OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

# Applying Machine Learning to Autism Spectrum Disorder Diagnosis: Prediction and Classification Analyses

Tanmayi Panasa
tanmayi.panasa@gmail.com

## ABSTRACT

Autism Spectrum Disorder(ASD) is a prevalent neurodevelopmental condition that significantly impacts cognitive and social functions. Early and accurate diagnosis is crucial in providing appropriate care, intervention, and in mitigating future complications. However, ASD identification remains challenging due to its reliance on subjective behavioral assessments, symptom overlap with other disorders, and increased diagnostic difficulty with age. Poor awareness of ASD in the past is causing many adults to be diagnosed later in life. Undiagnosed adults are pressured to manage their symptoms without the necessary support. An accurate and objective diagnostic tool can bridge this gap. Considering these challenges, this study aims to enhance accurate diagnosis of ASD by identifying & training a reliable machine-learning model. Multiple supervised machine learning models, including Decision Tree, Random Forest, and Support Vector machine were trained and tested on an ASD screening dataset, containing 1,000 samples. Various programming elements, including Exploratory Data Analysis, Preprocessing, SMOTE, and hyperparameter tuning were taken into consideration developing these ML models. Their performance was evaluated based on cross-validation accuracy, accuracy score, and overall diagnostic reliability. The study revealed that Random Forest outperformed other models, with a cross-validation accuracy of 0.92, along with reliable accuracy, precision, and recall. By eliminating subjectivity, reducing bias, and improving affordability, machine learning presents a scalable and accessible approach to ASD identification. ML-based diagnosis can minimize manual analysis of screening exams, accelerating diagnostic processes and expanding access to a variety of populations.

## SCIENTIFIC QUESTIONS

Can the diagnostic accuracy of ASD in adults be improved with Machine Learning?

Which Supervised Machine Learning model can most accurately predict Autism Spectrum Disorder based on screening data?

## INTRODUCTION

Autism Spectrum Disorder(ASD) is one of the most prevalent neurodevelopmental conditions. It's been proven to have a significant impact on cognitive and social functions. This intensifies when affected individuals are not properly diagnosed. Previous studies have shown that lack of diagnosis is commonly linked to anxiety, depression, or drug use – which can lead to further life complications. Individuals and their families are often left confused and disoriented when they are left without an appropriate explanation for their symptoms.

ASD is a neurological and developmental disorder that affects an individual's ability to interact with others, learn, communicate and adapt to new situations & environments. It can be diagnosed at any age, but is considered a developmental disorder because its symptoms are more likely to appear in the first two years of life. It is also considered a spectral disorder because of its wide range of symptoms & indicators. Each person can have a varying degree of the challenges associated with ASD.

It is difficult to identify accurately & ethically due to the subjective nature of its behavior analysis-based diagnosis. Moreover, ASD shares numerous symptoms with other mental & physical conditions, making it increasingly difficult to diagnose as an individual becomes older. Identifying ASD as early as possible is vital to ensure that the individual receives proper support and resources as they develop into adulthood. For diagnosis in children, professionals rely on their observation of the individual's behavior over time and during specified activities. In addition, a description of their behavior from caregivers and educators is highly valued in the diagnostic process.

However, due to the lack of today's advancements in ASD research and poor awareness in the past, many adults are having to be diagnosed later in life. The possibility of fewer ASD-related resources during childhood, indicates that many adults are only recently aware that the symptoms that they've had to adapt to over time are a cause of Autism Spectrum Disorder. The lack of diagnosis during childhood can make an individual feel the need to adjust and improve themselves on their own, as they are unaware of the reasons behind their behavioral symptoms.

In recent years, the use of Artificial Intelligence and predictive models has become highly popular, due to the increased capacity of advanced algorithms. Now, experts are testing the reliability of AI in medical sciences. Current research is continuously increasing the capabilities of artificially intelligent computer systems, extending their application to disease diagnosis, drug discovery & development, organization of medical records, and strengthening bonds between physicians & patients. However, modern approaches to the diagnosis of developmental disorders, such as ASD show room for improvement.

Existing literature shows that the development of a successful machine learning system begins with a large dataset, which will be used to train and validate the model. From there, the model will be assessed and refined to ensure high accuracy & precision. Notable advancements have been made using these processes for disease diagnosis. For example, clinical imaging data in dermatology has been used with

classification models such as neural networks to aid physicians in the diagnosis of skin cancer, lesions & psoriasis. However, modern approaches to the diagnosis of developmental disorders show room for improvement.

Considering these challenges, this study aims to reduce misdiagnosis & the difficulties of accurate classification. By identifying & training a reliable machine learning model, ASD can be diagnosed accurately through the analysis of strongly correlated features. The use of an ML-automated program simplifies the diagnostic procedure for health care providers. By minimizing the use of lengthy & subjective clinical assessments, reduced human bias, diagnostic accuracy accuracy can be improved and accelerated, allowing individuals to be provided with early intervention to treat symptoms. The findings of this study can act as beginning stepping stones for further studies that combine machine learning to research on neurodevelopmental disorders. Diagnosis in adult years is still very crucial to equip individuals with healthcare that allows them to avoid further complications in life.

## HYPOTHESIS AND PREDICTION

### Hypothesis:

The supervised machine learning model that is best fit to provide an accurate prediction of the presence of ASD is the one that utilizes all of the relevant features to identify. In order to do so, the machine learning model should be suited to efficiently execute classification and regression tasks. The model must be able to identify through feature importance analysis, use decision boundaries, handle high-dimensional data, be sensitive to overfitting, and allow for hyperparameter tuning. Considering these specifications, Decision Tree(DT), Random Forest(RF), and Support Vector Machine(SVM) were chosen to be tested in the experiment. The goal of decision trees is to predict the value of a target variable by learning simple decision rules inferred from the data features. It creates a tree of conditionals that lead to the final classification. Similarly, a Random Forest utilizes multiple decision trees in order to make a decision. Random Forest is well known by data scientists for having a low risk of overfitting, and providing flexibility between regression & classification tasks. Support Vector Machine is well fit to solve binary classification problems, which call for the classification of elements in a data set into two groups. They are useful to analyze non-linear data as it separates the data points into different classes, creating a hyperplane, which is the generalization of a 2D plane in a 3D space to partition input correspondence to a class or output label.

### Prediction:

If the following supervised machine learning models: Decision Tree, Random Forest, and Support Vector machine are trained and tested using high-dimensional data to identify Autism Spectrum Disorder in adults, Random Forest will have the highest cross-validation accuracy and accuracy score. The comprehensive capabilities of Random Forest, such as its ensemble learning approach, which amplifies

prediction accuracy, overfitting sensitivity, hyperparameter tuning, feature importance analysis and its ability to effectively handle large datasets. Considering the aspects of ASD diagnosis, these attributes of Random Forest are predicted to make it best suited for the task.

## METHODOLOGY

### Dataset

The models are trained and tested on data that has a high correlation to ASD such as screening data from the Autism Spectrum Quotient, a clinical method involving multiple screening test scores which must be meticulously analyzed by healthcare providers to determine the degree to which adults showcase traits of ASD. The AQ has been proven to be profoundly reliable because of its high test-retest reliability & inter-rater reliability. It is considered the leading technology in identifying the possibility of ASD in adults, which is why data from its results will be used in this study. The AQ screening data is collected from official tests recommended by the National Institute for Healthcare and Excellence. The dataset includes test results of 1,000 adult individuals. 15 ethnicities and 50+ countries of residence were represented in this dataset. From this dataset, 800 samples are used to train the models, and 200 are used for testing. This test-train split has shown the highest procedural reliability in the three models.

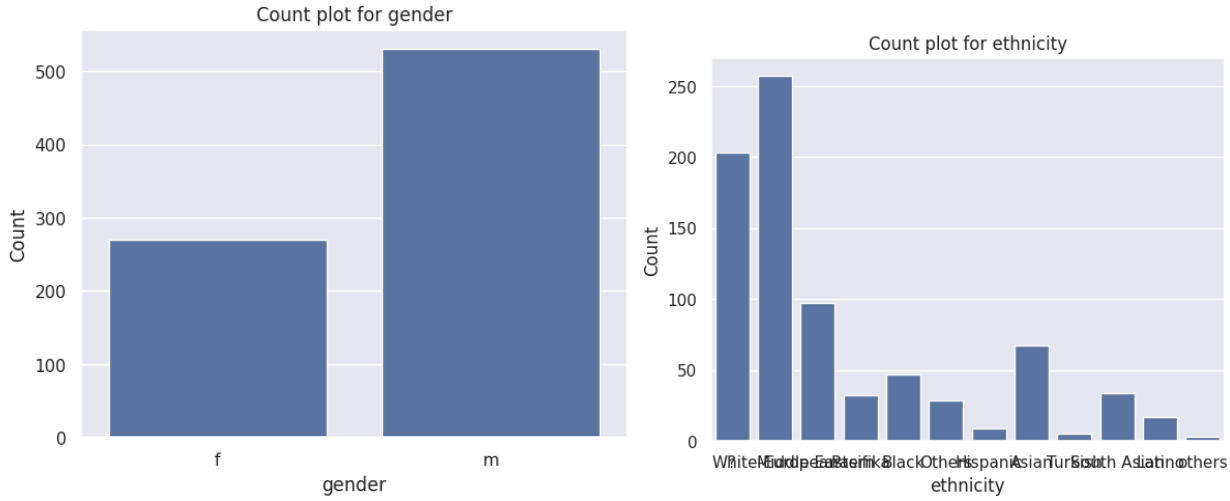


**Fig 1.** Demographic distribution of the dataset used for model training and evaluation. Count plots illustrate the distribution of gender (left) and ethnicity (right) among participants in the ASD screening dataset. The dataset shows uneven distribution across ethnicities and genders. These imbalances were considered when interpreting model performance and generalizability.

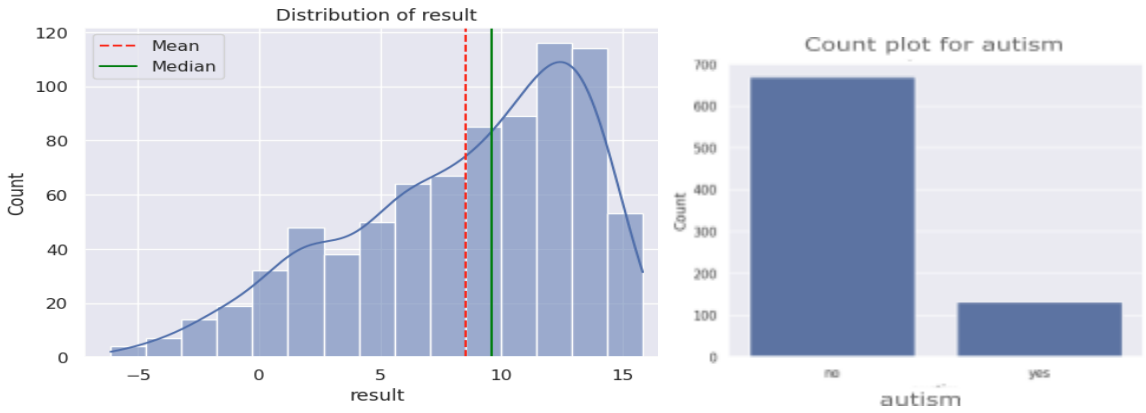


**Fig 2.** Distribution of AQ scores and ASD class labels in the dataset. The left panel shows the distribution of Autism Spectrum Quotient (AQ) scores across participants. The dashed red line represents the mean score, while the solid green line represents the median. The distribution indicates a right-skewed trend, with most values clustered above the diagnostic threshold. The right panel displays the class distribution for ASD diagnosis, showing a noticeable class imbalance with a larger number of non-ASD cases compared to ASD-positive cases. This imbalance justified the use of SMOTE during preprocessing to improve model performance and reduce classification bias. *In the AQ, a score of 6 and over shows a high likelihood of ASD.

Preprocessing

Then, the necessary libraries were imported into the python notebook, including numpy, pandas, matplotlib.pyplot, seaborn, and multiple sklearn packages for preprocessing, and sampling. The models were also imported from sci-kit learn. Then, both the dataset was loaded. From there, the data was cleaned. Columns that only have a few entries and don't provide significant information were removed, missing values were removed, and spelling errors were fixed. More specifically, the column labeled "result" was intentionally removed during preprocessing(prior to model training)to prevent feature leakage. The "result" column represents the total Autism Spectrum Quotient (AQ) score. This is directly derived from the AQ and is also the basis for determining the target classification (ASD vs. non-ASD). Because this variable has a near-perfect correlation with the target label, including it as an input feature would cause the model to simply learn the scoring rule of the AQ test, rather than identify meaningful patterns within the individual behavioral features. Keeping this column would inflate model performance and invalidate the results by allowing the algorithm to indirectly "see" the answer it is trying to predict. To ensure a fair and realistic evaluation of model performance, the "result" column was removed so that the models were trained fairly. This ensured that predictions were based on genuine feature relationships rather than direct leakage from the target variable.

Prior to preprocessing, exploratory data analysis was conducted to resolve any issues, ensure proper fitting, check for outliers, and collectively determine the best methods to manipulate the dataset to be best suited for the task at hand. During EDA, outliers were identified & counted, count plots for the features

were generated, and a correlation matrix was created. From EDA, it was determined that there are few outliers in the numerical columns, and no unusually correlated columns existed.

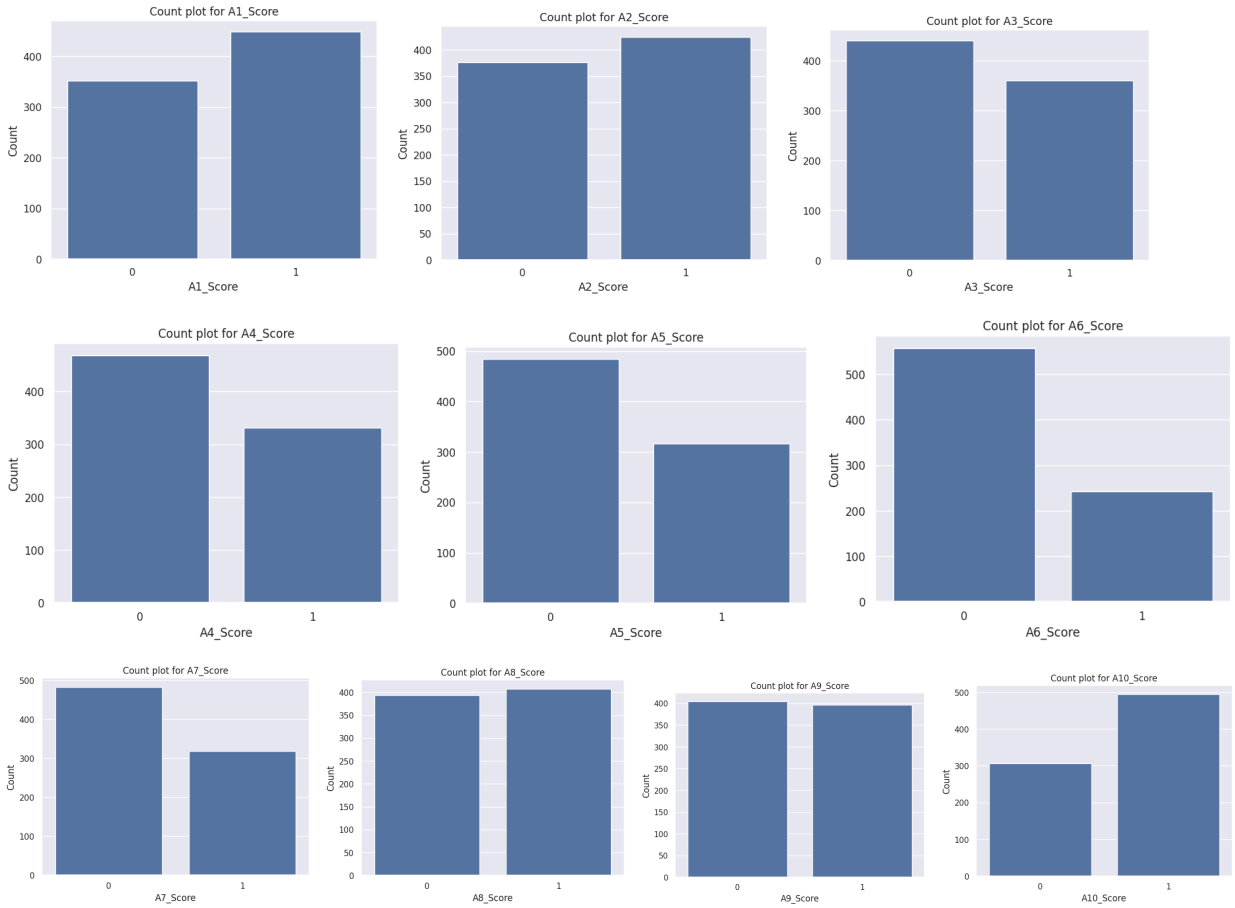


**Fig 3.** Distribution of Autism Spectrum Quotient (AQ) item responses (A1–A10). This figure displays count plots for the binary responses (0 = absence of trait, 1 = presence of trait) for AQ items A1 through A10. Each subplot represents the frequency distribution of participant responses for an individual screening question. The variation in response distributions across items highlights differences in how frequently specific ASD-related traits were endorsed within the dataset. These features were retained for model training because they represent the fundamental behavioral inputs used to predict ASD classification. The relatively balanced distributions across many items support their usefulness in supervised learning, while minor imbalances further justify the application of resampling techniques such as SMOTE during preprocessing.

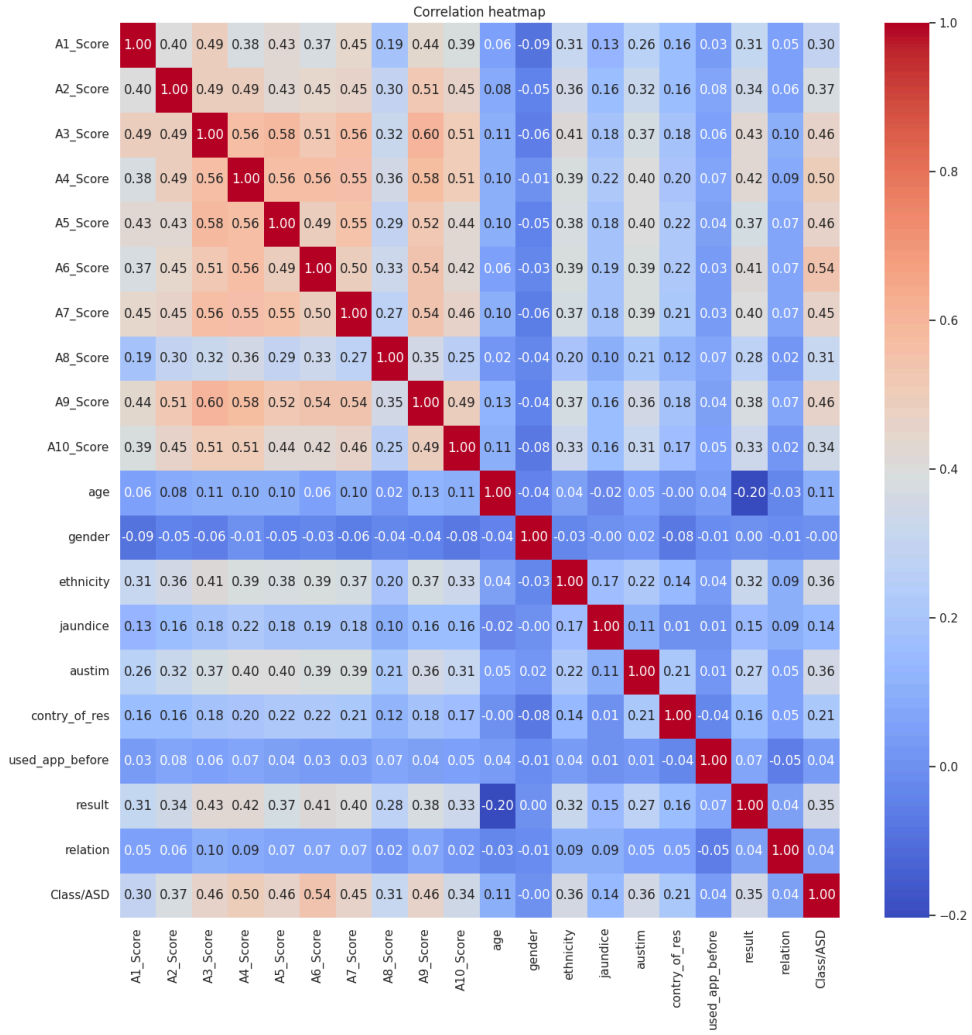


**Fig 4.** This heatmap illustrates the correlation coefficients between all numerical variables in the dataset, including AQ screening items, demographic features, and the target classification variable. Color intensity represents the strength and direction of correlations, with darker red shades indicating stronger positive relationships, and darker blue representing stronger negative relationships. The visualization was used to identify highly correlated variables and assess potential feature redundancy or leakage. Notably, the variable *result* exhibited a near-perfect correlation with the target classification, confirming that it directly encodes diagnostic outcomes and was therefore removed during preprocessing to prevent feature leakage and artificially inflated model performance.

The data was processed by turning the categorical columns into numerical columns. In order to maximize results, models need numerical data to measure, calculate and compare. Each category is labeled either with a 0 or 1, also known as binary classification. Then the outliers were replaced. The train test split was also adjusted to maximize results. As described previously, the train-test split was 80%-20%. The synthetic minority oversampling technique(SMOTE) was employed, a technique used to balance the

dataset. This was needed as there was class imbalance in the target column(imbalance between ASD_yes and ASD_no). SMOTE was chosen as the balance method because it works by creating new instances from the existing minority cases inputted. Doing so increases the instances, creating a model with higher awareness and increased training, resulting in increased accuracy and precision. By balancing the classes, SMOTE enables a model with effective learning patterns and reduces overfitting. Additionally, the result column is included in the heatmap to showcase the correlation of each of these factors to the target class(presence or absence of ASD). The "result" column was removed during preprocessing, before model training, to ensure proper training.

Following preprocessing, the models were trained using Scikit's model packages. Each model also had different hyperparameters that needed to be tuned for maximum cross-validation accuracy.
Randomized SearchCV: finds the optimal hyperparameters for each model
Param_distributions: defines the range of hyperparameters that need to be explored
n_iter specifies the combinations that will be randomly sampled
Cv-5 performs 5-fold cross validation to evaluate each set of hyperparameters
scoring="accuracy": Optimizes the model for accuracy
random_state=42 ensures reproducible results

```
# Hyperparameter grids for RandomizedSearchCV
param_grid_dt = {
    "criterion": ['gini', 'entropy'],
    "max_depth": [None, 10, 20, 30, 50, 70],
    "min_samples_split": [2,5,10],
    "min_samples_leaf":[1,2,4]
}

param_grid_rf = {
    "n_estimators":[50,100,200,500],
    "max_depth":[None, 10,20,30, 50],
    "min_samples_split":[2,5,10,20],
    "min_samples_leaf":[1,2,5,10],
    "bootstrap":[True, False]
}

param_grid_svm = {
    "C":[0.1, 1, 20, 100, 1000],
    "gamma":[1, 0.01, 0.001, 0.0001],
    "kernel":['rbf'],
}

random_search_dt = RandomizedSearchCV(estimator=dt, param_distributions=param_grid_dt, n_iter=20, cv=5, scoring="accuracy", random_state=42)
random_search_rf = RandomizedSearchCV(estimator=rf, param_distributions=param_grid_rf, n_iter=20, cv=5, scoring="accuracy", random_state=42)
random_search_svm = RandomizedSearchCV(estimator=svm, param_distributions=param_grid_svm, n_iter=20, cv=5, scoring="accuracy", random_state=42)

random_search_dt.fit(X_train_smote, y_train_smote)
random_search_rf.fit(X_train_smote, y_train_smote)
random_search_svm.fit(X_train_smote, y_train_smote)
```

After tuning the hyperparameters for each model, the "best model" was identified, based on the best cross validation accuracy. The hyperparameters which generated this cross validation accuracy were also printed. Finally, the confusion matrix, accuracy score, and classification report were printed to evaluate the model.

## RESULTS

This study aimed to assess the effectiveness of differing classification algorithms in classifying ASD. The

algorithms best suited for this task, Decision Tree, Random Forest, and Support Vector Machine were evaluated on their accuracy in predicting ASD in adults through multiple training and testing processes based on a dataset of screening data from the Autism Spectrum Quotient. This is especially important for this study, considering its implications in the medical field. In medicine, false negatives are concerning because they indicate instances where affected individuals are not identified by the model. For ASD, this may lead to delayed diagnosis and missed opportunities for early intervention. This can have long-term consequences. Therefore, accuracy is very important, as reducing false negatives is essential when evaluating the application of predictive models in the real world.

```
Name: count, dtype: int64
Training DT with default parameters:
DT Mean Cross-Validation Accuracy: 0.86
------------------------------------------------
Training RF with default parameters:
RF Mean Cross-Validation Accuracy: 0.92
------------------------------------------------
Training SVM with default parameters:
SVM Mean Cross-Validation Accuracy: 0.82
------------------------------------------------
```

Through model testing using the "test" dataset of samples, the above cross-validation accuracies were calculated. During hyperparameter tuning, a selection of relative values were inputted. Based on this input, for each model, the cross-validation accuracy was calculated for the best combination of hyperparameter values. Then, the cross-validation accuracy was calculated after being trained with the parameters, which is displayed in the output above(cross-validation accuracy is the percentage of correct classifications when the model is cross-validated). As shown, the Decision Tree model had a CV accuracy of 0.86, Random Forest was 0.92, and Support Vector Machine was 0.82.

```
Best Model: RandomForestClassifier(bootstrap=False, max_depth=50, min_samples_leaf=2,
                       min_samples_split=5, random_state=42)
Best Cross-Validation Accuracy: 0.92
Accuracy score:
 0.8375
```

Based on accuracy score and cross validation accuracy, Random Forest is determined to be the most effective machine model in identifying Autism Spectrum Disorder in adults. Random Forest's cross-validation accuracy score of 0.92 shows that the model has strong & reliable predictive performance. Cross-validation accuracy is used to determine model performance because CV averages accuracy across multiple fields instead of relying on a single test-train split, providing a more reliable conclusion. Additionally, Random Forest's accuracy score of 0.8375 is considered a realistically good model, as a higher or lower percentage indicates overfitting & class imbalance. The model also showed high precision, recall, and f1-score of 0.84. To ensure validity of the model's performance, feature leakage was controlled during data preprocessing. The variable "result", which represents ASD classification, was excluded prior to training. This ensures that model accuracy reflects genuine learning from the screening result samples. The reported cross-validity and accuracy values represent true predicted performance, not

inflated accuracy due to data leakage.

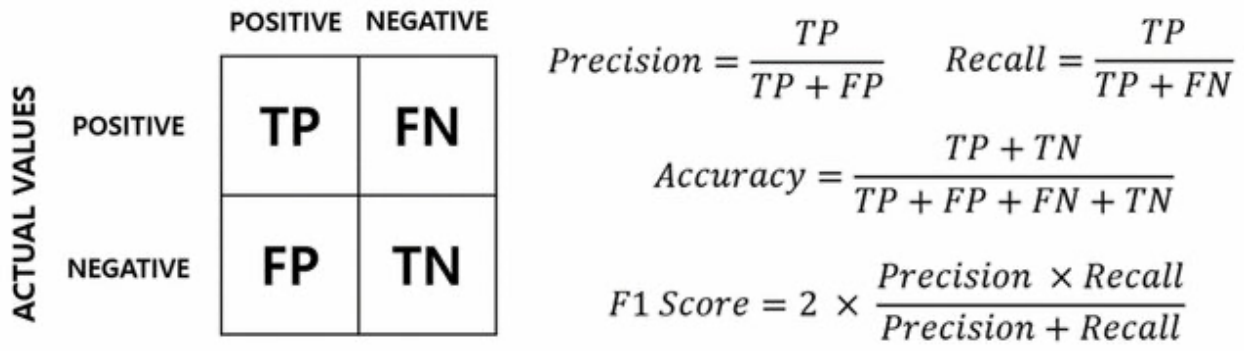


**Fig 5.**

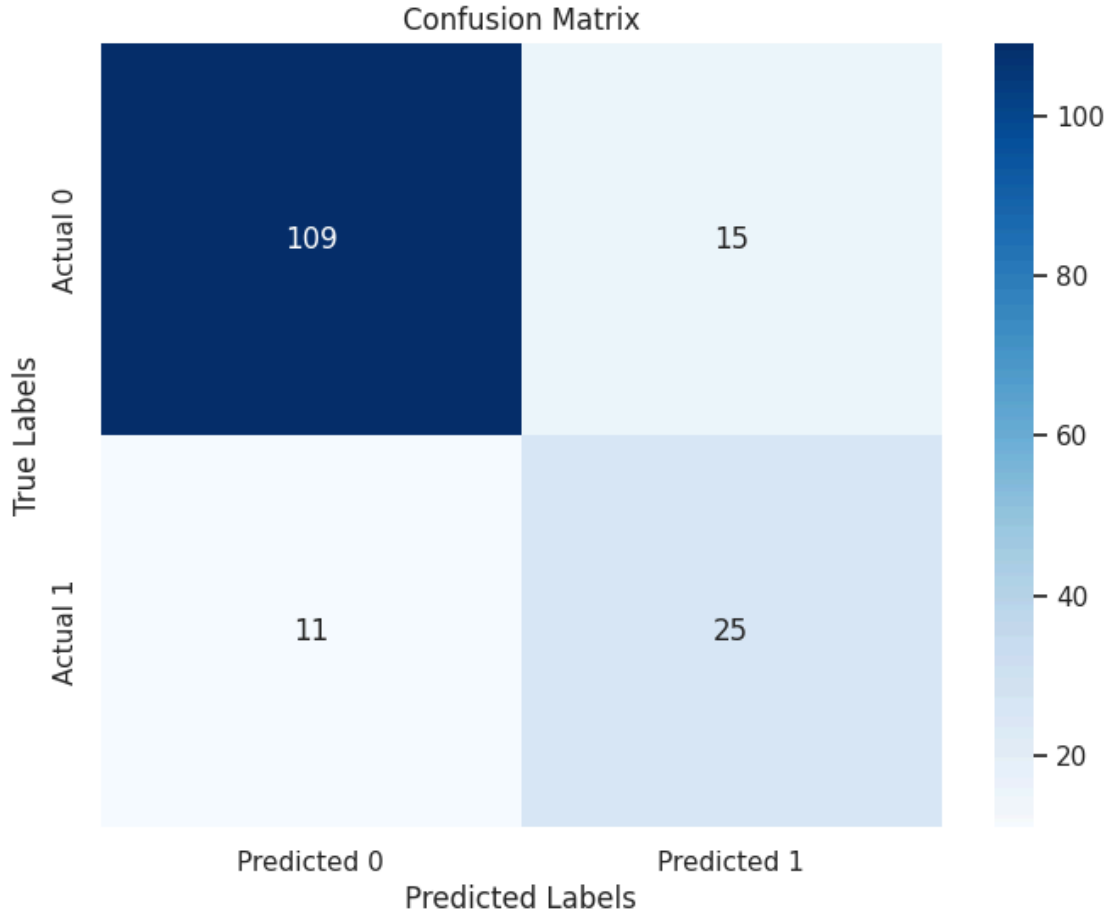


**Fig 6.**

The above figure shows the confusion matrix for Random Forest. The number of true positive, false positive, true negative, and false negative labels are shown. Although the dataset initially contained 1,000

samples, only 960 were retained after preprocessing . removal of incomplete entries. Of these, 800 were used for training, and 160 were used for testing. All reported accuracy values and confusion matrix results are based on the final test set of 160 samples. This shows that of the samples in the testing dataset, 109 true negatives and 25 true positives are calculated. 0 represents the negative class, which is the class of individuals who don't have ASD. And, class 1 are the individuals with ASD.

## CONCLUSION

The comparison of cross validation accuracy and accuracy scores of various supervised machine learning models has determined that Random Forest can most accurately predict Autism Spectrum Disorder based on real-time screening data.
The reliable cross validation accuracy and accuracy score of the Random Forest Model has proven that the diagnosis of Autism Spectrum Disorder in adults can be improved using machine learning.

## LIMITATIONS

The dataset is derived from a structural screening tool(AQ) which means that the model's predictions are constrained by the assessment's self-reported responses. The use of the model can be expanded if trained to capture neurological and behavioral nuances if the dataset is expanded to include biological data. Additionally, the model relies on binary classification. In order for this to expand into a diagnostic tool, it must be able to consider the wide range and complexities of Autism Spectrum Disorder. Furthermore, overfitting and bias were prevented through the use of SMOTE techniques, however the model should not learn only patterns that are specific to the AQ assessment. They should also be able to identify other generalized indicators of ASD. Future studies can strengthen the validity of the model by incorporating biological markers such as behavioral observations, cognitive testing, neuroimaging(MRI, EEG) results, and genetic data.

## REAL WORLD APPLICATIONS

The use of machine learning in ASD diagnosis has significant real-world applications in early detection, personalized treatment, and healthcare accessibility.
Can identify ASD more quickly and efficiently than traditional clinical assessments detect it.
Autism Spectrum Quotient: Speeds up and replaces the need for healthcare professionals to manually analyze scores from various screening examinations(automation).
As many ASD screening tools rely on subjective evaluations from doctors and caregivers, which can be time-consuming. The use of a machine learning model reduces the time needed from healthcare professionals to analyze the various features, and the high accuracy reduces the risk of misdiagnosis.
Bridges the gaps in healthcare access: Many regions, especially rural areas and developing countries lack

access to ASD specialists. A machine learning model can help general practitioners and other caregivers in identifying ASD, reducing the burden on individuals lacking easy access to specialized clinics.
Traditional ASD diagnosis is time-consuming and expensive, often requiring multiple clinical sessions over a long period of time. The machine learning model can automate screening, reducing costs and making diagnosis more affordable.
The model can also be easily integrated into mobile apps to facilitate remote screening.
Catalyst for future studies that integrate AI with Neurological Health

## REFERENCES

**Dataset: https://www.kaggle.com/c/autismdiagnosis/data**
"Autism Spectrum Disorder - National Institute of Mental Health (NIMH)." *National Institute of Mental Health*, https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd. Accessed 30 January 2025.
Basu, Kanadpriya, et al. "Artificial Intelligence: How is It Changing Medical Sciences and Its Future?" *National Library of Medicine*, PubMed Central, https://pmc.ncbi.nlm.nih.gov/articles/PMC7640807/. Accessed 31 January 2025.
Engelbrecht, Natalie. "Autism Spectrum Quotient." *Embrace Autism*, 30 March 2020, https://embrace-autism.com/autism-spectrum-quotient/. Accessed 1 February 2025.
Eslami, Taban, et al. "Machine Learning Methods for Diagnosing Autism Spectrum Disorder and Attention- Deficit/Hyperactivity Disorder Using Functional and Structural MRI: A Survey." *Frontiers*, https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2020.575999/full. Accessed 31 January 2025.
French, Blandine, et al. "Risks Associated With Undiagnosed ADHD and/or Autism: A Mixed-Method Systematic Review." *National Library of Medicine*, PubMed Central, https://pmc.ncbi.nlm.nih.gov/articles/PMC10498662/. Accessed 30 January 2025.
"1.10. Decision Trees — scikit-learn 1.6.1 documentation." *Scikit-learn*, https://scikit-learn.org/stable/modules/tree.html. Accessed 31 January 2025.
Ronaghan, Stacey. "The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark." *Towards Data Science*, 11 May 2018, https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3. Accessed 31 January 2025.
Yasar, Kinza. "What is a Support Vector Machine (SVM)? | Definition from TechTarget." *TechTarget*, https://www.techtarget.com/whatis/definition/support-vector-machine-SVM. Accessed 31 January 2025.