken into consideration, and more views were adopted. The social network visualized the interconnections among the choices of schools and the interconnections among the locations of the chosen schools. The social networks demonstrated that there are different universities with zero cases of rape which fitted in Alex's list of institutions. Graphs should help Alex understand the interconnections more intuitively. Thanks to Lukas' feedback, the team finished our creative EDM assignment in a comprehensive way.

Reference

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning analytics* (pp. 61-75). Springer, New York, NY.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of educational data mining*, *1*(1), 3-17.

Siemens, G., & Baker, R. S. D. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252-254).