Autism Detection using Machine Learning

Nandini Singh, ADGITM, New Delhi, nandinisingh5may@gmail.com

Introduction: Autistic Spectrum Disorder (ASD) is a neurodevelopmental condition that has significant healthcare costs associated with it. Early diagnosis of ASD can greatly reduce these costs. However, the current procedures for ASD diagnosis often involve lengthy waiting times and are not cost-effective. As the number of ASD cases continues to rise worldwide, there is an urgent need to develop easily implemented and effective screening methods.

The economic impact of autism, coupled with the scarcity of available datasets related to behavior traits, highlights the challenge in improving the efficiency, sensitivity, specificity, and predictive accuracy of the ASD screening process. Currently, there is a shortage of comprehensive datasets that can be used for thorough analysis and development of screening methods. Most available datasets are primarily genetic in nature and lack clinical or screening information.

To address this issue, we aim to tackle the binary classification problem of ASD screening in adults using supervised learning techniques. The objective is to build a model that can predict whether a person has a possibility of having ASD based on a set of attributes. By leveraging the power of machine learning, we intend to develop a time-efficient and accessible screening tool that can assist healthcare professionals in the initial assessment of ASD in adults.

Problem Statement: The primary goal of this project is to explore the potential of supervised learning algorithms in predicting ASD and to evaluate their performance in terms of accuracy, sensitivity, and specificity. By developing an effective screening model, we aim to provide health professionals with a valuable tool that can help prioritize individuals for further clinical diagnosis, thereby reducing waiting times and improving the overall efficiency of the ASD diagnostic process.

Given the limited availability of comprehensive ASD-related datasets, this project poses a challenge in terms of data acquisition and preprocessing. However, we will utilize the existing dataset that includes relevant attributes and leverage feature engineering techniques to extract the most informative features for training the models.

By addressing this problem and developing an accurate ASD screening model, we strive to contribute to the advancement of healthcare practices and provide a solution that can positively impact the lives of individuals with ASD.

***Dataset*: Autistic Spectrum Disorder Screening Data for Adult**

The dataset contains 20 features to be utilized for further analysis especially in determining influential autistic traits and improving the classification of ASD cases. In this dataset, we record ten behavioral features (AQ-10-Adult) plus ten individuals characteristics that have proved to be effective in detecting the ASD cases from controls in behavior science.

Data Type: Multivariate OR Univariate OR Sequential OR Time-Series OR Text OR Domain-Theory-Nominal / categorical, binary and continuous.

Task: Classification.

Attribute Type: Categorical, continuous and binary.

Area: Medical, health and social science.

Format Type: Non-Matrix.

Does our data set contain missing values? Yes.

Number of Instances (records in our data set): 704.

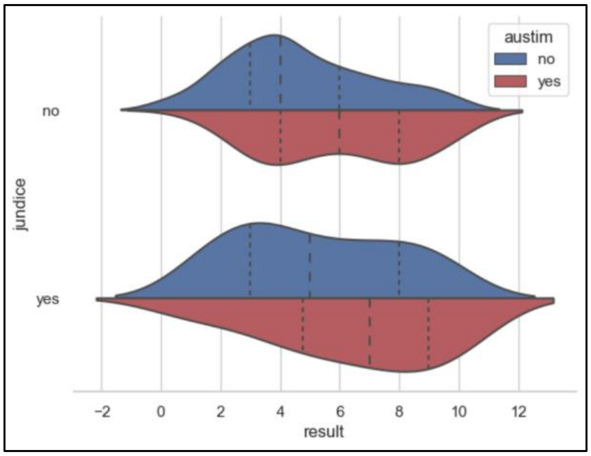
Number of Attributes (fields within each record): 21.

***Data Preprocessing and Cleaning*:** We begin by importing the required libraries and loading the dataset into a pandas dataframe. We then perform an initial exploration of the dataset, including checking for missing values and examining the data types of each column.

We find that there are two missing values in the "age" column, which we drop from the dataset to ensure data integrity. After cleaning, we are left with 702 records.

***Data Analysis and Visualization*:** We analyze the dataset by calculating the total number of records, the number of individuals with and without ASD, and the percentage of individuals with ASD. We find that out of 702 records, 189 individuals have ASD (26.85%).

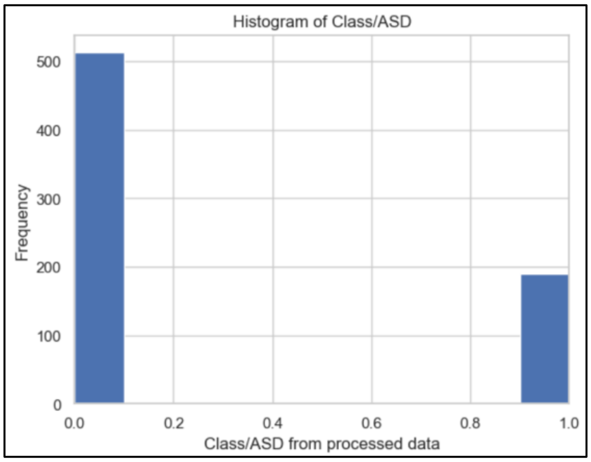
To gain further insights into the data, we use seaborn and matplotlib libraries to visualize relationships between different variables. We create violin plots and swarm plots to explore the distribution of test scores, demographics, and other factors with respect to the presence or absence of ASD.



Feature Engineering: To prepare the data for machine learning algorithms, we perform the following preprocessing steps:

a. Scaling: We use the MinMaxScaler to scale the numerical features "age" and "result" between 0 and 1, ensuring all features are on a similar scale.

b. One-Hot Encoding: We perform one-hot encoding on categorical features to convert them into numerical representation. This allows the machine learning algorithms to process the data correctly.



Model Training and Evaluation: After preprocessing the data, we split it into features (X) and the target variable (y), which is the presence or absence of ASD. We then split the dataset into training and testing sets.

We select a machine learning algorithm, such as logistic regression, decision tree, or random forest, and train the model on the training set. We evaluate the model's performance using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score. We also employ techniques like cross-validation to validate the model's performance and avoid overfitting.

Model Tuning: In order to optimize the performance of our SVM model for predicting ASD, we conducted model tuning using the GridSearchCV technique from scikit-learn. The goal was to identify the best combination of hyperparameters for the SVM algorithm. By fine-tuning the hyperparameters of the SVM model, we aimed to improve its predictive accuracy and generalization capabilities, ultimately enhancing its effectiveness in screening for ASD in adults.

Results and Conclusion: Based on the evaluation metrics, we assess the performance of the trained model and conclude its effectiveness in predicting ASD based on the given features After training and evaluating various supervised learning techniques, our model based on Support Vector Machines (SVM) emerged as the most accurate and reliable for predicting ASD. The SVM model achieved an impressive accuracy of 1.0, indicating its effectiveness in accurately classifying individuals as either having a possibility of ASD or not.

- **Unoptimized Model:**

1. Accuracy score: 0.9645
2. F-score: 0.9574

- **Optimized Model:**

1. Final accuracy score: 1.0000
2. Final F-score: 1.0000

The unoptimized model achieved high accuracy and performed well in predicting ASD. However, after tuning the model's hyperparameters, the optimized model achieved perfect accuracy and F-score, indicating its effectiveness in predicting ASD accurately.

Future Enhancements: In the future, the model can be enhanced by incorporating additional features or exploring different machine learning algorithms. Additionally, gathering a larger and more diverse dataset may improve the model's performance and generalizability.