Final Project Proposal- Autism Spectrum Disorder (ASD) Screening

Dataset Overview - https://archive.ics.uci.edu/dataset/426/autism+screening+adult

- **Dataset Focus:** Used for screening individuals for Autism Spectrum Disorder (ASD) based on demographic details, family history, and responses to a screening questionnaire.
- **Data Type:** Multivariate, categorical, continuous, and binary attributes.
- **Task:** Classification (predicting ASD likelihood).
- **Number of Records:** 704 instances.
- **Number of Features:** 21 attributes.
- **Key Features:**
    - Demographics: Age, Gender, Ethnicity, Country of Residence.
    - Medical & Family History: Born with Jaundice, Family Member with PDD (Pervasive Developmental Disorder).
    - Screening Information:
        - Who is completing the test (e.g., self, parent, medical staff).
        - Prior use of screening apps.
        - Screening Method Type (categorized for toddlers, children, adolescents, and adults).
        - Responses to 10 ASD-related questions (binary: 0 or 1).
        - Final screening score (computed automatically).

Machine Learning approach:

- **Loading data:** *Autism_Data.arff* – For this .arff format, I'd use the pandas library in Python to read the data efficiently. Exploratory data analysis (EDA) visualizes feature distributions and correlations to identify potential issues like class imbalance
- **Pre-processing:**. Missing values are imputed, categorical features (ethnicity, country) are one-hot encoded, and numerical features (age, screening score) are scaled or normalized for consistency.
- **Feature Extraction:** Interaction terms and aggregated features are generated, with PCA explored for dimensionality reduction.
- **Model Training:** In the model training phase, the primary goal is to build a predictive model that accurately classifies whether an individual is likely to have Autism Spectrum Disorder (ASD) based on the provided features. By training various machine learning models like XGBoost, AdaBoost, SVMs(various kernels), Naive Bayes, and neural networks, we aim to learn the complex relationships between medical history, and questionnaire responses and their association with the likelihood of ASD. The trained models will enable automated screening and risk assessment, potentially facilitating earlier diagnoses and interventions.
- **Evaluate Performance (CV, Accuracy, AUC, others related):** Cross-validation ensures robustness, with accuracy, AUC-ROC, and F1-score used for evaluation.

References:

1. Mashudi, N. A., et al. (2021). Classification of Adult Autistic Spectrum Disorder Using Machine Learning Approach. *International Journal of Artificial Intelligence*, PDF Link.
2. Exploration of Autism Spectrum Disorder using Classification Algorithms .Author links open overlay panelB Deepa, K.S Jeen Marseline, Link.
3. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , Link.*