**Autism Prediction**

Madhuri Basava

Bellevue University

Predictive Analytics 630

Professor Andrew Hua

**Project Milestone 5**

**Introduction:**

#### Autism, or autism spectrum disorder (ASD), is a complex neurodevelopmental disorder that affects individuals in different ways, particularly in social interaction, communication, and behavior. It is termed a "spectrum" disorder because it can manifest differently in each person, ranging from mild to severe symptoms, and may include a wide range of challenges and strengths. Common symptoms may include challenges in social interaction (difficulty understanding social cues, making eye contact, or forming relationships), communication difficulties (delayed speech development, trouble with language pragmatics), repetitive behaviors, or restricted interests (engaging in repetitive movements, fixation on specific topics). Early identification and intervention are crucial for improving outcomes.

#### While there is no cure for autism, early intervention and various therapies can help individuals with autism lead fulfilling lives. Treatment approaches may include behavioral therapy, speech and language therapy, occupational therapy, and educational interventions tailored to the individual's needs. Supportive services and accommodations, such as special education programs, social skills training, and assistive technologies, can also be beneficial.

It's important to recognize that each person with autism is unique, with their strengths, challenges, and experiences. Embracing neurodiversity and promoting acceptance and inclusion can contribute to creating a more supportive and inclusive society for individuals with autism.

#### I would like to sell this project to anyone who is related to healthcare.

#### The data source was taken from the Kaggle website: <https://www.kaggle.com/code/raselmeya/asd-predictions-with-8-different-models-85-7/notebook>

#### This data has various features with which we can predict if the individual is affected by autism and provide intervention for them.

**Methods/Results:**

The data source is taken from the Kaggle website**(**[ASD Predictions with 8 different models [85.7%] | Kaggle](https://www.kaggle.com/code/raselmeya/asd-predictions-with-8-different-models-85-7/notebook)). This dataset consists of 10 Questionnaire features and 12 personal info features. The training set is greatly imbalanced with 800 rows. This project focuses on improving Autism Screening by creating models that can predict the likelihood of the individual suffering from autism based on the given features.

**Features Information:**

1. **ID**- ID of the patient that is the unique Identifier.
2. **A1\_Score to A10\_Score** - Score based on Autism Spectrum Quotient (AQ) 10-item screening tool.

|  |  |  |
| --- | --- | --- |
| Question 1 (A1) | Binary (0, 1) | S/he often notices small sounds when others do not, (Child, Adolescent) S/he notices patterns in things all the time, (Adult) Does your child look at you when you call his/her name? (Toddler) |
| Question 2 (A2) | Binary (0, 1) | S/he usually concentrates more on the whole picture, rather than the small details, (child, Adolescent, Adults) How easy is it for you to make eye contact with your child? (Toddler) |
| Question 3 (A3) | Binary (0, 1) | In a social group, s/he can easily keep track of several different people’s conversations, (child, Adolescent) I find it easy to do more than one thing at once, (Adult) Does your child point to indicate that s/he wants something? (e.g. a toy that is out of reach) (Toddler) |
| Question 4 (A4) | Binary (0, 1) | S/he finds it easy to go back and forth between different activities, (child, Adolescent) If there is an interruption, s/he can switch back to what s/he was doing very quickly, (Adult) Does your child point to share interest with you? (e.g. pointing at an interesting sight) (Toddler) |
| Question 5 (A5) | Binary (0, 1) | S/he doesn’t know how to keep a conversation going with his/her peers, (child, Adolescent) I find it easy to read between the lines when someone is talking to me, (Adult) Does your child pretend? (e.g. care for dolls, talk on a toy phone) (Toddler) |
| Question 6 (A6) | Binary (0, 1) | S/he is good at social chit-chat, (child, Adolescent) I know how to tell if someone listening to me is getting bored, (Adult) Does your child follow where you’re looking? (Toddler) |
| Question 7 (A7) | Binary (0, 1) | When s/he is read a story, s/he finds it difficult to work out the characters intentions or feelings, (Child) When s/he was younger, s/he used to enjoy playing games involving pretending with other children, (Adolescent) When I’m reading a story, I find it difficult to work out the characters intentions, (Adult) If you or someone else in the family is visibly upset, does your child show signs of wanting to comfort them? (e.g. stroking hair, hugging them (Toddler) |
| Question 8 (A8) | Binary (0, 1) | When s/he was in preschool, s/he used to enjoy playing games involving pretending with other children, (Child) S/he finds it difficult to imagine what it would be like to be someone else, (Adolescent) I like to collect information about categories of things (e.g. types of car, types of bird, types of train, types of plant, etc.), (Adult) Would you describe your child’s first words as: (Toddler) |
| Question 9 (A9) | Binary (0, 1) | S/he finds it easy to work out what someone is thinking or feeling just by looking at their face, (Child) S/he finds social situations easy, (Adolescent) I find it easy to work out what someone is thinking or feeling just by looking at their face, (Adult) Does your child use simple gestures? (e.g. wave goodbye) (Toddler) |
| Question 10 (A10) | Binary (0, 1) | S/he finds it hard to make new friends, (Child, Adolescent) I find it difficult to work out people’s intentions, (Adult) Does your child stare at nothing with no apparent purpose? (Toddler) |

1. **age** - Age of the individual in years
2. **gender** - The gender of the individual
3. **ethnicity** - Ethnicity of the individual
4. **jaundice** - Whether the individual had jaundice at the time of birth.
5. **autism** - Whether an immediate family member was diagnosed with autism.
6. **country\_of\_res** - Country of residence of the individual
7. **used\_app\_before** - Whether the individual went through a screening test before
8. **result** - Score for AQ1-10 screening test.
9. **age\_desc** - Age of the individual
10. **relation** - Relation of an individual who completed the test.
11. **Class/ASD Class** - The result is either 0 or 1. Here 0 means No and 1 means Yes.

I followed the below procedure:

1. **Load the train and test data sets into data frames.**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. **Perform the EDA to understand the characteristics of the data set.**

**A screenshot of a computer

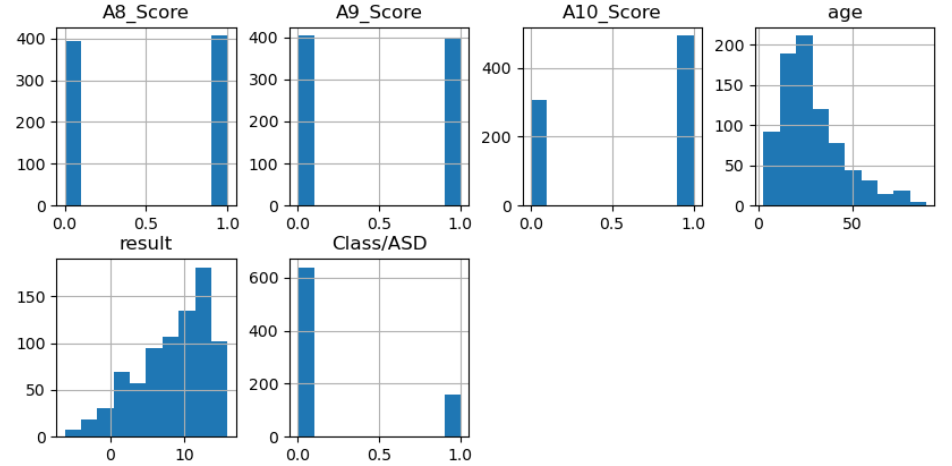
Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a graph

Description automatically generated**

****

**A screenshot of a computer screen

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a graph

Description automatically generated**

**A graph of different numbers

Description automatically generated with medium confidence**

1. **Clean the data set. Remove the unnecessary features.**

Data is taken from the Kaggle website. So, mostly the data is clean. Data is prepared as below:

* Checked the data for null values with the info() function. There are no null values.
* Checked the ‘age\_desc’ column for unique values with value\_counts() function. It has only one unique value. So, I dropped the ‘age\_desc’ column from the training set.
* Also ‘Id’ column is not needed for modeling. So, I dropped the ‘ID’ column also.

A screenshot of a computer

Description automatically generated

* Performed CHI Square to check which columns are not needed and dropped the 'gender' and 'used\_app\_before' columns.A screenshot of a computer code

  Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A close up of a box

Description automatically generated

* Converted Boolean yes/no variables to 0 (No) and 1 (Yes).

A screenshot of a computer program

Description automatically generated

A screenshot of a computer code

Description automatically generated• Scaled the ‘result’ and ‘age’ data using a RobustScaler.A black text on a white background

Description automatically generated

1. **Evaluate the correlation between the variables in the dataset.** **A screen shot of a graph

   Description automatically generated**

The target class distribution looks imbalanced, so need to resample before modeling.

**A screenshot of a graph

Description automatically generated**

From the above chart, we can conclude that the female percentage is approximately half of the male percentage in both those not having autism conditions and those with autism conditions.

**A screenshot of a graph

Description automatically generated**

From the above graph, we can conclude that at least half of the patients with autism have Jaundice.

**A screenshot of a graph

Description automatically generated**

From the above graph, we can conclude that most patients have relatives with autism in the family.

**A screenshot of a computer code

Description automatically generated**

**A graph with numbers and a bar

Description automatically generated**

From the above graph, we can conclude that autism does not depend on app usage.

**A graph of age vs autism

Description automatically generated**

**A screenshot of a computer screen

Description automatically generated**

**A group of graphs with numbers

Description automatically generated**

1. **perform the prediction analysis with Eight different models as below.**
2. **Create a confusion matrix to show the performance of each model to evaluate the predicted values from the model vs. the actual values from the test dataset.**
3. **Calculate the Accuracy for each model and choose the best model.**
4. **Perform the prediction analysis with Eight different models as below:**
5. K Nearest Neighbors
6. Decision Tree
7. LGBM (Light Gradient Boosting Machine)
8. XGBRF (Extreme Gradient Boost-Random Forest)
9. Cat Boost Classifier
10. Random Forest
11. Logistic Regression
12. SVC (Support Vector Classifier)

I chose all these models to improve the accuracy to reduce the false positives and false negatives (Inaccurate predictions). I calculated the accuracy and plotted the confusion Matrix for above 8 Models.

1. **KNearestNeighbours:**

**A screenshot of a computer

Description automatically generated**

A screenshot of a computer

Description automatically generated

A white background with text

Description automatically generated

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

5) Categorical Boost Classifier:

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**8) Support Vector Machine Model:**

**A screenshot of a computer screen

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**8) Plotting all the models to select the best model**

**A screen shot of a computer

Description automatically generated**

**A screen shot of a graph

Description automatically generated**

**Interpretation of Results:**

A higher ROC AUC score indicates better discrimination performance of the model. Specifically:

* ROC AUC = 0.5: The model performs no better than random guessing. It's essentially equivalent to a coin flip.
* 0.5 < ROC AUC < 0.7: The model has poor to fair discrimination ability.
* 0.7 < ROC AUC < 0.8: The model has good discrimination ability.
* 0.8 < ROC AUC < 0.9: The model has very good discrimination ability.
* ROC AUC ≥ 0.9: The model has excellent discrimination ability.

Based on the above scales:

The Support Vector Classifier has the lowest accuracy, which is 0.51. It is like random guessing.

The DecisionTree Model has good discrimination ability as the accuracy is 0.79.

KNearestNeighbours Model has very good discrimination ability as the accuracy is 0.88 and

All other models have excellent discrimination ability and from all of them, the Category Boost Model and Logistic Model are the best Models as their accuracy is 0.93.

Created the confusion Matrix for all 8 models. It depicted four crucial factors as below:

1. **True Positives (TP):** True positives represent the number of instances where the model correctly predicts individuals with autism. These are cases where individuals have autism, and the model correctly identifies them as such.
2. **True Negatives (TN):** True negatives represent the number of instances where the model correctly predicts individuals without autism. These are cases where individuals do not have autism, and the model correctly identifies them as not having autism.
3. **False Positives (FP):** False positives represent the number of instances where the model incorrectly predicts individuals as having autism when they do not. These are cases where individuals do not have autism, but the model mistakenly classifies them as having autism.
4. **False Negatives (FN):** False negatives represent the number of instances where the model incorrectly predicts individuals as not having autism when they do. These are cases where individuals have autism, but the model fails to classify them as such.

**Conclusion:**

After performing analysis on the above 8 different models, we saw that the Categorical Boost Classifier and Logistic Regression are the best Models which gave 93% accuracy.

Identified the most important features and factors contributing to the prediction of autism which is depicted in the above Feature Importance graph. Understanding these factors can provide valuable insights into the underlying mechanisms and risk factors associated with autism spectrum disorder (ASD).

We should consider potential applications of the predictive model in clinical practice, research, and public health initiatives. This may include early screening and intervention programs, personalized treatment approaches, or resource allocation for autism services.

There is a need to acknowledge the limitations and uncertainties associated with autism prediction, including the risk of false positives, false negatives, and the complexity of ASD diagnosis. It's essential to communicate these limitations transparently to stakeholders.

We should ensure that the predictive model respects individuals' autonomy, privacy rights, and cultural sensitivities. Also, raise awareness about autism, and promote understanding of predictive modeling in healthcare.

**References:**

1. *Autism prediction dataset retrieved from Kaggle website:* [*ASD Predictions with 8 different models [85.7%] | Kaggle*](https://www.kaggle.com/code/raselmeya/asd-predictions-with-8-different-models-85-7/input)
2. [*A new machine learning model based on induction of rules for autism detection - Fadi Thabtah, David Peebles, 2020 (sagepub.com)*](https://journals.sagepub.com/doi/10.1177/1460458218824711)
3. [*Detecting autism spectrum disorder using machine learning techniques - PMC (nih.gov)*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8024224/)
4. [*Autism Prediction with CatBoost | Kaggle*](https://www.kaggle.com/code/naveenkonam1985/autism-prediction-with-catboost)
5. [*A routine prenatal ultrasound can identify early signs of autism, study finds | ScienceDaily*](https://www.sciencedaily.com/releases/2022/02/220209112107.htm)
6. [*Machine learning analysis of pregnancy data enables early identification of a subpopulation of newborns with ASD - PMC (nih.gov)*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7994821/)