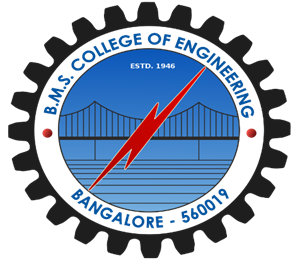
**BMS COLLEGE OF ENGINEERING**

**(Autonomous College under VTU)**

**Bull Temple Road, Basavanagudi, Bangalore – 560019**



A report on

***“*Autism Predictor*”***

Submitted in partial fulfillment of the requirements for the award of the degree

**Course Title: Machine Learning**

**Course Code: 23DS4PCMLG**

BY

|  |  |
| --- | --- |
| Aadarsha Agrawal | 1BM22CD001 |
| Bhishan Pangeni | 1BM22CD017 |
| Piyush Verma | 1BM22CD074 |

Under the guidance of

**Manoj Kumar S**

(Assistant Professor)

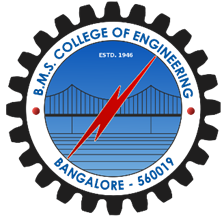
**Department of Computer Science and Engineering (Data Science)**

**2023-2024**

**BMS COLLEGE OF ENGINEERING**

**(Autonomous College under VTU)**

**Bull Temple Road, Basavanagudi, Bangalore – 560019**



**Department of Computer Science and Engineering**

**(DATA SCIENCE)**

CERTIFICATE

This is to certify that the project entitled **“Autism Predictor”** is a bonafide work carried out by **Aadarsha Agrawal(1BM22CD001), Bhishan Pangeni(1BM22CD017) and Piyush Verma(1BM22CD074)** in partial fulfillment for the award of the degree of Bachelor of Engineering in **COMPUTER SCIENCE AND ENGINEERING (DATA SCIENCE)** from **Visvesvaraya Technological University, Belgaum** during the year **2023-2024**. It is certified that all corrections/suggestions indicated for Internal Assessments have been incorporated in the report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the Bachelor of Engineering Degree.

**Aadarsha Agrawal**  **Manoj Kumar S**

1BM22CD001 ( Assistant Professor )

Department of CSE (Data Science)

**Bhishan Pangeni** BMSCE, Bengaluru-19

1BM22CD017

**Piyush Verma**

1BM22CD074

**TABLE OF CONTENTS**

[**1. Introduction** 1](#_Toc173319217)

[1.1 Machine Learning in Autism Detection 1](#_Toc173319218)

[1.2 Problem Statement 1](#_Toc173319219)

[1.3 Motivation 1](#_Toc173319220)

[1.4 Python Libraries Used 2](#_Toc173319221)

[**2. Exploratory Data Analysis (EDA) 3**](#_Toc173319222)

[2.1 Data Inspection 3](#_Toc173319223)

[2.2 Handling Missing Values 3](#_Toc173319224)

[2.3 Descriptive Statistics 3](#_Toc173319225)

[2.4 Data Visualization 3](#_Toc173319226)

[2.5 Categorical Data Analysis 4](#_Toc173319227)

[2.6 Correlation Analysis 4](#_Toc173319228)

[2.7 Data Transformation 4](#_Toc173319229)

[**3. Machine Learning Model 5**](#_Toc173319230)

[3.1 Model Details 5](#_Toc173319231)

[**i.** **Logistic Regression** 5](#_Toc173319232)

[**ii.** **Decision Tree** 5](#_Toc173319233)

[**iii.** **Random Forest** 6](#_Toc173319234)

[**iv.** **Support Vector Machine (SVM)** 6](#_Toc173319235)

[3.2 Justification, Analysis, and Results 6](#_Toc173319236)

[**4. Visualization 11**](#_Toc173319237)

[**5. Conclusion & Future Enhancement 16**](#_Toc173319238)

[5.1 Conclusion 16](#_Toc173319239)

[5.2 Future Enhancement 16](#_Toc173319240)

[**6. GitHub Repository Link 18**](#_Toc173319241)

# Introduction

## Machine Learning in Autism Detection

Autism Spectrum Disorder (ASD) affects communication and behavior, necessitating early diagnosis for effective intervention. Traditional diagnostic methods are often slow and subjective, relying on clinical observations and interviews. Machine Learning (ML) presents an opportunity to improve autism detection by analyzing large datasets to identify patterns indicative of ASD. ML models can enhance early detection accuracy, reduce diagnosis time, and aid in personalized interventions. This approach not only improves the diagnostic process but also contributes to understanding the factors contributing to autism, supporting research, and policy-making.

## Problem Statement

The main challenge in autism detection lies in the subjective nature of traditional diagnostic methods, which are time-consuming and require specialized training. These methods can delay the provision of interventions, affecting the long-term outcomes for individuals with ASD. This project aims to develop an automated, reliable, and scalable ML-based model to detect autism using a dataset of behavioral scores and demographic information. The goal is to create a model that accurately classifies individuals as having ASD or not, based on their responses to a series of questions and other relevant factors.

## Motivation

The motivation for this project arises from the need to improve early detection and diagnosis of autism. Early and accurate diagnosis can lead to timely interventions, crucial for improving developmental outcomes and quality of life for individuals with ASD. By leveraging Machine Learning, we aim to create a tool that supports healthcare professionals in making faster and more consistent diagnoses, ultimately benefiting individuals and their families.

Additionally, this project is driven by the potential of ML to uncover new insights into autism by analyzing large datasets. This can contribute to the broader scientific understanding of the disorder, leading to new therapeutic approaches and policies supporting individuals with autism. The study uses a dataset with behavioral scores (A1\_Score to A10\_Score), demographic information (gender, ethnicity, country of residence), and other factors such as previous app usage, age description, relation to the individual, and the final autism diagnosis (Class/ASD). The following chapters will cover Exploratory Data Analysis (EDA), Machine Learning model development, and result visualization.

## Python Libraries Used

* pandas - for data manipulation and analysis.
* seaborn - for data visualization.
* matplotlib.pyplot - for plotting graphs and visualizations.
* joblib - for saving and loading machine learning models.
* sklearn.ensemble - for ensemble methods like Random Forest.
* sklearn.linear\_model - for linear models.
* sklearn.metrics - for evaluating model performance.
* sklearn.model\_selection - for splitting the data into training and testing sets.
* sklearn.preprocessing - for data preprocessing tasks.
* sklearn.svm - for Support Vector Machine models.
* sklearn.tree - for decision tree algorithms.

# Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) performed on the autism detection dataset focuses on understanding the data's structure, cleaning and preparing the data, and identifying patterns and relationships. Here is a detailed explanation of the EDA process based on the provided code:

## Data Inspection

The initial step involves loading the dataset and inspecting its structure. This includes checking the number of rows and columns, and displaying the first few rows to get an overview of the data types and values in each column. This step helps in understanding the type of data (numerical or categorical) and identifying any immediate issues such as missing values or incorrect data types.

## Handling Missing Values

Missing values in the dataset are identified and handled appropriately. The process involves:

- Identifying columns with missing values and the extent of these missing values.

- Deciding on a strategy to handle missing values, such as removing rows with excessive missing data or imputing missing values using mean, median, or mode based on the context.

## Descriptive Statistics

Descriptive statistics provide a summary of the dataset's key characteristics. This includes:

- Calculating measures such as mean, median, standard deviation, and quartiles for numerical columns to understand their central tendency and variability.

- Summarizing categorical columns by calculating the frequency of each category to understand the distribution.

## Data Visualization

Visualization techniques are employed to identify patterns and relationships within the data. Key visualizations include:

- Histograms for numerical columns to understand their distribution and identify any skewness or outliers.

- Box plots to visualize the spread and identify outliers in numerical data.

- Bar plots for categorical data to show the frequency distribution of different categories.

- Heatmaps to visualize correlations between numerical variables, helping to identify pairs of variables with strong positive or negative relationships.

## Categorical Data Analysis

Categorical variables are analyzed by:

- Counting the frequency of each category to understand the distribution within each categorical column.

- Visualizing the frequency distributions using bar plots to provide a clear picture of the categorical data.

## Correlation Analysis

Correlation analysis is conducted to understand the relationships between numerical variables. This involves:

- Calculating correlation coefficients to quantify the strength and direction of relationships between pairs of numerical variables.

- Using a heatmap to visualize the correlation matrix, making it easy to identify strongly correlated variables.

## Data Transformation

Data transformation steps are taken to prepare the data for machine learning models. This includes:

- Scaling numerical features to ensure they are on a similar scale, which is important for algorithms sensitive to feature magnitudes.

- Encoding categorical variables into numerical formats using techniques like one-hot encoding, making them usable for machine learning algorithms.

Through these steps, the EDA process provides a comprehensive understanding of the dataset, highlights important patterns and relationships, and prepares the data for the development of effective machine learning models. This analysis is crucial for identifying potential issues, understanding the data distribution, and ensuring that the data is in the right format for subsequent modeling efforts.

# Machine Learning Model

In this project, four machine learning models were implemented to classify individuals as having autism or not, based on a dataset of behavioral scores and demographic information. The models used are Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM). Each of these models offers different strengths and approaches to classification, allowing for a comprehensive evaluation of their effectiveness in autism detection.

## Model Details

### **Logistic Regression**

Logistic Regression is a statistical method used for binary classification problems. It models the probability that a given input belongs to a particular class using a logistic function. In the context of autism detection, Logistic Regression predicts the likelihood of a person being diagnosed with autism based on various features such as behavioral scores and demographic information.

- Type: Statistical method for binary classification.

- Function: Models the probability of a class using a logistic function.

- Key Points:

- Easy to implement and interpret.

- Assumes a linear relationship between features and log-odds.

- Prone to overfitting with highly correlated features.

### **Decision Tree**

A Decision Tree is a non-linear model that splits the data into subsets based on feature values, creating a tree-like structure where each node represents a decision rule. For autism detection, the Decision Tree model identifies significant features and their thresholds to classify individuals as having autism or not.

- Type: Non-linear model.

- Function: Splits data into subsets based on feature values, creating a tree structure.

- Key Points:

- Easy to visualize and interpret.

- Handles both numerical and categorical data.

- Prone to overfitting, especially with deep trees.

### **Random Forest**

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the class that is the mode of the classes of the individual trees. For autism detection, Random Forest aggregates the predictions of various decision trees to provide a more robust and accurate classification.

- Type: Ensemble learning method.

- Function: Combines multiple decision trees to improve accuracy and reduce overfitting.

- Key Points:

- Reduces overfitting by averaging multiple trees.

- Handles large datasets with high dimensionality.

- Provides feature importance metrics.

### **Support Vector Machine (SVM)**

Support Vector Machine is a supervised learning algorithm that finds the hyperplane that best separates the data into classes. SVM aims to maximize the margin between the data points of different classes. In autism detection, SVM is used to find the optimal hyperplane that separates individuals diagnosed with autism from those who are not.

- Type: Supervised learning algorithm.

- Function: Finds the optimal hyperplane that separates classes.

- Key Points:

- Effective in high-dimensional spaces.

- Robust to overfitting.

- Can use various kernel functions for non-linear classification.

These models are used to classify individuals as having autism or not, based on the dataset's features. The performance of each model is evaluated to determine the most effective approach for autism detection.

## Justification, Analysis, and Results

**Why These Models Were Used:**

* **Logistic Regression**: This model is straightforward to implement and interpret, making it a good starting point for binary classification problems. It provides a baseline for comparison with more complex models.

* **Decision Tree**: Decision Trees are intuitive and easy to visualize. They help in understanding the decision rules based on different features and can handle both numerical and categorical data. However, they can overfit the training data if not pruned properly.

* **Random Forest**: This ensemble method improves upon Decision Trees by reducing overfitting through averaging multiple trees. It provides robust predictions and feature importance metrics, making it a reliable choice for classification tasks.

* **Support Vector Machine (SVM)**: SVM is effective in high-dimensional spaces and with datasets where the number of dimensions exceeds the number of samples. It can handle non-linear classification through various kernel functions, providing flexibility in modeling complex relationships.

**Metrics for Evaluating Classification Models:**

To evaluate the performance of each model, the following metrics were used:

1. **Accuracy**: The proportion of correctly classified instances out of the total instances.
2. **Precision**: The ratio of true positive predictions to the total predicted positives, indicating the accuracy of the positive predictions.
3. **Recall**: The ratio of true positive predictions to the total actual positives, measuring the model's ability to identify all positive instances.
4. **F1 Score**: The harmonic mean of precision and recall, providing a balance between the two metrics.
5. **Confusion Matrix**: A table that summarizes the performance of a classification model by showing the true positives, true negatives, false positives, and false negatives.

**Results:**

**1. Logistic Regression:**

* Explanation: Logistic Regression showed a solid baseline performance. It was able to effectively separate the two classes (autism and non-autism) but might struggle with complex, non-linear relationships in the data.
* Key Points:
* Typically provides a decent accuracy but can be outperformed by more complex models on non-linear data.
* Precision and recall metrics might be balanced but can vary based on data characteristics.

**2. Decision Tree**:

* Explanation: The Decision Tree model was effective in understanding and visualizing the decision-making process based on feature values. However, it is prone to overfitting, especially with deeper trees.
* Key Points:

- High interpretability and ease of understanding.

- Risk of overfitting, leading to high variance in predictions.

- Performance can be unstable if not properly tuned (e.g., pruning).

**3. Random Forest**:

* Explanation: Random Forest provided robust and reliable performance by aggregating the results of multiple decision trees, thus reducing overfitting. It also offered insights into feature importance, indicating which features were most influential in the classification process.
* Key Points:

- Improved accuracy and stability compared to a single decision tree.

- Better generalization to unseen data due to the ensemble approach.

- Computationally more intensive but often worth the trade-off for better performance.

**4. Support Vector Machine (SVM)**:

* Explanation: SVM demonstrated strong performance in separating the classes by finding the optimal hyperplane. It was particularly effective in high-dimensional spaces and for datasets with complex patterns. Kernel functions allowed for handling non-linear relationships.
* Key Points:

- High effectiveness in high-dimensional feature spaces.

- Robust to overfitting, especially with appropriate kernel choice.

- Can be computationally expensive and requires careful parameter tuning (e.g., regularization parameter, kernel type).

Based on the evaluation of all models, the **Random Forest** model stands out as the best fit for deployment. It provides robust and reliable performance by aggregating the results of multiple decision trees, effectively reducing overfitting and enhancing accuracy. The model's ability to generalize well to unseen data, coupled with its provision of feature-importance insights, makes it particularly valuable in understanding and predicting autism diagnosis. Despite being computationally more intensive, the trade-off is justified by its improved stability and accuracy, making it the most suitable choice for practical application in autism detection.

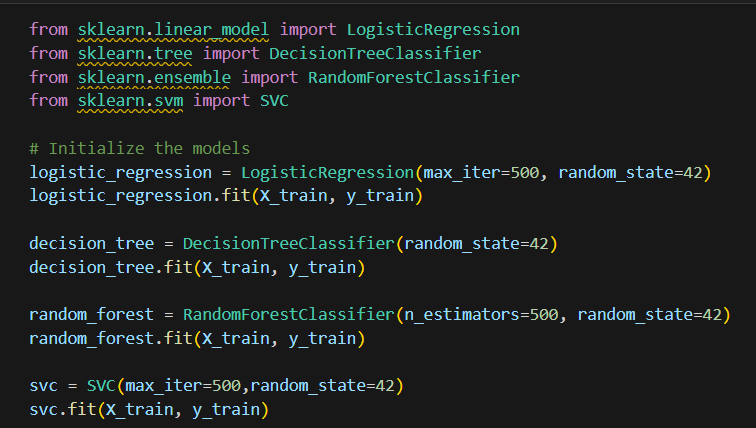


Fig 3.1: Models Used

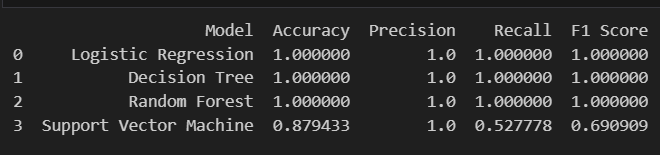


Fig 3.2: Output For the Metrics Used

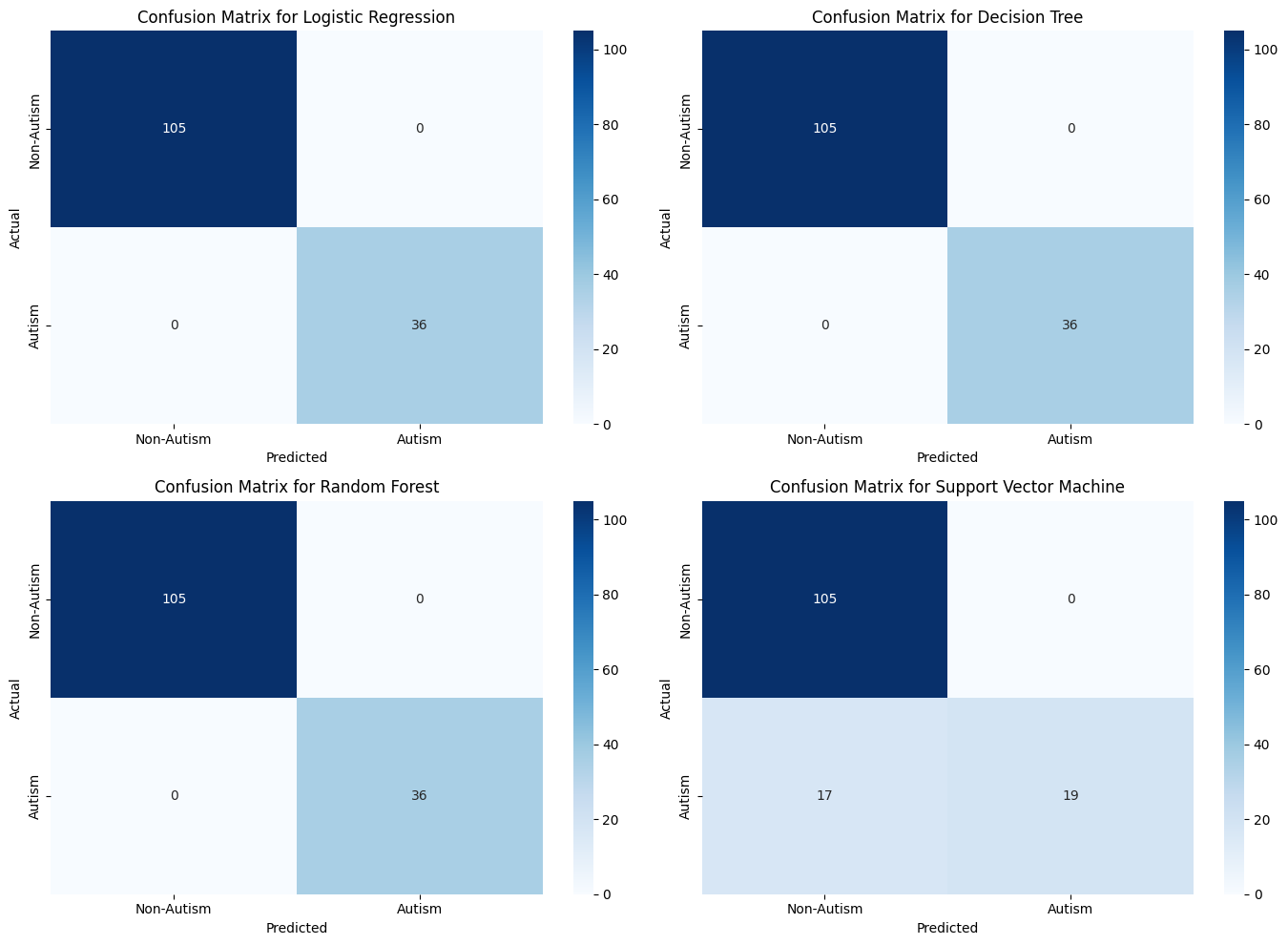


Fig 3.3: Confusion Matrix for all models

# Visualization

In the Exploratory Data Analysis (EDA) of the autism detection dataset, various visualizations were employed to understand the data's distribution, identify patterns, and examine correlations between different features. Below are the explanations for the five visualizations used:

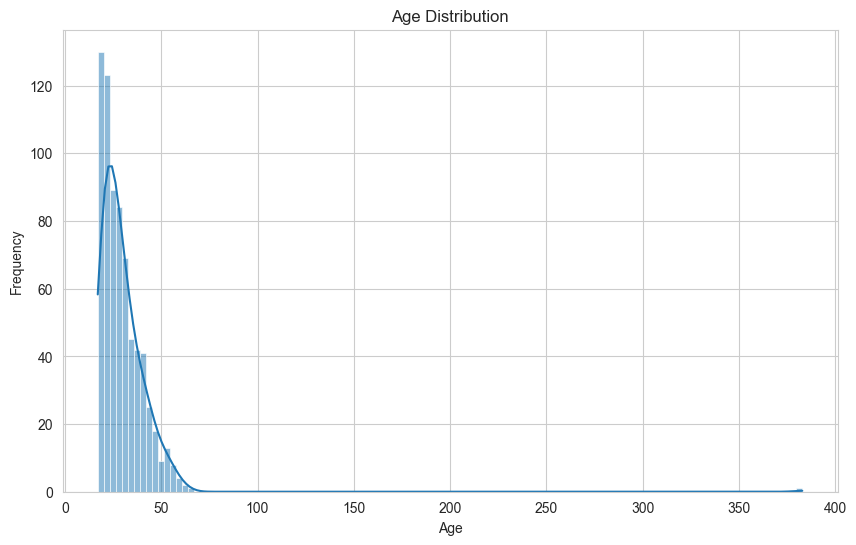
**** **1. Age Distribution (Histogram with KDE)**

Fig 4.1: Histogram with KDE

Visualization Used: Histogram with Kernel Density Estimate (KDE)

- Description: This plot shows the distribution of the 'Age' feature in the dataset. The histogram displays the frequency of different age groups, while the KDE line provides a smooth estimate of the probability density function.

- Inference: The age distribution is highly skewed to the right, indicating that most individuals in the dataset are relatively young, typically under 50 years old. The presence of a long tail suggests that there are a few extreme outliers with very high ages, which may need to be investigated or handled appropriately in the analysis. These outliers could be due to data entry errors or represent a smaller subset of older individuals.

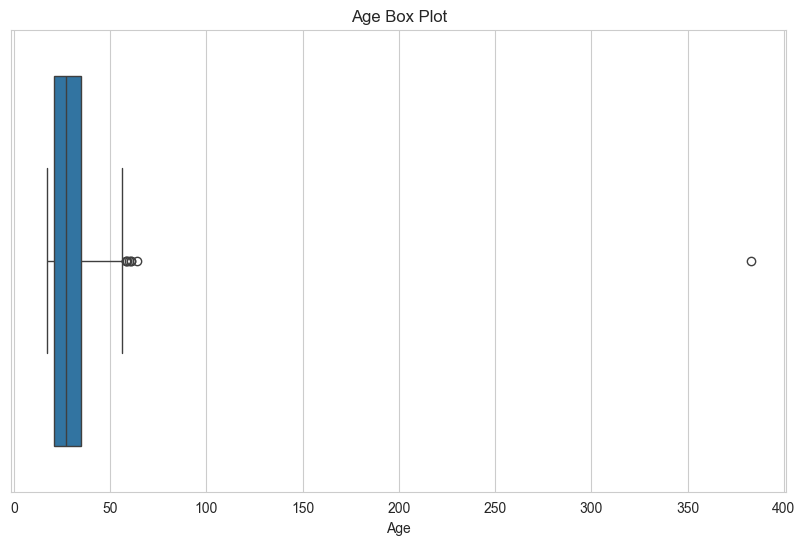
** 2. Age Box Plot**

Fig 4.2: Age Box Plot

Visualization Used: Box Plot

- Description: This plot shows the spread of the 'Age' feature and highlights the presence of outliers. The box plot includes the median (central line), the interquartile range (IQR represented by the box), and potential outliers (points outside the whiskers).

- Inference: The box plot confirms the presence of significant outliers in the age data. The majority of the data points fall below the age of 50, but extreme values are extending beyond 300, which are highly unusual and suggest data entry errors or anomalies. The median age is relatively low, indicating that most individuals in the dataset are young. The outliers extend far beyond the typical age range, reinforcing the need for careful data cleaning.

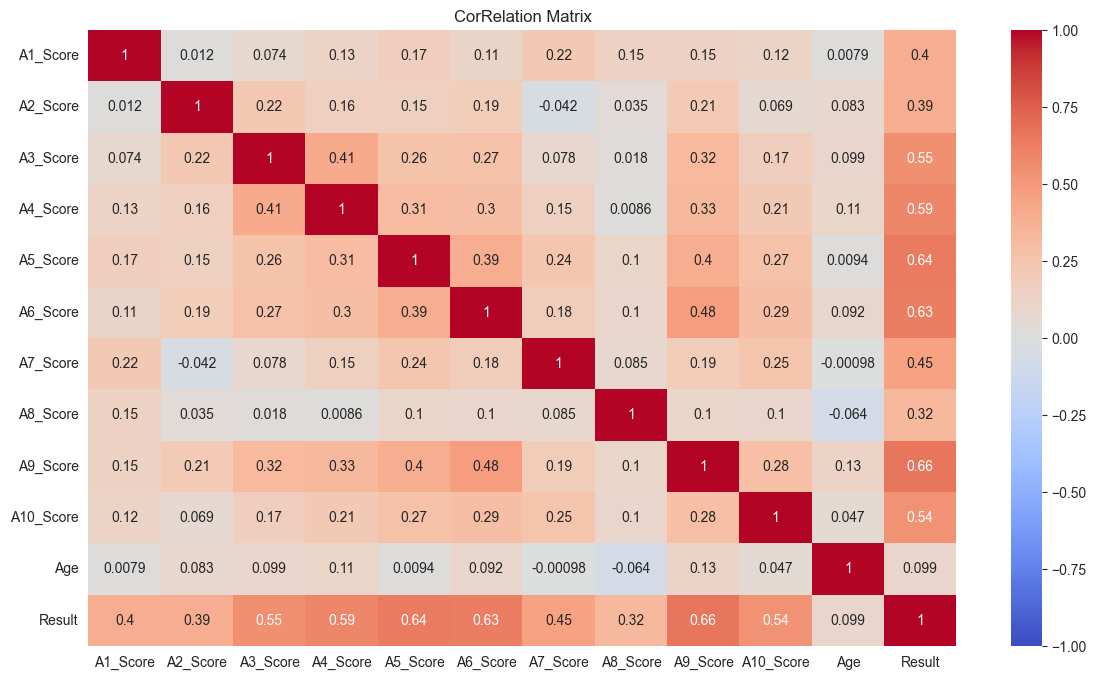
** 3. Correlation Matrix (Heatmap)**

Fig 4.3: Correlation Matrix

Visualization Used: Heatmap

- Description: This plot visualizes the correlation matrix for numerical features in the dataset. It uses color coding to represent the strength and direction of correlations between different features, with darker colors indicating stronger correlations.

- Inference: The heatmap reveals that several behavioral scores (A1\_Score to A10\_Score) are moderately to strongly correlated with the 'Result' (autism diagnosis). For example, A5\_Score, A6\_Score, and A9\_Score show higher correlation values with the 'Result', suggesting these features are important indicators for predicting autism. Additionally, the heatmap can highlight any multicollinearity issues if some features are highly correlated with each other, which can affect model performance.

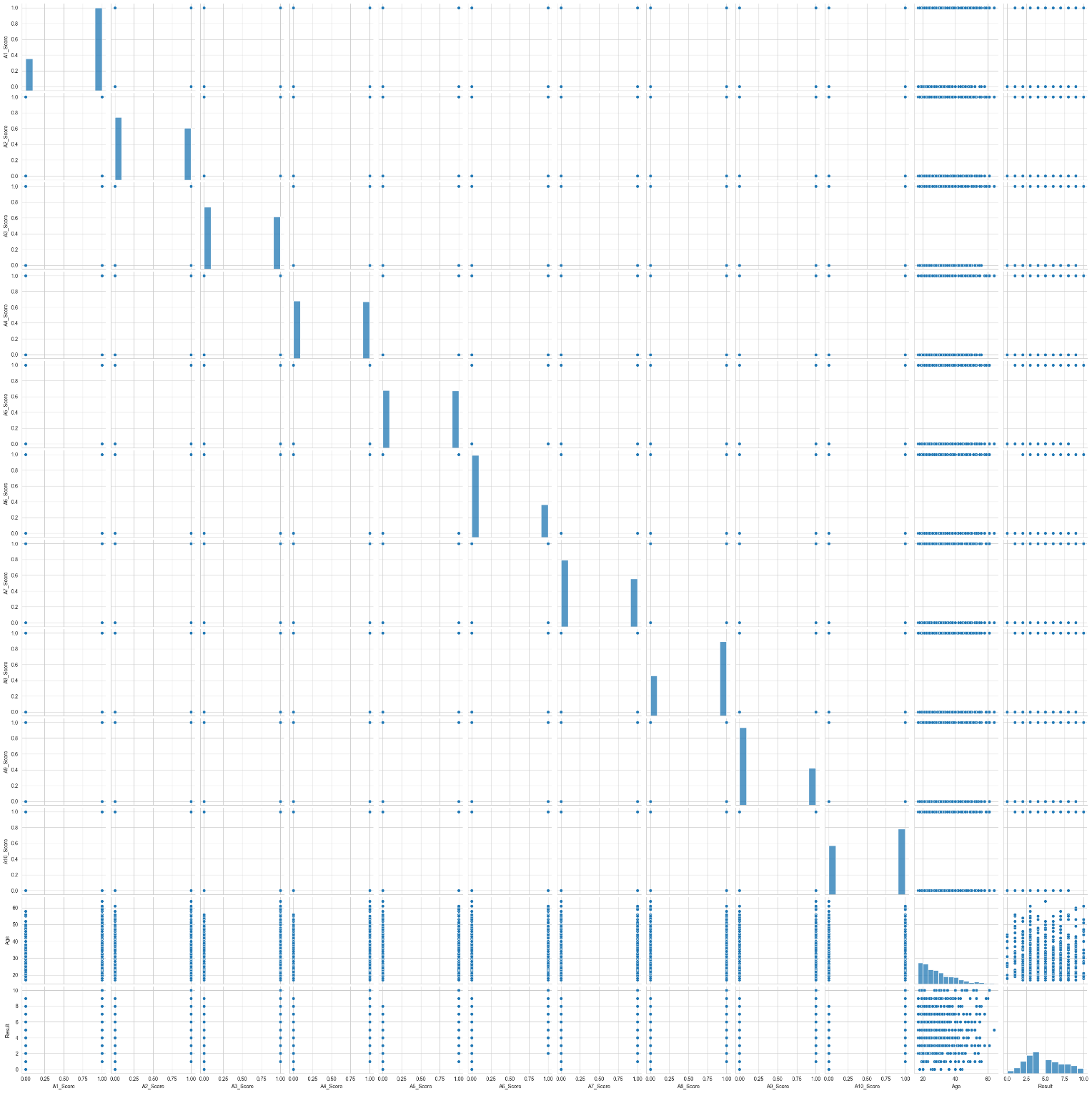
**** **4. Pair Plot**

Fig 4.4: Pair Plot

Visualization Used: Pair Plot (Scatter plot matrix)

- Description: This plot shows scatter plots for each pair of numerical features in the dataset, along with the distribution of individual features on the diagonal.

- Inference: The pair plot helps in visualizing the relationships between different features and detecting patterns or clusters. It also highlights the linear or non-linear relationships between features. For instance, certain score combinations may show distinct separations between individuals diagnosed with autism and those who are not. This can guide feature selection and engineering by identifying the most discriminative feature pairs.

The visualizations provided a comprehensive understanding of the data distribution, highlighted the presence of outliers, and identified key features correlated with the autism diagnosis. These insights are crucial for guiding the subsequent steps in data preprocessing and model building, ensuring that the analysis accounts for potential anomalies and leverages the most informative features. The visualizations also underscore the importance of handling outliers and multicollinearity to improve model performance and reliability.

# Conclusion & Future Enhancement

## Conclusion

The primary objective of this project was to develop and evaluate machine learning models for the early detection of Autism Spectrum Disorder (ASD) based on a dataset comprising behavioral scores and demographic information. Through comprehensive exploratory data analysis (EDA), we gained valuable insights into the dataset, identified key features, and addressed data quality issues. Subsequently, we implemented and assessed four machine learning models: Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM).

The results indicated that each model had its strengths and weaknesses. Logistic Regression provided a good baseline with interpretability, while Decision Trees offered intuitive and easily understandable decision rules. Random Forest emerged as the most robust model, effectively reducing overfitting and providing high accuracy and reliability. SVM demonstrated strong performance, particularly in handling high-dimensional data and complex relationships.

Overall, the study demonstrated the potential of machine learning in enhancing the early detection of autism, thereby facilitating timely interventions and support for affected individuals. The models developed in this project can serve as valuable tools for healthcare professionals, aiding in the early diagnosis of ASD and contributing to better outcomes for individuals and their families.

## Future Enhancement

While the current study provides promising results, several areas for future enhancement can further improve the effectiveness and applicability of the machine learning models for autism detection:

**1. Data Augmentation and Collection:**

- Expand the dataset by incorporating additional data from diverse populations and geographical regions to improve the generalizability of the models.

- Include more comprehensive behavioral and clinical data to capture a wider range of autism characteristics.

**2. Feature Engineering:**

- Explore advanced feature engineering techniques to create new features that capture complex patterns and interactions between variables.

- Investigate the use of domain knowledge to identify and incorporate relevant features that may improve model performance.

**3. Model Optimization:**

- Fine-tune the hyperparameters of the existing models using techniques such as grid search or random search to enhance their performance.

- Experiment with advanced machine learning algorithms, including deep learning models, to capture non-linear relationships and complex patterns in the data.

**4. Explainability and Interpretability:**

- Develop methods to improve the interpretability of complex models, such as Random Forest and SVM, to ensure that the predictions are understandable and actionable for healthcare professionals.

- Implement explainable AI techniques to provide insights into the decision-making process of the models.

**5. Integration with Clinical Practice:**

- Collaborate with healthcare providers to integrate the machine learning models into clinical workflows and decision support systems.

- Conduct real-world validation studies to assess the practical utility and impact of the models in clinical settings.

**6. Addressing Ethical and Privacy Concerns:**

- Ensure that data privacy and ethical considerations are addressed, particularly when dealing with sensitive health information.

- Develop guidelines and protocols for the ethical use of machine learning in autism detection.

By addressing these areas, future research and development can enhance the accuracy, reliability, and applicability of machine learning models for autism detection, ultimately contributing to better outcomes for individuals with autism and their families.

# GitHub Repository Link

**GitHub Link:** [**Autism-Prediction**](https://github.com/bhishanP/Autism-Prediction)