**Name: Jason Bard**

**Project Title: Determining Autism Diagnosis based on Phenotypic Data**

1. **Introduction**

Autism spectrum disorder, also known as ASD, is characterized by an “impairment in social interaction” and by “repetitive interests and behaviors.” The ABIDE dataset, also known as the Autism Brain Imaging Data Exchange, is a collection of more than 1000 individuals, about half of whom have been diagnosed with autism. This database contains not only scans of their brains, but also phenotypic data, including scores from several tests subjects may have taken during the course of the study. While not all subjects took the same tests, many subjects took similar tests and are of all ages.

1. **Dataset**

While the bulk of the dataset consists of resting-state functional MRI scans, this project will focus on the phenotypic descriptive data of the patients, such as age, gender, diagnosis, IQ, test scores, and other variables. This is stored in a CSV file through the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC) website at <https://www.nitrc.org/frs/downloadlink.php/9108>, although registration through an institution may be required.

1. **Data Preparation**

The first step in data mining is to prepare the data. This involves cleaning and transforming it to ensure that it is in a format that can be easily analyzed. This will involve combining several variables into one variable, inputting values in place of missing values, and turning categorical values into numeric values using one-hot encoding, in addition to other potential feature engineering.

1. **Exploratory Data Analysis**

To analyze the data and predict whether or not any individual patient has autism, we will first perform **exploratory data analysis (EDA)**. This will involve plotting the distribution of all necessary variables, creating scatter plots to visualize pairwise relationships, and creating summary and descriptive statistics for each variable. This may help intuitively identify patterns or relationships that may be pursued further with machine learning.

1. **Modeling**

For modeling, we will select three algorithms to be used in the data set: **k-nearest neighbors**, **decision trees**, and **logistic regression**. The goal of modeling will be to predict whether or not any individual patient has autism using the characteristics of their phenotypic data. We will split the data into training and testing sets, train the model on the training set, and then evaluate the model's performance on the testing set.

1. **Evaluation**

Model evaluation is an important step, and we will calculate performance metrics, such as binary cross-entropy, to determine how well the model performs. In general, a higher accuracy will indicate a better performing model. In addition, precision, recall, and F1 will also be measured in order to determine the most viable model.