**Aim of the Project:**

The aim of this project is to develop a statistical model to predict the likelihood of Autism Spectrum Disorder (ASD) diagnosis using AQ-10 screening scores and selected demographic features. Using a publicly available dataset, the analysis focuses on identifying key predictors and assessing model performance through robust statistical techniques.

**Introduction:**

The dataset used in this project is publicly available on Kaggle under the title “Autism Prediction Dataset”. It consists of 800 observations retained from the training set after initial cleaning.

The dependent variable is Class/ASD, a binary categorical variable indicating whether the individual has been diagnosed with Autism Spectrum Disorder (1) or not (0). The dataset includes screening-related variables such as binary responses to the AQ-10 questionnaire (A1\_Score to A10\_Score) and their total score (result). It also contains demographic features like age, gender, ethnicity, relation to the respondent, country of residence, family history of autism, prior app usage, and age categorization.

**Data Cleaning and Tidying:**

Although the dataset did not contain explicit NA values at first glance, inspection revealed that the ethnicity and relation columns contained "?" entries, which were treated as missing. These were later addressed during the imputation phase.

The result column, intended to reflect the total AQ-10 screening score (0–10), contained numerous invalid entries, including negative and non-integer values. Approximately 49% of these values were unusable. Fortunately, the dataset included responses to all ten AQ questions (A1\_Score to A10\_Score), allowing me to reconstruct the total score manually. I created a new correct\_result variable by summing these binary scores, which proved to be more reliable and aligned with expected values.

The age column presented similar issues. Around 26% of the entries were invalid or conflicted with the age\_desc column. Several values indicated ages below 18, despite metadata suggesting that all participants were above 18. To address this, all age entries are rounded to whole numbers and flagged below 18 as missing.

To explore the nature of missingness, I used missingness plots, including heatmaps and bar plots.

A diagram of a dna sequence

AI-generated content may be incorrect.A graph with numbers and lines

AI-generated content may be incorrect.

These revealed that while missing age values were scattered throughout the dataset, the missingness in ethnicity and relation showed structured patterns, suggesting the possibility of non-random mechanisms.

Lastly, since the test data lacked the Class label, it could not be used for model validation metrics like accuracy or AUC. Therefore, I restricted all analyses to the training data only.

**Exploratory Data Analysis (EDA):**

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.

The age variable is continuous, ranging from 18 to 89 years, with a mean of 34.29 and a median of 30. The distribution is moderately right-skewed, as confirmed by visualizations such as histograms and boxplots, which showed a concentration of participants in their 20s and 30s, and a long right tail extending into older age groups. A total of 210 values (26.3%) were missing, making age one of the most incomplete variables in the dataset.

A graph of a number of blue bars

AI-generated content may be incorrect.

The boxplot of age clearly highlights several upper-end outliers. These outliers appear well beyond the upper whisker, suggesting a distinct subgroup of older individuals in the dataset.

A graph with a bar graph

AI-generated content may be incorrect.

To assess these outliers more rigorously, I applied both interquartile range (IQR) and z-score methods. The IQR approach identified 24 outlier observations, all aged 72 or older, while the z-score method flagged 6 of these as extreme. These values were retained after review, as they were deemed clinically plausible and consistent with dataset metadata. Their inclusion was important given the potential relevance of later-life diagnoses and symptom recognition in ASD.

I then conducted a subgroup analysis comparing these older individuals to the rest of the dataset. While ASD diagnosis rates were slightly higher in the 72+ group (29.2% vs 21.2%), the difference was not statistically significant (χ² p = 0.4988). However, the difference in AQ scores was more pronounced: the older subgroup had a higher median AQ score (5.5 vs 4.0) and a wider distribution, with many individuals scoring 8 or above. This pattern may reflect increased self-awareness in older adults, delayed diagnosis, or more intentional participation by individuals with more pronounced traits.

**A graph showing different colored rectangles

AI-generated content may be incorrect.A graph showing a number of patients

AI-generated content may be incorrect.**

This is clearly shown in the density plot grouped by diagnosis, where non-ASD individuals cluster around the early 20s, while ASD-diagnosed participants display a broader and flatter distribution across the 25–50 range, including a pronounced right tail. The boxplot of age further supports this, showing a slightly higher median and broader spread for the ASD group compared to non-ASD individuals.

In summary, the age variable displays right-skewness, notable missingness, and meaningful subgroup variation in AQ score severity. Notably, age alone is not a strong predictor of ASD diagnosis in this dataset, but it appears linked to how symptoms are reported or recognized.

The correct\_result variable was constructed by summing binary responses to the ten AQ-10 screening items (A1\_Score to A10\_Score), as the original result column contained numerous invalid or inconsistent values. This recalculated score ranges from 0 to 10, with a mean of 4.67. The distribution is left-skewed, as most individuals in the dataset scored on the lower end of the spectrum. This pattern is consistent with expectations for a general screening population, where the majority are not expected to meet ASD criteria.

A graph of a bar graph

AI-generated content may be incorrect.A diagram of a diagram showing different colored squares

AI-generated content may be incorrect.

Visualizations, including a histogram and boxplot, confirmed the skewed distribution and a concentration of scores below 5. In contrast, individuals diagnosed with ASD typically had much higher scores. To investigate this relationship formally, I compared correct\_result values between ASD and non-ASD groups. The boxplot comparison revealed a clear separation: ASD cases were tightly clustered between 7 and 10, while the non-ASD group showed broader variability, generally scoring between 0 and 6.

To statistically confirm this observation, I first tested for normality using the Shapiro-Wilk test, which returned highly significant p-values (p < 2.2e-16) for both groups, indicating that AQ scores are not normally distributed. Consequently, I used both Welch’s t-test and the Wilcoxon rank-sum test to assess group differences. Both tests yielded highly significant results (p < 2.2e-16), confirming that individuals with ASD had significantly higher AQ scores than those without a diagnosis. This finding aligns with the intended use of the AQ-10 as a screening tool: high scores are strongly associated with ASD diagnosis. The strength and consistency of this association, both visually and statistically, establish correct\_result as the most informative predictor in the dataset.

The dataset includes both male and female participants, with a greater number of males overall. While more ASD diagnoses were observed among males in absolute terms, this reflected the gender imbalance in the dataset rather than a higher diagnostic rate. When the proportion of ASD cases was calculated within each gender, both groups showed nearly identical rates of diagnosis, approximately 22%.

**A graph showing a number of people with different colored squares

AI-generated content may be incorrect.A graph with a bar and a person's hand

AI-generated content may be incorrect.**

To test whether gender was associated with ASD diagnosis, I conducted a chi-squared test. The result was not statistically significant, indicating no meaningful relationship between gender and diagnosis outcome in this sample. This is also evident in the proportion bar plot, where both groups show similar rates of ASD diagnosis.

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.

The variable austim, indicating whether the individual has a family history of autism, showed one of the strongest associations with ASD diagnosis in the dataset. Among those without a reported family history, 13.8% were diagnosed with ASD; among those with a family history, the rate jumped to 52.7%. This large difference was confirmed by a chi-squared test (χ² = 100.82, p < 2.2e-16), providing strong statistical evidence of an association.

Visualizations supported these findings: count plots showed that while most participants reported no family history, ASD cases were disproportionately concentrated among those who did. Proportion plots offered an even clearer view, revealing that over half of the individuals with a reported family history received an ASD diagnosis, compared to a much lower rate in the group without such a history.

**A graph with a diagram

AI-generated content may be incorrect.A graph showing a number of autism history

AI-generated content may be incorrect.**

Clinically, this result aligns well with existing research suggesting a genetic or hereditary component in the development of autism. Individuals with a family history of autism may be more likely to seek screening and recognize symptoms in themselves. Overall, the analysis confirms that a reported family history of autism is a strong and statistically significant predictor of ASD diagnosis in this dataset.

Ethnicity emerged as another significant predictor of ASD diagnosis. Diagnosis rates varied widely across groups, with the White-European group showing the highest proportion (approximately 47%), and other groups, such as Pasifika, Black, Asian, and Middle Eastern participants, exhibiting much lower rates, some below 10%. A chi-squared test confirmed that this variation was statistically significant (χ² = 109.84, df = 6, p < 2.2e-16).

A graph of different colored bars

AI-generated content may be incorrect.A graph of a disease

AI-generated content may be incorrect.

These differences were clearly visible in both count and proportion plots, where the White-European group dominated in both absolute and relative ASD diagnoses. While these findings may suggest a meaningful association, they likely reflect a combination of cultural, diagnostic, and access-related factors. For example, ethnic minority groups may be underrepresented in clinical diagnoses due to lower screening rates, limited access to healthcare, or stigma around neurodevelopmental conditions. Thus, while ethnicity appears predictive in this dataset, it should be interpreted with caution and considered alongside systemic factors that shape health data.

Among the participants, those with a history of neonatal jaundice were significantly more likely to be diagnosed with ASD (30.3%) compared to those without (17.1%). A chi-squared test confirmed the association (χ² = 14.60, p = 0.00013), and this pattern was clearly illustrated in the proportion plot (Figure X). This finding may support existing hypotheses that early-life physiological stressors, such as jaundice, could be linked to neurodevelopmental outcomes like ASD, although the causal mechanisms remain under investigation.

A graph with red and blue squares

AI-generated content may be incorrect.A graph of a bar chart

AI-generated content may be incorrect.

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). 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.

I explored whether missingness in key variables was associated with ASD diagnosis, focusing on age, ethnicity, and relation. For age, the proportion of missing values did not differ significantly between ASD and non-ASD participants (χ² p = 0.120), suggesting that missingness is independent of the outcome and thus consistent with a Missing Completely at Random (MCAR) mechanism. This supported the use of imputation methods that assume randomness in the missing pattern.

**A graph of a bar chart

AI-generated content may be incorrect.**

In contrast, both ethnicity and relation displayed structured missingness patterns. A significantly larger portion of missing values in these variables was found in individuals without an ASD diagnosis. This pattern was visually evident in bar plots and confirmed by chi-squared tests, which returned highly significant p-values (ethnicity: p ≈ 1.39e-11, relation: p = 0.008). These results indicate that the missingness in ethnicity and relation is Missing Not at Random (MNAR) and likely correlated with the diagnosis outcome.

**A graph showing a missing person

AI-generated content may be incorrect.A graph of missing and missing

AI-generated content may be incorrect.**

The interpretation is critical for modeling: analyses using only complete cases could introduce bias due to the non-random nature of missingness in these predictors. Therefore, imputation was necessary. These findings also may reflect behavioral or psychological factors, such as discomfort disclosing personal information or disengagement from the survey, especially among individuals not seeking a clinical diagnosis.

Based on the identified missingness mechanisms and variable types, I used Predictive Mean Matching (PMM) to impute missing values in the age column. PMM is well-suited for continuous variables with real-world meaning, as it preserves the distribution of observed data and avoids introducing implausible outliers. For the categorical variables ethnicity and relation, which were identified as Missing Not at Random (MNAR), I used polytomous logistic regression (polyreg) imputation to respect their predictive relationships with the outcome.

A graph showing the age and the age of a person

AI-generated content may be incorrect.A diagram showing the difference between ethnicity and relation

AI-generated content may be incorrect.

A green lines with dots

AI-generated content may be incorrect.A graph showing the number of the same number

AI-generated content may be incorrect.A graph showing the number of imputations

AI-generated content may be incorrect.

To validate the imputations, I examined trace plots and strip plots across multiple imputation chains. These confirmed the stability of the imputed values and their similarity to the observed data. In particular, the distribution of imputed categories closely mirrored the original proportions, and no unusual shifts or distortions were observed. The dominance of the "Self" category in relation was maintained after imputation, suggesting that the modeling process respected the underlying structure of the data.

Together, these results indicate that the selected imputation strategies were appropriate and yielded realistic, unbiased values for use in later analysis.

**Cross Validation:**

To estimate predictive performance, I used 5-fold cross-validation. This approach balances computational efficiency and generalizability, especially with a moderate dataset size (n = 800). Compared to leave-one-out CV, it avoids overfitting and excessive variance, while offering more stability than a single train-test split.

The initial model was a logistic regression classifier using all available predictors. It performed reasonably well, achieving an AUC of 0.811, with high sensitivity (0.892) but lower specificity (0.584). While not perfectly balanced, this trade-off aligns with screening contexts where missing a true case is more problematic than a false alarm.

Multicollinearity checks revealed severe issues, especially in the relation variable, which exhibited extremely high variance inflation (VIF > 10⁸). Additionally, many categorical variables (like ethnicity and country\_of\_res) had sparse or rare levels. To address this, I dropped the relation and grouped infrequent levels in ethnicity and country\_of\_res, reducing noise and improving convergence.

Following these adjustments, the logistic regression model was re-evaluated using the same 5-fold cross-validation setup. Performance improved significantly: the AUC rose to 0.905, sensitivity reached 0.914, specificity increased to 0.634, and AIC dropped from 535.66 to 482.8. This indicated a more balanced and better-fitted model. Key predictors such as correct\_result (AQ score) remained highly significant, while others like country\_of\_residenceUnited States and Canada gained slight predictive relevance. Multicollinearity was also reduced substantially, with most VIFs falling below 3 except for a few borderline cases.

To complement the quantitative evaluation of model refinement, I compared model performance before and after data cleaning and simplification. The AUC increased from 0.81 to 0.91, and specificity improved from 58% to 63%, indicating a better-calibrated and more generalizable model. These improvements validate key modeling decisions, such as dropping variables with extreme multicollinearity and collapsing sparse categorical levels.

Visual diagnostics reinforced these findings. The cross-validated ROC curve clearly demonstrates the model’s excellent classification ability, with an AUC close to 0.91. The curve lies well above the 45-degree diagonal of a random classifier, illustrating a strong trade-off between sensitivity and specificity. This suggests that the model maintains reliable discriminative performance across folds, even under the variation introduced by resampling.

A variable importance plot was also generated to visualize each predictor’s relative contribution. As expected, correct\_result was by far the most influential variable, consistent with earlier EDA and clinical expectations. Several levels of country\_of\_residence, particularly the United States and Canada, also showed moderate influence, which may reflect regional differences in diagnostic access, cultural reporting norms, or dataset composition. In contrast, predictors such as age, jaundice, and gender contributed minimally to the model, again aligning with previous statistical tests that showed weak or non-significant associations.

A graph with a graph on it

AI-generated content may be incorrect.A graph showing a curve

AI-generated content may be incorrect.

**LASSO Regularization:**

To explore an alternative, regularized approach, I implemented LASSO logistic regression. LASSO applies an L1 penalty during model fitting, shrinking less informative coefficients toward zero and effectively performing variable selection. The optimal value of the regularization parameter (λ) was selected through cross-validation.

The final LASSO model retained a reduced subset of predictors: correct\_result (AQ score), austimyes (family history), and several country-of-residence indicators, such as the United States and Canada. All other variables (including age, gender, jaundice, and used\_app\_before) were shrunk to zero, reinforcing earlier statistical findings that these factors had limited predictive value.

Finally, I compared logistic regression and LASSO models under different classification thresholds. In terms of performance, the LASSO model achieved an AUC of approximately 0.915, slightly exceeding the refined logistic regression model. I also explored threshold tuning by lowering the classification cutoff from 0.5 to 0.4. This increased sensitivity (recall of ASD cases) from 0.696 to 0.752, though it reduced precision and specificity slightly. This trade-off can be valuable in clinical screening settings where minimizing false negatives is a priority.

| **Metric** | **LASSO (0.4 Threshold)** | **Logistic Regression (0.5)** |
| --- | --- | --- |
| Accuracy | 0.859 | 0.874 |
| Sensitivity | 0.752 | 0.696 |
| Specificity | 0.886 | 0.919 |
| Precision (PPV) | 0.624 | 0.683 |
| Balanced Accuracy | 0.819 | 0.807 |
| Kappa | ~0.592 | 0.610 |

While both models performed well, they serve different purposes. The logistic regression model offers better overall precision and interpretability, making it suitable for confirmatory analysis or diagnostic support. In contrast, the LASSO model, with higher sensitivity and automated variable selection, is well-suited for screening contexts where early detection is prioritized over precision.

These results provide a strong statistical foundation for exploring machine learning models in the next stage of the project.

**Şevval Günal 2429082**