**Background**

Project Source - [Kaggle](https://www.kaggle.com/c/autism-prediction/overview)

This project is a part of the ML Olympiad - an associated Kaggle Community Competitions hosted by ML GDEs or TFUGs, sponsored by Google Developers.

This dataset is composed of survey results for more than 700 people who filled an app form. There are labels portraying whether the person received a diagnosis of autism, allowing machine learning models to predict the likelihood of having autism, therefore allowing healthcare professionals to prioritize their resources.

The data files for this competition were developed synthetically by using CTGan on [Autism Screening on Adults Dataset](https://www.kaggle.com/andrewmvd/autism-screening-on-adults).

About Autism

It is mostly influenced by a combination of genetic and environmental factors. Because autism is a spectrum disorder, each person with autism has a distinct set of strengths and challenges. The ways in which people with autism learn, think and problem-solve can range from highly skilled to severely challenged.

Research has made clear that high quality early intervention can improve learning, communication and social skills, as well as underlying brain development. Yet the diagnostic process can take several years.

About AQ1 - 10 Screening Test

According to the National Institute of Health Care Excellence (NICE), the autism spectrum quotient (AQ-10) tool is recommended for use with adults with possible autism who do not have a moderate or severe learning disability. This may help identify whether an individual should be referred for a comprehensive autism assessment.

**Objective**

Predict the likelihood of an individual having autism.

**Evaluation**

The evaluation metric for this competition is AUC-ROC Score. The AUC-ROC Score is commonly used for classification tasks. It measures the degree or measure of separability between the classes.

The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis.

TPR or True Positive Rate can be defined as (True Positives / (True Positives + False Negatives))

Similarly FPR, or False Positive Rate, can be defined as (False Positives / (False Positives + True Negatives))

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings

**Give Feature Information**

1. ID - ID of the patient
2. A1\_Score to A10\_Score - Score based on Autism Spectrum Quotient (AQ) 10 item screening tool
3. age - Age of the patient in years
4. gender - Gender of the patient
5. ethnicity - Ethnicity of the patient
6. jaundice - Whether the patient had jaundice at the time of birth
7. autism - Whether an immediate family member has been diagnosed with autism
8. contry\_of\_res - Country of residence of the patient
9. used\_app\_before - Whether the patient has undergone a screening test before
10. result - Score for AQ1-10 screening test
11. age\_desc - Age of the patient
12. relation - Relation of patient who completed the test
13. Class/ASD - Classified result as 0 or 1. Here 0 represents No and 1 represents Yes. This is the target column, and during submission submit the values as 0 or 1 only.

**Plan of Action**

L1 Data Assessment Methods:

Training:

Train on AUC-ROC Score

Submission:

The sample submission contains ID and class. ID is the identification of an individual who has taken the survey and the class is the class prediction by the algorithm. Once we train the model, we need to use the test file to test the accuracy and then submit after deleting all features except those mentioned in the sample submission file.

**Level 1 - Data Assessment - General Quality and Tidiness Assessment**

Train Data Basic Details:

1. Memory - Train is just 0.08 MB
2. Observations - 800
3. Features - 22
4. Unit of an observation - An individual
5. Target Variable - Class/ASD
6. Problem Type - Classification

Test Data Basic Details:

1. Memory - Train is just 0.02 MB
2. Observations - 200
3. Features - 21

Sample Submission Data Basic Details:

1. Memory - Train is just 0.00 MB
2. Observations - 200
3. Features - 2 - ID, Class/ASD

Quality Issues

1. Age
   1. It has at least 6 decimal points - Reduce it to 1 decimal point
2. Result
   1. The result column is of float type - Ensure result is rounded to 6 decimal places
3. Convert to lower-case
   1. Ethnicity
4. Missing Values are depicted by ?
   1. Replace the missing values in ethnicity (18.88%) and relation (9.6%) with the word Unknown
5. Change the column names
   1. The ‘austim’ column is not descriptive and there is a typo error - Change it to ‘fam\_history\_autism’
   2. The ‘contry\_of\_res’ column is not descriptive and there is a typo error - Change it to ‘Country’
   3. Non descriptive column header result - Change it to ‘final\_score’
   4. Non descriptive column header relation - Change it to ‘survey\_done\_by’
6. Drop columns
   1. The AQ1-10 screening test recommended for adults without moderate or severe learning disability. Hence this model is useful only for adults. Also, the participants in the dataset are also adults - Drop age\_desc feature

Tidiness Issues - None

Feature Types

1. Quantitative - 8
   * 1. ['gender', 'ethnicity', 'jaundice', 'austim', 'contry\_of\_res', 'used\_app\_before', 'age\_desc', 'relation']
2. Qualitative - 14
   1. ['ID', 'A1\_Score', 'A2\_Score', 'A3\_Score', 'A4\_Score', 'A5\_Score', 'A6\_Score', 'A7\_Score', 'A8\_Score', 'A9\_Score', 'A10\_Score', 'age', 'result', 'Class/ASD']
3. Feature Category
   1. ID - Interval
   2. A1\_Score - A10\_Score - Nominal
   3. Gender - Nominal
   4. Ethnicity - Nominal
   5. Jaundice - Nominal
   6. Fam\_history\_autism - Nominal
   7. Country - Nominal
   8. Used\_app\_before - Nominal
   9. Survey\_done\_by - Nominal
   10. Class/ASD - Nominal
   11. Age - Ratio
   12. Final\_score - Ratio

Insights after Level 1 Assessment:

1. There are 8 quantitative columns and 14 qualitative columns
2. There are no duplicate rows and columns
3. The functional relationship between the AQ scores and the result is unknown. Even though on average, the result score above 6 is considered autistic, there are values below 6 as well that are classified autistic. Similarly, the result score seems very noisy for people classified as non autistic. Hence, there are other factors that are contributing to whether the person is autistic or not.
4. There are no NaN values. However, missing values are encoded as ‘?’ in the ‘ethnicity’ and ‘relation’ features
5. 0 Values are not indicators of missing values but are indicators of class values for A1\_Score to A10\_Score and Class/ASD
6. The missing values are occurring at random
7. The scores from A1 to A10 and the final result scores have positive values only - 0 and 1
8. Most of the participants have themselves filled out the form (77%)
9. The class proportion of people with autism and people without is disproportionate. About 77% do not have autism and only 23% have autism
10. Most users have not used the app before
11. About 18% of the users are from USA
12. Most participants do not have relatives with autism (85%)
13. 18% of ethnicity values are missing
    1. 27 from UAE
    2. 25 from India
    3. 22 from New Zealand
    4. 14 from USA
14. The gender participation is quite balanced (52% female and 48% male)
15. Most participants are in the age group of 15 to 30
16. The higher outliers in age feature are above 53.96. There are 51 such columns. These need not be removed since these cannot be considered as edge cases

**Level 1 - Data Cleaning**

Insights after Level 1 Cleaning:

1. Points for Feature Engineering -
   1. Contry\_of\_res - ['Burundi', 'Lebanon'] that are in test set are not present in train set
2. Feature Category
   1. Discrete
      1. ID - Interval
      2. A1\_Score - A10\_Score - Nominal
      3. Gender - Nominal
      4. Ethnicity - Nominal
      5. Jaundice - Nominal
      6. Fam\_history\_autism - Nominal
      7. Country - Nominal
      8. Used\_app\_before - Nominal
      9. Survey\_done\_by - Nominal
      10. Class/ASD - Nominal
   2. Continuous
      1. Age - Ratio
      2. Final\_score - Ratio

Output:

1. Encoded data - l1\_clean\_encoded.csv
2. Unencoded data - l1\_clean\_unecoded.csv

**Level 2 - Data Assessment - Relationships between variables**

Correlation Matrix:

1. Target Variable Class/ASD -
   1. Closest Intuitive Predictor - Final score - 0.5
   2. Scores - A9, A6, A4, A3 - 0.5+
   3. Fam\_history\_autism - A weak correlation of 0.38
   4. Gender - Negative weak correlation
2. Final score -
   1. A9, A5, A4 - 0.5+
   2. Fam\_history\_autism - A weak correlation of 0.32
3. Fam\_history\_autism - Weak correlation between scores
4. A1\_Score to A1\_Score - 0.5+
   1. Score to score correlation
      1. A3 - A4, A9
      2. A4 - A3, A5, A9
      3. A5 - A4
      4. A6 - A9
      5. A9 - A6, A4, A3
   2. A1 and A8 scores are weakly correlated with all other scores
5. Age is not linearly correlated with any other features

Questions and observations:

1. What does the class balance look like?
   1. The classes are unbalanced - About 77% of the participants do not have autism
2. How are age distributed in the dataset?
   1. There is nothing mentioned about the age limits of the participants of the test. However, according to the research I have done, the way in which the test is conducted is different for children and adults.
   2. In this dataset, there are 153 participants who are below 18, starting from 10 year olds. Most of the participants in this category are adolescents.
3. What is the distribution of people who have taken the tests for the participants?
   1. An overwhelming majority of participants have taken the test themselves
4. Divide the participants between who has taken the test for them and if they had autism
   1. Have autism
      1. 20% of the participants who has autism has taken the test by themselves
      2. 9% by parent
   2. Does not have autism
      1. 74% took the test themselves
      2. 5% by parent
5. What is the relationship between the people who have taken the test and the age of the participants?
   1. A wide range of people of all ages have taken the test by themselves - from 10 years to 70+
   2. Some participants of age ranging from mid 20’s to mid 30’s have taken the help of a healthcare professional, parent or relatives
   3. Unknown helpers and Others have helped participants whose ages mostly range from late teens to early 30’s
6. What does the plot of final\_scores look like?
   1. The scores range from -2 to 14
   2. Most scores are below 8
7. How are final scores distributed over the two classes?
   1. Class 0
      1. Ranges from -2.5 to 13
      2. Most of the values lie in between 4 and 8
   2. Class 1
      1. Ranges from -2.8 to 13
      2. Most values lie between 9 and 13
8. 10. How many participants have a family history of autism?
   1. 117 - 15%
9. When there is a family history of autism, people could be concerned about its inheritance. Of all the people who took the test, who took the test for people with a family history of autism?
   1. The majority of the participants took the test by themselves
10. How are final\_scores distributed between classes with respect to family history of autism and whether they have autism or not?
    1. There are no significant patterns here
11. How is the frequency of jaundice distributed among Class/ASD?
    1. There are no relationships
12. How is autism distributed across the age classes?
    1. Age class 4 (Ages 33 to 71) has the highest number of autism cases
    2. There are 70 people in this group, 37% of all the people who have autism
    3. These ages are quite varied - 9 standard deviations
13. When we encode the age\_class and recreate the correlation map, is there a difference in the correlations?
    1. The age class has weak correlation with any other variable

Insights:

1. 77% of the participants have not been diagnosed with autism. Therefore, to understand the characteristics of someone with autism, there is very little information to work with
2. People from ages 10 to 70+ have participated in this test.
3. 74% of the participants took the test by themselves
4. In the case of people diagnosed with autism, most values lie between 9 and 13

**Level 2 - Data Cleaning**

Cleaning Steps - Issue, Remedy, Result

**Level 3 - Data Assessment - Baseline Model (OLS or Logistic Regression)**

**Level 3 - Data Cleaning**

Cleaning Steps - Issue, Remedy, Result

**Level 4 - Data Assessment - Random Forest Based Assessment**

**Level 4 - Data Cleaning**

Cleaning Steps - Issue, Remedy, Result