**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.

# **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. Retrieved from http://www.autismalert.org/uploads/PDF/SCREENING--AUTISM--QCHAT-10%20Question%20Autism%20Survey%20for%20Toddlers.pdf

Ilyas, I. (2018, April). Data cleaning is a machine learning problem that needs data systems help. CHERITON SCHOOL OF COMPUTER SCIENCE. Retrieved from https://cs.uwaterloo.ca/news/data-cleaning-machine-learning-problem-needs-data-systems

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006, NUMBER 1). Data Preprocessing for Supervised Leaning. INTERNATIONAL JOURNAL OF COMPUTER SCIENCE, 111-117. Retrieved from https://www.researchgate.net/publication/228084519\_Data\_Preprocessing\_for\_Supervised\_Learning

Xiong, H., Pandey, G., Steinbach, M., & Kumar, V. (n.d.). Enhancing Data Analysis with Noise Removal. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING. Retrieved from http://datamining.rutgers.edu/publication/tkdehcleaner.pdf