**Reflection Paper**

Before I attended to HUDK 4050, I was from Economic and Education program. I took few courses of data mining for policy evaluation and policy analysis by STATA. However, I always struggled with the data analysis and confused about the programing due to lack of statistic knowledge. Thus, I decided to take HUDK 4050 to learn some fundamental knowledge of data analysis. Moreover, learning how to do the data mining through Python is another purpose of mine. Before I attended to HUDK 4050, I have none background knowledge on how to use Python programming. Since Python is one of the most convenient and useful tool of data mining, I believe it is necessary for me to learn how to do the data analysis by using Python. Furthermore, I used to believe data mining and programming is hard for me to learn. I even don’t have courage to take the programming course. In order to overcome my shortcoming, I push myself to take the HUDK 4050. At the beginning my goal is not only learning some knowledge on data mining but also overcoming myself and providing me some self-confidence on data mining.

I have learned numerous useful knowledge on data mining this semester. At the beginning, I learned how to deal with the tidy data. From the ACA 1, we filtered nearly 4000 data from the data set. Then, I learned how to do the data prediction by using the linear regression model, logistic regression model, classification model and Naïve Bayes model through Python programming. In the ACA2, we built the classifier for off-task behavior in classroom by using the machine learning. Then, I also learn the knowledge of clustering and how to make the k-mean clustering. In the ACA 3, we use cluster to identify the similar colleges and universities. I also learned how to compress the data by using PCA model. That’s one of the most interesting parts for me, since I never consider data also have different dimension. In the end of the semester, I also learned how to create the social network analysis. Due to the social network analysis, we can analyze the centrality and different relationship between each factor. The social network analysis is my favorite part. Since we can build the data set with our group members and find out the interesting correlation between each other through Python programming. The most challenge part of this course for me is how to clean up the tidy data and filter out the dataset we need. At the beginning, my group members and I have less skills on debug and dropping the empty data frame and the missing value. It took us few days to learn from the Pandas library. However, after we overcame this challenge, we feel very fulfilled.

Overall, I really appreciate this course which not only provide me knowledge but also help me built self-confidence on data analysis. I never could imagine I am the person who can overcome plenty of challenges on data mining. Although the process was tough, I still so proud of myself.