Tali Aharon 034791236

Data Science Fundamentals (90001) - Final Project - Report

ASD (Autism Spectrum Disorder) - Prediction based on Phenotypic Data from ABIDE II Dataset

Introduction:

Disclaimer -
This project is part of a larger research Helit Bauberg and I are doing on Autism Spectrum Disorder diagnosis from MRI scans. The first part of this research (submitted to Prof. Arriel Benis) was written in Orange and included an EDA of the dataset. The current project has different goals and is written in Python. Nonetheless, as I am using the same dataset - there could be some similarities between both project reports on the EDA part.

Goals

The goal of the full research is to assist in early diagnosis of ASD (Autism Spectrum Disorder).
Early diagnosis is crucial to ASD treatment (bearing in mind the high neuroplasticity at younger ages).
As babies can have harmless MRI scans (while asleep) - prediction of ASD from MRI data would be better suited for early diagnosis than the diagnosis made nowadays from many different kinds of tests done at later ages (tests which I'll describe shortly in the *Data* section).

In the current project, my goal is to predict ASD from the phenotypic data.
I wish to evaluate the accuracy one can get by using only this data (in order to compare it to the prediction scores we will get from the MRI scans, later in our research).

I propose the following framework:

1. EDA: ABIDE II is the largest and most global dataset for ASD research. It includes 1,152 subjects (about half of them with ASD. It contains sMRI data, rs-fMRI data, and phenotypic data. In this project I would like to explore the phenotypic data files, which contain a large set of test scores used by psychiatrists, psychologists, pediatrics etc. to diagnose ASD.
2. Feature Selection and Classification: I would like to go over the large amount of features on the phenotypic data files and select, by several different manners, only the ones that will help me get the best scores on the Classification models I intend to apply. I will train the classification models, and then try to predict for each record if the subject has ASD or not.
3. Dimensionality Reduction and Clustering: As an alternative to the supervised classification models, I will try and see if I can get better results by applying some "brute force" on the same data - using only PCA (Principal Component Analysis) and Clustering.

Data

The data related to the subjects in the ABIDE II dataset is split into several files.
I downloaded them from: https://www.nitrc.org/frs/downloadlink.php/9108

The files are grouped as follows:

- **Phenotypic data -** 2 CSV data files and one Data legend file (which may be found in this GoogleDrive link)

    1. ABIDEII_Composite_Phenotypic: My main data source for subject classification, containing 1114 subjects.
    2. ABIDEII_Long_Composite_Phenotypic: Includes data related to additional 38 subjects. All subjects were scanned twice a few years apart (baseline + followup scans). I added all baseline phenotypic data of these 38 subjects to the 1114 subjects from the first file (resulting in 1152 subjects for my analysis).
    3. ABIDEII_Data_Legend: showing per each feature its label, type, description and range of values.

- **MRI related data files:** Will not be used in the scope of this project.

The Phenotypic data includes tests for ASD diagnosis alongside tests from other domains with some relevance to ASD. To get more knowledge of the domain, I discussed the relevance of each test with a child psychologist specialising in ASD diagnosis. I also read several articles discussing the different tests. A full description of each feature/test, its relevance to ASD and supporting links can be found in this GoogleDrive link under ABIDEII_Data_Legend.xlsx

The features/tests that I found to be the most relevant to ASD are the following:

ADI-R - The **Autism Diagnostic** Interview-Revised (ADI-R) is a structured interview conducted with the parents of individuals who have been referred for the evaluation of possible autism or autism spectrum disorders.

ADOS_G and **ADOS_2** - The **Autism Diagnostic** Observation Schedule (ADOS) is considered "gold standard" assessment measures in the evaluation of ASD. It is a semi-structured, standardized assessment of communication, social interaction, play, and restricted and repetitive behaviors.

AQ - The **Autism Spectrum** Quotient (AQ) is a 50 item self-report measure used to assess traits of autism in adults and adolescents aged 16 years and over. The measure is suitable for men and women who have normal intellectual functioning. The AQ measures 5 symptom clusters important in understanding the profile of strengths and weaknesses for individuals with Autism: social skills, attention switching, attention to detail, communication and imagination.

BASC - The **Behavior** Assessment System for Children, Second Edition (BASC-2) is a commonly used behavior rating scale.

BRIEF - The **Behavior** Rating Inventory of Executive Functions (BRIEF) screens for **executive function** deficits in 5- to 18-year-olds. Many people with autism have difficulty with executive functioning. They may have trouble with certain skills like planning, staying organized, sequencing information, and self-regulating emotions.

CASI - A 37-item instrument in Hindi with dichotomous yes/no responses [Chandigarh **Autism Screening** Instrument (CASI)] was developed to be applied on children aged 1.5-10 yr.

CBCL - The Child **Behavior** Checklist (CBCL) is a well established and widely used parent-completed measure of emotional, behavioral, and social problems in children aged 1.5–5 years and 6–18 years. It was developed to assess a range of different behavior problems, but the test developer proposed that CBCL is also useful for ASD-specific screening within clinical settings.

CELF - The Clinical Evaluation of Language Fundamentals (CELF) is a standardized measure, commonly used for clinical assessment of **language** in autism.

CPRS - The Comprehensive Psychopathological Rating Scale (CPRS) is a scale for rating the severity of **psychiatric** symptoms and observed **behaviour**. Autistic children were rated on the CPRS for evaluating psychopathology in autistic children.

CSI - Child Symptom Inventory-4 (CSI-4) are scoring algorithms **differentiating** children with **ASD from** youngsters with **ADHD**.

DSM 4 and **DSM 5** - The Diagnostic and Statistical Manual of **Mental Disorders** (DSM) is the handbook used by healthcare professionals as the authoritative guide to the diagnosis of mental disorders. The American Psychiatric Association (APA) published the DSM-5 in 2013, replacing DSM-4. DSM-5 **ASD diagnosis** is made on the basis of difficulties in 2 areas – 'social communication', and 'restricted, repetitive and/or sensory behaviours or interests'.

HANDEDNESS - Children with ASD have a less definitive **hand preference** for certain actions as opposed to neurotypical ones.

MASC - Theory of Mind (**ToM**) is one of the most relevant concepts in the field of social cognition in ASD. The Movie for the Assessment of Social Cognition (MASC) is a sensitive video-based test for the evaluation of subtle mind reading difficulties.

RBSR - A key feature of autism is **restricted repetitive behavior** (RRB). The Repetitive Behavior Scale-Revised (RBS-R) is a questionnaire that captures the breadth of RRB in autism.

SCQ and **SRS** - Both the **Social Communication** Questionnaire (SCQ) and **Social Responsiveness** Scale (SRS) are questionnaires designed for detecting risk for ASD.

VINELAND - Vineland-3 is a standardized measure of **adaptive behavior**: Whereas ability measures focus on what the examinee can do in a testing situation, Vineland-3 focuses on what he/she **actually does in daily life**.

Methods:

Exploratory Data Analysis (EDA)

I've made an Exploratory Analysis of the phenotypic data using python (the code notebook can be found in this GoogleDrive link under ASD Prediction.ipynb)
In addition, a link to Sweetviz html file can be found in this GoogleDrive link under ASD_sweetviz.html - showing a quick view of the basic attributes and statistics of the features (plus and Red-Yellow-Green indication of the missing data percentage).

Phenotypic CSV:
1114 Instances
348 Features (~ 82% missing data) - Mostly numerical features (only 5 are strings)

Long Phenotypic CSV:
38 Instances
349 Features (~ 93% missing data) - Same features as on the Phenotypic CSV + an additional SESSION feature (string).

Select Rows (Long Phenotypic):
SESSION == 'baseline' (and then drop the SESSION feature and concatenate to the Phenotypic dataframe)

Concatenated dataframe (df):
1152 Instances (1114+38)
348 Features (~ 83% missing data)

df.describe():
We can already see some main characteristics of the data:

- *count is very low on many of the features (since the percentage of missing data is very high)*
- *By comparing between min, 25%, 50%, 75% and max values per column - we can spot some skewed features (such as Age at scan, Sex, Handedness Category etc.)*

Note:
Additional Data Analysis will be done in the Feature Selection section where I will look at some tables, plots and heatmap of the most relevant features.

Feature Selection

Selecting a Target Feature

In this project I selected DX_GROUP as my target feature for classification (where: 1=Autism; 2=Control).

DX_GROUP is the final and actual diagnosis of ASD for the subjects and it is well balanced: 521 with ASD; 593 with No ASD.

1st phase of Feature Selection - using domain knowledge relevance to ASD

Seeing that the dataset contains many features that are related to a variety of diagnostics tools and tests, my first step included consultations with a domain expert – which resulted in looking only at total scores of the most relevant tests. This way, I was able to drop 213 columns from the dataframe.

Feature Selection phase I : 348 features => 135 features

2nd phase of Feature Reduction - aggregating several feature into one (using mean)

Some relevant tests were not supplied with a Total score. In order to avoid any biased predictions by having several features per one specific test, I aggregated features per test category (using nanmean since some subjects had more/less missing data per test category than others) . By doing that, I ended up with 31 features to analyze.

Column Aggregations by Test Category:
ADI (5 Columns), BRIEF (11 Columns), CELF (6 Columns), BASC2 (25 Columns), CPRS (14 Columns), CASI (29 Columns), CSI (21 Columns)

Feature Reduction phase II : 135 features => 31 features

3rd phase of Feature Selection - removing features where 90% or more of the data is missing

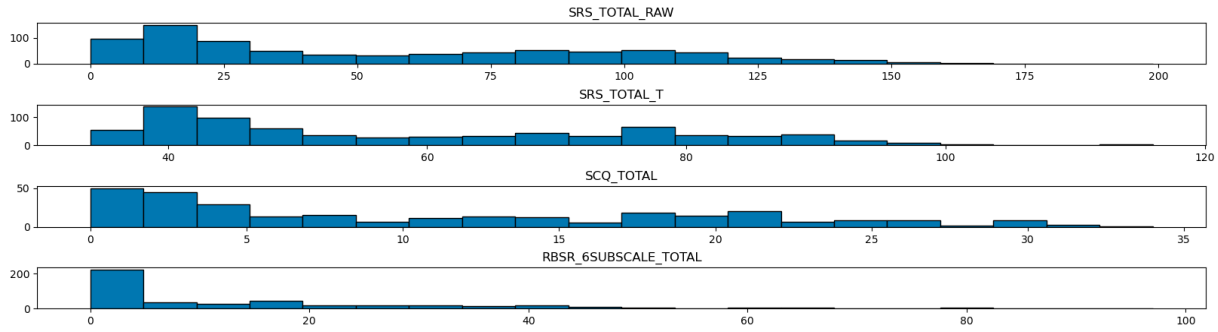
The following features were dropped since they had 90% or more missing data (which means that even if the correlation to our target feature - DX_GROUP - is great, we will not be able to get enough data from these features to predict ASD):
'AQ_TOTAL', 'VINELAND_SUM_SCORES', 'CBCL_6-18_TOTAL_COMPETENCE_T', 'CBCL_1.5-5_TOTAL_T', 'CELF_MEAN', 'CPRS_MEAN', 'CASI_MEAN', 'CSI_MEAN'

Feature Reduction phase III : 31 features => 23 features

Additional Data Analysis - plots and heatmap of the most relevant features

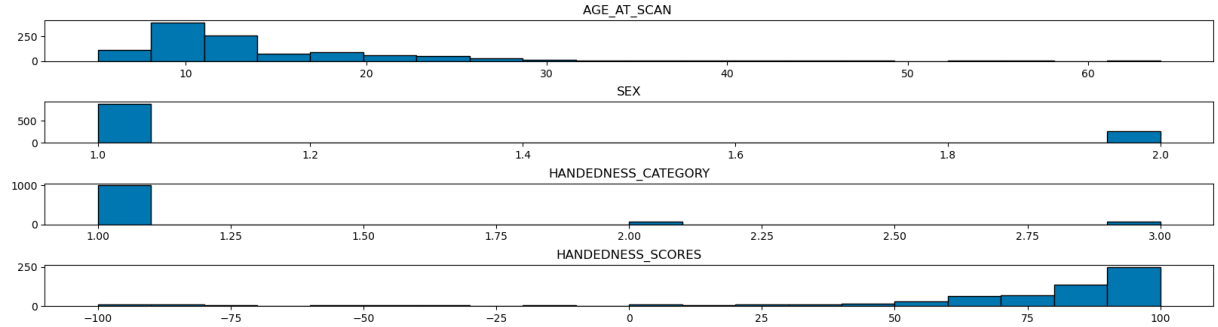
Histograms: I created a histogram for each numeric feature (from the 23 features that were left).
In the following images, you can see the distribution of 4 different features as an example.
Some observations can be easily made from this visualization:

- SRS_TOTAL_RAW and SRS_TOTAL_T look quite similar and I need to run some correlation analysis to see if they are indeed highly correlated (which means one of them should be dropped).
- RBSR_6SUBSCALE_TOTAL is skewed and should be dealt with in the Preprocessing phase.



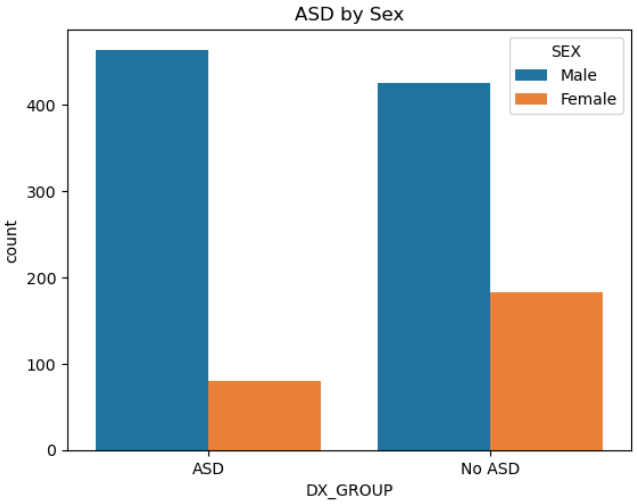
Bar Plots: Looking at the numeric features histograms, I understood that some of them can be viewed as categorical features (as they only have 2 or 3 possible values) - the ones I was interested to get further insight into were - SEX, HANDEDNESS_CATEGORY and AGE_AT_SCAN (divided into age groups).

Histograms for these main categorical features showed that they are all very skewed (should I drop them? - we will see later on):

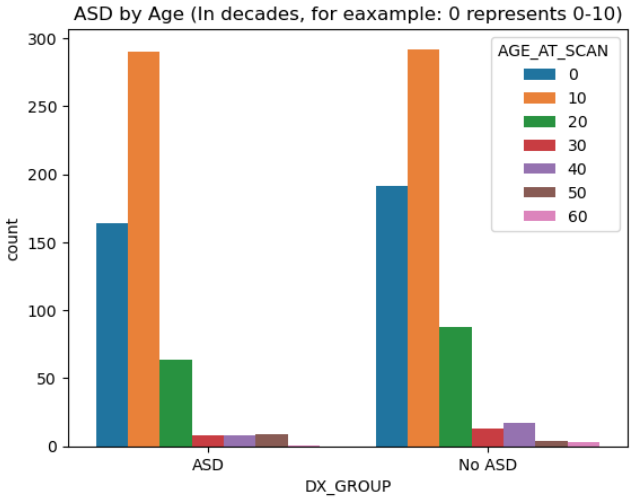


I added Bar Plots for these 3 features to see their distribution between the ASD and No-ASD categories:

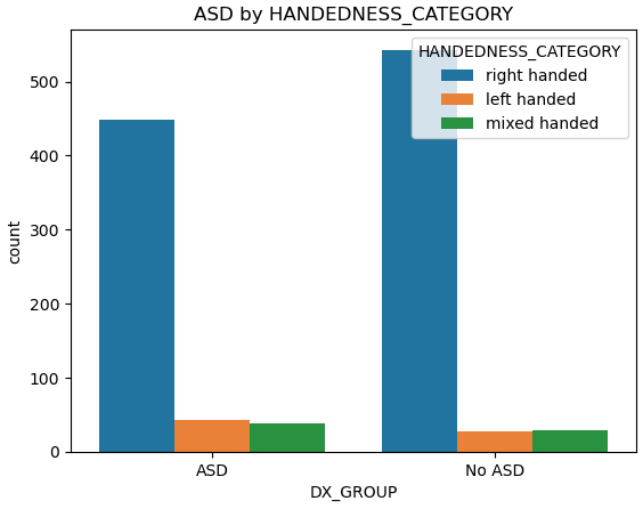
- SEX: Clearly, the ABIDE II dataset has more male than female subjects with ASD, which represents the proportion in the general population (ASD is more than 4 times more common among males than among females). This may explain why the dataset was built containing more male subjects in general.



- AGE_AT_SCAN: The vast majority of the subjects were scanned between ages 0-30. This feature will be important when we get to look at the actual MRI scans. For the scope of the current project, since ASD is diagnosed once and for life, I think it can be dropped without causing any critical bias.



- HANDEDNESS_CATEGORY: This feature is very skewed (representing the proportion of right vs. left handed in the general population). I don't see enough correlation between handedness and ASD from this bar plot and I should probably drop it (I'll look at the Heatmap first to get some numbers regarding the actual correlation)



Heatmap: I created a heatmap showing linear correlation between each pair of features. I was interested to find the following:

- Correlations between non-target features: identify features that are over correlated (and drop one to avoid bias). The pairs I found to be highly correlated were:
  - ADOS_2_TOTAL vs. ADOS_2_SEVERITY_TOTAL (the latter has less granularity)
  - HANDEDNESS_CATEGORY vs. HANDEDNESS_SCORES (the latter has a lot of missing data)
  - SRS_TOTAL_T vs. SRS_TOTAL_RAW (it appears SRS_TOTAL_T is the one used for actual diagnosis)
  - RBSR_5SUBSCALE_TOTAL vs. RBSR_6SUBSCALE_TOTAL (5th version of the test vs. a younger 6th version)

- Correlations between target and non-target features: to find the best features for the classification models, I looked for the features that had the highest positive/negative correlations to ASD (DX_GROUP column). See 5th phase of feature reduction.

Note: Although the number of features was considerably reduced, the heatmap is still too big to be viewed properly from a PDF file and I recommend to view it directly from the notebook found in this GoogleDrive link under ASD Prediction.ipynb)

4th phase of Feature Selection - removing one feature from each pair of highly correlated non-target features

I decided to drop the following features:
'ADOS_2_SEVERITY_TOTAL', 'HANDEDNESS_SCORES', 'SRS_TOTAL_RAW', 'RBSR_5SUBSCALE_TOTAL'
Feature Reduction phase IV : 23 features => 19 features

5th phase of Feature Selection - selecting the final features which best correlate to ASD

I created the final dataframe (df_ASD_Pred) which will serve as input to the classification models, containing only the features with the highest correlation to DX_GROUP (from the remaining features):
'PDD_DSM_IV_TR', 'ASD_DSM_5', 'SRS_TOTAL_T', 'SCQ_TOTAL', 'RBSR_6SUBSCALE_TOTAL', 'CBCL_6-18_TOTAL_PROBLEM_T',

'BRIEF_MEAN', 'ADOS_G_TOTAL'

Feature Reduction phase V - FINAL: 19 features => 8 features + 1 target

Preprocessing of the selected columns

- We saw in the histograms that the RBSR, SCQ and SRS features were skewed. I handled this by applying log1p.
- I then standardized the numeric values (with StandardScaler)
- and filled in NaN values (using KNN imputer as there are many missing values)

Models

Supervised - Classification Models

After splitting the data into train (75%) and test (25%), I applied 5 different classification models:

- Logistic Regression
- KNN
- Support Vector Machine
- Decision Tree
- Random Forest

I first trained each model on the training data, then tried to predict ASD/No-ASD for the test data.
I then evaluated my results applying classification_report on the predicted target class vs. the actual DX_GROUP value of each subject on the test data.

Note: I used a range of different hyperparameters' values and found the ones that worked best for each model.
I will share the best hyperparameters and the detailed results for each model in the *Results* section.

Unsupervised - PCA and Clustering

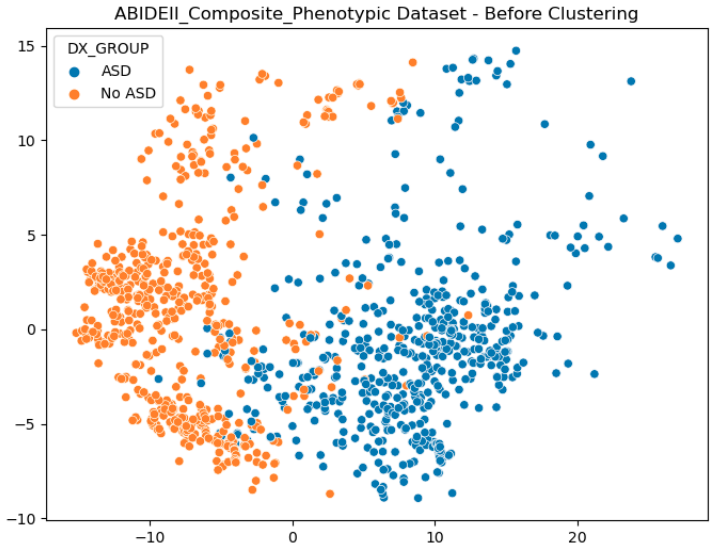
At this stage, I wanted to see if there was a faster way (which didn't need any research) to reach the same accuracy of results.

First, I needed to deal with the high dimensionality of the data:

I took the initial data (before any feature selection), this time using a fast and automatic method to reduce the dimensionality:

I applied Principal Component Analysis which is known to convert a set of correlated features in a high dimensional space into a new set of uncorrelated features in a lower dimensional space.

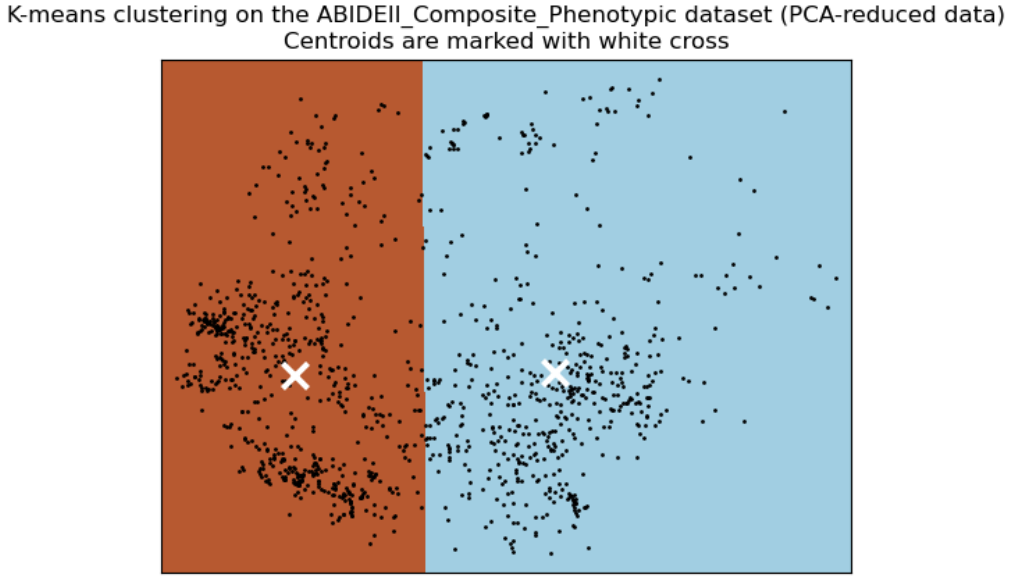
The following image shows the reduced data and how it is distributed (colors were added to show DX_GROUP for each subject):



Looking at the scatter plot above it was clear that if I now try to find 2 clusters on this reduced dimensionality data, I would most likely end up with 2 clusters highly resembling the blue and orange "clusters" (ASD/No-ASD) seen on the scatter plot, since most of the blue points were situated close to each other on the right side and most of the orange points were close-together on the left side.

Next, I applied KMeans on the reduced data looking for 2 clusters. I got exactly what I was expecting -

One cluster on the right side (representing predicted ASD subjects) and another one on the left side (representing predicted No-ASD subjects). Note that this is only a speculated representation since no actual labels were given to the algorithm (as it works in an unsupervised manner):



Results

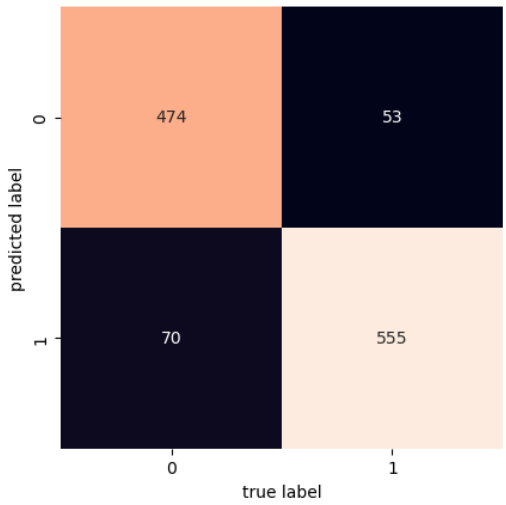
Supervised Classification Models:

Algorithm	Best hyperparameters	Precision vs. recall	Training accuracy	Test accuracy
Logistic Regression	'classifier__C': 1 'classifier__max_iter': 500	precision recall 1 0.96 0.84 2 0.87 0.97 A good precision-recall balance	0.93	0.91
KNN	'classifier__leaf_size': 6 'classifier__n_neighbors': 8	precision recall 1 0.95 0.88 2 0.89 0.96 A good precision-recall balance	0.94	0.92
Support Vector Machine	'classifier__C': 10 'classifier__kernel': 'rbf	precision recall 1 0.97 0.87 2 0.89 0.97 A good precision-recall balance	0.94	0.92
Decision Tree	'classifier__max_depth': 5 'classifier__min_samples_leaf': 2	precision recall 1 0.96 0.84 2 0.87 0.97 A good precision-recall balance	0.95	0.91
Random Forest	'classifier__max_depth': 7 'classifier__min_samples_leaf': 1	precision recall 1 0.98 0.87 2 0.88 0.99 A good precision-recall balance	0.95	0.93

The table above (representing the classification report) clearly shows that the **Random Forest** Model reached the best scores with test accuracy of 0.93 and a good balance between precision and recall on both classes.

Unsupervised Clustering:

I decided to evaluate the Clustering prediction results using a Confusion Matrix:



From this matrix, I calculated the following scores:

Accuracy = (474 + 555) /1152 = **0.89**

A good accuracy (even though it is lower than what we got on the supervised classification models).

Recall = 474 / (474 + 53) = 0.9

Precision = 474 / (474 + 70) = 0.87

There is a good balance between Recall and Precision.

Discussion

Looking at the Supervised Classification, the best accuracy was reached using **Random Forest**:
Prediction accuracy score of 0.93
But these results were obtained after a lot of research, careful feature selection and trying different hyperparameters until finding the best ones for each model.

On the other hand, the PCA-Clustering process took almost no work and yielded an **accuracy of 0.89**.

In this case, the hard work paid off. But, it would still be interesting to experiment with different feature reduction and clustering algorithms to see if we can reach the level of accuracy obtained by the Random Forest model.

Looking again at our general goal:

We saw that it is possible to predict ASD with high accuracy from the phenotypic data of a given subject.

Now, the next question is - How long would it take for a child to undergo these critical phenotypic tests, bearing in mind the overload on child development centers and the fact that the diagnosis process at these centers starts at the age of 2.

If we can reach the same level of accuracy from MRI scans and diagnose ASD on babies - even before they show speech impairment, social difficulties and other ASD related characteristics - therapists may be able to address and treat issues before they become actual problems in the child's life.