Evaluating the Use of Machine Learning to Predict Autism in Adults

Milestone 5: Final Term Project Paper

Benjamin Bartek

Bellevue University

DSC 630-T302 Predictive Analytics (2245-1)

Prof. Andrew Hua

June 1, 2024

I. Introduction

Autism Spectrum Disorder is a condition that affects both children and adults. According to the National Institute of Mental Health, "Autism spectrum disorder (ASD) is a neurological and developmental disorder that affects how people interact with others, communicate, learn, and behave."[1]. While the degree to which Autism Spectrum Disorder affects an individual's life varies, the fact that ASD can negatively impact multiple aspects of an individual's life underscores the need for effective treatment of ASD. Moreover, as stated by Fadi Thabtah in the abstract for the selected dataset, "Autistic Spectrum Disorder (ASD) is a neurodevelopment condition associated with significant healthcare costs, and early diagnosis can significantly reduce these[,]" and "[t]he economic impact of autism and the increase in the number of ASD cases across the world reveals an urgent need for the development of easily implemented and effective screening methods."[2].

Patients, mental health professionals, and insurance companies could all potentially benefit from an effective ASD screening method. Though an effective screening method is not itself a replacement for a diagnosis of being on the autism spectrum, effective screening methods could potentially help streamline the process for individuals to obtain an ASD diagnosis from a mental health professional. From a data analysis perspective, if predictive machine learning models could successfully identify which variables are most predictive of an individual having Autism Spectrum Disorder, a high accuracy model could assist those who suspect that they may have ASD in obtaining a diagnosis. Such a data-driven model could also help mental health professionals have greater confidence in evaluating patients for the characteristics of ASD. Faster, higher confidence

diagnoses would lead to patients getting quicker access to beneficial resources for the treatment of ASD, with the potential to also reduce the costs associated with the diagnostic process.

A. Dataset

In furtherance of the goal to use predictive analytics to determine whether there are characteristics associated with Autism Spectrum Disorder, I selected a dataset by Fadi Thabtah called "Autism Screening Adult", which is housed in the University of California Irvine Machine Learning Repository [2]. It should also be noted that Thabtah also has a separate dataset for children that will not be used for this project. Thabtah's dataset contains 704 rows and 21 columns of data for adults, including 'age', 'gender', 'ethnicity', 'jaundice', 'autism', 'country_of_res', 'used_app_before', 'result', ''age_desc', 'relation', 'Class/ASD', and binary values representing the answer to 10 questions [2].

II. Methods/Results

A. Data Preparation

The data preparation process involved importing necessary libraries, loading the dataset from a CSV file, renaming columns to correct typos, providing summary statistics and descriptions for both categorical and numerical variables, applying label encoding to convert categorical columns into numerical values, checking data types to ensure they are appropriate for analysis, and verifying the dimensions of the DataFrame to understand the size of the dataset. Additionally, the age and result columns were converted to integers, and misspellings in the autism and jaundice columns were corrected. Missing values in the age column were replaced with the mean age, and an outlier (383 years) was also adjusted to the mean age. The ethnicity and relation columns had

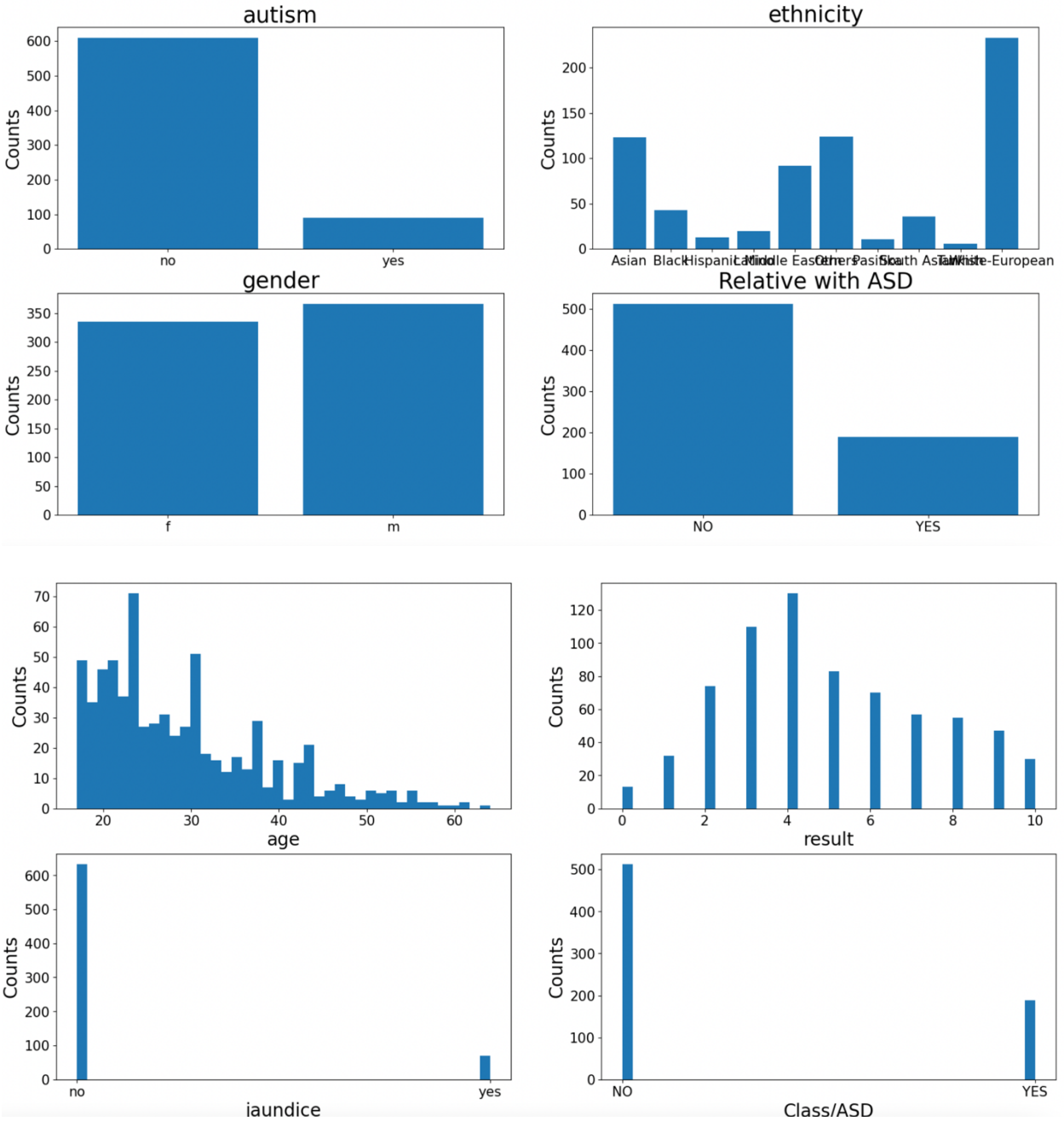
categorical values of "?" which were recategorized and combined with "Others" for ethnicity and "Unknown" for relation. The unnecessary age_desc column was dropped. After exploring the data with visualizations, the Class/ASD target value was separated from the features, label encoding was applied, and the data was split into training and testing sets with a test size of 0.2. These steps ensured the data was properly cleaned, encoded, and ready for further analysis and visualization.

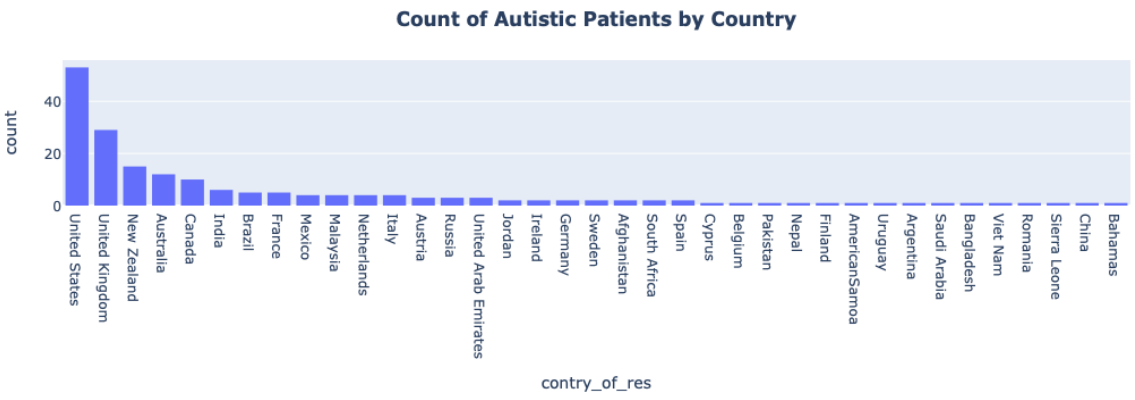
B. Data Exploration

By exploring the data, several key insights were discovered. The majority of the respondents (approximately 87%) did not have autism, while a smaller proportion (about 13%) did. The dataset contained responses from respondents from various countries, including a significant number from the United States, United Arab Emirates, India, New Zealand, and the United Kingdom. Slightly more respondents were male than female, and the prevalent ethnicity was White-European. Additionally, the data showed that the vast majority of respondents (98%) had not used Thabtah's autism screening app before, and the most common relation of the respondent was 'Self'. Most respondents did not have a relative with autism.

C. Visualizations

Visualizations were employed to explore and analyze the dataset, including bar plots to display the distribution of categorical variables like gender, ethnicity, and relation; histograms to show the distribution of numerical variables such as age and result scores; and confusion matrixes to display the results following model implementation. These visualizations collectively tell a story by illustrating the demographic makeup of the dataset, highlighting key differences and similarities between groups, and revealing patterns and correlations that provide deeper insights into the characteristics and prevalence of ASD.


autism
Counts
no
yes
ethnicity
Counts
Asian
Black
Hispanic
Middle Eastern
Others
Pasifika
South Asian
White-European
gender
Counts
f
m
Relative with ASD
Counts
NO
YES
Counts
age
Counts
result
Counts
no
yes
jaundice
Counts
NO
YES
Class/ASD



D. Models and Metrics

Two machine learning models were employed to analyze the dataset: a logistic regression model and a random forest model. Both models were built using default parameters. To evaluate the models, accuracy scores, confusion matrices, and classification reports (including precision, recall, and F1 scores) were used. These metrics helped assess the performance and effectiveness of the models in predicting ASD classification from the provided dataset.

Accuracy, precision, recall, and F1 scores were chosen because they provide a comprehensive evaluation of the model's performance from different perspectives. Accuracy gives an overall measure of correctness, though it can be misleading with imbalanced datasets. Precision is crucial for assessing the exactness of positive predictions, especially when the cost of false positives is high. Recall measures the model's ability to identify all relevant instances, which is important when the cost of false negatives is high. The F1 score balances precision and recall, offering a single metric that is particularly useful for imbalanced datasets or when both false positives and false negatives are important. Together, these metrics ensure a thorough assessment of the model's predictive capabilities[3].

```
logreg = LogisticRegression(random_state=16)

logreg.fit(X_train, y_train)

y_pred = logreg.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

```
Accuracy: 1.0
```

```
from sklearn.metrics import classification_report
target_names = ['No Autism', 'Autism']
print(classification_report(y_test, y_pred, target_names=target_names))
```

```
              precision    recall  f1-score   support

   No Autism       1.00      1.00      1.00        98
      Autism       1.00      1.00      1.00        43

    accuracy                           1.00       141
   macro avg       1.00      1.00      1.00       141
weighted avg       1.00      1.00      1.00       141
```

Because the machine learning models have 100% accuracy and 100% metrics (precision, recall, and F1 score), it means the model is perfectly predicting the target variable, correctly classifying every instance in the dataset. This could indicate perfect prediction, but it more commonly suggests overfitting, where the model has learned the training data too well, including noise and outliers. It might also point to data issues like data leakage. Such perfect metrics are rare in real-world scenarios and may indicate that the model won't generalize well to new data, however, the source of the overfitting was not identified prior to the conclusion of this project.

```
rf = RandomForestClassifier()
rf.fit(X_train, y_train)

y_pred = rf.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

```
Accuracy: 1.0
```

```
from sklearn.metrics import classification_report
target_names = ['No Autism', 'Autism']
print(classification_report(y_test, y_pred, target_names=target_names))
```

```
              precision    recall  f1-score   support

   No Autism       1.00      1.00      1.00        98
      Autism       1.00      1.00      1.00        43

    accuracy                           1.00       141
   macro avg       1.00      1.00      1.00       141
weighted avg       1.00      1.00      1.00       141
```

III. Conclusion

A. Model Feasibility

My data analysis indicates that it is potentially feasible to create a predictive tool to diagnose autism using machine learning models. The logistic regression and random forest models used in the analysis both showed exceptionally strong performance. The data preparation steps ensured that the dataset was clean and well-structured, while the visualizations and exploratory analysis provided insights that informed model building. The evaluation metrics demonstrated the models' ability to predict autism effectively, suggesting that a predictive tool based on these models could be reliable. However, as previously indicated, the exceptional performance of the models suggests overfitting, indicating that the model learned the training data too well or that there was leakage.

Accordingly, the models are certainly not ready for deployment, and I would recommend further testing and evaluation. Additional analysis would be required to determine for certain whether the models' metrics are correct. Future work with these models should include verifying the results by checking for data leakage, using cross-validation to validate performance on different subsets, examining the dataset for errors or duplicates, testing the model on an external dataset to assess generalizability, and ensuring the model isn't overly complex and prone to overfitting. These steps help confirm whether the model's perfect performance is legitimate or a result of overfitting or data issues.

B. Ethical Considerations

When dealing with data related to healthcare generally, and in this instance data for diagnosing autism, it is essential to ensure privacy and confidentiality to protect individuals' identities, obtain informed consent from participants, and check for and mitigate any biases that could lead to unfair predictions. Additionally, robust data security measures should be in place to safeguard sensitive information, and transparency about the data usage and predictive tool operation should be maintained. Thabtah's dataset was already anonymized, eliminating immediate concerns with respect to using the data for this project.

With respect to the model and presentation of results, the model must be checked for and mitigated against biases to ensure fairness and equity across different demographics. Transparency and explainability of the model's predictions are important factors to make the model's predictions meaningful and trustworthy for healthcare providers and patients alike. One must also consider the potential harm that false positives and negatives might have on individuals' lives. Robust data security and confidentiality measures help to protect sensitive information from becoming public.

However, the possibility of false positive or false negatives also underscores the necessity of ensuring that the machine learning models themselves are built correctly and tested across multiple datasets.

IV. References

1. National Institute of Mental Health. (2024). *Autism Spectrum Disorder*. https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd.
2. Thabtah, Fadi. (2017). *Autism Screening Adult*. UCI Machine Learning Repository. https://doi.org/10.24432/C5F019.
3. Nicholson, Chris V. (2023). *Evaluation Metrics for Machine Learning - Accuracy, Precision, Recall, and F1 Defined*. Pathmind. https://wiki.pathmind.com/accuracy-precision-recall-f1.