Prediction of Autism Spectrum Disorder Using Screening Scores and Demographic Indicators

Şevval Günal
Middle East Technical University
Ankara, Turkey

Abstract— Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition that often goes underdiagnosed due to limited access to clinical specialists and variability in symptom presentation. This study investigates the use of data-driven predictive modeling to support early ASD screening. Responses from the AQ-10 screening tool and key demographic variables were used to develop and evaluate multiple classification models, including logistic regression, LASSO, Support Vector Machines (SVM), Artificial Neural Networks (ANN), Random Forests (RF), and XGBoost. Each stage, from imputation to model tuning, was approached with the goal of optimizing prediction while preserving interpretability.

Keywords — Autism Spectrum Disorder, AQ-10, Predictive Modeling, Screening Tools, Machine Learning, Classification, Feature Selection

I. Introduction

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by difficulties in social interaction, communication, and repetitive behaviors. Early and accurate detection is essential for timely intervention and improved outcomes. However, ASD remains underdiagnosed in many populations due to limited access to specialists, stigma, and variable symptom presentation. The use of digital screening tools, combined with predictive modeling, offers a viable path for scalable and accessible ASD detection.

This project aims to build a reliable classification model to predict ASD diagnosis using responses from the AQ-10 screening tool and demographic information. The analysis incorporates both traditional statistical methods and machine learning algorithms. Key preprocessing steps, such as handling missing data and recalculating AQ scores, ensured data quality. Model performances are compared, and feature importance is examined to gain further insight into the most predictive factors for ASD.

II. METHODOLOGY

A. Dataset

The dataset used in this study is sourced from Kaggle under the title Autism Prediction Dataset by Shivam Shinde. It contains responses from individuals who completed the AQ-10 screening questionnaire, along with additional demographic and contextual features. After initial cleaning and filtering, the dataset includes 800 observations.

The variable description is given below with the id being removed:

- Class/ASD (Target Variable) – Binary, 1 = Diagnosed with ASD, 0 = Not diagnosed
- A1_Score to A10_Score – Binary, Scores based on AQ-10 screening questions, 1 or 0
- correct_result – Discrete, total score computed from A1–A10 responses
- age – Continuous, in years (with invalid entries corrected or imputed)
- gender – Binary categorical, f = Female, m = Male
- ethnicity – Categorical, (e.g., White-European, Latino, Black, Others)
- country_of_res – Categorical, country of residence (e.g., United States, India, etc.)
- relation – Categorical, respondent's relation to the individual (e.g., Parent, Self, Others)
- jaundice – Binary, history of neonatal jaundice, yes & no
- austim – Binary, family history of autism, yes & no
- used_app_before – Binary, whether the patient has undergone a screening test before, yes & no
- age_desc – Categorical, description of age group (e.g., above 18)

Initial inspection revealed issues such as invalid entries, ambiguous missing values (e.g., "?"), and inconsistencies in calculated totals. These issues were addressed through specific imputation methods (e.g., PMM and polyreg) and AQ-score recalculations, described in the subsequent sections.

B. Descriptive / Summary Statistics

Before modeling, an initial analysis was conducted to summarize the structure and characteristics of the cleaned dataset, which includes 800 observations and a mix of numerical, binary, and categorical variables relevant to ASD diagnosis.

Variable	**Min**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**	**Missing**
age	18	23	30	34.29	42	89	210
correct_result	0	1	4	4.67	8	10	0

Table 1. Summary Statistics of Numerical Variables

Table 1 shows that the age variable ranges from 18 to 89 and is moderately right-skewed, with a concentration of participants in their 20s and 30s. Outliers (mostly 72+) were identified but retained due to clinical plausibility. Age also had the highest missingness in the dataset (26.3%). The correct_result variable, recalculated from AQ-10 items, ranges from 0 to 10 and is left-skewed, with most individuals scoring on the lower end, consistent with expectations for a general screening population.

Variable	Levels	Missing	ASD Diagnosis Rate (%)
gender (Biological sex of the participant)	m (530) / f (270)	0	20.4% (f) vs 20.0% (m)
jaundice (History of neonatal jaundice)	yes (185) / no (615)	0	30.3% (yes) vs 17.1% (no)
austim (Family history of autism)	yes (131) / no (669)	0	52.7% (yes) vs 13.8% (no)
used_app_before (Previously used the screening application)	yes (50) / no (750)	0	26.0% (yes) vs 19.7% (no)
Class/ASD (ASD diagnosis (1 = ASD, 0 = Non-ASD))	0 (639) / 1 (161)	0	79.9% (No ASD), 20.1% (ASD)

Table 2. Summary Statistics of Binary Variables

As we can see on Table 2, males were more frequent in the dataset, but ASD diagnosis rates were nearly equal between genders (~20%). Family history of autism (austim) showed the strongest association with ASD: 52.7% of those with a family history were diagnosed, compared to 13.8% without. A history of neonatal jaundice was also linked to a higher diagnosis rate (30.3% vs 17.1%). Conversely, prior app usage (used_app_before) was not associated with ASD status.

Variable	Type	Missing
ethnicity (Participant's ethnic background)	Categorical	203
country_of_res (Country of residence)	Categorical	0
age_desc (Age descriptor (e.g., above 18))	Categorical	0
relation (Relationship of respondent to the individual)	Categorical	40

Table 3. Summary Statistics of Categorical Variables

Table 3 demonstrates that ethnicity showed clear disparities in diagnosis rates. The White-European group had the highest ASD proportion (~47%), while other groups were significantly lower, often under 10%. These differences may reflect screening access rather than biological risk. Missing values in ethnicity (25%) and relation (5%) were handled using imputation methods discussed later in the report. Also, detailed ASD diagnosis proportions by ethnicity and relation are discussed in the EDA section. Due to large class imbalance and sparsity across levels, they were not included directly in the table.

C. Exploratory Data Analysis

To better understand the structure and distribution of the dataset, I conducted a comprehensive Exploratory Data Analysis (EDA), including both numerical summaries and visual techniques.

C.1 *How does age relate to AQ scores and ASD diagnosis, especially among older adults?*

The age distribution in the dataset is moderately right-skewed, with a concentration of participants in their 20s and 30s and a long tail into older age groups. A subgroup analysis of individuals aged 72 and above revealed slightly higher ASD diagnosis rates (29.2% vs 21.2%), but this difference was not statistically significant ($\chi^2 = 0.458$, $p = 0.4988$). However, these older individuals tended to have higher and more widely distributed AQ scores, with many scoring 8 or above.

Despite this pattern, statistical tests did not support age as a strong independent predictor of ASD diagnosis. Still, the observed variation may reflect meaningful behavioral patterns, such as delayed diagnosis, increased self-awareness, or selective participation by older individuals with pronounced traits. Therefore, while age contributes some contextual insights, it was not retained as a key variable in subsequent modeling.

C.2 Does the corrected AQ-10 total score (correct_result) significantly differ between ASD and non-ASD groups?

The correct_result variable, computed as the sum of responses to the ten AQ-10 screening items, exhibited a clear and statistically significant association with ASD diagnosis. A histogram of AQ scores revealed a left-skewed distribution, with most participants scoring between 0 and 5.

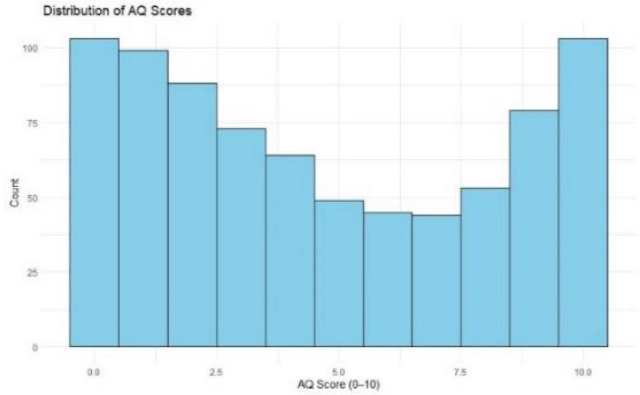

Figure 1. Distribution of AQ-10 Total Scores Among Participants

To examine group differences, a violin plot was created, combining the interpretability of a boxplot with a visualization of the full distribution shape. The plot showed a striking contrast between diagnostic groups: individuals with ASD were densely concentrated around scores of 8–10, while those without ASD showed a wider and flatter spread, typically between 0 and 6. The medians and interquartile ranges were clearly distinct, with minimal overlap.

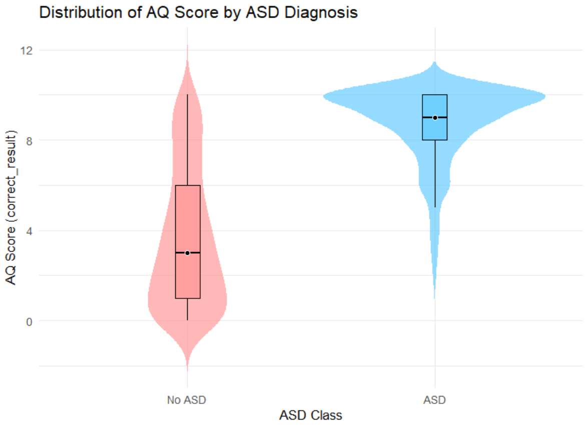

Figure 2. Violin Plot, Distribution of AQ-10 Scores by ASD Diagnosis

Given the non-normal distribution of AQ scores (confirmed via Shapiro-Wilk test, $p < 2.2e-16$ for both groups), both Welch's t-test and the Wilcoxon rank-sum test were applied. These tests yielded highly significant results ($p < 2.2e-16$), reinforcing the visual interpretation.

These findings strongly support the AQ-10's intended purpose as a screening instrument, where higher scores indicate increased likelihood of ASD. Statistically and visually, correct_result emerged as the most informative predictor in the dataset.

C.3 Does gender influence ASD diagnosis rates?

Although males were more frequent in the dataset, ASD diagnosis rates were nearly identical across genders, with both male and female participants showing a diagnosis rate of approximately 20%. A chi-squared test confirmed that this difference was not statistically significant ($p = 0.878$).

Interestingly, this finding contrasts with clinical literature, which typically reports higher ASD prevalence among males. The lack of disparity here may be due to characteristics of the sample, such as voluntary self-reporting, or the possibility that females with ASD in this dataset are more likely to seek screening than in the general population. While gender remains an important contextual variable, it does not appear to influence ASD prediction in this dataset.

C.4 Are any demographic or medical and behavioral history variables (e.g., austim, ethnicity, jaundice, used_app_before) significantly associated with ASD diagnosis?

The austim variable, indicating whether the individual has a family history of autism, showed one of the strongest associations with ASD diagnosis. Among those without a family history, 13.8% were diagnosed with ASD, compared to 52.7% of those who did report a family history. A chi-squared test confirmed the strength of this association ($\chi^2 = 100.82$, $p<2.2e-16$).

This is clearly illustrated in the side-by-side proportion bar chart, which shows that ASD diagnoses are substantially more frequent among those reporting a family history. Clinically, this aligns with genetic research and suggests heightened awareness or screening intent in these individuals.

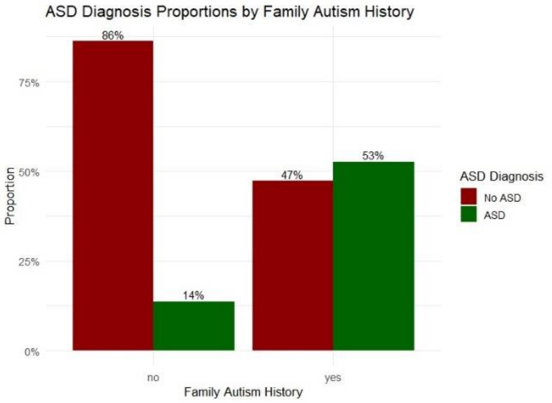

Figure 3. Bar Plot, ASD Diagnosis Proportions by Family History of Autism

Ethnicity also demonstrated a statistically significant association with ASD diagnosis ($\chi^2 = 109.84$, $df = 6$, $p < 2.2e-16$). Diagnosis rates varied widely across groups, with White-European individuals showing the highest diagnosis rates (~47%), while other groups, including Asian, Black, and Middle Eastern participants, had lower observed rates (often under 15%).

To improve readability across multiple ethnicity categories, a horizontal bar chart was used. This visualization makes it clear that most ASD diagnoses occurred among the White-European group, while minority groups showed comparatively fewer diagnoses. However, these patterns should be interpreted with caution, as they may reflect diagnostic access disparities, cultural stigma, or sample selection bias rather than true underlying prevalence.

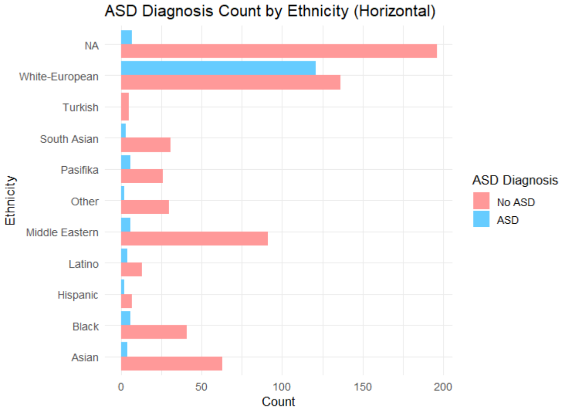

Figure 4. Horizontal Bar Plot, ASD Diagnosis Count by Ethnicity

Participants with a history of neonatal jaundice had significantly higher rates of ASD diagnosis than those without (30.3% vs. 17.1%). This association was confirmed statistically ($\chi^2 = 14.60$, $p = 0.00013$). A side-by-side bar chart with percentage labels highlights this relationship, showing a visibly higher proportion of ASD cases among those who experienced jaundice at birth. While the exact mechanisms are still unclear, the pattern supports ongoing hypotheses about the relationship between early-life physiological stress and neurodevelopment.

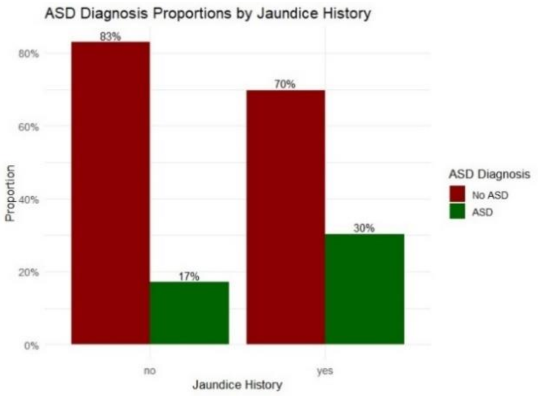

Figure 5. Bar Plot, ASD Diagnosis Proportions by Jaundice History

In contrast, the variable used_app_before, which indicates whether a participant had previously completed the screening, did not show a statistically significant relationship with ASD diagnosis ($p = 0.375$). But when we look at the proportions, the patients who has undergone a screening test before were slightly more diagnosed with ASD (26.0% yes vs 19.7% no). This small increase may be explained by self-selection bias; individuals who already suspect ASD may be more likely to retake the test, but it does not appear to reflect a strong or reliable diagnostic pattern.

D. Missingness & Imputation

Two key variables contained missing data: ethnicity (203 entries, 25.3%), and relation (40 entries, 5%). These were initially masked using "?" and only identified after data inspection.

Additionally, the result column, intended to reflect total AQ-10 scores, contained numerous invalid values (e.g., negatives, non-integers), affecting nearly half the entries. Since the individual AQ items (A1_Score to A10_Score) were complete and valid, I created a new variable, correct_result, by summing those binary responses.

For age, some entries also contradicted the dataset's metadata (e.g., values below 18 in an adult sample). All values were rounded and any entries under 18 were flagged as missing (210 entries, 26.3%).

To visualize the overall extent of missingness, a bar plot was created showing the number of missing entries per variable. This highlighted age, ethnicity, and relation as the only variables with substantial missing data.

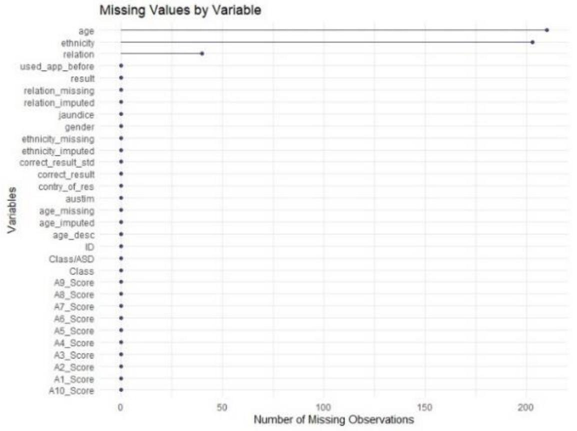

Figure 6. Number of Missing Values Plot

Next, a missingness heatmap was used to inspect patterns across observations. This revealed that age missingness was scattered throughout the dataset, while ethnicity and relation appeared to follow more structured, possibly non-random patterns.

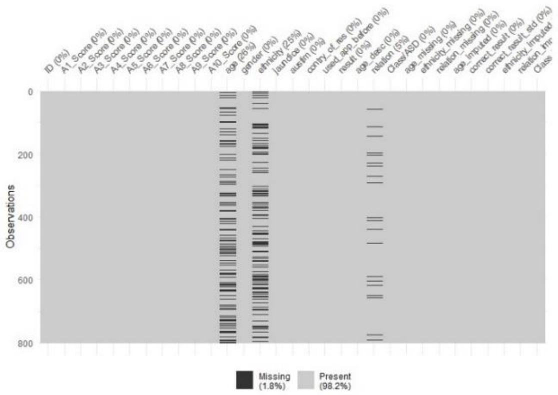

Figure 7. Missingness Heatmap

To test the assumption of random vs. non-random missingness, I examined whether missingness in each variable was related to ASD diagnosis using chi-squared tests:

- Age: Missingness was not significantly associated with diagnosis (χ^2 $p = 0.120$), suggesting a Missing Completely at Random (MCAR) mechanism.
- Ethnicity: Missing values were significantly more common among non-ASD individuals (χ^2 $p \approx 1.39e-11$).
- Relation: Similarly, missing values were more frequent in non-ASD cases (χ^2 $p = 0.008$).

Since complete-case analysis would likely introduce bias due to the non-random patterns in ethnicity and relation, multiple imputation was applied using the mice package:

- Predictive Mean Matching (PMM) was used for age to preserve distributional realism and avoid outliers.
- Polytomous logistic regression (polyreg) was used for ethnicity and relation, respecting their predictive relationships with diagnosis.

These imputations were validated using trace and distributional diagnostics, which confirmed the stability of imputed values across chains and their consistency with observed data. While these plots are not shown here for conciseness, no substantial distortions were observed.

E. Modelling

i. Cross Validation:

To evaluate how well each model would generalize to unseen data, I used 5-fold cross-validation. This method offers a good balance between computational efficiency and reliable performance estimates, especially for a dataset of moderate size like this ($n = 800$). Unlike leave-one-out CV, which can be overly sensitive to individual data points and slow to run, 5-fold CV reduces variance without being too computationally intense. It also avoids the instability of a single train-test split, making it a practical and robust choice for this project.

ii. Logistic Regression

The first model I built was a logistic regression classifier using all available predictors. Performance was promising out of the gate, with an AUC of 0.811 and notably high sensitivity (0.892), though specificity lagged behind at 0.584. This kind of trade-off isn't unexpected in early screening scenarios, where missing true ASD cases can be far more costly than occasional false alarms.

However, as I examined the model diagnostics, several issues came to light. Multicollinearity was a major concern, particularly with the relation variable, which had a staggering variance inflation factor (VIF > 10^8). Additionally, categorical variables like ethnicity and country_of_res contained many rare or sparsely populated levels, which increased noise and complicated convergence. To resolve this, I removed the problematic relation variable and grouped low-frequency levels in ethnicity and country_of_res under an "Other" category.

After simplifying the model, I re-estimated the 5-fold cross-validation. The performance improved substantially:

- AUC increased from 0.811 to 0.905
- Sensitivity improved from 0.892 to 0.914
- Specificity rose from 0.584 to 0.634

Multicollinearity also dropped to acceptable levels (most VIFs < 3), suggesting a cleaner, more stable model. These improvements not only boosted discriminative performance but also maintained high sensitivity, ideal for ASD screening tools, where early detection is key.

The final model included only the most meaningful predictors:

$$\begin{aligned}\text{logit(ASD)} = \; & -6.74 \\ & + 0.58 \times \text{Correct\_AQ\_Score} \\ & + 1.56 \times \text{Country\_United\_States} \\ & + 1.66 \times \text{Country\_Canada} \\ & + 1.37 \times \text{Country\_Australia} \\ & + 1.04 \times \text{Country\_United\_Kingdom} \\ & + 0.88 \times \text{Country\_Other} \\ & + 0.31 \times \text{Used\_App\_Before} \\ & + 0.18 \times \text{Family\_History\_of\_Autism}\end{aligned}$$

[Only predictors with $p < 0.1$ are shown. Coefficients are on the log-odds scale.]

Visual diagnostics supported these improvements:

- The cross-validated ROC curve showed strong separability between classes.
- A variable importance plot confirmed that correct_result (AQ-10 score) was by far the most predictive feature.
- Country of residence also played a moderate role, possibly reflecting differences in healthcare access, awareness, or diagnostic pathways across regions.

Altogether, this refined logistic model became more interpretable and reliable, striking a useful balance between performance and simplicity, exactly what's needed in scalable ASD screening systems.

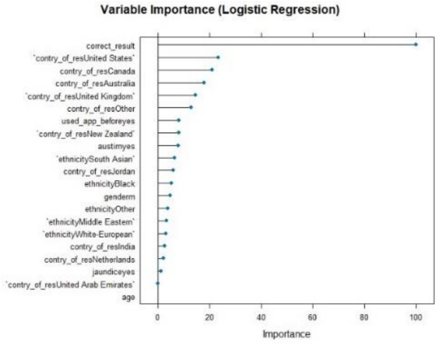

Figure 8. Variable Importance Plot

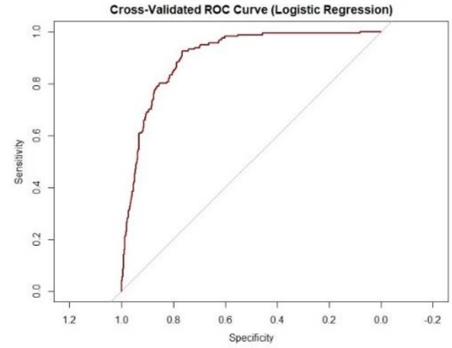

Figure 9. Cross-Validated ROC Curve

iii. LASSO Regularization

As a regularized alternative, I implemented a LASSO logistic regression model, which applies an L1 penalty to shrink less informative coefficients toward zero. This approach is especially useful when working with a high number of categorical variables, many of which in this dataset had sparse or unbalanced levels. By simplifying the model, LASSO helps reduce overfitting while enhancing interpretability.

I tuned the regularization parameter ($\lambda = 0.0103$) using 5-fold cross-validation, selecting the value that maximized the area under the ROC curve.

The final model achieved strong overall performance:

- AUC: 0.915
- Accuracy: 0.8725
- Sensitivity: 0.6522
- Specificity: 0.9280
- Kappa: 0.594
- Precision (PPV): 0.6954
- Balanced Accuracy: 0.7901

These results are comparable to the refined logistic regression model, but with the added benefit of automated variable selection. Several weak predictors were excluded entirely (their coefficients reduced to zero), which aligns with earlier EDA findings and further supports their limited value.

LASSO Model (Log-Odds of ASD Diagnosis):

$$\begin{aligned}
\text{Logit(ASD)} = \; & -2.49 \\
& + 1.85 \times \text{Correct\_AQ\_Score} \\
& + 0.06 \times \text{Age} \\
& + 0.12 \times \text{Ethnicity\_WhiteEuropean} \\
& + 0.06 \times \text{Family\_History\_of\_Autism} \\
& + 0.00 \times \text{Country\_Australia} \\
& + 0.07 \times \text{Country\_Canada} \\
& - 0.10 \times \text{Country\_India} \\
& - 0.05 \times \text{Country\_Netherlands} \\
& - 0.06 \times \text{Country\_NewZealand} \\
& + 0.01 \times \text{Country\_UnitedKingdom} \\
& + 0.24 \times \text{Country\_UnitedStates}
\end{aligned}$$

[Only non-zero coefficients are shown. Coefficients are on the log-odds scale.]

The sparse structure makes the model easier to interpret. As expected, the AQ-10 score (correct_result) and family history of autism emerged as the strongest predictors. Country-specific effects were also visible, especially for the United States and Canada, possibly reflecting systemic or cultural differences in diagnosis. On the other hand, variables like gender and app usage were excluded entirely, reinforcing the earlier observation that they lacked strong independent predictive value.

iv. Support Vector Machine

To explore a nonlinear classification alternative, a Support Vector Machine (SVM) with a radial basis function (RBF) kernel was employed. SVMs are well-suited for classification tasks involving complex, high-dimensional relationships, and the RBF kernel is particularly effective at capturing nonlinear decision boundaries without extensive feature engineering.

Model tuning was performed using the caret package, which optimized the cost parameter (C) while holding the kernel's sigma fixed at 0.032. Among the tested values of C, the model achieved its best performance at C = 1, balancing sensitivity and specificity effectively. This configuration used 305 support vectors, indicating a moderately complex decision boundary.

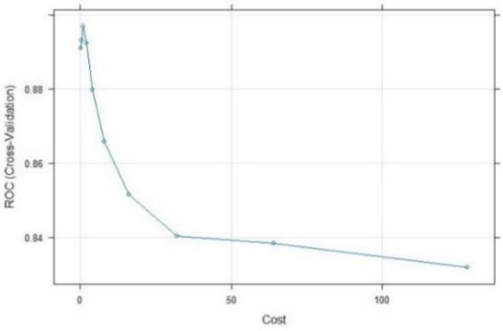

Figure 10. SVM Plot

The cross-validated results for the best SVM model (C = 1, σ = 0.032) are summarized below:

Metric	Value
AUC	0.897
Accuracy	0.863
Sensitivity (Recall)	0.516
Specificity	0.950
Balanced Accuracy	0.733
Kappa	0.521
Precision (PPV)	0.722

Table 4. Performance Metrics of SVM

These metrics paint a clear picture: this SVM is excellent at identifying non-ASD cases, but less reliable in detecting true positives. Its high specificity and precision make it a confident classifier when it predicts the absence of ASD, but with a relatively low sensitivity, it's prone to missing actual ASD cases. That makes it better suited to confirming a diagnosis than for broad screening.

Another limitation is interpretability. Unlike logistic regression, SVMs operate as black-box models, making it difficult to explain individual predictions. Still, the high precision and specificity indicate strong reliability when the model does predict ASD absence. For settings where explanatory power is crucial (e.g., clinical use), this trade-off must be carefully considered.

v. Artificial Neural Network:

Among all models tested, the Artificial Neural Network (ANN) offered the most flexibility in how it learned from the data. Its layered structure allowed it to capture complex, nonlinear relationships without relying on predefined feature interactions or decision splits. This flexibility made it especially well-suited to uncovering subtle patterns, something particularly valuable in the context of ASD, where symptom presentation can be highly varied.

To train the model, I used 5-fold cross-validation with ROC as the optimization metric. I tuned two key hyperparameters:

- Hidden units: 3, 5, 7
- Weight decay: 0.01, 0.1, 0.5

The best performance (ROC ≈ 0.899) was achieved with 3 hidden units and decay = 0.5, as shown in the figure below.

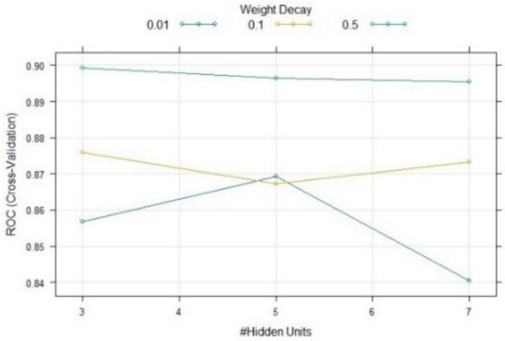

Figure 11. ANN Plot

Metric	**Value**
Accuracy	0.856
AUC (Cross-Validated)	0.899
Sensitivity (Recall)	0.627
Specificity	0.914
Precision (PPV)	0.647
Balanced Accuracy	0.771
Kappa	0.548

Table 5. Performance Metrics of ANN

The ANN model demonstrated solid classification performance, especially in terms of specificity (91.4%), indicating strong ability to correctly identify non-ASD cases. Sensitivity was moderate (62.7%), suggesting a balanced but cautious approach in detecting true ASD cases. The relatively high AUC (~0.899) confirms that the model effectively distinguishes between classes across different thresholds. Additionally, the tuning results indicate that stronger regularization (decay = 0.5) improved generalization performance, particularly when paired with a smaller network (3 hidden units), helping to avoid overfitting.

vi. Random Forest:

The Random Forest (RF) model brought stability and simplicity to the mix. As a tree-based ensemble, it handles both categorical and numerical features well, and it's especially effective in reducing overfitting through its built-in randomness.

During tuning, the best results were achieved when the number of variables randomly selected at each split (mtry) was set to 2. Increasing mtry beyond this point led to a slow decline in performance, suggesting that smaller, more diverse trees generalized better in this context.

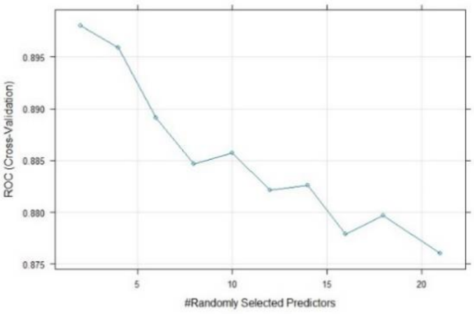

Figure 12. Random Forest Plot

Metric	**Value**
Accuracy	0.859
Sensitivity (Recall)	0.491
Specificity	0.951
Precision (PPV)	0.718
Balanced Accuracy	0.721
AUC (ROC)	0.898
Kappa	0.502

Table 6. Performance Metrics of Random Forest

Despite its reliability, RF showed the lowest sensitivity of any model tested, meaning it missed more actual ASD cases than others. This likely reflects its averaging nature: while it suppresses noise, it can also mute subtle but important patterns, especially in imbalanced or heterogeneous data.

In short, Random Forest offered a hands-off modeling experience with decent overall performance, but lacked the flexibility or sharpness needed for tasks where detecting true positives is crucial. It's a great starting point for baseline comparisons, but not the top choice for ASD screening on its own.

vii. XGBoost:

XGBoost was the final ensemble model explored, and one of the strongest performers overall. As a gradient boosting framework, it builds decision trees in sequence, with each new tree correcting the errors of the last. This allows it to capture subtle patterns and complex nonlinear relationships, while regularization keeps it from overfitting.

In tuning, a surprisingly effective combination emerged: a higher learning rate (eta = 0.1) paired with very shallow trees (max_depth = 2) and just 50 rounds of boosting. Simpler configurations like this tend to train faster and generalize better, especially when the data has a few dominant predictors.

Cross-validated performance metrics:

Metric	**Value**
Accuracy	0.866
Sensitivity (Recall)	0.621
Specificity	0.928
Precision (PPV)	0.685
Balanced Accuracy	0.775
Kappa	0.569
AUC (Cross-validated)	0.913

Table 7. Performance Metrics of XGBoost

The XGBoost model delivered the highest AUC (0.913) among all tested algorithms, demonstrating excellent overall discriminatory power. It maintained strong specificity (0.928), indicating reliable identification of non-ASD cases, while achieving moderate sensitivity (0.621) in detecting ASD. Notably, this high performance was achieved using a very simple configuration (shallow trees with a depth of just 2) which suggests that even relatively simple decision boundaries were sufficient to capture meaningful patterns in the data. Overall, XGBoost offered a compelling balance between efficiency, flexibility, and performance, making it a strong candidate for scalable deployment in ASD screening contexts.

To assess potential overfitting, the final XGBoost model was also evaluated on the training data. The table below summarizes the performance metrics for both the training and cross-validated results:

Metric	**Training Set**	**Cross-Validated**
AUC	0.931	0.913
Accuracy	0.880	0.866

Sensitivity (Recall)	0.634	0.621
Specificity	0.942	0.928
Precision (PPV)	0.734	0.685
Balanced Accuracy	0.788	0.775

Table 8. Training and CV Performance Metrics of XGBoost

The training AUC (0.931) and accuracy (88%) are only slightly higher than the cross-validated results, indicating that the model generalizes well and is not substantially overfitting. The consistency across sensitivity and specificity further supports this conclusion.

To better understand which variables contributed most to the model's decisions, feature importance was calculated using the gain metric from the final model. The figure below illustrates the ten most influential predictors.

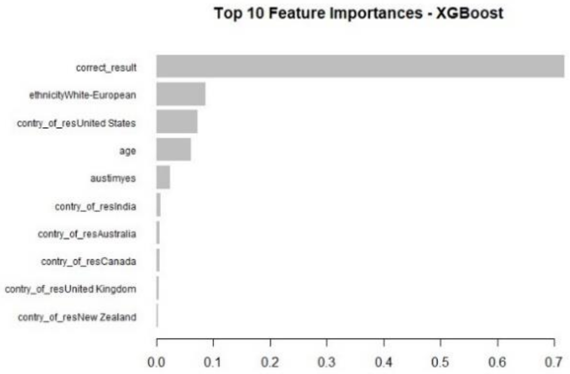

Figure 13. Top 10 Feature Importances from XGBoost

As expected, the AQ-10 total score (correct_result) emerged as the most powerful predictor, by a wide margin. This aligns with the tool's design, as it was specifically developed to screen for behavioral traits linked to ASD.

Several demographic variables also contributed meaningfully to the model's predictions, particularly ethnicity (White-European) and country of residence, with individuals from the United States showing higher predictive weight. These effects likely reflect sociocultural differences, such as access to diagnosis or awareness, rather than inherent biological variation.

Interestingly, age also had a moderate impact, which may be related to differences in diagnosis timing or cohort-specific traits. While family history of autism (austimyes) was a strong predictor in earlier models, it ranked lower in the XGBoost model, possibly because its signal was absorbed by overlapping variables. Still, its appearance among the top 10 features reinforces its relevance in screening decisions.

Overall, this ranking of features aligns well with the findings from the exploratory data analysis, confirming that the AQ-10 score remains the most consistent and powerful driver in predicting ASD outcomes.

F. Model Comparison

To evaluate generalizability, all models were trained using 5-fold cross-validation, and their predictive performances were compared using AUC, accuracy, sensitivity, specificity, precision, balanced accuracy, and Kappa:

Metric	Logistic Reg.	LASSO	SVM
AUC	0.905	0.915	0.897
Accuracy	0.874	0.873	0.863
Sensitivity	0.696	0.652	0.516
Specificity	0.919	0.928	0.950
Precision (PPV)	0.683	0.695	0.722
Balanced Accuracy	0.807	0.790	0.733
Kappa	0.610	0.594	0.521

Table 9. Performance Comparison Table

Metric	ANN	Random Forest	XGBoost
AUC	0.899	0.898	0.913
Accuracy	0.856	0.859	0.866
Sensitivity	0.627	0.491	0.621
Specificity	0.914	0.951	0.928
Precision (PPV)	0.647	0.718	0.685
Balanced Accuracy	0.771	0.721	0.775
Kappa	0.548	0.502	0.569

Table 10. Performance Comparison Table

[Note: All metrics are based on 5-fold cross-validation.]

Each model presents a different trade-off, reflecting the priorities and constraints of the screening context.

Logistic Regression proved to be a strong and interpretable baseline. Its high specificity (0.919) makes it good at ruling out non-ASD cases, though its sensitivity (0.696) is more modest; acceptable for screening, but with room to improve.

LASSO Logistic Regression achieved the highest AUC (0.915) and shared the top specificity (0.928). By shrinking weak predictors, it achieved a leaner, more generalizable model. Although its sensitivity dropped slightly (0.652), the trade-off may be worth it in real-world deployment where model simplicity is also valued.

Support Vector Machine (SVM) offered the highest precision (0.722) and specificity (0.950), making it the most conservative classifier. However, its low sensitivity (0.516) means it missed many ASD cases, making it less suited for broad screening but potentially helpful in confirmatory stages.

Artificial Neural Network (ANN) performed reliably, with an AUC of 0.899 and balanced accuracy of 0.771. It detected ASD cases more effectively than SVM and Random Forest, while maintaining strong specificity (0.914). Although it didn't outperform LASSO or XGBoost, its ability to model complex patterns without manual tuning is a notable strength, though its limited interpretability may be a drawback in clinical use.

Random Forest showed the clearest preference for avoiding false positives, with the highest specificity (0.951) but the lowest sensitivity (0.491). While accurate overall, this imbalance could make it less suitable for early screening, where catching potential ASD cases is critical.

XGBoost stood out as the most well-rounded model. It combined strong AUC (0.913), balanced sensitivity (0.621) and specificity (0.928), and solid precision. It slightly trailed LASSO in precision and interpretability but made up for it in flexibility and pattern recognition, especially given its shallow decision trees and low overfitting risk.

Ultimately, model selection depends on screening priorities. If interpretability and ease of communication are top concerns, LASSO is an ideal choice. If the goal is to maximize predictive performance in a complex real-world context, XGBoost emerges as the most promising option.

III. Conclusion

This study developed and compared multiple statistical and machine learning approaches to predict Autism Spectrum Disorder (ASD) using AQ-10 screening scores and demographic indicators. Among them, LASSO regression stood out for its interpretability and lean design, filtering out noise while retaining predictive power. XGBoost, on the other hand, delivered the most balanced and accurate performance, especially in capturing nonlinear patterns, making it a compelling candidate for scalable screening tools.

Across all models, the AQ-10 score and family history of autism consistently emerged as the strongest predictors, underscoring their central role in early ASD detection. While logistic regression offered transparency and ease of explanation, ensemble models like XGBoost and Random Forest showed strength in modeling complex decision boundaries. Depending on the use case, each model offers unique benefits: LASSO is ideal for interpretable deployment, ANN for uncovering subtle relationships in exploratory analysis, and Random Forest for conservative second opinions in clinical pipelines.

Looking ahead, evaluating these models in real-world settings (across diverse populations and age groups) will be crucial. But performance alone isn't enough. To truly support equitable and ethical diagnosis, future tools must prioritize transparency, trust, and inclusiveness alongside accuracy. To ensure accessibility across populations, the final model could be deployed as a mobile-first screening assistant, particularly in under-resourced regions where access to diagnostic services is limited. Models that not only work well but work responsibly will shape the next generation of ASD screening systems.

References

[1] S. Shinde, "Autism Prediction Dataset", Kaggle, [Online]. Available: https://www.kaggle.com/datasets/shivamshinde123/autismprediction

[2] Taylor, E. C., Livingston, L. A., Clutterbuck, R. A., & Shah, P. (2020). Psychometric concerns with the 10-item Autism-Spectrum Quotient (AQ10) as a measure of trait autism in the general population. Experimental Results, 1, e3. doi:10.1017/exp.2019.3

[3] Booth, T., Murray, A.L., McKenzie, K. et al. Brief Report: An Evaluation of the AQ-10 as a Brief Screening Instrument for ASD in Adults. J Autism Dev Disord 43, 2997–3000 (2013). https://doi.org/10.1007/s10803-013-1844-5