STUDY ON AUTISM SPECTRUM DISORDER

EKPO Bognan Etienne

17 March 2019

To obtain all ressources related to this report, kindly visit https://github.com/beteko/asd-datamining (https://github.com/beteko/asd-datamining)

PROBLEM UNDERSTANDING

Autism Spectrum Disorder (ASD) is a developmental disorder that affects communication and behavior. People with ASD have difficulty with social communication and interaction, restricted interests, and repetitive behaviors.[5]Although autism can be diagnosed at any age, it is said to be a “developmental disorder” because symptoms generally appear in the first two years of life [5]. It is Known as also “spectrum” disorder because there is a wide variety and intensity of the symptoms that can be observed in a subject. The causes of that disorder are still unclear despite several researches and hypotesis made over the year. Some research suggests that genes can act together with influences from the environment to affect development in ways that lead to ASD. Other studys claim that newborns who have Jaundice are more likely to be diagnosed with Autism.[6]

In view of the rapid growth in the number of ASD cases worldwide, it is crucial to investigate on the possible factors and cause of this disorder. Our study will therefore focus on exploring and mining available records collected from various subjects evaluated at an early stage in order to come up a pattern or association rules that might provide insight on the possible factors that leads to ASD as well as proposing a model that predicts the risks of a toddler deveoloping the disoder based on certain criteria. The study will seek to answer the following questions :

- Is there any proof from the collected dataset that ASD is linked to genes or/and Jaundice ?
- Can we exploit any significant pattern between the symptoms and the gender/sex , family history and Jaundice status of the subjects ?
- Can we predict the risk level of a child/toddler developing ASD based of his on his sex/gender, origin/ethnicity and family history ?

For more information on ASD, Please visit https://en.wikipedia.org/wiki/Autism_spectrum (https://en.wikipedia.org/wiki/Autism_spectrum)

DATA UNDERSTANDING

The medical dataset used for this study was extracted from **Kaggle** and is composed of 1054 records with 19 features about toddlers across the world ranging from ages 1 to 3 years who have undergone some diagnostics to determine whehter or not they are in one of the Spectrum of Autism. This diagnostics is a 10 Multiple Choice Questions (Q-Chat-10) that evaluate several aspects of the mental condition of the subject. Beside the questionaires of the diagnostic, other individuals characteristics that have proved to be effective in detecting the ASD cases from controls in behaviour science have also been collected using the ASDTests app. Find below sample and summary of the ASD dataset

```
##      A1 A2 A3 A4 A5 A6 A7 A8 A9 A10 Age_Mons Qchat.10.Score Sex
## 1053  1  0  0  0  0  0  0  1  0   1       19              3   m
## 1054  1  1  0  0  1  1  0  1  1   0       24              6   m
##           Ethnicity Jaundice Family_mem_with_ASD Who.completed.the.test
## 1053 White European       no                 yes          family member
## 1054          asian      yes                 yes          family member
##      Class.ASD.Traits.
## 1053                No
## 1054               Yes
```

```
##     Age_Mons     Qchat.10.Score   Sex              Ethnicity   Jaundice
##  Min.   :12.00   Min.   : 0.000   f:319   White European:334   no :766
##  1st Qu.:23.00   1st Qu.: 3.000   m:735   asian         :299   yes:288
##  Median :30.00   Median : 5.000           middle eastern:188
##  Mean   :27.87   Mean   : 5.213           south asian   : 60
##  3rd Qu.:36.00   3rd Qu.: 8.000           black         : 53
##  Max.   :36.00   Max.   :10.000           Hispanic      : 40
##                                           (Other)       : 80
##  Family_mem_with_ASD              Who.completed.the.test Class.ASD.Traits.
##  no :884             family member           :1018       No :326
##  yes:170             Health care professional:   5       Yes:728
##                      Health Care Professional:  24
##                      Others                  :   3
##                      Self                    :   4
##
##
```

Let us describe the the columns, their data types and descriptions

- A1, A2 … , A10 : ten behavioural features From (Q-Chat-10) Attributes: A1-A10: Items within Q-Chat-10 in which questions possible answers : “Always, Usually, Sometimes, Rarly & Never” items’ values are mapped to “1” or “0” in the dataset. For questions 1-9 (A1-A9) in Q-chat-10, if the respose was Sometimes / Rarly / Never “1” is assigned to the question (A1-A9). However, for question 10 (A10), if the respose was Always / Usually / Sometimes then “1” is assigned to that question.

- Age_Mons : Age of the toddler in months
- Qchat-10-Score: Score collected, If your child scores more than 3 (Q-chat-10- score) then there is a potential ASD traits otherwise no ASD traits are observed
- Sex : Categorical (m / f)
- Ethnicity : Categorical (asian, Hispanic, Black, White European …)
- Jaundice: (True/ False) Evaluating presence or absence of Jaundice on the subject
- Family_mem_with_ASD : (Yes/No) Is there a family member with ASD ?
- Who completed the test : Categorical (Health Care Professional / Family member/ Self / Others)
- Class/ASD Traits: Result of the diagnostic. Yes in case the subject is in the Spectrum and No if Not in the spectrum

It should be noted that the class variable was assigned automatically based on the score obtained by the user while undergoing the screening process using the ASDTests app.

For more information on the data used in our study, Please visit https://www.kaggle.com/fabdelja/autism-screening-for-toddlers (https://www.kaggle.com/fabdelja/autism-screening-for-toddlers)

DATA PREPARATION

The dataset does not contain any missing values. However, we will assess the quality and validity of some records.

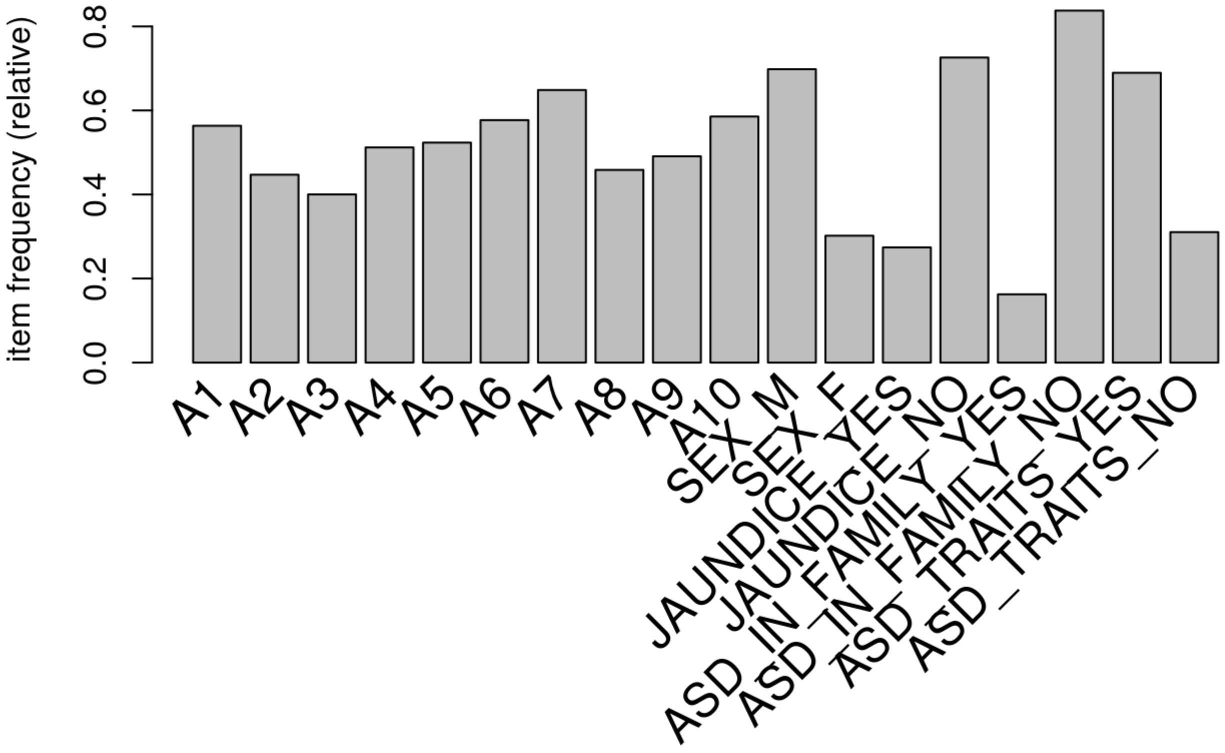
While Exploring the dataset, we noticed that some records were completed by the children thenseves (eg. case No 50). Given the age of the toddlers (1 to 3 years) we deem right to clean/invalidate those records along with all other entities (Others) that completed the form aside the family members and health Care professionals

Also in the same column(Who.completed.the.test), let’s attend to a typo error by changing “Health care professional” to “Health Care Professional”

Prior to the description of the dataset during the previous section (DATA UNDERSTANDING), We know that the feature/Column (Qchat-10-Score) is just the sum of all the scores obtained from question A1 to A10. This derived score is therefore highly correlated with the Class/ASD-Traits since a Toddler is dignosed with ASD whenever he/she obtained a score greater than three (3). We will therefore remove the QChat-10-Score from our dataset so as to prevent the model to overfit due to this strong assumption.

Let us build an itemset(ASD.ITEMSET) based on the ASD dataframe in which caseNo feature represents the [Transaction Ids] and all other selected attributes (A1 ..A10, Sex, Jaundice Status, Family memeber with ASD, ASD status) the items. This transformation will allow us to mine association rules between various features.

Below is the ASD.ITEMSET Frequencies for every attributes



And a sample of the ASD itemset after transformation

```
##     items             transactionID
## [1] {SEX_M,
##      JAUNDICE_YES,
##      ASD_IN_FAMILY_NO,
##      ASD_TRAITS_NO}              296
```

Finally we will split the dataset into Training and test set (78% Training set, 22% Testing Set)that will be used during the Modeling et Evaluation step

```
## [1] "nrow of ASD.TRAINSET : 818 | nrow of ASD.TESTSET : 229"
```

MODELING

In this section, we will apply the Apriori Algorithm on the generated ASD.ITEMSET with some contraints (number_of_items >= 4, support >= 0.04 and confidence >= 40 %) in order to deduce the most relevant associations rules as shown below. The first 10 association rule will be selected being sorted by LIFT in decreasing order.

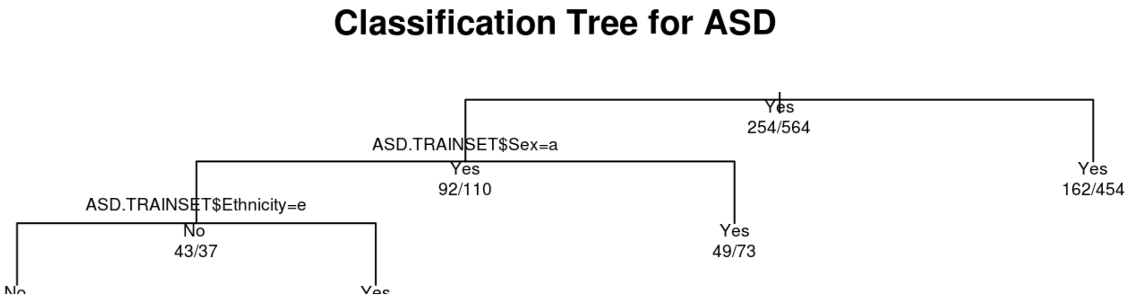
```
##     lhs                      rhs                 support confidence      lift count
## [1] {SEX_M,
##      JAUNDICE_YES,
##      ASD_IN_FAMILY_YES} => {ASD_TRAITS_YES} 0.04011461  0.8235294 1.1942317    42
## [2] {SEX_M,
##      JAUNDICE_YES,
##      ASD_IN_FAMILY_NO}  => {ASD_TRAITS_YES} 0.11843362  0.7898089 1.1453323   124
## [3] {SEX_M,
##      JAUNDICE_NO,
##      ASD_IN_FAMILY_NO}  => {ASD_TRAITS_YES} 0.29894938  0.7033708 1.0199851   313
## [4] {SEX_M,
##      JAUNDICE_NO,
##      ASD_IN_FAMILY_YES} => {ASD_TRAITS_YES} 0.04871060  0.6538462 0.9481675    51
## [5] {SEX_F,
##      JAUNDICE_YES,
##      ASD_IN_FAMILY_NO}  => {ASD_TRAITS_YES} 0.04202483  0.6285714 0.9115156    44
## [6] {SEX_F,
##      JAUNDICE_NO,
##      ASD_IN_FAMILY_NO}  => {ASD_TRAITS_YES} 0.12034384  0.6146341 0.8913046   126
```

We will focus on Rule [2] and [3] since they have a significant support and positive correlation (lift > 1) and Rules [5] and [6] for their relatively significant supports and negative correlations (lift < 1) . Those association rules suggest that male toddlers with a Positive Jaunice status and no family history of ASD are more likely to be diagnosed with ASD than female with the same attributes. Also, Male with no trace of Jaunice and no family history of ASD are likewise more likely to be diagnosed with ASD than their female counterpart with the same observations.

```
##     lhs                     rhs                support confidence     lift count
## [1] {A3,
##      A6,
##      SEX_M,
##      JAUNDICE_NO,
##      ASD_IN_FAMILY_NO} => {ASD_TRAITS_YES} 0.1251194          1 1.450139   131
## [2] {A3,
##      A7,
##      SEX_M,
##      JAUNDICE_NO,
##      ASD_IN_FAMILY_NO} => {ASD_TRAITS_YES} 0.1270296          1 1.450139   133
```

Per the above result (after observing more than 500 parttens), we could not confidently deduce any significant association rules between the symptoms (A1 - A10) and the ASD Positive Subject attributes (Gender, Jaundice Status and Family History).

Based on previous surveys we will select the following features (gender, Jaundice_status , family_member_with_ASD and Ethnicity) in order to propose a classifier that will predict the risk of a toddler before even taking the ASD Test/dianosis. We deem right to select a Classification and Regression Tree (CART) for this purpose firt of all because of its intepretability as we are in the medical field as well as its performance in classification task under limited number of features. We will initially train our model on the training set (ASD.TRAINSET) and visualize the initial decision tree.

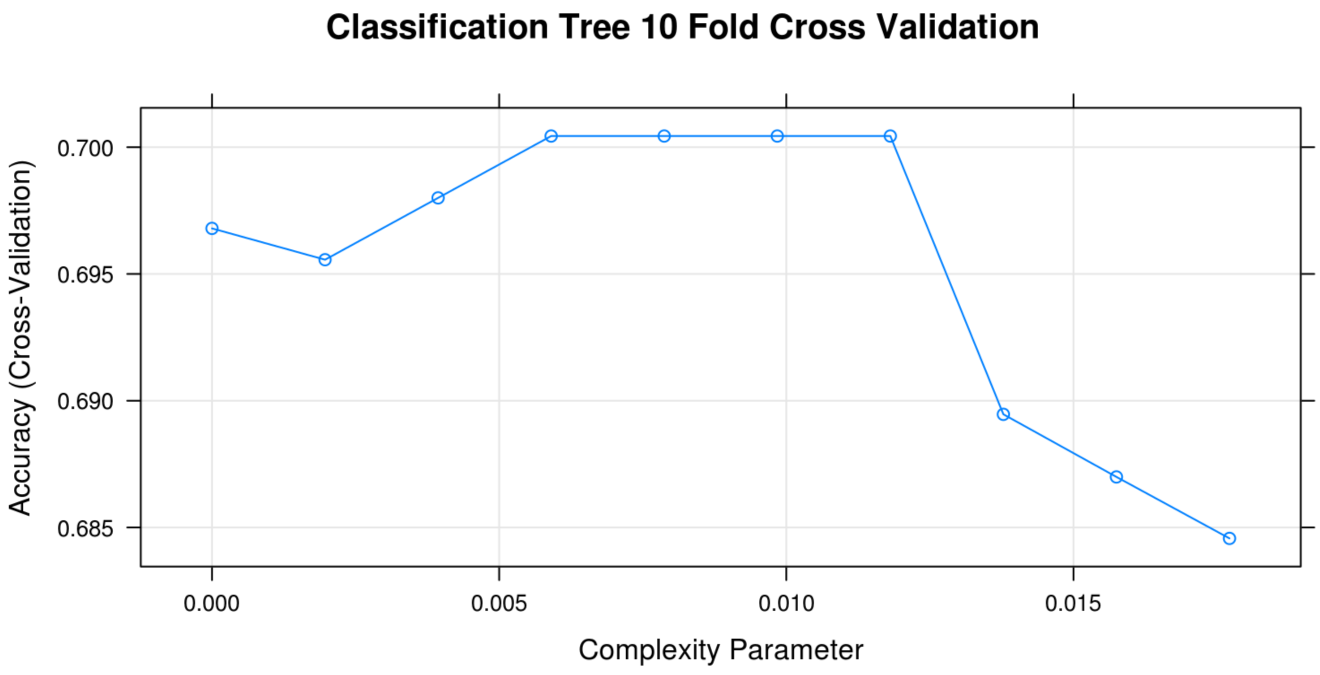


MODEL EVALUATION

In order to obtain an accurate measure of the performance of our Decision Tree Classifier and select the optimal hyperparameters, we applied a 10-Fold Cross Validation on our training (ASD.TRAINSET) set while tuning the complexity parameter of the model with the seed set to 3 .

Find below the details associated with the model evaluation and hyperparemeter selection.

```
## CART
##
## 818 samples
##   4 predictors
##   2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 736, 737, 735, 737, 737, 737, ...
## Resampling results across tuning parameters:
##
##   cp           Accuracy   Kappa
##   0.000000000  0.6967964  0.13062878
##   0.001968504  0.6955618  0.12137877
##   0.003937008  0.6980009  0.12366692
##   0.005905512  0.7004399  0.11897218
##   0.007874016  0.7004399  0.11897218
##   0.009842520  0.7004399  0.11897218
##   0.011811024  0.7004399  0.11897218
##   0.013779528  0.6894632  0.06485362
##   0.015748031  0.6869940  0.04681579
##   0.017716535  0.6845697  0.02878535
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.01181102.
```



Per the result above, the best complexity parameter selected is 0.0118 which yield a score of 0.70. We will now apply the best-performing classifier on the test dataset and generate the confusion matrix to observe the specificity and sensitivity of the model towards test subjects (ASD.TESTSET).

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction  No Yes
##        No   11   5
##        Yes  60 153
##                                          
##                Accuracy : 0.7162         
##                  95% CI : (0.653, 0.7736)
##     No Information Rate : 0.69           
##     P-Value [Acc > NIR] : 0.217          
##                                          
##                   Kappa : 0.1567         
##  Mcnemar's Test P-Value : 2.115e-11      
##                                          
##             Sensitivity : 0.15493        
##             Specificity : 0.96835        
##          Pos Pred Value : 0.68750        
##          Neg Pred Value : 0.71831        
##              Prevalence : 0.31004        
##          Detection Rate : 0.04803        
##    Detection Prevalence : 0.06987        
##       Balanced Accuracy : 0.56164        
##                                          
##        'Positive' Class : No             
## 
```

The specificity and sensitivity of the model demonstate the accuracy of the model in classifying the subject with ASD. Again,one need to note that a new subject classified as ASD positive by the model does not imply that the subject has the disorder but rather is at risk of developping ASD.

CONCLUSION

This studies allowed us to explore various patterns and possible association between Toddlers with ASD Traits and their family history, sex/gender, ethnicity , Jaundice status , Symptoms and many others. Per the result obtained in the section above ,we can counclude that ASD is strongly associated with genes and the likelihood of male with Jaundice developping the disorder is more important than in female. Exploration of the dataset did not provide any clear association betwen the symptoms (A1 … A10) and the subject attributes.

Beside the above findings, we also proposed a relatively accurate Descision Tree classifier (CART) based on the ASD dataset that could predict whether a child might be at risk of having ASD based on his gender, family history, jaunice status and origin. This aims at helping doctors around the world to know the likelihood of a toddler developping the disorder before even undergoing the ASD test.

REFERENCES

1. Tabtah, F. (2017). Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment. Proceedings of the 1st International Conference on Medical and Health Informatics 2017, pp.1-6. Taichung City, Taiwan, ACM.

2. Thabtah, F. (2017). ASDTests. A mobile app for ASD screening. www.asdtests.com [accessed December 20th, 2017].

3. Thabtah, F. (2017). Machine Learning in Autistic Spectrum Disorder Behavioural Research: A Review. Informatics for Health and Social Care Journal.

4. Thabtah F, Kamalov F., Rajab K (2018) A new computational intelligence approach to detect autistic features for autism screening. International Journal of Medical Infromatics, Volume 117, pp. 112-124.

5. https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd/index.shtml (https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd/index.shtml)

6. https://www.webmd.com/brain/autism/news/20101011/jaundice-in-newborns-may-be-linked-to-autism#1 (https://www.webmd.com/brain/autism/news/20101011/jaundice-in-newborns-may-be-linked-to-autism#1)

7. https://www.kaggle.com/fabdelja/autism-screening-for-toddlers/version/1 (https://www.kaggle.com/fabdelja/autism-screening-for-toddlers/version/1)