I remembered that at the beginning of this course, Prof. Lukas assigned us an introduction about pre-knowledge and expectations. Now, look back after taking this class. I could say I am benefited a lot from this EDM class a lot. I assumed after taking the next HUDK4051 class. I am totally capable to have a data analyzing job in an education company.

Looking through the HUDK 4050 Fall 2021 Schedule, I learned the basic knowledge of EDM, then Lukas taught us the basic data sources and manipulation approaches. Then the core knowledge was introduced from Regression to predict, and we did the first ACA assignment. I remember I was stuck with the Crime_data, because I did not fully understand the ACA1 main problem. There is definitely an easier way to handle it and rank the colleges. Then we learned the basic knowledge about the classification and clustering techniques:

Clustering:

Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large `n_samples`, medium `n_clusters` with *MiniBatch code*	General-purpose, even cluster size, flat geometry, not too many clusters	Distances between points
Affinity propagation	damping, sample preference	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	bandwidth	Not scalable with `n_samples`	Many clusters, uneven cluster size, non-flat geometry	Distances between points
Spectral clustering	number of clusters	Medium `n_samples`, small `n_clusters`	Few clusters, even cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	number of clusters	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints	Distances between points
Agglomerative clustering	number of clusters, linkage type, distance	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints, non Euclidean distances	Any pairwise distance
DBSCAN	neighborhood size	Very large `n_samples`, medium `n_clusters`	Non-flat geometry, uneven cluster sizes	Distances between nearest points

Even though we haven't learned too many techniques but the basic usages are similar but different applications.

Classification:

1. Naive Bayes
2. Logistic
3. Decision Tree
4. SVM(not covered)

Then I learned the basic knowledge about PCA principal component analysis, it is fundamentally a dimensionality reduction algorithm.ICE6 is very detailed about PCA however, I still don't know when to apply it. Then the next important topic is Social Network Analysis. I like this part particularly since it could show interesting figures to show the relations. Then we talked about the privacy issues and ethical concerns in EDM, but what I think is that more private data could contribute the higher accuracy. We just need to protect the data we collected.

Lastly, we did a final project to define problems and solve them by ourselves. Honestly speaking, our analysis is too simple. Can't even put it on the resume. However, the good point is we could use that knowledge in our final project easily like we totally absorbed that knowledge and stored them in our minds.

All in all, I really appreciate the lectures Mr. Lukas gave and I learned a lot about EDM. I look forward to learning more advanced and challenging EDM projects and knowledge.

Again, thank you Lukas.