Università degli studi di Milano
M.Sc. Data Science for Economics
Statistical Learning Project Report
Written by Arina Lopukhina

Autism Spectrum Disorder Diagnosing: Comparative Analysis

1. Introduction

Autism spectrum disorder (ASD) currently remains one of the least understood mental health illnesses. ASD can manifest itself a wide range of symptoms, including but not limited to difficulties with socializing, focusing, sensitivity to lifestyle changes, meltdowns, and many more, as they might vary from one person to another, even for patients been diagnosed with the same type of ASD. Not only there is a lack of guidance regarding treatment possibilities since every case requires a tailored approach, but there are also a lot of debates on how ASD can be or should be diagnosed. Medical professionals usually rely on surveying patients to verify if they possess certain traits or symptoms, as well as considering one's family medical history, namely, whether close relatives were affected by ASD. However, when it comes to children with suspected ASD, it might get trickier. Similar to many other medical conditions, diagnostic processes for children and adults are carried out differently due to the inability of the former to properly express themselves and explain their feelings or communicate their thoughts clearly. For adults, there is the AQ-10-Adult survey, which is usually filled out by a patient, while for children there is a different survey with simpler and more straightforward questions called Q-Chat-10. Both are widely used for medical research and medical practice.

This project aims to assess which features make one more likely to be diagnosed with autism in both young age (below 3 years old) and adulthood, using the datasets published by UC Irvine in 2018. Both datasets contain very similar features, so it was interesting to not only compare the exploratory analysis but also see how the same models treat similar data but for such different age groups. It was particularly intriguing to see whether it is feasible to make an assumption that one's demographic factors such as sex and ethnicity, as well as the family medical history, have the same effects among children and adults.

The initial hypothesis is that one's family medical history is going to be the main factor for having autism in both toddlers and adults, while ethnical background, sex, and age play little to no role. To access this, both supervised and unsupervised statistical learning methods were employed. Among supervised learning techniques, different types of regression were used (linear, logistic, Naïve Bayes, Ridge, Lasso, and Elastic net). Also, kNN and a random forest analysis was performed. In terms of unsupervised techniques, factor analysis was chosen as k-means clustering or PCA were not suitable for the datasets with over 90% of categorical data. Additionally, Apriori algorithm was used to try to map relationships between certain features and their link to the final outcome of diagnosing.

2. **Exploratory data analysis**

2.1.Dataset information

Both datasets are predominantly based on a 10-question survey, which has two versions – one for children, one for adults. You can see the sample AQ-10-Adult created by the British National Institute for Health Research below:

Figure 1:AQ-10-Adult survey

	Please tick one option per question only:	Definitely agree	Slightly agree	Slightly disagree	Definitely disagree
1	I often notice small sounds when others do not.				
2	I usually concentrate more on the whole picture, rather than the small details.				
3	I find it easy to do more than one thing at once				
4	If there is an interruption, I can switch back to what I was doing very quickly				
5	I find it easy to 'read between the lines' when someone is talking to me				
6	I know how to tell if someone listening to me is getting bored				
7	When I'm reading a story I find it difficult to work out the characters' intentions				
8	I like to collect information about categories of things (e.g. types of car, bird, train, plant etc.)				
9	I find it easy to work out what someone is thinking or feeling just by looking at their face				
10	I find it difficult to work out people's intentions				

The adult dataset originally contains 704 observations and 20 variables, ten of which are binary responses to the above-mentioned questions. The other 10 being the: result column (0 to 10, where 6< means likely autistic), sex, age, country of residence, ethnicity, jaundice, autism (whether these conditions have been present in the individual themselves or their close family member), whether one has taken this test before, whether one filled out the survey themselves or with a help of a doctor/family member/etc., and, finally, whether one was suspected to have autism based on the results of the survey.

Out of approximately 700 samples, roughly one thirds was people with suspected autism and the remaining two-thirds were neurotypical subjects, hence the ratio is 2:1 (neurotypical to autistic). Upsampling tactics were applied, however they were noted to have a negative impact on the results as they have been associated with overfitting as well as manipulating coefficients.

The children dataset contains responses to 10 questions from Q-Chart-10 presented below:

Figure 2: Q-Chart-10 survey

For each item, please circle the response which best applies to your child:

		A	B	C	D	E
1	Does your child look at you when you call his/her name?	Always	Usually	Sometimes	Rarely	Never
2	How easy is it for you to get eye contact with your child?	Very easy	Quite easy	Quite difficult	Very difficult	Impossible
3	Does your child point to indicate that s/he wants something? (e.g. a toy that is out of reach)	Many times a day	A few times a day	A few times a week	Less than once a week	Never
4	Does your child point to share interest with you? (e.g. pointing at an interesting sight)	Many times a day	A few times a day	A few times a week	Less than once a week	Never
5	Does your child pretend? (e.g. care for dolls, talk on a toy phone)	Many times a day	A few times a day	A few times a week	Less than once a week	Never
6	Does your child follow where you're looking?	Many times a day	A few times a day	A few times a week	Less than once a week	Never
7	If you or someone else in the family is visibly upset, does your child show signs of wanting to comfort them? (e.g. stroking hair, hugging them)	Always	Usually	Sometimes	Rarely	Never
8	Would you describe your child's first words as:	Very typical	Quite typical	Slightly unusual	Very unusual	My child doesn't speak
9	Does your child use simple gestures? (e.g. wave goodbye)	Many times a day	A few times a day	A few times a week	Less than once a week	Never
10	Does your child stare at nothing with no apparent purpose?	Many times a day	A few times a day	A few times a week	Less than once a week	Never

Overall, 1024 children were surveyed as well as the same additional data points were collected as for the adults, such as age, sex, ethnicity, family medical history, and so on. Additionally, around 650 of them were suspected to have autism, while the remaining 375 were classified as neurotypical. Similarly to the adult dataset, the attempt to even out the sampling has only created additional difficulties for model construction and result interpretation.

2.2. Exploratory data analysis

Initially, a z-test was performed to establish whether the features in the dataset have any effect on the target variable of having autism, and it showed statistically significant results in both datasets. Afterwards, exploratory data analysis was performed to asses whether there are any trends that can be seen before applying the models.

As can be seen from Figures 3 and 4, there is no major difference between the distributions of autistic and neurodivergent respondents with respect to presence of close relatives diagnosed with ASD as well as whether one has experienced jaundice at birth. However, one can notice that there are way more toddlers with jaundice at birth, while adults barely have any cases of jaundice. However, this could be attributed to the fact that toddlers' questionnaires were filled out mainly by their parents or medical professionals who are more aware of whether their child/patient had jaundice, while an average adult might not know such details about themselves and hence is more inclined to respond with a 'no'.

Figure 3: Percentages of respondent with and without ASD and having relatives with ASD

(left: toddlers, right: adults)

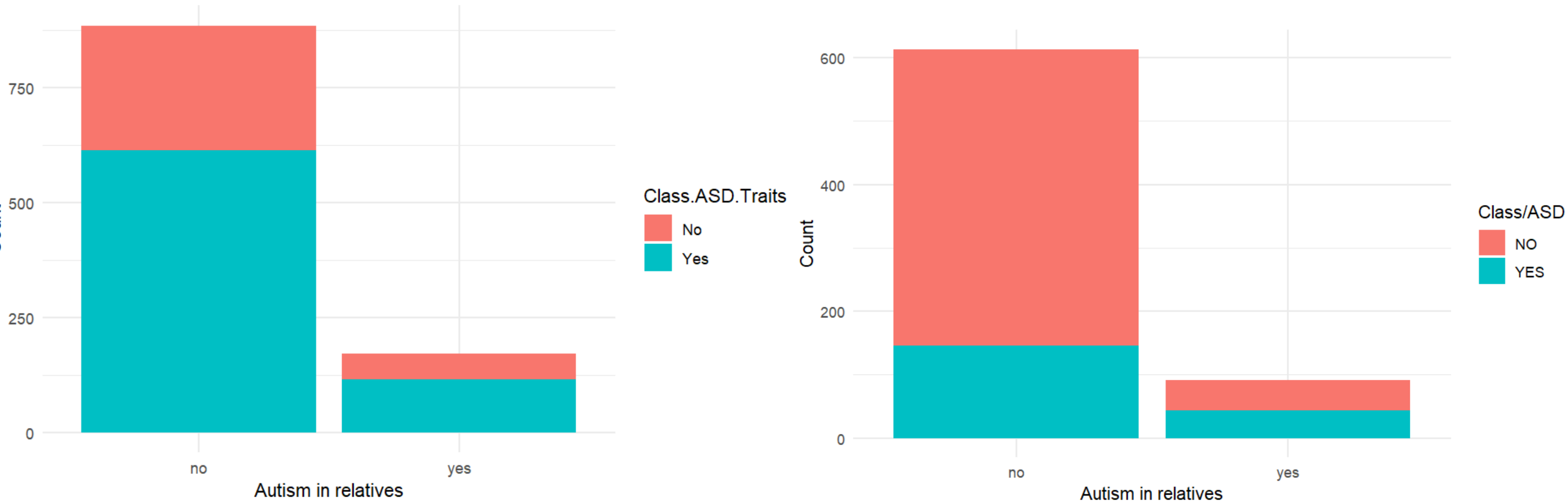


Figure 4: Percentages of respondent with and without ASD and being born with jaundice

(left: toddlers, right: adults)

Additionally, Figures 5 and 6 show the distribution of diagnosis by ethnicity. It can be noted that most respondents are predominantly White, Asian, or Middle Eastern among both populations. Also, White patients seem to be diagnosed with autism more frequently than representatives of other ethnicities. Although, it must be noted that in the case of the adults dataset, there were 95 rows with missing ethnicity data, so in order to avoid reducing the sample size, it was decided to place them in a separate category called ‘missing ethnicity’, which brings greater ambiguity.

Figure 5: ASD diagnosis by ethnicity (Toddlers).

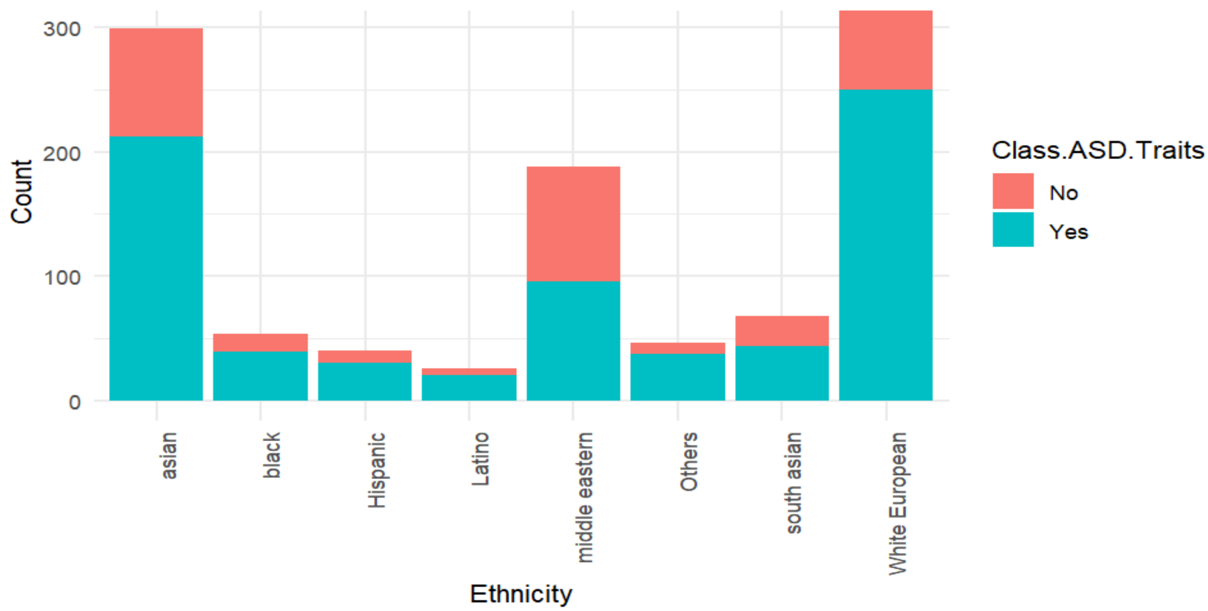
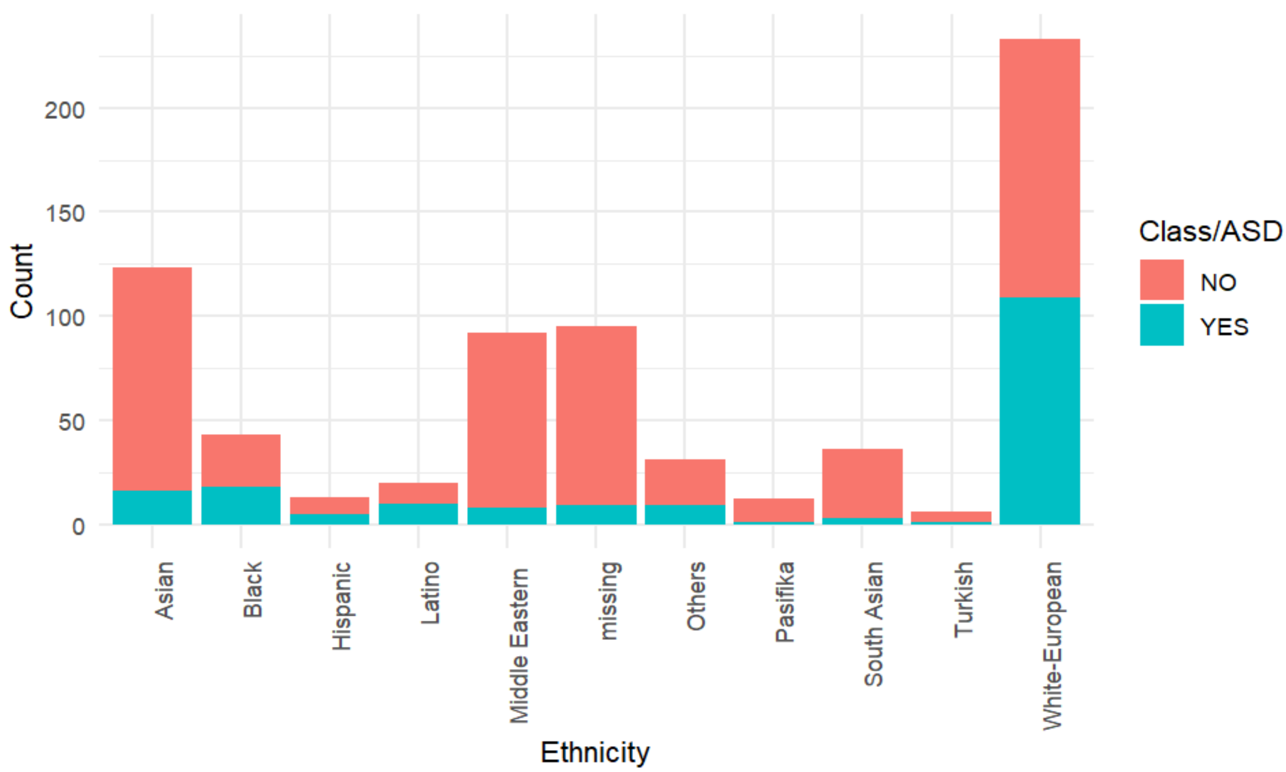
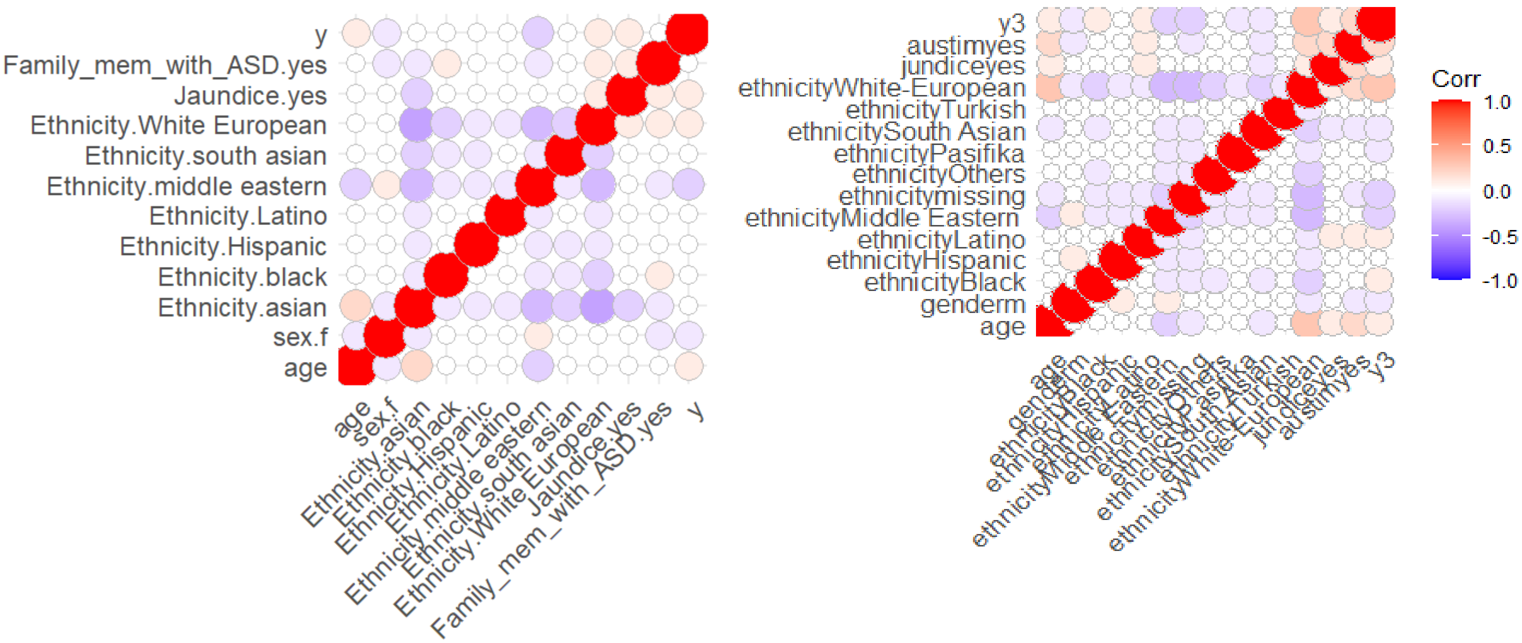


Figure 6: ASD diagnosis by ethnicity (Adults).



In terms of correlation coefficients (see Figure 7), there are no variables that immediately stand out. However in both populations, there several variables which seem mildly correlated with the target variable and shall be investigated further, such as age, jaundice, family member with ASD, etc.

Figure 7: Correlation matrices (left: toddlers, right: adults)



2.3. Data cleaning

In the adults dataset, there were some missing values among several variables, so they had to be deleted or, in case of 'age', replaced with a median since the data was not normally distributed. Also, log-age was created to compensate for the Gamma-like distribution. Additionally, age outliers were identified (likely a data entry mistake) and erased. Afterwards, one-hot-encoding (OHE) was performed for the categorical variables.

After OHE, there were very few variables per country of residence which created multicollinearity, so it was decided to exclude them from analysis and rely on ethnicities only. Consequently, a few other variables like whether the patient responded to the survey on their own or whether they have done it before, etc., were excluded as they seemed too loosely related to the target variable. Most importantly, question responses were also excluded as they had created linear dependence between the target variable since the result was determined by summing up the points received for each question. Thus, any model could 100% accurately predict whether one has ASD or not just from the questions alone, but this does not correspond the goals of this paper.

At the end, the adult dataset was comprised of 704 observations across 14 variables, one of which is numeric and the rest are categorical. In the toddlers dataset, the same procedures were performed (although it did not contain as many missing variables) and at the end there were 1054 observations for 11 categorical variables and 1 numeric.

3. Supervised learning techniques

The main goal of supervised techniques was identifying the most relevant features in both datasets. As previously stated, predictive power of the model without the test scores was expected to be very low, so in this analysis, focus was laid mostly on identifying the statistically significant factors among the ones given.

3.1 Regression

The initial idea behind performing regression was to see which features would have a low-enough p-value in both datasets. With simple multivariate linear regression, the toddler dataset suggested the following results: being a girl slightly reduces chances of having ASD, so does coming from a Middle Eastern ethnicity, whereas experiencing jaundice at birth slightly contributes to being diagnosed with autism.

Figure 8: Multivariate linear regression - Toddlers

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.8650 -0.5364  0.2278  0.2961  0.6270

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                   0.60674    0.18957   3.201  0.00142 **
age                           0.05884    0.04982   1.181  0.23793
sex.f                        -0.08083    0.03479  -2.324  0.02039 *
Ethnicity.asian              -0.08725    0.07934  -1.100  0.27177
Ethnicity.black              -0.01953    0.10102  -0.193  0.84672
Ethnicity.Hispanic           -0.05401    0.10516  -0.514  0.60769
Ethnicity.Latino             -0.03265    0.12722  -0.257  0.79751
`Ethnicity.middle eastern`   -0.25975    0.08362  -3.106  0.00196 **
`Ethnicity.south asian`      -0.13693    0.09454  -1.448  0.14788
`Ethnicity.White European`   -0.03472    0.07901  -0.439  0.66050
Jaundice.yes                  0.07203    0.03597   2.002  0.04556 *
Family_mem_with_ASD.yes      -0.05632    0.04258  -1.323  0.18636
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4517 on 832 degrees of freedom
Multiple R-squared:  0.05222,   Adjusted R-squared:  0.03969
F-statistic: 4.167 on 11 and 832 DF,  p-value: 5.135e-06
```

With respect to logistic regression, the results were confirmed for toddlers. The same three features were identified as statistically significant, with roughly similar coefficients (see Figure 9):

Figure 9: Logistic regression - Toddlers

```
Call:
glm(formula = y_train ~ ., family = binomial(link = "logit"),
    data = X_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9332  -1.2390   0.7156   0.8267   1.4549

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                  0.4748     0.9356   0.507  0.61184
age                          0.2765     0.2414   1.145  0.25205
sex.f                       -0.3865     0.1669  -2.316  0.02057 *
Ethnicity.asian             -0.4660     0.4267  -1.092  0.27476
Ethnicity.black             -0.1138     0.5420  -0.210  0.83370
Ethnicity.Hispanic          -0.2978     0.5508  -0.541  0.58870
Ethnicity.Latino            -0.1847     0.6625  -0.279  0.78045
`Ethnicity.middle eastern`  -1.2045     0.4390  -2.744  0.00608 **
`Ethnicity.south asian`     -0.6990     0.4887  -1.430  0.15260
`Ethnicity.White European`  -0.1886     0.4283  -0.440  0.65971
Jaundice.yes                 0.3734     0.1845   2.024  0.04298 *
Family_mem_with_ASD.yes     -0.2825     0.2070  -1.365  0.17226
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1039.14  on 843  degrees of freedom
Residual deviance:  995.89  on 832  degrees of freedom
AIC: 1019.9
```

For adults, there was more significance among certain ethnicities: for instance, coming from Black, Latino, or White background was associated with higher chances of developing ASD, and so was having relatives with autism.

Figure 10:Multivariate linear regression - Adults

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.6225 -0.2723 -0.1004  0.3755  0.9672

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 0.19776    0.20762   0.953   0.3412
age                        -0.01693    0.06069  -0.279   0.7804
genderm                    -0.03597    0.03460  -1.040   0.2990
ethnicityBlack              0.34504    0.08194   4.211 2.97e-05 ***
ethnicityHispanic           0.26172    0.11941   2.192   0.0288 *
ethnicityLatino             0.34956    0.10812   3.233   0.0013 **
`ethnicityMiddle Eastern `  -0.07588    0.06168  -1.230   0.2192
ethnicitymissing           -0.03841    0.06125  -0.627   0.5308
ethnicityOthers             0.15780    0.08669   1.820   0.0693 .
ethnicityPasifika          -0.03119    0.14054  -0.222   0.8244
`ethnicitySouth Asian`     -0.06986    0.08375  -0.834   0.4046
ethnicityTurkish            0.05725    0.18485   0.310   0.7569
`ethnicityWhite-European`   0.28381    0.05291   5.364 1.20e-07 ***
jundiceyes                  0.06257    0.05833   1.073   0.2839
austimyes                   0.11221    0.05410   2.074   0.0385 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4022 on 549 degrees of freedom
Multiple R-squared:  0.168,     Adjusted R-squared:  0.1468
F-statistic: 7.921 on 14 and 549 DF,  p-value: 1.887e-15
```

Logistic regression for the adult dataset showed very similar results:

Figure 11: Logistic regression - Adults

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)    
(Intercept)                 -1.5314     1.2590  -1.216 0.223851    
age                         -0.0969     0.3658  -0.265 0.791074    
genderm                     -0.2074     0.2168  -0.957 0.338789    
ethnicityBlack               1.8109     0.4713   3.842 0.000122 ***
ethnicityHispanic            1.4926     0.6554   2.277 0.022765 *  
ethnicityLatino              1.7850     0.5937   3.007 0.002641 ** 
`ethnicityMiddle Eastern `  -0.8436     0.5586  -1.510 0.131031    
ethnicitymissing            -0.3472     0.4866  -0.713 0.475575    
ethnicityOthers              1.0142     0.5221   1.943 0.052041 .  
ethnicityPasifika           -0.2506     1.1110  -0.226 0.821525    
`ethnicitySouth Asian`      -0.7925     0.7974  -0.994 0.320333    
ethnicityTurkish             0.4620     1.1658   0.396 0.691905    
`ethnicityWhite-European`    1.5445     0.3525   4.382 1.18e-05 ***
jundiceyes                   0.3577     0.3359   1.065 0.286950    
austimyes                    0.5744     0.3022   1.901 0.057337 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 638.67  on 563  degrees of freedom
Residual deviance: 539.51  on 549  degrees of freedom
AIC: 569.51
```

Naturally, both regression models have a very low R-squared (around 0.06) since the explored features alone cannot explain the occurrence of ASD. After exploring Ridge, Lasso, and then Elastic regression and experimenting with the regularization coefficients, there was no notable improvement or change, so it was decided to move on to the next method.

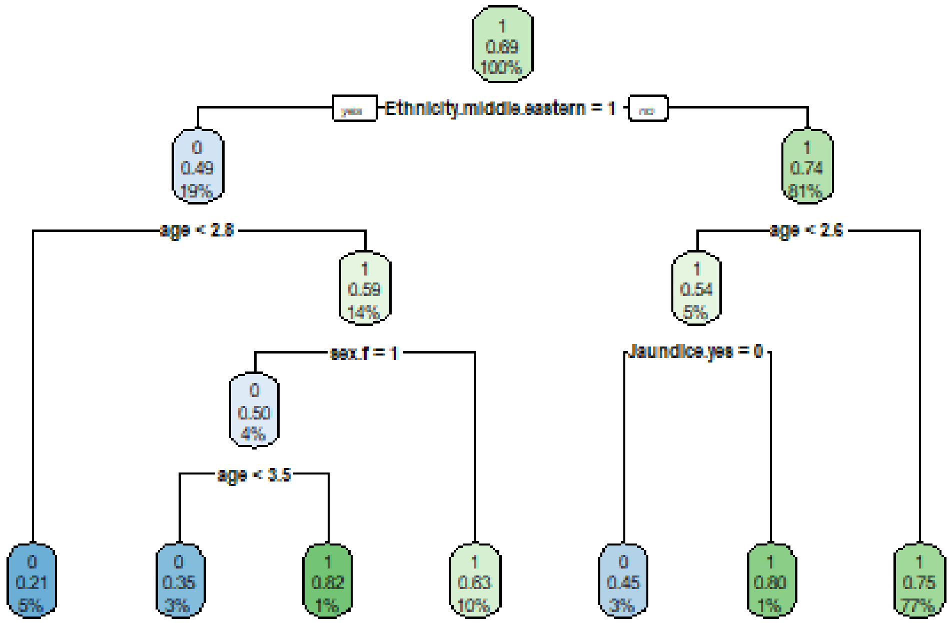
3.2 k-NN

K-NN was performed (both including the question variables and then excluding them) to attempt to define groups of similar patients. However, adding question responses brought about overfitting as the algorithm understood the pattern behind the result and thus any choice of k resulted in over 90% accuracy. Without the questions, however, accuracy was way lower, to the point where it was deemed unreliable for coming to any conclusions.

3.3 Decision trees

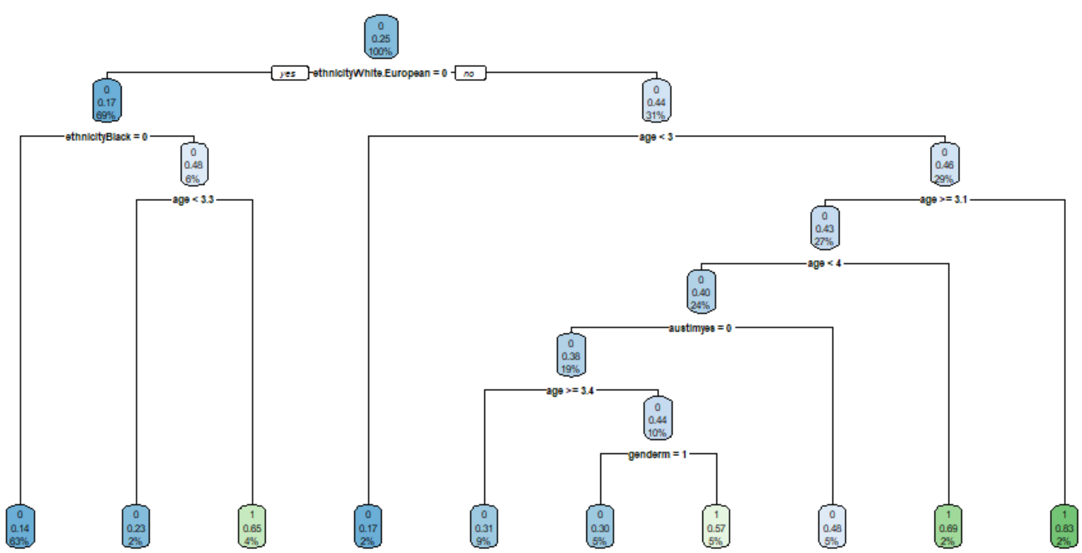
In the toddlers dataset, after excluding questions, the main decision tree features used for splitting were: ethnicity (Middle eastern), age, jaundice, and sex. Thus, relying on similar characteristics as the regression models, but also with ‘age’.

Figure 12: Decision tree - Toddlers



In adults, the results were similar apart from the ethnicity – instead of splitting by Middle Eastern ethnicity, in the adults dataset White ethnicity has more significance, as well as age.

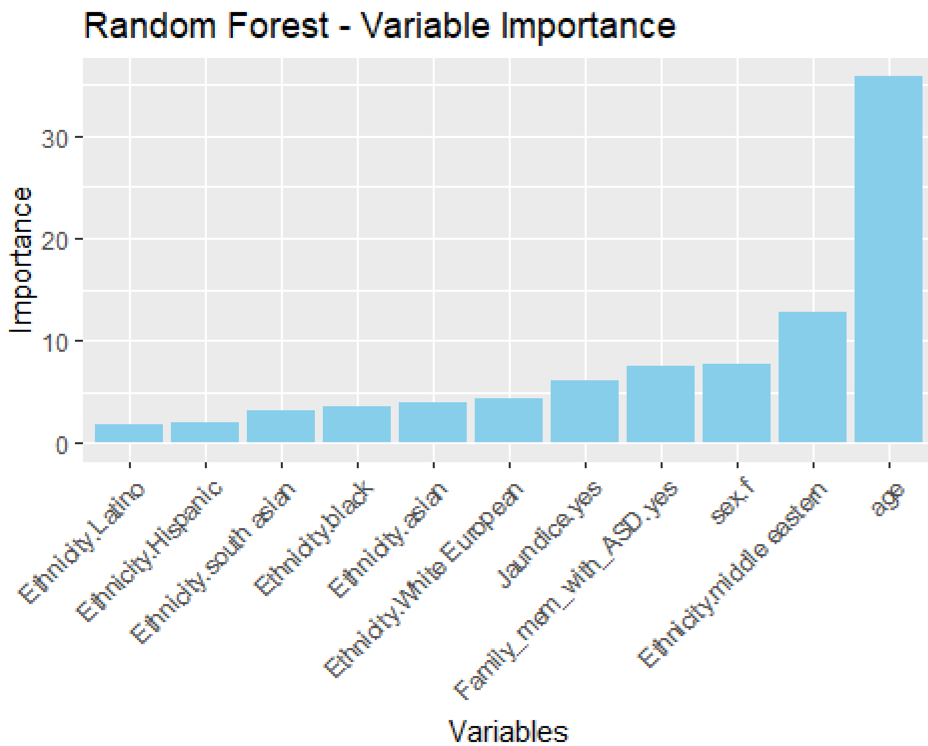
Figure 13: Decision tree - Adults



3.4 Random forest

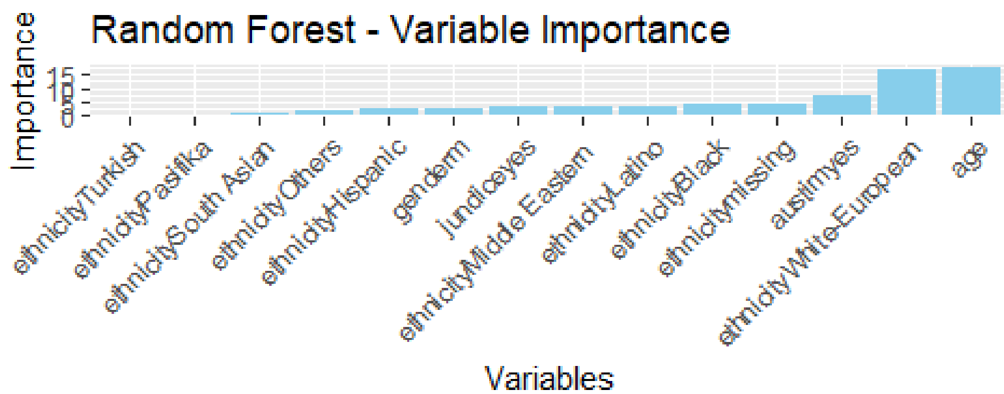
The random forest results were similar to the ones of the decision tree method, with age playing the most significant role (the older a toddler is, the more likely they are to be diagnosed since, allegedly, there are more symptoms showing). Age is followed in importance by association with the Middle Eastern ethnicity, as well as being a female and having jaundice or/and family members with ASD.

Figure 14: Random Forest - Toddlers



Similarly, age, White ethnicity, and ASD running in the family showed most relevance in adults. However, low rates of accuracy has been shown in both models (~0.30), so the splitting decision cannot be fully trusted.

Figure 15: Random Forest - Adults



4. Unsupervised learning techniques

Given the peculiar data structure, the unsupervised learning options were quite limited as it was challenging to find methods that would perform well with mostly categorical data. Finally, factor analysis was chosen as a mean of identifying commonalities among the significant variables as it provides a different outlook from the correlation matrix or regression coefficients. Also, Apriori algorithm technique was used to see if there are any associations rules between variables and the target.

4.1 Factor analysis

To select the relevant variables, the Kaiser–Meyer–Olkin (KMO) test was performed, which is a statistical measure to identify data which is suitable for factor analysis. Later on, these factors are attempted to be grouped up into the chosen number of factors.

In both toddler and adult datasets without question responses, the KMO test determined that only 4 variables had a score of 0.50 and above and thus were significantly influential – age, sex (F), jaundice, and family members with ASD. The rest of the variables had a very low MSA score and thus were discarded, as can be seen below.

Figure 16: KMO test results - Toddlers

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = X_train)
Overall MSA =  0.08
MSA for each item =
                  age                 sex.f         Ethnicity.asian
                 0.56                  0.58                    0.12
      Ethnicity.black    Ethnicity.Hispanic        Ethnicity.Latino
                 0.03                  0.03                    0.02
Ethnicity.middle eastern  Ethnicity.south asian Ethnicity.White European
                 0.09                  0.04                    0.11
         Jaundice.yes  Family_mem_with_ASD.yes
                 0.81                  0.81
```

Unfortunately, the 4 remaining variables could not be grouped up into factors, so the best conclusion that can be drawn is that in toddlers, jaundice and family cases of ASD are the most significant reasons for having ASD.

Surprisingly, the adults dataset showed similar results, only that the coefficients were slightly lower (see Figure 17). Nonetheless, just like in the case of toddlers, there was not enough data to construct factors. However, it was attempted to re-include the question variables and build factors with them. However, the results were not too fascinating as it only showed which questions were used to determine ASD presence (response – 1), and this information could have been retrieved from the rules of the survey.

Figure 17: KMO test results - Adults

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = X_train3)
Overall MSA =  0.2
MSA for each item =
                  age                 genderm          ethnicityBlack
                 0.72                    0.57                    0.10
    ethnicityHispanic          ethnicityLatino ethnicityMiddle Eastern
                 0.12                    0.11                    0.15
     ethnicitymissing          ethnicityOthers        ethnicityPasifika
                 0.15                    0.10                    0.08
  ethnicitySouth Asian        ethnicityTurkish  ethnicityWhite-European
                 0.12                    0.10                    0.24
            jundiceyes               austimyes
                 0.64                    0.61
```

4.2 Apriori algorithm

After the conversion of the data frame into a specific format, the Apriori algorithm detects patterns and dependencies between all the variables and the given target. In toddlers dataset, the five most-present links identified were between family member with ADS, Asian or White-European backgrounds, jaundice, and being a female. These factors were claimed to be associated with ASD traits of 1.

Figure 18: Apriori algorithm results - Toddlers

```
     lhs                             rhs                         support confidence
coverage      lift count
[1]  {Family_mem_with_ASD.yes}    => {Class.ASD.Traits.Yes} 0.1091082  0.6764706
0.1612903 0.9793956   115
[2]  {Ethnicity.asian}            => {Class.ASD.Traits.Yes} 0.2011385  0.7090301
0.2836812 1.0265353   212
[3]  {Jaundice.yes}               => {Class.ASD.Traits.Yes} 0.2039848  0.7465278
0.2732448 1.0808246   215
[4]  {Ethnicity.White European}   => {Class.ASD.Traits.Yes} 0.2371917  0.7485030
0.3168880 1.0836843   250
[5]  {sex.f}                      => {Class.ASD.Traits.Yes} 0.1840607  0.6081505
0.3026565 0.8804816   194
```

Interestingly, in the adult dataset the connections identified were multi-linked (see Figure 19), this could be attributed to the smaller sample size. As can be seen in the coverage column, most of these connections apply to 3% of the dataset, roughly 20 observations.

Figure 19: Apriori algorithm results - Adults

```
     lhs                              rhs     support confidence   coverage
lift count
[1]  {ethnicityLatino}            => {y3} 0.01420455     0.5000 0.02840909
1.862434    10
[2]  {age,
      ethnicityLatino}            => {y3} 0.01420455     0.5000 0.02840909
1.862434    10
[3]  {jundiceyes,
      austimyes}                  => {y3} 0.01562500     0.5500 0.02840909
2.048677    11
[4]  {ethnicityWhite-European,
      jundiceyes}                 => {y3} 0.02556818     0.5625 0.04545455
2.095238    18
[5]  {ethnicityWhite-European,
      austimyes}                  => {y3} 0.03835227     0.5400 0.07102273
2.011429    27
[6]  {age,
      jundiceyes,
      austimyes}                  => {y3} 0.01562500     0.5500 0.02840909
2.048677    11
[7]  {age,
      ethnicityWhite-European,
      jundiceyes}                 => {y3} 0.02556818     0.5625 0.04545455
```

5. Key findings & Conclusion

It has been interesting to experiment with different models, however, all of them have to be assessed critically due to the dataset limitations. Here are the key take-aways:

- The initial hypothesis regarding the prevailing importance of family history of ASD and jaundice has been partially confirmed by most methods used – various regression types, decision tree, and the factor analysis. However, these factors were more present in toddlers than adults, which can be attributed to adult's lack of awareness or neglect, as in the case of toddlers, parents or doctors were responsible for filling out the survey.
- There was a moderate correlation between ethnicities and the likelihood of developing ASD traits, while Middle Eastern ethnicity was associated with negative likelihood, coming from other ethnic backgrounds yielded a greater chance of testing ASD positive. However, such conclusions have to be drawn cautiously as both populations did not have balanced distribution of ethnicities and hence it could be the effect of sampling.

- Overall, the differences between results in toddlers and adults were not so drastic, which suggests a rather similar development of ASD traits across different ages.
- Likely due to the small sample size and oversampling (more ASD positive respondents among toddlers and more ASD negative respondents in adults), the regression models for adult have shown more sensitivity to ethnicities, which was not supported by the unsupervised methods, the KMO testing and the Apriori algorithm. Unfortunately, upsampling techniques led to discrepancies in regression coefficients so this issue can be only resolved with gathering additional data.
- Unfortunately, the linear dependency between question variables (A1 – A10) and the target variable largely limited the classification potential of the algorithms, as high accuracy would be deemed redundant and supposedly biased.

Overall, I could see continuing this research in the future as it would be interesting to carry out similar analysis but comparing the accuracy of diagnosing the patients using the Q10 surveys with the other diagnostic methods. I hope that more steps will be taken in the direction of studying autism in both children and adults and we will have more quality data to perform not only feature analysis but also more complex classification.

References

Fadi, Thabtah. (2017). Autism Screening Adult. UCI Machine Learning Repository. https://doi.org/10.24432/C5F019.

Kaggle User (2018). Autism Screening data for Toddlers. Kaggle Repository. https://www.kaggle.com/datasets/fabdelja/autism-screening-for-toddlers.

Appendix

As the markdown files were 20 and 42 pages long, I attach them as links to documents:

adult%20markdown.docx toddler%20markdown.docx