Keiuntae Smith
DSC550
Data Mining
May 31, 2022

Introduction

Introduce the problem

The subject I chose for my final project is one that hits close to home. I am a proud father of two autistic children that are nonverbal. There are many challenges that come with dealing with individuals on the autism spectrum. Autism spectrum disorder (ASD) is a complex developmental condition involving persistent challenges with social communication, restricted interests, and repetitive behavior. While autism is considered a lifelong disorder, the degree of impairment in functioning because of these challenges varies between individuals with autism. Early signs of this disorder can be noticed by parents/caregivers or pediatricians before a child reaches one year of age. Typically, most symptoms develop more consistently noticeable by the time a child is 2 or 3 years old. In some cases, the functional impairment related to autism may be mild and nonexistent until the child attends school, after which their deficits may be pronounced when amongst their peers.

Justify why it is important/useful to solve this problem

Being able to improve predicting autism traits will definitely help parents get ahead of the curve set up future services that will benefit their children in the future. Although there are parents who don't like their child to be labeled, an accurate diagnosis of autism can help to prevent delays in getting the support and help needed. Knowledge becomes the tool. This is also beneficial to the healthcare providers as well by aiding them in prioritizing their recourses. Overall, the screening and model prediction will allow parents and providers to devise a long-term comprehensive plan that will be beneficial to the child and their future.

How would you pitch this problem to a group of stakeholders to gain buy-in to proceed?

Detecting and treating Autism Spectrum Disorder in its early stages are extremely crucial as this helps to decrease or alleviate the symptoms to a certain extent, thus improving the overall quality of life for the individual. Machine learning offers a way for computer systems to learn and improve from experience continuously. Data is becoming more and more prevalent in the field of business and health. Using data-driven decisions to be able to prevent or detect autism can be a huge breakthrough for many families and healthcare providers abroad. This model can help pave the way for this to become a reality.

Explain where you obtained your data

The dataset used for this analysis is the Autism Screening dataset from Kaggle. This dataset is composed of survey results from approximately 700 people who completed an app form. There are labels portraying whether the person received a diagnosis of autism, allowing machine learning models to predict the likelihood of having autism, therefore allowing healthcare professionals prioritize their resources. The following variables are represented in the dataset:

- A1_Score I often notice small sounds when others do not

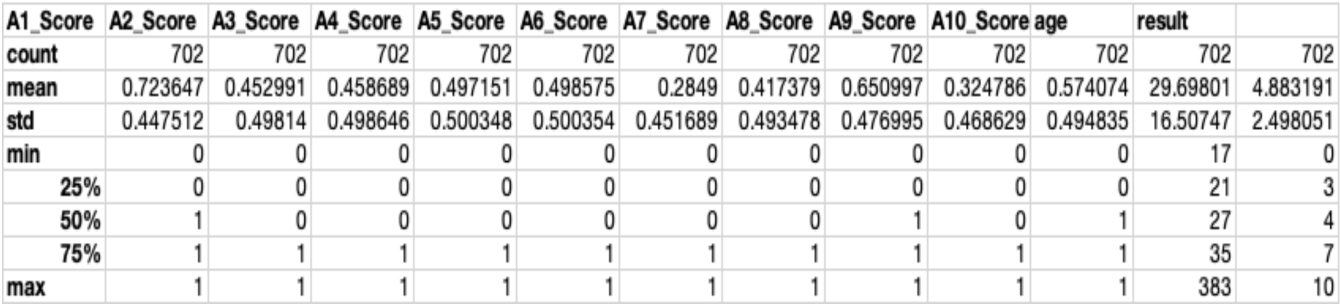
- A2_Score I usually concentrate more on the whole picture, rather than the small details
- A3_Score I find it easy to do more than one thing at once
- A4_Score If there is an interruption, I can switch back to what I was doing very quickly
- A5_Score I find it easy to 'read between the lines' when someone is talking to me
- A6_Score I know how to tell if someone listening to me is getting bored
- A7_Score When I'm reading a story I find it difficult to work out the characters' intentions
- A8_Score I like to collect information about categories of things (e.g. types of cars, birds, trains, etc)
- A9_Score I find it easy to work out what someone is thinking or feeling just by looking at their face
- A10_Score I find it difficult to work out people's intentions
- age: Participant age in years
- gender: Participant gender
- ethnicity: Participant ethnicity
- jundice: Whether or not the participant was born with jaundice?
- austim: Whether or not anyone in tbe immediate family has been diagnosed with autism?
- contry_of_res: Countries
- used_app_before: Whether the participant has used a screening app
- result: The total score from the AQ-10 screen
- age_desc: Age as categorical
- relation: Relation of person who completed the test
- Class/ASD: Participant classification

AQ-10 SCORING: Only 1 point can be scored for each question. Score 1 point for Definitely or Slightly Agree on each of items 1, 7, 8, and 10. Score 1 point for Definitely or Slightly Disagree on each of items 2, 3, 4, 5, 6, and 9. If the individual scores 6 or above, consider referring them for a specialist diagnostic assessment.

Organized and detailed summary of Milestones 1-3

EDA; include any visuals you think are important to your project

The original dataset consisted of 704 rows and 21 columns. Below is table of statistical information of the data.

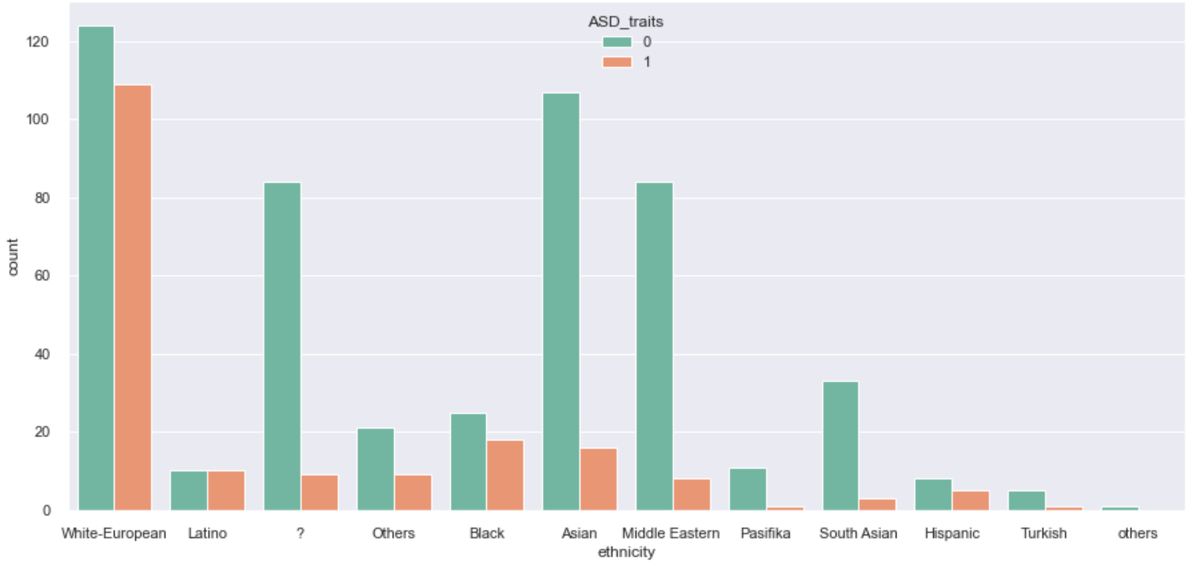


A1_Score	A2_Score	A3_Score	A4_Score	A5_Score	A6_Score	A7_Score	A8_Score	A9_Score	A10_Score	age	result	
count	702	702	702	702	702	702	702	702	702	702	702	702
mean	0.723647	0.452991	0.458689	0.497151	0.498575	0.2849	0.417379	0.650997	0.324786	0.574074	29.69801	4.883191
std	0.447512	0.49814	0.498646	0.500348	0.500354	0.451689	0.493478	0.476995	0.468629	0.494835	16.50747	2.498051
min	0	0	0	0	0	0	0	0	0	0	17	0
25%	0	0	0	0	0	0	0	0	0	0	21	3
50%	1	0	0	0	0	0	0	1	0	1	27	4
75%	1	1	1	1	1	1	1	1	1	1	35	7
max	1	1	1	1	1	1	1	1	1	1	383	10

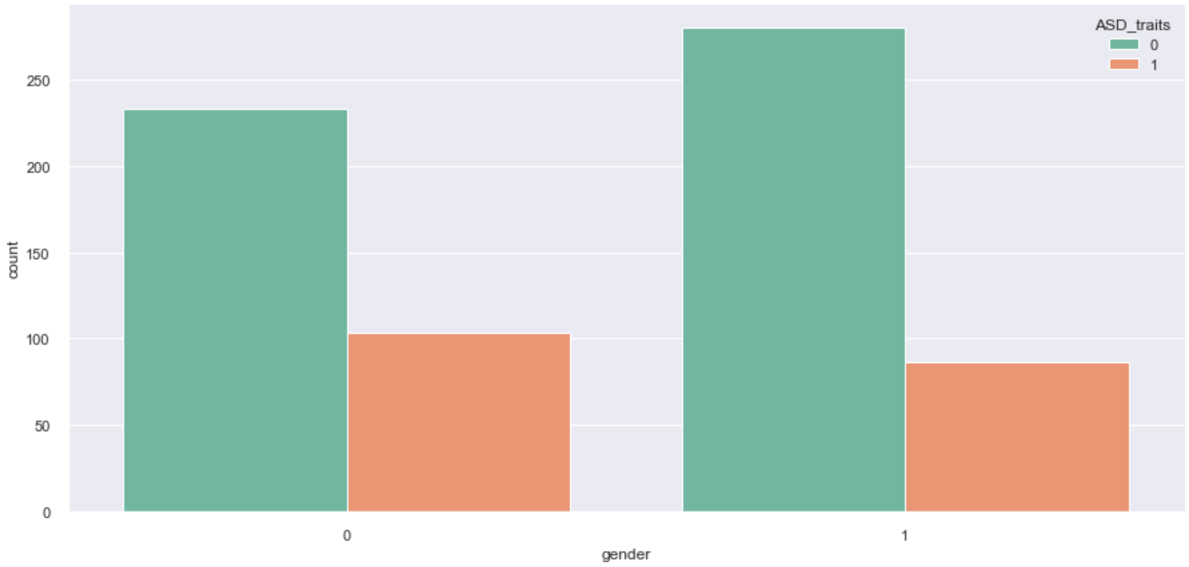
Keiuntae Smith
DSC550
Data Mining
May 31, 2022

Below are a few graphical visualizations of dataset:

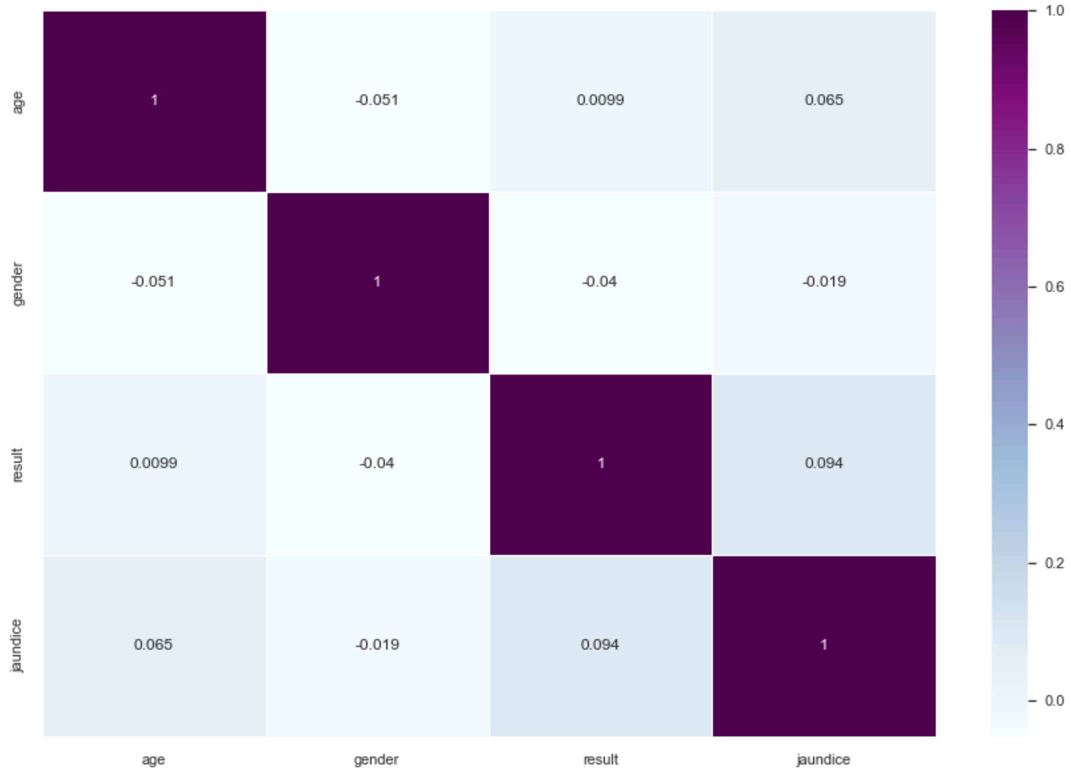
Visualization 1: Bar graph to depict autism spectrum disorder traits amongst different races.



Visualization 2: Bar graph to visualize autism spectrum disorder traits among gender.



Visualization 3: Heatmap visualization to display the correlation of variables across the data set



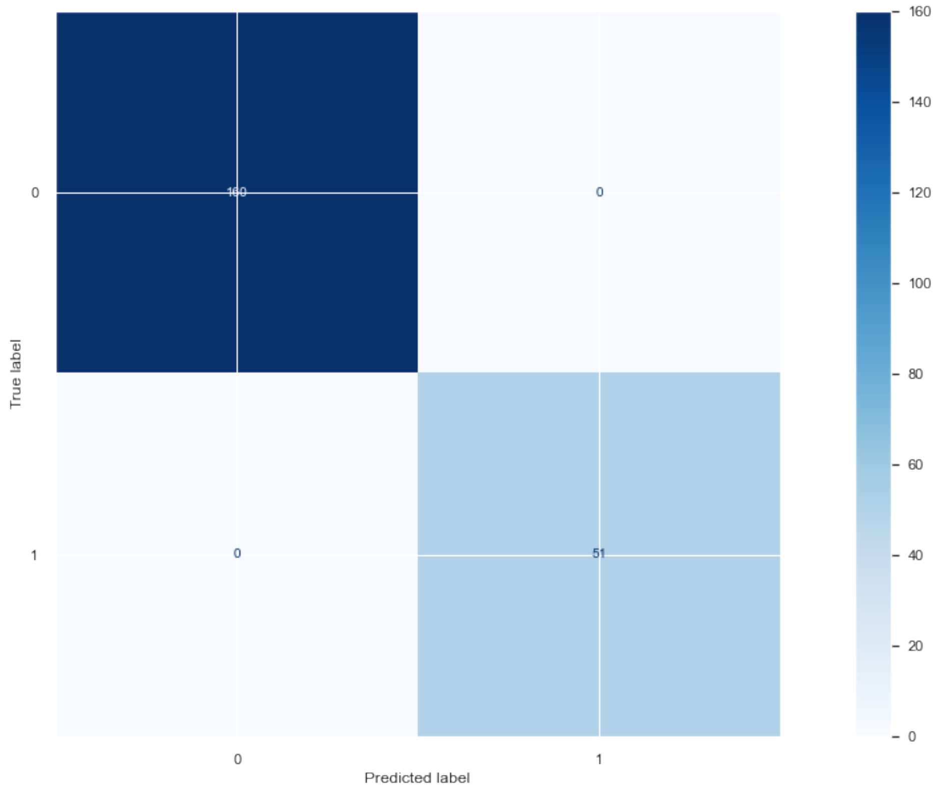
Data preparation

There were a few conversions that needed to be made to the data set to change some of the string values to numerical values. All of the yes and no were converted to 0s and 1s to facilitate implementation of the analysis. This conversion was applied to the following variables: autism, jaundice, gender and class/asd. I also renamed some of the misspelled categories to eliminate any future confusions this may cause. The word jaundice was misspelled ‘jundice’ and the word autism was misspelled ‘austim’. I also dropped a few columns because they would bear of no importance to the analysis. The columns dropped were 'contry_of_res','used_app_before', 'age_desc','relation'. Once these transformations were completed and the dataset was transformed to numerical values, I was able to come up with a final cleaned data set that was prepared for the analysis process.

Model building and evaluation

Now that the data had been prepared and cleaned, the next process was to build a model for evaluation. To do this, the data had to be split between a test and train, meaning that the test set consisted of 20% of the data and the training entailed the remaining 80%. This was accomplished by using the Scikit-learn library. This library also has the Random Forest classifier to which I applied to the data. The Random forest is a collaborative learning method for classification and regression. This model works by constructing a multitude of decision trees at training time and outputting the class that is the comprised of the classes or mean prediction of the individual trees.

Once we instantiated and fit the model, it produced an outcome of a 100% accuracy. The confusion matrix displays the following:



Conclusion

What does the analysis/model building tell you?

I also wanted to experiment with other machine learning algorithms to see how they fared in comparison to the Random Forest Classifier algorithm. The findings were as follows: The accuracy of prediction using the Random forest classification algorithm was 100%. The accuracy of prediction using the k-nearest neighbors algorithm was 97.1%. The accuracy of prediction using the Naïve Bayes classifier algorithm was 96.9%. The accuracy of prediction using the Support Vector Machine algorithm was 100%.

Is this model ready to be deployed?

I believe this model is a good start and has good qualities that could lead to implementation but I do not believe it is actually ready for deployment. I believe that more data is needed in order to make the model more reliable and trustworthy. Accuracy is vital to any machine learning model and is the most often spoken about. Without accurate predictions, there is no purpose for deploying the algorithm. With that being said, I am not fully confident in the 100% accuracy of the model.

What are your recommendations?

My recommendation of this analysis is to acquire more knowledge and input of the data set. The dataframe should include even more screening questions to fine tune the analysis process. Before deploying the machine learning model, it is essential to evaluate the performance of your model on real-world data. If your model is trained on a clean and robust dataset, it is important to generate a dataset that imitates real-world data as closely as possible and evaluate the model on this. This is especially significant when the dataset does not come from exactly the same source as the production data.

What are some of the potential challenges or additional opportunities that still need to be explored?

One of the potential challenges that could derail the analysis is the honesty of the screening questions of the participants. Entering erroneous answers or misleading entries can contaminate the final dataset.