Tali Aharon 034791236

Helit Bauberg 027466002

Ron Shternberg 034466862

Machine Learning – 99004 – Project:

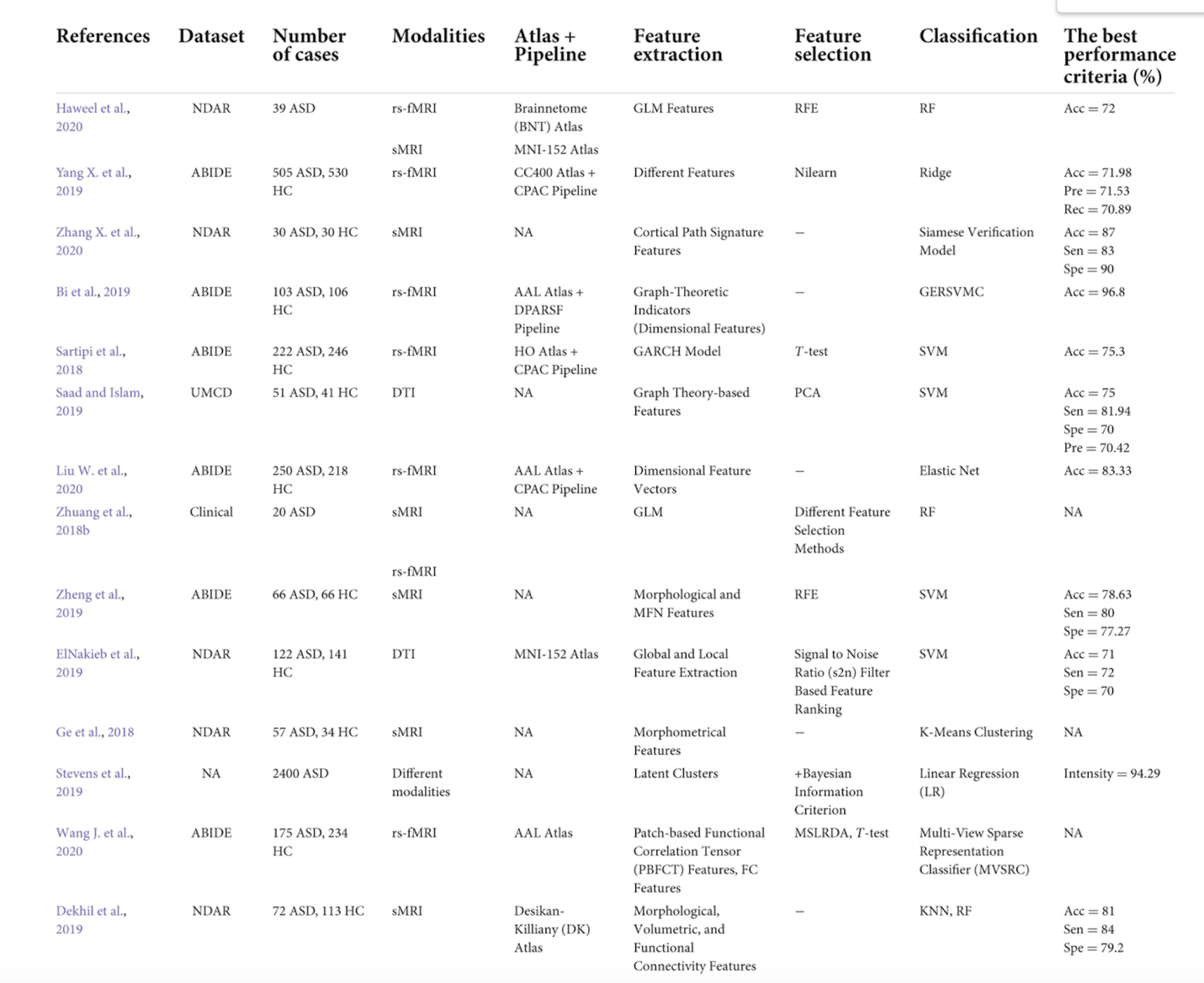
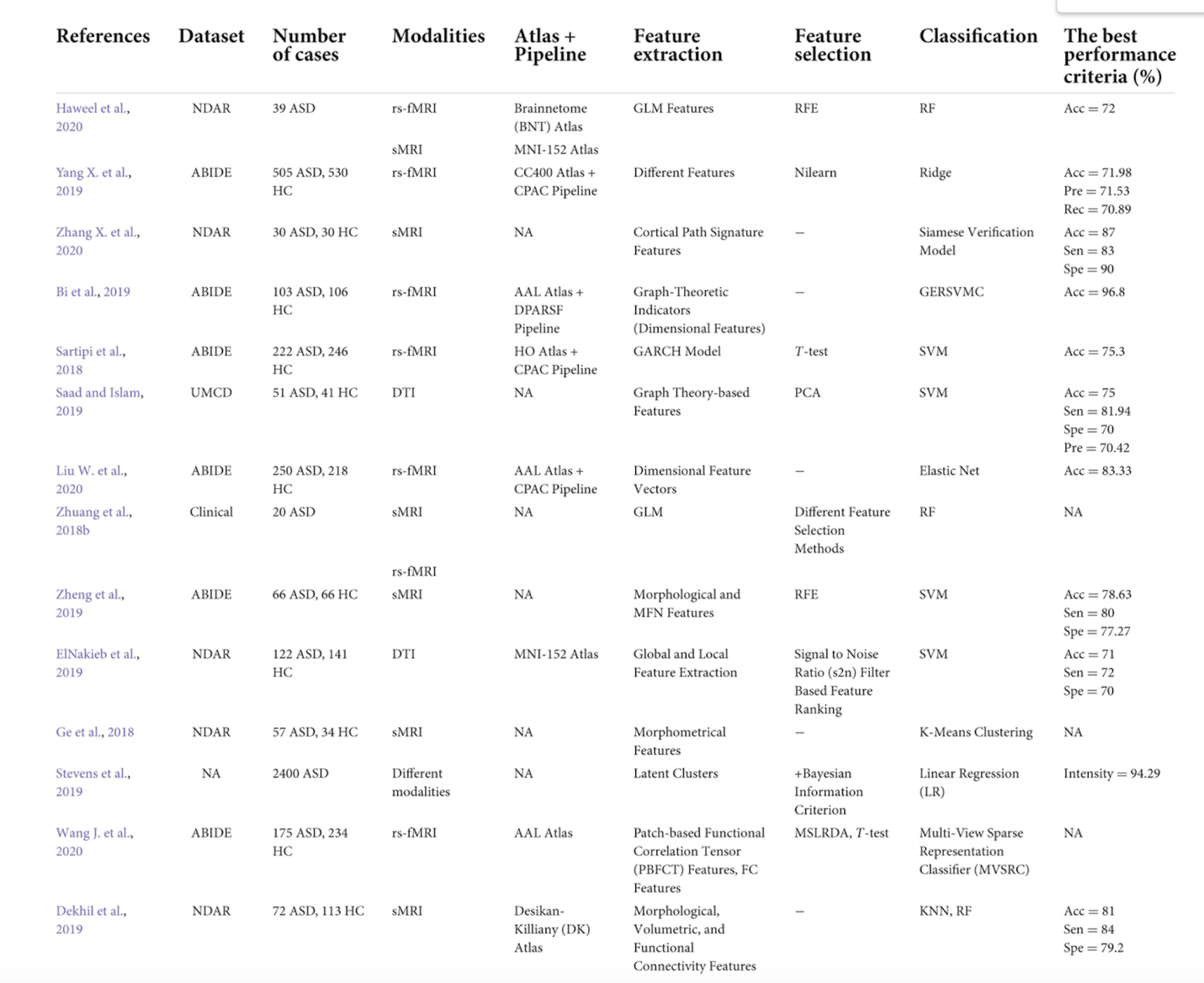
**Preprocessing** ABIDE dataset –   
Phase I in **preparations** for Comparative Analyses – applying CAD System for Personalized Autism Diagnosis Using sMRI and fMRI Data

Scope:

Our focus of interest is to compare, customize and provide additional insights for a Computer-Aided Diagnosis System (CADS) for diagnosis of Autism Spectrum Disorder (ASD).

The system was developed by Dekhil, O., Ali, M., El-Nakieb, Y., Shalaby, A., Soliman, A., Switala, A., et al. and published in their article in 2019 - link to the full article:   
[*https://doi.org/10.3389/fpsyt.2019.00392*](https://doi.org/10.3389/fpsyt.2019.00392)

Using a fusion of structural MRI and resting-state functional MRI data, Dekhil et al extracted various statistical features for ASD diagnosis from MRI data and then applied KNN and SVM algorithms in the classification step. The authors reported 81% accuracy:



Notable from the above research summary are three points:

1. Original study used National Database of Autism Research (NDAR) dataset, and the total number of subjects used is 185 subjects (72 of them with ASD).
2. Dekhil et al only took as subject patients who have ADOS score (ADOS is the golden standard of ASD diagnostic tests).
3. There was no feature selection in their study, nor did they use any Unsupervised Learning algorithms.

In our proposed project, we would like to explore the following:

1. We would like to use a different dataset, from the Autism Brain Imaging Data Exchange ([ABIDE II](http://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html)).  
   ABIDE is recognized as the most complete and freely available database of MRI modalities for the automatic diagnosis of ASD. ABIDE II includes 1,152 subjects and contains sMRI data, rs-fMRI data, and phenotypic data that includes a large set of test scores and diagnostic tools used by psychiatrists, psychologists etc. to diagnose ASD. While Dekhil used **ADOS-G** standard as the target variable to predict ASD, the standard has changed since 2012 and **ADOS-2** is now the adopted standard. We would like to factor both ADOS scores and compare our scores to Dekhil’s.
2. Dekhil has dropped all records that did not include ADOS score. Yet, the ABIDE II dataset includes other scores that can be used to calculate and predict what would be the matching ADOS score. There are approximately 600 records with neither ADOS-G nor ADOS-2 score. We would like to predict ADOS scores and include those records’ MRI images in our analyses.
3. We would like to explore and apply various feature selection algorithms, as well as experiment with association rule algorithms.

The goal of the full project is to assist in the field of research and early diagnosis of ASD. Early diagnosis is crucial to ASD treatment (bearing in mind the high neuroplasticity at younger ages) and thus be able to contribute further insight into the Autistic brain.

We hope that by using bigger datasets, various feature selection and classification algorithms, and with further fine tuning (e.g., comparing age ranges) we might contribute to a CADS that has already shown a very high prediction rate.

**Part I**

**In this semester** we will focus on analyzing and preprocessing the data in preparation for the actual image data analyses.

Data preprocessing include two main focuses:

1. MRI data – filter out any scans that failed quality tests; filter out any patients missing either f-MRI or s-MRI scans.
2. Phenotypic data - There are over 340 features in the ABIDE phenotypic dataset. These can be grouped into subcategories based on the related tests. As our target feature is ADOS, we will try to predict ADOS score for records missing that score using the other tests scores and features provided in the dataset.

### Data Source

The data related to the subjects in the ABIDE dataset is split into a few files as follows:

* **Phenotypic data**

1. ABIDE II Composite Phenotypic File => Our main data source for subject classification and filtering, containing 1114 subjects.
2. ABIDE II - Longitudinal Composite Phenotypic File => Includes data related to additional 38 subjects. All subjects have baseline + followup scans data a few years apart. We will include all baseline data for these subjects in our analyses
3. ABIDE II Phenotypic Data Legend => showing per each feature its label, short description, variable type and range of values.

The files can be found here <https://www.nitrc.org/frs/downloadlink.php/9108>

Out of the 340+ features in the phenotypic dataset, we dropped all features that are either not related to ASD, or general characteristics that we don’t foresee adding value to our analyses.

The xls file that includes the full list of features and whether we included or excluded each, a description of each feature, its relevance to ASD and supporting links is in [this](https://drive.google.com/drive/folders/13C_HiP6--kg5Wx32iunwkt-O6yhnWadA?usp=sharing) GoogleDrive link under ABIDEII\_Data\_Legend.xlsx

The Phenotypic file includes tests for ASD diagnostic alongside tests from other domains with relevance to ASD.

The test categories that we will look at 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. The ADI-R and the ADOS (see below) are used either independently or combined.

**ADOS** - The **Autism Diagnostic** Observation Schedule (ADOS) was 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. Since its release in May 2012, ADOS-2 replaced ADOS. With updated protocols, revised algorithms, a new Comparison Score, and a Toddler Module, the ADOS-2 provides a highly accurate picture of current symptoms, unaffected by language. It can be used to evaluate almost anyone suspected of having ASD—from 1-year-olds with no speech to adults who are verbally fluent.

**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 five 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 problem behaviors rather than ASD in particular. The instrument developer proposed that the 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 for **differentiating** children with **ASD** **from** youngsters with **ADHD**.

**DSM 4 and 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 autism spectrum disorders (ASD) have a less definitive **hand preference** for certain actions as opposed to neurotypical children.

**MASC** - Theory of Mind (**ToM**) is one of the most relevant concepts in the field of social cognition, in the case of ASD. The Movie for the Assessment of Social Cognition (MASC) is a sensitive video-based test for the evaluation of subtle mindreading 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**: the things that people do to function in their everyday lives. 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**.

### Exploratory Data Analysis (EDA) – Cycle I

A link to Sweetviz html file can be found in [this](https://drive.google.com/drive/folders/13C_HiP6--kg5Wx32iunwkt-O6yhnWadA?usp=sharing) GoogleDrive link under ASD\_sweetviz.html

*Phenotypic CSV:*

*1114 Instances*

*345 Features (82.3% missing data)*

*3 Meta Attributes (41.7% missing data):*

1. *NDAR\_GUID*
2. *NONASD\_PSYDX\_LABEL*
3. *CURRENT\_MEDICATION\_NAME*

*Long Phenotypic CSV:*

*38 Instances*

*346 Features (93.2% missing data)*

*2 Meta Attributes (85.5% missing data):*

1. *NDAR\_GUID*
2. *CURRENT\_MEDICATION\_NAME*

*Select Rows (Long Phenotypic):*

*SESSION == ‘baseline’*

*Concatenated Table:*

*1152 Instances (1114+38)*

*350 Features (82.9% missing data)*

*3 Meta Attributes (43.3% missing data):*

1. *NDAR\_GUID*
2. *NONASD\_PSYDX\_LABEL*
3. *CURRENT\_MEDICATION\_NAME*

Preprocessing

### Preliminary Feature Selection – utilizing domain expertise

Seeing that the dataset contains many features that are related to a variety of diagnostics tools and tests, our first step included consultations with a domain expert – a psychologist specialized in ASD and ASD-ADOS diagnostics with a vast experience in the field. We were given a detailed explanation on all the data that is included in the dataset, and based on those discussions and further research, we have narrowed down our feature list to 140 features. Those features are based on either total scores of a specific test, or a new column calculated using other relevant columns based on domain knowhow.

The full list of features and whether we dropped/kept each is in [this](https://drive.google.com/drive/folders/13C_HiP6--kg5Wx32iunwkt-O6yhnWadA?usp=sharing) GoogleDrive link under ABIDEII\_Data\_Legend.xlsx

***Feature Reduction Cycle I : 346 features 🡺 138 features***

### Feature aggregation

### In order to further reduce our dataset’s dimensions, we will aggregate features per test category in order to use one score instead of many. By doing that, we will now have a total of 34 features to analyze.

### This also allows us to avoid any biased predictions by having one feature per test.

*Column Aggregations (Mean) of various Tests:*

1. *ADI (5 Columns)*
2. *BRIEF (11 Columns)*
3. *CELF (6 Columns)*
4. *BASC2 (9 Columns)*
5. *CPRS (9 Columns)*

*Column Aggregations (Sum) of various Tests:*

1. *CASI (10 Columns)*
2. *CSI (21 Columns)*

*Phenotypic Dataset Preprocessed:*

*145 Features (138 + 7 [Aggregated Mean]) (82.6% missing data)*

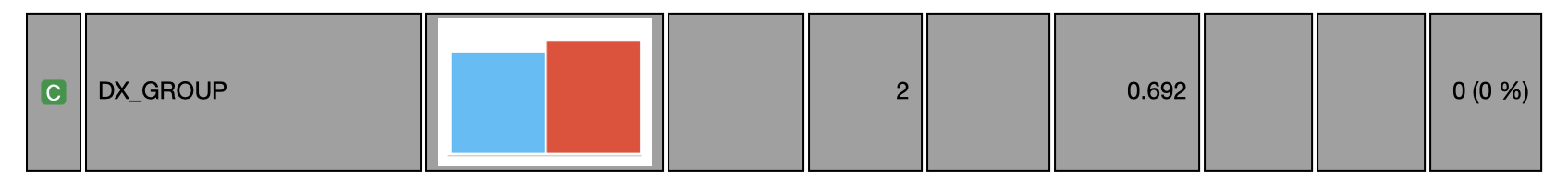
***Feature Reduction Cycle II : 138 features 🡺 31 features + 1 target + 3 Meta***

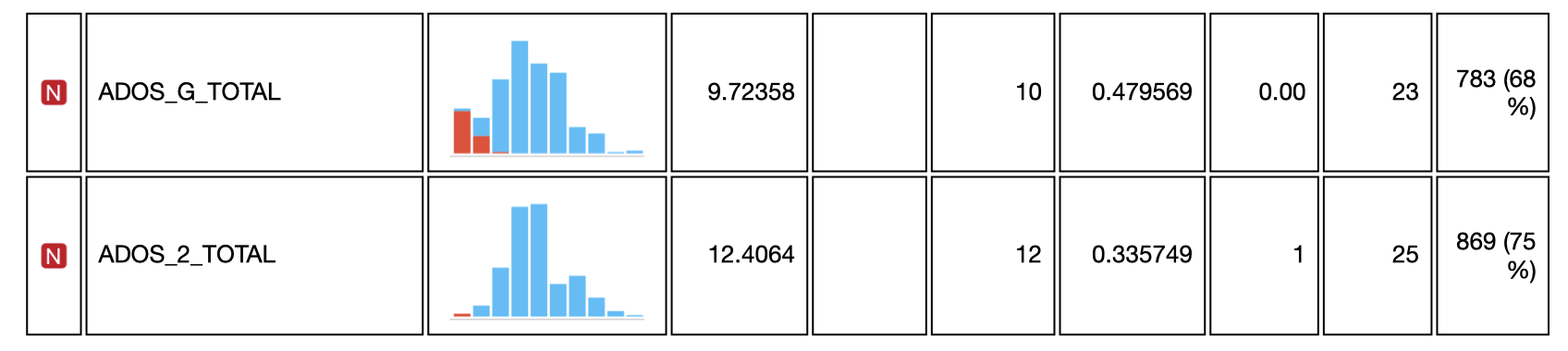
### Exploratory Data Analysis (EDA) – Cycle II

For the remaining 34 left features, we intend to apply:

* Classic feature statistics to identify min/max values, missing values percentage and distributions.
* Identify missing value percentage for each column in relation to missing ADOS 2/G values
* Skew: from initial observations, two important features are very skewed: the percentage of female subjects is very small compared to male subjects. As the tests scores are indigenous to gender, we will not take this fact into account. Same goes for the age feature, that is not distributed evenly amongst the subjects, with little to no representation of early ages. Again, test scores are not affected by this, and we will not take this fact into account. Once the EDA is done, we may identify other features that will need our attention.





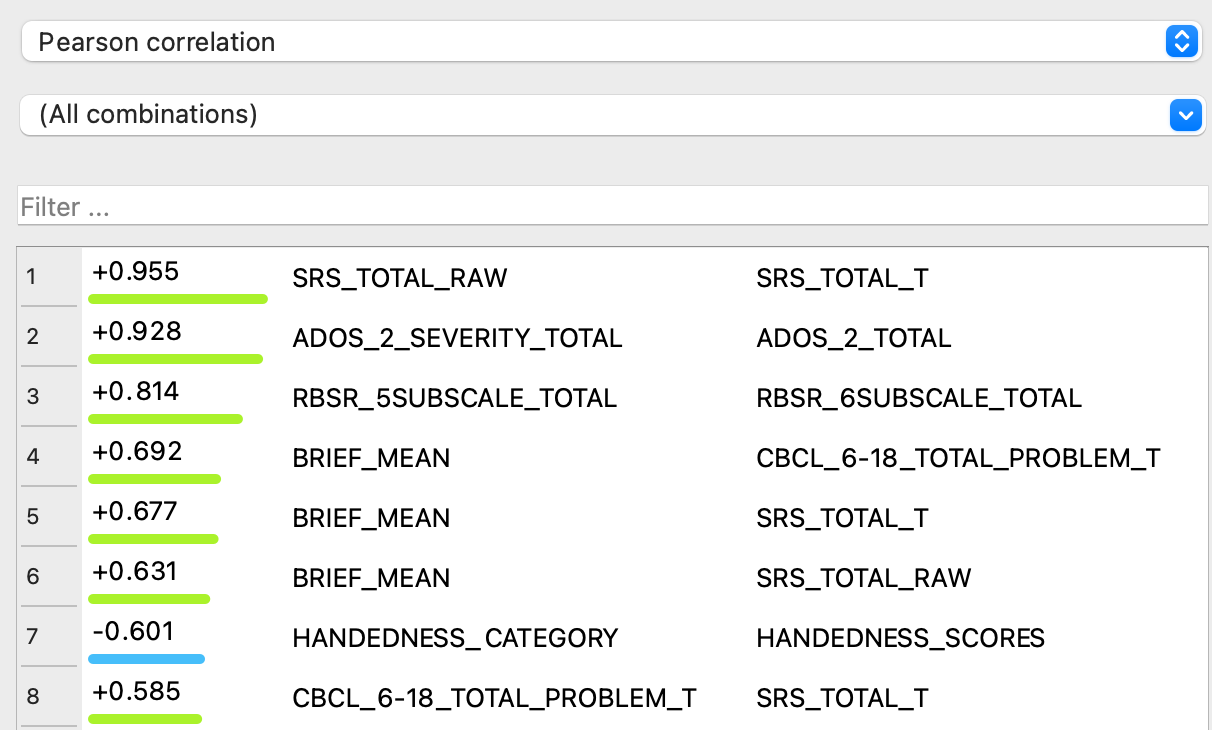


Analyze:

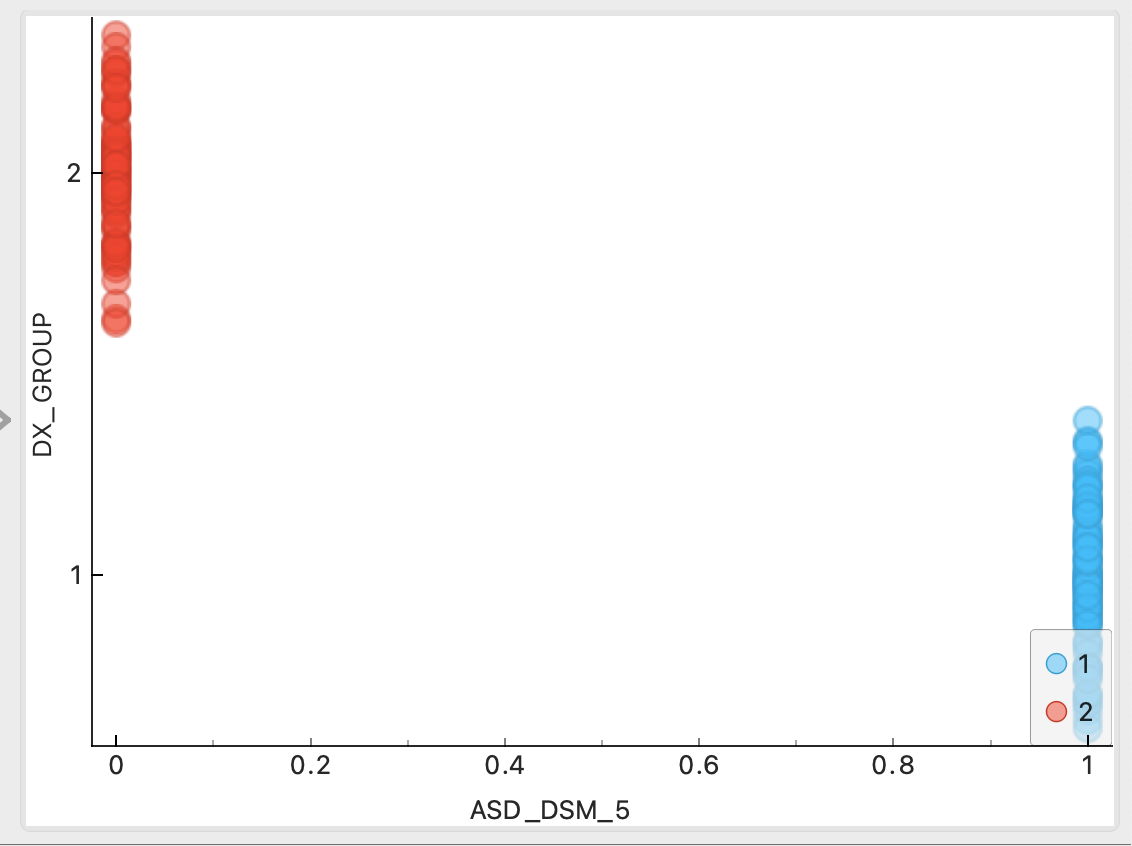
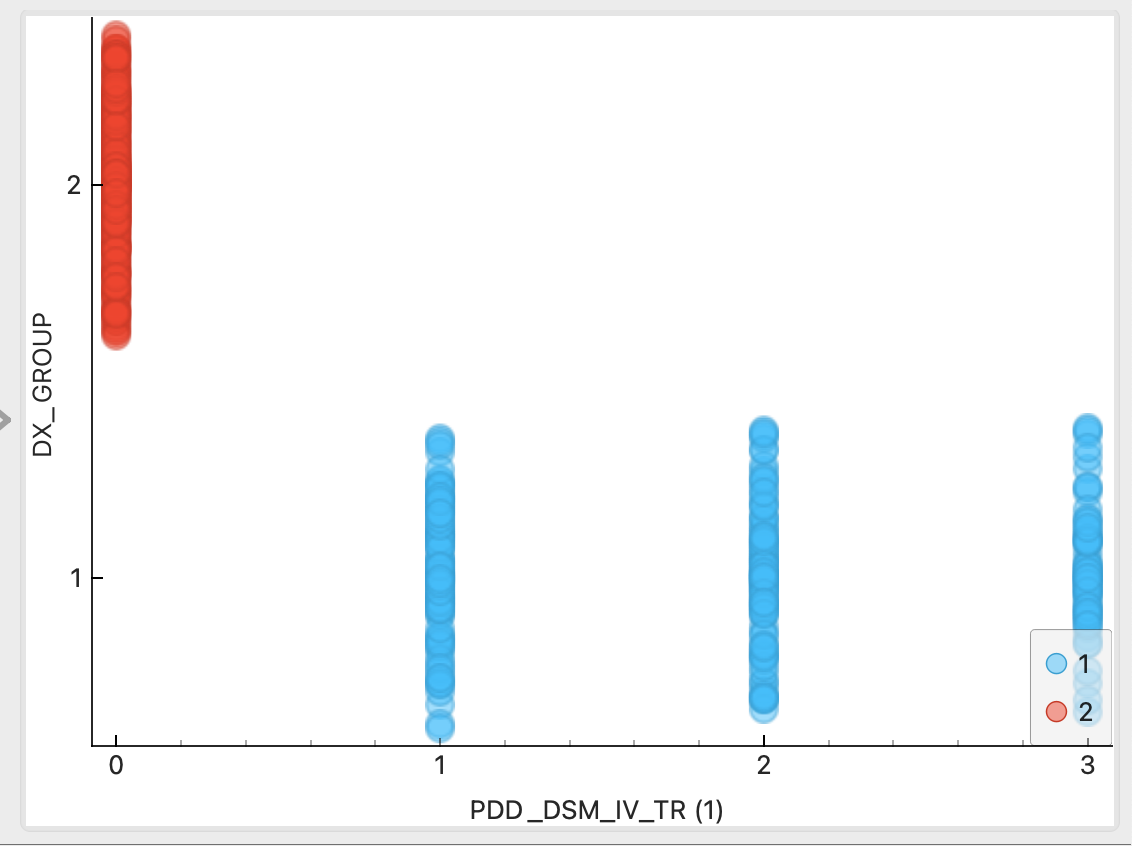
* Correlations between target and non-target features: to find the best features for our algorithms, we will look at the features that have the highest correlations to ASD (ADOS\_G, ADOS\_2 features)
* Correlations between non-target features: to identify features that are too correlated between themselves and drop one.

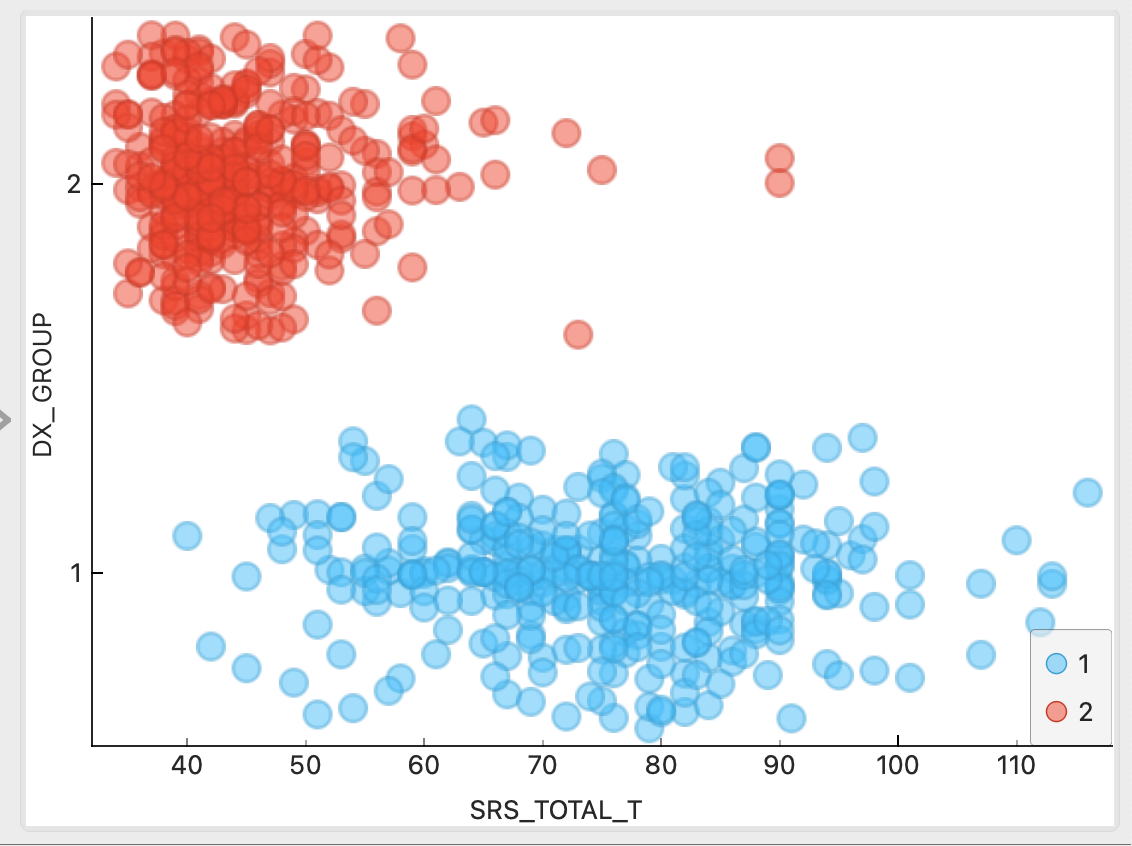
|  |  |
| --- | --- |
|  | Graphical user interface, application  Description automatically generated |

Linear correlations – ADOS\_G Linear correlations – ADOS\_2



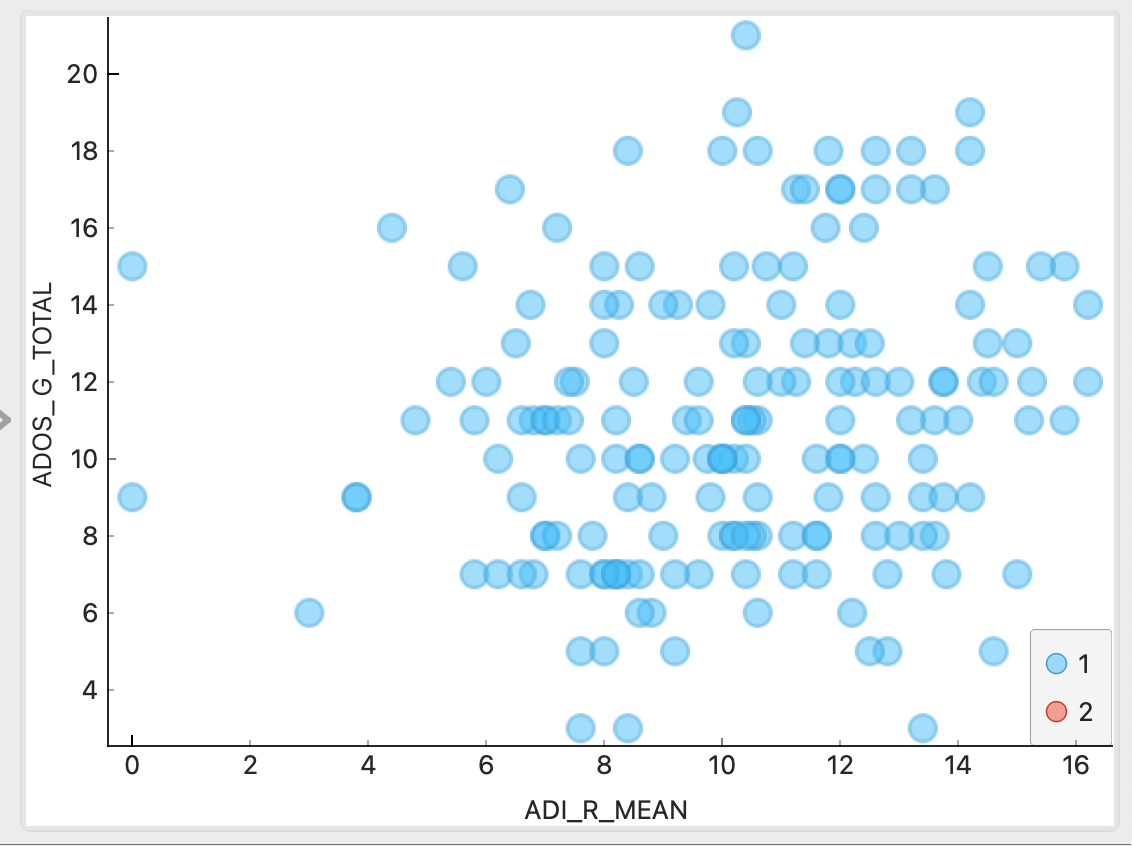
Linear correlations – All Combinations





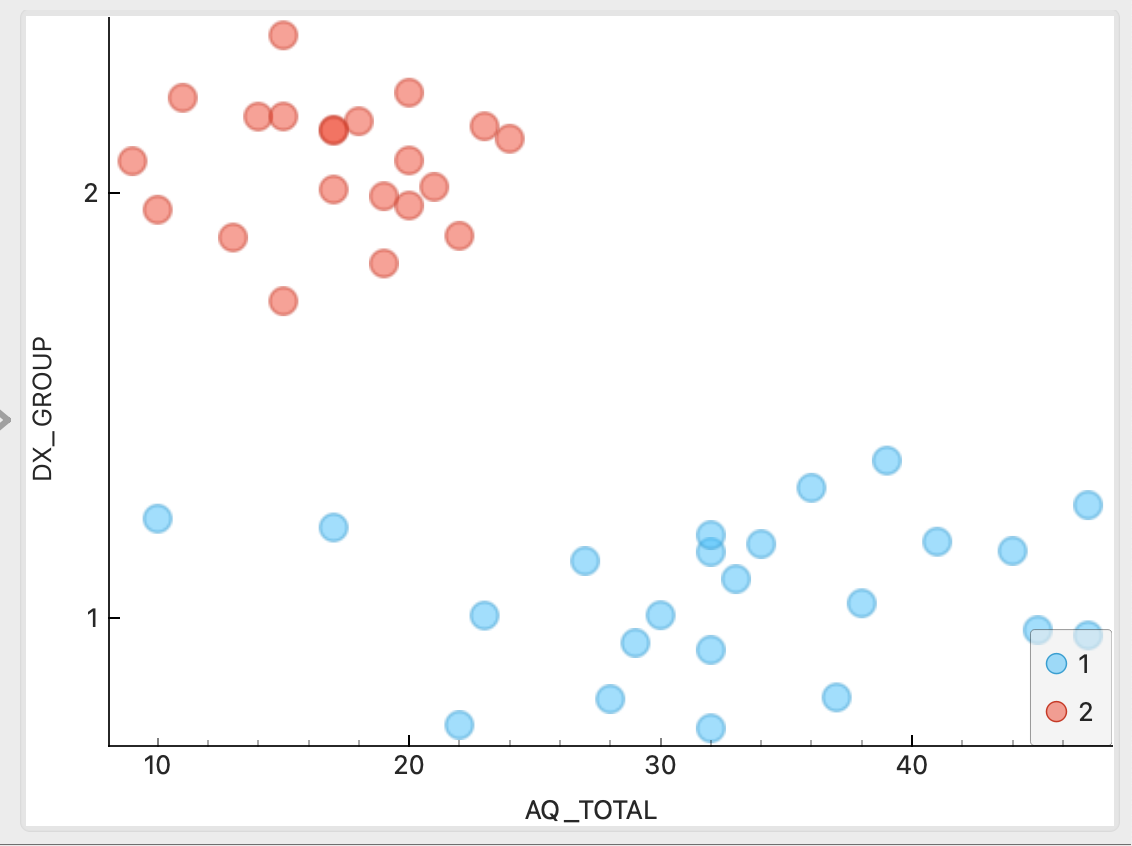
Conclusions from the above correlations/scatter plots:

* 7 Features dropped based on over correlation
* All features with low correlation to ADOS\_G / ADOS\_2 targets were dropped.
* DSM\_IV, DSM\_5, SRS and SCQ scores are highly correlated with our targets, so we keep the features.



Conclusion:

* ADI\_MEAN values, although values exist only for ASD diagnosed records, shows high correlation with ADOS scores, hence we decided to keep it.



Conclusion:

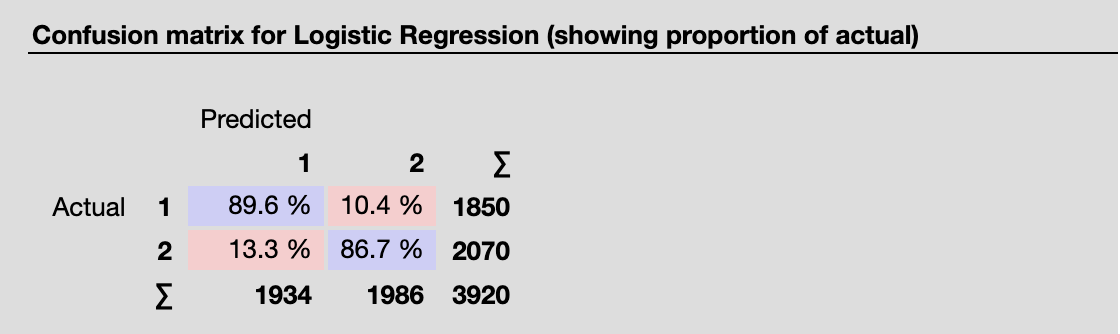
* Even though we are seeing good correlation between DX\_GROUP and AQ scores, we suspect we should drop the feature because of its sparseness – we only have 4% of the data available, 96% missing.

### **Classification using Logistic Regression**

Once we identified the important features, we wanted to evaluate how those subsets of features correlate as a whole, compared to our DX\_GROUP target. We used Logistic Regression for binary classification (made sense as DX\_GROUP is a binary classifier). The results yielded good accuracy as follows:

Graphical user interface, text, application

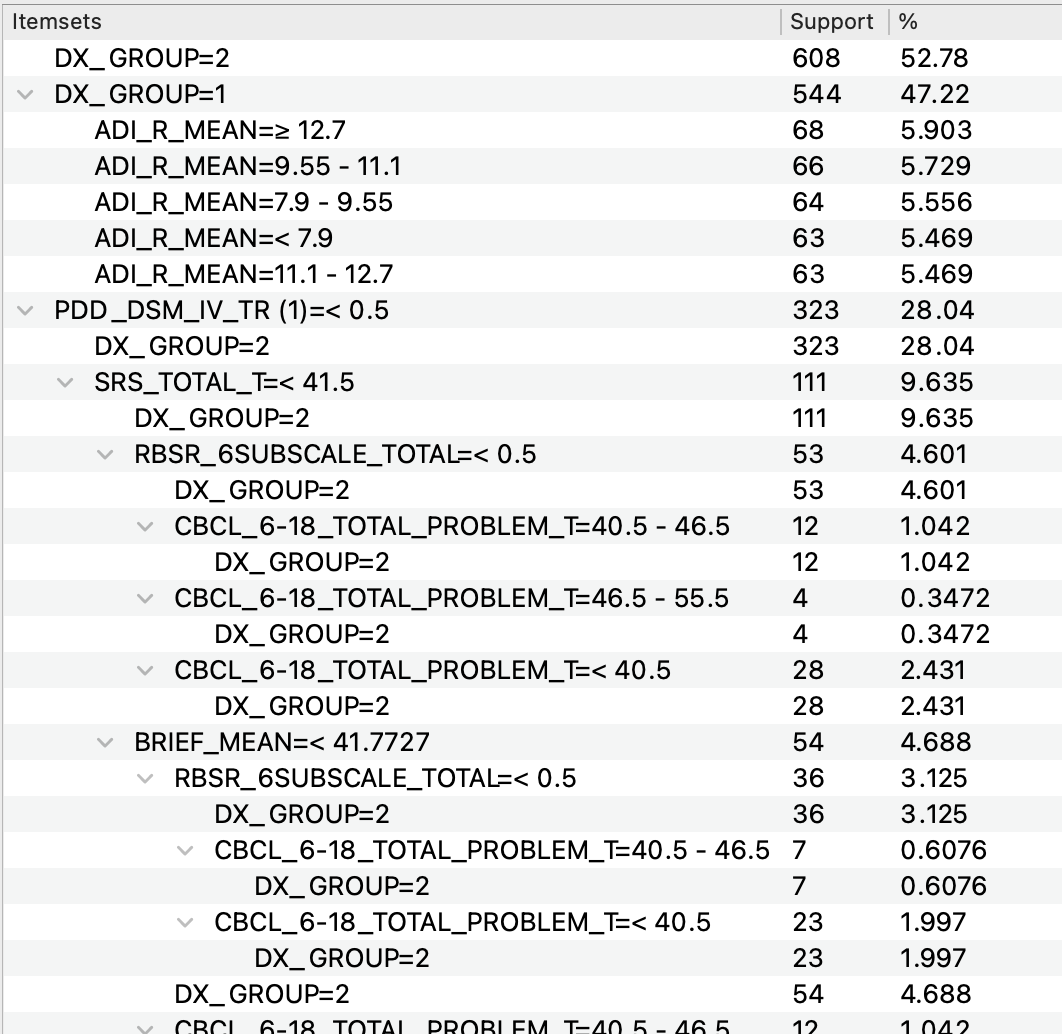
Description automatically generated



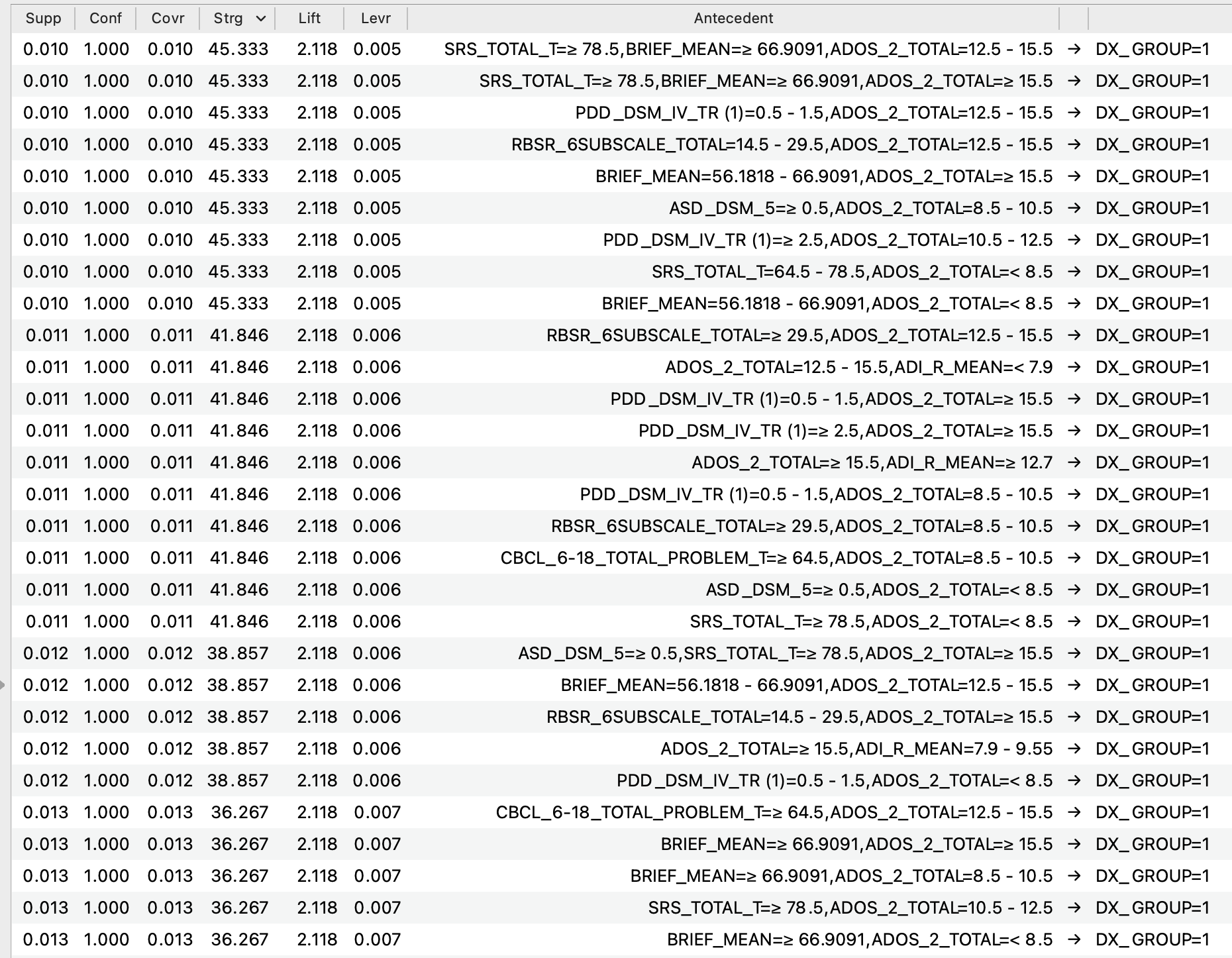
We managed to get to precision level of 88% in predicting both target values. The results are balanced (precision vs. recall) and clearly demonstrate that our pre-assumptions about the relations between the selected feature set make sense and we can continue in this direction.

### **Using Association Rules:**

* We applied equal discretization on all categorized features before sending the data to the Association Rules Engine. Results as follows:



Frequent Itemset



**Conclusions:**

* BRIEF\_MEAN, CPRS\_MEAN, AQ\_TOTAL, CBCL\_1.5-5\_TOTAL\_T dropped as there is no support for associations between those features and our targets.

***Feature Reduction Cycle III : 31 features 🡺 7 features + 3 targets + 3 Meta***

**Data discovery and Predictive Analyses :**

**High Level**: As we wrote in the project [scope](#SCOPE), whereas our overall target is to predict DX\_GROUP, i.e., ASD / Non-ASD, we intend to rely on ADOS (ADOS\_2/ADOS\_G) score prediction as an intermediary predictor for the final target.

ADOS\_G is the retired score version, followed by ADOS\_2.

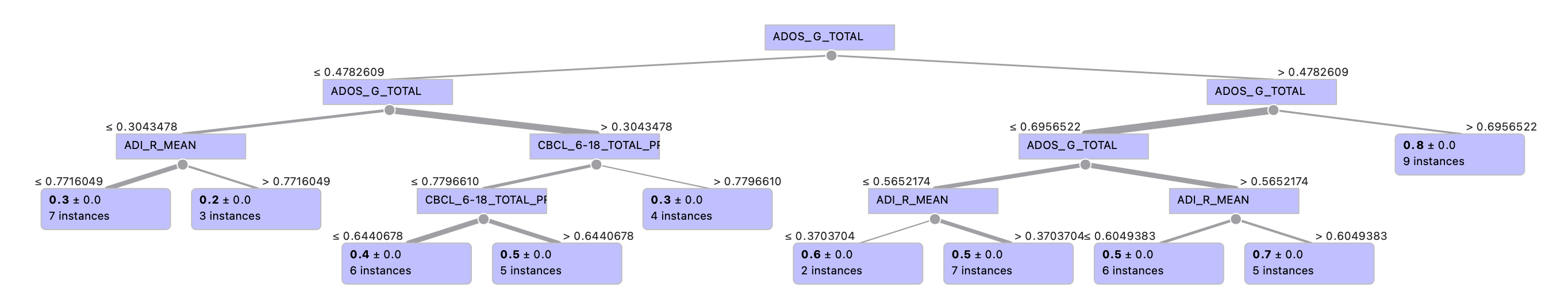
Both targets are sparse (68% / 75% missing values), but there is a clear transformation scheme between the scores, based on an algorithm proposed by Gotham et al. (Gotham K, Risi S, Pickles A, Lord C. The autism diagnostic observation schedule: revised algorithms for improved diagnostic validity. J Autism Dev Disord. 2007;37(4):613–627. doi:10.1007/s10803-006-0280-1). It made sense to run the same learners and prediction algorithms on both targets, and by this enlarge the number of records we can feed our final target (DX\_GROUP) prediction accuracy.

At this stage of the project, we have identified and reshaped the set of features we believe we can use for predictions, and used the following learning models for ADOS\_2, ADOS\_G.

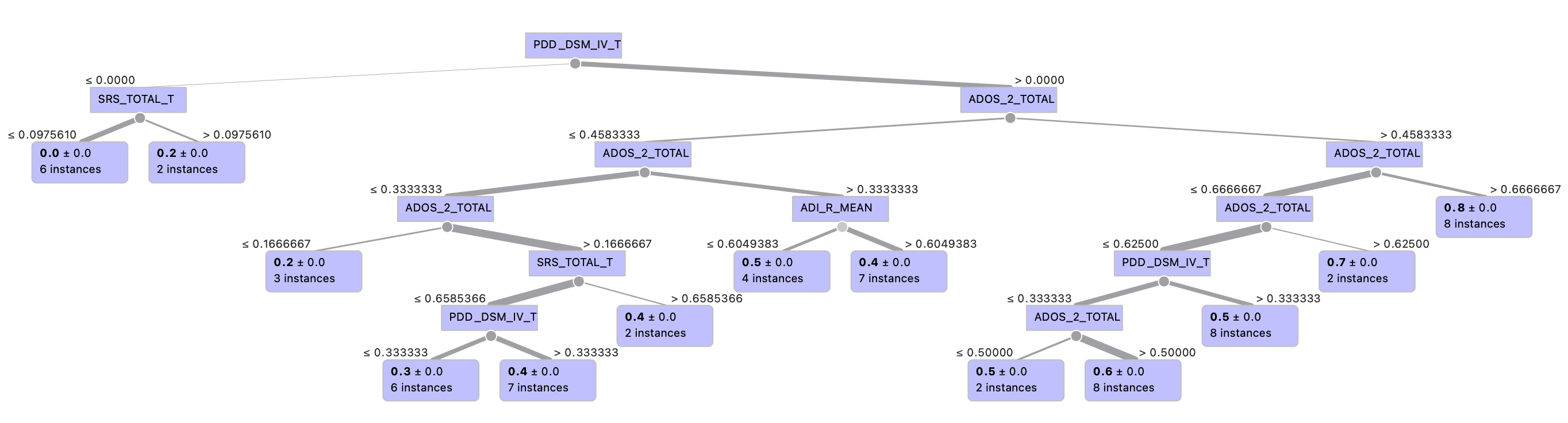
All score features were transformed back to their original numeric values, and normalized so we can apply the regression models.

Summary of Learning Models:

|  |  |  |
| --- | --- | --- |
|  | **ADOS\_G** | **ADOS\_2** |
| Data | Data instances: 1152 Features: ASD\_DSM\_5, PDD\_DSM\_IV\_TR (1), SCQ\_TOTAL, SRS\_TOTAL\_T, RBSR\_6SUBSCALE\_TOTAL, CBCL\_6-18\_TOTAL\_PROBLEM\_T, ADOS\_G\_TOTAL, ADI\_R\_MEAN Meta attributes: NDAR\_GUID, SUB\_ID, SITE\_ID, SEX, AGE\_AT\_SCAN Target: **ADOS\_2\_TOTAL** | Data instances: 1152 Features: ASD\_DSM\_5, PDD\_DSM\_IV\_TR (1), SCQ\_TOTAL, SRS\_TOTAL\_T, RBSR\_6SUBSCALE\_TOTAL, CBCL\_6-18\_TOTAL\_PROBLEM\_T, ADOS\_2\_TOTAL, ADI\_R\_MEAN Meta attributes: NDAR\_GUID, SUB\_ID, SITE\_ID, SEX, AGE\_AT\_SCAN Target: **ADOS\_G\_TOTAL** |
| Learner | **Decision Tree** | |
| Reason we chose the model: | Classic model for classification and regression. Easy to draw conclusions from visual representation of the tree on feature importance. | |
| Model  Parameters | Pruning: at least two instances in leaves, at least 10 instances in internal nodes, maximum depth 100 Splitting: Stop splitting when majority reaches 95% (classification only) Binary trees: Yes | |
| Learner | **SVM** | |
| Reason we chose the model: | We wanted to provide extra flexibility to the linear regression models, in order to identify correlations between a wider distribution of values. We are aware of the load SVM computations may put on system resources, yet since we are looking at relatively small dataset with only ~1,100 records, and since we managed to substantially reduce the number of dimensions, we decided to test this also. | |
| Model  Parameters | SVM type: SVM, C=1.0, ε=0.1 Kernel: Linear Numerical tolerance: 0.001 Iteration limt: 2000 | |
| Learner | **Random Forest** | |
| Reason we chose the model: | We wanted to compare the results of 10 trees to our Decision Tree Learner, and see if any provides better insight. | |
| Model  Parameters | Number of trees: 10 Maximal number of considered features: unlimited Replicable training: No Maximal tree depth: unlimited Stop splitting nodes with maximum instances: 5 | |
| Learner | **Linear Regression** | |
| Reason we chose the model: | Classic model for linear correlations. | |
| Model Parameters | Regularization: Ridge Regression (L2) with α=0.0001 Fit intercept: Yes | |



Tree view – ADOS\_2 – Decision Tree Learner Output

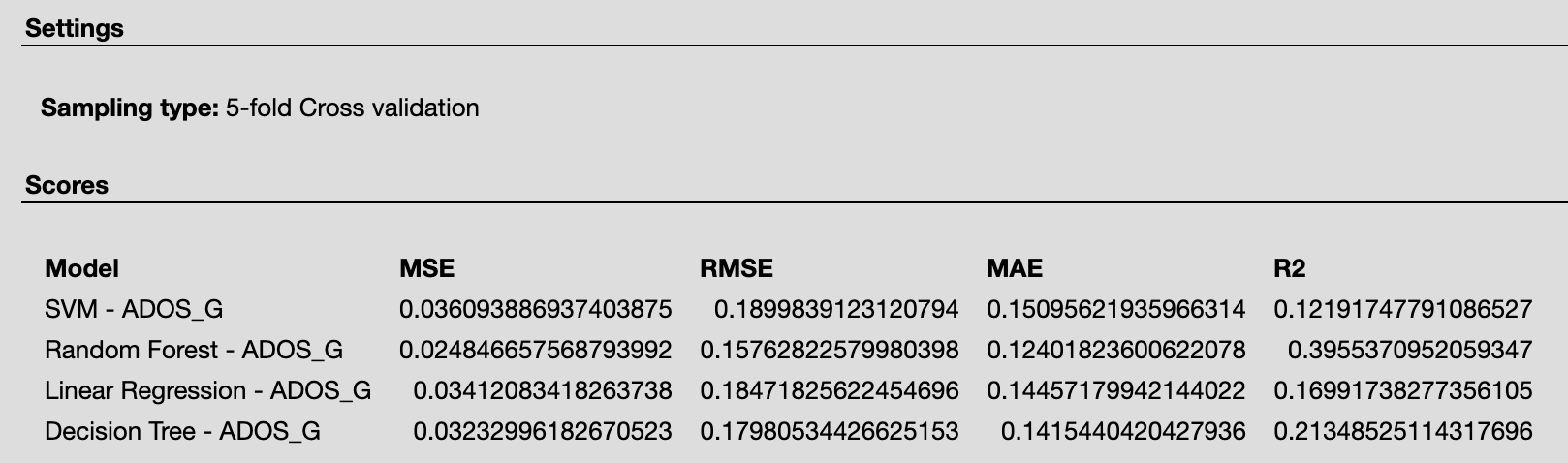


Tree view – ADOS\_G – Decision Tree Learner Output

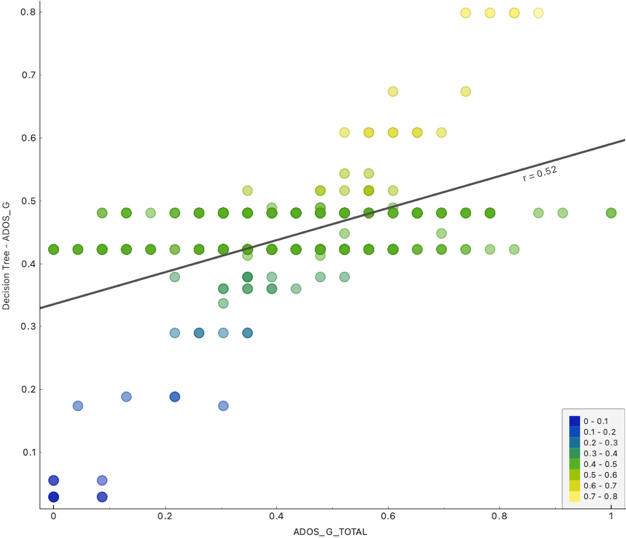
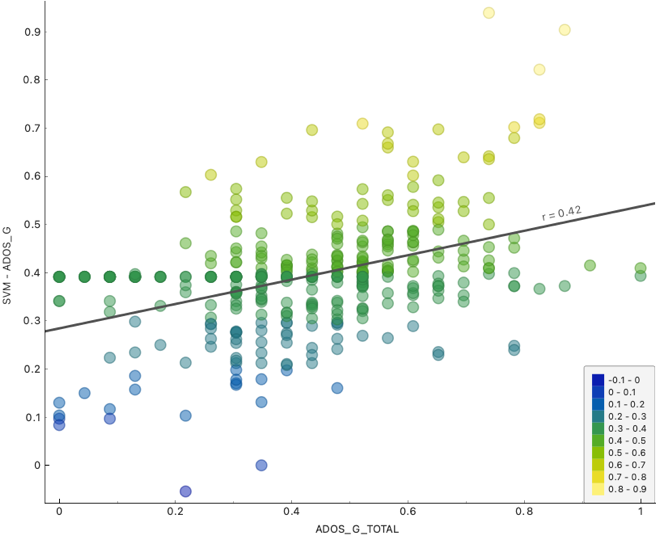
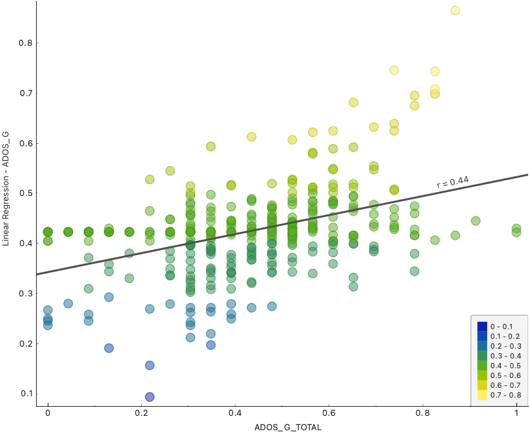
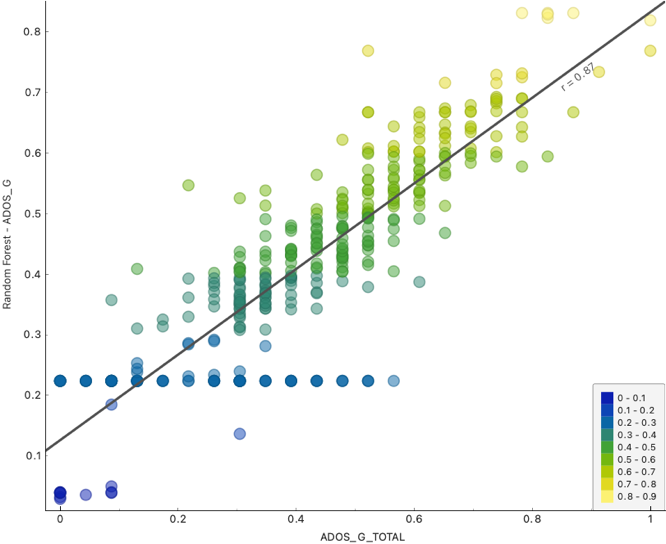
### **Comparison of Learning Models**

Test and Predictions – ADOS G

Results: Random Forest scores exceed the other models, showing   
the highest correlation (R^2 = 0.395 🡺 R = 0.628) with minimal MSE (0.0248), and the best prediction correlation.



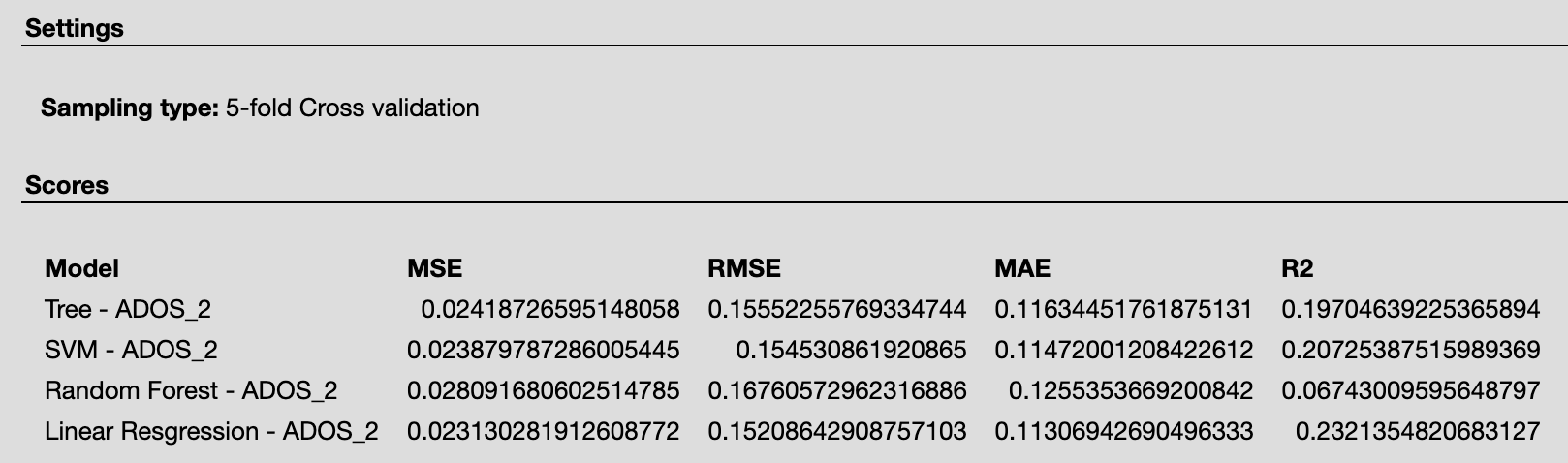
Test Scores - ADOS\_G

Correlations between prediction models and actual values – ADOS\_G

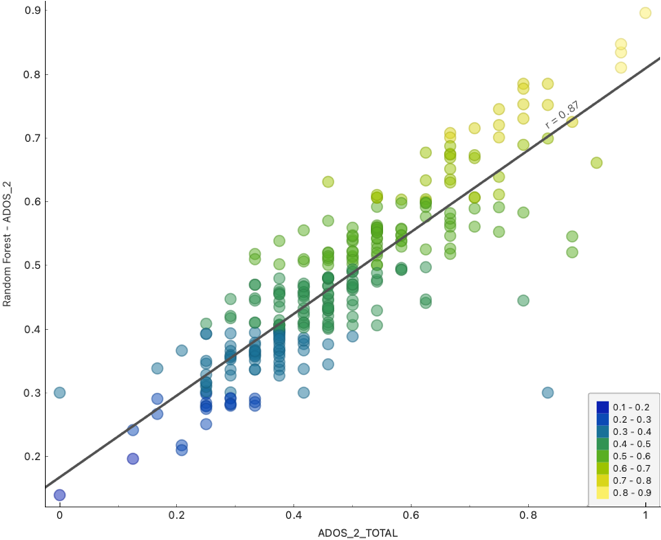
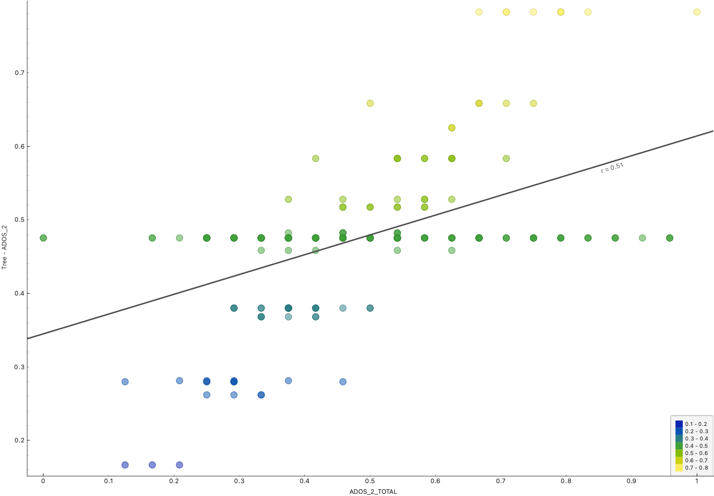
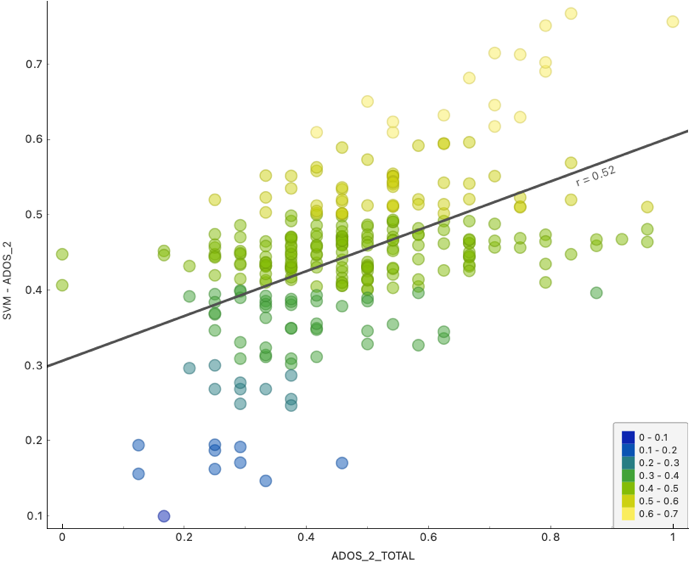
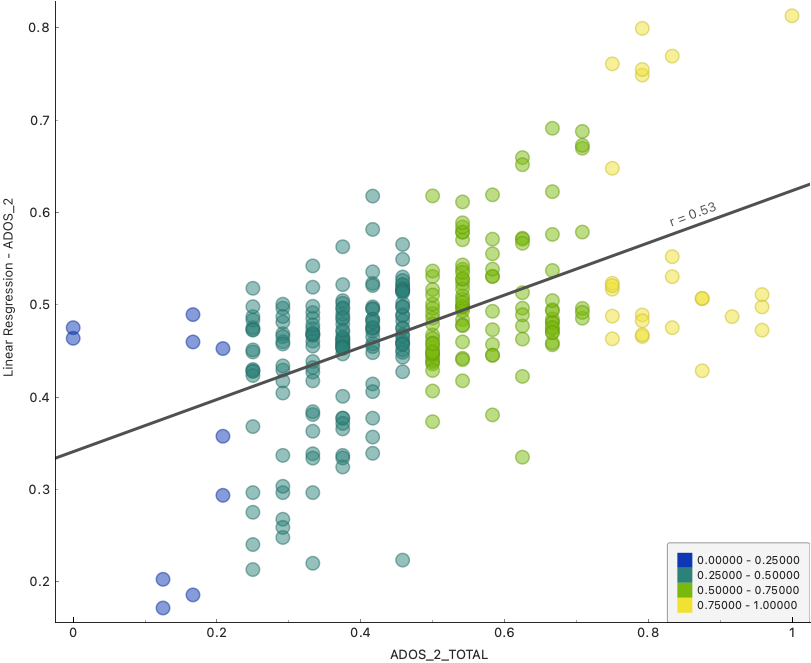
Test and Predictions – ADOS 2

Results: Linear Regression scores exceed the other models, showing   
the highest correlation (R^2 = 0.232 🡺 R = 0.481) with minimal MSE (0.0231).



Test Scores - ADOS\_2

Best prediction correlation model in this case differs from the best test correlation model, and like in ADOS\_G, is Random Forest:

Correlations between prediction models and actual values – ADOS\_2

### **Conclusions:**

Using the ABIDEII dataset we have shown that we can use either ADOS\_G or ADOS\_2 scores to predict our target value (ASD/Non ASD) with high accuracy.

Where ADOS scores were missing, we have shown that using a small set of features that combine other tests’ scores is a good alternative predictor for ADOS scores.

This will allow us to use all the sample records in the dataset for further ASD research and include those missing ADOS scores with high confidence predicted values.

### **Next Steps:**

For further analyses of the ABIDEII, we plan to:

* Reconvert normalized predicted ADOS\_2 values to original score scale
* Convert all ADOS\_G scores (predicted and actual) into ADOS\_2 scores, in order to remain only with the latest version of the ADOS test – as ADOS\_G was retired.
* Identify interesting clusters in the dataset that may give us directions when analyzing MRI scans
* Analyze and compare MRI quality metrics produced by different MRI analyses packages, in order to identify which scans will be included in our next step.