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 multifaceted neurodevelopmental condition that manifests differently across individuals and throughout various stages of life. Although general awareness has improved, adult diagnoses remain significantly underrepresented due to subtler symptomatic expressions and the absence of widespread screening initiatives. Delays in diagnosis can result in academic challenges, professional underachievement, psychological stress, and social exclusion from essential support systems [1,2].

## 1.2 Societal Significance

This project will investigate the feasibility of predicting adult autism using structured screening instruments in conjunction with demographic variables, leveraging supervised machine learning methodologies. A successfully deployed solution may contribute to early identification efforts by supporting self-administered platforms such as mobile applications. This initiative is aligned with modern healthcare paradigms that emphasise proactive intervention, scalable diagnostics, and data-augmented clinical decision-making frameworks [3,4].

# 2: Big Data Characteristics and Dataset Choice

## 2.1 Dataset Overview

• Name: Autism Screening Adult Dataset  
• Source: Kaggle (Fedesoriano, 2022)  
• Volume: ~1,050 instances with 20 features  
• Format: CSV  
• Key Attributes: Demographics, AQ-10 responses, scoring outcome, familial history, self-reported ASD diagnosis

## 2.2 Suitability in a Big Data Framework

Despite its modest volume, the dataset is structurally rich and highly relevant for proof-of-concept modeling. As it mirrors formal clinical screeners, the dataset allows for meaningful validation of preliminary classifiers. In subsequent stages, we plan to augment the data scope through open-source health repositories and textual inputs, fulfilling the broader 'big data' criteria of volume, variety, and velocity [1,3].

## 2.3 Addressing the 4 Vs

• Volume: Expansion through auxiliary public datasets  
• Variety: Integration of structured tabular data and unstructured forum/user data  
• Velocity: Real-time inputs via mobile diagnostics in future iterations  
• Veracity: Preprocessing and feature filtering to handle self-reporting biases [2,3]

## 2.4 Planned Supplementary Datasets

• CDC Autism Surveillance Datasets [10]  
• Reddit discussions via Pushshift API  
• ABIDE Neuroimaging Repository [11]

# 3: Industry Context and Value Proposition

## 3.1 Industry Application and Impact

The adoption of artificial intelligence in preventative and diagnostic healthcare has accelerated markedly in recent years. This project aligns with that trajectory, especially in neurodiversity-informed tools. Sectors such as digital therapeutics, mobile health, and telepsychiatry stand to benefit from intelligent pre-screeners. With the AI healthcare market projected to surpass USD 100 billion by 2030, innovations targeting adult mental health disparities hold growing significance.

## 3.2 Relevance to Public Health and Policy

By contributing to earlier ASD identification in adults, this project will address a critical gap identified by clinicians and policymakers. The CDC has reported widespread underdiagnosis among adults, leading to inequities in care access and resource distribution [10]. Our proposed lightweight, explainable classifier aims to empower users to initiate help-seeking behaviours based on personalised triage.

## 3.3 Feedback Loops and Personalisation Potential

We intend to conceptualise a mobile-based triage interface capable of learning from aggregate user inputs. Iterative updates—driven by new data streams such as symptom clusters—will enable the model to adapt. This approach supports the demand for scalable, personalised healthcare systems [3,4].

## 3.4 Cross-Industry Value

• Healthcare: Preliminary screening aids clinical intervention decisions  
• Human Resources: Consent-based screening can inform neurodiversity accommodations  
• Education: Digital tools may assist in identifying students with learning or engagement difficulties

# 4: Data Processing and Risk Mitigation

## 4.1 Ethical Framework

Handling sensitive mental health data requires adherence to privacy standards. Although the dataset is anonymised and publicly available, we will follow ethical frameworks such as the GDPR and Australian Privacy Act [12], and secure informed consent where applicable.

## 4.2 Responsible Development Practices

• Bias Audits: Evaluation of demographic fairness via equal opportunity and disparate impact analysis  
• Interpretability: SHAP (SHapley Additive exPlanations) will clarify model decisions [6]  
• User Feedback: Collection of anonymised feedback on model predictions to enable refinement

## 4.3 Preprocessing Strategy

1. Remove metadata unless analytically useful  
2. Apply encoders for categorical variables  
3. Normalise continuous variables using MinMaxScaler  
4. Balance imbalanced labels with SMOTE-Tomek hybrid [9]  
5. Select features using Recursive Feature Elimination and correlation filtering

## 4.4 Advanced Methodologies

• SHAP for interpretability [6]  
• Ensemble models and stacking [7]  
• GridSearchCV for hyperparameter tuning [8]  
• SMOTE + Tomek links for class balancing [9]

## 4.5 Anticipated Limitations

The dataset omits factors like socioeconomic status and comorbidities. These limitations will be addressed through the addition of supplementary datasets in future iterations.

# 5: Question Refinement and Analytical Plan

## 5.1 Exploratory Analysis Strategy

Initial visual analyses will evaluate AQ-10 distributions and familial ASD history correlations. We expect positive skewness in diagnosed cases. Heatmaps will highlight correlations between features, boxplots will explore distributional differences by gender and age, and KDE plots will help detect density overlaps. Additionally, violin plots and pair plots will support pattern recognition across features. This strategy will help identify key contributors, guide feature selection, and detect hidden biases or data sparsity.

## 5.2 Refined Research Questions

• To what extent can ASD in adults be predicted using self-assessment responses and demographic variables?  
• Can machine learning models provide fair and interpretable predictions across gender and age subgroups?  
• How do different classification algorithms (e.g., Logistic Regression, Random Forest, SVM) compare in predicting ASD from structured survey data?  
• What features contribute most significantly to the accurate identification of ASD in adults using SHAP explainability techniques?  
• Can model fairness be maintained across underrepresented demographic groups (e.g., females, ethnic minorities)?  
• How does class imbalance in ASD datasets affect predictive model performance and calibration?  
• What preprocessing pipeline yields the most robust model performance when dealing with real-world, self-reported ASD screening data?  
• Can unsupervised learning methods (e.g., clustering) reveal latent symptom groupings not captured by diagnostic labels?  
• How might real-time screening tools deployed via mobile apps impact ASD detection rates in adult populations?  
• What ethical considerations must be addressed to ensure responsible deployment of machine learning tools in mental health diagnostics?

## 5.3 Planned Methodological Roadmap

• **Data Visualization**: Firstly, we will employ matplotlib and seaborn to produce high-quality visualizations, including correlation heatmaps, pair plots, and kernel density estimates (KDEs). These tools will allow us to uncover latent patterns, assess multicollinearity, and explore feature distributions across diagnostic classes.

• **Classifiers**: Secondly, we aim to compare a suite of classification algorithms. Baseline models will include Logistic Regression, Random Forest, and Support Vector Machines (SVM). Moreover, we will implement advanced gradient-boosting models such as XGBoost and LightGBM, which are well-suited to handling class imbalance and non-linear interactions.

• **Explainability**: Furthermore, to ensure transparency and trustworthiness in predictions, we will apply SHAP (SHapley Additive exPlanations) to interpret model outputs. In addition, LIME (Local Interpretable Model-Agnostic Explanations) may be incorporated to triangulate SHAP findings and facilitate explanation of model behavior to non-technical stakeholders.

• **Evaluation Metrics**: To conduct a rigorous performance analysis, we will evaluate models using multiple metrics. These include ROC-AUC, F1-score, accuracy, recall, and precision. Additionally, we will compute the Matthews Correlation Coefficient (MCC), which provides balanced measure even in the presence of imbalanced class distributions.

• **Cross-Validation**: We will implement k-fold cross-validation with stratified sampling to ensure robust generalization and to preserve the original class distribution in each fold. This will reduce variance in performance estimates across models.

• **Hyperparameter Tuning**: To fine-tune the performance of each classifier, both GridSearchCV and RandomizedSearchCV will be applied. These techniques will explore hyperparameter spaces efficiently and systematically to identify optimal configurations.

• **Pipeline Automation**: Additionally, we will construct modular and reusable data pipelines using scikit-learn’s Pipeline class. This will streamline preprocessing, transformation, and model training steps, ensuring reproducibility and reducing human error.

• **Model Comparison Dashboard**: Finally, for comparative analysis, we plan to design an interactive dashboard using tools such as Plotly or Dash. This dashboard will present model metrics, feature importance rankings, and fairness indicators in a user-friendly format suitable for both technical reviews and stakeholder presentations.

## 5.4 Alternative Pathway

If classifiers do not perform well, we will pivot to unsupervised learning. Clustering techniques such as k-means and DBSCAN may expose latent patterns, further enhanced by contextual linking with forum datasets [10,11].

# 6: Contribution and Conclusion

This initiative addresses a diagnostic void by focusing on adult ASD populations, who are often excluded from early intervention programs. By designing an accessible and explainable model, we aim to democratise ASD screening. Unlike conventional diagnostic pipelines that rely on extensive clinical evaluation, this solution will offer a lightweight, modular approach to support mental health awareness and timely care.

In conclusion, this proposed project will harness the power of machine learning to support early detection of adult autism in a responsible, scalable, and ethical manner. The outlined methods—ranging from model explainability to fairness auditing—will not only ensure robustness but also enhance public trust in data-driven diagnostic tools. As societal awareness of neurodiversity grows, our system could serve as an enabler of inclusive healthcare strategies, allowing for tailored interventions that respect the individuality of those on the spectrum. By rigorously evaluating data sources, refining analytical strategies, and addressing known limitations, this work aspires to deliver meaningful insights with practical applicability across healthcare, education, and employment domains.

# 7: References

[1] Soriano, F. (2022). Autism Screening Adult Dataset. Kaggle. https://www.kaggle.com/datasets/fedesoriano/autism-screening-on-adults

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

[3] Erl, T., Khattak, W., & Buhler, P. (2016). Big Data Fundamentals. Pearson.

[4] Marr, B. (2016). Big Data in Practice. Wiley.

[5] Thomas, G. (2017). How to Do Your Research Project. SAGE.

[6] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. NIPS.

[7] Wolpert, D. H. (1992). Stacked Generalization. Neural Networks.

[8] Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. JMLR.

[9] Batista, G. E., Prati, R. C., & Monard, M. C. (2004). Balancing Machine Learning Training Data. SIGKDD.

[10] 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. Molecular Psychiatry.

[12] Australian Government OAIC. (2023). Privacy Act. https://www.oaic.gov.au/privacy/the-privacy-act