Individual Project Pt.2

COMP 3710

April 3rd 2020

Camila Castillo

T00564151

Camila Castillo
T00564151
Machine Learning
Individual Project

Introduction

Autism is a range of conditions characterized by challenges with social skills, repetitive behaviours, speech and nonverbal communication, as well as by unique strengths and differences. It tends to appear in the first 3 years of age of an infant. There are many signs that babies and toddlers show that may indicate autism, what they all have in common though is communication related. Some examples of common behaviours are:

- By 6 months, limited or no eye contact, no social smiles or other warm, joyful expressions directed at people
- By 9 months, no sharing of vocal sounds, smiles or other nonverbal communication
- By 12 months, no babbling, no use of gestures to communicate (e.g. pointing, reaching, waving etc.), no response to name when called
- By 16 months, no words
- By 24 months, no meaningful, two-word phrases
- Any loss of any previously acquired speech, babbling or social skills

For children showing this or similar symptoms, a screening is recommended to be done by a parent or a health care provider. If the screening shows "a high chance" of autism a formal evaluation is ordered to have an official diagnosis; this takes time, money and resources some families do not even have access to.

Therefore, a time-efficient and accessible ASD screening is imminent to help health professionals and inform individuals whether they should pursue formal clinical diagnosis.

Camila Castillo
T00564151
Machine Learning
Individual Project

The purpose of this study

The purpose of this project is to determine through the basic screening (see appendix A) and some general questions if a toddler has autistic traits and if a formal clinical diagnosis should be pursued. This is important because the sooner the toddler starts receiving therapy and using the correct methods for their education the better and more improvements will be noticed. The program should ultimately be able to predict this and show the correlation between factors and the likelihood of autistic traits.

Research questions

Some of the research questions I explore in this project are:

- What gender is more likely to show/have autistic traits ?
- Does Jaundice have a relation to toddlers showing autistic traits ?
- Can the program successfully learn and predict if a child should pursue a formal clinical diagnosis?

About the data

Before deciding to study computer science I had an enormous interest in psychology and education. I volunteered for five years at a non-profit foundation that help kids with down syndrome and autism. I learned about their behaviours, how they learn the best and assisted the teachers, therapist and psychologist perform classes and therapies. So when thinking about this project I thought it would be interesting in combining two of my passions. This became possible when I came across the “Toddler_Autism_dataset.csv” thanks to Kaggle.com.

The data at first seems difficult to understand so the dataset came with ‘readme’ document that explains the dataset with more detail (see appendix B). Even with my previous experience and this thorough explanation I still had to do a lot of research on autism, the screenings and how the symptoms/traits are portrayed by children to have a better understanding of the entire topic and what I was trying to do.

Summary of what I did

After understanding the data this is what I did on a nutshell. I formatted the data to make it quantifiable (replaced “yes” with 1 and “no” with 0). I then created a heatmap to identify which factors besides the screening questions had a higher correlation to the “diagnosis”. Based on those results I plotted some graphs to answer the first set of research questions. I then proceeded to train my data using supervised learning and then to see if it learned I removed the target answer.

Next, I performed some regression techniques including: Linear regression, Logistic regression, Lasso, Naïve Bayes and Stochastic Gradient descent. Finally, I studied the progress of the predictions comparing predictions to the actual results and calculated how accurate my program is.

Camila Castillo
T00564151
Machine Learning
Individual Project

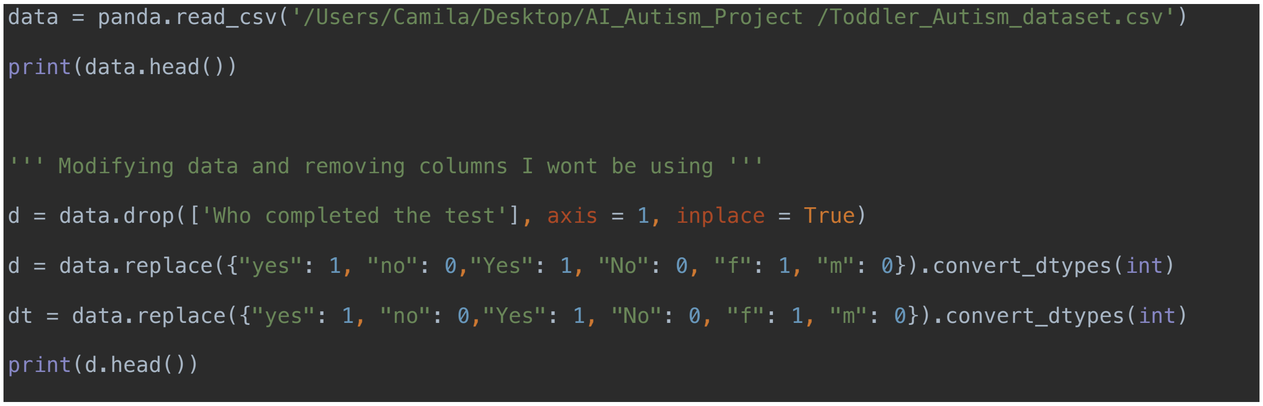
Background

There have been several works done similar to mine. In Kaggle itself users have use the same dataset to come up with correlation graphs and regression values but never together and not training the data. After some more research I found out that in 2019 there was a paper published where they talk about how they created a ML project to try and predict if a kid is autistic based on the same screening model I am using. The project is described as “automated machine learning (ML) method for overcoming barriers to ASD screening, specifically using the feedforward neural network (fNN).” (Achenie, Luke E K, et al).

From all of these projects developed I took inspiration to come up with my own. I tried to combine as many factors as possible and thought it was a great idea to test several regression models.

Project details

First of all, I used pandas to analyze the dataset. I noticed that the data was not binary so I modified it to make it quantifiable and remove a column I knew I was not going to use as it had no relevancy to my project.



```
data = panda.read_csv('/Users/Camila/Desktop/AI_Autism_Project /Toddler_Autism_dataset.csv')
print(data.head())

''' Modifying data and removing columns I wont be using '''
d = data.drop(['Who completed the test'], axis = 1, inplace = True)
d = data.replace({"yes": 1, "no": 0,"Yes": 1, "No": 0, "f": 1, "m": 0}).convert_dtypes(int)
dt = data.replace({"yes": 1, "no": 0,"Yes": 1, "No": 0, "f": 1, "m": 0}).convert_dtypes(int)
print(d.head())
```

Once my data was formatted and ready to use I plotted a heatmap to be able to visualize which factors besides the screening questions had a higher correlation to autistic traits. I did this using seaborn.

```
'''Heat Map Correlation'''

plt.figure(figsize=(14,14))

cor = d.corr().abs()

cor_target = abs(cor["Class/ASD Traits "])

sns.heatmap(data = cor, cmap=plt.cm.Blues, annot = True, square = True, cbar = True )
```

I also decided to list the values with their corresponding correlations to have a clearer idea of what I was going to plot.

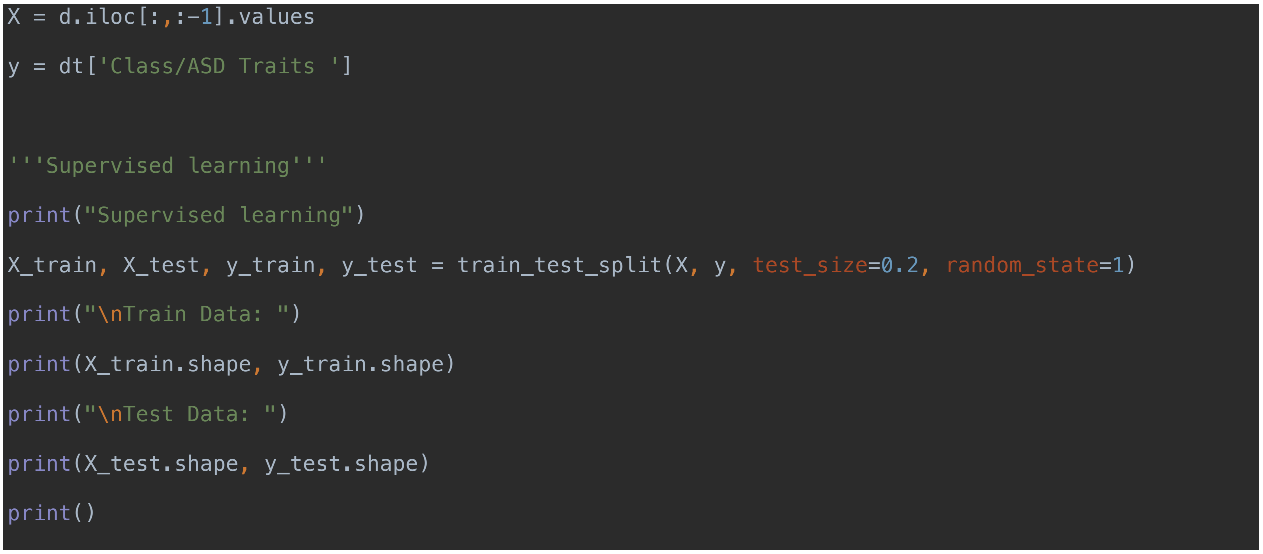
```
print("\nFrom highest to lowest correlation\n")

print(cor_target[cor_target>-1].sort_values(ascending=False))
```

based on this list I plotted the following graphs:

- The count of toddlers with autistic traits and their ethnicity
- Jaundice against the results of autistic traits to see the relationship.
- And Sex/Gender against the results of autistic traits to determine who is more likely to show/have autistic traits

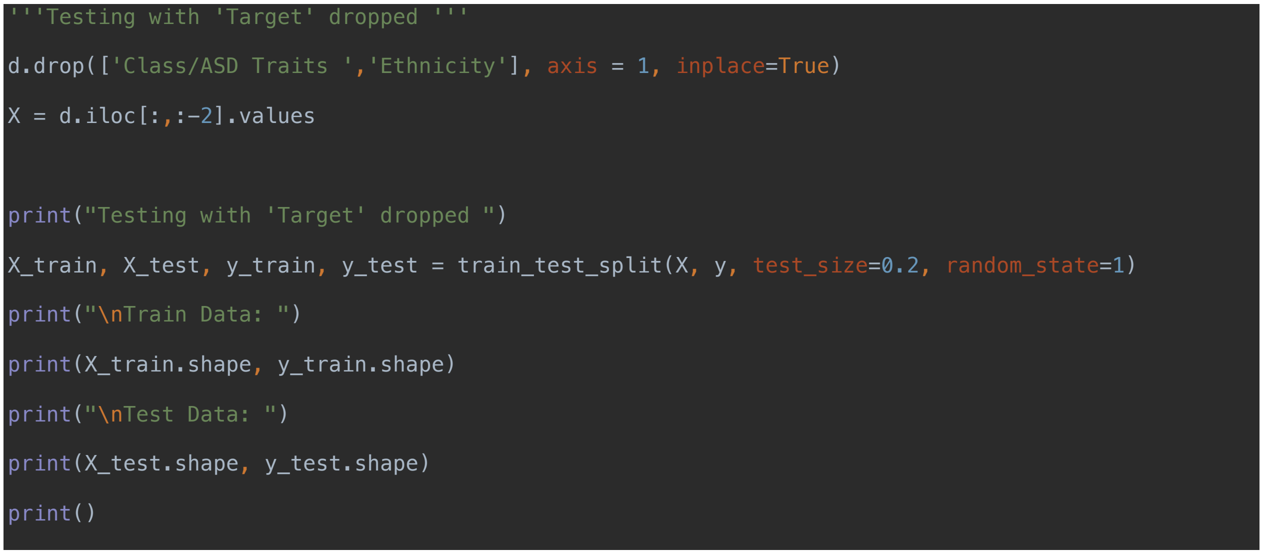
```
print("\nCorrelation between variables and if toddlers presented ASD traits")

plt.figure(figsize = (12,8))

sns.countplot(x='Ethnicity', data=data)

plt.figure(figsize = (12,8))

sns.countplot(x = 'Class/ASD Traits ', hue = 'Jaundice', data = data)

plt.figure(figsize = (12,8))

sns.countplot(x = 'Class/ASD Traits ', hue = 'Sex', data = data)

print(plt.show())
```

Next, I assigned my data to X and Y and performed the supervised learning training. I set my test size to 0.2 and trained and tested the data.



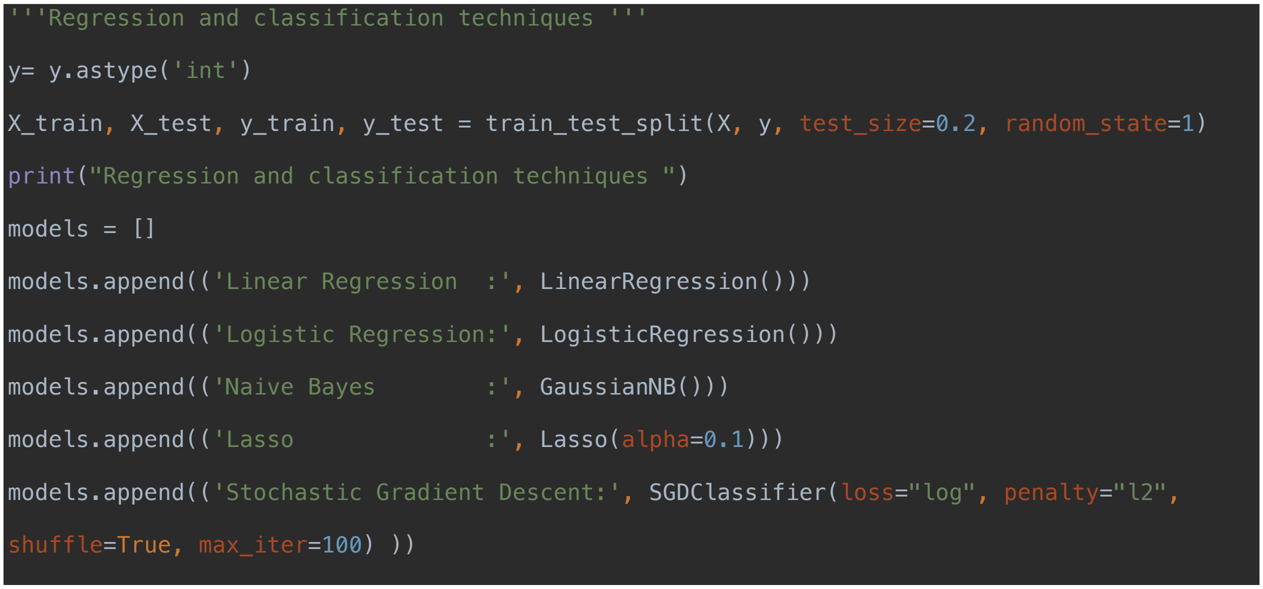
```
X = d.iloc[:,:-1].values
y = dt['Class/ASD Traits ']

'''Supervised learning'''
print("Supervised learning")
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)
print("\nTrain Data: ")
print(X_train.shape, y_train.shape)
print("\nTest Data: ")
print(X_test.shape, y_test.shape)
print()
```

I then did the unsupervised learning training. Meaning that I dropped the target 'Class/ASD Traits’. I was getting an error when I did this, so I decided to also remove 'Ethnicity' as I was not using it anymore and this also seemed to solve the error.



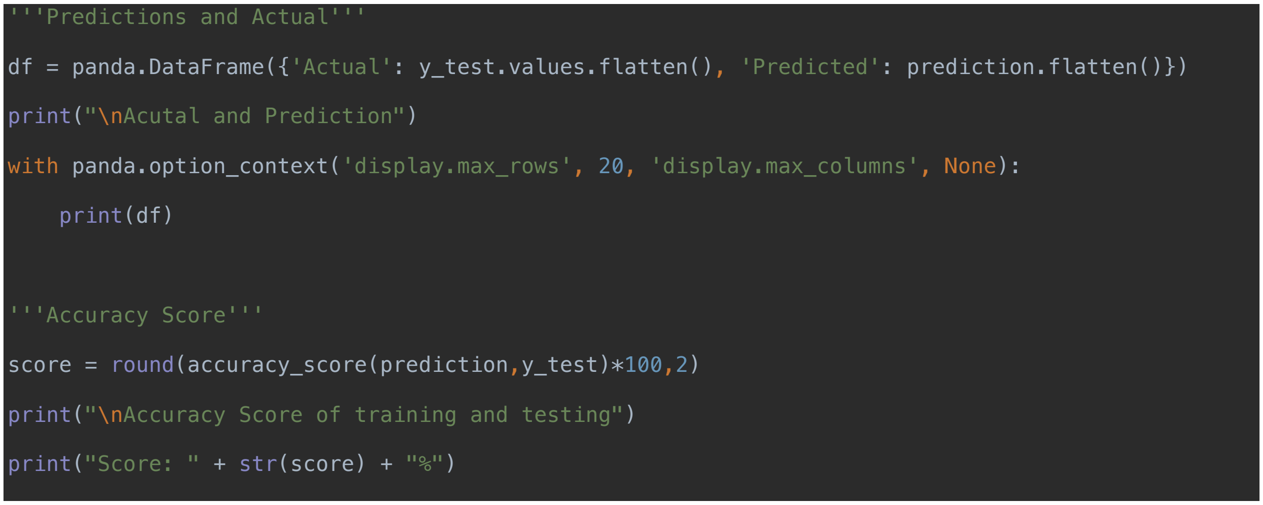
```
'''Testing with 'Target' dropped '''
d.drop(['Class/ASD Traits ','Ethnicity'], axis = 1, inplace=True)
X = d.iloc[:,:-2].values

print("Testing with 'Target' dropped ")
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)
print("\nTrain Data: ")
print(X_train.shape, y_train.shape)
print("\nTest Data: ")
print(X_test.shape, y_test.shape)
print()
```

I then applied regression and classification techniques. I was able to optimize my code by putting all the techniques into an array and go through them with a loop.



```
'''Regression and classification techniques '''
y= y.astype('int')
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)
print("Regression and classification techniques ")
models = []
models.append(('Linear Regression  :', LinearRegression()))
models.append(('Logistic Regression:', LogisticRegression()))
models.append(('Naive Bayes        :', GaussianNB()))
models.append(('Lasso              :', Lasso(alpha=0.1)))
models.append(('Stochastic Gradient Descent:', SGDClassifier(loss="log", penalty="l2",
shuffle=True, max_iter=100) ))
```

Finally, I ran the programs predictions and compared them with the actual results in the dataset to see how accurate my program is. Based on that, I calculated an accuracy score for my program overall.

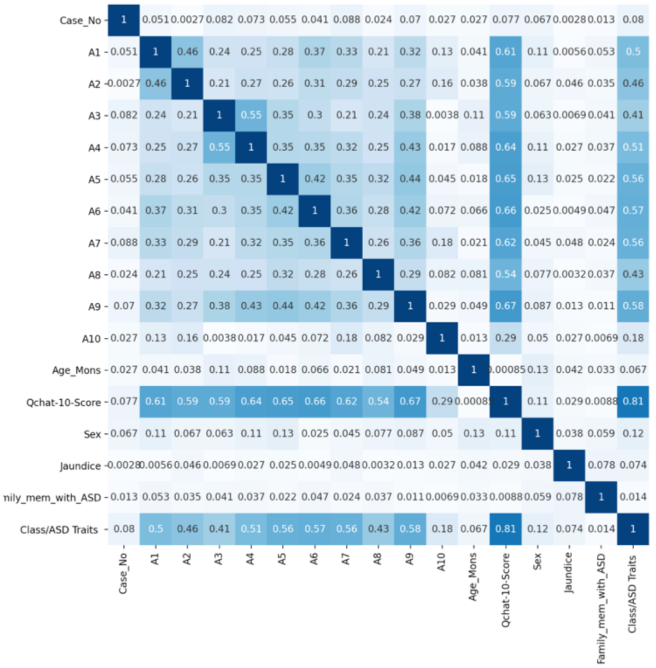


```
'''Predictions and Actual'''
df = panda.DataFrame({'Actual': y_test.values.flatten(), 'Predicted': prediction.flatten()})
print("\nAcutal and Prediction")
with panda.option_context('display.max_rows', 20, 'display.max_columns', None):
    print(df)

'''Accuracy Score'''
score = round(accuracy_score(prediction,y_test)*100,2)
print("\nAccuracy Score of training and testing")
print("Score: " + str(score) + "%")
```

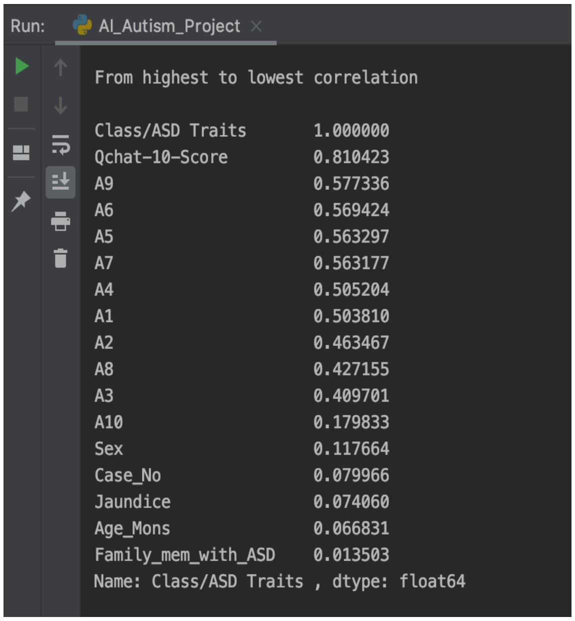
Camila Castillo
T00564151
Machine Learning
Individual Project

Results

As mentioned before, I first focused on the relation between certain factors like gender and Jaundice and if the toddlers showed autistic traits or not. The plotting of the graphs and the correlation results allow the visualization of these relationships. The following graphs also allow me to answer a couple of my original research questions.

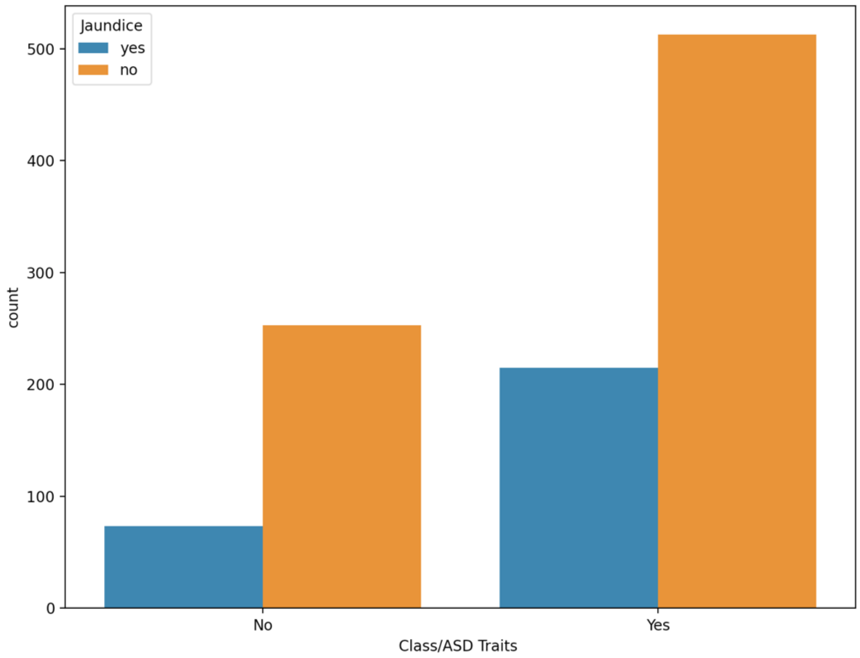


	Case_No	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Age_Mons	Qchat-10-Score	Sex	Jaundice	Family_mem_with_ASD	Class/ASD Traits
Case_No	1	0.051	0.0027	0.082	0.073	0.055	0.041	0.088	0.024	0.07	0.027	0.027	0.077	0.067	0.0028	0.013	0.08
A1	0.051	1	0.46	0.24	0.25	0.28	0.37	0.33	0.21	0.32	0.13	0.041	0.61	0.11	0.0056	0.053	0.5
A2	0.0027	0.46	1	0.21	0.27	0.26	0.31	0.29	0.25	0.27	0.16	0.038	0.59	0.067	0.046	0.035	0.46
A3	0.082	0.24	0.21	1	0.55	0.35	0.3	0.21	0.24	0.38	0.0038	0.11	0.59	0.063	0.0069	0.041	0.41
A4	0.073	0.25	0.27	0.55	1	0.35	0.35	0.32	0.25	0.43	0.017	0.088	0.64	0.11	0.027	0.037	0.51
A5	0.055	0.28	0.26	0.35	0.35	1	0.42	0.35	0.32	0.44	0.045	0.018	0.65	0.13	0.025	0.022	0.56
A6	0.041	0.37	0.31	0.3	0.35	0.42	1	0.36	0.28	0.42	0.072	0.066	0.66	0.025	0.0049	0.047	0.57
A7	0.088	0.33	0.29	0.21	0.32	0.35	0.36	1	0.26	0.36	0.18	0.021	0.62	0.045	0.048	0.024	0.56
A8	0.024	0.21	0.25	0.24	0.25	0.32	0.28	0.26	1	0.29	0.082	0.081	0.54	0.077	0.0032	0.037	0.43
A9	0.07	0.32	0.27	0.38	0.43	0.44	0.42	0.36	0.29	1	0.029	0.049	0.67	0.087	0.013	0.011	0.58
A10	0.027	0.13	0.16	0.0038	0.017	0.045	0.072	0.18	0.082	0.029	1	0.013	0.29	0.05	0.027	0.0069	0.18
Age_Mons	0.027	0.041	0.038	0.11	0.088	0.018	0.066	0.021	0.081	0.049	0.013	1	0.00085	0.13	0.042	0.033	0.067
Qchat-10-Score	0.077	0.61	0.59	0.59	0.64	0.65	0.66	0.62	0.54	0.67	0.29	0.00085	1	0.11	0.029	0.0088	0.81
Sex	0.067	0.11	0.067	0.063	0.11	0.13	0.025	0.045	0.077	0.087	0.05	0.13	0.11	1	0.038	0.059	0.12
Jaundice	0.0028	0.0056	0.046	0.0069	0.027	0.025	0.0049	0.048	0.0032	0.013	0.027	0.042	0.029	0.038	1	0.078	0.074
nily_mem_with_ASD	0.013	0.053	0.035	0.041	0.037	0.022	0.047	0.024	0.037	0.011	0.0069	0.033	0.0088	0.059	0.078	1	0.014
Class/ASD Traits	0.08	0.5	0.46	0.41	0.51	0.56	0.57	0.56	0.43	0.58	0.18	0.067	0.81	0.12	0.074	0.014	1



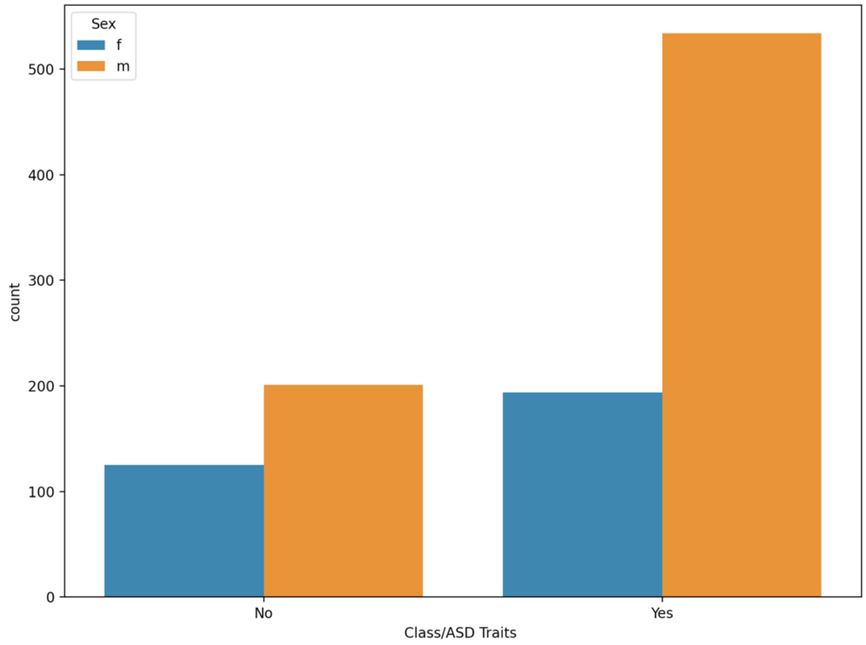
```
From highest to lowest correlation

Class/ASD Traits       1.000000
Qchat-10-Score         0.810423
A9                     0.577336
A6                     0.569424
A5                     0.563297
A7                     0.563177
A4                     0.505204
A1                     0.503810
A2                     0.463467
A8                     0.427155
A3                     0.409701
A10                    0.179833
Sex                    0.117664
Case_No                0.079966
Jaundice               0.074060
Age_Mons               0.066831
Family_mem_with_ASD    0.013503
Name: Class/ASD Traits , dtype: float64
```

From the heat map and the correlation list we can see that the variables with higher correlation besides the results of the screening are Sex and Jaundice. This is why I decided to plot and analyze the following graphs:



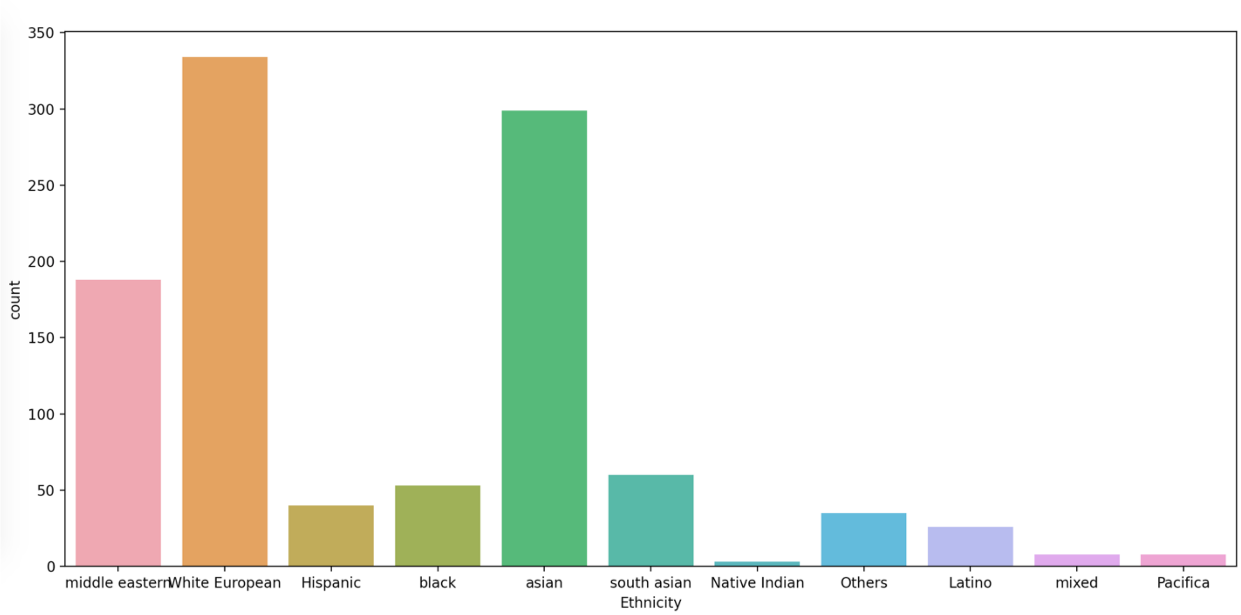
The graph to the left show the relationship between having autistic traits and if a child had jaundice at birth or not. Jaundice is when a baby has too much bilirubin this makes their skin and eyes yellowish colour. It tends to disappear in a week or two after birth (Caring for kids).

From the graph we can see that Jaundice has no effect on whether a kid has autistic traits or not. In any case kids that did not have Jaundice at birth where more likely to develop autism.

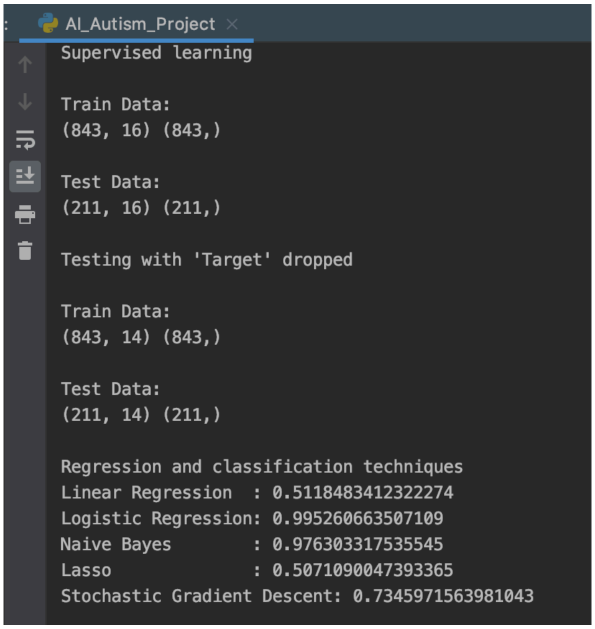


This next graph shows that boys are more likely to develop autism rather than girls.

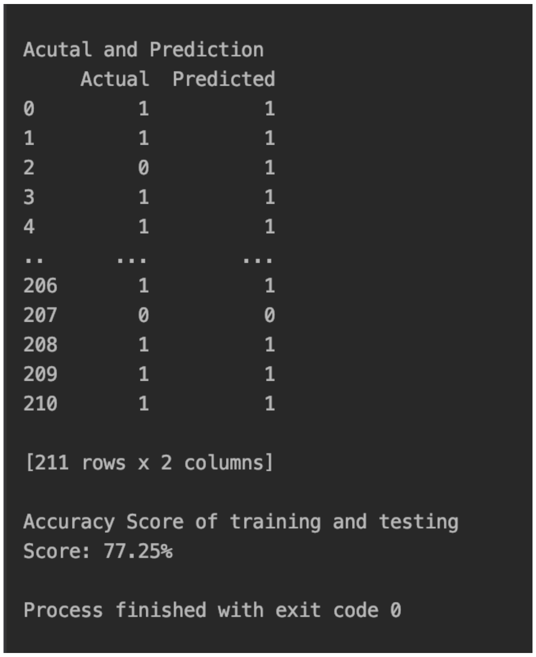
I also decided to plot a bar graph based on the ethnicity of the toddlers; mainly just to show that the dataset is diverse and that essentially ethnicity has no impact on whether a child is autistic or not



The results for regression and classification techniques where as followed:



```
Supervised learning

Train Data:
(843, 16) (843,)

Test Data:
(211, 16) (211,)

Testing with 'Target' dropped

Train Data:
(843, 14) (843,)

Test Data:
(211, 14) (211,)

Regression and classification techniques
Linear Regression  : 0.5118483412322274
Logistic Regression: 0.995260663507109
Naive Bayes        : 0.976303317535545
Lasso              : 0.5071090047393365
Stochastic Gradient Descent: 0.7345971563981043
```



```
Acutal and Prediction
     Actual  Predicted
0         1          1
1         1          1
2         0          1
3         1          1
4         1          1
..      ...        ...
206       1          1
207       0          0
208       1          1
209       1          1
210       1          1

[211 rows x 2 columns]

Accuracy Score of training and testing
Score: 77.25%

Process finished with exit code 0
```

We can see that the most accurate was Naïve Bayes. These values ended up being what I expected specially because logistic regression is used when the dependent variable is binary in nature which is the case of the dataset. Finally, we can see that the data learned pretty good.

At the end the program ended up having a 77.25% accuracy which I believe is very good. As we can also see from the result of the predictions there was only one mistake in the information shown.

Discussion and Conclusion

Working with this dataset has taught me that this program has not reached it's potential. I got some accurate results considering the time I had and the visualizations are useful to see the trends in the data.I believe better results could be achieved by focusing in less or more specific questions. The improvement and further development of this project could result in negating the need for labor-intensive follow-up and circumvents human error, providing an advantage over previous screening methods.

Future plan

My future plans consist on improving the prediction models. Take another dataset and combine them so that the program has more to learn from. I believe this will make predictions more accurate. I would also like to try other libraries like Deeplearning4j to work with this dataset. I believe this topic and dataset have so much potential that only experimenting and "playing around with it" would lead us to incredible finding; that who knows may help screenings and diagnosis in the future.

Camila Castillo
T00564151
Machine Learning
Individual Project

Appendix A

Table 1: Details of variables mapping to the Q-Chat-10 screening methods

Variable in Dataset	**Corresponding Q-chat-10-Toddler Features**
A1	Does your child look at you when you call his/her name?
A2	How easy is it for you to get eye contact with your child?
A3	Does your child point to indicate that s/he wants something? (e.g. a toy that is out of reach)
A4	Does your child point to share interest with you? (e.g. poin9ng at an interes9ng sight)
A5	Does your child pretend? (e.g. care for dolls, talk on a toy phone)
A6	Does your child follow where you're looking?
A7	If you or someone else in the family is visibly upset, does your child show signs of wan9ng to comfort them? (e.g. stroking hair, hugging them)
A8	Would you describe your child's first words as:
A9	Does your child use simple gestures? (e.g. wave goodbye)
A10	Does your child stare at nothing with no apparent purpose?

Appendix B

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. If the user obtained More than 3 Add points together for all ten questions. If your child scores more than 3 (Q-chat-10- score) then there is a potential ASD traits otherwise no ASD traits are observed.

The remaining features in the datasets are collected from the "submit" screen in the ASDTests screening app. It should be noted that the class varaible was assigned automatically based on the score obtained by the user while undergoing the screening process using the ASDTests app.

References

Achenie, Luke E K, et al. “A Machine Learning Strategy for Autism Screening in Toddlers.” *Journal of Developmental and Behavioral Pediatrics : JDBP*, U.S. National Library of Medicine, June 2019, www.ncbi.nlm.nih.gov/pubmed/30985384.

Caring for Kids. “Jaundice in Newborns.” *Jaundice in Newborns - Caring for Kids*, Oct. 2017, www.caringforkids.cps.ca/handouts/jaundice_in_newborns.