IST 707 – Project Proposal

Benjamin Harwood, Kevin Vogel, Kyle Wass, Nina Williams

DATA MINING PROBLEM AND DATA SET:

2015 US Census Demographic Data retrieved from Kaggle at the following link: [https://www.kaggle.com/muonneutrino/us-census-demographic-data/data#](https://www.kaggle.com/muonneutrino/us-census-demographic-data/data)

The data set consists of 74,001 instances with 37 separate variables. The data is organized by a primary key of Census Track ID’s. A typical census track contains approximately 5,000+ residents. Each census track belongs to a county within one of the 50 states. The data contains information pertaining to population, gender, race, voting age, income, poverty levels, professional industries, commuting methods and employment summaries.

EXPERIEMENT DESIGN:

The data selected will be analyzed and interpreted through many different algorithms and data science methods and models. Our team is not going into the research in search of a supervised learning outcome but rather approaching the dataset from an unsupervised learning position. We plan to examine the variables by tuning the data through different levels of discretization and applying those variables to the models and algorithms to find meaningful conclusions to the census data.

Our focus for examining the data will be through models such as Classification, Clustering, Association Rule Mining, Decision Trees and Naïve Bayes computed through RStudio. Since our approach is unsupervised, we may discover through our research that additional datasets may need to be implemented and combined with our selected census dataset to support our findings. Throughout our research, we will be looking for correlations and patterns as well as disconnects and data outliers. Our team is keeping an open mind of what the outcomes may be. Census results and conclusions can often result in controversial assumptions that we are eager and willing to discover.