**Term End Project Milestone-5**

Gyan Kannur

DSC, Bellevue University

DSC550 Data Mining

 Professor Brett Werner

04/05/2025

ABSTRACT:

This term-end project aims to evaluate the use of Data Mining techniques. The objective of this projetct is to ilustrate use of learned techniques to mine and analyze large datasets to discover useful knowledge. Text mining, unstructured data, social networks, and other types of unsupervised data mining methods for data science are included

**Term End project Milestone -5**

* **INTRODUCTION**

Analyze data to predict the traits to detect Autistics disease amoung the toddlers

Abstract:

Problem :

Autistic Spectrum Disorder (ASD) is a neurodevelopmental condition associated with significant healthcare costs, and early diagnosis can significantly reduce these. Unfortunately, waiting times for an ASD diagnosis are lengthy and procedures are not cost effective. The economic impact of autism and the increase in the number of ASD cases across the world reveals an urgent need for the development of easily implemented and effective screening methods.

Solution:

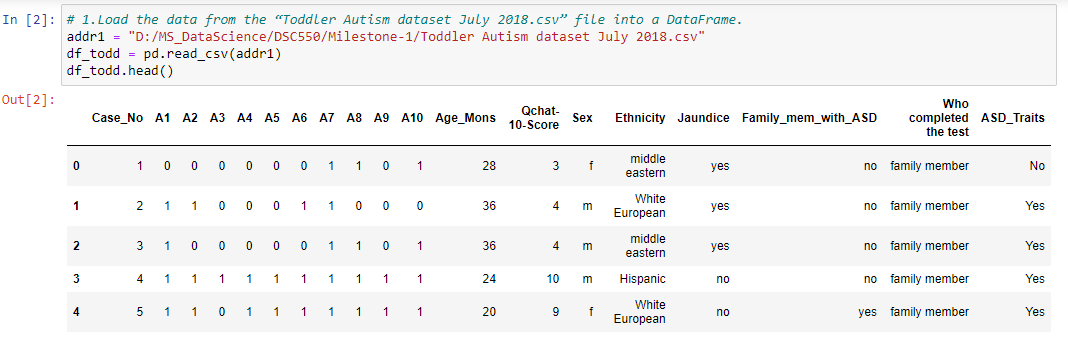
Therefore, a time-efficient and accessible ASD screening is imminent to help health professionals and inform individuals whether they should pursue formal clinical diagnosis. The rapid growth in the number of ASD cases worldwide necessitates datasets related to behaviour traits

Data Source : <https://www.kaggle.com/fabdelja/autism-screening-for-toddlers?select=Toddler+Autism+dataset+July+2018.csv>

Description about Dataset:

The dataset was developed by Dr Fadi Fayez Thabtah (fadifayez.com) using a mobile app called ASDTests (ASDtests.com) to screen autism in toddlers.we can use it to estimate the predictive power of machine learning techniques in detecting autistic traits.

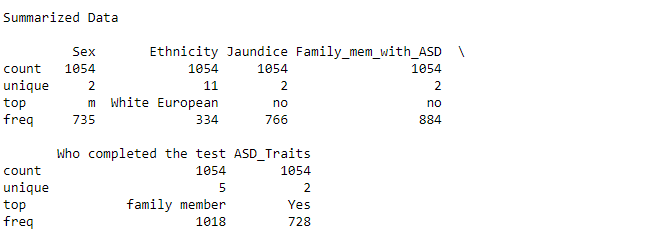
Screen print showing the used raw dataset:

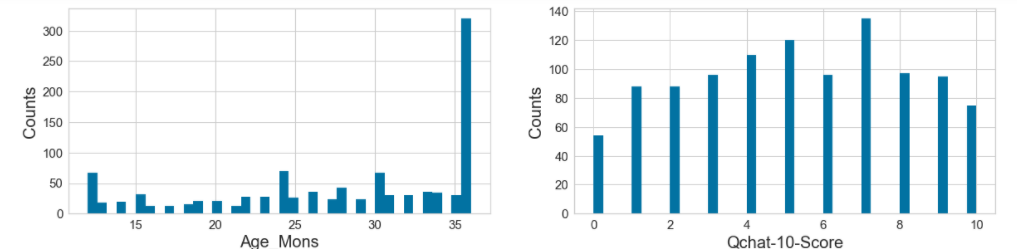


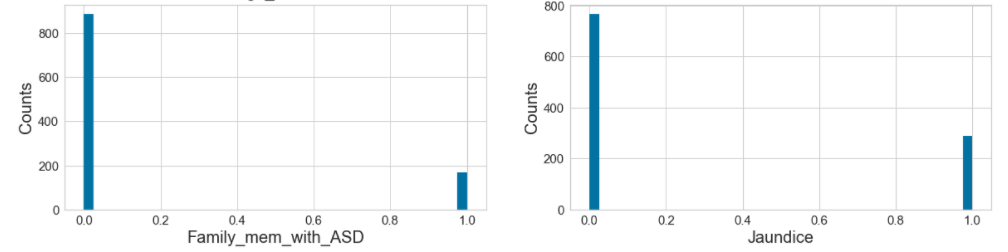
**Summary of the Process:**

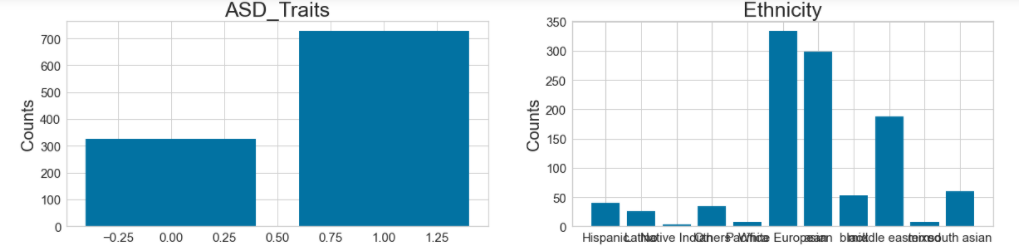
Please find below Steps involved in creating the ASD Screener Model.

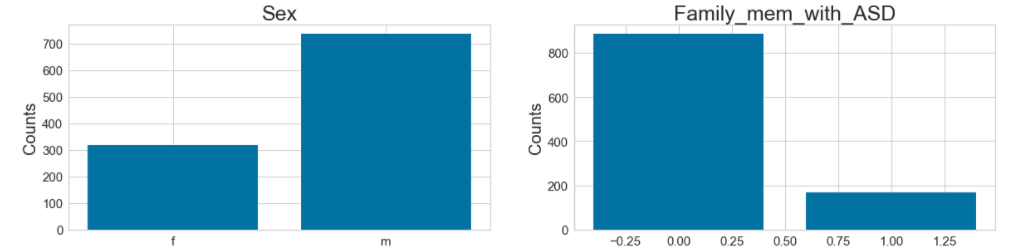
Begin Milestone 1 : Completed the graphical analysis of data by creating a minimum of four graphs.



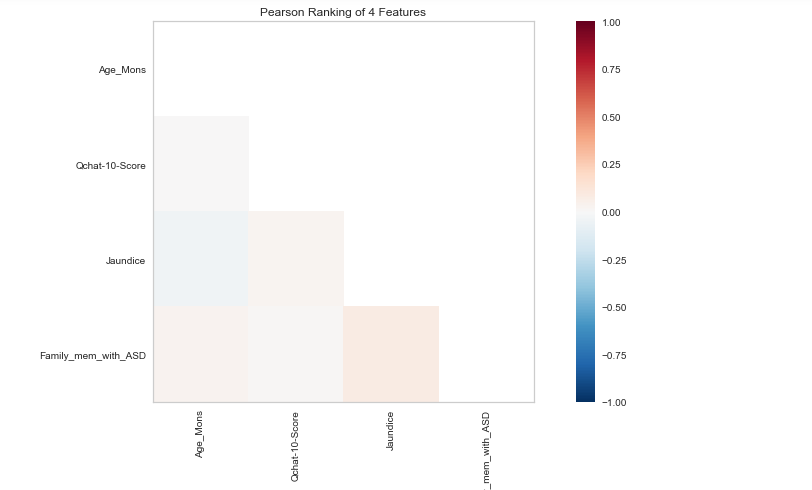








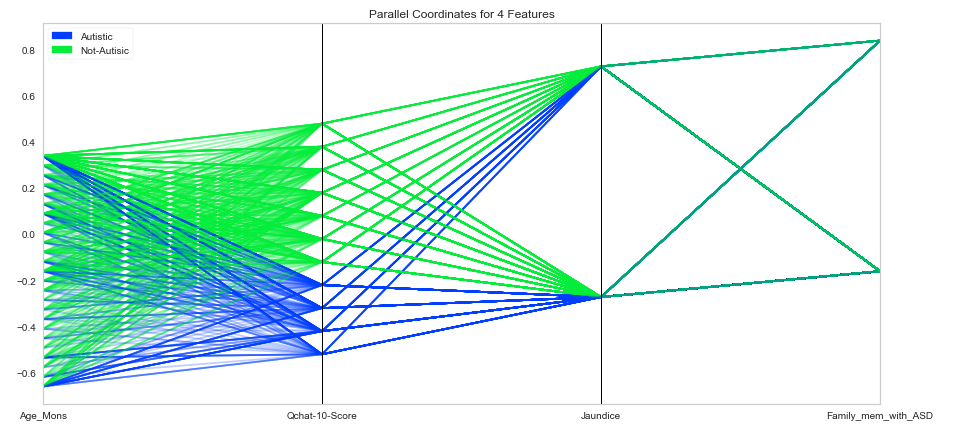
Determine the Pearson Ranking amoungs features:



The correlation between the variables is low (1 or -1 is high positive or high negative, 0 is low or no correlation)

Here These results show there is positive correlation between 'ASD\_Traits' & 'Qchat-10-Score', but it’s not a high correlation amoung other variables.

Finding out the Parallel correlation amoung the selected features:

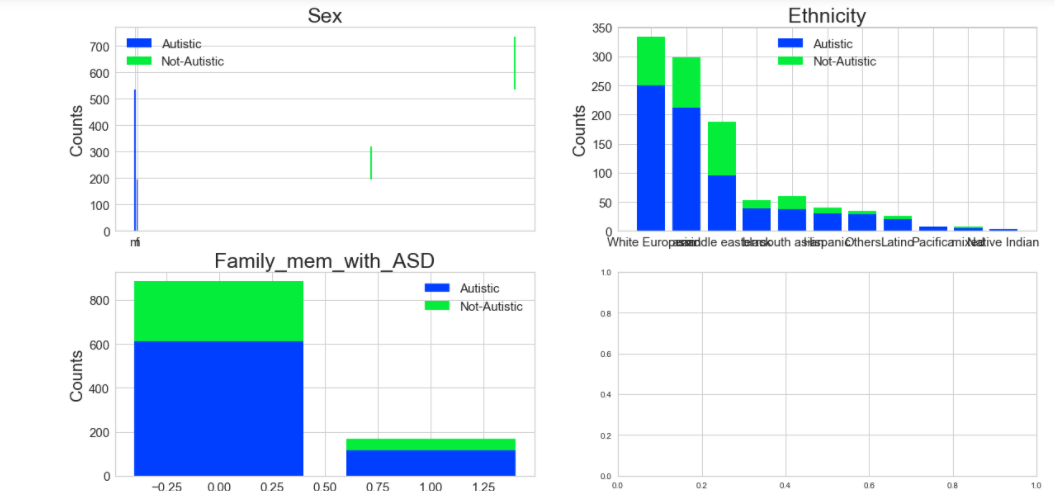


Parallel cordinate for 4 features shows below information:

For Austics patients : we have a seen relationship between age ,score and jaundice variables,how ever family men with ads feature is not having any relation with other features

For Non Austics: we have seens relationship beteen all listed 4 features

Used Stack Bar Charts to compare toddlers who is having ASD & who didn’t have ASD based on the other variables.



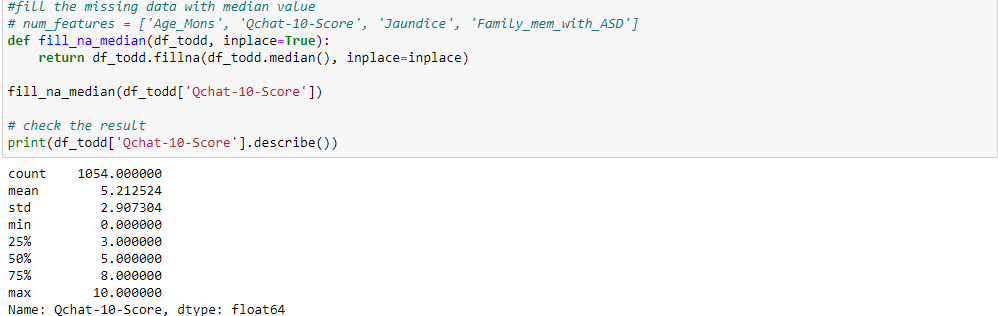
less females have ASD as compared to MEN, white europian is having more rate for impact sue to ASD fasmily menber history, as compared with Non ASD history.

Milestone 2

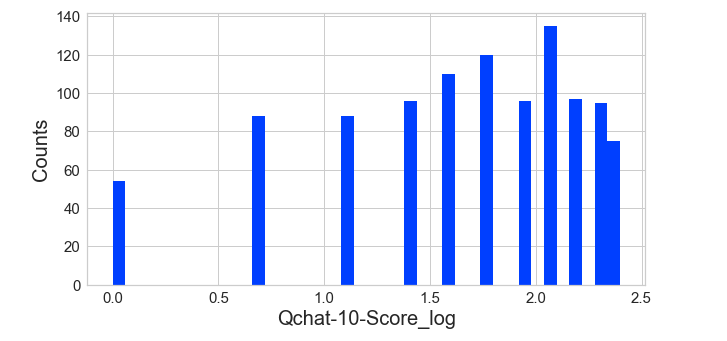
1) drop any features that are not useful for your model building. You should explain and justify why the feature dropped is not useful

2) address any missing data issues.

3) Build any new features that you need for your model, e.g., create dummy variables for categorical features if necessary. Explain your process at each step. You can use any methods/tools you think are most appropriate.

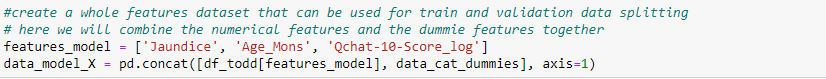


log transformation is showing high skewed values & its counts in historgram :

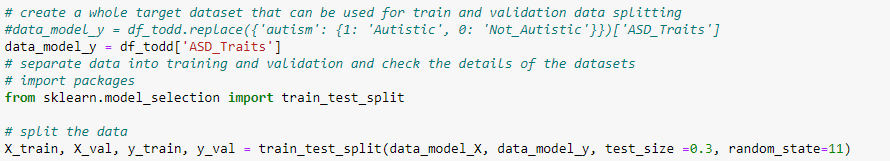


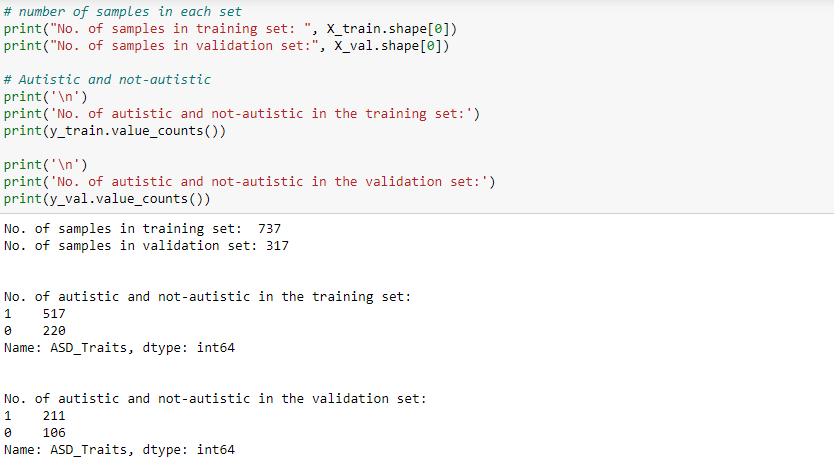
Milestone 3, Build and evaluate at least one model.

Selected the features for the model



splitting the datasets into training & test datasets

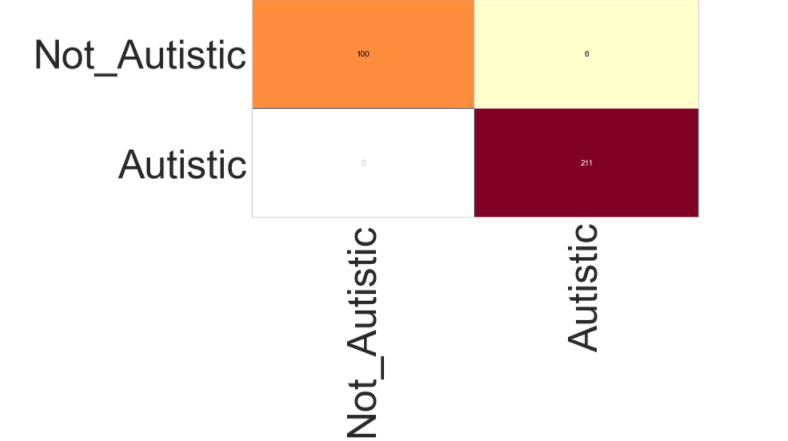




Here i have used Classification technique , since we are categorizing data into a given number of classes like 'Not\_Autistic','Autistic'.

The main goal of a classification problem is to identify the category/class to which a new data will fall under.

# Eval Metrics by using ConfusionMatrix



There are two possible predicted classes: "Autistic" and "Not\_Autistic". If we were predicting the presence of a disease, for example, "yes" would mean they have the disease, and "no" would mean they don't have the disease.

The classifier made a total of 317 predictions

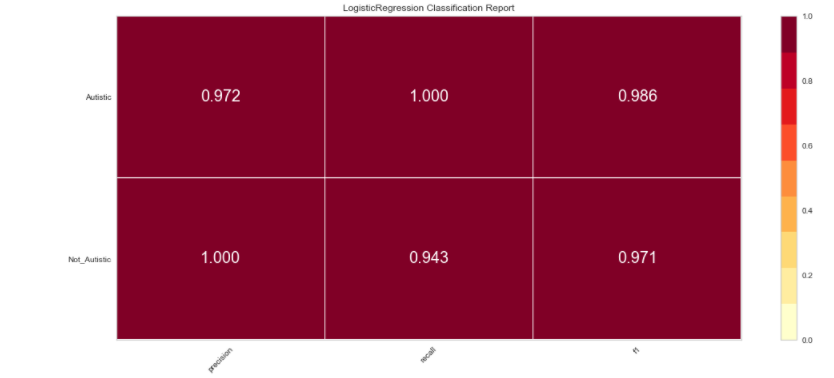
Out of those 317 cases, the classifier predicted "yes" 217 times, and "no" 100 times.

In reality, 211 patients in the sample have the disease, and 106 patients do not.

Accuracy: Overall, how often is the classifier correct?

(TP+TN)/total = (211+100)/317 = 0.98

Classification Report:



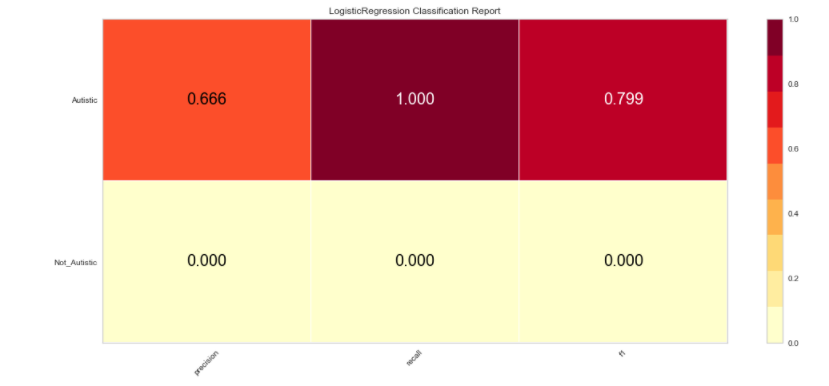
Found below reports from logistic regression classification report:

Precision – What percent of your predictions were correct? 97.2%

Recall – What percent of the positive cases did you catch? 100%

F1 score – What percent of positive predictions were correct? 98.6%

before concluding the final result, I have computed the classification reportes by using others features too (['Jaundice', 'Age\_Mons'])



Precision – What percent of your predictions were correct? 66.2%

Recall – What percent of the positive cases did you catch? 100%

F1 score – What percent of positive predictions were correct? .79%

data is biased it is not all combination autistic and not autistic datas

seems 'Jaundice', 'Age\_Mons' are not good candidates as individuals predictors

**Conclusions/Recommendations:**

**The ASD Pre-Screening traits data is very useful, by using above created model , the health professionals can predict accurately about the possibility of a toadler to be an ADS patient.**

**when we build the model by using ['Jaundice', 'Age\_Mons', 'Qchat-10-Score\_log'] features then the accuracy rate of trained model is arount 97% , looks like Qchat-10-Score\_log score feature data is one of the major contributer of good accuracy.**

Precision – What percent of your predictions were correct? 97.2%

Recall – What percent of the positive cases did you catch? 100%

F1 score – What percent of positive predictions were correct? 98.6%

**However when i have dropped the Qchat-10-Score\_log feature from the model and rebuild the model by using ['Jaundice', 'Age\_Mons'] features , then its accurecy scores got decreased,**

Precision – What percent of your predictions were correct? 66.2%

Recall – What percent of the positive cases did you catch? 100%

F1 score – What percent of positive predictions were correct? .79%

**Recommendations: Toddler ASD predection model is one of the useful tool, which is going to reduce the expenses related to unnecessary ADS diagnostics for all todlers.**

**by using this model , The healthcare professional can accurately predict , which toddler required ASD diagnostics and which dosent required such type of diagnostics tests.**

**This Model will predict perfectly , only if its will get the accurate traits details from toddlers parents.**

**References**

Chris Albon, Machine Learning with Python

Benjamin Bengfort, Applied Text Analysis with Python