Problem Statement

Autism Spectrum Disorder Screening using Machine Learning

Introduction:

Autism Spectrum Disorder (ASD) is a neurodevelopmental 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.

About this Data

Description

Domain: Healthcare

Hence, we propose a new dataset related to Autism screening of adults that contains 20 features to be utilised for further analysis especially in determining influential Autism 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.

In recent times, the application of Machine Learning to cross-disciplinary subjects have been very active and successful, especially in the fields of biology and neurology.

Many researchers are interested in creating computational frameworks for automatically generating patterns and trends in large medical data-sets. A learned data representation can help visualise data to assist humans in clinical decision making and predict a target variable from a set of input features.

The data selected for this project is the Autism Spectrum Disorder (ASD) screening data for adults. ASD refers to several related disorders that normally begin in childhood and continue in adulthood. There is no cure for ASD, but treatments can help to improve symptoms. As per HSE (2017), the symptoms can include:

- Social interaction where it is difficult to understand situations and other people's feelings and emotions.
- Difficulties to communicate, which can involve delayed language development, also not being able to take part in conversations properly.
- Unusual physical behaviour such as doing repetitive physical movements, which becomes a routine, then the behaviour becomes routine and the individual can get upset if the routine is disrupted.
- The ASD symptoms can vary from person to person, and it can classify in three main types. The most typical type is "Autism disorder", followed by "Asperger syndrome" and "pervasive developmental disorder" (PDD).
- The third one is also known as 'atypical Autism '. ASD are estimated to affect 1 in every 100 children and boys are more likely to develop ASD than girls by four times.

Pursuing such research necessitates working with datasets that record information related to behavioural traits and other factors such as gender, age, ethnicity, etc. 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. At present, very limited autism datasets associated with clinical or screening are available and most of them are genetic in nature. These data are extremely sensitive and hard to collect for social and personal reasons and the regulations around them.

Problem Statement:

Data Type: Multivariate OR Univariate OR Sequential OR Time-Series OR Text OR Domain-Theory

Nominal / categorical, binary and continuous

Task: Classification

Attribute Type: Categorical, continuous and binary

Area: Medical, health and social science

Format Type: Non-Matrix

Does your data set contain missing values? Yes

Number of Instances (records in your data set): 704

Number of Attributes (fields within each record): 21

Attribute Information:

- **Age**: Number Age in years
- **Gender:** String Male or Female
- **Ethnicity:** String List of common ethnicities in text format
- **Born with jaundice Boolean (yes or no):** Whether the case was born with jaundice.
- **Family member with PDD Boolean (yes or no):** Whether any immediate family member has a PDD.
- **Who is completing the test:** String Parent, self, caregiver, medical staff, clinician ,etc.

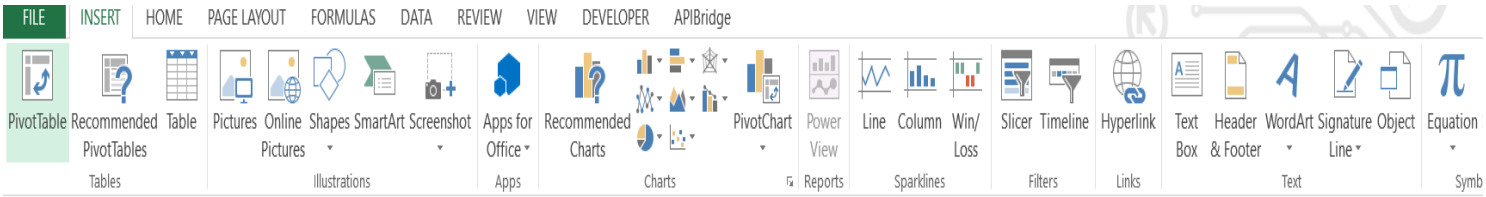
- **Country of residence:** String List of countries in text format
- **10 Questions:** The answer code (0, 1) of the question based on the screening method used.
- **Screening Score**: Integer

Attribute Information:

Table 1: Features and their descriptions

Attribute	Type	Description
Age	Number	Age in years
Gender	String	Male or Female
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean (yes or no)	Whether the case was born with jaundice
Family member with PDD	Boolean (yes or no)	Whether any immediate family member has a PDD
Who is completing the test	String	Parent, self, caregiver, medical staff, clinician ,etc.
Country of residence	String	List of countries in text format
Used the screening app before	Boolean (yes or no)	Whether the user has used a screening app
Screening Method Type	Integer (0,1,2,3)	The type of screening methods chosen based on age category (0=toddler, 1=child, 2= adolescent, 3= adult)
Question 1 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 2 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 3 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 4 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 5 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 6 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 7 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 8 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 9 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 10 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Screening Score	Integer	The final score obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner

Here is the preview of the CSV file:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	A1_Score	A2_Score	A3_Score	A4_Score	A5_Score	A6_Score	A7_Score	A8_Score	A9_Score	A10_Score	age	gender	ethnicity	jaundice	autism	country_o	used_app_	result	age_desc	relation	Class_ASD
2	1	1	1	1	0	0	1	1	0	0	26	f	White-Eur	no	no	'United Sta	no	6	'18 and mc	Self	NO
3	1	1	0	1	0	0	0	1	0	1	24	m	Latino	no	yes	Brazil	no	5	'18 and mc	Self	NO
4	1	1	0	1	1	0	1	1	1	1	27	m	Latino	yes	yes	Spain	no	8	'18 and mc	Parent	YES
5	1	1	0	1	0	0	1	1	0	1	35	f	White-Eur	no	yes	'United Sta	no	6	'18 and mc	Self	NO
6	1	0	0	0	0	0	0	1	0	0	40	f	?	no	no	Egypt	no	2	'18 and mc	?	NO
7	1	1	1	1	1	0	1	1	1	1	36	m	Others	yes	no	'United Sta	no	9	'18 and mc	Self	YES
8	1	1	1	1	0	0	0	0	1	0	64	m	White-Eur	no	no	'New Zeala	no	5	'18 and mc	Parent	NO
9	1	1	0	0	1	0	0	1	1	1	29	m	White-Eur	no	no	'United Sta	no	6	'18 and mc	Self	NO
10	1	1	1	1	0	1	1	1	1	0	17	m	Asian	yes	yes	Bahamas	no	8	'18 and mc	'Health ca	YES
11	1	1	1	1	1	1	1	1	1	1	33	m	White-Eur	no	no	'United Sta	no	10	'18 and mc	Relative	YES
12	1	0	0	0	0	0	1	1	0	1	17	m	?	no	no	Austria	no	4	'18 and mc	?	NO
13	1	0	0	0	0	0	1	1	0	1	17	f	?	no	no	Argentina	no	4	'18 and mc	?	NO
14	1	1	0	1	1	0	0	1	0	1	18	m	'Middle Ea	no	yes	'New Zeala	no	6	'18 and mc	Parent	NO
15	1	0	0	0	0	0	1	1	1	1	31	m	'Middle Ea	no	no	Jordan	no	5	'18 and mc	Self	NO
16	1	1	1	1	0	0	0	1	0	0	43	m	?	no	no	Lebanon	no	5	'18 and mc	?	NO
17	1	0	0	1	1	1	1	1	0	1	37	f	White-Eur	no	yes	'United Sta	no	7	'18 and mc	Self	YES
18	1	1	0	0	0	0	1	0	0	1	55	f	Others	no	no	'New Zeala	no	4	'18 and mc	Self	NO
19	1	1	1	1	1	1	1	1	1	1	18	f	White-Eur	yes	no	'South Afri	no	10	'18 and mc	Self	YES
20	1	1	1	1	1	1	1	1	1	1	18	f	White-Eur	no	no	'South Afri	no	10	'18 and mc	Self	YES
21	1	0	1	1	1	1	0	0	1	0	34	f	White-Eur	no	no	'New Zeala	no	6	'18 and mc	Self	NO
22	1	0	0	1	1	1	1	0	1	1	53	f	White-Eur	no	no	'New Zeala	no	7	'18 and mc	Self	YES
23	1	0	1	1	0	1	1	1	1	1	35	f	White-Eur	no	yes	'United Sta	no	8	'18 and mc	Self	YES
24	1	0	1	1	1	0	1	1	0	1	20	f	Latino	yes	no	Italy	no	7	'18 and mc	Self	YES
25	1	1	1	1	1	1	0	1	0	1	27	f	White-Eur	no	no	'New Zeala	no	8	'18 and mc	Self	YES
26	1	0	1	1	1	1	0	1	1	0	53	f	White-Eur	no	no	'New Zeala	no	7	'18 and mc	Relative	YES
27	1	1	1	1	0	1	0	0	0	0	24	f	Pasifika	no	no	'New Zeala	no	5	'18 and mc	Relative	NO
28	1	1	0	0	0	1	1	1	0	1	30	f	Asian	no	no	Banglades	no	6	'18 and mc	Self	NO
29	1	0	0	1	0	0	0	1	0	0	21	f	Latino	no	no	Chile	no	3	'18 and mc	Self	NO
30	1	1	1	1	1	1	1	1	1	1	35	f	Black	no	no	France	no	10	'18 and mc	Parent	YES

Task to Do:

With the available ASD data on individuals our goal is to make predictions regarding new patients and classify them into one of two categories: “patient has ASD” or “patient does not have ASD ”.

In other words, we are working on a binary classification problem with the ultimate goal of being able to classify new instances, i.e.
When we have a new adult patient with certain characteristics we would like to able to predict whether or not that individual has high probability of having ASD.

We will use supervised machine learning to refer to creating and using models that are learned from data, i.e., there is a set of data labelled with the correct answer for the model to learn from.
Also apply a feature selection algorithm to figure out which of the 20 variables are most important in determining whether an individual has ASD or not.

This work aims to explore several competing supervised machine learning classification techniques namely:

- Decision Trees
- Logistic Regression

In addition to being able to predict whether an individual with the given characteristics will have ASD or not, we would also like to be able to identify the most influential autistic traits. The hope is to identify individuals who have a high chance to be diagnosed with ASD and provide them with relevant treatment, therapy and counselling in a time sensitive fashion.

1. Import the required libraries and read the Dataset.

```
In [40]: dataset = pd.read_csv('dataset.csv')
         data = dataset.dropna()
         data.columns = ['A1_Score', 'A2_Score', 'A3_Score', 'A4_Score', 'A5_Score', 'A6_Score', 'A7_Score', 'A8_Score', 'A9_Score', 'A10_
         data.head()
```

Out[40]:

	A1_Score	A2_Score	A3_Score	A4_Score	A5_Score	A6_Score	A7_Score	A8_Score	A9_Score	A10_Score	...	gender	ethnicity	jaundice	autism	cour
0	1	1	1	1	0	0	1	1	0	0	...	f	White-European	no	no	'Ur
1	1	1	0	1	0	0	0	1	0	1	...	m	Latino	no	yes	
2	1	1	0	1	1	0	1	1	1	1	...	m	Latino	yes	yes	
3	1	1	0	1	0	0	1	1	0	1	...	f	White-European	no	yes	'Ur
4	1	0	0	0	0	0	0	1	0	0	...	f	?	no	no	

5 rows × 21 columns

```
In [41]: data.describe()
```

Out[41]:

	A1_Score	A2_Score	A3_Score	A4_Score	A5_Score	A6_Score	A7_Score	A8_Score	A9_Score	A10_Score	result
count	704.000000	704.000000	704.000000	704.000000	704.000000	704.000000	704.000000	704.000000	704.000000	704.000000	704.000000
mean	0.721591	0.453125	0.457386	0.495739	0.498580	0.284091	0.417614	0.649148	0.323864	0.573864	4.875000
std	0.448535	0.498152	0.498535	0.500337	0.500353	0.451301	0.493516	0.477576	0.468281	0.494866	2.501493
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	3.000000
50%	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	1.000000	4.000000
75%	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	7.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	10.000000

2. Data preparation & exploratory data analysis

Before data can be used as input for machine learning algorithms, it must be cleaned, formatted, and maybe even restructured — this is typically known as pre-processing. Unfortunately, for this dataset, there are many invalid or missing entries.
We must deal with, moreover, there are some qualities about certain features that must be adjusted. This pre-processing can help tremendously with the outcome and predictive power of nearly all learning algorithms.

3. Import the label Encoder, LabelEncoder can be used to normalize labels. It can also be used to transform non-numerical labels (as long as they are hashable and comparable) to numerical labels.

4.Split the Dataset into Train and Test

Split Your Dataset With scikit-learn's train_test_split()

- The Importance of Data Splitting. Training, Validation, and Test Sets.
- Prerequisites for Using train_test_split()
- Application of train_test_split()
- Supervised Machine Learning With train_test_split()
- Other Validation Functionalities.

5.Cleaning and preprocessing of data for training a classifier, Data preprocessing is crucial in any data mining process as they directly impact success rate of the project. This reduces complexity of the data under analysis as data in real world is unclean.

6.Import roc_curve, auc, confusion_matrix, classification_report, Accuracy_score,

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. ... The rows represent the predicted values of the target variable.

ROC stands for curves receiver or operating characteristic curve. It illustrates in a binary classifier system the discrimination threshold created by plotting the true positive rate vs false positive rate.

Recall is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives.

Accuracy represents the number of correctly classified data instances over the total number of data instances. In this example, Accuracy = (55 + 30)/(55 + 5 + 30 + 10) = 0.85 and in percentage the accuracy will be 85%

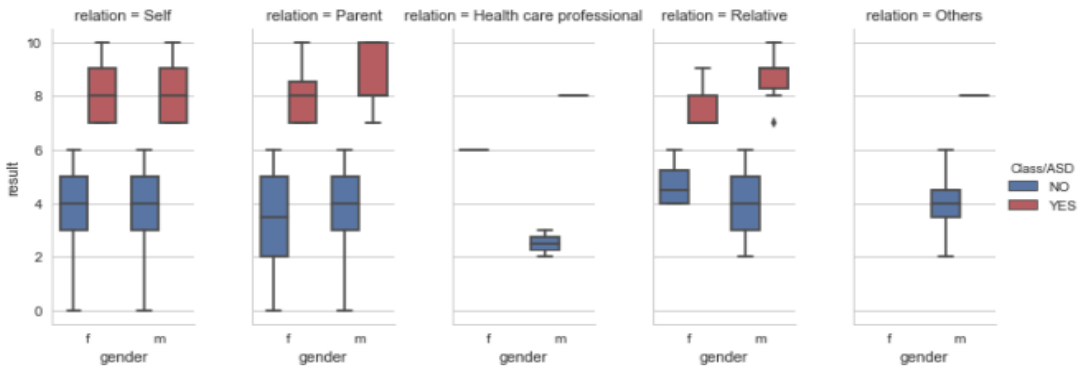
7. Now you Must have to Plot Roc Curve and An **ROC curve** (receiver operating characteristic **curve**) is a **graph** showing the performance of a classification model at all classification thresholds. This **curve** plots two parameters: True Positive Rate. False Positive Rate

8.Calculate the fpr and tpr for all the thresholds of the classifications,

The **TPR** defines how many correct positive results occur among all positive samples available during the test. **FPR**, on the other hand, defines how many incorrect positive results occur among all negative samples available during the test.

The true positive rate will be 1 (**TPR** = TP / (TP + FN) but FN = 0, so **TPR** = TP/TP = 1) The false positive rate will be 1 (**FPR** = FP / (FP + TN) but TN = 0, so **FPR** = FP/FP = 1)

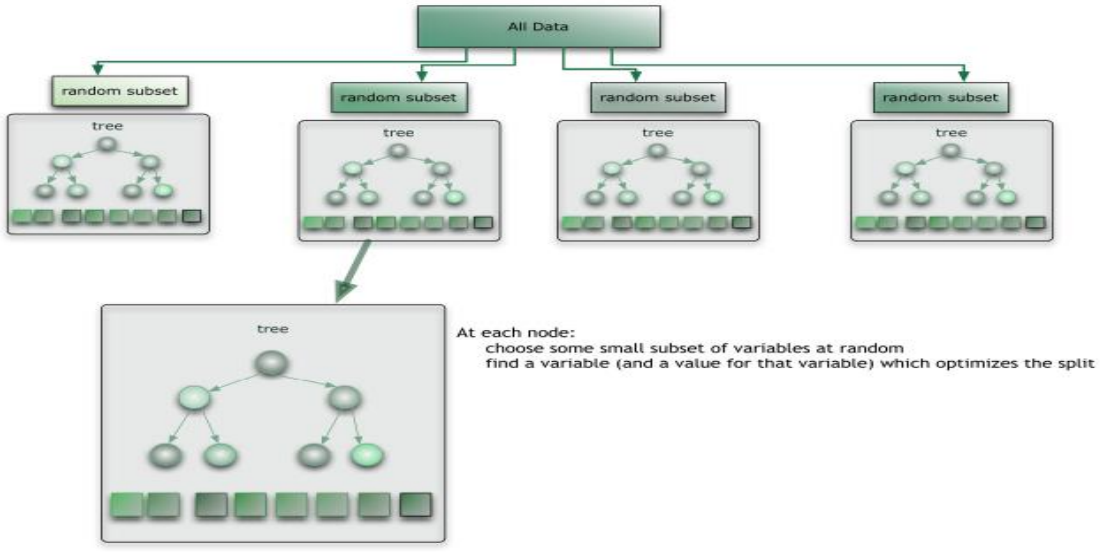
9. Exploratory Visualization: Before proceeding to apply any algorithm, we take a moment to visualize the ASD data set using the Seaborn module of Python



10. Algorithms and Techniques:

Decision Trees: A Decision Tree uses a tree structure to represent a number of possible decision paths and an outcome for each path. It can also perform regression tasks and they form the fundamental components of Random Forests which will be applied to this data set in the next section.

A Decision Tree model is a good candidate for this problem as they are particularly adept at binary classification, however it may run into problems due to the high number of features so care will have to be taken with regards to feature selection. Due to these advantages and the ease of interpretation of the results, make the use of Decision Tree Classifier as the benchmark model



11. Logistic Regression: Logistic regression is a classic predictive modelling technique and still remains a popular choice for modelling binary categorical variables.

Model Evaluation:

Output:

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 15)                285
_________________________________________________________________
dropout (Dropout)            (None, 15)                0
_________________________________________________________________
dense_1 (Dense)              (None, 15)                240
_________________________________________________________________
dropout_1 (Dropout)          (None, 15)                0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 16
=================================================================
Total params: 541
Trainable params: 541
Non-trainable params: 0
_________________________________________________________________
```

```
plot_roc(classifier, X_test, y_test)
```