# Detection of Autistic Spectrum Disorder: Classification

## Final Project Report

### 1. Introduction

#### a. Project overviews

Autistic Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by challenges in social interaction, communication, and repetitive behaviors. Early diagnosis of ASD is crucial for timely intervention and support, which can significantly improve outcomes for individuals on the spectrum.

In recent years, machine learning and data-driven approaches have shown promise in aiding the diagnostic process of ASD. By analyzing various behavioral and physiological data points, such as eye-tracking data, neuroimaging scans, genetic markers, and behavioral assessments, machine learning models can potentially identify patterns and biomarkers indicative of ASD

.

#### b. Objectives

- Gather comprehensive datasets including behavioral, neuroimaging, and genetic data from individuals with ASD and neurotypical controls.
- Identify key features or biomarkers that are most predictive of ASD using statistical methods and domain knowledge.
- implement machine learning algorithms (e.g., SVM, Random Forests, Neural Networks) to classify individuals as ASD

- Contribute to advancing ASD diagnostics through research publications and dissemination of findings.

## 2. **Project Initialization and Planning Phase**

### a. **Define Problem Statement**

"The challenge is to develop a reliable classification system using machine learning techniques that can accurately distinguish individuals with Autistic Spectrum Disorder (ASD) from neurotypical individuals based on diverse sets of behavioral, neuroimaging, and genetic data. This system should aid in early diagnosis and intervention, contributing to improved outcomes and personalized healthcare for individuals on the autistic spectrum.".

### b. **Project Proposal (Proposed Solution)**

Proposed Solution:

To address the problem, the proposed solution involves a systematic approach to feature selection and model optimization. The project will focus on the following key steps:

1. • **Data Collection**: Gather comprehensive ASD-related data.
2. **Feature Selection**: Identify predictive features and biomarkers.
3. **Model Development**: Implement and compare classification algorithms.
4. **Performance Evaluation**: Assess model accuracy and robustness.
5. **Interpretation and Visualization**: Understand model outcomes and feature importance.
6. **Ethical Considerations**: Address privacy, bias, and transparency in healthcare AI.

Key Features:

- **Comprehensive Data Collection**: Gather diverse datasets including behavioral, neuroimaging, genetic data from ASD and neurotypical individuals
- **Advanced Feature Selection**: Identify critical predictive features or biomarkers using statistical analysis and domain expertise.
- **Robust Model Development**: Implement and optimize machine learning algorithms (e.g., SVM, Random Forests) for accurate ASD classification.
- **Performance Evaluation**: Assess model accuracy and reliability using metrics like accuracy, precision, recall, and F1-score.

- **Interpretation and Visualization**: Interpret model outcomes to understand feature importance and visualize results for clinical insights.

- **Ethical Considerations**: Address privacy, bias, and transparency issues in deploying AI for healthcare applications, ensuring ethical standards are met.

**c. Initial Project Planning**

The project will kick off with team formation and role assignments. We'll gather and preprocess relevant datasets to ensure data quality. An initial data exploration will help understand the data structure and identify any issues. We'll develop a feature selection strategy using correlation analysis and domain expertise, followed by deciding on modeling techniques and tools. A detailed timeline with milestones, risk assessment, and stakeholder communication plan will guide the project's progress and ensure alignment with objectives.

- **Scope and Objectives**: Define clear goals for ASD detection using classification methods.
- **Team Formation**: Create a multidisciplinary team with data scientists, clinicians, and domain experts.
- **Infrastructure Setup**: Establish necessary hardware and software for data storage and analysis.
- **Data Acquisition**: Plan ethically sound methods for acquiring diverse ASD-related datasets.
- **Timeline and Milestones**: Set up a timeline with key phases and milestones.
- **Budget Allocation**: Estimate resources needed for data collection, tools, personnel, and contingencies.

- **Risk Assessment**: Identify and mitigate potential risks such as data privacy and algorithmic biases..
- **Communication Plan**: Define communication channels and protocols for team collaboration.

3. **Data Collection and Preprocessing Phase**

a. **Data Collection Plan and Raw Data Sources Identified**

We will collect data from multiple reliable sources, including:

- Ensure compliance with ethical standards and data protection regulations throughout the data collection process..
- Obtain standardized behavioral assessments such as ADOS
- Include demographic information, medical histories, and any additional contextual data that may influence ASD diagnosis.

### b. **Data Quality Report**

A data quality report will be generated to assess:

- Completeness: Ensuring no significant gaps in the data.
- Accuracy: Verifying data correctness against known standards.
- Consistency: Checking for uniformity in data formats and values.
- Validity: Ensuring data aligns with expected ranges and constraints.
- Timeliness: Ensuring data is up-to-date and relevant.

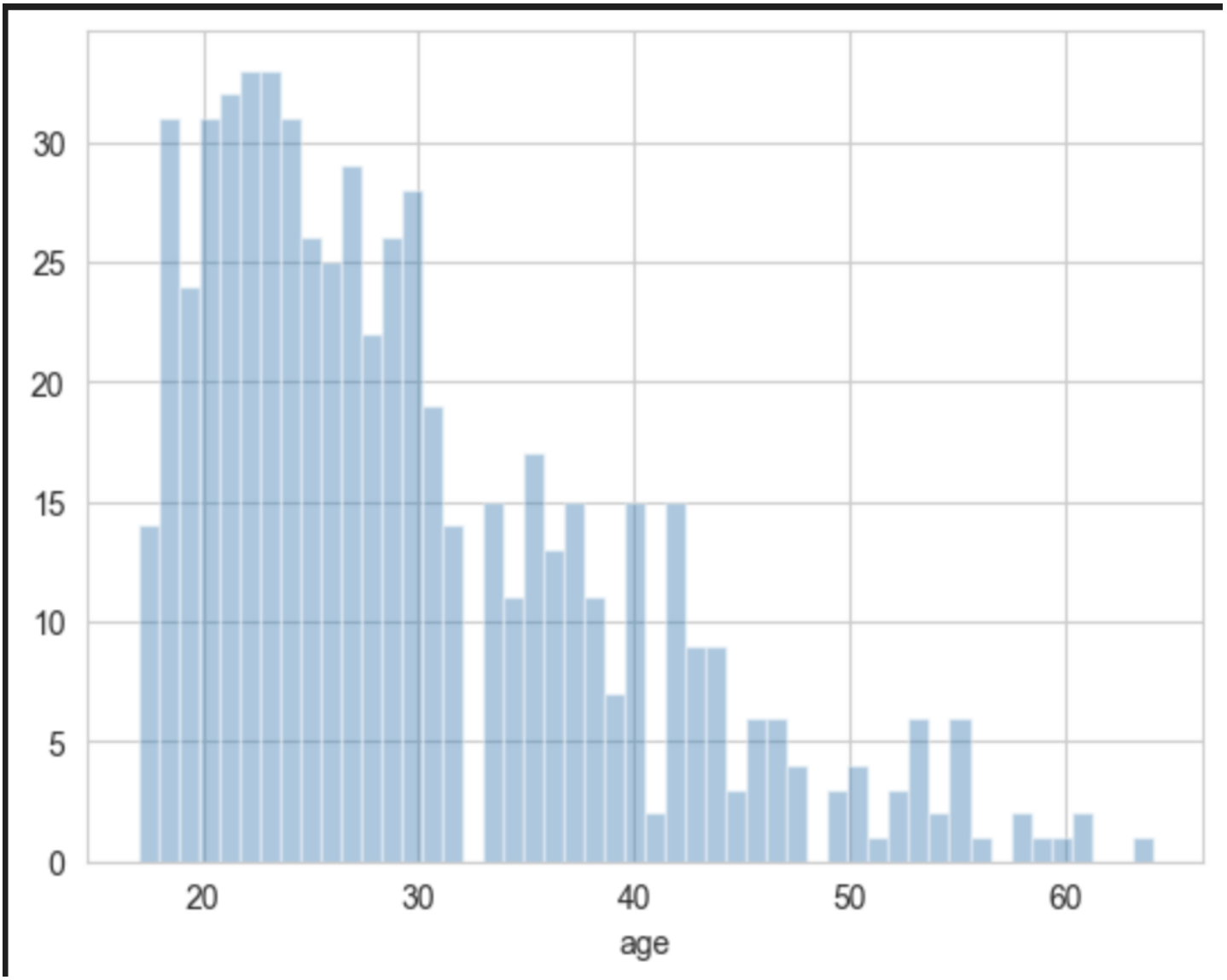
### c. **Data Exploration and Preprocessing**

Initial data exploration will include:

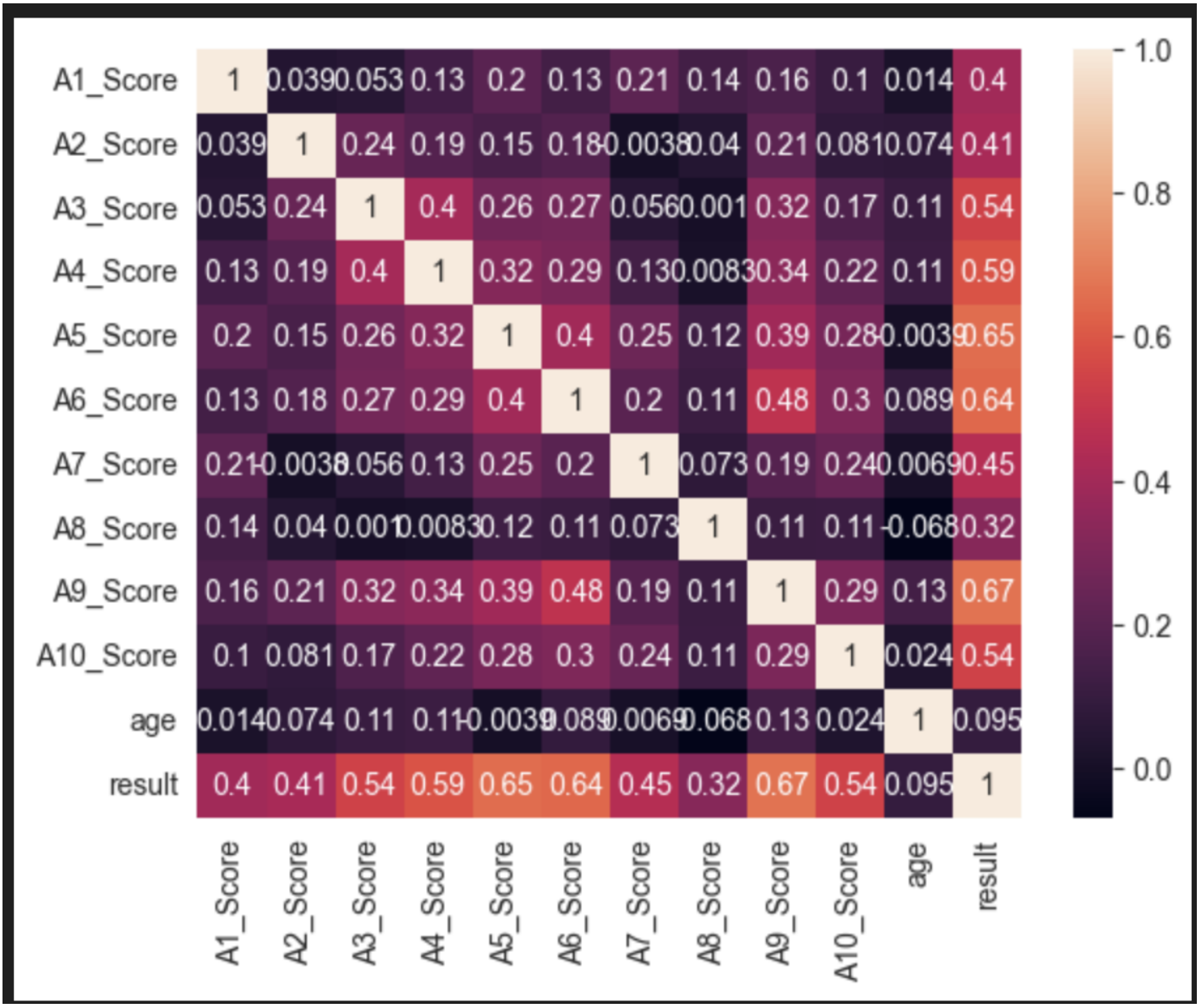
- Explore the dataset to understand its structure, variables, and distributions..
- Visualize relationships between variables using charts (e.g., histograms, scatter plots).:
- Handle missing data through imputation or deletion based on analysis.
- Transformation: Normalize or standardize numerical data to ensure uniformity and improve model performance..
- Feature engineering: Feature selection to identify relevant features and reduce dimensionality if necessary..
- Splitting: Conduct quality checks to ensure data integrity and consistency. data:

| | A1_Score | A2_Score | A3_Score | A4_Score | A5_Score | A6_Score | A7_Score | A8_Score | A9_Score | A10_Score | age | result |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 609.000000 | 609.000000 | 609.000000 | 609.000000 | 609.000000 | 609.000000 | 609.000000 | 609.000000 | 609.000000 | 609.000000 | 609.000000 | 609.000000 |
| mean | 0.740558 | 0.469622 | 0.481117 | 0.520525 | 0.525452 | 0.307061 | 0.428571 | 0.665025 | 0.341544 | 0.597701 | 30.215107 | 5.077176 |
| std | 0.438689 | 0.499487 | 0.500054 | 0.499989 | 0.499762 | 0.461654 | 0.495278 | 0.472370 | 0.474617 | 0.490765 | 17.287470 | 2.522717 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 17.000000 | 0.000000 |
| 25% | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 22.000000 | 3.000000 |
| 50% | 1.000000 | 0.000000 | 0.000000 | 1.000000 | 1.000000 | 0.000000 | 0.000000 | 1.000000 | 0.000000 | 1.000000 | 27.000000 | 5.000000 |
| 75% | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 35.000000 | 7.000000 |
| max | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 383.000000 | 10.000000 |

**Pairplot :**



**Heatmap**


A1_Score 1 0.039 0.053 0.13 0.2 0.13 0.21 0.14 0.16 0.1 0.014 0.4
A2_Score 0.039 1 0.24 0.19 0.15 0.18 0.0038 0.04 0.21 0.081 0.074 0.41
A3_Score 0.053 0.24 1 0.4 0.26 0.27 0.056 0.001 0.32 0.17 0.11 0.54
A4_Score 0.13 0.19 0.4 1 0.32 0.29 0.13 0.0083 0.34 0.22 0.11 0.59
A5_Score 0.2 0.15 0.26 0.32 1 0.4 0.25 0.12 0.39 0.28 0.0039 0.65
A6_Score 0.13 0.18 0.27 0.29 0.4 1 0.2 0.11 0.48 0.3 0.089 0.64
A7_Score 0.21 0.0038 0.056 0.13 0.25 0.2 1 0.073 0.19 0.24 0.0069 0.45
A8_Score 0.14 0.04 0.001 0.0083 0.12 0.11 0.073 1 0.11 0.11 -0.068 0.32
A9_Score 0.16 0.21 0.32 0.34 0.39 0.48 0.19 0.11 1 0.29 0.13 0.67
A10_Score 0.1 0.081 0.17 0.22 0.28 0.3 0.24 0.11 0.29 1 0.024 0.54
age 0.014 0.074 0.11 0.11 -0.0039 0.089 0.0069 0.068 0.13 0.024 1 0.095
result 0.4 0.41 0.54 0.59 0.65 0.64 0.45 0.32 0.67 0.54 0.095 1
A1_Score
A2_Score
A3_Score
A4_Score
A5_Score
A6_Score
A7_Score
A8_Score
A9_Score
A10_Score
age
result
1.0
0.8
0.6
0.4
0.2
0.0

Loading Data :

```
1  data=pd.read_csv("Autism_Data.arff")
```

```
1  data.head(10)
```

Handling Missing Data:

```
1 X=data_featured[['A1_Score', 'A2_Score', 'A3_Score', 'A4_Score', 'A5_Score', 'A6_Score',
2         'A7_Score', 'A8_Score', 'A9_Score', 'A10_Score', 'age', 'result', 'm',
3         'Had_jaundice_yes', 'Rel_had_yes']]
4 y=data_featured['Detected_YES']
```

```
Click to add a breakpoint e()

count    609.000000
mean       5.077176
std        2.522717
min        0.000000
25%        3.000000
50%        5.000000
75%        7.000000
max       10.000000
Name: result, dtype: float64
```

```
1  y

0       False
1       False
2        True
3       False
5        True
       ...
698      True
699      True
700     False
702     False
703      True
Name: Detected_YES, Length: 609, dtype: bool
```

4. Model Development Phase

a. Feature Selection Report:

| Feature | Description | Selected (Yes/No) | Reasoning |
|---|---|---|---|
| data_featured | Data features for feature selection | No | while feature selection identifies the most relevant subset for model optimization. |
| Gender | Applicant's gender | No | Relevant for assessing diversity and potential bias in disorder |
| Accuracy_lgr | Accuracy of Logistic Regression model | Yes | measuring the proportion of correctly predicted outcomes. |
| rel_autism | features associated with (ASD) | Yes | denotes the degree of relevance or correlation of factors associated with Autism Spectrum Disorder (ASD). |

b. Model Selection Report :

**Model Selection Report:**

| Model | Description | Hyperparameters | Performance Metric (e.g., Accuracy, F1 Score) |
|---|---|---|---|
| Random Forest | Ensemble of decision trees ;robust, handles complex relationships ,reduces overfitting,andprovides feature importance for disorder prediction | n_estimators, max_depth, max_features | Accuracy score = 100% |
| Decision Tree | Simple tree structure; ,interpretable, captures non-linear relationships, suitable for initial insights into disorder approval patterns. | Criterion, splitter, splitter | Accuracy score = 100% |
| **KNN** | **Classifies based on nearestneighbors; adapts well to data patterns, effective** | n_neighbors, weights, metric | **Accuracy score = 96.17** |

c. Initial Model Training Code, Model Validation and Evaluation Report

Training code :

```
sex=pd.get_dummies(data['gender'],drop_first=True)
jaund=pd.get_dummies(data['jundice'],drop_first=True,prefix="Had_jaundice")
rel_autism=pd.get_dummies(data['austim'],drop_first=True,prefix="Rel_had")
detected=pd.get_dummies(data['Class/ASD'],drop_first=True,prefix="Detected")
```

✓ 0.0s

```
data=data.drop(['gender','jundice','austim','Class/ASD'],axis=1)
data_featured=pd.concat([data,sex,jaund,rel_autism,detected],axis=1)
data_featured.head()
```

✓ 0.0s

| | A1_Score | A2_Score | A3_Score | A4_Score | A5_Score | A6_Score | A7_Score | A8_Score | A9_Score | A10_Score | age |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 26.0 |
| 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 24.0 |
| 2 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 27.0 |
| 3 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 35.0 |
| 5 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 36.0 |

```
sns.distplot(data_featured['age'],bins=50,kde=False)
```

✓ 0.1s

```
C:\Users\farde\AppData\Local\Temp\ipykernel_12116\219094534.py:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751
```

```
from sklearn.metrics import classification_report
```

✓ 0.0s

```
accuracy_lgr = accuracy_score(y_test,y_pred_lgr)
print('Accuracy LGR:', accuracy_lgr*100)
```

✓ 0.0s

```
Accuracy LGR: 100.0
```

```
print(classification_report(y_true=y_test,y_pred=pred))
```

✓ 0.0s

```
              precision    recall  f1-score   support

       False       1.00      1.00      1.00       132
        True       1.00      1.00      1.00        51

    accuracy                           1.00       183
   macro avg       1.00      1.00      1.00       183
weighted avg       1.00      1.00      1.00       183
```

Model Validation and Evaluation Report:

Random Forest Regressor :

```
predictionRF = rand_forest.predict(X_test)

print('Training set: ',rand_forest.score(X_train, y_train))
print('Testing set: ',rand_forest.score(X_test, y_test))
```

0.0s

```
Training set:  1.0
Testing set:  1.0
```

```
accuracy_RF=rand_forest.score(X_test, y_test)
print ("Accuracy_RF:",accuracy_RF*100)
```

0.0s

```
Accuracy_RF: 100.0
```

Linear

```
accuracy_lgr = accuracy_score(y_test,y_pred_lgr)
print('Accuracy LGR:', accuracy_lgr*100)
```

0.0s

```
Accuracy LGR: 100.0
```

5. **ModelOptimization and TuningPhase**

During the Model Optimization and Tuning phase for ASD detection, several crucial steps are undertaken to refine and improve the classification models. Hyperparameter tuning plays a pivotal role, involving the adjustment of parameters such as learning rates or tree depths to optimize model performance. Feature selection is another critical aspect, aiming to identify the most relevant features that contribute significantly to ASD prediction, thereby enhancing model efficiency and interpretability.

b. Final Model Selection Justification

Model : Random Forest Regressor

Reasoning :

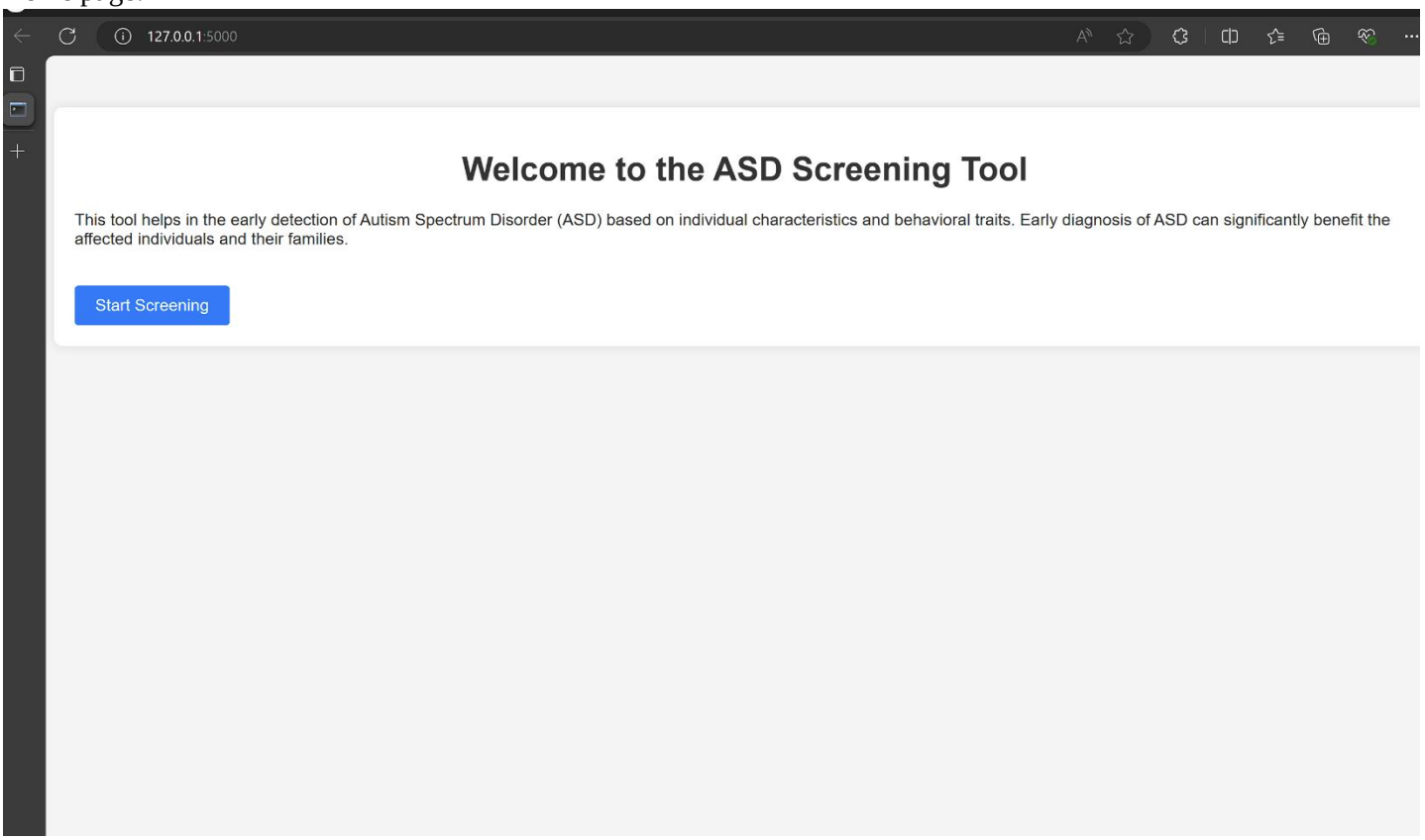
The Random Forest Regressor is a suitable choice for modeling Autism Spectrum Disorder (ASD) detection for several reasons. First, Random Forests are robust and versatile ensemble learning methods that excel in handling complex datasets with high-dimensional feature spaces. This is particularly advantageous in ASD detection, where diverse types of data such as behavioral observations, genetic markers, and neuroimaging results may be integrated.

Secondly, Random Forests are capable of capturing non-linear relationships and interactions between features, which is crucial in understanding the multifaceted nature of ASD. This helps in accurately predicting the continuum of ASD severity and its varied manifestations across individuals.

## 6. **Results**

a. Output Screenshots

Home page:


127.0.0.1:5000
Welcome to the ASD Screening Tool
This tool helps in the early detection of Autism Spectrum Disorder (ASD) based on individual characteristics and behavioral traits. Early diagnosis of ASD can significantly benefit the affected individuals and their families.
Start Screening

Login Page :

# ASD Screening

A1_Score:

A2_Score:

A3_Score:

A4_Score:

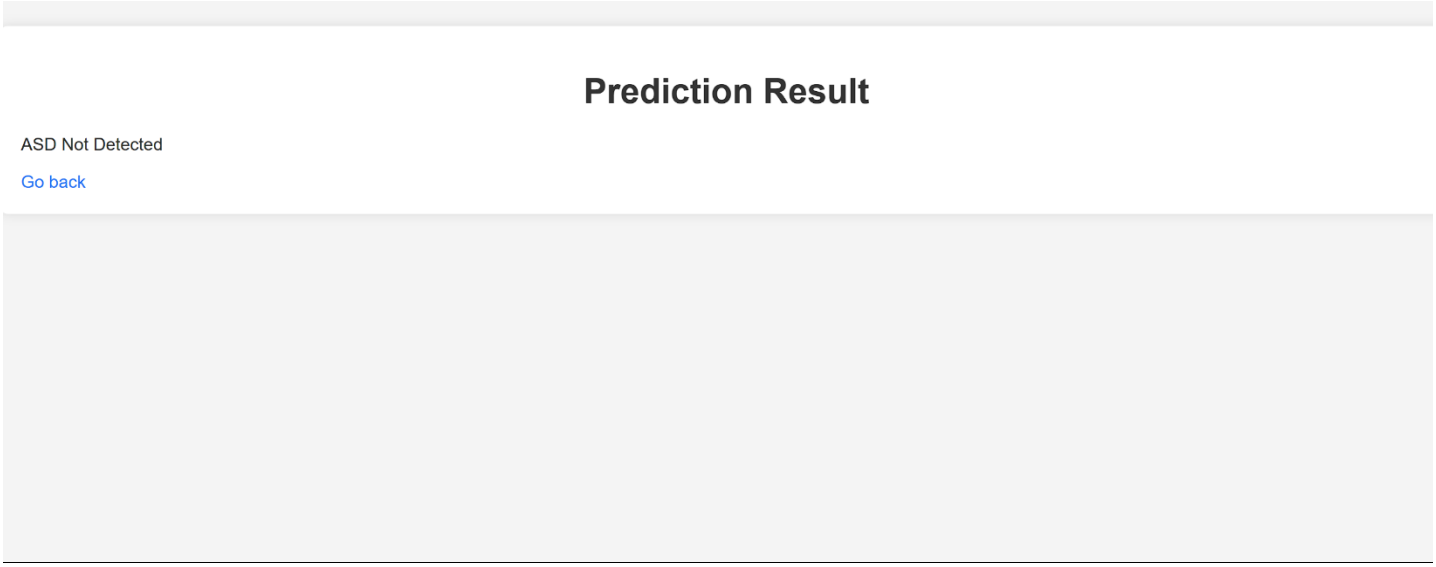
A5_Score:

A6_Score:

A7_Score:

Input :


Output :



7. **Advantages & Disadvantages:**

Advantages:

1. Provides objective assessment of ASD likelihood.
2. Potential for early detection and intervention.
3. Feature Importance: Integrates various types of data for improved accuracy

Disadvantages:

1. Models can be complex and challenging to interpret..
2. Accuracy heavily depends on the quality of input data

---

## Conclusion :

while classification techniques offer objective assessment and potential for early detection of ASD by integrating diverse data types, their complexity and dependence on data quality underscore the need for careful interpretation and validation to ensure accurate and ethically sound applications in clinical settings.

9. **Future Scope**

- Enhanced Accuracy: Improving models with larger and more diverse datasets to enhance accuracy and reduce biases..
- Personalized Medicine: Tailoring interventions based on individualized predictions from machine learning models.
- Integration of Biomarkers: Incorporating biomarkers from genetics, neuroimaging, and other fields to refine diagnostic precision.
- Real-time Monitoring: Developing systems for continuous monitoring and early detection in everyday environments..
- Ethical Considerations: Addressing privacy concerns and ensuring equitable access to diagnostic technologies across diverse populations.

## 10. **Appendix**

a. Source Code:

b. GitHub & Project Demo Link :

https://github.com/gauravf6/Detection-of-Autistic-Spectrum-Disorder-Classification.git