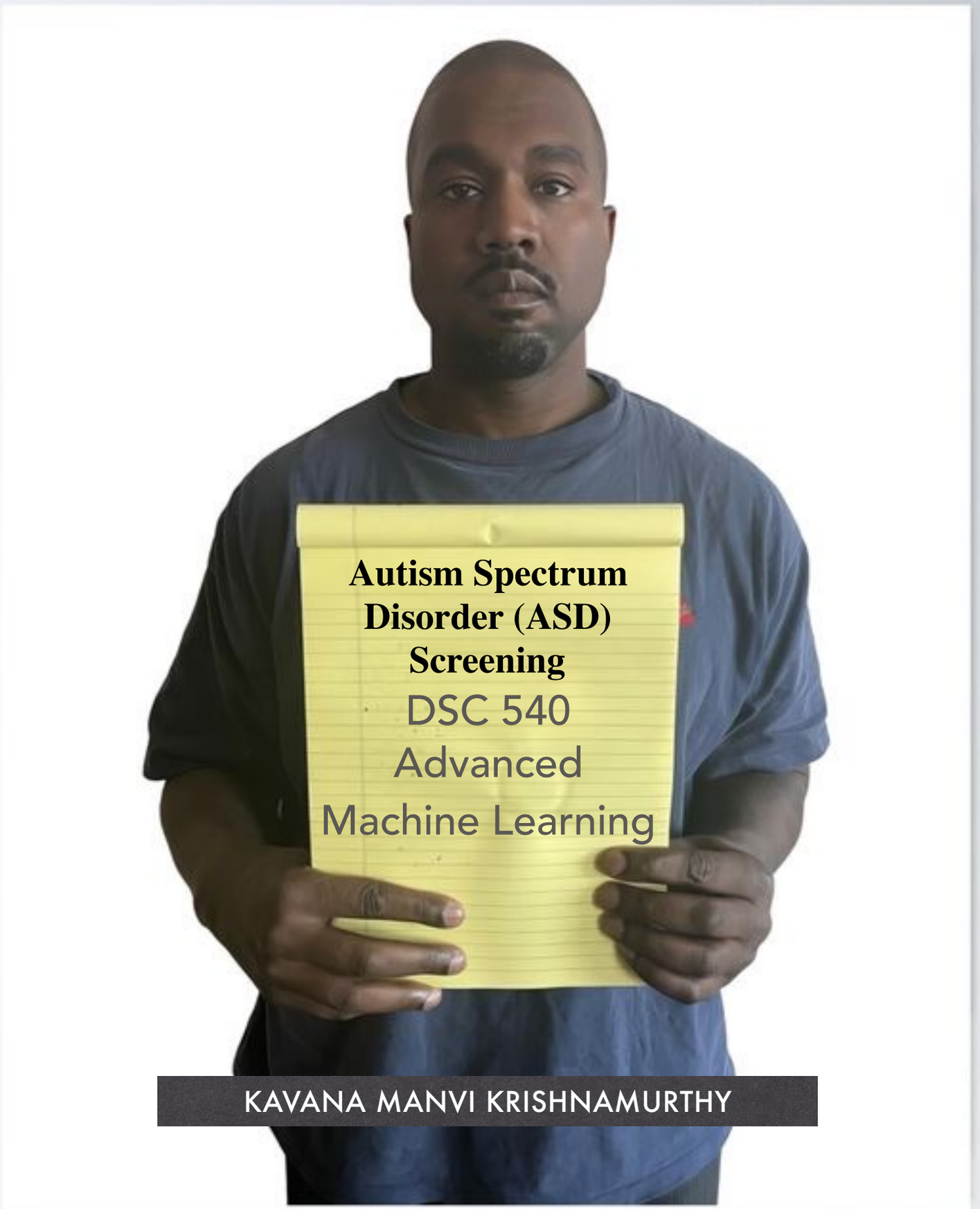

Autism Spectrum
Disorder (ASD)
Screening
DSC 540
Advanced
Machine Learning
KAVANA MANVI KRISHNAMURTHY

INTRODUCTION

- Autism Spectrum Disorder (ASD) is a neurodevelopmental condition with significant healthcare costs.
- Early diagnosis can reduce costs and improve outcomes.
- Current ASD diagnosis processes are costly and have lengthy waiting times.
- There is an urgent need for time-efficient and accessible ASD screening methods.

DATASET OVERVIEW

New dataset developed for adult ASD screening.

Contains 20 features:

- 10 behavioral features (AQ-10-Adult).
- 10 individual characteristics.
- 609 instances(after cleaning)

Designed to assist in identifying influential autistic traits and improve ASD classification.

Attribute	Type	Description
Age	Number	Age in years
Gender	String	Male or Female
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean (yes or no)	Whether the case was born with jaundice
Family member with PDD	Boolean (yes or no)	Whether any immediate family member has a PDD
Who is completing the test	String	Parent, self, caregiver, medical staff, clinician, etc.
Country of residence	String	List of countries in text format
Used the screening app before	Boolean (yes or no)	Whether the user has used a screening app
Screening Method Type	Integer (0,1,2,3)	Type of screening methods chosen based on age category (0=toddler, 1=child, 2=adolescent, 3=adult)
Screening Score	Integer	The final score obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner
A_1	Binary (0, 1)	Difficulty noticing small details in surroundings
A_2	Binary (0, 1)	Struggles with social interactions and making friends
A_3	Binary (0, 1)	Finds it challenging to understand others' emotions
A_4	Binary (0, 1)	Prefers routine and feels distressed when routines change
A_5	Binary (0, 1)	Strong focus on specific interests or hobbies
A_6	Binary (0, 1)	Finds it difficult to engage in conversations or follow social cues
A_7	Binary (0, 1)	Notices patterns or details that others may overlook
A_8	Binary (0, 1)	Prefers solitude over social gatherings
A_9	Binary (0, 1)	Feels overwhelmed in noisy or crowded environments
A_10	Binary (0, 1)	Has difficulty adapting to unexpected changes

METHODOLOGY

1. DATA CLEANING - MISSING VALUES
2. DATA EXPLORATION - CORRELATION
3. DATA PREPROCESSING - NORMALISATION
4. FEATURE SELECTION - 3 TYPES
5. MODEL - 5 ML ALGORITHMS,
6. EVALUATION - CV, AUC, ACCURACY

DATA CLEANING

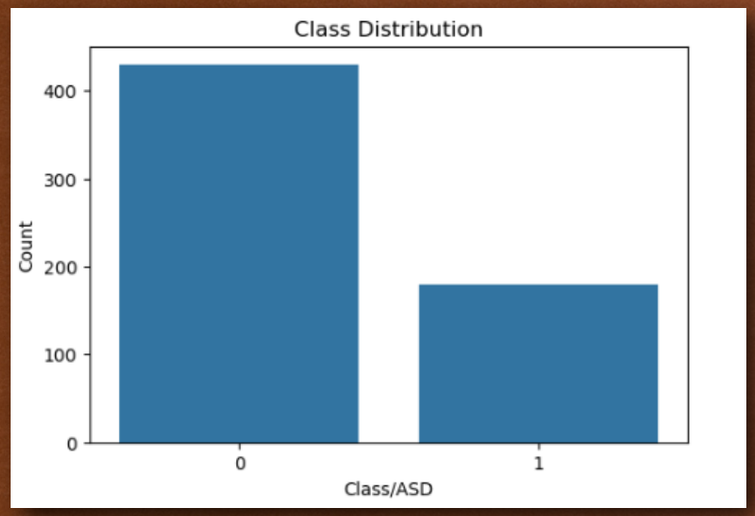
Handling Missing Values

- Age Missing Values- MODE
  - Replaced missing **age** values with its **most frequent value** (mode).
  - Preserves data distribution while maintaining relevance for model training.
- Ethnicity Missing Values- DROPPED
  - Dropped **ethnicity** values due to their association with the "No autism" class.
  - Ensures focus on autism-related features, improving model sensitivity.
  - Reduces potential **bias** without compromising model performance.

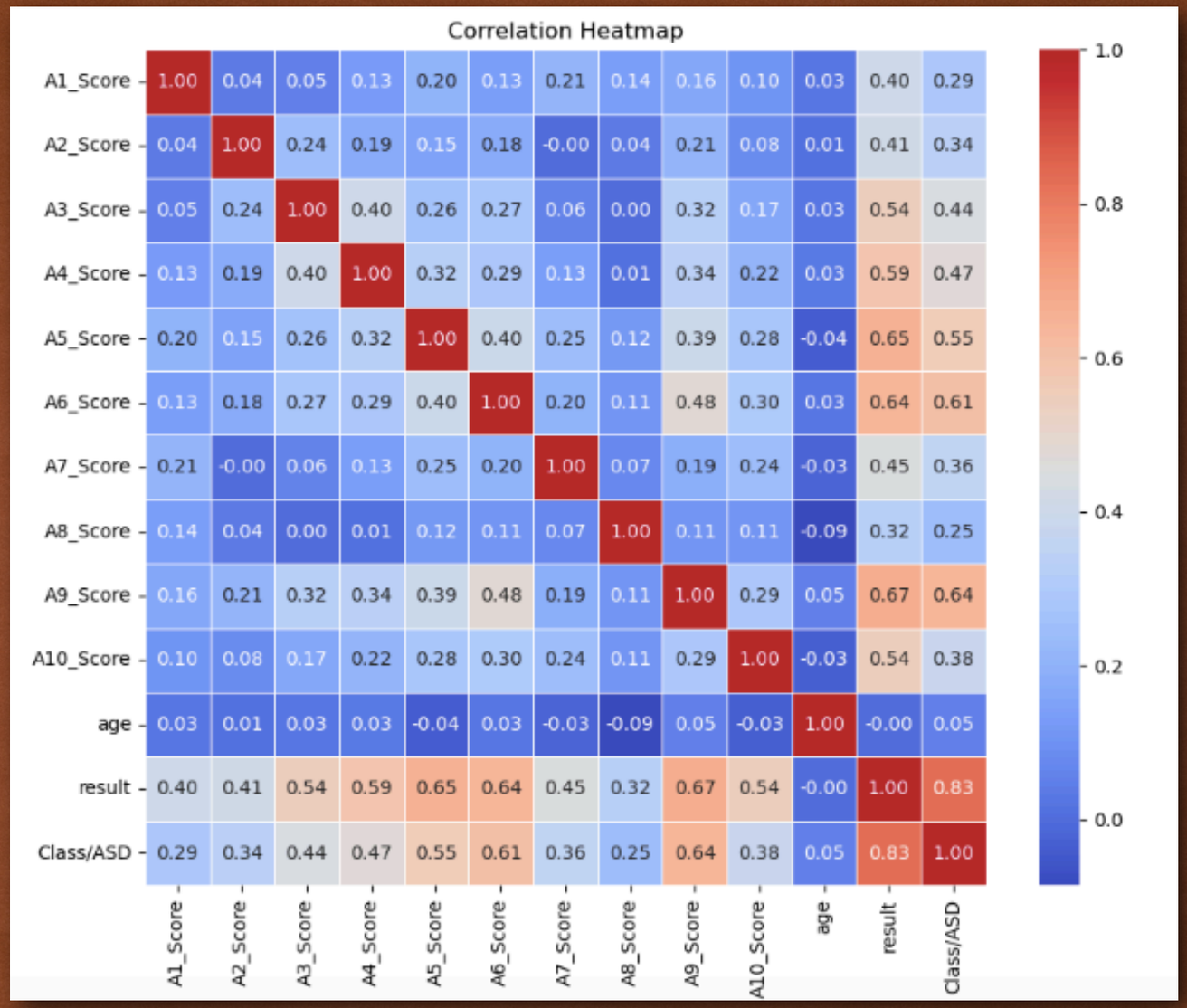
EXPLORATION

Imbalanced dataset!

609 cases

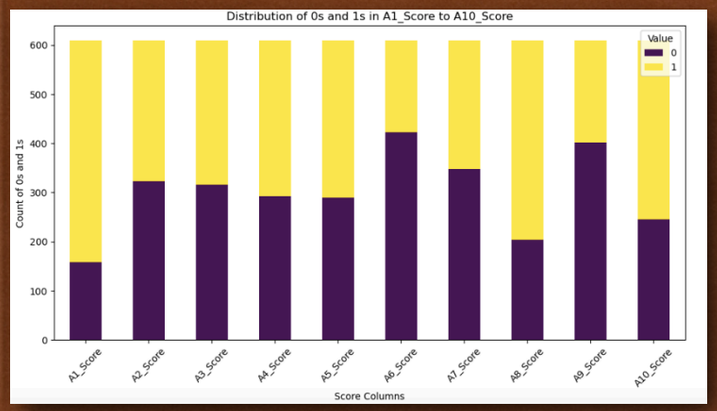


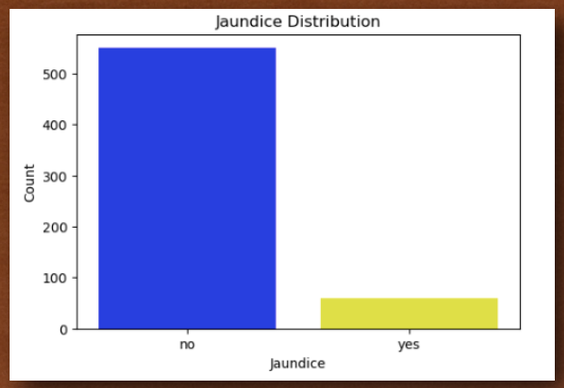
Class/ASD	
NO	429
YES	180

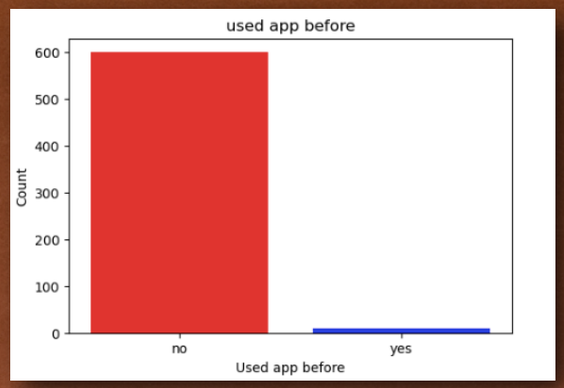


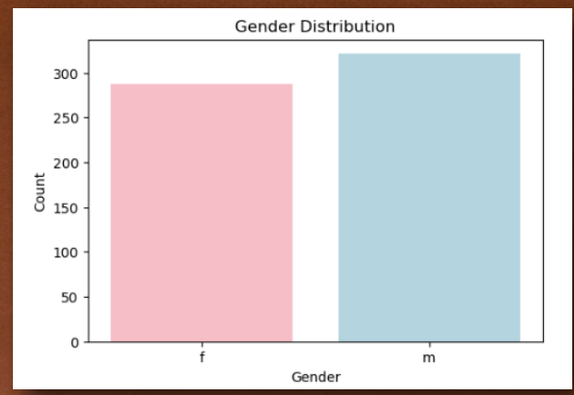
Removed ‘result’

CATEGORICAL: BINARY

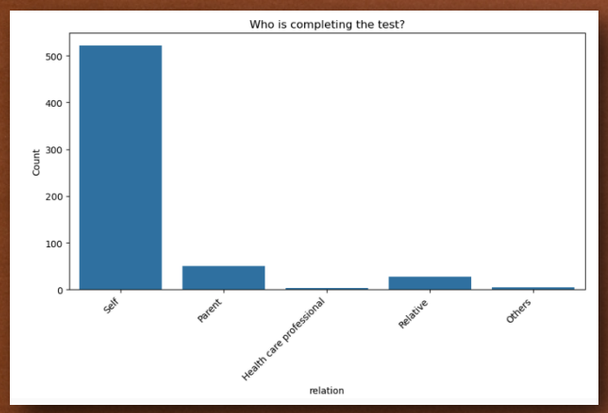


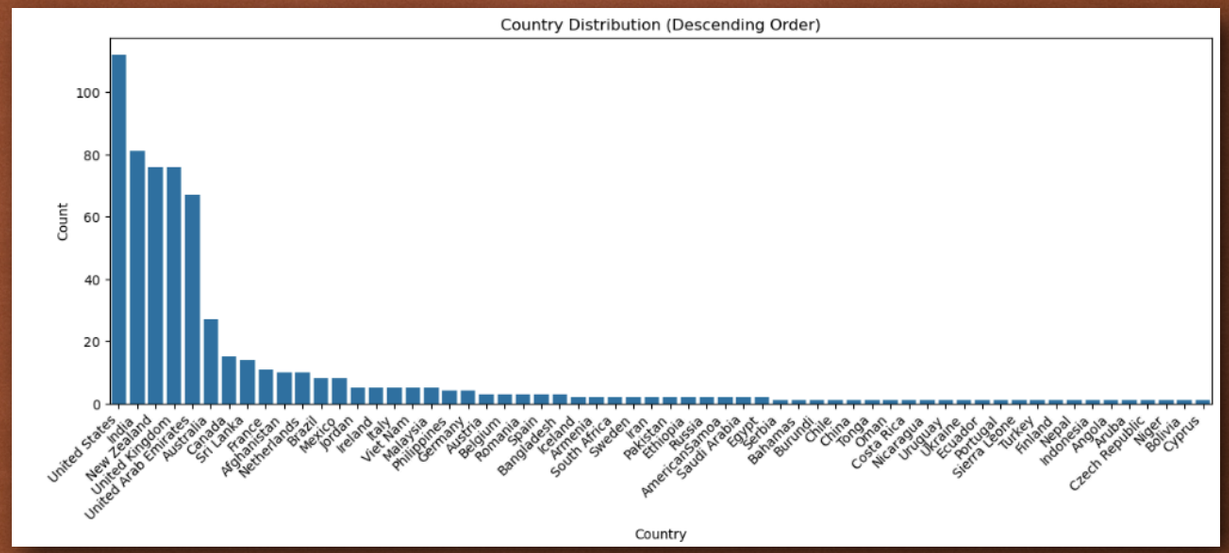




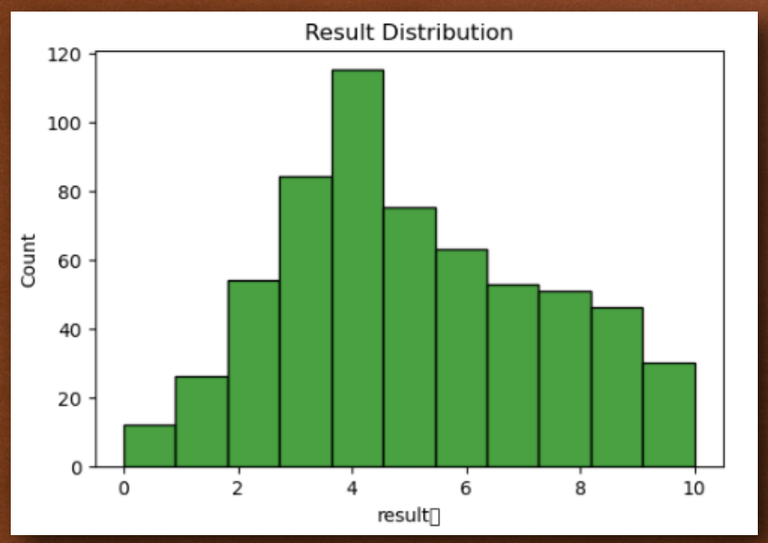


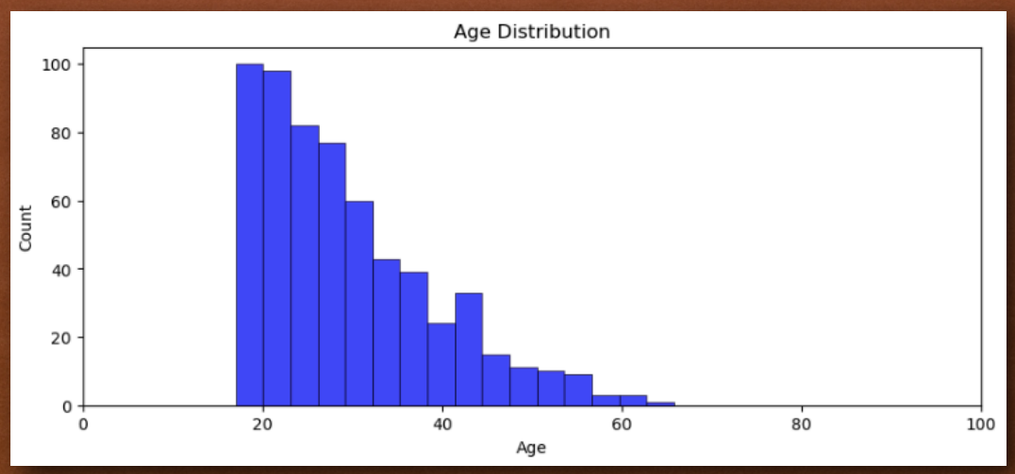
CATEGORICAL:MULTI VALUED ATTRIBUTES





NUMERICAL:





DATA PREPROCESSING

Convert Categorical Variables:

- Convert "yes/no" and "male/female" into binary values (0/1) for compatibility with models like logistic regression and neural networks.

Apply One-Hot Encoding:

- Use one-hot encoding for categorical features with multiple distinct values to avoid introducing incorrect ordinal relationships.

Normalize Numerical Features:

- Apply Min-Max scaling on numerical columns (e.g., "age" and "result") to ensure they are scaled within a fixed range (0 to 1), improving model convergence in gradient-based models.

FEATURE SELECTION

```
num_columns = df_normalized.shape[1]
print(f"Number of columns: {num_columns}")
```

```
Number of columns: 91
```

1. **Mutual Information (MI):** Captures non-linear relationships between features and ASD diagnosis, highly effective for behavioral and clinical data.

2. **Recursive Feature Elimination (SVC):** Iteratively removes the weakest features, commonly used with Support Vector Machines (SVM) for optimal feature selection.

3. (RF)**Conditional Mutual Information Maximization (CMIM):** Enhances diagnostic accuracy, showing superior performance when combined with Random Forest

TRAINING AND EVALUATION

TRAINING

- Models Tested :
  Why? Diverse algorithms to capture linear/non-linear patterns.
- **Run Grid Search first** to find the best hyperparameters for each model.

1. XGBoost,
2. SVM (RBF),
3. Logistic Regression,
4. Random Forest,
5. MLP Neural Network

EVALUATION METRICS

- Stratified 5-Fold CV :Maintains class balance across folds → critical for imbalanced medical data.

1. **Accuracy**: Overall correctness (risk: misleading for imbalanced classes).
2. **AUC-ROC**: Model's ability to distinguish classes (threshold-independent).
3. **Specificity**: True Negative Rate (avoid misdiagnosing healthy individuals).
4. Sensitivity (Recall): True Positive Rate → Most critical for ASD.

XGBOOST

What is XGBoost? A gradient boosting algorithm optimized for speed and performance.

Why is XGBoost good for ASD data? Handles missing values, feature importance ranking, and class imbalance well.

Hyperparameters Used: `max_depth`, `learning_rate`, `subsample`, `colsample_bytre`

Robust Against Overfitting: It includes built-in regularization techniques to prevent overfitting, ensuring robust performance on complex datasets.

Feature Selection	Accuracy (+/-)	AUC (+/-)	Sensitivity (+/-)	Specificity (+/-)	CV Runtime (sec)
All	0.9705 (+/- 0.02)	0.9629 (+/- 0.03)	0.9444 (+/- 0.07)	0.9814 (+/- 0.03)	1.139
Mutual Information (MI)	0.9392 (+/- 0.03)	0.9230 (+/- 0.05)	0.8833 (+/- 0.12)	0.9627 (+/- 0.05)	0.455
Recursive Feature Elimination (RFE)	0.9721 (+/- 0.03)	0.9657 (+/- 0.04)	0.9500 (+/- 0.08)	0.9814 (+/- 0.04)	0.400
CMIM-inspired RF-based Selection	0.9705 (+/- 0.02)	0.9629 (+/- 0.03)	0.9444 (+/- 0.07)	0.9814 (+/- 0.03)	0.502

```
XGBCLASSIFIER(USE_LABEL_ENCODER=FALSE,
EVAL_METRIC='LOGLOSS',MAX_DEPTH=NONE,N_ESTIMATORS=100,RANDOM_STATE=RAND_ST)
```

SVM(RBF KERNEL)

What is SVM? A classification algorithm that finds the optimal decision boundary using kernels.

Why is SVM good for ASD data? RBF kernel captures non-linear patterns, and it performs well on small datasets.

Hyperparameters Used: `C`, `gamma`, `class_weight`.- imbalance

Feature Selection	Accuracy (+/-)	AUC (+/-)	Sensitivity (+/-)	Specificity (+/-)	CV Runtime (sec)
All	0.9573 (+/- 0.02)	0.9681 (+/- 0.02)	0.9944 (+/- 0.02)	0.9417 (+/- 0.02)	0.472
Mutual Information (MI)	0.9343 (+/- 0.04)	0.9356 (+/- 0.06)	0.9389 (+/- 0.12)	0.9324 (+/- 0.05)	0.172
Recursive Feature Elimination (RFE)	0.9721 (+/- 0.03)	0.9802 (+/- 0.02)	1.0000 (+/- 0.00)	0.9604 (+/- 0.05)	0.186
CMIM-inspired RF-based Selection	0.9622 (+/- 0.02)	0.9700 (+/- 0.02)	0.9889 (+/- 0.03)	0.9510 (+/- 0.02)	0.307

```
SVC(KERNEL="RBF", PROBABILITY=TRUE, CLASS_WEIGHT='BALANCED', C=1.0,GAMMA='AUTO')
```

LOGISTIC REGRESSION

What is Logistic Regression? A simple, interpretable model for binary classification.

Why is Logistic Regression good for ASD data? Effective baseline model and provides feature importance insights.

Hyperparameters Used: `C`, `solver`, `penalty`.

Feature Selection	Accuracy (+/-)	AUC (+/-)	Sensitivity (+/-)	Specificity (+/-)	CV Runtime (sec)
All	0.9819 (+/- 0.01)	0.9856 (+/- 0.01)	0.9944 (+/- 0.02)	0.9767 (+/- 0.02)	0.088
Mutual Information (MI)	0.9392 (+/- 0.02)	0.9472 (+/- 0.02)	0.9667 (+/- 0.04)	0.9278 (+/- 0.03)	0.065
Recursive Feature Elimination	0.9918 (+/- 0.03)	0.9942 (+/- 0.02)	1.0000 (+/- 0.00)	0.9884 (+/- 0.04)	0.087
CMIM-inspired RF-based Selectio	0.9836 (+/- 0.01)	0.9883 (+/- 0.01)	1.0000 (+/- 0.00)	0.9767 (+/- 0.02)	0.083

`LOGISTICREGRESSION(MAX_ITER=1000, CLASS_WEIGHT='BALANCED', SOLVER= 'LBFGS', C=1.0)`

RANDOM FOREST

What is Random Forest? An ensemble method that combines multiple decision trees.

Why is Random Forest good for ASD data? Handles missing values, reduces overfitting, and ranks feature importance.

Hyperparameters Used:
`n_estimators`, `max_depth`, `min_samples_split`.

Feature Selection	Accuracy (+/-)	AUC (+/-)	Sensitivity (+/-)	Specificity (+/-)	CV Runtime (sec)
All	0.9491 (+/- 0.04)	0.9268 (+/- 0.06)	0.8722 (+/- 0.12)	0.9813 (+/- 0.01)	1.204
Mutual Information (MI)	0.9409 (+/- 0.06)	0.9242 (+/- 0.06)	0.8833 (+/- 0.08)	0.9651 (+/- 0.07)	1.360
Recursive Feature Elimination (RFE)	0.9672 (+/- 0.04)	0.9606 (+/- 0.04)	0.9444 (+/- 0.04)	0.9767 (+/- 0.06)	1.109
CMIM-inspired RF-based Selection	0.9491 (+/- 0.02)	0.9268 (+/- 0.03)	0.8722 (+/- 0.06)	0.9813 (+/- 0.01)	1.181

```
RANDOMFORESTCLASSIFIER(N_ESTIMATORS=100,MAX_DEPTH=20,
CRITERION='ENTROPY',MIN_SAMPLES_SPLIT=3,RANDOM_STATE=RAND_ST)
```

MLP

What is MLP? A neural network that captures complex relationships in data.

Why is MLP good for ASD data? Learns non-linear patterns and adapts to various data distributions.

Hyperparameters Used: `hidden_layer_sizes`, `activation`, `alpha`.

Feature Selection	Accuracy (+/-)	AUC (+/-)	Sensitivity (+/-)	Specificity (+/-)	CV Runtime (sec)
All	0.9556 (+/- 0.02)	0.9540 (+/- 0.04)	0.9500 (+/- 0.08)	0.9580 (+/- 0.01)	5.334
Mutual Information (MI)	0.9655 (+/- 0.03)	0.9594 (+/- 0.03)	0.9444 (+/- 0.04)	0.9744 (+/- 0.03)	4.285
Recursive Feature Elimination (RFE)	0.9918 (+/- 0.01)	0.9926 (+/- 0.02)	0.9944 (+/- 0.02)	0.9907 (+/- 0.02)	4.213
CMIM-inspired RF-based Selection	0.9573 (+/- 0.02)	0.9520 (+/- 0.03)	0.9389 (+/- 0.06)	0.9650 (+/- 0.03)	5.558

- `MLPCLASSIFIER(MAX_ITER=300, ACTIVATION='RELU', LEARNING_RATE=0.01,RANDOM_STATE=RAND_ST)`

RESULTS

The combination that produces the best sensitivity is:(Reduces false positives)

- SVM (RBF Kernel) with Recursive Feature Elimination (RFE)

Model ◦	Feature Selection	Accuracy (+/-)	AUC (+/-)	Sensitivity (+/-)	Specificity (+/-)	CV Runtime (sec)
SVM (RBF Kernel)	Recursive Feature Elimination (RFE)	0.9721 (+/- 0.03)	0.9802 (+/- 0.02)	1.0000 (+/- 0.00)	0.9604 (+/- 0.05)	0.186
Logistic Regression	Recursive Feature Elimination (RFE)	0.9918 (+/- 0.03)	0.9942 (+/- 0.02)	1.0000 (+/- 0.00)	0.9884 (+/- 0.04)	0.087
MLP Neural Network	Recursive Feature Elimination (RFE)	0.9918 (+/- 0.01)	0.9926 (+/- 0.02)	0.9944 (+/- 0.02)	0.9907 (+/- 0.02)	4.213
Logistic Regression	All	0.9819 (+/- 0.01)	0.9856 (+/- 0.01)	0.9944 (+/- 0.02)	0.9767 (+/- 0.02)	0.088
MLP Neural Network	All	0.9556 (+/- 0.02)	0.9540 (+/- 0.04)	0.9500 (+/- 0.08)	0.9580 (+/- 0.01)	5.334

MITIGATE OVERFITTING

Cross-Validation: Stratified K-Fold ensures robust performance estimation.

Hyperparameter Tuning: GridSearchCV for optimal parameters in XGBoost, SVM, Logistic Regression, Random Forest, and MLP.

Regularization: L2 (Logistic Regression), Alpha (MLP), and Class Weights (SVM, Random Forest) to prevent overfitting.

Feature Selection: RFE and Mutual Information to reduce noise and irrelevant features.

Model Complexity Control: Limiting depth, estimators, and splits in tree-based models (XGBoost, Random Forest).

Ensemble Learning: Combines models for improved generalization.

CONCLUSION

RBF Kernel SVM emerged as the best-performing model.

- **Sensitivity**: Highest at 0.99, indicating strong detection of positive cases.
- **AUC**: Achieved an AUC of 0.97, signifying excellent model discrimination.
- **Specificity**: 0.94, ensuring a balanced performance with low false positives.
- **Accuracy**: 0.96, with consistent and reliable predictions across different subsets.

Feature Selection Impact:

- The combination of **Mutual Information (MI)** and **Recursive Feature Elimination (RFE)** significantly improved model performance.

Model Tuning:

- **Hyperparameter Optimization** (via GridSearchCV) enhanced model performance and reduced overfitting.

FUTURE WORK

Hyperparameter Tuning:

- Further optimization using **randomized search** or **Bayesian optimization** for fine-tuning model performance.

Advanced Feature Selection Techniques:

- Experiment with L1 regularization or mutual information-based feature selection to enhance model performance.

Ensemble Learning:

- Combine models like **Random Forest** and **XGBoost** for **stacking** or **boosting** to reduce bias and variance.

THANK YOU