Previously, many researchers had used a different number of datasets to classify Autism. The data set we choose is quite new and not many people work on this. We believe with our work many future researchers will be benefited.

We collected the dataset [1] from the UCI Machine Learning Repository. There are 21 attributes in the dataset. There are in total 704 rows in the dataset. The main aim of the dataset is to distinguish the NON ASD people from, according to the “Class/ASD” column where “NO” was set for NON-ASD and “YES” for ASD. Table 1 presents the description of the columns and Variable type.

TABLE 1: Description of columns

|  |  |  |  |
| --- | --- | --- | --- |
| No | Feature Name | Description | Variable Type |
| 1 | A1\_Score | "The answer code of the question based on the screening  method used" | Binary (0/1) |
| 2 | A2\_Score | The answer code of the question : "I usually concentrate more on the whole picture, rather than the small details." | Binary (0/1) |
| 3 | A3\_Score | The answer code of the question : "I find it easy to do more than one thing at once." | Binary (0/1) |
| 4 | A4\_Score | The answer code of the question : "If there is an interruption, I can switch back to what I was doing very quickly." | Binary (0/1) |
| 5 | A5\_Score | The answer code of the question : "I find it easy to ‘read between the lines’ when someone is talking to me." | Binary (0/1) |
| 6 | A6\_Score | The answer code of the question : "I know how to tell if someone listening to me is getting bored" | Binary (0/1) |
| 7 | A7\_Score | The answer code of the question : "When I’m reading a story I find it difficult to work out the characters’ intentions." | Binary (0/1) |
| 8 | A8\_Score | The answer code of the question : "I like to collect information about categories of things (e.g. types of car, types of bird, types of train, types of plant etc)." | Binary (0/1) |
| 9 | A9\_Score | The answer code of the question : "I find it easy to work out what someone is thinking or feeling just by looking at their face." | Binary (0/1) |
| 10 | A10\_Score | The answer code of the question : "I find it difficult to work out people’s intentions." | Binary (0/1) |
| 11 | age | Age in years | Numeric |
| 12 | gender | Male or Female | Categorical |
| 13 | ethnicity | List of common ethnicities in text format | Categorical |
| 14 | jundice | Whether the case was born with jaundice | Categorical |
| 15 | autism | Whether any immediate family member has a PDD | Categorical |
| 16 | Country\_of\_res | List of countries in text format | Categorical |
| 17 | Used\_app\_before | Whether the user has used a screening app | Categorical |
| 18 | result | The final score obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner | Numeric |
| 19 | age\_desc | Age above 18 or more | Categorical |
| 20 | relation | Parent, self, caregiver, medical staff, clinician ,etc | Categorical |
| 21 | Class/ASD | Whether the case has autism | Categorical |

Three different types of columns presented above. Binary, Categorical and numeric. “age”, “ethnicity” and “age\_desc” contain missing value as .28%,13.49% and 13.49% respectively. Table 2 shows some descriptive statistical analysis for each attribute. From the descriptive statistical analysis, we found that,  
except MDVP.FoHz, HNR, RPDE, and DFA all other attributes contain outliers.  
Fig. 1 demonstrates the distribution of classes for the feature “status” which contains the classes (Healthy (0)  
and PD (1)

From Fig. 1, we can see that in this dataset, among 704 instances, there were 189 instances with ASD and only 515 instances with NON ASD people. That means, only 26.48% people are labeled as ASD and the rest 73.15% people are labeled with NON-ASD; which arises the class imbalance problem for this dataset. Therefore, if we classify all the instances as ASD, still we would get 100% accuracy by doing feature engineering.

REFERENCES

[1] <https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult>