Group members

Name	Sec	BN
Habiba Mohamed Hanafy	1	25
Zeiad Ayman Mohammed	1	31
Omar Ayman Mohammed	2	2
Yara Hossam El-Din Mostafa	2	48

Problem we're tackling

Autistic spectrum disorder (ASD) is a common neurodivergence disorder among kids and adults. It affects the way the brain functions and represents a huge hurdle in everyday tasks. Early diagnosis of ASD is a major key to help patients lead an easy life without the pressure of blending in with society. Diagnosis can take a long time and extensive therapy sessions. It can be an exhausting process for both parents and kids. Our objective is to create a prediction model for ASD in toddlers, in an ambitious trial to lessen the burden of the lengthy diagnosis process.

Dataset

The dataset we're going to use is this one that's collected through a survey performed by "**Fayez Thabtah**" on a screening app for autism.

The dataset contains 17 features in total (some will be disregarded later on) and one label to classify whether the patient is autistic or not. The number of records is 1054.

Features of interest include 10 answers to behavioral questions that are usually used to diagnose autistic kids along with 5 other individual characteristics that were proven before to help in the diagnosis process (age-gender-ethnicity-jaundice-whether family members have ADS).

Preprocessing

The preprocessing needed for this dataset seems very basic and minimal as it doesn't contain any missing values. It doesn't need segmentation. But we need to use label encoding to quantify our categorical features for example. We might also need to explore more for any misspellings or duplicates.

Feature Extraction

There seems to be two irrelevant features in our data such as:
- the case number of the patient
- the person who completed the survey for the toddler (family, healthcare staff or other)

We plan on dropping these two columns from our dataset.
Also one feature seems redundant (score) as it's used directly to classify the patients. So we might need to drop it in case of overfitting.

Proposed Solution Techniques

The problem is vividly a classification one, we plan on using multiple classification techniques then compare the results and scores of our models to choose the best one for our data.

Classifiers we plan on trying are:
- Logistic regression
- Knn classifier
- SVM
- Decision trees
- Random forest