**Note on Variable Selection for Autism Prediction Model**

For my autism prediction model using the 2023 National Survey of Children’s Health (NSCH) dataset (autism.csv), I selected K2Q35A (Ever Diagnosed with Autism) as the target variable and 34 predictors to capture health, physical, socioeconomic, demographic, and other relevant factors, including screen time, language spoken, and race. Below is the rationale for these choices:

* **Target Variable: K2Q35A**

I chose K2Q35A (binary: 1=Yes, 2=No, encoded as 1=Yes, 0=No) because it inclusively captures children ever diagnosed with autism or autism spectrum disorder (ASD), including those with past or current diagnoses. This broad scope is ideal for early autism prediction, as it encompasses cases with autism traits, even if not currently diagnosed, aligning with my goal of identifying at-risk children for early intervention. Unlike AutismInd\_23 (current diagnosis only), K2Q35A provides a more comprehensive target by including all diagnosed cases, ensuring relevance for both clinical and predictive purposes. Its binary nature simplifies classification modeling.

* **Predictor Variables (34 Total)**

From the 895 features in autism.csv, I selected 34 predictors to balance comprehensive coverage of relevant factors with model efficiency and user-friendliness for a family-facing web app. The predictors were chosen based on autism research, emphasizing health, physical, socioeconomic, demographic, and behavioral/environmental factors, with specific attention to language spoken, race, and screen time, as these are linked to autism risk and diagnostic disparities. The predictors are grouped as follows:

* 1. **Demographic (6)**: SC\_AGE\_YEARS, SC\_SEX, HHLANGUAGE, SC\_RACE\_R, SC\_HISPANIC\_R, BORNUSA. These capture age, sex, household language, race, ethnicity, and U.S. birth status. Notably, HHLANGUAGE addresses higher autism risk in non-English-speaking families due to diagnostic access barriers, and SC\_RACE\_R allows exploration of racial disparities in diagnosis rates.
  2. **Socioeconomic (6)**: FPL\_I1, FAMCOUNT, ACE2more\_23, AdultEduc\_23, HOUSINGINSTAB\_23, FoodSit\_23. These reflect poverty, family size, adverse childhood experiences, parental education, housing instability, and food insecurity, which influence healthcare access and developmental stress.
  3. **Health (8)**: K2Q31A (ADHD), K2Q33A (anxiety), K2Q36A (developmental delay), K6Q40 (breastfeeding), ALLERGIES, K2Q34A (seizures), HEART, ASTHMA\_23. These include comorbidities and chronic conditions strongly associated with autism risk.
  4. **Physical (3)**: K2Q05 (prematurity), BIRTHWT\_L (low birth weight), BMICLASS (BMI). These physical factors are linked to developmental challenges.
  5. **Behavioral/Developmental (9)**: K6Q13A (speech concerns), K6Q13B (interaction concerns), K6Q14A (words/phrases concerns), K6Q14B (behavior concerns), ONEWORD, TWOWORDS, ASKQUESTION, SCREENTIME, HRSLEEP. These capture early developmental signs (e.g., speech, social delays) and behavioral factors like screen time and sleep, which are linked to autism-like behaviors.
  6. **Parental/Environmental (3)**: A1\_MENTHEALTH (maternal mental health), K10Q31 (neighborhood support), ParAggrav\_23 (parental aggravation), FamResilience\_23 (family resilience). These reflect parental and environmental influences on child development and diagnosis.
* **Why a Subset of 34 Predictors?**

With 895 features, including all would lead to overfitting, redundancy (e.g., FPL\_I1 to FPL\_I6), and impracticality for a web app. I prioritized predictors with evidence-based links to autism (e.g., speech delays, prematurity, socioeconomic stressors) and ensured they are intuitive for families to input (e.g., SCREENTIME, HHLANGUAGE). Variables like DENTALSERV1 or VAPE were excluded due to weak autism relevance. Special values (e.g., 99, 95, 90) were cleaned to ensure data quality, as seen in the dataset overview.

This selection ensures a robust, interpretable model that addresses key autism risk factors while supporting a practical web app for families, aligning with my project’s focus on early autism detection using socioeconomic, behavioral, and demographic data.