Chapter 4: Data Analysis and Machine Learning in ASD Diagnosis

Autism Spectrum Disorder (ASD) is progressively dependent on machine-learning approaches to enhance the accuracy of diagnosing the disorder at an early age and explaining its natural complexities. To conduct the current study, a dataset acquired at Kaggle is used to test the predictive potential of various machine-learning algorithms in relation to ASD-related characteristics, considering both demographic, clinical, and behavioral factors. The above dataset, which combines both categorical and numerical variables, was analyzed using a thorough exploratory data analysis (EDA) and then underwent a systematic pre-processing process. After these preparation activities, several machine-learning models, including Decision Trees, Random Forests, and Support Vector Classifiers, among others, were instantiated and their performance was thoroughly tested. The results show that the Random Forest and AdaBoost implementations are considered the most accurate and precise, and thus may be employed to predict ASD traits.

4.1 Dataset and Exploratory Data Analysis

The data (https://www.kaggle.com/datasets/uppulurimadhuri/dataset/data) used in this study includes a substantial sample of demographic, clinical, and behavioural variables that will be gathered on a total population of 1,074 individuals suspected of having autism spectrum disorder characteristics. One of the key goals of the exploratory data analysis was to identify any potential patterns and correlations in the data that could serve as informative predictors of ASD characteristics. The preliminary examination revealed a strong gender bias in the dataset.
Figure 1 shows the distribution of gender in the dataset, illustrating the significant male predominance.

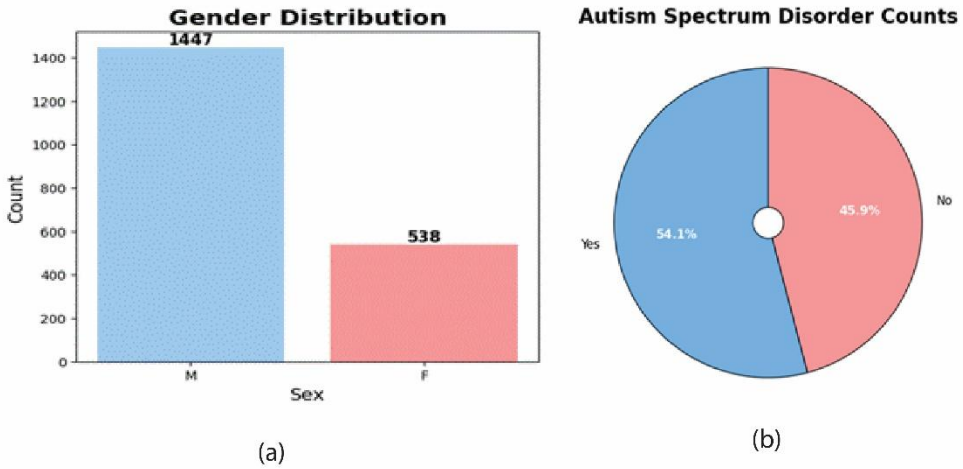

Figure 1: he distribution of gender in the dataset, illustrating the significant male predominance

The cohort under analysis included persons who were evaluated in terms of ASD. Descriptive statistics showed that there were some major imbalances in demographics. Among the entire sample of ASD cases that have been examined, 963 persons were classified as male, or 89.7 percent of the whole sample, and only 111 persons were female, or 10.3 percent. Figure 2 shows the Gender Distribution of ASD Patients (N = 1074).

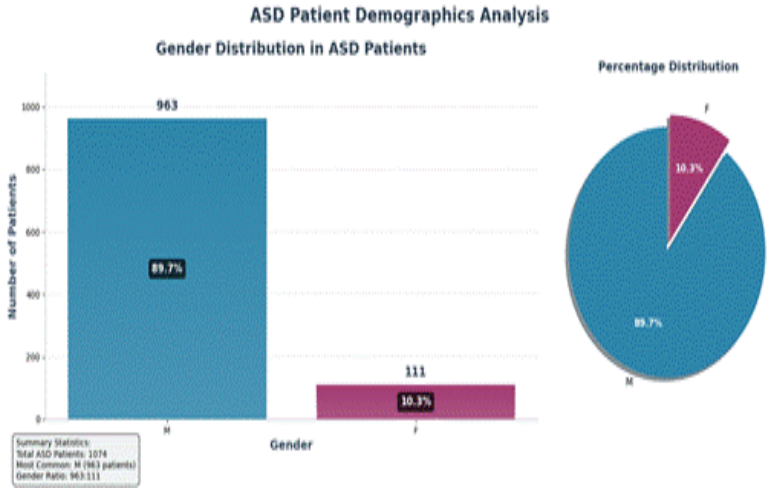

Figure 2: Gender Distribution of ASD Patients (N = 1074).

The result of this distribution is a strong male-to-female gender ratio of about 8.7:1 in the population under study, which is significantly higher than the generally quoted epidemiological estimates of ASD prevalence (typically between 2:1 and 4:1). 1 This ratio implies that the sample is a specialized clinical referral sample or a sample chosen to represent individuals with more severe and classic phenotypes, which are historically associated with male phenotypes. Such a consistent male dominance in ASD research has been noted to be primarily a source of debate in relation to etiological theories. 2 One, notable theory, the Extreme Male Brain (EMB) theory, holds that ASD is an extreme expression of typical male cognitive processes, which may be mediated by biological factors, including fetal testosterone levels. The Social Responsiveness Scale (SRS) Scores in a histogram.

A histogram showing the frequency of scores of the participants on the Social Responsiveness Scale (SRS). The textual analysis that comes with it revolves around the high prevalence of Jaundice (77% positive: 1,536 Yes vs. 449 No). Figure 3 presents a bar chart depicting the distribution of jaundice across the sample population. Figure 4 shows distribution of Social Responsiveness Scale (SRS) Scores.

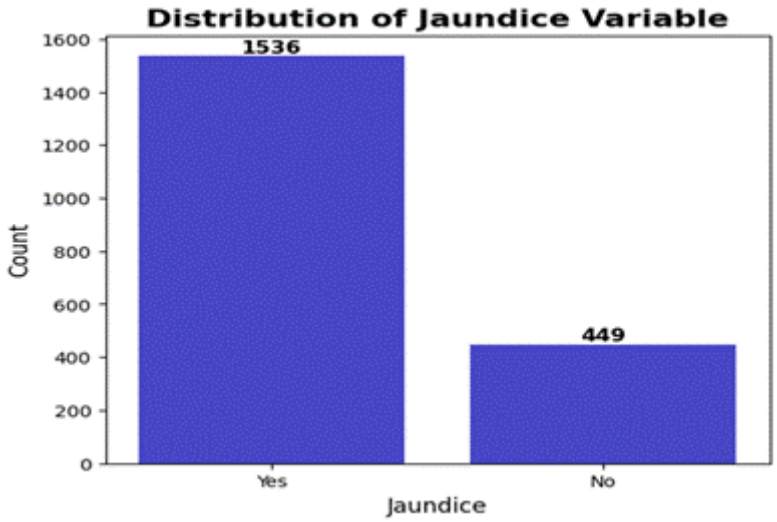

Figure 3: A bar chart depicting the distribution of jaundice across the sample population.

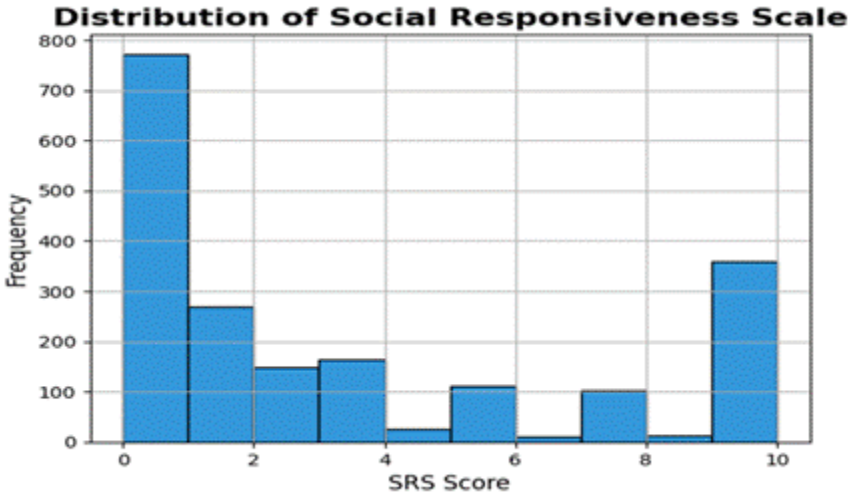

Figure 4: Distribution of Social Responsiveness Scale (SRS) Scores

Anxiety Disorder is also a common comorbidity, and its distribution was quantified. The data indicate the presence of a comparatively equal proportion of participants who experience an anxiety disorder (1,051 individuals reported that they have an anxiety disorder) and who do not experience an anxiety disorder (934 individuals reported that they do not have an anxiety disorder). Such an approximately equal distribution highlights that anxiety disorders are among the most prevalent in cohorts of ASD characteristics, and it can be stated that it is an important clinical factor that should be incorporated into predictive outcomes. Figure 5 visualizes the distribution of anxiety disorders within the sample population.

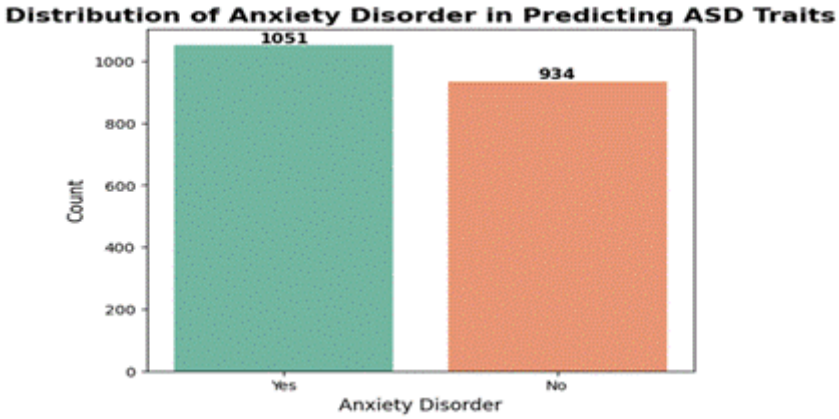

Figure 5: Visualizes the distribution of anxiety disorders within the sample population.

The scatter plot shows age against scores on the Childhood Autism Rating Scale (CARS) among the participants. Each of the points is a representation of a person, and it demonstrates the changing severity of their autism over age. Through the plot, we can see possible patterns, such as whether the higher CARS scores are more prevalent in some age groups or the scores are generally evenly distributed across age groups. In general, it gives a visual insight into the potential of age to be associated with the severity of autism in the population under study. Figure 6 shows the scatter plot illustrating the relationship between age and CARS scores, indicating no clear correlation between age and severity. Figure 7 represents comparison of Qchat-10 Scores between Male and Female Patients.

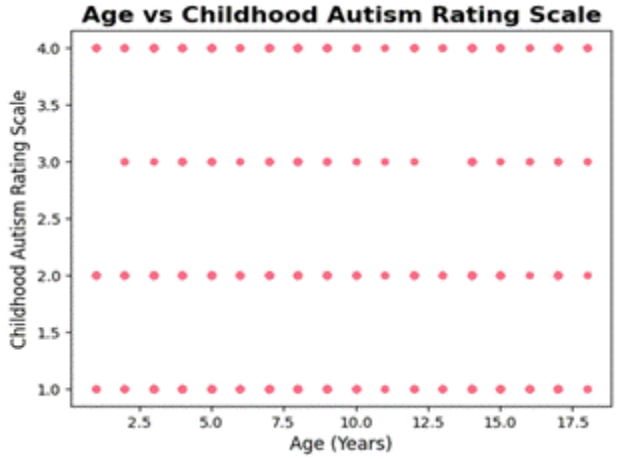

Figure 6: The scatter plot illustrating the relationship between age and CARS scores, indicating no clear correlation between age and severity.

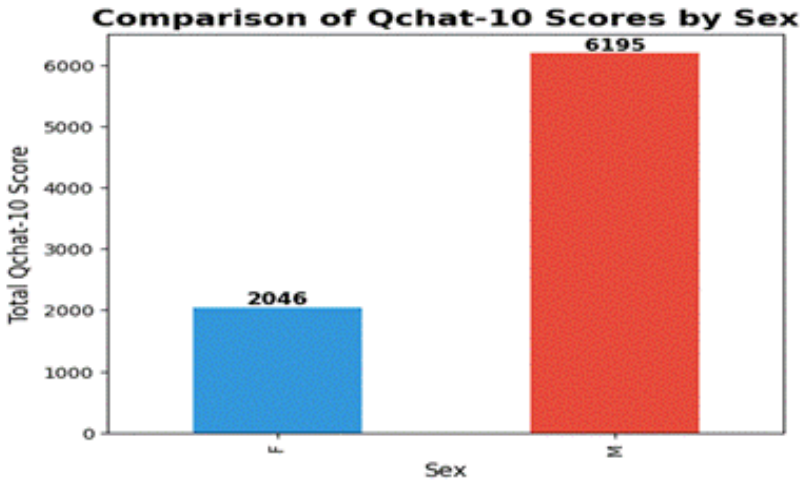

Figure 7: Comparison of Qchat-10 Scores Between Male and Female Patients.

The comparison between sex and the QCHAT-10 score (a questionnaire that is used to screen children at a young age about any signs of early autism) can be further analyzed using a boxplot or strip plot to examine how there might be gender differences in early-autism screening. Since the dataset is largely male-dominated, the plot shows whether the male and female participants differ in their patterns of QCHAT-10 scores. In most cases, males are generally more likely to score higher in QCHAT-10, and this can mean that they are more autistic at an earlier age, which is consistent with the diagnostic patterns where males are more commonly diagnosed with ASD. These results, however, should be interpreted with the possibility of gender bias in the event of early diagnosis. The plot assists in picturing how gender can be played out with screening scores indicating the importance of considering gender variance in predictive models of ASD.

4.2 Ethnicity vs. Autism Spectrum Quotient (AQ)

The difference between ethnic traits in different groups of people with autism was measured using an ethnicity vs Autism Spectrum Quotient (AQ) comparison graph- probably a bar chart or grouped boxplot. The AQ is a self-report scale that is usually employed to determine the level of how someone displays autistic characteristics. The plot exposes how various ethnicities demonstrate different scores on the AQ score, with some groups demonstrating higher scores on average. This difference may indicate that there exist pre-dispositional differences in ASD among ethnic groups, or it may be the presence of a sociocultural phenomenon that shapes the expression or diagnosis of autism characteristics. The need to ensure that predictive models are not biased to some ethnic groups can be explained by this finding because data representation in machine learning models should reflect the diverse population. There is a familiar problem of ethnic bias in the diagnostic criteria and interpretation of the distribution. Figure 9 shows composition of Autism Spectrum Quotient by Anxiety Disorder. Figure 10 shows social responsiveness scale by depression.

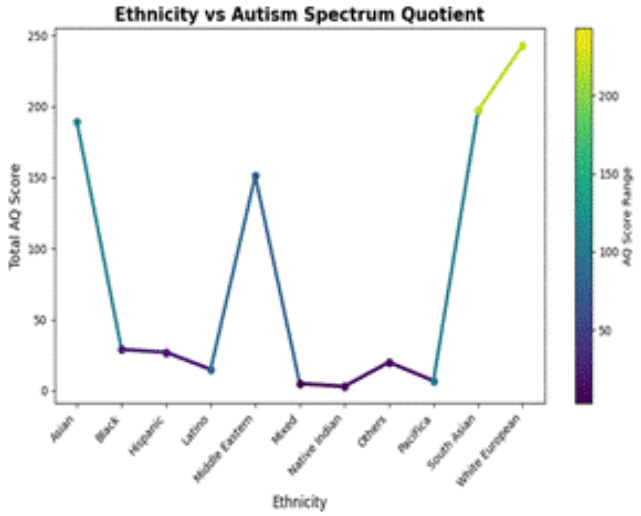

Figure 8: different scores on the AQ score.

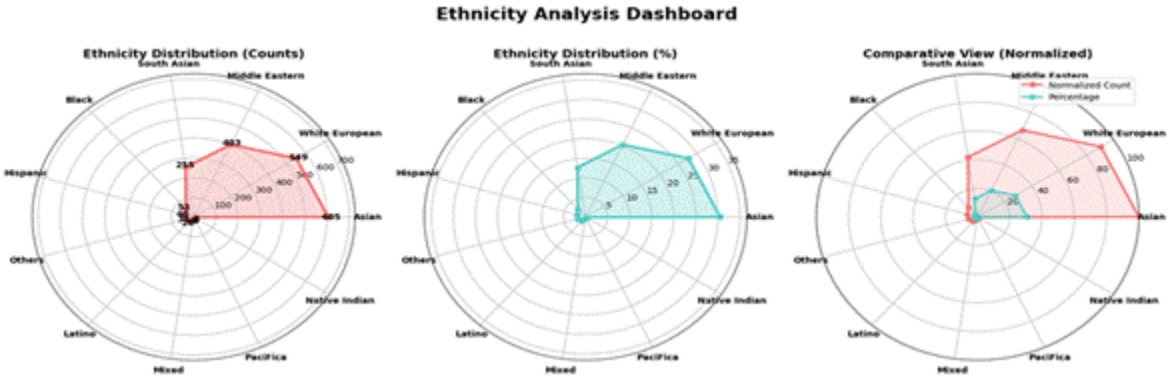

Figure 9: Composition of Autism Spectrum Quotient by Anxiety Disorder.

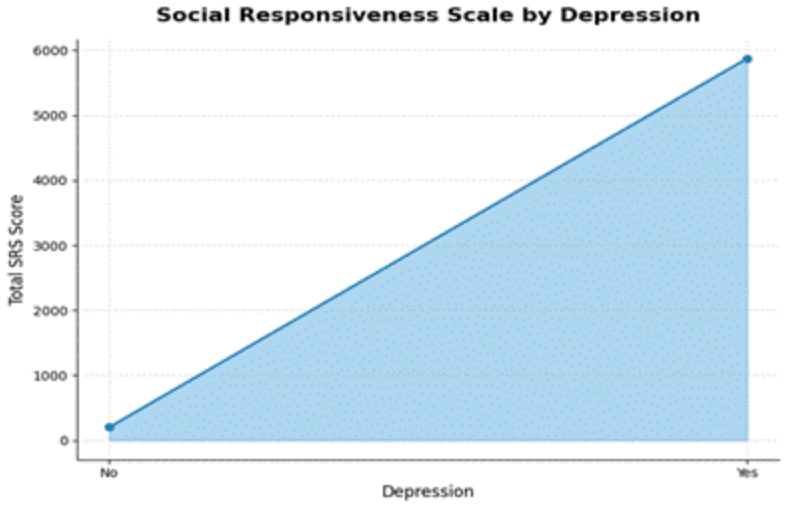

Figure 10: Social responsiveness scale by depression.

4.3 ASD Traits Heatmaps vs Sex- Family History- Genetic Disorders

A variety of heatmaps were created to investigate the dependencies among ASD traits and other categorical variables, e.g., sex, family history, and genetic disorders. With these visualizations, one could gain a sense of how these factors overlap with the occurrence of ASD traits. Figure 11 presents associations between ASD Traits and Demographic/Genetic Factors.

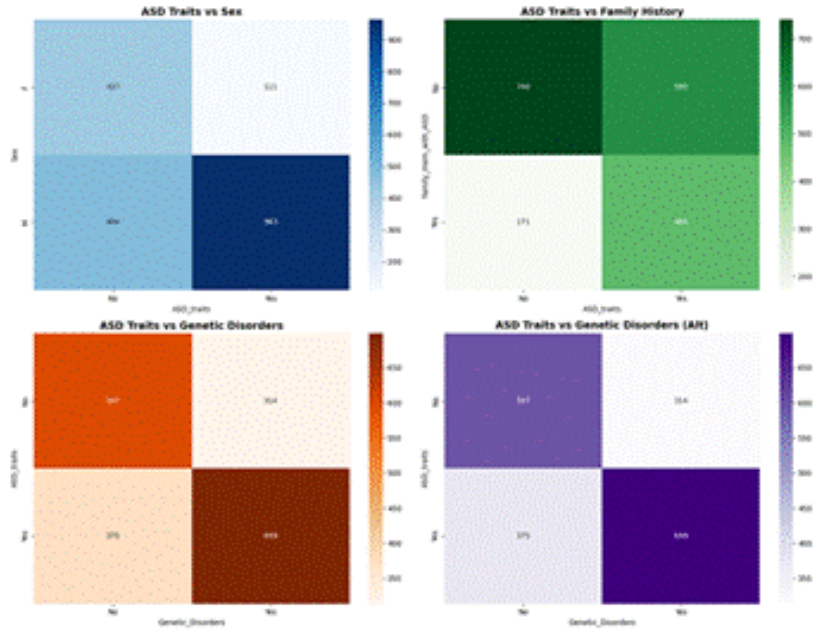

Figure 11: Associations between ASD Traits and Demographic/Genetic Factors.

The heatmap in Figure 12 shows ASD characteristics vs. sex, supporting the earlier reported case of male predominance in ASD, where more males were found to have ASD traits than females. The heatmap of ASD traits vs. family history indicated that subjects with a family history of ASD were more prone to ASD traits, which confirms the long-established genetic aspect of autism. Lastly, the ASD traits vs. genetic disorders heatmap showed that the genetic disorders (e.g., fragile X syndrome, Rett syndrome) were strongly correlated with the occurrence of ASD traits. These heatmaps prove useful not only to reveal the interdependence of different aspects but also to highlight the complexity of the etiology of ASD, which is probably due to the interplay of genetic, environmental, and familial factors.

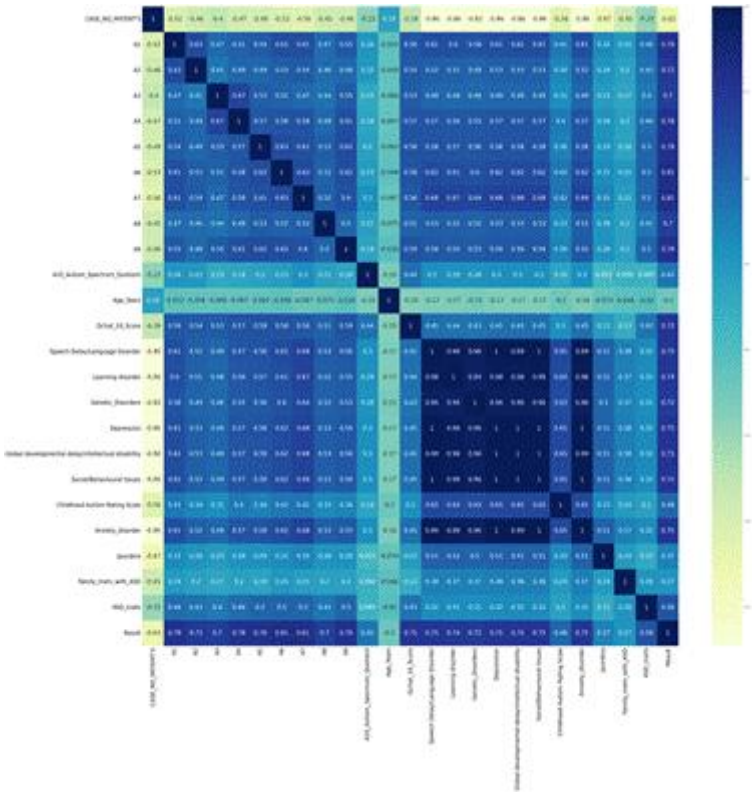

Figure 12: ASD characteristics vs. sex supported the earlier reported case of male predominance in ASD.

4.4Pre-processing and Feature Engineering of Data.

Machine learning is not a process that can happen without data preprocessing because raw data is, in many cases, characterized by inconsistencies, missing values and irrelevant features that can harm the performance of a model. In this paper, the preprocessing pipeline used was systematic in cleaning and transforming the data so as to make the best use in machine learning models. The first process was to solve the issue of missing values (a typical issue in medical datasets). The numerical characteristics, which contained missing data, were transformed by random sampling of the non-null rows of the corresponding columns, so that the imputed values were within the distribution of the observed data. Categorical features by contrast, were originally investigated based on placeholder values like “?” or blank entries. They were then changed to NaN values and then imputed with the most frequent value within each column, a method called mode imputation. This will be useful in retaining the categorical characteristic of the data and also in dealing with gaps without causing serious bias. Additionally, binary categorical data (e.g., whether the participant had jaundice or an anxiety disorder) were converted into numerical values, where "Yes" would be assigned a value of 1 and "No" would be assigned a value of 0. This conversion was to make sure that the models worked effectively with categorical data, which they regarded as binary features. The categorical columns, such as Sex, Ethnicity, and Test Completer, had to be encoded in order to be used in the models. Some of the categorical variables, including Ethnicity, were encoded using ordinal encoding, whereas others were coded using frequency encoding. Ordinal encoding puts a numerical value in categories with meaningful order, and frequency encoding puts a numerical value in categories according to the frequency

of occurrence of each category and is especially effective with variables with many categories. The feature engineering followed the encoding of categorical variables and correction of missing values and produced new composite features. Result: the aggregate of the top ten behavioral indicators, which reflected the main ASD features, was one such new variable. The Other_signs was also obtained by adding up the rest of the clinical features, which is a more comprehensive way of describing the clinical profile of the individual. Lastly, the information was divided into test and training sets in a standard 80/20 split, where 80% of the information was used to train the models and the other 20% was used to assess the model. The goals of this preprocessing and feature engineering pipeline were to maximize the utility of the dataset and provide the predictive models with high-quality and consistent input data for training.

4.5 Algorithms and Model Training

After the data preprocessing and feature engineering were done, the next step was to train a few machine learning classifiers to determine their predictive capacity in detecting ASD characteristics. Many different classifiers were chosen to be compared, and each of them offered various advantages in data complexities. The models that were selected to be used in this analysis were Decision Tree Classifier, Random Forest Classifier, Extra Trees Classifier, Support Vector Classifier (SVC), AdaBoost Classifier, K-Nearest Neighbors (KNN), Gaussian Naive Bayes, Stochastic Gradient Descent (SGD) Classifier, Logistic Regression and the Bernoulli Naive Bayes. These models were chosen so as to indicate the various categories of algorithms, including tree-based, linear and probabilistic models, and thereby enable a holistic analysis of their strengths and weaknesses when it comes to predicting ASD traits. The steps used in training the models were organized as follows: the preprocessing ended with defining of the features (X) and the target variable (y), where the target variable was the presence of ASD traits. The models were then trained using the training data and then predictions on the test data were made. Each model was evaluated using metrics to assess performance, such as accuracy, precision, recall, F1-score, Cohen's Kappa, and mean absolute error (MAE). These measures have been chosen to get a strong idea of the effectiveness of each model, paying attention to not only the classification performance (precision, recall, F1-score) but also to the general reliability of the model (accuracy, Kappa, and MAE). Using these classifiers on the set and testing their results on the unknown data, we tried to find out which models were able to most accurately predict the ASD characteristics using the demographic, clinical, and behavioral factors.

4.6 Model Evaluation

The performance of both models was measured using various measures that aimed at giving a holistic picture of the predictive power of each model on ASD traits. The main findings of this analysis are summarized in Table 2, where the models were measured in terms of their accuracy, precision, recall, the F1-score, Cohen's Kappa, and the mean absolute error (MAE). Table 1 shows a Comparison of machine learning models on ASD trait prediction.

Table 1: Comparison of machine learning models on ASD trait prediction.

Table 2: Comparison of Machine Learning Models on ASD Trait Prediction

Model	Accuracy	Precision	Recall	F1	Kappa	MAE
DecisionTree	0.973	0.973	0.971	0.972	0.946	0.027
RandomForest	0.998	0.998	0.998	0.998	0.997	0.002
ExtraTrees	0.990	0.990	0.990	0.990	0.980	0.010
SVC	0.787	0.834	0.787	0.785	0.585	0.213
AdaBoost	0.997	0.997	0.997	0.997	0.993	0.003
KNN	0.928	0.929	0.928	0.928	0.854	0.072
GaussianNB	0.716	0.728	0.716	0.717	0.436	0.284
SGDClassifier	0.824	0.863	0.824	0.823	0.656	0.176
LogisticRegression	0.844	0.868	0.844	0.844	0.692	0.156
BernoulliNB	0.680	0.685	0.680	0.681	0.357	0.320

The findings indicated that the Random Forest and AdaBoost classifiers performed better than the other models with almost perfect scores on all the metrics. In particular, the accuracy of Random Forest was 0.998, and the F1-score is 0.998, which indicates an outstanding level of predictive accuracy and equalization of classification. In the same vein, AdaBoost showed high performance with an accuracy of 0.997 and a Cohen Kappa of 0.993, which shows a high degree of agreement between actual and predicted. These findings imply that ensemble-based approaches, which involve the use of several base learners to arrive at a choice, are particularly good at this kind of predictive problem, probably because they are able to capture intricate feature interplays. Decision Tree and Extra Trees, on the other hand, were also rather good and showed a slightly lower metric, but their accuracy scores were 0.973 and 0.990, correspondingly. Although still powerful, these models might not have been able to model the same interaction between features as the ensemble methods. Another model that recorded low performance was models like Support Vector Classifier (SVC) and Gaussian Naive Bayes, which had 0.787 and 0.71,6, respectively, as their accuracy. These results indicate that some algorithms may be difficult to apply to this data, especially when there are non-linear interactions between features or other conditions are not met (as in Naive Bayes, where the features are assumed to be independent of each other). The comprehensive performance indicators indicate that ensemble classifiers, especially Random Forest and AdaBoost, are the best in situations that require predictive tasks with complex and heterogeneous data, like in the case of ASD prediction.