Johnny Lin – The Data Incubator Proposal

1. What is the motivation for tackling this problem?

**I am interested in predicting with high accuracy whether a person will be diagnosed with autism spectrum disorder as it is one of the highest behavioral disorder affecting 1 out of 68 newborns in the US. Knowing whether someone is likely to be diagnosed with autism spectrum disorder can provide early treatment strategies before they face the barriers of social communication that impact the community participation and employment.**

1. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise?

**The clients would be parents with transition-age youth, health care practitioners, and employers. The goal is that if we can reliably predict whether someone will have higher risk of autism spectrum disorder in ample time, then the parents and health care practitioner can help implement a treatment plan.**

1. What data are you going to use for this? How will you acquire this data?

**Adults:**[**https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult**](https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult)

**Adolescents:**[**https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Adolescent+++**](https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Adolescent+++)

**Child:**[**https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Children++**](https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Children++)

**I will be acquiring these 3 unprocessed data sets and combining into one cohort with 1392 sample size with 21 attributes from the UCI repository.**

1. In brief, outline your approach to solving this problem (knowing that this might change later).

**First I need to figure out how to import the 3 datasets. The format is unstructured so I will need to first figure out how to wrangle the data into structured and clean manners. Next I need to merge the 3 data sets. Following this I will identify the missing values or pre-process the values that don’t make sense to make the better decision on how to deal with these values. Next I will explore the data to look for the relevant features. I will then explore different machine learning algorithms – such as logistic or linear regressions would be appropriate here. I will also explore clustering as I think it is possible there will be ‘clusters’ of people with certain characteristics more likely to be classified as autism spectrum disorder over another.**

1. What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.

**My deliverables are jupyter notebook along with slides outlining the motivation, methods, and results. Additionally, I’d like to create a risk level assessment based on the predicted probabilities of autism spectrum disorder for each person in the dataset.**