A COMPARATIVE STUDY ON AUTISM SPECTRUM DISORDER DETECTION BY LEVERAGING MACHINE LEARNING

A.D.V.N. Murthy[1], G. Neeraja[2], G.Venkata Karthik[3], B. Samanvitha[4], G. Pavan Ganesh Naidu[5]

[1] Assistant Professor
[2-5] B. Tech Student, LIET
[1,2,3,4,5] Computer Science & Engineering, Lendi Institute of Engineering and Technology, Vizianagaram

---***---

Abstract - We propose a study to compare machine learning (ML) algorithms for early detection of Autism Spectrum Disorder (ASD). Our approach involves four feature scaling methods and eight ML algorithms: Ada Boost, Random Forest, Decision Tree, K-Nearest Neighbours, XGBoost, Logistic Regression, Support Vector Machine, and Linear Discriminant Analysis. We evaluate these algorithms on datasets representing different age groups: Toddlers, Adolescents, Children, and Adults. After assessing performance using accuracy, precision, recall, F1-score, MCC, and Kappa, XGBoost emerges as the top performer, followed closely by Random Forest. Linear Discriminant Analysis and AdaBoost exhibit respectable scores, while Support Vector Machine and Logistic Regression offer moderate performance. K-Nearest Neighbors and Decision Tree perform weaker. Our study helps identify key factors contributing to ASD risk and ranks them using different techniques. This enables healthcare professionals to prioritize their assessments when screening for ASD. Overall, our method shows promise for early ASD detection, offering a valuable tool for healthcare providers.

Key Words: Autism Spectrum Disorder, Machine Learning, Feature Scaling Technique, Feature Selection Technique, Classifications.

1. INTRODUCTION

Autism Spectrum Disorder (ASD) is a complex condition that affects individuals across their lifespan, impacting social interaction, communication, and behavior. Despite increased awareness and research, there remains a critical need for innovative approaches, therapies, and support mechanisms to enhance the quality of life for individuals with ASD and their families. Machine learning (ML) is a crucial tool in this endeavor, as it can identify patterns, risk factors, and subtypes of ASD, facilitating early detection, accurate diagnosis, and personalized treatment planning. ML algorithms can also predict treatment responses, optimize interventions, and advance our understanding of the neurobiological underpinnings of ASD.

In this study, we employed eight ML algorithms: Ada Boost, Random Forest, Decision Tree, K-Nearest Neighbors, Gaussian Naïve Bayes, Logistic Regression, Support Vector Machine, and Linear Discriminant Analysis. These algorithms were used to tackle various aspects of the problem domain effectively, including classification tasks, proximity-based methods, and dimensionality reduction.

To enhance the performance of these ML models, we utilized various feature selection and feature scaling techniques. Feature selection techniques, such as Information Gain Attribute Evaluation and Correlation-based Feature Selection, identified the most relevant features for predicting ASD across different age groups. Feature scaling techniques, including Quantile Transformation and Normalization, ensured that all features contributed equally to the model's training process and prevented attributes with larger scales from dominating the learning algorithm. By combining feature selection and feature scaling techniques, we aimed to develop robust and accurate predictive models for ASD detection across various age groups. These techniques contributed to the optimization of model performance, ensuring that the ML algorithms could effectively leverage the available data to identify early signs of ASD.

2. PROPOSED SYSTEM

The framework utilizes various machine learning (ML) classifiers trained on different feature-scaled datasets representing different age groups: Toddlers, Children, Adolescents, and Adults. The classifiers are evaluated based on several performance metrics such as accuracy, precision, recall, ROC, F1-score, kappa, log loss, and MCC. The study demonstrates promising results, with different ML classifiers achieving high-performance metrics across the different age groups.

ALGORITHMS:

Random Forest (RF)

Random Forest (RF) is a decision tree-based ensemble classification method that combines multiple decision trees to generate a forest. The algorithm works in two steps: first, it constructs a decision tree for each random sample from the training data set, and then it makes predictions for each test sample based on a majority vote from the decision trees. The workflow involves selecting a random sample, constructing a decision tree, and repeating the process to create a forest of 'N' decision trees. Finally, the class value is assigned to the test sample based on the majority vote.

Decision Tree (DT)

To create a predictive model for class values, DT uses training data-inducing decision-making rules in a top-down manner. The information gain method was employed in this study to determine the optimal characteristic.

Logistic Regression (LR)

Using a dataset of independent variables, logistic regression determines the probability of an event, like voting or not, occurring. Since the outcome is a probability, the range of the dependent variable is 0 to 1. In logistic regression, the probability of success, or chances, Utilizing the logit formula, the values are converted and split by the likelihood of failure.

Linear Discriminant Analysis (LDA)

Although LDA reduces dimensionality, it may also be used to classify data by examining the linear combination of characteristics. The Bayes theorem is used by LDA to estimate the probability.

Ada Boost (AB)

An ensemble classifier based on trees; AB reduces misclassification errors by combining many weak classifiers. The algorithm is retrained by choosing the training set and assigning weights iteratively based on the accuracy of the prior training. Using an arbitrary subset of the entire training set, AB gives weights to each occurrence and classifier to train any weak classifier.

K-Nearest Neighbors (KNN)

KNN uses the training data directly to classify the test data by determining the K value, which represents the number of KNNs. It calculates the distance for every occurrence. sorting the distance between each training instance. In addition, the test data's final class label is assigned using a majority vote procedure. The distances between the cases in this study are computed using Euclidean distance.

XGBoost (XGB)

XGBoost is a robust machine-learning algorithm that can help you understand your data and make better decisions. XGBoost is an implementation of gradient-boosting decision trees. It has been used by data scientists and researchers worldwide to optimize their machine-learning models. XGBoost is designed for speed, ease of use, and performance on large datasets. It does not require optimization of the parameters or tuning, which means that it can be used immediately after installation without any further configuration.

Support Vector Machine (SVM)

In general, SVM performs well when applied to high-dimensional data that has nonlinear mappings and can be used to classify both linear and nonlinear data. It investigates the best hyperplane or decision boundary for classifying data. This research Uses the Radial Basis Function (RBF) as the kernel function, and SVM lowers the upper bound of the predicted test error while automatically defining centers, weights, and thresholds.

Feature Scaling Techniques:

Four different types of feature scaling techniques are used: Quantile Transformer (QT), Power Transformer (PT), normalizer, and Mean Absolute Scaling (MAS).

Quantile Transformer (QT): QT transforms features to follow a Gaussian distribution, mapping data to a uniform distribution and then to a Gaussian distribution, mitigating the impact of outliers.

Power Transformer (PT): PT applies a power transformation to make data more Gaussian-like, stabilizing variance and improving normality, and can handle both positive and negative data values.

Normalizer: Normalization rescales features to a fixed range (0-1) by subtracting the minimum value and dividing by the range, ensuring each feature contributes equally to the analysis.

Mean Absolute Scaling (MAS): MAS scales features based on their mean absolute deviation from the median, resulting in standardized features with a median absolute deviation of 1, robust to outliers and skewed distributions.

3. SYSTEM ARCHITECTURE

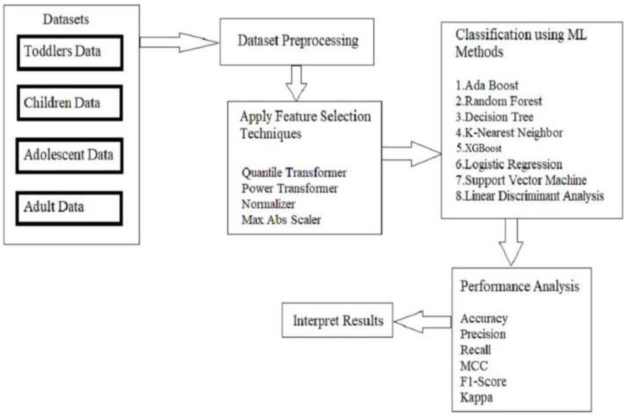
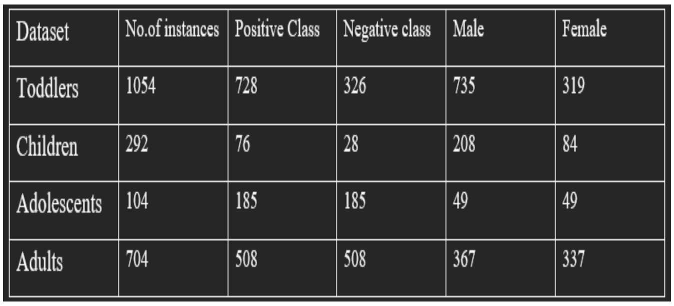


Fig -1: System Architecture of ASD Proceeding

4. RESULTS AND DISCUSSIONS

We collect four ASD datasets (Toddlers, Adolescents, Children, and Adults) from the publicly available repositories: Kaggle and UCI ML. The authors in [13] created the ASDTests smartphone app for Toddlers, Children, Adolescents, and Adults ASD screening using QCHAT-10 and AQ-10. The application computes a score of 0 to 10 for every individual, with which the final score is 6 out of 10 which indicates an individual has positive ASD. In addition, ASD data is obtained from the ASDTests app while open-source databases are developed in order to facilitate research in this area. The detailed description of the Toddlers, Children, Adolescents, and Adults ASD datasets are given in the below tables:



Dataset	No.of instances	Positive Class	Negative class	Male	Female
Toddlers	1054	728	326	735	319
Children	292	76	28	208	84
Adolescents	104	185	185	49	49
Adults	704	508	508	367	337

Table-1: Datasets Description

We develop a generalized ML framework for early-stage detection of ASD in people of different ages. We solve the imbalanced class distribution issue through the Random Over Sampler to avoid the ML models being biased towards the majority class samples. We select the best Feature Scaling (FS) method to map individual ASD dataset's feature values to improve the prediction performance. We investigate eight simple but effective ML approaches on each feature-scaled ASD dataset, analyze their classification performances, and identify the best FS techniques for each ASD dataset. Furthermore, we also calculate and analyze the feature importance values on each best feature-scaled ASD.

RANDOM FOREST:

The Random Forest classifier was evaluated on four datasets: toddlers, children, adolescents, and adults, using various scaling techniques. For toddlers, Quantile Transformer (QT) yielded high accuracy (97.6%) and all-round performance, while Normalizer showed the best performance except for precision. Among children, QT and Power Transformer performed best across metrics, with QT showing superior accuracy (98.3%) and recall (100%). In the adolescent dataset, Power Transformer with Yeo-Johnson method exhibited the highest performance, followed by QT and Normalizer, while MaxAbsScaler had the lowest performance. For adults, both Quantile Transformer and MaxAbsScaler achieved perfect scores across all metrics, indicating their suitability for the Random Forest model, followed closely by Power Transformer. Although these findings underscore the importance of scaling techniques in model performance, it's crucial to validate these results across diverse datasets and models for robust generalization.

DECISION TREE:

In analyzing the Decision Tree classifier across four datasets, it's evident that scaling techniques play a varying role in performance. For toddlers, all scaling methods exhibit similar effectiveness, with QuantileTransformer yielding slightly higher accuracy (94.8%) and metrics. Conversely, in the children's dataset, Quantile Transformer and PowerTransformer outshine MaxAbsScaler and Normalizer due to their preservation of data distribution, with accuracy at 89.8%. Adolescents show the best performance with MaxAbsScaler and PowerTransformer, indicating their suitability over Normalizer. However, for adults, perfect scores across all metrics suggest potential overfitting, especially with QuantileTransformer, MaxAbsScaler, and PowerTransformer, highlighting the Decision Tree Classifier's susceptibility to training data noise and the need for caution in model complexity.

LOGISTIC REGRESSION:

The Logistic Regression model's performance on four datasets targeting different age groups highlights the influence of scaling techniques. For toddlers, Quantile Transformer (QT) and MaxAbsScaler exhibit perfect accuracy, precision, recall, and F1 scores, while Normalizer underperforms significantly. Power Transformer shows promise but requires further investigation. In the children's dataset, Power Transformer demonstrates the highest metrics, potentially indicating overfitting. Adolescents' data favor MaxAbsScaler, with Quantile Transformer also performing well. However, different classifiers may yield different results. For adults, all scaling techniques yield perfect scores, suggesting successful model training, but generalization to new data warrants caution. Regularization and cross-validation are recommended to guard against overfitting.

LINEAR DISCRIMINANT ANALYSIS:

Analyzing Linear Discriminant Analysis (LDA) across four datasets reveals the impact of scaling techniques on performance. For toddlers, all scaling methods perform well, with Normalizer slightly outperforming others in most metrics. However, differences are minor, indicating the choice may depend on specific dataset characteristics. In the children's dataset, PowerTransformer yields the highest accuracy and F1 score, suggesting its superiority. Adolescents also benefit from MaxAbsScaler or PowerTransformer, enhancing separability and reducing outlier impact. In adults, PowerTransformer demonstrates the best performance across all metrics, followed closely by QuantileTransformer, while Normalizer exhibits the lowest performance. These findings underscore the significant influence of scaling techniques on LDA performance, emphasizing the importance of selecting appropriate methods based on dataset characteristics to optimize classifier effectiveness.

ADABOOST:

Across toddlers, children, adolescents, and adults datasets, the AdaBoostClassifier demonstrates robustness to scaling techniques, performing nearly perfectly with all methods. For toddlers and children, QuantileTransformer, MaxAbsScaler, and PowerTransformer achieve flawless metrics, with Normalizer showing minor decreases in recall and F1 score. Similarly, in adolescents, scaling techniques perform comparably, with QT, maxabs, and power exhibiting similar accuracy, F1 score, MCC, and kappa. Conversely, normalizer exhibits lower performance. In adults, all scaling techniques achieve perfect accuracy, with QT and maxabs performing slightly better in precision, recall, F1 score, MCC, and kappa compared to normalizer and power. These findings suggest that while scaling techniques may not significantly impact AdaBoost classifier performance, maintaining feature similarity through scaling remains beneficial.

K-NEAREST NEIGHBOUR:

In evaluating K-Nearest Neighbors (KNN) across toddlers, children, adolescents, and adults datasets, the impact of scaling techniques on performance varies. For toddlers, MaxAbsScaler and Normalizer show slightly higher scores in most metrics, suggesting their effectiveness. Conversely, in children, QuantileTransformer outperforms other techniques, while Normalizer demonstrates the poorest performance, potentially due to its scaling method. Adolescents exhibit varying impacts of scaling techniques, with 'maxabs' and 'power' scalers showing higher accuracy, precision, recall, and F1 score, while 'QT' scaler achieves perfect recall but slightly lower precision. Finally, in adults, 'QT' scaler performs the best across all metrics, followed closely by 'maxabs', with 'power' scaler slightly trailing behind. Notably, 'normalizer' scaler consistently performs the worst across all datasets. These results emphasize the significant influence of scaling techniques on KNN classifier performance and underscore the importance of selecting appropriate methods based on dataset characteristics for optimal performance.

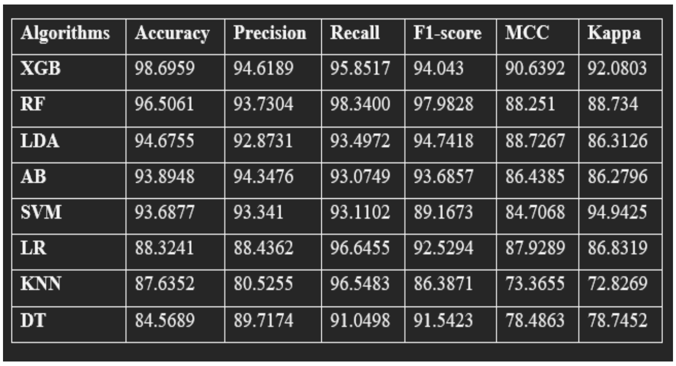
XGBOOST:

Analyzing XGBoost across toddlers, children, adolescents, and adults datasets reveals varying impacts of scaling techniques on performance. For toddlers, all scaling methods perform well, with Normalizer exhibiting slightly higher accuracy, recall, F1 score, MCC, and kappa. In children, QuantileTransformer, MaxAbs, and Power scalers outperform Normalizer, showing consistently high metrics. Adolescents show similar performance across QuantileTransformer, MaxAbsScaler, and PowerTransformer, while Normalizer lags behind significantly. Similarly, in adults, all scaling techniques achieve perfect scores on the training set, indicating potential overfitting, especially considering the small dataset size. These findings suggest that while scaling techniques may have varying impacts on XGBoost classifier performance, they are essential for achieving good results, especially in datasets with limited samples, and careful consideration should be given to selecting appropriate methods to avoid overfitting.

SUPPORT VECTOR MACHINE:

Analyzing the Support Vector Machine (SVM) performance across toddlers, children, adolescents, and adults datasets reveals the impact of different scaling techniques. For toddlers, Quantile Transformer and MaxAbsScaler show the highest accuracy, precision, recall, F1 score, MCC, and kappa, while Normalizer lags behind significantly. Similarly, in children, Quantile Transformer scaling exhibits the best performance, followed closely by MaxAbsScaler, with Normalizer significantly trailing. Adolescents demonstrate perfect classification with Quantile Transformer and MaxAbsScaler, while PowerTransformer also performs well but slightly lower. However, Normalizer shows poorer performance compared to other techniques. Finally, in adults, 'QT' and 'maxabs' scalers achieve perfect classification across all metrics, indicating their suitability for this dataset. In contrast, 'power' scaling also yields perfect classification, albeit slightly lower than 'QT' and 'maxabs'. Conversely, 'normalizer' scaling leads to less accurate results. In summary, the choice of scaling technique significantly impacts SVM classifier performance, with 'QT' and 'maxabs' scalers emerging as optimal choices across all age groups. It's essential to note that these are general observations, and the performance of feature scaling techniques may vary depending on the dataset and algorithm used. Therefore, it's always a good practice to experiment with different feature scaling techniques to determine the most effective one for each use case.

Performance Evaluation Between Models:



Algorithms	Accuracy	Precision	Recall	F1-score	MCC	Kappa
XGB	98.6959	94.6189	95.8517	94.043	90.6392	92.0803
RF	96.5061	93.7304	98.3400	97.9828	88.251	88.734
LDA	94.6755	92.8731	93.4972	94.7418	88.7267	86.3126
AB	93.8948	94.3476	93.0749	93.6857	86.4385	86.2796
SVM	93.6877	93.341	93.1102	89.1673	84.7068	94.9425
LR	88.3241	88.4362	96.6455	92.5294	87.9289	86.8319
KNN	87.6352	80.5255	96.5483	86.3871	73.3655	72.8269
DT	84.5689	89.7174	91.0498	91.5423	78.4863	78.7452

Table-2: Performance Evaluation between Models

The XGB (Proposed) algorithm performs the best overall, achieving the highest accuracy, precision, recall, F1-score, MCC, and Kappa among the listed algorithms.Random Forest (RF) also demonstrates strong performance across all metrics. Linear Discriminant Analysis (LDA) and AdaBoost (AB) (Proposed) exhibit similar performance, with slightly lower accuracy compared to RF and XGB. Support Vector Machine (SVM) and Logistic Regression (LR) (Proposed) achieve moderate performance, with lower accuracy but higher recall. K-Nearest Neighbors (KNN) and Decision Tree (DT) show comparatively weaker performance among the listed algorithms, with lower accuracy, precision, and MCC.

Overall, XGB (Proposed) and RF are recommended for tasks where high performance across multiple metrics is desired, while considering the specific requirements and trade-offs of each algorithm.

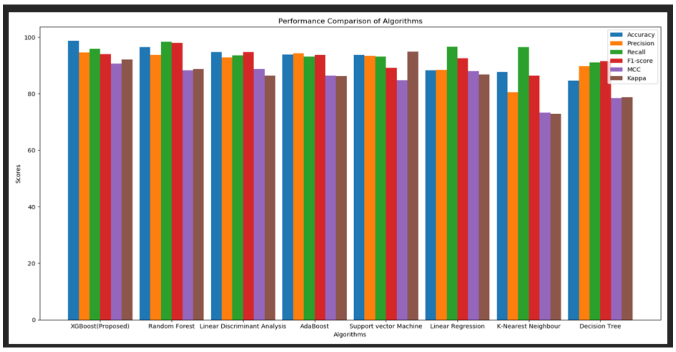


Fig-2: Graphical Representation of Performance Comparision of Algorithm

5. CONCLUSIONS

In conclusion, our study presents an innovative approach to early detection of Autism Spectrum Disorder (ASD) using machine learning techniques. By comparing various ML algorithms and employing different data preparation methods, we aimed to improve the accuracy and efficiency of ASD detection across different age groups.

After evaluating algorithms based on accuracy, precision, recall, F1-score, MCC, and Kappa, XGBoost (XGB) emerges as the top performer, showcasing exceptional performance across all metrics. Random Forest (RF) closely follows with strong performance, while Linear Discriminant Analysis (LDA) and AdaBoost (AB) exhibit respectable scores but slightly trail XGB and RF. Support Vector Machine (SVM) and Logistic Regression (LR) offer moderate performance with higher recall but lower accuracy. K-Nearest Neighbors (KNN) and Decision Tree (DT) show weaker performance. In conclusion, XGBoost (Proposed) emerges as the top choice for its exceptional performance across all evaluated metrics, closely followed by Random Forest, while other algorithms also offer viable alternatives depending on specific task requirements and constraints.

Furthermore, our study provides valuable insights into the factors contributing to ASD risk by ranking their importance using different techniques. This information can aid healthcare professionalsin prioritizing assessments during ASD screening, leading to more targeted and efficient interventions.

Overall, our method shows promise for enhancing early ASD detection compared to existingapproaches. By leveraging machine learning techniques and comprehensive data analysis, we offer a valuable tool for healthcare providers to improve outcomes and quality of life for individuals with ASD. However, further research and validation are necessary to fully assess the feasibility and effectiveness of our approach in clinical settings.

ACKNOWLEDGEMENT

We would like to thank the Department of Computer Science & Engineering Lendi Institute of Engineering and Technology, Vizianagaram for helping us to carry out the work and supporting us throughout the research.

REFERENCES

1. Mousumi Bala, Mohammad Hanif Ali, Md.Shahriare Satu, Khondokar Fida Hasan and Mohammad Ali Moni "Efficient Machine Learning Models for Early Stage Detection of Autism Spectrum Disorder," Algorithms, vol. 15, p. 166, May 2022. https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-ml/
2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10352530/
3. S. B. Shuvo, J. Ghosh, and A. S. Oyshi, ''A data mining-based approach to predict autism spectrum disorder considering behavioral attributes,'' in Proc. 10th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT), Jul. 2019, pp.1–5.
4. F. N. Buyukoflaz and A. Ozturk, ''Early autism diagnosis of children with machine learning algorithms,'' in Proc. 26th Signal Process. Commun. Appl. Conf. (SIU), May 2018, pp. 1–4.
5. M. F. Misman, A. A. Samah, F. A. Ezudin, H. A. Majid, Z. A. Shah, H. Hashim, and M. F. Harun, ''Classification of adults with autism spectrum disorder using deep neural network,'' in Proc. 1st Int. Conf. Artif. Intell. Data Sci. (AiDAS), Sep. 2019, pp. 29–34.
6. https://www.geeksforgeeks.org/support-vector-machine-algorithm/
7. F. Z. Subah, K. Deb, P. K. Dhar, and T. Koshiba, ''A deep learning approach to predict autism spectrum disorder using multisite resting-state fMRI,'' Appl. Sci., vol. 11, no. 8, p. 3636, Apr. 2021.
8. H. Abbas, F. Garberson, E. Glover, and D. P. Wall, ''Machine learning approach for early detection of autism by combining questionnaire and home video screening,'' J. Amer. Med. Informat. Assoc., vol. 25, no. 8, pp. 1000–1007, 2018.
9. Kaggle. (2022). Autism Spectrum Disorder Detection Dataset for Toddlers.[Online]. Available: https://www.kaggle.com/fabdelja/autism-screeningfor-toddlers
10. UCI. (2022). UCI Machine Learning Repository: Autistic Spectrum Disorder Screening Data for Adolescent dataset [Online]. Available: https://shorturl.at/fhxCZ
11. UCI. (2022). UCI Machine Learning Repository: Autism Screening Adult dataset [Online]. Available:https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult
12. UCI. (2022). UCI Machine Learning Repository: Autistic Spectrum Disorder Screening Data for Children dataset [Online]. Available: https://shorturl.at/fiwLU
13. https://www.kaggle.com/datasets/uppulurimadhuri/dataset?select=data_csv.csv Dataset for Children
14. https://archive.ics.uci.edu/dataset/420/autistic+spectrum+disorder+screening+data+for+adolescent Dataset for Adolescent
15. https://www.kaggle.com/datasets/andrewmvd/autism-screening-on-adults Dataset for Adults
16. https://www.kaggle.com/datasets/fabdelja/autism-screening-for-toddlers?select=Toddler%2BAutism%2Bdataset%2BJuly%2B2018.csv Dataset for Toddlers
17. Autism Screening Adult dataset Source by Fadi Fayez Thabtah, Department of Digital Technology, Manukau Institute of Technology, Auckland, New Zealand. https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult