Lasso-LDA-based Adult Autism Recognition (Project Proposal)

**Group name: Project 3**

**Group Member: Liuxi Wang (Student ID: 222025005)**

**Shule Ge (Student ID: 222025024)**

**Chenxing Liu (Student ID:222025026)**

**Chun Wang (Student ID:222025007)**

**Xiaoningsi Wang (Student ID:222025002)**

**Zhuang Chen (Student ID:222025021)**

# Project Background

According to the "2021 Blue Book of Developmental Disabilities Rehabilitation Industry", there are currently more than 10 million autistic patients in China, of which about 8 million are over 14 years old; There are only 100,000 teachers engaged in autism rehabilitation, and the number of professional talents is obviously insufficient. For autism, the most critical thing is "early identification". On the one hand, most adults ignore their own emotional problems, do not understand and do not pay attention to them, resulting in delayed diagnosis. On the other hand, the pressure of work in modern society is huge, and it is difficult for young people to have the time and energy to conduct timely diagnosis and treatment offline.

Therefore, Touch Autism launched Autism Apps, a search application that can help people understand themselves early and carry out relevant educational interventions through user behavior, directory, and star rating matching corresponding content functions. However, the market application of Autism Apps is not ideal, through the survey, we found that its root cause is that there are many factors affecting autism, and the platform does not have a good way to classify reasonable people and push effective content and services.

 Available data sets: UCI Database/Autism Screening Adult's dataset

# Problem and Objectives

Problem: 1) Because the cause of autism is still uncertain, and there are many possible influencing factors, It's hard for people to measure whether they have the symptoms of autism. 2) The amount of data extracted from clinical medical trials is small and not representative. The existing literature on the application of autism diagnosis in related fields is less, and the methods used are relatively simple. It is difficult for platform to make diagnosis and treatment plans according to their conditions

Objectives: 1) Help users find out whether there are disease characteristics in time and carry out relevant educational intervention. 2)The platform classifies people with autism and pushes effective content and services.

# Proposed Solution

For the classification model, each patient can be considered as a data source. We first deal with missing and outlier values in the patient data by means of multiple interpolation and AVF algorithms and balance the data using the SMOTE algorithm. The specific data covers the patient's personal information as well as the test results. Then three variable selection methods, Lasso variable selection, optimal subset selection and random forest, were combined with CART, random forest and XG-Boost to construct a nine-group integrated model to classify whether a patient has autism. the Lasso variable selection method was able to be compressed in and LDA gave topic-based topic clustering. K-means was chosen for multi-classification decisions and scatter plots were drawn. Finally, common evaluation metrics of classification models (e.g. check accuracy metrics, AUC, and F1 values and plotted using AUC curves) were used for analysis. A comparison is made as to which combination has higher accuracy and validity.

# Expected Impacts

Autism is a more serious developmental disorder that can seriously affect the quality of life of patients and also have a certain socioeconomic impact. Currently, the number of autistic patients is increasing, but there are many factors that cause autism, autism diagnosis is more difficult, and it is difficult for service platforms to make reasonable population classification and targeted treatment. Therefore, our product aims to accurately screen autism patients, provide reliable theoretical references for clinical diagnosis, and push effective content and services for different patient populations.

# Required Resources

(1)Packages such as numpy, pandas, vim, varSelRF, lattice, Mass, nnet, DMwR, Matrix

(2)Python 3.7

(3)Anaconda 2.0.3

(4)Server with Win11/macOS 12.6

# Timeline

