Machine Learning Report

Z0183990

Submitted: February 20, 2023

1. INTRODUCTION

In 2022 the Centres for Disease Control and Prevention released a shocking estimate that 5.5 million adults in the United States alone are living with ASD [3]. In addition, that same year WHO also released an estimate that 1 in 100 children worldwide are born with ASD [6]. Autism Spectrum Disorder (ASD) is a serious neurodevelopmental disorder that spans all ages and comes with high healthcare costs. Yet, due to societally embedded stigma, a general lack of knowledge, long wait times, and a lack of access to formal assessment processes, a significant proportion of autistic adults are not diagnosed until later in life [1]. A mixed method study published in the Journal of Autism and Developmental Disorders concluded that the general quality of lives of those living with undiagnosed autism is substantially improved upon official diagnosis, though this is often incredibly hard to obtain [1].

In recent times, the application of Machine Learning to cross-disciplinary subjects in the fields of biology and neurology has been incredibly successful. In clinical decision-making and diagnostics, machine learning methods are often employed to visualise patterns and trends present in large sets of medical data, aiding data interpretation.

At present there are various studies dedicated to aiding the early diagnosis system of ASD in young children. However, the majority of these studies revolve around medical professionals, collecting data, and analysing brain imaging [7], leaving a gap in the research for a quick, accessible method of diagnosis, potentially through machine learning methods.

This report will focus on pre-processing and classifying a likely ASD diagnosis based off of the ASD in Adults dataset from the UCI database [4]. Gathered by Dr. Fadi Fayez of the Nelson Marlborough Institute, some of the data records are optional, and thus the dataset is incomplete, making it unfit to directly influence medical decisions.

1.1. Data Exploration

The relevant dataset consists of 704 observations with 21 variables; 2 of which are numeric, with the rest being either categorical or binary in nature.

For the purposes of predicting a likely diagnosis, the target class of this dataset will be what is known as the "Class_ASD". This variable is a factor with two levels: 'YES' (for a likely diagnosis) and 'NO'. Fig. 1 depicts that the data is unbalanced, with proportionality showing a large bias towards the 'NO' predictor.

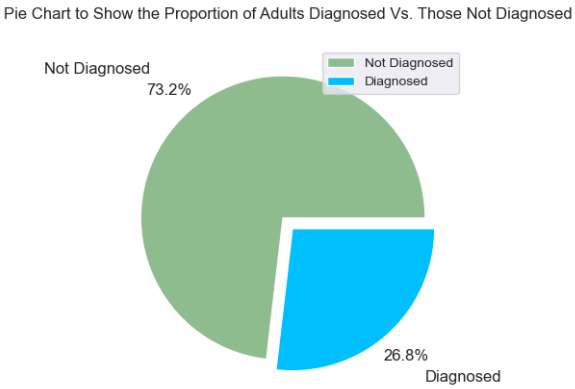

FIG. 1: Proportions of the Target Class in the Adult ASD Dataset

The dataset contains 95 rows with missing values, predominantly from categorical variables. As it is difficult to create replacement values these observations were removed. The remaining dataset is then made up of 609 observations.

The dataset required various other minor cleaning tasks. The 'age' feature contained an impossibly high maximum of 383. This was assumed a mistake and converted to 38. The new max age is 64. The 'ethnicity' feature contained duplicate categories 'Others' and 'others'. These were combined into one: 'Others' category.

1.2. Data Visualisation

Fig. 2 illustrates the relationship between various categorical features and the target class (ASD). The boxplot on the left highlights the distribution of gender with age alongside the respective target class. Both genders share similar age ranges, though men tend to have a wider distribution of individuals suspected of having ASD. This distribution follows the 4:1 men to female ratio in medical diagnosis for ASD and is expected.

The second boxplot takes a closer look at the spread of men and women against the resultant overall score of the survey. Perhaps unexpectedly the figure shows that there is a balanced distribution between the genders, meaning that men and women are treated equally by this dataset. Interestingly, the positive ASD classes show that a score of 7+ in the result feature tends to classify a suspected case of ASD.

The bar chart on the left shows that there is no significant relationship in this dataset between jaundice and having ASD, as the number of diagnosis' is approximately the same for both classification levels. The second barchart shows the same result but for having a family member with ASD. The second result, while true for the dataset, does not match our understanding of ASD generally. The risk of ASD when having a sibling with ASD is increased 8-fold to the general population. This suggests that the 'autism' category may not be representative of real-world findings.

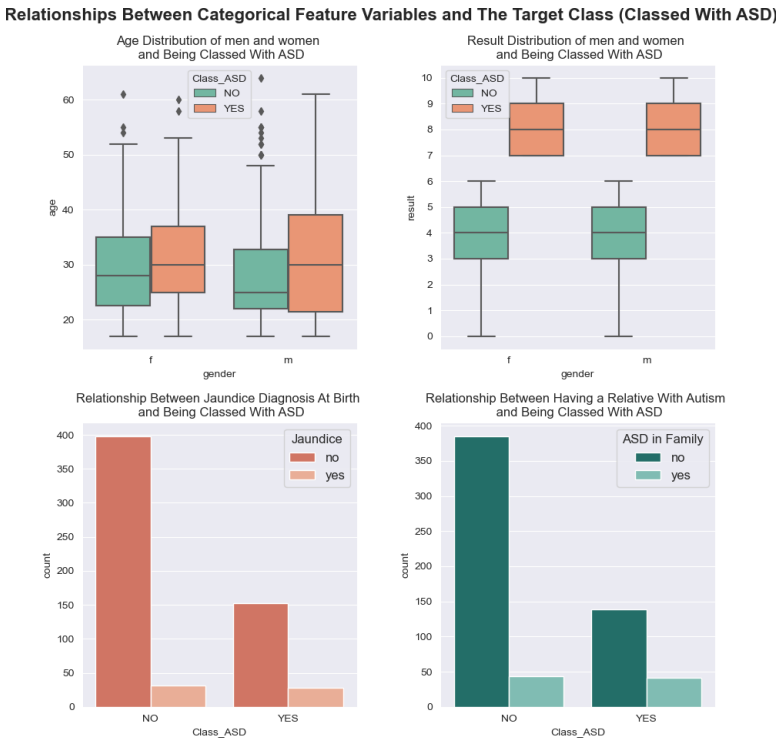

FIG. 2: Relationships Between Categorical Features and the Target Class

1.3. Pre-Processing

Before modelling the data, various irrelevant or unsustainable features were removed from the dataset. The 'used_app' feature provides no relevant insight into the target variable and is provided merely as a personal survey for the developer. In addition, the 'age description' feature also carries a singular factor and thus can have no significant bearing on the predicted outcomes.

The 'country of relevance' feature could potentially provide useful insight on the target class diagnosis, however, the feature contains over 60 potential answers which is too large to be reasonably processed by certain classification functions. This variable is then removed for the sake of model compatibility.

Finally, the 'result' feature carries too much information as to the resultant target class. This report is interested in potentially providing an accessible diagnosis based on symptoms, and thus the model training process may be influenced too heavily by the result feature, undermining the importance of various symptomatic features.

As 15 of the 16 predicting features left are either binary or categorical, to facilitate the training process, these features will be made numeric through the process of one-hot-encoding. Though this form of data processing does inflate the number of variables significantly (from 16 to 42), the dataset is large enough to accommodate such a transition, especially when paired with the application of k-fold cross validation when evaluating model performance.

The fully processed dataset then consists of 609 observations, 42 features, and a target class with 2 factor levels; 'NO': 0, 'YES':1. These observations were then split into features and results to be used in Cross-Validation.

2. ANALYSIS AND MODELLING

In order to classify the dataset, three popular machine learning methods were applied to the classification problem: Decision Tree, Random Forests, Sequential Neural Network.

All of the three models' predictive performance will be evaluated using an Accuracy and F1 score statistic calculated through an application of 10-fold cross validation. Accuracy measures model performance in terms of true positives and true negatives to all positive and negative observations, whilst the F1 score is a function of precision and recall.

It is perhaps a preferable measure to accuracy in datasets where the target class is heavily imbalanced, as is true in the ASD dataset. Furthermore, the F-statistic takes into account the False Positives and False Negatives, which are crucial to a model used in clinical decision-making. Diagnosing someone falsely, or missing an ASD diagnosis is, in this case, just as important as a correct diagnosis. Nevertheless, both statistical scores will therefore be used to evaluate the models presented in this paper.

To create a robust evaluation scheme, 10-Fold Cross-Validaiton has been selected to ensure that the scores of the following models do not depend on the manner in which training and test sets have been selected. Cross-Validation will ensure that every observation from the original dataset has a chance of appearing in both sets to help ensure that no patterns are overlooked in the data.

For reproducibility purposes, training and test splits, alongside model creation and evaluation are done under a set seed of (704) in Python.

2.1. Decision Tree

A decision tree is a non-parametric, supervised learning algorithm which takes a hierarchical tree structure consisting of a root node, branches, internal nodes and leaf nodes. Decision trees then organise a series of rules to partition a feature space into a number of smaller regions using identified splitting points.

To identify optimal splitting points within a tree structure, Decision trees employ a greedy algorithm whereby an optimal, local choice is made at each node. The splitting process is then repeated in a top-down, recursive manner until all (or a majority) of the data points have been classified under specific class labels.

There are various algorithms which one can adopt to select which attributes from a dataset should constitute the root or an internal node. The model presented in this paper employs the 'entropy' criteria. Entropy is a measure of the purity of a sub split and always lies between 0 and 1. Generally, the Gini index is the preferred algorithm, however, after trial and error, it was concluded that entropy yielded a higher mean validation accuracy and f1 score on the overall model. Perhaps this is to do with Entropy's consideration of the uncertainty involved in a choice.

In order to combat over-fitting, this model employs a backwards pruning method where non-significant branches are removed using a cost-complexity pruning technique.

Fig. 3 shows the accuracy for the training and validation sets by each value of a potential pruning parameter α. The larger the value of α, the more branches will be pruned.

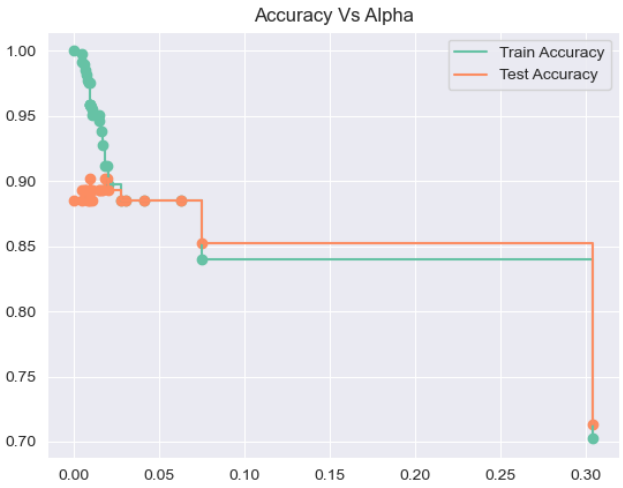

FIG. 3: Training and Test Value Accuracies According To Varying Values of Alpha.

Choosing the optimal value of (alpha) was in this instance rather a matter of trial and error through the final value chosen was 0.04.

Fig. 4 depicts the accuracy scores for the Decision Tree model before and after pruning. According to the graph, the pruning technique appears rather successful. Whereas the first bar chart shows clear over-fitting where the training accuracy often reaches the optimal while the validation accuracy suffers, the second shows more balance across the folds.

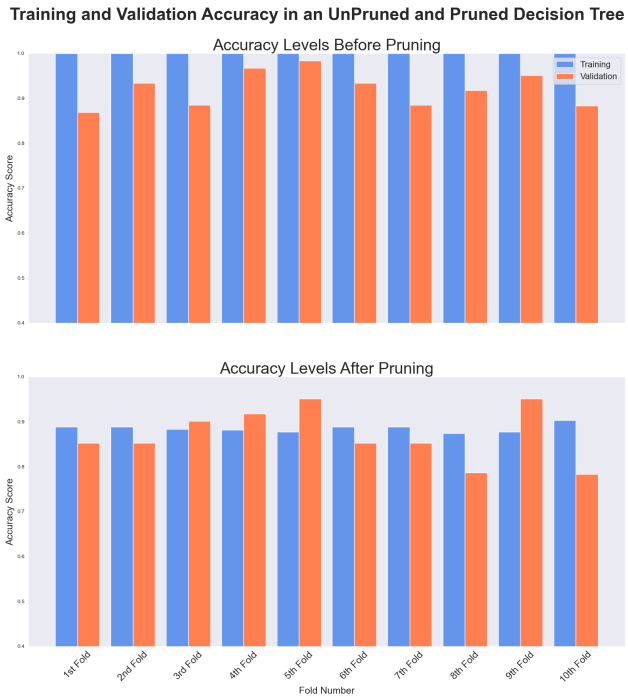

FIG. 4: Accuracy Levels Before and After Pruning Application.

Rather counter-intuitively, however, after pruning the mean validation accuracy and f1 scores

of the decision tree model noticeably reduced. That is, though the model was trained to be better applicable to generalised application, towards this dataset the accuracy was much reduced.

Model	Validation Accuracy Result	Validation F1 Score
Unpruned	91.78%	86.01%
Pruned	87.02%	75.05%

TABLE I: A table showing the overall accuracy and F1 scores for all three models.

Which model is preferable relies heavily upon its intended application. As this report is concerned with medical predictions, the primary concern lies in predictive power and as such will prefer the unpruned model.

2.2. Random Forests

Random Forests is an ensemble supervised learning method that trains using Bootstrap Aggregation to combine predictions from multiple machine learning algorithms. This 'bagging' method whereby Aggregation reduces random sample datasets from the Bootstrap into summary statistics drastically reduces the problem of over-fitting in what would be a high-variance model. This is because the Random Forest model combines multiple uncorrelated models, picks the classification with the most 'votes' and averages the output of all trees.

Optimising a Random Forests model relies heavily on the selection of hyperparameters before training to improve predictive power and speed. To choose the appropriate hyperparameters, I utilised 10-fold Cross Validation to train and test numerous model parameters, selecting from them the parameters which performed the best. This can be done as part of Scikit-Learn's Random Hyperparameter Grid. By combining the grid results with trial and error, the model used in this report is trained using 70 trees with a max depth of 5. Both the minimal number of samples required to split a node and be at leaf node were left at the skelearn default '2' and '1' respectively. Unlike with the decision tree model, the evaluation criteria is set to the Gini index due to its lesser computation cost with no noticeable decrease in performance.

These hyperparameters were chosen as the model seems to hit a point of diminishing returns above this point.

2.3. Sequence Classification

A sequential network traces data from input to output through a series of linear neural layers, one after the other. The Keras sequential model allows the user to create models layer-by-layer through the use of dense networks. As this dataset does not contain multiple inputs or outputs, this structured form of network building is ideal for a binary classification problems.

There are several parameters to tune for optimisation in a neural network. Similarly to the other models, the model below largely results from a combination of cross validation grid searching alongside consequent trial and error.

The final model used in this paper consists of three layers, with a ReLu activiaiton function in all but the output layer which utilises a 'Sigmoid' function. The mixture of ReLu and Sigmoid

allows for easier computation: ReLu does not activate every neuron in the network and rectifies the vanishing gradient problem. Whereas, because SIGMOID exists between (0,1), it acts as the ideal output for binary classification problems. By using a mixture of the two we therefore get the advantage of each.

The number of neurons at each layer is 45, apart from the last where it decreases to 1 to fit the output. To avoid over-fitting, there are also three dropout layers attached to the three layers set to (0.5). By randomly eliminating some inputs and hidden neurons in training, dropout reduces the computational cost whilst improving generalised application of the model.

Binary cross entropy is used as the loss function of this model, with the optimiser being set to 'adam'. Binary cross entropy compares each of the predicted probabilities to actual class output which can be either 0 or 1 and then penalises the probabilities based on their distance from the actual value. Due to its output class of 0 or 1, this loss function is ideal for a binary classification problem. Adam was chosen as the network optimiser due to its fast computation time and a low requirement of parameters for tuning.

FIG. 5: Training Curve of the Sequential Model By Epoch.

Fig. 4 shows that the dropout layers are working as they should to avoid over-fitting the model, though they are not perfect. There is, however, a steady decrease in loss as the model learns which is as expected, until it rounds off at around 80 epochs.

2.4. Model Comparison

Fig. 5 shows the Accuracy and F1 scores of all three models across the 10 Folds they were evaluated on. Overall, the models seem to have generally performed as would be expected with the sequential network performing the best, followed by the tree models. The closeness in performance between Random Forests and Decision Tree could be considered rather unexpected.

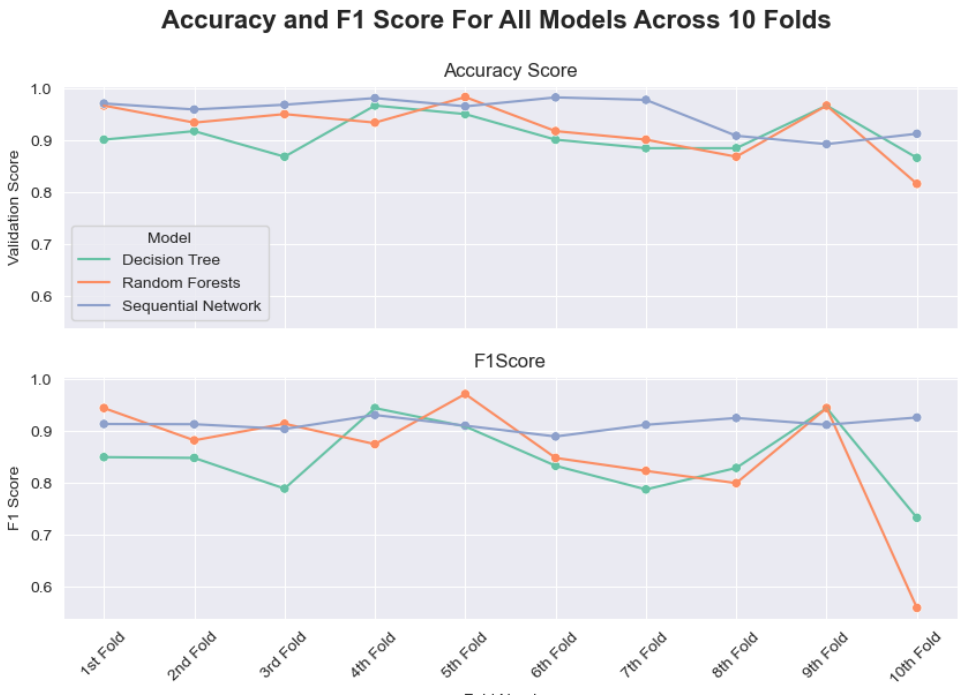

FIG. 6: Scatter-Line Comparison of All Model's Final Accuracy and F1 Score.

In the F1 plot, the Sequential Network immediately appears much more stable than the other networks. Once again, both Tree models appear rather interchangeable and erratic, and tend to follow the same pattern of peaks and valleys.

Both the Accuracy and F1 plot show a uniform decrease in predictive power across all three models in the later folds, apart from the Sequential Network in the latter graph. The final figures

Model	Validation Accuracy Result	Validation F1 Score
Decision Tree	91.78%	86.01%
Random Forests	92.43%	85.66%
Sequential Network	95.31%	91.12%

TABLE II: A table showing the overall accuracy and F1 scores for all three models.

in Table II Illustrate the superiority of the Sequential Network, with it clearly outperforming both tree models by a respectable margin of 2% - 5% . Sequential Networks therefore retain the highest score for Accuracy and F1 with 95.31% and 91.12% respectively.

On the other hand, the results of the tree models are largely interchangeable with no model clearly outperforming the other. In terms of Accuracy Random Forests outperformed the Decision Tree model by 0.65%, but in F1 this was reversed with the Decision Tree outperforming by 0.45%.

As expected, all models performed better overall on the Accuracy Score than on the F1 Statistic. As the F1 statistic takes into account the crucial nature of False Positives and False Negatives, it is generally considered the harsher metric.

3. DISCUSSION

As can be seen in the above section, the ASD Adult dataset was classified with an accuracy of 90% and an F1 statistic of 85% and above. For a first round of testing on a largely incomplete and un-processed dataset these results are promising for future machine clearing endeavours concerning an accessible ASD diagnosis, however, the immediate results are not fit to inform medical decisions.

Fig. 5 showed a general trend whereby accuracy and f1 scores dipped rather dramatically in the later evaluation folds. This may suggest that the training is unrepresentative of validation sets, and that the models are still, despite pruning and dropout methods, overfitting. In the future, stratified cross validation could be used in attempts to mitigate this drop-off. Stratified cross-validation would ensure that the items in each fold are sampled by considering their target labels. This means that both target labels would appear with the same relative frequency as in the larger dataset. The models might therefore be less sensitive to rare positive diagnosis cases.

Considering the comparative performance of the tree models, it might have been expected that Random Forests would outperform the Decision Tree and even come to rival the Sequential Network. However, it is not the case that ensemble methods are universally better than single models. With over 600 rows and numerous parameters, it might be the case that the Decision Tree model learned to a point of reasonable stability, and it is only in times of instability that ensemble methods thrive. That is, if the predictions of singular Decision Trees are stable, then all submodels in an ensemble model would return the same prediction. Random Forest would then simply make the same prediction of each singular tree. This could potentially describe why the models follow similar trends and perform largely the same with less than 1% differences between their overall scores.

Going forward, a larger dataset with less bias in the target class would drastically improve the performance and sensitivity of all three models. Both Random Forests and deep learning models tend to thrive with large quantities of data, and thus an increase in observations would reduce any problems of overfitting and reduce variance in performance rates across folds. This would largely increase F1 scores and improve upon accuracy simultaneously. General predictive power would therefore improve and move closer to a desired medical robustness of at least 99%.

One immediate way to increase the dataset size would be to re-introduce the Adult ASD dataset back into its original larger framework which also included data from children and adolescents. This would both enlarge the observation count, and provide another potentially influential feature ‘age description’ which this report removed. Isolated the feature provides no significant information, but set within a larger framework it has the potential to provide further insight to ASD diagnosis.

Finally, to further improve upon predictive power and robustness, one might adopt an approach using PCA to find the feature’s variance explanation. Using this information, it would be possible to exercise an effective feature selection; providing the means to act upon features that appeared to have little influence in the Data Visualisation section. This report found the point of diminishing returns for a Decision and Random Forests tree made up from all 42 features, but in reducing the number of features one could create entirely new models with reduced complexity and increased performance. If the dataset is enlarged, this would also increase training time.

4. EXECUTIVE SUMMARY

An estimated 2.21% of population of the United States are living with ASD [1]. A significant proportion of those adults did not receive their diagnosis until later in life and are thus dubbed "the lost generation". Generally, as we age, it is expected that we form a coherent identity and gain positive feelings about the self [5]. Autistic adults, however, as they age are forced to 'mask' their symptoms of ASD in order to camouflage themselves into what is considered 'normal'. For those adults who are forced to hide there is an increased risk of developing co-occuring mental conditions such as anxiety and depression as they age compared to the general population [2]. An earlier diagnosis has been suggested to be crucial to mitigating the negative effects of ageing with ASD.

The process to get a diagnosis as an adult is, however, incredibly long, frustrating, emotional, and sometimes very disappointing. This report is then dedicated to taking the preliminary steps to provide a quick, accessible path to diagnosis through the use of machine learning.

Using data from a survey concerning the prevalence of various ASD symptoms conducted on adults 18 and over, the algorithms created in this report tried to correctly assign each individual as either having ASD or not. How well each model performed was then based on their predictive performance - how well they accurately assigned each diagnosis(accuracy), as well as how often they misassigned one(F-statistic).

4.1. Results

Typically it takes 1 to 3 years on the UK's NHS waitlist to receive a letter for an autism screening alone. All three of the models presented in this report processed 609 observations and 42 columns of symptom information within minutes, returning a diagnosis prediction for each case. All three models scored above 90% in correctly predicting the right diagnosis, and above 85% when also taking into account the times they predicted wrong. One method - the Sequential Network - performed particularly well compared to others; scoring 95% in accuracy, and 91% as its F1 statistic.

Though not up to clinical standards by any means, the conclusions of this report highlight that it absolutely is possible for an automated method informed by various professionals to provide a fast and most importantly accessible ASD diagnosis system.

4.2. Going Forward

The Sequential Method shows that a quick, accessible and accurate ASD diagnosis is possible. However, the accuracy of that diagnosis relies upon access to large, detailed, and complete datasets on individuals with potential ASD, their symptoms, and their final diagnosis. Going forward, efforts should be focused upon curating such datasets and continuing to fit automated methods to new observations. Furthermore, building upon the success of the rather simple methods presented in this report, newer and more complex methods should be built to accommodate larger datasets.

REFERENCES

[1] Atherton, G., Edisbury, E., Piovesan, A. et al. *Autism Through the Ages: A Mixed Methods Approach to Understanding How Age and Age of Diagnosis Affect Quality of Life.* J Autism Dev Disord 52. 3639–3654 (2022).

[2] Cage, E., Di Monaco, J., Newell, V. *Experiences of autism acceptance and mental health in autistic adults.* Journal of Autism and Developmental Disorders, 48(2). 473–484 (2018).

[3] Dietz, P. M., Rose, C. E., McArthur, D., Maenner, M. *National and State Estimates of Adults with Autism Spectrum Disorder.*Journal of autism and developmental disorders, 50(12). 4258–4266 (2020). %book

[4] Dua, D. and Graff, C. *UCI Machine Learning Repository.* University of California, School of Information and Computer Science. (2019).

[5] Miner-Rubino, K., Winter, D. G., Stewart, A. J. *Gender, social class, and the subjective experience of aging: Self-perceived personality change from early adulthood to late midlife.* Personality and Social Psychology Bulletin, 30(12). 1599–1610 (2004).

[6] World Health Organization[WHO]. *Autism* (2022).

[7] Peral, J.; Gil, D.; Rotbei, S.; Amador, S.; Guerrero, M.; Moradi, H. *A Machine Learning and Integration Based Architecture for Cognitive Disorder Detection Used for Early Autism Screening.*Electronics,(9). 516 (2020).