Autism Classification in Individuals

Department of Applied Data Science, San Jose State University

DATA-240 SEC 11

Prof. Seunjoon Lee

Final report

Team 3- Bhakti Raichura, Khushee Thakker, Shreya Srirama, Varsha Srinivasan

May 19, 2023

Objective:

Autism screening refers to the process of identifying the presence of Autism Spectrum Disorder (ASD) in individuals. It has been a topic of great interest and research in the medical field for a long time. The primary objective of developing an autism classification function is to establish a reliable and standardized method for diagnosing ASD in individuals. By accurately identifying autism, this classification function can contribute to the development of personalized treatment strategies to address specific challenges associated with autism. To effectively determine whether an individual has autism or not, the autism classification project focuses on creating and evaluating classification models. These models include Multinomial Naive Bayes, Random Forest Classifier, Decision Tree Classifier, and Logistic Regression. By utilizing these models and analyzing the autism dataset, the project aims to develop a screening tool that is both reliable and effective in diagnosing autism in adults. The research also seeks to enhance our understanding of the disorder by exploring how these categorization models can be used to identify subgroups within the autism spectrum.

In summary, the goal of the autism screening project is to establish a standardized method for diagnosing ASD and develop a reliable screening tool for autism in adults. By utilizing classification models and analyzing the autism dataset, the project aims to create an accurate screening process while also gaining insights into the subgroups within the autism spectrum.

Motivation:

Adult autism classification has gained significant attention in recent years as it has become increasingly recognized that autism is not limited to childhood but can also be diagnosed in adulthood. While autism is commonly associated with early childhood, many individuals may not receive a diagnosis until later in life. Diagnosing autism in adults requires a comprehensive approach that combines cognitive, developmental, and medical tests, as the disorder manifests differently in each individual. To determine the prevalence of autism in the adult population and improve the process of autism screening, researchers conduct projects that utilize classification models. These models leverage machine learning techniques and advancements to enhance the accuracy and effectiveness of identifying individuals on the autistic spectrum. By collecting data through autism screening projects, researchers can gain valuable insights into the number of adults affected by autism. Having an understanding of the prevalence of autism in the adult population is essential for healthcare practitioners, decision-makers, and support groups. This knowledge enables them to better allocate resources, develop targeted programs, and prepare for the future needs of individuals with autism. By identifying the number of people expected to have autism, these stakeholders can plan and provide appropriate services and support, improving the overall quality of life for individuals with autism. Autism screening plays a vital role in the diagnosis process and provides individuals with a better understanding of their own experiences. For those who have faced various challenges throughout their lives without an autism diagnosis, screening can offer a sense of self-awareness and understanding. It can help individuals make sense of their difficulties, validate their experiences, and provide a framework for accessing appropriate support and interventions. Moreover, early detection of autism through screening allows for timely intervention and tailored support. By identifying autism in adulthood, individuals can access services and therapies that can address their unique needs and challenges.

This can significantly enhance their quality of life by providing them with the necessary tools and support to navigate their daily lives effectively.

In summary, adult autism classification is a crucial topic that recognizes the prevalence of autism beyond childhood. Autism screening projects utilizing classification models and advancements in machine learning techniques contribute to more accurate identification of individuals on the autistic spectrum. This knowledge aids healthcare practitioners, decision-makers, and support groups in effectively managing resources, developing programs, and providing necessary support and services to improve the lives of individuals with autism.

Literature review:

Thabtah et al. (2019) in their research paper proposed a ML framework using logistic regression for autism screening in adolescents and adults for feature analysis and predictive analysis. The research focuses on two datasets based on the AQ-10 data adult screening method and AQ-10 adolescents screening method using ASDTests. The generated logistic regression classifiers showed acceptable levels of sensitivity, accuracy and specificity based on the feature sets selected by IG and CHI. The results also showed that CHI and IG filtering methods consistently infer common autistic traits from both adult and adolescent datasets. Thus, their research paper concluded that machine learning techniques like logistic regression, can be useful in developing accurate autism classification models that can help with early diagnosis and intervention.

Radzi et al. (2022) published an article to compare different classification algorithms used to predict ASD using WEKA. They used algorithms like Random Forest, KNN, Naive Bayes, J48, Logistic Regression, SVM, and Neural Networks on the ASD screening dataset by Dr. Fadi Fayez Thabtah . Data was preprocessed and the missing values were imputed with the mean value. The authors then compared the results of these classifiers' for both approaches i.e. with missing values and without missing values based on significant and commonly used parameters like sensitivity, specificity, accuracy, and ROC. They concluded that the J48 algorithm gave the best results for both cases.

Raj and Masood (2019) proposed analysis to detect autism using various machine learning techniques like use Naïve Bayes, Support Vector Machine (SVM) , Logistic Regression, K- nearest Neighbours (KNN), Convolutional Neural networks (CNN) and others for prediction of ASD problems for all age groups. They used three different publicly available datasets for the analysis. After preprocessing the data, training and testing, the authors compared results for all the ML models. Figure below shows sensitivity, specificity, and accuracy was calculated for each of the models and they concluded that better results were achieved for SVM and CNN.

In the paper authors A. Kanchana and R. Khilar, (2022) presents a study on predicting autism spectrum disorder (ASD) in adults using a random forest classifier. The authors collected data from adults with and without ASD, including demographic and behavioral features. The authors used different classification methods such as logistic regression, random tree, naive bayes, decision tree and random forest classifier. Their results shows that the random forest classifier was trained and tested on the dataset, achieving an accuracy of 93.44% in predicting ASD. The study suggests that the random forest classifier can be a useful tool in the early identification and diagnosis of ASD in adults. The papers describe there was limited data available for the test and in future this could be improved.

Data Source and Summary

To contribute to this ongoing research, this project aims at building an effective classifier using the data provided by UCI Machine Learning Repository. This dataset contains ten behavioral features (AQ-10-Adult) plus ten individual characteristics that have been proved to be effective in detecting ASD cases from controls in behavior science. Table below provides attribute information such as features, type and their description.

Attribute	**Type**	**Description**
Age	Number	Age in years
Gender	String	Male or Female
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean (yes or no)	Whether the case was born with jaundice
Family member with PDD	Boolean (yes or no)	Whether any immediate family member has a PDD
Who is completing the test	String	Parent, self, caregiver, medical staff, clinician ,etc.
Country of residence	String	List of countries in text format
Used the screening app before	Boolean (yes or no)	Whether the user has used a screening app
Screening Method Type (Type of screening method chosen based on age category)	Integer (0,1,2,3)	The type of screening methods chosen based on age category (0=toddler, 1=child, 2= adolescent, 3= adult)
Question 1 Answer (I often notice small sounds when others do not)	Binary (0, 1)	The answer code of the question based on the screening method used
Question 2 Answer (I usually concentrate more on the whole picture, rather than the small details)	Binary (0, 1)	The answer code of the question based on the screening method used
Question 3 Answer (I find it easy to do more than one thing at once)	Binary (0, 1)	The answer code of the question based on the screening method used

Question 4 Answer (If there is an interruption, I can switch back to what I was doing very quickly	Binary (0, 1)	The answer code of the question based on the screening method used
Question 5 Answer (I find it easy to read between the lines when someone is talking to me)	Binary (0, 1)	The answer code of the question based on the screening method used
Question 6 Answer (I know how to tell if someone listening to me is getting bored)	Binary (0, 1)	The answer code of the question based on the screening method used
Question 7 Answer (When I'm reading a story I find it difficult to work out the character's intentions)	Binary (0, 1)	The answer code of the question based on the screening method used
Question 8 Answer (I like to collect information about categories of things (e.g. types of cars, types of bird, types of train, types of plant, etc)	Binary (0, 1)	The answer code of the question based on the screening method used
Question 9 Answer (I find it easy to work out what someone is thinking or feeling just by looking at their face)	Binary (0, 1)	The answer code of the question based on the screening method used
Question 10 Answer (I find it difficult to work out people's intentions)	Binary (0, 1)	The answer code of the question based on the screening method used
Screening Score	Integer	The final score was obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner

Data Preprocessing:

Data preprocessing is the process of cleaning, transforming, and preparing raw data before it is used for data modeling The quality of the data used in analysis or models greatly affects the accuracy and reliability of the results. Therefore, it is important to carefully preprocess the data to ensure that it is accurate, complete, and suitable for the intended purpose. Following steps were performed as part of data preprocessing

- Convert from .arff to .csv- Source file was downloaded in .arff format and then converted to.csv for further python processing

- Dropping irrelevant columns like ID, age_desc- Unnecessary columns were dropped off. Age_desc had only 1 value for all the instances and hence was dropped
- Correcting columns names (austim -> autism , contry_of_res - country_of_res)- Basic spell checks were performed
- Performed data cleaning by replacing '?' as 'others'- Basic data sanity checks were performed

```
[ ] data['ethnicity'] = data['ethnicity'].replace('?', 'Others')
    data['ethnicity'] = data['ethnicity'].replace('others', 'Others')

    data['ethnicity'].value_counts()

    White-European     257
    Others             235
    Middle Eastern      97
    Asian               67
    Black               47
    South Asian         34
    Pasifika            32
    Latino              17
    Hispanic             9
    Turkish              5
    Name: ethnicity, dtype: int64
```

- Data Discretization - Converted continuous feature 'age' into multiple age groups as defined below
    - 0 - child (<13)
    - 1 - adolescent (<21)
    - 2 - adult (<40)
    - 3 - middle aged (<60)
    - 4 - elderly (>60)

```
def create_age_group(age):
    """Determine age group and return an integer indicating the category."""
    if age < 13:
        return 0 #child
    elif age < 21:
        return 1 #adolescent
    elif age < 40:
        return 2 #adult
    elif age < 60:
        return 3 #middle-aged
    else:
        return 4 #elderly
```

Data Transformation:

Data transformation is the process of converting raw data into a suitable format for analysis or modeling. This can involve a variety of techniques, including label encoding, normalization, and data balancing.

- Label Encoding-
  This is done to convert all the categorical data in the dataset to numerical data for ease in analysis as shown below

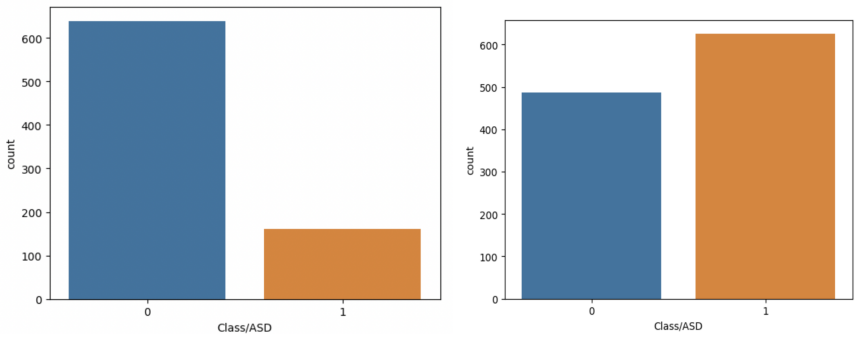
gender	ethnicity	jaundice	autism	country_of_res	used_app_before	result	relation	Class/ASD	age_group
f	Others	no	no	Austria	no	6.351166	Self	0	2
m	Others	no	no	India	no	2.255185	Self	0	3
m	White-European	no	yes	United States	no	14.851484	Self	1	0
f	Others	no	no	United States	no	2.276617	Self	0	2
m	Others	no	no	South Africa	no	-4.777286	Self	0	3

gender	ethnicity	jaundice	autism	country_of_res	used_app_before	result	relation	Class/ASD	age_group
0	5	0	0	7	0	6.351166	4	0	2
1	5	0	0	25	0	2.255185	4	0	3
1	9	0	1	54	0	14.851484	4	1	0
0	5	0	0	54	0	2.276617	4	0	2
1	5	0	0	46	0	-4.777286	4	0	3

- Data Normalization-
  The purpose of this scaling is to make sure that all features contribute equally to the analysis or model building process. In min-max scaling, each feature is scaled to a range between 0 and 1.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.555556	0.0	0.0	0.454545	0.0	0.381655	1.0	0.75
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.555556	0.0	0.0	0.981818	0.0	0.382630	1.0	0.50
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.555556	0.0	0.0	0.836364	0.0	0.061865	1.0	0.75
3	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	1.0	0.444444	0.0	0.0	0.563636	0.0	0.713926	1.0	0.50
4	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	0.666667	0.0	0.0	0.963636	0.0	0.642190	1.0	0.50

- Unbalanced data-
  Since the data was unbalanced, SMOTE-ENN was performed to balance the data. It is a combination of two data preprocessing techniques: Synthetic Minority Over-sampling Technique (SMOTE) and Edited Nearest Neighbor (ENN). This is used to handle class imbalance in datasets. This ensures balance between both the classes for autism as shown below



Data Models without Feature Selection

Multinomial Naive Bayes

Multinomial Naive Bayes (MNB) works well with discrete data: The MNB is chosen over Gaussian or Bernoulli because Gaussian is well suited for continuous input features and Bernoulli is well suited for binary input features. But as our input features are discrete in nature, MNB is well suited for our dataset.

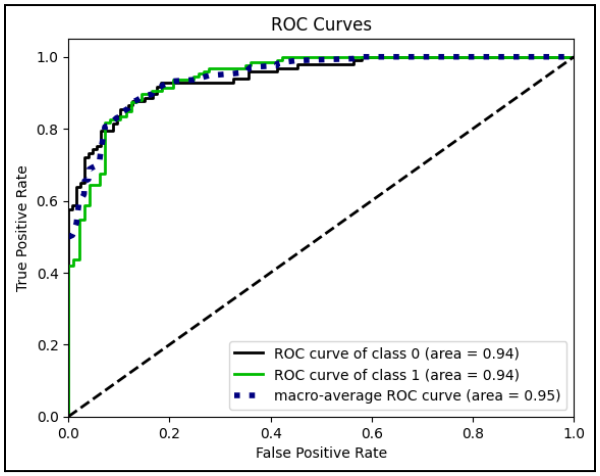
In addition the autism data involves a large number of features, in such a situation MNB can handle high-dimensional data. Further to note Naive Bayes is a simple and fast algorithm. For MNB after hyperparameter tuning, the parameters which gave the best accuracy are alpha = 1.0 and class_prior = [0.5,0.5] and the test accuracy is 86.54%. And the confusion matrix is as shown below

	Predict[0]	Predict[1]
True[0]	77	20
True[1]	10	116

And the calculated performance metrics are as shown in the below table.

Multinomial Naive Bayes	
Accuracy	0.8654
Recall	0.9206
Specificity	0.7938
Precision	0.8529
False Negative Rate	0.07

Below is the ROC curve with AUC = 0.94



Decision Tree Classifier

The main advantage of any tree classifier is its understandability and interpretability. As questionnaire is a part of our dataset, tree classification is better. And also this parametric approach can predict better if our problem is a non linear classification.

When first the DecisionTreeClassifier was trained with entropy as ‘gini’ and max_depth to be default, the testing accuracy was 93% but when checked for train accuracy, it was 100% which was a pure example of an overfit model. The depth of the full grown tree was 12.

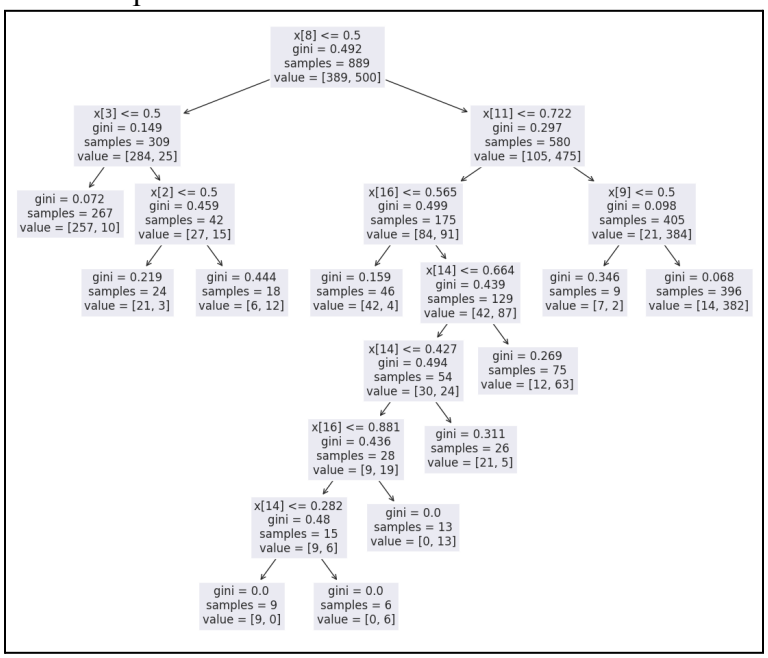
To avoid the model from overfitting, pruning the tree based on cost complexity function was performed. Multiple ccp_alpha values for the training data were found as shown below.

```
array([0.        , 0.00074991, 0.00098425, 0.00105456, 0.00106862,
       0.00107129, 0.0010832 , 0.00111044, 0.00112074, 0.00114721,
       0.00140799, 0.00149981, 0.00149981, 0.00168729, 0.0019685 ,
       0.00199975, 0.00200552, 0.00216937, 0.00218638, 0.00226311,
       0.00265866, 0.00272743, 0.00299963, 0.00385666, 0.00394854,
       0.00420817, 0.0043597 , 0.00449944, 0.00520667, 0.00678933,
       0.00686968, 0.00717149, 0.00834425, 0.01091227, 0.01105118,
       0.02632676, 0.0503943 , 0.24705608])
```

By training the model for these ccp_alpha , a plot of train and test accuracy vs ccp_alpa was drawn as shown below

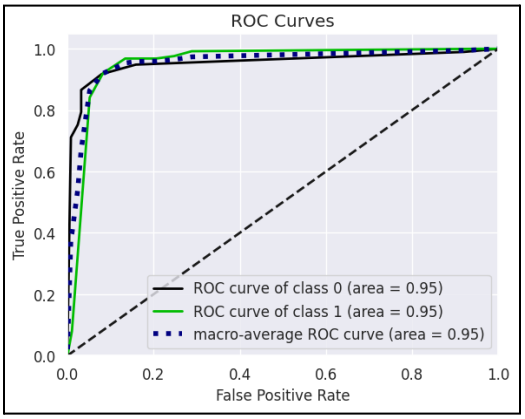
Analyzing this graph, it can be seen that the highest test accuracy of around 92% occurred approximately near ccp_alpha value of 0.00520667. Hence this value was chosen as a complexity parameter in further modeling of DecisionTreeClassifier.

Now the optimized DecisionTreeClassifier gave the test accuracy as 92.37% whereas train accuracy was 93.70%. Hence the model was not overfitting anymore. It is also observed that the depth of the tree had reduced to 7. Below is the final tree prediction.



So our final confusion matrix, performance matrix and ROC curve is as shown below.

	Predict[0]	Predict[1]
True[0]	84	13
True[1]	4	122

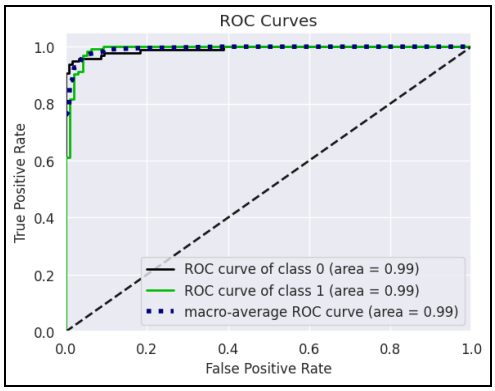


Decision Tree Classifier	
Accuracy	0.9237
Recall	0.9683
Specificity	0.8659
Precision	0.9037
False Negative Rate	0.0317

Random Forest Classifier

Random forest being a tree based classifier has the same advantages as decision trees but with the additional advantage of providing us with the insights of important features. The RandomForestClassifier with 100 estimators and max_depth default, a testing accuracy of 97% was achieved whereas the train accuracy was 100%, which was again a case of overfitting. But with the information obtained from the cost complexity pruning of decision tree about max_depth of the post pruned tree to be 7, the same max_depth was applied to the RandomForestClassifier to stop the tree from growing. Hence a better performance without overfitting was achieved with a test accuracy of 96.41% and training accuracy of 98.42%. Below is our final confusion matrix, performance matrix and ROC curve with AUC = 0.99 for RandomForestClassifier.

	Predict[0]	Predict[1]
True[0]	90	7
True[1]	1	125



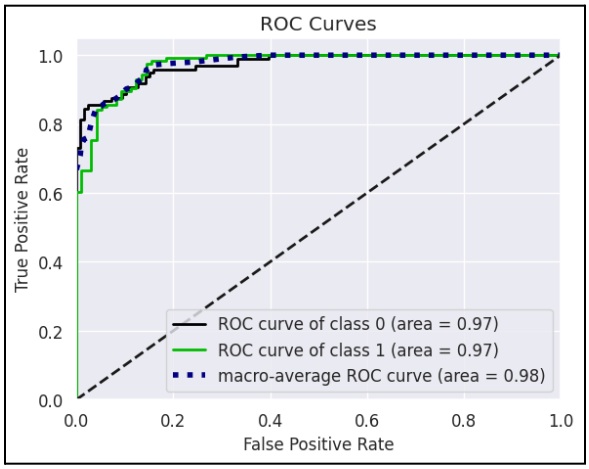
Random Forest Classifier	
Accuracy	0.9641
Recall	0.9920
Specificity	0.9278
Precision	0.9469
False Negative Rate	0.0079

Logistic Regression

To try out a parametric approach of classification, logistic regression was chosen. Also logistic regression is suitable for binary classification with an advantage of having inbuilt optimizer & regularization term. After hyperparameter tuning, it was found that logistic regression performed better for our dataset with the default penalty parameter 'l2' and solver = liblinear. It was also confirmed that this model did not overfit unlike our tree classifiers.

Below is our final confusion matrix, performance matrix and ROC curve with AUC = 0.97 for logistic regression.

	Predict[0]	Predict[1]
True[0]	83	14
True[1]	6	120



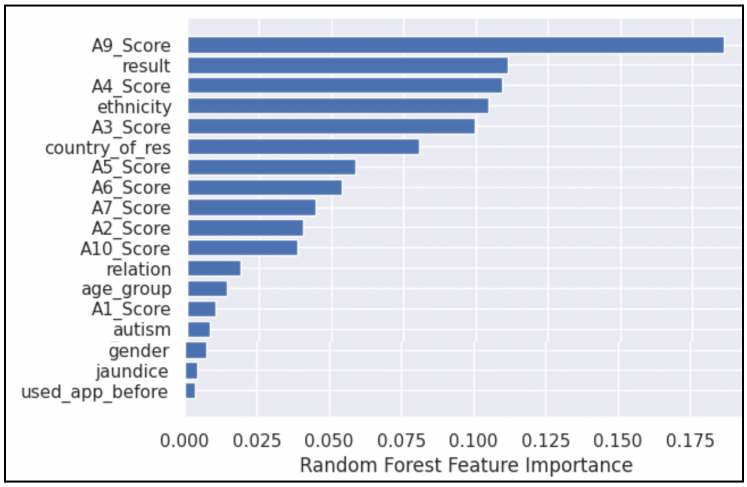
Random Forest Classifier	
Accuracy	0.9103
Recall	0.9523
Specificity	0.8556
Precision	0.8955
False Negative Rate	0.0476

Feature Selection

Now to perform feature selection we performed four different techniques and combined their results along with our domain knowledge to identify important features and eliminate least important features.

1. **Feature Importance by Random forest**

As mentioned earlier, random forest provides information about the important features. This information was utilized as one of the feature selection techniques. Below is the feature importance graph obtained from random forest.



From this graph the last four features were chosen as candidates to eliminate as they are the least important feature according to the Random forest's feature importance.

2. **90% Confidence Interval for p-value.**

P-value for each of the features was found using the logistic regression. The 95% confidence interval for p-value gave many features to be not important for the prediction but which was not the case based on the domain knowledge. Hence we decreased the confidence interval to 90% and then obtained features which have p-value less than 0.1. These were considered to be candidates to eliminate features. Below is the output of 90% Confidence Interval for p-value.

A1_Score	False
A2_Score	True
A3_Score	True
A4_Score	True
A5_Score	True
A6_Score	True
A7_Score	False
A8_Score	False
A9_Score	True
A10_Score	False
gender	True
ethnicity	True
jaundice	True
autism	False
country_of_res	False
used_app_before	True
result	True
relation	True
age_group	False

3. Chi square

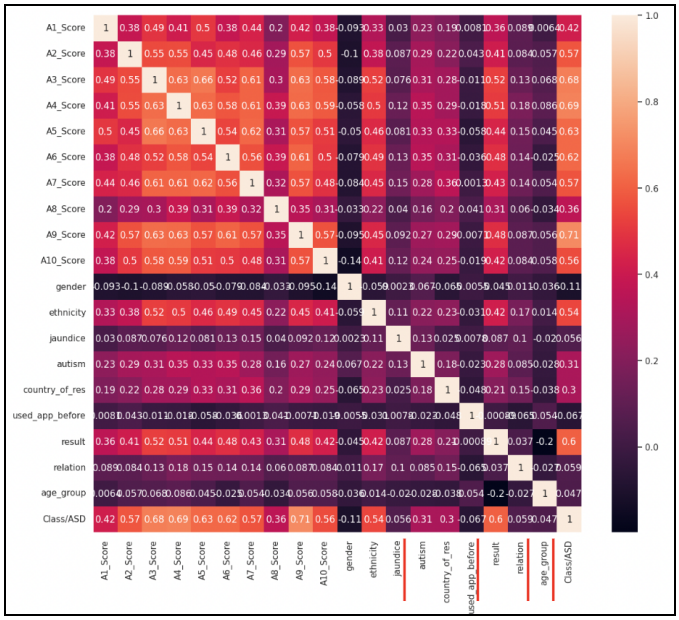
Chi square is one of the statistical methods to find the independence of a variable. So this method was used to find which features were independent from the target feature. Below is the chi square values of all feature variables wrt the target features.

	P value
Class/ASD	6.213423e-243
A9_Score	1.958013e-123
ethnicity	1.492661e-115
A4_Score	1.601689e-115
A3_Score	6.608902e-113
A5_Score	4.026656e-96
A6_Score	9.492527e-96
country_of_res	1.809455e-93
A2_Score	2.391242e-81
A7_Score	2.267194e-80
A10_Score	4.607808e-76
A1_Score	5.447364e-44
A8_Score	1.868653e-33
autism	6.699007e-25
relation	3.977709e-10
gender	2.290753e-04
used_app_before	3.689858e-02
jaundice	7.080787e-02
age_group	7.501013e-02

Higher the value, it is more independent to the target feature. Hence we are selecting the last four features as the candidates to eliminate.

4. Correlation Matrix

The last method for feature selection was correlation matrix. This matrix describes how well each feature is correlated with each other. Hence lower the correlation value, lesser the impact. So we selected the least four correlated features wrt to the target feature ‘Class/ASD’ as shown below.



Finally wrapping all the candidates into the below table.

Correration	Feature Imp	odds ratio	Chi square
Age_group	used_app_before	A1_score	Age_group
Jaundice	Jaundice	A7_score	Jaundice
relation	gender	A10_score	used_app_before
used_app_before	Autism	Autism	gender
		Country	
		Age_group	

First the features common in 3 methods were chosen. They are

- Age_group : Through our domain knowledge, we know that age of an individual does not contribute to finding if the individual has ASD or not. Hence we decided to drop it
- Used_app_before : This feature describes if the screening test was taken prior or not. Clearly this does not help in finding ASD in an individual. Hence we decided to drop it
- Jaundice : Though this feature is present as a candidate for a low important feature, we doubted this finding through our domain knowledge. Hence through our research we found that the research published in the journal Pediatrics, full-term babies with jaundice had a 56 percent higher likelihood than babies without jaundice of later developing an autism spectrum disease (article by Reuters).
    - So rather than dropping this feature, we considered it as one of the important features.
    - We also inspected the questionnaire to find if any questions were related to jaundice, as jaundice being one of the important preconditions, to our surprise there were no questions related to this but they were purely behavioral questions.

    - Next concentrating on features common in 2 methods. They are
- Gender : As our initial research on gender, we found that more males were diagnosed with ASD compared to females.
    - But digging deeper into this, the findings of May et al., (2014) was quite insightful. According to the authors males were more hyperactive and got more integration-aide assistance. While lower levels of hyperactivity in females may be a factor in their underrepresentation. Thus, it can be inferred that there are no gender differences in the symptoms of autism.
    - Hence we decided to drop this feature for better prediction.
- Autism : This feature informs if any immediate family members were diagnosed with ASD. This information was quite important according to our knowledge on ASD. Hence more research was done to find more information on this.
    - We found that the Centers of Disease Control and Prevention have publicly mentioned in their site that an individual is more likely to have developed ASD if present in their family health history.
    - So rather than dropping this feature, we considered it as one of the important features.

Now by dropping these three least important features, we ran the same model over this updated dataset and compared their results. This comparison has been discussed in the result comparison section below.

Result Comparison:

The results of the models after feature selection are presented in Table 7.

	Models	**MNB**	**DT**	**RF**	**LR**
Without Feature Selection	**Accuracy**	0.86	0.93	0.96	0.92
	FNR	0.07	0.03	0.007	0.04
With Feature Selection	**Accuracy**	0.89	0.93	0.96	0.91
	FNR	0.04	0.03	0.007	0.03

In the analysis, different classification models were evaluated in the context of autism screening. Without feature selection, Multinomial Naive Bayes exhibited the lowest accuracy (86%) and the highest false-negative rate (0.07) among the compared models. This suggests that Multinomial Naive Bayes may not be well-suited for accurately predicting autism in the dataset without considering the most relevant features. On the other hand, Random Forest showed the highest accuracy and the lowest false-negative rate (0.007) after feature selection. Feature selection helped improve the performance of the Random Forest model by focusing on the most informative features. The ability of the Random Forest model to provide a feature importance measure further assists in identifying critical features for predicting autism. This feature importance analysis can guide future research efforts and help identify potential risk factors

associated with autism. Random Forest models offer several advantages in handling the specific challenges present in the ASD dataset, such as high dimensionality, skewed data, missing values, and outliers. These models are capable of effectively capturing complex relationships and interactions among variables, leading to improved accuracy in predicting autism. In the context of medical datasets, specificity is an important statistic for evaluating model efficiency. Specifically, in the case of ASD data, specificity measures the ability of a model to correctly recognize negative cases. It calculates the percentage of true negatives accurately categorized as negative by the model. Analyzing specificity provides crucial insights into the model's effectiveness in correctly ruling out individuals who do not have the illness.

While Multinomial Naive Bayes showed higher results in textual categorization, it may be unsuitable for non-textual variables present in the ASD dataset, such as age, gender, ethnicity, and medical history. Decision Tree models, as indicated in Table 6, provided a very low specificity score, suggesting potential limitations in accurately identifying negative cases. Logistic regression underperformed in the context of ASD diagnosis because it assumes a linear relationship between predictors and the outcome, potentially missing non-linear correlations and interactions that are essential in understanding autism. In contrast, Random Forest exhibited the best accuracy (96%) and the lowest false-negative rate, making it the most effective method for predicting ASD in the dataset. The model demonstrated high accuracy and reliability, with only one incorrect prediction out of 126 cases, highlighting its potential in accurately identifying individuals with autism.

Overall, the results emphasize the importance of feature selection and the suitability of the Random Forest model for autism screening. By considering relevant features and leveraging the strengths of Random Forest models, researchers can improve the accuracy of predictions and gain valuable insights into the factors associated with autism.

	Random Forest	Multinomial Naive Bayes	Decision Tree	Logistic Regression
Accuracy	**0.96**	0.89	0.93	0.91
FNR	**0.007**	0.04	0.03	0.03
Specificity	**0.927**	0.86	0.81	0.84

Discussions

Why is one method better than other methods in your data?

In the analysis of the autism screening data, different classification models were compared based on the evaluation metrics of accuracy and false negative ratio (FNR). Accuracy represents the overall correctness of the model's predictions, while FNR specifically measures the percentage of positive cases that were incorrectly predicted as negative. Among the compared models, the Random Forest model emerged as the top performer in terms of both accuracy and FNR. It achieved the highest accuracy of 96% and the lowest FNR of 0.007. These results indicate that the Random Forest model provided the most accurate and reliable

predictions for identifying individuals with autism, minimizing the occurrence of false negatives, and reducing the risk of missing important diagnoses.

The Random Forest model's superior performance can be attributed to its ability to handle complex relationships and interactions between features in the dataset. By combining multiple decision trees and utilizing ensemble learning techniques, the Random Forest model can effectively capture the diverse patterns and variability present in the data. This enables the model to make more accurate predictions and better distinguish individuals with autism from those without. Furthermore, the Random Forest model's ability to handle high-dimensional data, skewed distributions, missing values, and outliers makes it well-suited for addressing the challenges commonly encountered in medical datasets. These characteristics contribute to the model's robust performance in autism screening.

The Random Forest model's success in accurately predicting autism suggests that the features used in the model hold valuable information related to the disorder. By analyzing the importance of these features within the model, researchers and practitioners can gain insights into the factors that play a significant role in autism diagnosis. This knowledge can contribute to a better understanding of the underlying characteristics and risk factors associated with autism.
Overall, the Random Forest model's superior performance in terms of accuracy and FNR demonstrates its effectiveness as a predictive tool in autism screening. Its ability to handle complex relationships, effectively utilize ensemble learning, and handle various challenges of medical datasets makes it a promising approach in the field. The results obtained from this analysis provide valuable insights and contribute to improving the accuracy and reliability of autism screening processes.

Difference between all features and selected features. Why are selected features important?

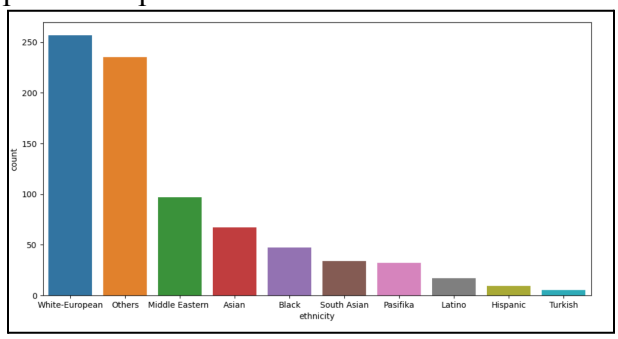
Feature selection is a part of identifying the most relevant and informative features from the available dataset. In the analysis, two scenarios were compared: one where all features were used, and another where feature selection was performed to determine the subset of important features. The selected features are important because they contribute the most to the prediction performance of the model. By selecting relevant features, the model can focus on the most discriminative and impactful variables, improving its ability to accurately classify individuals with autism. Feature selection helps address several challenges commonly found in medical datasets. High dimensionality, skewed data, missing values, and outliers can negatively impact model performance and interpretability. By selecting the most informative features, the model can better capture the underlying patterns and relationships within the data, leading to improved accuracy and performance. As part of this project, multiple feature selection methods were used like chi-square, correlation matrix etc.

Selected Features	Not Selected Features
Ethnicity	Age_group
Result	Used_app_before
Question 1 - 10	Gender

Family member with PDD	Screening Method Type (Type of screening method chosen based on age category)
Born with jaundice	
Relation	
Country of residence	

The selected features mentioned above are important for several reasons:

Ethnicity/Country of residence: Based on the below results,we see that individuals at certain geographical locations and with a certain heritage are more likely to be diagnosed as autistic. However, upon further research, we identified that the poor results are just an indicator of low testing at such locations. For example, the results show that a person with Hispanic or Turkish background is more likely to develop ASD when compared to some of the developed countries. However, this could just be on account of low testing opportunities in this country. While ethnicity does not play a very significant role, it still cannot be ruled out and helps in improved predictive performance.



Autism (Family member with PDD) : As spoken earlier in the feature selection process, the Centers of Disease Control and Prevention (CDC) have mentioned that any individual is more likely to have developed ASD if present in their family health history. So it's always been a step in collecting family health history during screening of autism or even during pregnancy for early detection of the possibility of ASD. Hence this feature is one of the important attributes in detecting ASD. This feature helps in Improved Predictive Performance and Enhanced Interpretability.

Question 1 - 10: These behavioural questions play an extremely important role in identification of ASD as they are direct responses from an individual. ASD has been studied in depth by researchers and the AQ-10 data adult screening method has proven to be highly efficient. Questions like reading between lines and attention span are very well able to capture the brain maturity and hence are an excellent parameter for ASD identification.

Jaundice: Jaundice is a condition where excessive bilirubin in bile juice is accumulated in the liver. Such a condition is quite common in children during their birth. In rare scenarios this bilirubin can cause damage in brain cells causing ASD in future. With this domain knowledge, it can be justified that feature jaundice is one of the important attributes. This is also supported by an article in Reuters where it speaks about a research published in the journal Pediatrics. This

research says that full-term babies with jaundice had a 56 percent higher likelihood than babies without jaundice of later developing an autism spectrum disease.

Meaning of the results and interpretability based on domain knowledge and references

The analysis of the autism screening dataset revealed that the Random Forest model, when equipped with a subset of selected features, outperformed other models in terms of accuracy and false negative ratio (FNR). This indicates that the Random Forest model is highly proficient in predicting autism when provided with the most relevant features.

To gain deeper insights from these results, researchers and practitioners can incorporate domain knowledge and existing references in the field of autism research. By examining question 3 which talks about if an individual finds it easy to multitask, one can easily analyze the symptom by just observing if an individual can multitask. This is one of the key and highly correlated questions in the prediction of ASD. And another essential behavior of an individual is quickly switching between tasks even if an interruption is presented. But such behavior is observed to be difficult in individuals with ASD. Our feature importance method highlighted that question 4 which answers this question is highly correlated with prediction of ASD.

Questions 5 and question 9 talk a lot about the Emotional Quotient (EQ) of an individual. Based on historic research, a person who finds it easy to read in between lines or understand the underlying tone or emotions when someone is talking can explain a lot. This contributes to a better understanding of the underlying characteristics of the disorder and provides researchers with additional support and reinforcement for existing knowledge.

Furthermore, considering the limitations of other models, such as Multinomial Naive Bayes, Decision Tree, and Logistic Regression, based on domain knowledge can enhance the interpretation of the results. Multinomial Naive Bayes may not be well-suited for handling non-textual variables present in the autism spectrum disorder (ASD) dataset. Decision Tree models may have limitations in terms of specificity, potentially affecting their ability to accurately identify negative cases. Logistic Regression assumes a linear relationship between predictors and outcomes, potentially overlooking non-linear correlations and interactions that are crucial in ASD diagnosis.

In conclusion, the results provide strong evidence supporting the effectiveness of the Random Forest model with selected features for autism screening. Its high accuracy and ability to focus on the most relevant features contribute to its predictive power. By integrating the model's findings with domain knowledge and references, researchers can enhance their understanding of the factors influencing autism diagnosis, thereby improving the overall interpretation and applicability of the results in the field.

References

Thabtah, F., Abdelhamid, N., & Peebles, D. (2019). A machine learning autism classification based on logistic regression analysis. *Health Information Science and Systems*, *7*(1). https://doi.org/10.1007/s13755-019-0073-5

Raj, S., & Masood, S. (2020). Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques. *Procedia Computer Science*, *167*, 994-1004. https://doi.org/10.1016/j.procs.2020.03.399

Mohd Radzi, S.F., Hassan, M.S. & Mohd Radzi, M.A.H. Comparison of classification algorithms for predicting autistic spectrum disorder using WEKA modeler. *BMC Med Inform Decis Mak* 22, 306 (2022). https://doi.org/10.1186/s12911-022-02050-x

A. Kanchana and R. Khilar, "Prediction of Autism Spectrum Disorder using Random Forest Classifier in Adults," *2022 IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*, Goa, India, 2022, pp. 242-249, doi: 10.1109/ICCCMLA56841.2022.9989304

A. S. Halibas, L. B. Reazol, E. G. T. Delvo and J. C. Tibudan, "Performance Analysis of Machine Learning Classifiers for ASD Screening," *2018 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, Sakhier, Bahrain, 2018, pp. 1-6, doi: 10.1109/3ICT.2018.8855759