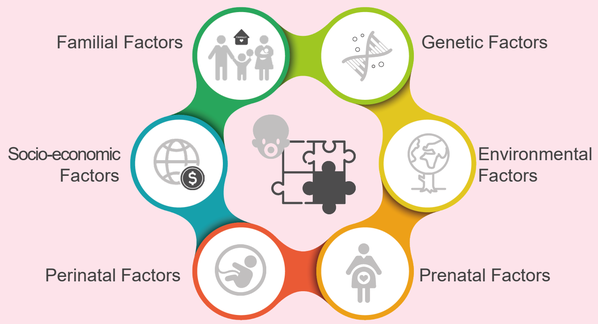
# Predictive Modeling of Autism Spectrum Disorder : Socioeconomic, Prenatal, and Environmental Influences

**Seyoung Park**
*The Webb Schools*

## INTRODUCTION

- Autism Spectrum Disorder (ASD) is a neurodevelopmental a condition characterized by difficulties in social communication interaction, and repetitive behaviors (CDC, 2022). ASD results from a complex interplay of genetic, environmental, pre natal, socioeconomic, and familial factors, yet these relationships are not fully understood (Bai et al., 2019; Rossignol & Genuis, 2014)
- This study applies machine learning techniques to data from the National Survey of Children's Health (NSCH) to identify critical ASD risk factors, aiming to enhance early detection and guide effective interventions.

**[ Understanding Autism Spectrum Disorder ]**




## METHOD

- **Data Source**
  Utilized data from the **National Survey of Children's Health** (NSCH), which provides comprehensive information on socio -economic status, prenatal/perinatal conditions, environmental exposures, family history, and child health outcomes (Child and Adolescent Health Measurement Initiative, 2023).

## METHOD

### Data Analysis and Model Evaluation Process


**Data Preprocessing**
- Missing values handled using imputation methods(mice, miss Forest in R).
- Categorical variables converted into numerical format via dummy encoding (caret package).
- Dataset split into training (70%) and testing sets (30%) to validate model accuracy.

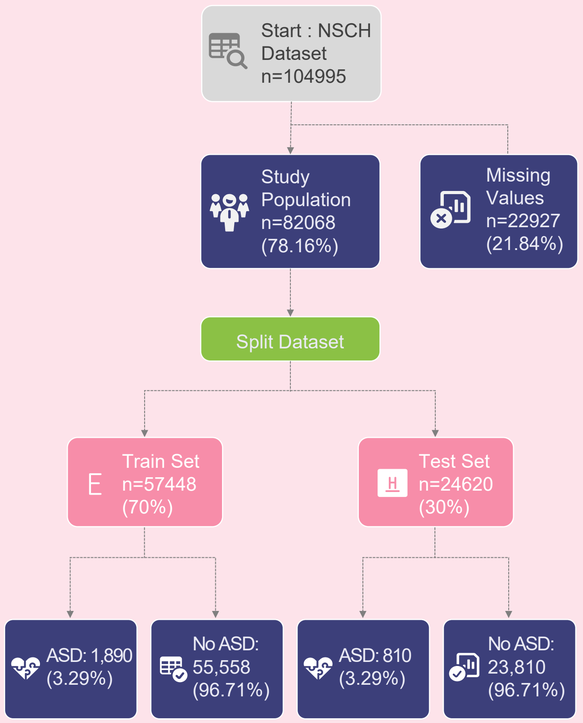
**Exploratory Data**
- Conducted descriptive statistics and visualizations to explore data characteristics and variable associations.
- Statistical tests (Chi-square, t-tests) were performed to identify significant relationships.

**Feature Selection**
- Recursive Feature Elimination (RFE) and Lasso regression applied to determine essential predictive variables..

**Machine Learning Models Evaluated**
- Logistic Regression
- Random Forest
- Support Vector Machine (SVM)
- Gradient Boosting Machines (GBM)
- Neural Networks

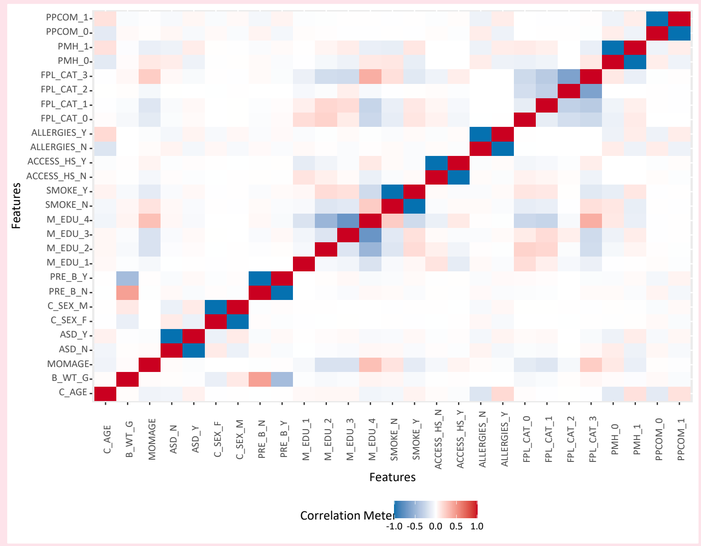
**Model Evaluation**
- Applied cross-validation and evaluated model performance using *Accuracy, Precision, Recall, F1-score, and AUC-ROC.*

### Data Processing Flowchart for NSCH Study




### Table 1. Baseline Characteristics of Study Population (ASD vs. Non-ASD)

| Variable | Levels | no ASD | ASD | p-value | Standardized Mean Difference (SMD) |
|---|---|---|---|---|---|
| n | | 101009 | 3446 | | |
| Age | | 8.33 ± 5.30 | 9.98 ± 4.66 | <0.001 | 0.330 |
| Sex (%) | Female | 49555 (49.1) | 799 (23.2) | <0.001 | 0.559 |
| | Male | 51454 (50.9) | 2647 (76.8) | | |
| Birth Weight (grams) | | 3307.04 ± 544.08 | 3270.76 ± 599.67 | 0.001 | 0.063 |
| Prematurity (%) | No | 89251 (89.3) | 2847 (83.6) | <0.001 | 0.167 |
| | Yes | 10724 (10.7) | 560 (16.4) | | |
| Ever Breast-fed (%) | No | 6015 (15.3) | 236 (28.6) | <0.001 | 0.325 |
| | Yes | 33197 (84.7) | 588 (71.4) | | |
| Parental Age at Birth (years) | | 30.38 ± 5.69 | 29.90 ± 6.07 | <0.001 | 0.083 |
| Maternal Education Level (%) | Less than high school | 2663 ( 2.6) | 91 ( 2.6) | <0.001 | 0.2 |
| | High school (including vocational, trade, or business school) | 13035 (12.9) | 554 (16.1) | | |
| | Some college or Associate Degree | 21081 (20.9) | 942 (27.3) | | |
| | College degree or higher | 64230 (63.6) | 1859 (53.9) | | |
| Smoke (%) | No | 87079 (88.3) | 2815 (83.1) | <0.001 | 0.150 |
| | Yes | 11503 (11.7) | 573 (16.9) | | |
| Access to Healthcare Services (%) | No | 4325 ( 4.3) | 81 ( 2.4) | <0.001 | 0.108 |
| | Yes | 96209 (95.7) | 3349 (97.6) | | |
| Allergies history (%) | No | 74364 (73.8) | 2112 (61.4) | <0.001 | 0.267 |
| | Yes | 26437 (26.2) | 1329 (38.6) | | |
| Family Poverty Ratio (%) | 0-99% | 10259 (10.2) | 497 (14.4) | <0.001 | 0.236 |
| | 100-199% | 16650 (16.5) | 768 (22.3) | | |
| | 200-399% | 36685 (36.3) | 1196 (34.7) | | |
| | ≥400% | 37415 (37.0) | 985 (28.6) | | |
| Family History of Mental Disorders (%) | No | 75107 (89.4) | 2114 (74.2) | <0.001 | 0.402 |
| | Yes | 8909 (10.6) | 735 (25.8) | | |
| Prenatal/Perinatal Complications (%) | No | 93478 (92.5) | 2978 (86.4) | <0.001 | 0.201 |
| | Yes | 7531 ( 7.5) | 468 (13.6) | | |

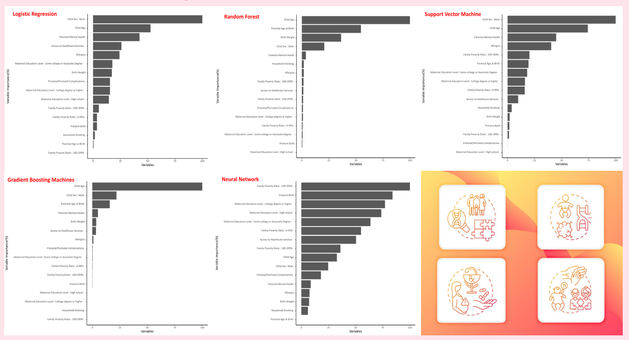



### Figure 1. Correlation Heatmap of Variables Associated with Autism Spectrum Disorder (ASD)

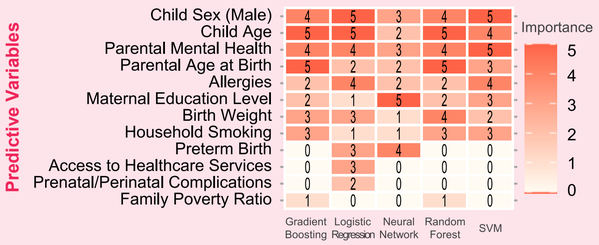
- This correlation matrix shows that ASD risk is influenced by multiple interrelated socio-economic, familial, prenatal/perinatal, and demographic factors, highlighting the complexity and multifactorial nature of ASD etiology.

| Variable | Description | Variable | Description |
|---|---|---|---|
| ASD | Autism ASD | M_EDU | Highest Level of Education among Reported Adults |
| C_AGE | Age of Selected Child - In Years | SMOKE | Anyone in Household Use Cigarettes |
| C_SEX | Sex of Selected Child | ACCESS_HS | Health Insurance Coverage - Currently Covered |
| B_WT_G | Standardized Birth Weight, Grams | ALLERGIES | Allergies |
| PRE_B | Born 3 or More Weeks Before Due Date | FPL_CAT | Family Poverty Ratio |
| EBF | Ever Breast-fed | PMH | Parental Mental Health |
| MOMAGE | Age of Mother - Years | PPCOM | Prenatal/Perinatal Complications |

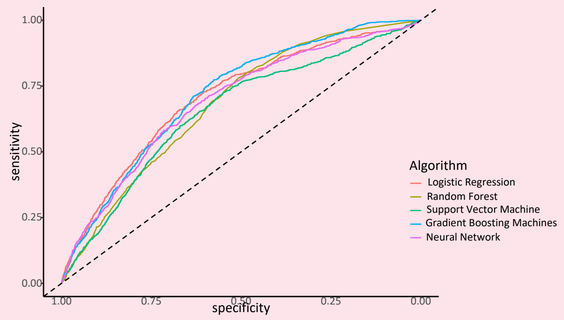
### Figure 2. Variable Importance Across Machine Learning Models Predicting Autism Spectrum Disorder (ASD)




### Figure 3. Heatmap of Variable Importance Scores Across Five Machine Learning Models Predicting Autism Spectrum Disorder (ASD)



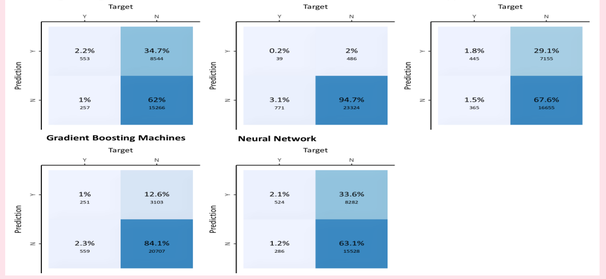

### Figure 4. ROC Curves Comparing Machine Learning Models for Predicting Autism Spectrum Disorder (ASD)




### Table 2. Performance Metrics of Machine Learning Models for Predicting Autism Spectrum Disorder (ASD)

| Machine Learning Model | Accuracy | Cohen's Kappa | ROC (Receiver Operating Characteristic) | Sensitivity | Specificity | Area Under Curve (AUC) | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|---|---|---|
| Logistic Regression | 0.643 | 0.055 | 0.705 | 0.641 | 0.683 | 0.983 | 0.983 | 0.641 | 0.776 |
| Random Forest | 0.949 | 0.033 | 0.680 | 0.980 | 0.048 | 0.773 | 0.968 | 0.980 | 0.974 |
| Support Vector Machine | 0.695 | 0.049 | 0.654 | 0.699 | 0.549 | 0.979 | 0.979 | 0.699 | 0.816 |
| Gradient Boosting Machines | 0.851 | 0.071 | 0.715 | 0.870 | 0.310 | 0.986 | 0.974 | 0.870 | 0.919 |
| Neural Network | 0.652 | 0.052 | 0.690 | 0.652 | 0.647 | 0.982 | 0.982 | 0.652 | 0.784 |

### Figure 5. Confusion Matrix Heatmaps for Autism Spectrum Disorder (ASD) Prediction by Machine Learning Models




## CONCLUSION

This study identified critical factors influencing the occurrence of Autism Spectrum Disorder (ASD) among children using machine learning techniques.

Key predictive variables included child sex (male), child age, parental mental health, birth weight, preterm birth, parental age at birth, and family socio-economic status.

Among the five machine learning models evaluated, Gradient Boosting Machines and Random Forest demonstrated superior predictive performance.

These findings provide valuable insights that can inform early detection, targeted interventions, and preventive strategies for ASD.

## REFERENCES


- Bai, D., Yip, B. H. K., Windham, G. C., Sourander, A., Francis, R., Yoffe, R., ... Sandin, S. (2019). Association of genetic and environmental factors with autism in a 5-country cohort. JAMA Psychiatry, 76(10), 1035–1043. https://doi.org/10.1001/jamapsychiatry.2019.1411
- Centers for Disease Control and Prevention (CDC). (2022). Autism Spectrum Disorder: Risk Factors and Characteristics. Retrieved from https://www.cdc.gov/ncbddd/autism/facts.html
- Child and Adolescent Health Measurement Initiative (CAHMI). (2023). National Survey of Children's Health (NSCH). Data Resource Center for Child and Adolescent Health. Retrieved from https://www.childhealthdata.org/learn-about-the-nsch/NSCH
- Gardener, H., Spiegelman, D., & Buka, S. L. (2011). Perinatal and neonatal risk factors for autism: A comprehensive meta-analysis. Pediatrics, 128(2), 344–355. https://doi.org/10.1542/peds.2010-1036
- Rossignol, D. A., & Frye, R. E. (2014). Environmental toxicants and autism spectrum disorders: A systematic review. Translational Psychiatry, 4(2), e360. https://doi.org/10.1038/tp.2014.4