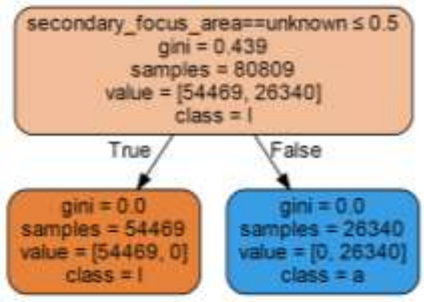
Machine Learning for Public Policy: Unsupervised Learning Analysis

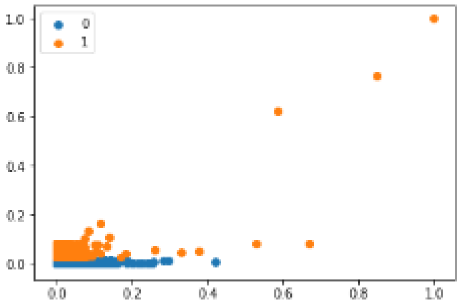
The Prediction of Donor Choices using data from DonorsChoose.org

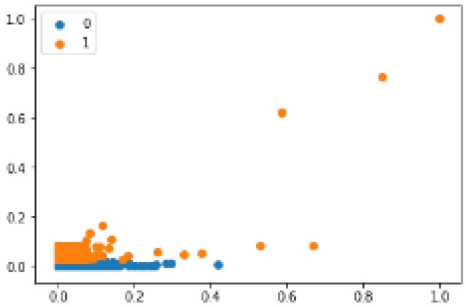
Ta-Yun Yang

After the donor data is pre-processed in the previous assignment, we will like to discover the potential distribution and characteristics of the data before we move on to supervised learning process. Unsupervised learning method might be helpful for us to find out some of the important characteristics which could be useful in supervised learning. To simplify the task, I only select the 3 most frequent categories in every categorical features and transform them to binary features. In the end, I include 2 scaled continuous features and 50 categorical features in the analysis.

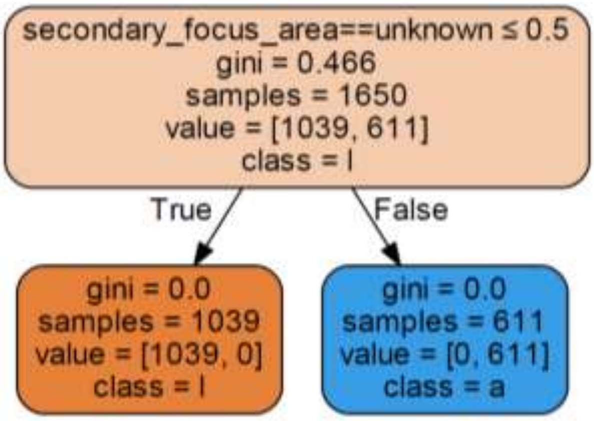
First of all, I use the largest subset of the donor data from 2012-2013 which contains cases within one and half year from the beginning of 2012. I then specify the K-mean approach in this task with 2 clusters (which is the same number as the binary classification we want to do in supervised learning). After comparing the descriptive statistics between groups of different labels, I discover that features like **secondary_focus_area==Literacy & Language, secondary_focus_area==Math & Science, secondary_focus_subject==unknown, secondary_focus_subject==Literature & Writing, secondary_focus_subject==Literacy** are significantly different between groups. Especially, according to the single layer tree plot, we can discover that whether the information of secondary focus area is missing is influential in clustering. Two scatter plots at the right hand side depict the distribution of two continuous variable given labels with different colors. We find out there is almost no different between two plots where they are labelled by the clustering labels and secondary focus area respectively.

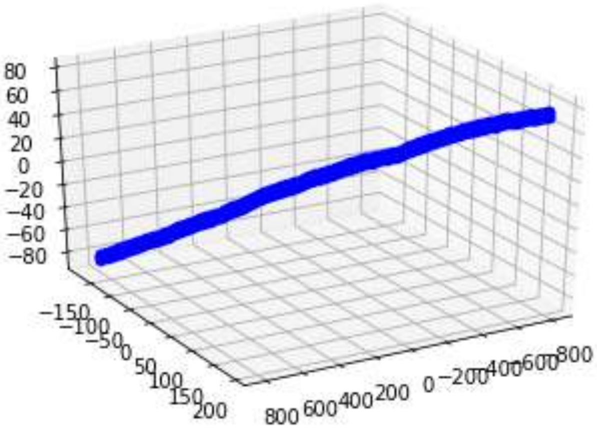






We then move on to explore the distribution of the data which are classified as the 5% riskiest cases. Features like **secondary_focus_area==others, secondary_focus_area==Literacy & Language, secondary_focus_area==Math & Science, secondary_focus_subject==unknown, secondary_focus_subject==Literacy** are significantly different between clusters, and features like **school_metro==urban, school_metro==others, grade_level==others, grade_level==Grades 3-5, grade_level==Grades 6-8, poverty_level==highest poverty** are slightly different between clusters. This is reasonable that the 5% subset has smaller sample size and larger variance comparing to the previous dataset. As previous case, whether secondary focus area is known is the most influential features in clustering. In order to get the general distribution of all the features, I implement Multidimensional Scaling Methods and project 52 features on 3-dimensional graphs where different labels are colored with blue and green. However, since the labelled is perfectly determined by whether secondary focus area is missing, the variation exclusively relies on the non-missing values of secondary focus area (marked as blue in the graph). We can observe a linear combination of three latent features with small variance in the graph, the trend of the data is relatively clear.





Secondary focus area is the most influential factors in clustering, it could possibly be used to improve the classification (further clustering true positive) in our riskiest 5% sub-group.