Assignment 1: Part A – Question Formation and Exploratory Analysis  
Course: 4536\_COMP\_SCI\_7209 (Trimester 2, 2025)  
Student Name: Sadman Sharif

1. Introduction and Societal Relevance

1.1 Problem Statement

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition with wide variability in manifestation, particularly in adult populations. While diagnostic awareness in children has improved globally, adults, especially those with high-functioning autism—remain significantly underdiagnosed due to the subtler expression of symptoms and a lack of standardized screening mechanisms tailored to adult neurodiversity. This diagnostic gap contributes to delayed interventions, social isolation, underemployment, and elevated mental health risks. Addressing this issue is of critical societal relevance, necessitating scalable and data-driven tools to support early identification and personalized care [1, 2].

1.2 Societal Significance

This project investigates the use of machine learning (ML) to predict ASD in adults based on structured questionnaire data and demographic variables. The end goal is to design a lightweight, interpretable, and scalable screening model that can be embedded into mobile applications or online platforms. Such innovations could help identify high-risk individuals, reduce diagnostic delays, and support the goals of proactive and equitable healthcare aligned with the WHO’s public health objectives [2–4].

2. Big Data Characteristics and Dataset Choice

2.1 Dataset Overview

* Dataset Name: Autism Prediction Dataset
* Source: Kaggle (Shinde, 2022) [1]
* Files Included: train.csv (2,800 records), test.csv (923 records)
* Total Size: 3,723 records × 20 features
* Target Variable: Binary classification (ASD / Not ASD)

The dataset includes 10 binary responses to screening questions (A1–A10), and demographic features such as age, gender, ethnicity, country of residence, family ASD history, and auxiliary flags (e.g., used\_app\_before, jundice). It provides a well-structured input for a supervised classification pipeline with opportunities for subgroup and feature interaction analysis.

2.2 Suitability in Big Data Context

While not at petabyte scale, this dataset exemplifies several Big Data characteristics:

* Volume: Nearly 4,000 instances—a sufficient size for meaningful analysis and future scaling.
* Variety: Mix of categorical, binary, and continuous attributes.
* Velocity: Potential use in real-time screening apps.
* Veracity: Includes self-reporting bias, mitigated through preprocessing [3, 4].

This structured format and the potential to augment unstructured data (e.g., Reddit discussions, ABIDE neuroimaging) make the dataset an excellent starting point for a predictive healthcare task.

2.3 Supplementary Datasets for Expansion

* CDC Autism Surveillance Data [10]
* ABIDE: Autism Brain Imaging Data Exchange [11]
* Reddit (Pushshift API): For sentiment analysis and language markers in ASD communities

3. Industry Context and Value Proposition

3.1 Industry Application

This work aligns with growing demand in mobile health (mHealth), digital diagnostics, and telemedicine. Embedding ASD screening into digital tools could streamline triage workflows in mental health clinics and human resources departments, where early detection facilitates intervention and support.

3.2 Public Health Relevance

National surveys show a rise in adult ASD referrals and diagnoses, but delays persist. Tools developed in this project could help clinicians prioritize those most in need of full evaluations, especially given the high burden of undiagnosed ASD among women and minorities [10].

3.3 Cross-Sector Use Cases

* Healthcare: Pre-diagnostic screening in general practice
* Education: Student support services
* Employment: HR-based neurodiversity assessments

3.4 Ethical and Legal Considerations

Privacy is paramount in mental health AI. The project will adhere to GDPR and the Australian Privacy Act [12]. Explainability and fairness audits will ensure ethical deployment across diverse populations.

4. Data Processing and Risk Mitigation

4.1 Preprocessing Pipeline

1. Categorical Encoding: Categorical variables such as gender and ethnicity must be numerically encoded for compatibility with most ML algorithms, which cannot natively process non-numeric values. Binary categorical variables (like gender) are label encoded for simplicity, while multi-class features (e.g., ethnicity, relation) are one-hot encoded to prevent the model from falsely assuming ordinal relationships between categories. This approach helps the model treat each category as distinct and non-hierarchical, reducing the risk of introducing bias from incorrect assumptions.
2. Scaling: Continuous features, especially age, vary in scale compared to binary and categorical inputs. Using MinMax scaling transforms such values to a standard [0,1] range, ensuring that algorithms sensitive to magnitude differences (e.g., SVMs with RBF kernels or logistic regression) do not disproportionately weigh certain features. This also improves convergence during training and facilitates better distance-based decision boundaries in models like k-NN or SVM.
3. A screen shot of a computer

   AI-generated content may be incorrect.Imputation: Missing data, if ignored, can lead to reduced training samples or biased learning. Instead of dropping rows with null values—which could compromise the sample size and representativeness—we use Miss Forest, a robust imputation method that preserves relationships between features by training decision trees iteratively. This is particularly suitable when both categorical and continuous features are present and when missingness is non-random.
4. Class Balancing: The dataset exhibits class imbalance (fewer ASD-positive cases), which risks biasing the model toward predicting the majority class. To address this, we use SMOTE (Synthetic Minority Over-sampling Technique) to synthetically generate minority class instances, combined with Tomek link removal to eliminate borderline noise. This hybrid method not only balances the dataset but also improves class separability, enhancing the model’s ability to correctly classify ASD-positive cases without overfitting.
5. Feature Selection: Recursive Feature Elimination (RFE) is employed to identify and retain only the most predictive features. By recursively removing the least important attributes based on model weights, RFE simplifies the model, improves generalizability, and can reduce overfitting. This is especially useful when dealing with high-dimensional data where some inputs may be redundant or irrelevant. A leaner model is also more interpretable and efficient in deployment scenarios such as mobile health apps.

4.2 Advanced Machine Learning Techniques

* Classification Models: A diverse set of algorithms—Logistic Regression, SVM, Random Forest, XGBoost, and LightGBM—are employed to capture both linear and nonlinear relationships in the data. Logistic Regression offers interpretability, while ensemble models like Random Forest and gradient-boosted trees are powerful for complex interactions. This breadth ensures robust benchmarking and allows us to select the model that offers the best trade-off between performance and interpretability for ASD prediction.
* Hyperparameter Tuning: Models have internal settings (e.g., tree depth, regularization strength) that must be fine-tuned for optimal performance. GridSearchCV and RandomSearchCV are used to explore these settings efficiently across folds of the dataset, avoiding overfitting and ensuring model robustness across unseen data. These techniques systematically identify configurations that yield high validation scores, making the model more generalizable [8].
* Interpretability: Transparency is crucial in healthcare-related AI. SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) help reveal how specific features contribute to a prediction, both globally (across the dataset) and locally (for individual cases). This is essential to build trust among clinicians and end-users, and to ensure that the model bases its decisions on meaningful factors rather than spurious correlations [6, 7].
* Fairness Audits: ML models can unintentionally perpetuate existing social biases if not evaluated fairly. By auditing for equal opportunity (e.g., ensuring true positive rates are comparable across genders), we reduce the risk of discriminatory outputs. Fairness checks also help ensure compliance with ethical standards and regulatory requirements, which are increasingly important in AI governance frameworks.
* Automation: The entire preprocessing and modeling process is encapsulated in scikit-learn Pipelines, ensuring that data transformations and model evaluations are consistent and reproducible. Pipelines help prevent data leakage by ensuring that operations like scaling and imputation are fit only on training data within cross-validation folds. This promotes scientific rigor and simplifies deployment in real-world applications.

4.3 Anticipated Limitations

* Lack of socioeconomic and comorbidity data
* Limited geographic diversity
* Self-reporting biases (mitigated via external validation and interpretability tools)

5. Question Refinement and Analytical Plan

5.1 Refined Research Questions

1. **How accurately can ASD in adults be predicted using demographic and screening questionnaire data?**  
   *This question focuses on the core predictive modeling objective—whether meaningful patterns in self-reported data can reliably identify ASD status.*
2. **Can explainable ML models ensure fairness across age, gender, and ethnicity subgroups?**  
   *Given known diagnostic disparities, especially among females and ethnic minorities, this question addresses equitable model behaviour and subgroup analysis.*
3. **Which features have the highest predictive contribution toward ASD identification?**  
   *This explores feature importance, especially whether the model relies on key screening items (AQ-10) or demographic risk factors like family history.*
4. **How does class imbalance influence classification outcomes, and how can it be managed?**  
   *This assesses the effect of imbalanced data (fewer ASD-positive cases) and evaluates the effectiveness of rebalancing strategies like SMOTE-Tomek.*
5. **What role can unsupervised learning play in uncovering latent behavioral clusters?** *(Backup question)*  
   *If supervised learning models do not yield satisfactory accuracy, we will explore clustering algorithms such as DBSCAN or k-means to detect natural groupings within ASD-positive cases. This may uncover hidden symptom profiles or subtypes of autism not captured by binary screening outcomes.*
6. **Can publicly available forum data (e.g., Reddit discussions) be integrated to enrich contextual insights about self-reported autistic traits?** *(Backup dataset)*  
   *As an extension, we may collect unstructured data from ASD forums via the Pushshift API to identify linguistic or sentiment-based features that complement the structured dataset. This supports the “Variety” and “Volume” dimensions of big data and may improve generalisability.*

5.2 Visual and Analytical Tools

* Exploratory Visualization: Boxplots, violin plots, KDEs (Kernel Density Estimates), and correlation heatmaps will be used to explore the underlying distribution of variables and their interrelationships. These visual tools will help us uncover potential patterns, anomalies, and biases in the dataset. Specifically, they allow us to:
  + Boxplots: Identify outliers and compare spread across groups (e.g., ASD vs non-ASD)
  + Violin plots: Visualize distribution density and variability of scores across demographics
  + KDEs: Examine the shape of variable distributions to detect skewness or multimodality
  + Correlation heatmaps: Detect multicollinearity or hidden associations between features

By applying these methods early in the pipeline, we gain insights that directly inform feature selection, preprocessing decisions, and the choice of suitable ML models.

* Metrics: ROC-AUC, F1-score, precision, recall, MCC
* Fairness Dashboards: To track performance variance across demographic subgroups
* Alternative Modelling: Use of DBSCAN or k-means if supervised classifiers underperform

5.3 Planned Roadmap

* Pipeline Creation: Automate using scikit-learn's Pipeline class
* Model Selection: Sequential comparison of base and ensemble classifiers
* Evaluation: Stratified k-fold cross-validation
* Explainability Integration: Combine SHAP + LIME visualization for interpretability
* Deployment Planning: Prototype a mobile-friendly dashboard using Streamlit or Dash

6. Conclusion

This project seeks to bridge the adult autism diagnostic gap through scalable, responsible AI. By leveraging a well-structured, diverse dataset and robust ML methodologies, we aim to deliver a predictive model that enhances early detection efforts and supports self-directed assessments. Furthermore, the inclusion of explainability, fairness audits, and deployment planning ensures alignment with ethical AI principles and societal needs. This work will contribute to both academic knowledge and real-world impact in public health, education, and employment.

7. References

**[1]** Shinde, S. (2022). *Autism Prediction Dataset*. Kaggle.  
 <https://www.kaggle.com/datasets/shivamshinde123/autismprediction>

**[2]** World Health Organization. (2023). *Autism Spectrum Disorders – Fact Sheet*.  
 <https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders>

**[3]** Erl, T., Khattak, W., & Buhler, P. (2016). *Big Data Fundamentals: Concepts, Drivers & Techniques*. Pearson.  
 <https://www.pearson.com/en-us/subject-catalog/p/big-data-fundamentals-concepts-drivers-and-techniques/P200000003656>

**[4]** Marr, B. (2016). *Big Data in Practice: How 45 Successful Companies Used Big Data Analytics*. Wiley.  
 <https://www.wiley.com/en-us/Big+Data+in+Practice%3A+How+45+Successful+Companies+Used+Big+Data+Analytics+to+Deliver+Extraordinary+Results-p-9781119231387>

**[5]** Thomas, G. (2017). *How to Do Your Research Project: A Guide for Students*. SAGE Publications.  
<https://uk.sagepub.com/en-gb/eur/how-to-do-your-research-project/book245009>

**[6]** Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. NeurIPS.  
<https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>

**[7]** Wolpert, D. H. (1992). *Stacked Generalization*. Neural Networks.  
 <https://www.sciencedirect.com/science/article/abs/pii/S0893608005800231>

**[8]** Bergstra, J., & Bengio, Y. (2012). *Random Search for Hyper-Parameter Optimization*. Journal of Machine Learning Research (JMLR).  
 <https://jmlr.csail.mit.edu/papers/v13/bergstra12a.html>

**[9]** Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). *A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data*. SIGKDD Explorations.  
 <https://dl.acm.org/doi/10.1145/1007730.1007735>

**[10]** Centers for Disease Control and Prevention (CDC). (2023). *Autism Data and Statistics*.  
 <https://www.cdc.gov/ncbddd/autism/data.html>

**[11]** Di Martino, A., et al. (2014). *The Autism Brain Imaging Data Exchange: Towards a Large-Scale Evaluation of the Intrinsic Brain Architecture in Autism*. Molecular Psychiatry.  
 <https://www.nature.com/articles/mp201410>

**[12]** Australian Government, Office of the Australian Information Commissioner (OAIC). (2023). *The Privacy Act*.  
 <https://www.oaic.gov.au/privacy/the-privacy-act>