The data is a subset of data from Dr. Fadi Fayez Thabtah, Department of Digital Technology, Manukau Institute of Technology, Auckland, New Zealand. His abstract is reproduced below.

Abstract: Autistic Spectrum Disorder (ASD) is a neurodevelopment condition associated with significant healthcare costs, and early diagnosis can significantly reduce these. Unfortunately, waiting times for an ASD diagnosis are lengthy and procedures are not cost effective. The economic impact of autism and the increase in the number of ASD cases across the world reveals an urgent need for the development of easily implemented and effective screening methods. Therefore, a time-efficient and accessible ASD screening is imminent to help health professionals and inform individuals whether they should pursue formal clinical diagnosis. The rapid growth in the number of ASD cases worldwide necessitates datasets related to behaviour traits. However, such datasets are rare making it difficult to perform thorough analyses to improve the efficiency, sensitivity, specificity and predictive accuracy of the ASD screening process. Presently, very limited autism datasets associated with clinical or screening are available and most of them are genetic in nature. Hence, we propose a new dataset related to autism screening of adults that contained 20 features to be utilised for further analysis especially in determining influential autistic traits and improving the classification of ASD cases. In this dataset, we record ten behavioural features (AQ-10-Adult) plus ten individuals characteristics that have proved to be effective in detecting the ASD cases from controls in behaviour science.

Below is the DATA DICTIONARY – Feature names and description is as follows

**X** Unique Identifier

**X.1** Unique Identifier

**A01** The answer code of the question based on the screening method used

**A02** The answer code of the question based on the screening method used

**A03** The answer code of the question based on the screening method used

**A04** The answer code of the question based on the screening method used

**A05** The answer code of the question based on the screening method used

**A06** The answer code of the question based on the screening method used

**A07** The answer code of the question based on the screening method used

**A08** The answer code of the question based on the screening method used

**A09** The answer code of the question based on the screening method used

**A10** The answer code of the question based on the screening method used

**Age** Age in Years

**Gender** Male or Female

**Ethnicity** List of common ethnicities in text format

**Jaundic** Whether the case was born with jaundice

**App** Whether the user has used a screening app previously

**Res** The final score obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner

**Rel** Who completed the test. (E.g. Parent, self, caregiver, medical staff, clinician ,etc.)

**Autism** Diagnosed with Autism

Transforming data-

Below are the transformations performed on the data-

1. gender\_DC->transformed to numeric by setting the value 1 for male and 0 for female
2. Ethnicity\_DC -> transformed to numeric by setting values from 1 to 9 for all unique values. Below are the values assigned-
   1. 'Middle Eastern' <- 1
   2. 'South Asian' <- 2
   3. 'Asian' <- 3
   4. 'Hispanic' <- 4
   5. 'Black' <- 5
   6. 'Turkish' <- 6
   7. 'Latino' <- 7
   8. 'Pasifika' <- 8
   9. 'Others' / ‘others’ <- 9

Created a new column ‘Ethnicity\_DC\_missing’ to treat missing values. Value of 1 indicates that the data is missing.

1. Jaundic\_DC is transformed to numeric , 0 indicates ‘no’ and 1 indicates ‘yes’
2. Rel\_DC is transformed to numeric by setting values from 1 to 5 for all unique values.

Below are the values assigned-

* 1. 'Self' <- 1
  2. 'Relative' <- 2
  3. 'Parent' <- 3
  4. 'Others' <- 4
  5. 'Health care professional' <- 5

Created a new column ‘Rel\_DC\_missing’ to treat missing values. Value of 1 indicates that the data is missing.

1. Transformed Autism\_DC to numeric, 0 indicates ‘NO and 1 indicates ‘YES’.