**Introduction**

Autism is an inescapable formative issue that influences a great many kids in the world. To watch babies with a mental imbalance is to watch individuals taking part in practices that can appear to be weird to the vast majority: hand-fluttering, toe-strolling, a fixation on specific hues or protests the avoidance of everything else, and an appearing neglectful to their general surroundings. (Kukoff, Z. 2013).

Autism is an extraordinary self-preoccupation is a therapeutic condition demonstrating the abnormal lead in the external condition and correspondence crippling appeared by the dull exercises and innovative capacity. A psychological irregularity is similarly a range issue as the children can be affected by a variety of reactions, be it smooth to extraordinary. Thusly, this appearance can be perceived by this given application assumed ASD-Quiz, while there are 10 one of a kind sorts of request in the application that may distinguish the outrageous introspection as per age gathering. It is a puzzled helpful condition in which the necessities of the individual are varied (Edelson, 2018). As indicated by the investigation surveys drove in the United States, in every 10000 live births, there has been recognizable proof 4-5 greatly withdrawn new born children (Olson, 2010).

A mental imbalance range issue which is an Autism Spectrum Disorders (ASD) is the name for a gathering of situations that influence the manner in which individuals understand the world. ASD's are deep rooted issues that influence social and relational abilities and additionally the path those with ASD think and carry on.

Some might encounter noteworthy disability with great a mental imbalance, while those with Asperger's might be profoundly smart yet socially awkward. There is a tremendous variety in how individuals are influenced in the middle of these two extremes, for instance, some may have going with learning incapacities, some not really. Be that as it may, whatever their general level of insight, everybody with the condition shares a trouble in understanding the world.

Individuals with ADS will impart, communicate, act, and learn in ways that are not the same as most other individuals, with some being extremely tested in critical thinking and considering. A few people with ASD require a ton of assistance in their day by day lives; while others require less. (Janet, P. (2014, September). Autism spectrum disorders)

ASDs are normally seen from an early age, from birth to three years of age. On the off chance that you speculate your kid has an ASD, early recognizable proof is critical to guarantee they get suitable medications and support at the earliest opportunity.

Custom curriculum projects and support of family/whanau have a critical impact in helping the kid accomplish their maximum capacity as they develop into adolescents and grown-ups. Without this, it has been demonstrated that kids encountering ASD won't create and additionally they could. This will influence them for whatever is left of their lives.

Despite the fact that a ton of investigation into ASD has gone ahead around the globe, the correct reason is as yet obscure. It's realized that qualities some way or another have an impact with certain quality examples making it more probable that a tyke will have ASD.

Studies have additionally demonstrated that siblings or sisters of youngsters with ASD have a more prominent possibility of having it also. Likewise, kids destined to more seasoned guardians are had a higher shot of creating ASD than those destined to more youthful guardians. (Janet, P. (2014, September). Autism spectrum disorders)

The quick development in the quantity of ASD cases overall requires datasets identified with conduct qualities. Be that as it may, such datasets are uncommon making it hard to perform intensive examinations to enhance the effectiveness, affectability, specificity and prescient exactness of the ASD screening process. Directly, exceptionally constrained a mental imbalance dataset a mental imbalance screening strategy in addition to different attributes that have turned out to be viable in identifying the ASD, have been gathered.

In medical data sets, data are predominately composed of “normal” samples with only a small percentage of “abnormal” ones, leading to the so-called class imbalance problems. In class imbalance problems, inputting all the data into the classifier to build up the learning model will usually lead a learning bias to the majority class.

The challenge confronted in analysing the autism data in this project as the class is supposedly is imbalanced. In this manner, the issue will be extremely hard to get a best outcome as it will for the most part lead a learning bias to the majority class.

**DATA PRE-PROCESSING**

Machine Learning (ML) has several stages in that most important step and time-consuming step was data pre-processing. There are many features influences ML to achieve required task to complete successfully. In Data Analytics the utmost important task is data pre-processing or clean the data because the data we get will have noises or undependable data. This might be because of missing data, duplicate data or imbalanced data on the class variable. All these issues will mislead the learning face and leads to wrong or biased prediction. Data pre-processing face conceives majority of the time. Data pre-processing consists of various tasks such as cleaning, normalization, transformation, feature extraction and selection, and so on (Kotsiantis, Kanellopoulos, & Pintelas, 2006).

**Why Pre-processing?**

Data we get for processing will be one of the following such as, one of the cause used to be dataset might be incomplete, there might be attributes which has missing relevant data, some time required relevant attribute itself missing, some time we have only aggregated data. Sometimes data we got might be noisy it means data may have error or not usable or outliers but, in some cases, outliers might be useful especially in healthcare sector, rather it plays an important role on prediction. Data might be inconsistent it may contain incorrect data in names or codes (Kotsiantis, Kanellopoulos, & Pintelas, 2006).

There are different tasks to avoid above issues, that is know as data pre-processing tasks. First tasks are **Data cleaning**, it involves update missing values with appropriate values, noisy data need to be smoothened, search and find out outliers and depends on the type of data take decision on avoiding it or keeping it. Because in medical filed outliers plays a vital role. Then solving the irregularities in data. Next task is **Data Integration**, this may or may not be needed because data might be lying in different data sources such as multiple Relational Database Management systems and databases or flat files and so on. We need to integrate these data because one may not have same type or format as another. Next step is more critical which is **Data Transformation** which involves transforming the data to recognised by data analysis tool such as normalizing sometimes aggregating and so on. **Data Reduction**, which means dropping the volume of data without loosing the similar result. This normally does in case of big data analytics in terms of dimensionality for example we may have 500 instances but 5000 attributes, in such cases reducing the attributes plays a vital role. **Data discretization**, which means sometimes data to fit in a mode we may need to convert the numerical data into nominal ones (Xiong, Pandey, Steinbach, & Kumar).

(C, B, & S, 2012) In this project we have got the dataset produced from ASD Test screening app for toddler consists of question as per Q-Chat-10 Quantitative checklist for Autism. This chart consists of 10 questions and how the result needs to be grouped. Questioner are developed in the form of mobile app for ease of use[http://asdtests.com](http://asdtests.com/). The given dataset Toddler dataset consists of 1054 instances and 18 variables which includes class variable. The test is conducted in Medical, health and social science. The dataset dose not contains any missing values. Attributes of the dataset explains as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Attributes** | **Question** | **Datatype** | **Description** |
| A1 | Does your child look at you when you call his/her name? | Binary (1,0) | Sometimes / Rarly / Never “1” is mapped or else 0 |
| A2 | How easy is it for you to get eye contact with your child? | Binary (1,0) | Sometimes / Rarly / Never “1” is mapped or else 0 |
| A3 | Does your child point to indicate that s/he wants something? (e.g. a toy that is out of reach) | Binary (1,0) | Sometimes / Rarly / Never “1” is mapped or else 0 |
| A4 | Does your child point to share interest with you? (e.g. pointing at an interes9ng sight) | Binary (1,0) | Sometimes / Rarly / Never “1” is mapped or else 0 |
| A5 | Does your child pretend? (e.g. care for dolls, talk on a toy phone) | Binary (1,0) | Sometimes / Rarly / Never “1” is mapped or else 0 |
| A6 | Does your child follow where you’re looking? | Binary (1,0) | Sometimes / Rarly / Never “1” is mapped or else 0 |
| A7 | If you or someone else in the family is visibly upset, does your child show signs of waning to comfort them? (e.g. stroking hair, hugging them) | Binary (1,0) | Sometimes / Rarly / Never “1” is mapped or else 0 |
| A8 | Would you describe your child’s first words as | Binary (1,0) | Sometimes / Rarly / Never “1” is mapped or else 0 |
| A9 | Does your child use simple gestures? (e.g. wave goodbye) | Binary (1,0) | Sometimes / Rarly / Never “1” is mapped or else 0 |
| A10 | Does your child stare at nothing with no apparent purpose? | Binary (1,0) | Always / Usually / Sometimes are mapped to “1” else 0 |
| Age\_Mons | N/A | Number | Toddler in months |
| Score | N/A | Number | Possible value 1-10 (1-3 Means no ASD traits and greater than 3 means ASD traits) |
| Sex | N/A | Character | m - Male and f - Female |
| Ethnicity | N/A | String | Ethnicity of the toddler can be selected list will show common ethnicity |
| Jaundice | N/A | Boolean | yes/no, when toddler born with jaundice or not |
| Family\_ASD | N/A | Boolean | Any history with Autism in immediate family members who are directly related to toddler |
| Who completed the test | N/A | String | Who filled the questioner, example: Parent, self, Health Care Professional, family member and so on |
| Class | N/A | String | Whether the toddler has ASD traits or not which is automatically filled by the ASDTests app as (Yes/No) |

**Data Cleaning**

(Ilyas, 2018) Data Cleaning processes is the primary process in data analysis. Once we find the irregularities and inappropriate data, we need to do a careful cleaning to make it right. The cleaning comprises of filling missing values but our dataset did not contain any missing values, outlier detection, transformation, deduplication are the main thing some time there will be wrong entry such as from our dataset we have found 4 instances that ‘Who completed the test’ was answered ‘Self’ with ‘Health Care professional’ because toddler of 36 months or less will not be able to fill that information. Then as per the instruction and our experiment we have found that the attribute ‘Score’ make the result biased because of overfit. It is difficult to find duplicate data, but it is only possible by mapping the data. There is no definite line or steps for cleaning.

**Data Evaluation:**

Data evaluation involves better understanding and visualisation of data for finding out patterns. By performing descriptive analysis on data help us to build right predictive model.

For Data Evaluation we used Weka and Excel

1) Out of total 1054 intense in Class 326 is No and 728 is Yes.

2) Attribute A1 has 460 intense for score 0 where 264 is belongs to no and 196 belongs to yes class and score 1 has 594 intense where 62 belongs to no and 532 belongs to yes class.

3) Attribute A2 has 581 intense for score 0 where 289 are belong to no and 292 belong to yes class and score 1 has 473 intense where 34 belong to no and 439 belong to yes class.

4) Attribute A3 has 631 intense for score 0 where 293 are belong to no and 338 belong to yes class and score 1 has 423 intense where 33 belong to no and 390 belong to yes class.

5) Attribute A4 has 514 intense for score 0 where 282 are belong to no and 232 belong to yes class and score 1 has 540 intense where 44 belong to no and 496 belong to yes class.

6) Attribute A5 has 601 intense for score 0 where 292 are belong to no and 209 belong to yes class and score 1 has 553 intense where 34 belong to no and 519 belong to yes class.

7) Attribute A6 has 446 intense for score 0 where 275 are belong to no and 171 belong to yes class and score 1 has 608 intense where 52 belong to no and 556 belong to yes class.

8) Attribute A7 has 369 intense for score 0 where 245 are belong to no and 124 belong to yes class and score 1 has 685 intense where 81 belong to no and 604 belong to yes class.

9) Attribute A8 has 570 intense for score 0 where 280 are belong to no and 290 belong to yes class and score 1 has 484 intense where 46 belong to no and 438 belong to yes class.

10) Attribute A9 has 538 intense for score 0 where 307 are belong to no and 231 belong to yes class and score 1 has 516 intense where 19 belong to no and 497 belong to yes class.

11) Attribute A10 has 436 intense for score 0 where 178 are belong to no and 258 belong to yes class and score 1 has 618 intense where 148 belong to no and 470 belong to yes class.

12) Attribute sex has 319 intense for Female where 125 are belong to no and 194 belong to yes class and male has 735 intense where 201 belong to no and 534 belong to yes class.

13) Attribute Jaundice has 288 intense for Yes where 73 are belong to no and 215 belong to yes class and No has 766 intense where 253 belong to no and 513 belong to yes class.

14) Attribute Family with ASD has 884 intense for No, where 271 are belong to no and 613 belong to yes class and Yes has 170 intense where 55 belong to no and 116 belong to yes class.

15) ) Attribute Who completed the test has 1018 intense for family member, where 316 are belong to no and 702 belong to yes class ,For Health Care Professional total 33 intense where no class has 10 and yes class has 23 intense and For Others total has 3 intense all belong to yes class.

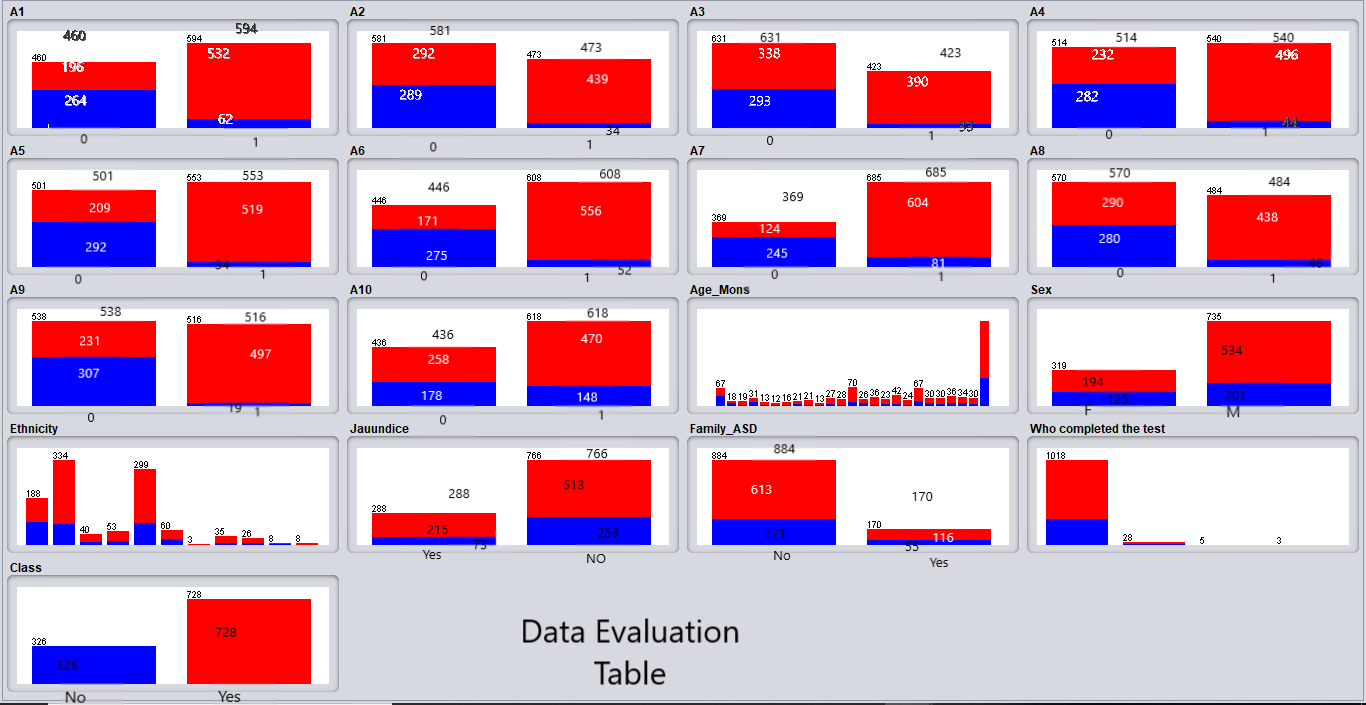
16) Attribute Ethnicity out of total 1054 observations break down is given under in table

|  |  |  |  |
| --- | --- | --- | --- |
|  | Class | |  |
|  | No | Yes | total |
| asian | 87 | 212 | 299 |
| black | 14 | 39 | 53 |
| Hispanic | 10 | 30 | 40 |
| Latino | 6 | 20 | 26 |
| middle eastern | 92 | 96 | 188 |
| mixed | 3 | 5 | 8 |
| Native Indian | 0 | 3 | 3 |
| Others | 6 | 29 | 35 |
| Pacifica | 1 | 7 | 8 |
| south asian | 23 | 37 | 60 |
| White European | 84 | 250 | 334 |
| Total | 326 | 728 | 1054 |

17) Attribute Age\_Mons Total 1054 observation break down is below

|  |  |  |  |
| --- | --- | --- | --- |
|  | Class | |  |
| Age in Months | No | Yes | total |
| 12 | 37 | 30 | 67 |
| 13 | 10 | 8 | 18 |
| 14 | 3 | 16 | 19 |
| 15 | 16 | 15 | 31 |
| 16 | 2 | 11 | 13 |
| 17 | 5 | 7 | 12 |
| 18 | 4 | 12 | 16 |
| 19 | 8 | 13 | 21 |
| 20 | 4 | 17 | 21 |
| 21 | 5 | 8 | 13 |
| 22 | 6 | 21 | 27 |
| 23 | 5 | 23 | 28 |
| 24 | 15 | 55 | 70 |
| 25 | 11 | 15 | 26 |
| 26 | 5 | 31 | 36 |
| 27 | 6 | 17 | 23 |
| 28 | 9 | 33 | 42 |
| 29 | 4 | 20 | 24 |
| 30 | 19 | 48 | 67 |
| 31 | 10 | 20 | 30 |
| 32 | 8 | 22 | 30 |
| 33 | 11 | 25 | 36 |
| 34 | 10 | 24 | 34 |
| 35 | 6 | 24 | 30 |
| 36 | 107 | 213 | 320 |
| Total | 326 | 728 | 1054 |

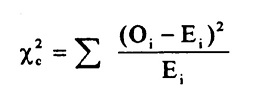
By Data Evaluation table it can be more easily visualize



**Feature Evaluation**:

Finding of correlation between dependent variable and independent variable we choose chi sq. test. Chi-square test is one of the most useful testing methods when variables are nominal or categorical. We want to test whether the incidence of a specific feature and the incidence of specific class are independent or not. If two feature is dependent than we can use the incidence of the feature to predict the incidence of the class.

Formula for chi-square test

The (Hopkins., 2008) chi-square 1

Where

c= the degrees of freedom(Df)

O= observed value

E = expected value

Df=(total row-1)(total col-1)[[1]](#endnote-1)

R Studio is used for conducting Chi-Square test.

Chi Square Test Table for Class Vs. Every variable

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | X-squared Value | Degree Of Freedom(df) | p-value |
| A1 | 265.34 | 1 | < 2.2e-16 |
| A2 | 224.39 | 1 | < 2.2e-16 |
| A3 | 175.12 | 1 | < 2.2e-16 |
| A4 | 266.83 | 1 | < 2.2e-16 |
| A5 | 332 | 1 | < 2.2e-16 |
| A6 | 339.26 | 1 | < 2.2e-16 |
| A7 | 331.75 | 1 | < 2.2e-16 |
| A8 | 190.46 | 1 | < 2.2e-16 |
| A9 | 348.82 | 1 | < 2.2e-16 |
| A10 | 33.301 | 1 | = 7.894e-09 |
| Age\_Mons | 56.625 | 24 | = 0.0001883 |
| Sex | 14.044 | 1 | = 0.0001786 |
| Ethnicity | 43.571 | 10 | = 3.93e-06 |
| Jauundice | 5.427 | 1 | = 0.01983 |
| Family\_ASD | 0.12094 | 1 | **= 0.728** |
| Who.completed.the.test | 1.3554 | 2 | **= 0.5078** |

**Result Analysis of Chi-square test:**

Setting up p-value >= 0.05 for accepting Null hypothesis means there is no association between variables

P-value <0.05 means rejecting Null hypothesis and accepting alternating hypothesis means there is association or relationship between variables.

So for ASD toddler data set by p-value we can say that variable family with ASD and who completed the test has no association or correlation with Class. By chi-sq. test we can say that these two variables are not making any significance for predicting class. Thus for building predictive model these two variables can be dropped.

# **References**

C, A., B, A., & S, B.-C. (2012). Q-CHAT-10 Quantitative Checklist for Autism in Toddlers. *Journal of the American Academy of Child and Adolescent Psychiatry* , 1.

Hopkins., J. (2008, August 01). *Use of the Chi-Square Statistic.* Retrieved August 10, 2018, from The Johns Hopkins University: http://ocw.jhsph.edu/courses/fundepiii/PDFs/Lecture17.pdf

Ilyas, I. (2018, April). Data cleaning is a machine learning problem that needs data systems help. *CHERITON SCHOOL OF COMPUTER SCIENCE*.

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006, NUMBER 1). Data Preprocessing for Supervised Leaning. *INTERNATIONAL JOURNAL OF COMPUTER SCIENCE*, 111-117.

Xiong, H., Pandey, G., Steinbach, M., & Kumar, V. (n.d.). Enhancing Data Analysis with Noise Removal. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*.

(Hopkins., 2008)

1. [↑](#endnote-ref-1)