Autism Spectrum Disorder Prediction

Yimiao Ou

Abstract

Autism Spectrum Disorder (ASD or autism) is a neurodevelopmental disorder that impacts brain development, affecting individuals' learning, interaction and communication with others. Early diagnosis of autism is crucial for better treatment of the disease and lower its cost to Society. A timely clinical assessment of autism may be hindered by resource limitations. As a supplement, Autism-Spectrum Quotient (AQ) test enables individuals to self-test their risks of autism. However, the accuracy and applicability of this test is unclear. In this report, we used data mining techniques to study dataset based on AQ test and found that autism can be correctly predicted with attributes extracted from AQ test using machine learning algorithms. Moreover, we identified social interaction and communication as the two most important factors for adult autism prediction.

Introduction

Autism is a developmental disorder that begins early in childhood and lasts throughout an individual's life [1]. The number of children diagnosed with autism has grown steadily in recent years [2,3]. According to a recent report by Public Health Agency of Canada, 1 in 66 children has been identified with autism [3], which leads to huge cost in healthcare, education and social services systems; and early diagnosis and treatment can significantly reduce these.

As a behaviorally defined disorder, autism is characterised by impairments in intellectual ability, social interaction and communication, with repetitive behaviors, restricted interests, sensory hyposensitivities or hypersensitivities to the environment and motor defects [4]. Diagnosis of autism is difficult since no medical test exists to diagnose the disorder. Instead, medical professionals evaluate autism based on a combination of behavior and development symptoms. Two commonly used clinical diagnosis methods for autism are Autism Diagnostic Interview (ADI) and Autism Diagnostic Observation Schedule (ADOS) [5,6]. However, resource limitations can delay or even prevent an individual from receiving a needed autism assessment. Therefore, a non-clinical tool such as the Autism-Spectrum Quotient (AQ), a self-report questionnaire that measures autistic traits, is helpful for individuals to determine whether they need referral to specialists for further clinical services [7].

The AQ test consists of 50 questions, covering 5 subdomains that are characteristic of individuals with autism: social interaction, communication, attention to detail, attention switching and imagination [8]. Each question is scored based on how strongly the subject agrees or disagrees with each statement. A total score is obtained by simply summing up the score for each question. Although higher total score indicates higher risk of autism, a threshold that clearly distinguishes autism individuals from non-autism individuals is difficult to determine since AQ test score has gender and discipline difference, with male scored higher than female and scientists scored higher than non-scientists [7]. Moreover, other neurological diseases also share some symptoms with autism [9,10,11]. Furthermore, the autistic symptoms vary widely from person to person [4]. Even in a single person with autism, the change of symptoms occurs over time as autism is a developmental disorder [12]. Therefore, the accuracy of AQ test in autism evaluation is unclear. Whether the 5 subdomains covered by AQ test are sufficient to accurately identify autism individuals still needs to be assessed.

To address the above questions, we employ Data Mining technique to study a dataset that based on AQ test. We found that autism can be correctly identified using a subset of the AQ test with different machine learning algorithms. In addition, the 5 subdomains of AQ test are sufficient to distinguish autism from non-autism. Among these domains, impairments in social interaction and communication are the two core symptoms of adult autistic patients.

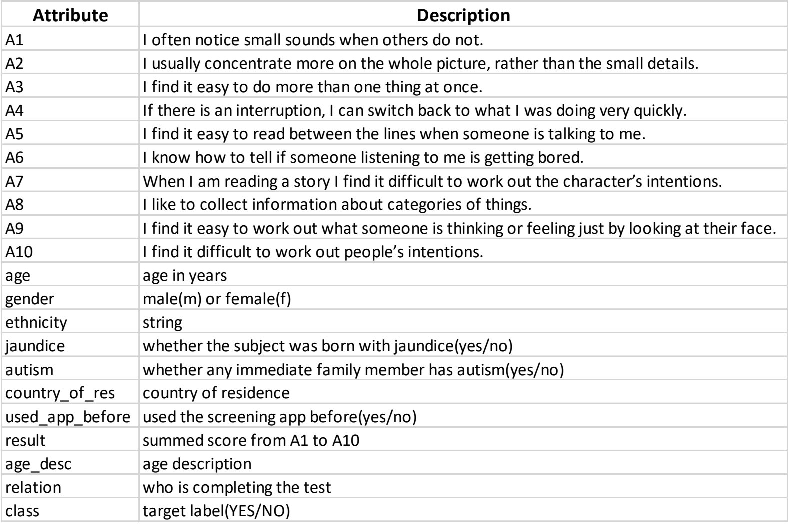
Problem Statement

The goal of this project is to build machine learning models for correctly classifying an individual into autism or non-autism category using dataset based on AQ test and identify the most important symptoms (attributes) for autism prediction. To accomplish this task, we used Python Scikit-Learn library together with RapidMiner to do data mining because this two software can complement each other in the process of data learning.

Dataset Description

In this project, we focused on the dataset "*Autistic Spectrum Disorder Screening Data for Adult*", which was downloaded from the UCI Machine Learning Repository [13]. The dataset contains 704 samples with 21 attributes, including Binary, Integer, String and Boolean values. The attributes and their description are listed in table 1. It is noteworthy that attribute *A1* to *A10* are questions selected from the AQ test and were scored either 1 or 0 depending on whether the answer agreed or disagreed with autism symptoms.

Table 1: Dataset attributes and their description



Attribute	Description
A1	I often notice small sounds when others do not.
A2	I usually concentrate more on the whole picture, rather than the small details.
A3	I find it easy to do more than one thing at once.
A4	If there is an interruption, I can switch back to what I was doing very quickly.
A5	I find it easy to read between the lines when someone is talking to me.
A6	I know how to tell if someone listening to me is getting bored.
A7	When I am reading a story I find it difficult to work out the character's intentions.
A8	I like to collect information about categories of things.
A9	I find it easy to work out what someone is thinking or feeling just by looking at their face.
A10	I find it difficult to work out people's intentions.
age	age in years
gender	male(m) or female(f)
ethnicity	string
jaundice	whether the subject was born with jaundice(yes/no)
autism	whether any immediate family member has autism(yes/no)
country_of_res	country of residence
used_app_before	used the screening app before(yes/no)
result	summed score from A1 to A10
age_desc	age description
relation	who is completing the test
class	target label(YES/NO)

Methods

I. Data Preprocessing

Missing Data

The dataset is incomplete with 95 samples contain missing data. It is difficult to fill in the missing values using substitution methods or model-based methods since majority of the missing occurred in attributes *ethnicity* and *relation* with a nominal type. Therefore, we applied the simple deletion method by removing all samples that contain missing values using Python *pandas* library. To do this, we first used *pandas* to read the input file, then replaced the missing values with NaN from the Python *numpy* library, later called the *dropna* function to delete all NaN values in place.

Feature Selection

In the dataset, some attributes seem irrelevant and redundant, which may introduce noise to the data and cause overfitting problem. Therefore, we applied feature selection to increase the predictive accuracy of the models. The attribute *result* is redundant because it is a summed outcome of the scores from attribute A1 to A10, so we removed it from the dataset. Then we used Decision Tree to select the most relevant and informative attributes by virtue of its built-in feature selection mechanism. Since Decision Tree in Scikit-Learn library can only be applied on numeric values and our data contained non-numeric attributes, we used the Decision Tree operator in RapidMiner to select the features because it accepts non-numeric values.

Outlier

In the dataset there is one sample with *age* value equals to 383, which obviously is an outlier. Since this is a single abnormal value, we just replaced it with the average *age* value calculated from all other samples using *pandas loc* function.

Attribute Type Change

The machine learning methods from Scikit-Learn library only accept numeric values, therefore we need to convert all attributes of our data to numeric type before applying methods from this library. After feature selection, the non-numeric attributes in the dataset were all binominal type, so we directly converted these values to 1 or 0 using *pandas replace* function.

II. Data Mining

Data Mining Functions

The data mining functions we used in this project include Association Rules from RapidMiner for investigating attributes relationship and other 6 different algorithms from Scikit-Learn and RapidMiner for autism prediction. The 6 classification algorithms are as follows.

1. Decision Tree: A very popular classification algorithm that has a built-in feature selection mechanism, which allows us to select relevant and informative attributes to increase classification accuracy.

2. Naïve Bayes: An easy to implement generative model that estimates the probability of a given input belonging to certain class using Bayes' rule.

3. Logistic Regression: A discriminative model for binary classification which directly estimates the probability of a given input belonging to one of the two classes.

4. Support Vector Machine: A commonly used supervised learning technique that is capable of producing different decision boundaries using different kernel functions. It is therefore can be applied to lots of classification problems.

5. Feed-forward Neural Network: A powerful learning model that has the capability of approximating any continuous functions and thus can be applied to various classification scenarios.

6. K-Nearest neighbors: An instance-based learning method that classify samples using distance.

Data Mining Procedures

RapidMiner had been covered in great detail in the class, therefore this part of the report will only focus on Scikit-Learn. Parameter setting for RapidMiner operators used in this project can be seen in supplement Table 1. The whole process of data mining with Scikit-Learn involves the following steps:

1. Import corresponding methods from Scikit-Learn library.

2. Use *pandas* to read data into two separate files, one file contains all attribute values without target labels and the other includes only target labels.

3. Split data into training set and test set using *train_test_split* function. 30% of the data were set aside as testing data (The same ratio of test data was also used in RapidMiner mining).

4. For algorithms that require gradient descent to reach optimal solution such as Logistic Regression and Feed-forward Neural Network or algorithms that use distance as metrics like Support Vector Machine and K-Nearest neighbors, we normalized the data using *StandardScaler* before applying the algorithms.

5. Use 10-fold cross-validation to optimize algorithm parameters with Random Search and select the best parameters set for each model.

6. Fit the model with training data, then predict the target labels for test data and report the results.

Results

Feature Selection Result

After deleting missing values and removing the redundant *result* attribute in the data preprocessing step, we applied Decision Tree on the data to do feature selection using RapidMiner. With prepruning and pruning, the values of the accuracy, recall and precision were: 91.26%, 81.48% and 88% respectively. The attributes used to build this tree were: *A1* to *A10*, *age*, *gender* and *jaundice* (supplement Figure 1). In contrast, with pruning alone, the corresponding values of the tree were: 92.35%, 92.59% and 83.33%, which included attributes: *A1* to *A10*, *age*, *jaundice*, *autism* and *relation* (supplement Figure 2). Although the recall of first tree is lower than the second tree the precision shows the opposite change, indicating that their attribute sets are complementary to each other. After combining the attributes from both trees and excluding the *relation* attribute because it is irrelevant to the data, we obtained the features *A1* to *A10*, *age*, *gender*, *jaundice* and *autism*. These features will be used to fit the models for autism prediction and find association rules.

Classification Result

The values of Accuracy, Precision, Recall and F1-score (YES in *class* attribute indicates autism positive) for the 6 different algorithms we tested are listed below, with table 2 recorded results from Scikit-Learn and table 3 recorded results from RapidMiner.

Table 2: Classification results using Scikit-Learn (with attributes *A1 to A10, age, gender, jaundice* and *autism*)

	Decision Tree	Naïve Bayes	Logistic Regression	Support Vector Machine	Neural Network	k-NN
Accuracy	0.92	0.94	1	1	0.99	0.98
Precision	0.94	0.89	1	1	0.98	0.98
Recall	0.81	0.93	1	1	1	0.95
F1-score	0.87	0.91	1	1	0.99	0.97

As we can see from Table 2, all 6 algorithms performed very well on autism classification as they scored very high on the values we measured. Among them, Logistic Regression, Support Vector Machine and Feed-forward Neural Network produced 100% or close to 100% accuracy, indicating that these three models are able to correctly predict autism using attributes *A1* to *A10*, *age*, *gender*, *jaundice* and *autism*.

Although it is hard to compare the same algorithm in Scikit-Learn and RapidMiner since they may be implemented differently and use distinct parameters, the similar results obtained by RapidMiner in Table 3 below consolidate the above conclusion we made.

Table 3: Classification results using RapidMiner (with attributes *A1* to *A10*, *age*, *gender*, *jaundice* and *autism*)

	Decision Tree	Naïve Bayes	Logistic Regression	Support Vector Machine	Neural Network	k-NN
Accuracy	0.91	0.95	1	1	1	0.95
Precision	0.88	0.92	1	1	1	0.89
Recall	0.81	0.91	1	1	1	0.94
F1-score	0.84	0.92	1	1	1	0.91

Important Attributes (Core Autistic Symptoms)

If we look at the structure of Decision Tree we built for Table 3 (supplement Figure 3), we can see that attributes *A5*, *A6*, *A9* and *A10* resided on the top part of the tree with the highest information gain, indicating that these 4 attributes are the most important attributes for autism prediction. From the association rules mining result (supplement Table 2), we also noticed that these 4 attributes were associated with each other. Based on these observations, we wondered whether models built with these 4 attributes alone are sufficient to produce a good prediction on autism. To test this idea, we applied the 6 classifiers on data contained attributes *A5*, *A6*, *A9* and *A10* using Scikit-Learn. The classification results are listed in Table 4.

Table 4: Classification results using Scikit-Learn (with attributes *A5, A6, A9* and *A10*)

	Decision Tree	Naïve Bayes	Logistic Regression	Support Vector Machine	Neural Network	k-NN
Accuracy	0.91	0.81	0.9	0.9	0.9	0.87
Precision	0.83	0.67	0.81	0.81	0.81	0.95
Recall	0.9	0.81	0.92	0.92	0.92	0.64
F1-score	0.86	0.73	0.86	0.86	0.86	0.77

As we can see from Table 4, Decision Tree, Logistic Regression, Support Vector Machine and Feed-forward Neural Network all achieved over 90% accuracy even only 4 attributes were used to build the models.

From previous section we know that attributes *A1* to *A10* were questions selected from the AQ test and was scored 1 if the answer agreed with adult autism symptom. If we checked the property of each attribute, we can see that they represent the 5 subdomains of AQ test [8]. Interestingly, attribute *A5* and *A6* test an individual's communication skill while attribute *A9* and *A10* test the social interaction skill. The high performance of classifiers using these 4 attributes alone suggests that impairments in communication and social interaction are the core symptoms of adult autistic patients. Based on this finding, if an adult scores 1 for each of these 4 attributes, he/she is at high risk of autism.

Conclusions

In this project, we found that:

1. With attributes *A1*, *A2*, *A3*, *A4*, *A5*, *A6*, *A7*, *A8*, *A9*, *A10*, *age*, *gender*, *jaundice* and *autism* we can predict adult autism with high accuracy using Decision Tree, Naïve Bayes, Logistic Regression, Support Vector Machine, Feed-forward Neural Network and K-Nearest Neighbors. Among these classifiers, Logistic Regression and Support Vector Machine produced a 100% accuracy, indicating that adult autism can be correctly identified using data based on AQ test.

2. With attributes *A5, A6, A9 and A10* alone, Decision Tree, Logistic Regression, Support Vector Machine and Feed-forward Neural Network all achieved over 90% accuracy on adult autism classification. This finding suggests that communication and social interaction are two important factors for accurate prediction of adult autism.

Discussion

All 6 classifiers performed very well on the dataset, with Logistic Regression and Support Vector Machine demonstrated to be perfect models. This may be the consequence that the attributes (symptoms) included in this dataset are characteristic of adult autism patients. Moreover, using feature selection to remove redundant or irrelevant attributes helps to lower the noise and increase accuracy.

Although the classification results are exciting, we need to be cautious when trying to draw conclusions. In this report we focused on only one set of data and received promising results on the classifiers we tested, but we don't know whether they are specific to this dataset or can be applied to other data we have not seen. Especially for the two classifiers that achieved perfect classification, it may be caused by overfitting. To exclude this possibility, we need to test our models using more data from other sources. However, the dataset we studied is clinical data and obtaining other similar dataset to serve our purpose might not be possible due to privacy protection. Even we were lucky to obtain other clinical dataset to assess our models, how to handle missing data is another challenge we need to face because clinical data normally contain lots of missing values. Although removing missing data is simple and worked well for us in this project, it may not be a good practice when the dataset is small and deleting missing value will cause data waste and prediction bias if features are not independent.

Remark

After completing the data mining with adult autism dataset, we were also intrigued to find out whether the same conclusions can be applied to autistic children. To address this question, we used similar procedures to study the dataset "Autistic Spectrum Disorder Screening Data for Children Data Set" [14]. From this dataset, we found that Logistic Regression, Support Vector Machine and Feed-forward Neural Network can classify autistic children with 100% accuracy using attributes *A1* to *A10* and *age*. Moreover, classification using attributes *A3*, *A4*, *A7*, *A9* and *A10* alone can reach 88% accuracy in 5 of the algorithms. *The results were appended at the last page of this report.* Since attribute *A3* and *A4* test an individual's attention switching skill, *A7* tests imagination and *A9* and *A10* test social interaction skill [8], these results indicate that attention switching, imagination and social interaction are the three key factors for predicting autism in children, which is different from the conclusion drawn for adult autism.

References

1. Landa RJ. Diagnosis of autism spectrum disorders in the first 3 years of life. *Nat Clin Pract Neurol*. 2008; 4 (3): 138–47

2. https://www.cdc.gov/ncbddd/autism/data.html

3.https://www.canada.ca/en/public-health/services/publications/diseases-conditions/autism-spectrum-disorder-children-youth-canada-2018.html

4. Baird G, Cass H, Slonims V. Diagnosis of autism. *BMJ.* 2003; 327(7413): 488-93

5. Lord C, Rutter M, Le Couteur A. Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord.* 1994; 24(24):659–85

6. Lord C, Risi S, Lambrecht L, Cook EH Jr, Lambrecht BL, DiLavore PC, Pickles A, Rutter M. et al. The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J Autism Dev Disord.* 2000; 30(30):205–23.

7. Baron-Cohen S, Wheelwright S, Skinner R, Martin J, Clubley E. The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *J Autism Dev Disord.* 2001; 31(31):5–17

8. Allison C, Auyeung B, Baron-Cohen S. Toward brief "red flags" for autism screening: the short Autism Spectrum Quotient and the short quantitative checklist for autism in toddlers in 1,000 cases and 3,000 controls. *J Am Acad Adolesc Psychiatry.* 2012; 51(2):202–12

9. Määttä, T, Tervo-Määttä, T, Taanila, A, Kaski, M, and Iivanainen, M. Mental health, behavior and intellectual abilities of people with Down syndrome. *Down Syndrome Research and Practice*, 2006; 11(1), 37-43

10. Lubs HA, Stevenson RE, Schwartz CE. Fragile X and X-linked intellectual disability: four decades of discover. *Am J Hum Genet.* 2012; 90(4):579-90

11. Dunn W, Bennett D. Patterns of Sensory Processing in Children with Attention Deficit Hyperactivity Disorder. *OTJR: Occupation, Participation and Health.* 2002; 22(1): 4-15

12. Matson JL, Nebel-Schwalm MS. Comorbid psychopathology with autism spectrum disorder in children: An overview. *Res Dev Disabil.* 2007; 28(4):341-52

13. https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult

14.https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Children++

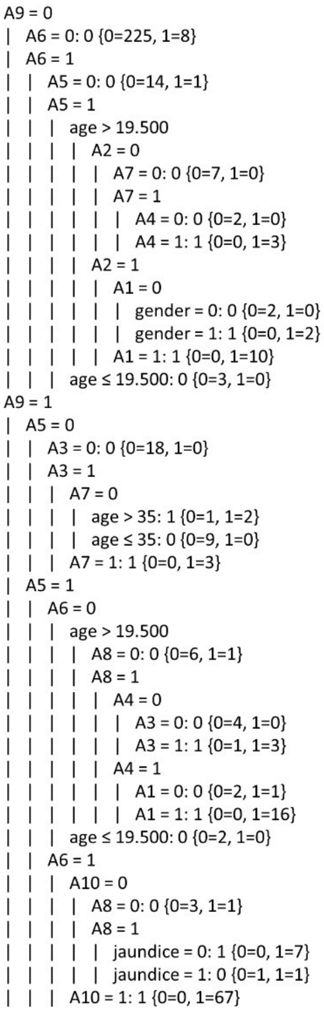
Supplement

Table 1: Parameter setting for RapidMiner operators

Decision Tree	criterion: gain_ratio, maximal depth: 20, confidence: 0.25, minimal gain: 0.1, minimal leaf size: 2, minimal size for split 4, number of prepruning alternatives: 3
Naïve Bayes	laplace correction
Logistic Regression	solver: IRLSM, standardize, add intercept
Support Vector Machine(LibSVM)	svm type: C-SVC, kernel type: linear, C: 1.0, cache size: 200, epsilon: 0.001, shrinking
Neural Net	hidden layers: 1 layer with 10 units, training cycles: 1000, learning rate: 0.05, momentum: 0.1, shuffle, normalize
k-NN	k: 10, weighted vote, measure types: MixedMeasures, mixed measure: MixedEuclideanDistance
Associattion Rules	used attributes A1 to A10, gender, jaundice, autism and class; min confidence: 0.8, min support: 0.95

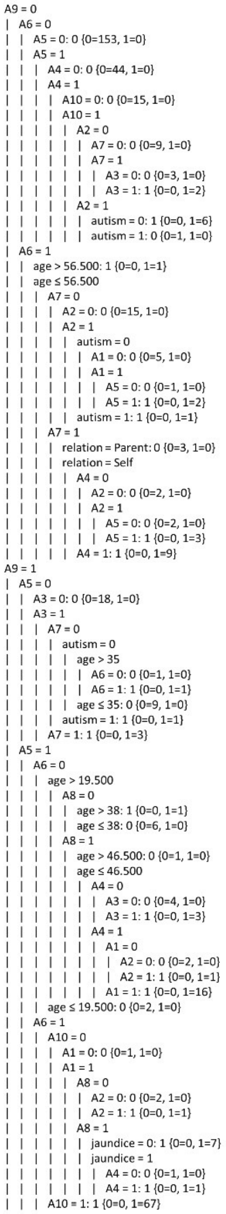
Figures: Decision Tree structures for feature selection

Figure 1: with prepruning and pruning



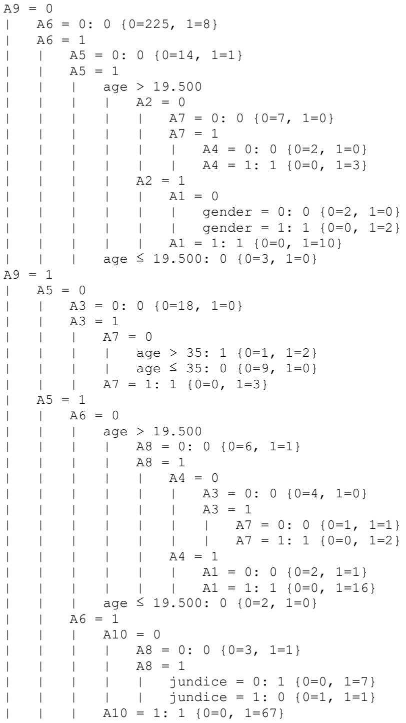
```
A9 = 0
|   A6 = 0: 0 {0=225, 1=8}
|   A6 = 1
|   |   A5 = 0: 0 {0=14, 1=1}
|   |   A5 = 1
|   |   |   age > 19.500
|   |   |   |   A2 = 0
|   |   |   |   |   A7 = 0: 0 {0=7, 1=0}
|   |   |   |   |   A7 = 1
|   |   |   |   |   |   A4 = 0: 0 {0=2, 1=0}
|   |   |   |   |   |   A4 = 1: 1 {0=0, 1=3}
|   |   |   |   A2 = 1
|   |   |   |   |   A1 = 0
|   |   |   |   |   |   gender = 0: 0 {0=2, 1=0}
|   |   |   |   |   |   gender = 1: 1 {0=0, 1=2}
|   |   |   |   |   A1 = 1: 1 {0=0, 1=10}
|   |   |   age ≤ 19.500: 0 {0=3, 1=0}
A9 = 1
|   A5 = 0
|   |   A3 = 0: 0 {0=18, 1=0}
|   |   A3 = 1
|   |   |   A7 = 0
|   |   |   |   age > 35: 1 {0=1, 1=2}
|   |   |   |   age ≤ 35: 0 {0=9, 1=0}
|   |   |   A7 = 1: 1 {0=0, 1=3}
|   A5 = 1
|   |   A6 = 0
|   |   |   age > 19.500
|   |   |   |   A8 = 0: 0 {0=6, 1=1}
|   |   |   |   A8 = 1
|   |   |   |   |   A4 = 0
|   |   |   |   |   |   A3 = 0: 0 {0=4, 1=0}
|   |   |   |   |   |   A3 = 1: 1 {0=1, 1=3}
|   |   |   |   |   A4 = 1
|   |   |   |   |   |   A1 = 0: 0 {0=2, 1=1}
|   |   |   |   |   |   A1 = 1: 1 {0=0, 1=16}
|   |   |   age ≤ 19.500: 0 {0=2, 1=0}
|   |   A6 = 1
|   |   |   A10 = 0
|   |   |   |   A8 = 0: 0 {0=3, 1=1}
|   |   |   |   A8 = 1
|   |   |   |   |   jaundice = 0: 1 {0=0, 1=7}
|   |   |   |   |   jaundice = 1: 0 {0=1, 1=1}
|   |   |   A10 = 1: 1 {0=0, 1=67}
```

Figure 2: with pruning



```
A9 = 0
|   A6 = 0
|   |   A5 = 0: 0 {0=153, 1=0}
|   |   A5 = 1
|   |   |   A4 = 0: 0 {0=44, 1=0}
|   |   |   A4 = 1
|   |   |   |   A10 = 0: 0 {0=15, 1=0}
|   |   |   |   A10 = 1
|   |   |   |   |   A2 = 0
|   |   |   |   |   |   A7 = 0: 0 {0=9, 1=0}
|   |   |   |   |   |   A7 = 1
|   |   |   |   |   |   |   A3 = 0: 0 {0=3, 1=0}
|   |   |   |   |   |   |   A3 = 1: 1 {0=0, 1=2}
|   |   |   |   |   A2 = 1
|   |   |   |   |   |   autism = 0: 1 {0=0, 1=6}
|   |   |   |   |   |   autism = 1: 0 {0=1, 1=0}
|   A6 = 1
|   |   age > 56.500: 1 {0=0, 1=1}
|   |   age ≤ 56.500
|   |   |   A7 = 0
|   |   |   |   A2 = 0: 0 {0=15, 1=0}
|   |   |   |   A2 = 1
|   |   |   |   |   autism = 0
|   |   |   |   |   |   A1 = 0: 0 {0=5, 1=0}
|   |   |   |   |   |   A1 = 1
|   |   |   |   |   |   |   A5 = 0: 0 {0=1, 1=0}
|   |   |   |   |   |   |   A5 = 1: 1 {0=0, 1=2}
|   |   |   |   |   autism = 1: 1 {0=0, 1=1}
|   |   |   A7 = 1
|   |   |   |   relation = Parent: 0 {0=3, 1=0}
|   |   |   |   relation = Self
|   |   |   |   |   A4 = 0
|   |   |   |   |   |   A2 = 0: 0 {0=2, 1=0}
|   |   |   |   |   |   A2 = 1
|   |   |   |   |   |   |   A5 = 0: 0 {0=2, 1=0}
|   |   |   |   |   |   |   A5 = 1: 1 {0=0, 1=3}
|   |   |   |   |   A4 = 1: 1 {0=0, 1=9}
A9 = 1
|   A5 = 0
|   |   A3 = 0: 0 {0=18, 1=0}
|   |   A3 = 1
|   |   |   A7 = 0
|   |   |   |   autism = 0
|   |   |   |   |   age > 35
|   |   |   |   |   |   A6 = 0: 0 {0=1, 1=0}
|   |   |   |   |   |   A6 = 1: 1 {0=0, 1=1}
|   |   |   |   |   age ≤ 35: 0 {0=9, 1=0}
|   |   |   |   autism = 1: 1 {0=0, 1=1}
|   |   |   A7 = 1: 1 {0=0, 1=3}
|   A5 = 1
|   |   A6 = 0
|   |   |   age > 19.500
|   |   |   |   A8 = 0
|   |   |   |   |   age > 38: 1 {0=0, 1=1}
|   |   |   |   |   age ≤ 38: 0 {0=6, 1=0}
|   |   |   |   A8 = 1
|   |   |   |   |   age > 46.500: 0 {0=1, 1=0}
|   |   |   |   |   age ≤ 46.500
|   |   |   |   |   |   A4 = 0
|   |   |   |   |   |   |   A3 = 0: 0 {0=4, 1=0}
|   |   |   |   |   |   |   A3 = 1: 1 {0=0, 1=3}
|   |   |   |   |   |   A4 = 1
|   |   |   |   |   |   |   A1 = 0
|   |   |   |   |   |   |   |   A2 = 0: 0 {0=2, 1=0}
|   |   |   |   |   |   |   |   A2 = 1: 1 {0=0, 1=1}
|   |   |   |   |   |   |   A1 = 1: 1 {0=0, 1=16}
|   |   |   age ≤ 19.500: 0 {0=2, 1=0}
|   |   A6 = 1
|   |   |   A10 = 0
|   |   |   |   A1 = 0: 0 {0=1, 1=0}
|   |   |   |   A1 = 1
|   |   |   |   |   A8 = 0
|   |   |   |   |   |   A2 = 0: 0 {0=2, 1=0}
|   |   |   |   |   |   A2 = 1: 1 {0=0, 1=1}
|   |   |   |   |   A8 = 1
|   |   |   |   |   |   jaundice = 0: 1 {0=0, 1=7}
|   |   |   |   |   |   jaundice = 1
|   |   |   |   |   |   |   A4 = 0: 0 {0=1, 1=0}
|   |   |   |   |   |   |   A4 = 1: 1 {0=0, 1=1}
|   |   |   A10 = 1: 1 {0=0, 1=67}
```

Figure 3: Decision Tree built for Table 3



```
A9 = 0
|   A6 = 0: 0 {0=225, 1=8}
|   A6 = 1
|   |   A5 = 0: 0 {0=14, 1=1}
|   |   A5 = 1
|   |   |   age > 19.500
|   |   |   |   A2 = 0
|   |   |   |   |   A7 = 0: 0 {0=7, 1=0}
|   |   |   |   |   A7 = 1
|   |   |   |   |   |   A4 = 0: 0 {0=2, 1=0}
|   |   |   |   |   |   A4 = 1: 1 {0=0, 1=3}
|   |   |   |   A2 = 1
|   |   |   |   |   A1 = 0
|   |   |   |   |   |   gender = 0: 0 {0=2, 1=0}
|   |   |   |   |   |   gender = 1: 1 {0=0, 1=2}
|   |   |   |   |   A1 = 1: 1 {0=0, 1=10}
|   |   |   age ≤ 19.500: 0 {0=3, 1=0}
A9 = 1
|   A5 = 0
|   |   A3 = 0: 0 {0=18, 1=0}
|   |   A3 = 1
|   |   |   A7 = 0
|   |   |   |   age > 35: 1 {0=1, 1=2}
|   |   |   |   age ≤ 35: 0 {0=9, 1=0}
|   |   |   A7 = 1: 1 {0=0, 1=3}
|   A5 = 1
|   |   A6 = 0
|   |   |   age > 19.500
|   |   |   |   A8 = 0: 0 {0=6, 1=1}
|   |   |   |   A8 = 1
|   |   |   |   |   A4 = 0
|   |   |   |   |   |   A3 = 0: 0 {0=4, 1=0}
|   |   |   |   |   |   A3 = 1
|   |   |   |   |   |   |   A7 = 0: 0 {0=1, 1=1}
|   |   |   |   |   |   |   A7 = 1: 1 {0=0, 1=2}
|   |   |   |   |   A4 = 1
|   |   |   |   |   |   A1 = 0: 0 {0=2, 1=1}
|   |   |   |   |   |   A1 = 1: 1 {0=0, 1=16}
|   |   |   age ≤ 19.500: 0 {0=2, 1=0}
|   |   A6 = 1
|   |   |   A10 = 0
|   |   |   |   A8 = 0: 0 {0=3, 1=1}
|   |   |   |   A8 = 1
|   |   |   |   |   jundice = 0: 1 {0=0, 1=7}
|   |   |   |   |   jundice = 1: 0 {0=1, 1=1}
|   |   |   A10 = 1: 1 {0=0, 1=67}
```

Table 2: Association Rules mining result

[A5, class] --> [A9] (confidence: 0.807)
[class] --> [A9] (confidence: 0.811)
[A6] --> [A10] (confidence: 0.818)
[A6] --> [A5] (confidence: 0.824)
[A10, A9] --> [A5] (confidence: 0.831)
[A5, A9] --> [A10] (confidence: 0.831)
[A5, A9] --> [class] (confidence: 0.831)
[class] --> [A10, A5] (confidence: 0.844)
[class] --> [A10] (confidence: 0.883)
[A5, class] --> [A10] (confidence: 0.889)
[A9, class] --> [A5] (confidence: 0.945)
[class] --> [A5] (confidence: 0.950)
[A10, class] --> [A5] (confidence: 0.956)

Appendix (children autism dataset description and results)

Table 1: Dataset attributes and their description

Attribute	Description
A1	S/he often notices small sounds when others do not.
A2	S/he usually concentrates more on the whole picture, rather than the small details.
A3	In a social group, s/he can easily keep track of several different people's conversations.
A4	S/he finds it easy to go back and forth between different activities.
A5	S/he doesn't know how to keep a conversation going with his/her peers.
A6	S/he is good at social chit-chat.
A7	When s/he is reading a story, s/he finds it difficult to work out the character's intentions or feelings.
A8	When s/he was in preschool, s/he used to enjoy playing games involving pretending with other children.
A9	S/he finds it easy to work out what someone is thinking or feeling just by looking at their face.
A10	S/he finds it hard to make new friends.
age	age in years
gender	male(m) or female(f)
ethnicity	string
jaundice	whether the subject was born with jaundice(yes/no)
autism	whether any immediate family member has autism(yes/no)
country_of_res	country of residence
used_app_before	used the screening app before(yes/no)
result	summed score from A1 to A10
age_desc	age description
relation	who is completing the test
class	target label(YES/NO)

Table 2: Classification results using Scikit-Learn (with attributes *A1* to *A10* and *age*)

	Decision Tree	Naïve Bayes	Logistic Regression	Support Vector Machine	Neural Network	k-NN
Accuracy	0.9	0.76	1	1	1	0.9
Precision	0.86	0.66	1	1	1	0.81
Recall	0.9	0.9	1	1	1	1
F1-score	0.88	0.76	1	1	1	0.89

Table 3: Classification results using RapidMiner (with attributes *A1* to *A10* and *age*)

	Decision Tree	Naïve Bayes	Logistic Regression	Support Vector Machine	Neural Network	k-NN
Accuracy	0.88	0.94	1	1	1	0.88
Precision	0.85	1	1	1	1	0.81
Recall	0.92	0.88	1	1	1	1
F1-score	0.88	0.94	1	1	1	0.9

Table 4: Classification results using Scikit-Learn (with attributes *A3, A4, A7, A9, A10*)

	Decision Tree	Naïve Bayes	Logistic Regression	Support Vector Machine	Neural Network	k-NN
Accuracy	0.88	0.68	0.88	0.88	0.88	0.88
Precision	0.8	0.58	0.83	0.8	0.8	0.83
Recall	0.95	0.86	0.9	0.95	0.95	0.9
F1-score	0.87	0.69	0.86	0.87	0.87	0.86