PROJECT REPORT

Detecting Autism Spectrum Disorder in Adults and Children

Introduction

Autism Spectrum Disorder (ASD) is a developmental disorder that affects communication and behavior. Although autism can be diagnosed at any age, it is said to be a "developmental disorder" because symptoms generally appear in the first two years of life (Association, 2013; NIMH, October 2016). The diagnosis rate of ASD has increased almost four times since 2000s, which leads to great debate whether we missed lots of diagnoses in the past, the modern environment is toxic for children's development, or it's more about misdiagnosis (Parry, 2016).

ASD is known as a "spectrum" disorder because of the wide variation in the type and severity of symptoms that patients may experience. ASD occurs in all ethnic, racial, and economic groups. Long term issues may include difficulties in performing daily tasks, creating and keeping relationships, and maintaining a job (Comer, 2007). Although ASD can be a lifelong disorder, treatments and services can improve a person's symptoms and ability to function. The American Academy of Pediatrics (AAP) recommends that all children be screened for autism.

One of the important problems in ASD research is to make ASD diagnosis more accessible to more and younger individuals to help health professionals and to inform individuals whether they should pursue formal clinical diagnosis and treatment. However, waiting times for a traditional ASD diagnosis are lengthy and procedures are not cost effective. The 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.

Over the past decade, many extensive and/or shortened questionnaires have been created and tested to help families determine whether further professional assessment is necessary for a child (Parry, 2016). The implementation that researchers have done using machine learning methods for the diagnosis of ASD is nascent. The classification/diagnosis for ASD is particularly difficult is because there are other explanations for similar behavioral symptoms such as learning disabilities. Some healthy individuals might exhibit the same behavior pattern.

In this project, I have built and compared several classifiers to help with the diagnosis of ASD using the *Autistic Spectrum Disorder Screening Data for Adults and Children* data sets. These classifiers could be useful for proposing possible new methods of ASD screening, which should be time-efficient and accessible to help health professionals and inform individuals whether they should pursue formal clinical diagnosis.

Data Description

The data set for the project are two data sets called *Autistic Spectrum Disorder Screening Data for Adults and Children*, respectively. They are available on the University of California, Irvine (UCI) Machine Learning Repository (F. Thabtah, 2017).

The data sets are related to autism screening of adults and children that contained 20 features (see table 1 for detailed descriptions) to be utilized for further analysis, especially in determining influential autistic behavioral traits and improving the classification of ASD cases. Ten behavioral features (from AQ-10-Adult and AQ-10-Child)(Allison, Auyeung, & Baron-Cohen, 2012; NICE, 2012) were recorded plus ten individuals' characteristics that have been proved to be effective in detecting ASD from research in behavior science.

The data set for children contains 292 instances for children aged 4 to 11, 141 with ASD (48.28%) and 151 without (51.72%). The data set for adults contains 704 instances for adults, 189 with ASD (26.85%) and 515 without (73.15%). The data sets are stored as *.arff files. The missing data are marked as NumPy NaN in python and may convert into numerical values if needed.

Table 1. Features and their descriptions (F. F. Thabtah, 2017)

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	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	Binary (0, 1)	The answer code of the question based on the screening method used
Question 2 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 3 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 4 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 5 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 6 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 7 Answer	Binary (0, 1)	The answer code of the question based on the screening method used

Question 8 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 9 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 10 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Screening Score	Integer	The final score obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner

These two data sets are established for the ASD classification problem. It has appropriate amount of data instances and relatively balanced class distributions. Most of the features are demographical information, medical terms or concepts related to behavioral science. Therefore, they should give good interpretation if needed.

One possible approach with assumption that there is no difference in behavioral traits in adults and children with ASD for classification is to use the adult data set for feature selection and classifier training because it has a larger sample size and apply the selected features and classifier to children data set. This assumption might hold because that the attributes in both data sets are the same and the behavioral questions are the same semantically. But they still might present different behavioral patterns. For this approach, results comparison between the two data sets is an interesting problem to look at and should be performed to verify the assumption. Another possible approach is to combine the two data sets into a larger data set while keeping adult/child as an extra feature. Then we can apply classifier to this new data set with train/test split. The latter approach is selected in the project as it provides more flexibility and no assumption is needed. The combined data set contains 996 instances for adults or children, 330 with ASD (33.13%) and 151 without (66.87%).

Figure 1 displays a T-SNE dimensions for adult data set. Figure 2 displays a T-SNE dimensions for children data set. Both seem to be separable after dimension reduction. Figure 3 displays a T-SNE dimensions for combined data set. As shown in figure 3, adults and children data are separable, i.e. they exhibit different patterns. Therefore, the latter approach is indeed better.

Figure 1. t-SNE visualization for adult data set

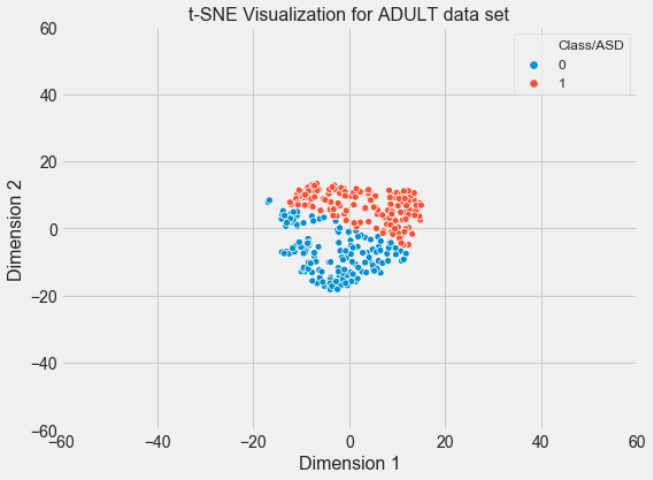

Figure 2. t-SNE visualization for children data set

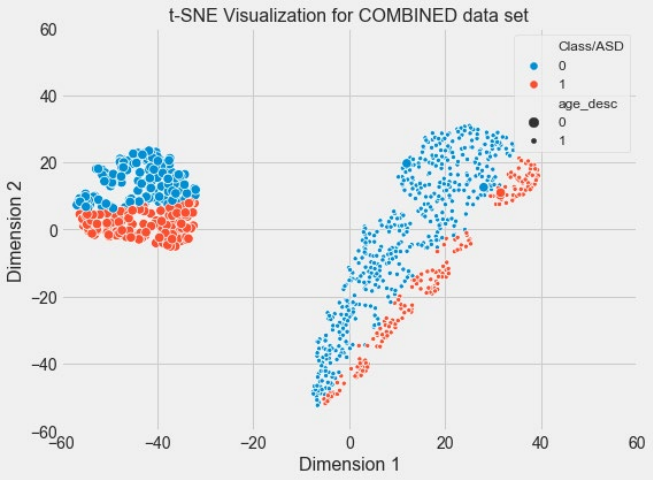

Figure 3. t-SNE visualization for combined data set, data are labeled by age category and ASD class, both of them are separable from visual inspection

Figure 4 displays a correlation heatmap for adult data set. Figure 5 displays a correlation heatmap for children data set. Figure 6 displays a correlation heatmap for combined data set.

We can see that the correlation between result and Class/ASD is strong. In figure 6, the correlation between age category and Class/ASD is weak, which reflects that adults and children behaves differently. Table 2 shows all the statistics for the combined data set. Age and results columns are more meaningful because other columns are binary values.

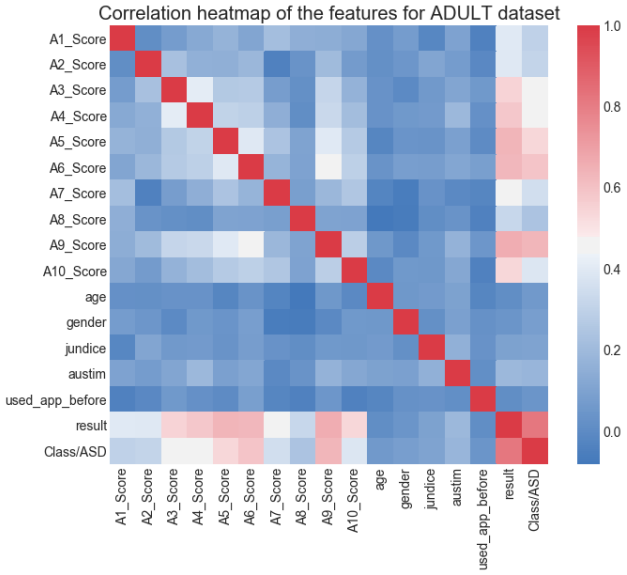

Figure 4. Correlation map of the features for adult data set

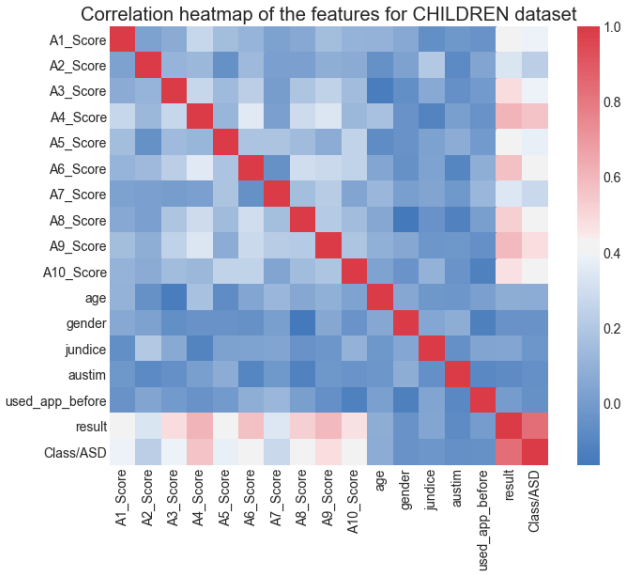

Figure 5. Correlation map of the features for children data set

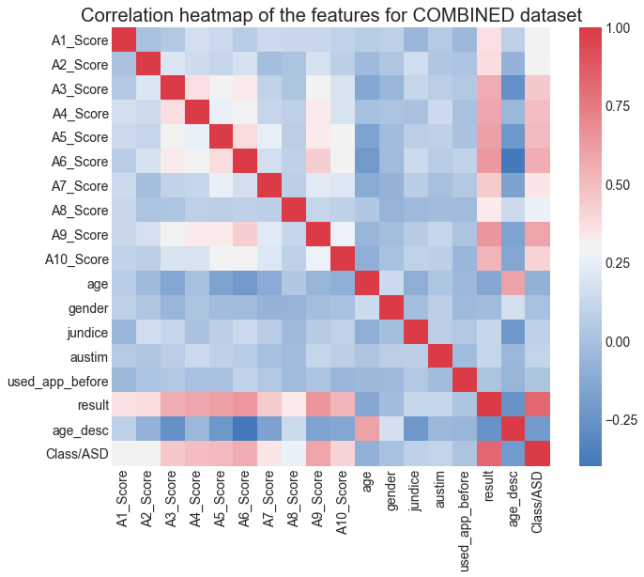

Figure 6. Correlation heatmap for combined data set

Table 2. Descriptive statistics of the combined data set

Features	count	mean	std	min	25%	50%	75%	max
A1_Score	996	0.695783	0.460306	0	0	1	1	1
A2_Score	996	0.476908	0.499717	0	0	0	1	1
A3_Score	996	0.541165	0.498553	0	0	1	1	1
A4_Score	996	0.512048	0.500106	0	0	1	1	1
A5_Score	996	0.570281	0.495285	0	0	1	1	1
A6_Score	996	0.409639	0.492014	0	0	0	1	1
A7_Score	996	0.472892	0.499515	0	0	0	1	1
A8_Score	996	0.604418	0.489221	0	0	1	1	1
A9_Score	996	0.373494	0.483975	0	0	0	1	1
A10_Score	996	0.618474	0.486005	0	0	1	1	1
age	996	22.90161	17.47694	4	10	22	30	383
gender	996	0.422691	0.494235	0	0	0	1	1
jundice	996	0.149598	0.356857	0	0	0	0	1
austim	996	0.140562	0.347744	0	0	0	0	1
used_app_before	996	0.023092	0.150272	0	0	0	0	1
result	996	5.2751	2.516802	0	3	5	7	10
age_desc	996	0.706827	0.455446	0	0	1	1	1
Class/ASD	996	0.331325	0.470926	0	0	0	1	1

Approach/Methods

Preprocessing

The initial step is to clean the data set and create model matrix. The original data set is saved as *.*arff* file and all the variables are bytes in python. Missing data are denoted as '?' in the original data set. There are both categorical and numerical columns. The problem is that these data types cannot be interpreted by the classifier. Thus, the data type needs to be converted to numerical values. Binary classes are replaced with 0/1 and for more classes creating dummy variables is a more feasible method. The missing values are filled with the mean of the given column, which is performed after the training and testing set are split. In the meantime, indications of missing values are added as new features, these columns have suffix "_mv". Model matrix is all numerical (except for certain null values) and ready for classifiers to process.

After the model matrix is created, it should be split into training and testing set with an appropriate testing size (e.g. 0.33). After the splitting, fill null values with the column mean for training set, as doing so before the splitting will "leak" testing set information to the training process. Finally, standardization or normalization can be applied to the features. Standardize features by removing the mean and scaling to unit variance. The standard score of a sample x is calculated as:

$$z = (x - u) / s \quad (1)$$

where u is the mean of the training samples, and s is the standard deviation of the training samples. Centering and scaling happen independently on each feature by computing the relevant statistics on the data in the training set(scikit-learn, 2018). Mean and standard deviation are then stored to be used on later data using the transform method. Standardization of a data set is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data (e.g. Gaussian with 0 mean and unit variance).

After the standardization, the data is used to train and test the classifiers. No feature selection is performed. The reason is that the number of features is not high therefore Principle Component Analysis (PCA) method is not necessary. Another reason is that we need to interpret the results because this classification is for medical diagnosis.

Classifier and Performance

As shown in Table 3, we can use the classifiers covered in class to do classification, like linear regression, logistic regression, Naive Bayes, k Nearest Neighbor(kNN), Support Vector Machine (SVM). The implement of these classifiers is through Scikit-learn package in python(Pedregosa et al., 2011).

Linear Regression fits a linear model with coefficients

$$w = (w_1, \dots, w_p) \quad (2)$$

to minimize the residual sum of squares between the observed responses in the data set, and the responses predicted by the linear approximation. The target value is expected to be a linear combination of the input variables. Mathematically it solves a problem of the form(Seber & Lee, 2012):

$$\min_{\mathrm{w}} \left|\left|\mathrm{Xw}-\mathrm{y}\right|\right|_2^2 \quad (3)$$

if $\hat{y}$ is the predicted value, then

$$\hat{y}(w,x) = w_0 + w_1x_1 + \ldots + w_px_p$$

Logistic regression is a linear model for classification, the probabilities of the possible outcomes are modeled using a logistic function. The cost function for binary class L2 penalized logistic regression is(Bishop, 2006):

$$\min_{w,c} \frac{1}{2}w^Tw + \sum_{i=1}^{n} \log\left(\exp\left(-y_i(X_i^Tw + c)\right) + 1\right) \quad (4)$$

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable.

The classification rule is

$$\begin{gathered} P(y \mid x_1, \ldots, x_n) \propto P(y)\prod_{i=1}^{n} P(x_i \mid y) \\ \Downarrow \\ \hat{y} = \arg\max_y P(y)\prod_{i=1}^{n} P(x_i \mid y) \end{gathered} \quad (5)$$

The different naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of $P(x_i \mid y)$. Here it is assumed to be Gaussian:

$$\mathrm{P}(\mathrm{x_i} \mid \mathrm{y}) = \frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left(-\frac{(\mathrm{x_i}-\mu_y)^2}{2\sigma_y^2}\right) \quad (6)$$

SVMs are a set of supervised learning methods used for classification and regression. The advantages of support vector machines are: (a) Effective in high dimensional spaces; (b)Still effective in cases where number of dimensions is greater than the number of samples. Given training vectors $x_i \in \mathbb{R}^p$, I = 1, … , n, in two classes, and a vector y∈ [1, −1] ^n, SVC solves the following primal problem: The decision function is:

$$\mathrm{sgn}\left(\sum_{i=1}^{n} y_i\alpha_i K(x_i,x) + \rho\right) \quad (7)$$

where $K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$ is the kernel. Here training vectors are implicitly mapped into a higher (maybe infinite) dimensional space by the function φ.

kNN classification is computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the

nearest neighbors of the point(Bentley, 1975). The optimal choice of the value k is highly data-dependent: in general, a larger k suppresses the effects of noise, but makes the classification boundaries less distinct. The decision statistic is calculated as follows (Note I(x) is the indicator function which evaluates to 1 when the argument x is true and 0 otherwise):

$$P(y = j|X = x) = \frac{1}{K} \sum_{i \in \mathcal{A}} I(y^{(i)} = j) \quad (8)$$

Cross validation is also implemented by separating the training set into training set and validation set and resampling the data set. Because the number of features is not two-dimensional, it's difficult to plot decision statistic surface, we can only evaluate the classifier through other performance metrics.

After the classifier is trained, we can test the classifier on the testing set to get accuracy. Other performance metrics like confusion matrix, Receiver operating characteristic (ROC), F1 score, precision-recall curve (PRC) can be used to assess the model.

Design trade-offs are crucial for ASD detection. As the damage for both types of mistakes are high, we want sensitivity and specificity stay balanced. Precision-recall curve is helpful to consider the problem. Normally ROC is a poor metric for medical data. However, this specific data set is not extremely imbalanced thus ROC is still worthy of consideration. After all the procedures above, tune the parameters of the classifier and repeat the process multiple times until a classifier with good performance is acquired. Figure 7 illustrates such process.

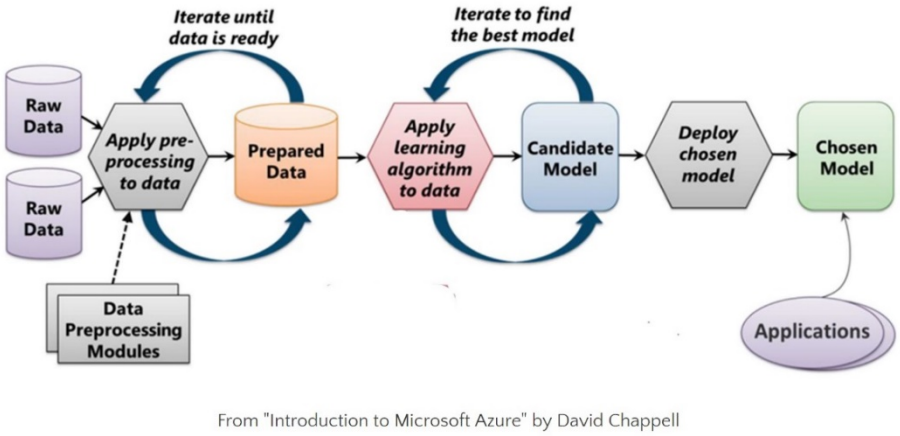

Figure 7. The process of classification

Results and Discussion

The model matrix initially has 996 rows and 23 columns/features. After training and testing split, the training set has 667 rows and 23 columns/features. It's all numerical without null values and the values are standardized.

Table 3 shows is the accuracy and F1 score for all the classifiers. Logistic Regression and SVM have the highest accuracy while kNN, Naïve Bayes and Linear regression are not as good but still high.

Table 3. The accuracy and F1 score for all classifiers

Classifiers	Accuracy (%)	F1 Score (%)
Logistic Regression	100.0	100.0
SVM	100.0	100.0
Linear Regression	94.53	92.04
kNN(k=7)	94.22	91.77
Naïve Bayes	94.22	91.48

Figure 8 is the ROC for all the classifiers and figure 9 is the PRC. SVM and Logistic Regression have perfect performance. The rest have good scores as well.

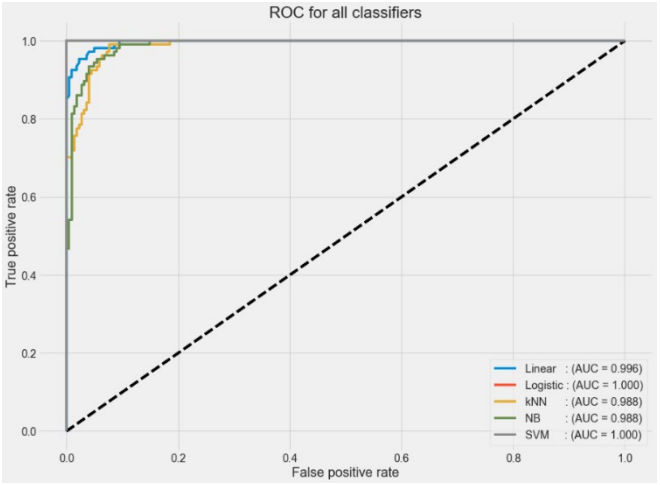

Figure 8. The ROC for all classifiers

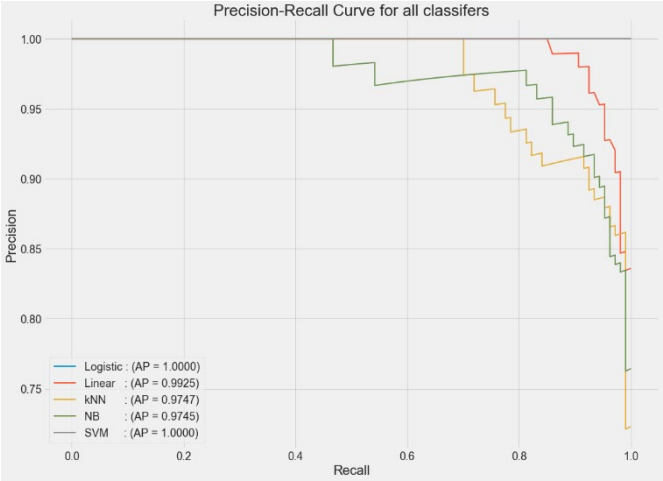

Figure 9. The PRC for all classifiers

Figure 10 is the confusion matrices for all classifiers. SVM and Logistic Regression are great. KNN has low true negative rate which is not expected given the good performance

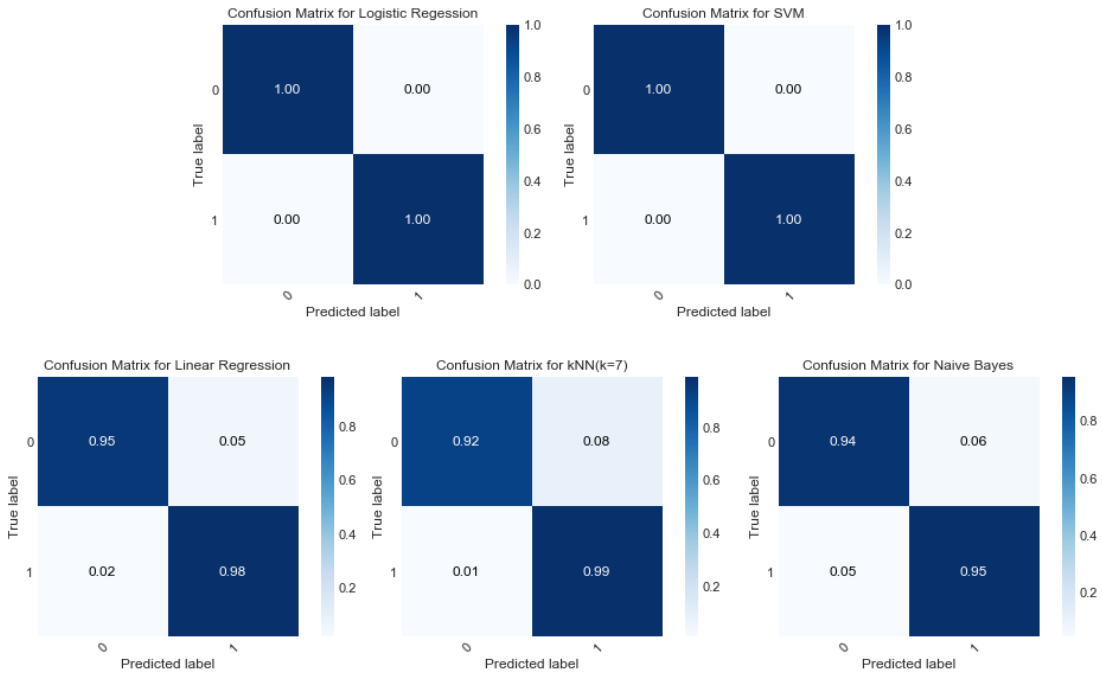

Figure 10. The Confusion Matrices for all classifiers

Figure 11 shows the feature importance for all the classifiers. All the classifiers have the same top one feature: result. This agrees with our expectation from the strong correlation between result and Class/ASD we saw in the Figure 6. Logistic Regression and SVM both show high importance for the almost only the scores and other information are not important. However, for other classifiers, they value information from other columns as well. Naïve Bayes view age group, missing values, autism as important feature after the test scores. kNN treat features that are not result almost the same. Linear Regression treats age as an important factor. All three columns containing age information have relatively high importance. The comparison between

the two best performance classifiers and the other three less good classifiers indicates that the diagnosis of ASD is most defined by the behavioral patterns. The differences in age, gender, ethnicity might not be minimal. Perhaps demographic information does not serve as good features.

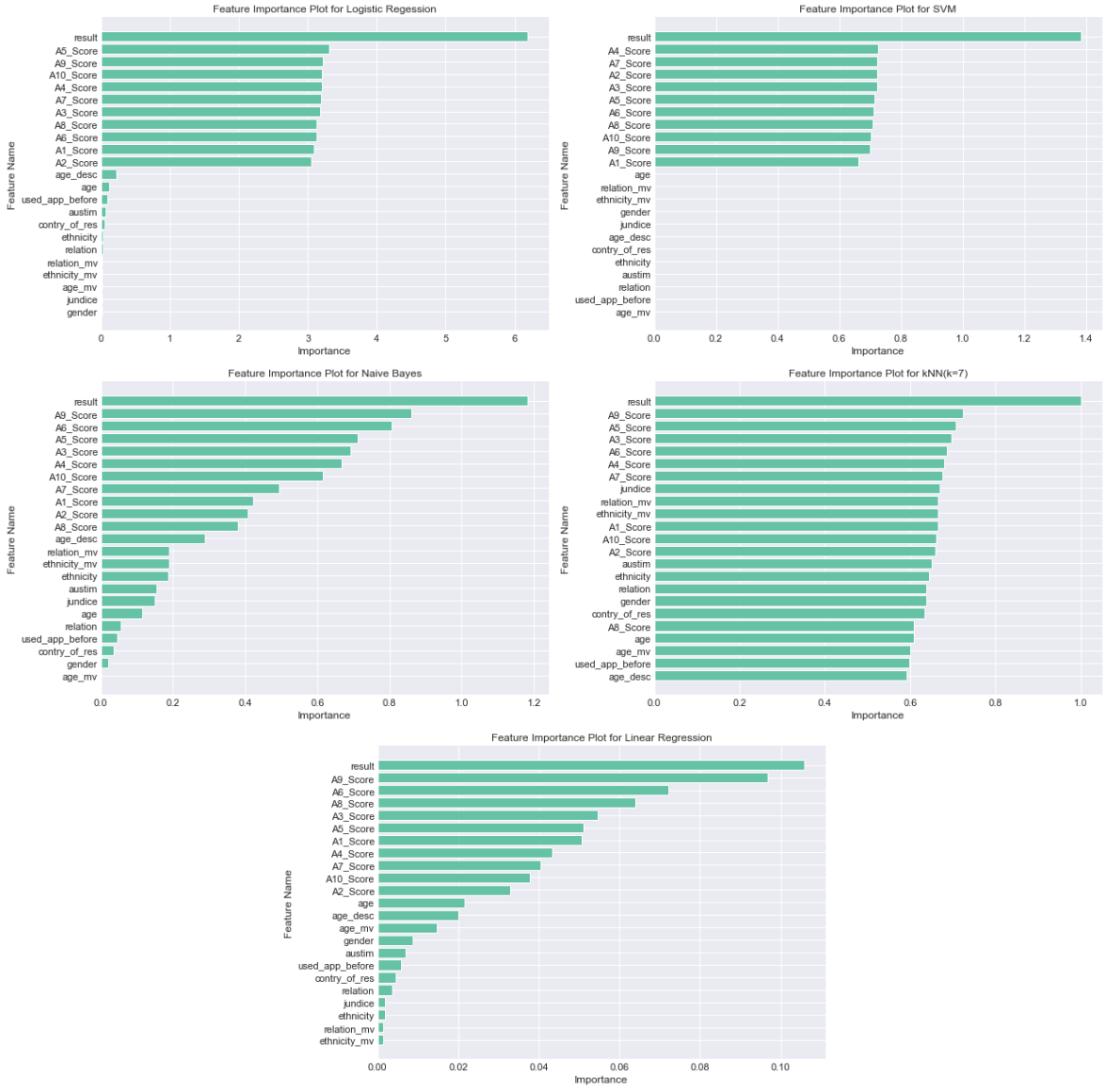

Figure 11. The Feature Importance Plot for all classifiers

Considering all the performance metrics, the classifiers have good performances in general. Not only they achieve high accuracy, their ROCs and PRCs are also very good. The classifiers sorted by the performances are: Logistic Regression = SVM > Linear Regression > kNN > Naïve Bayes.

Conclusion

Comparing all the classifiers, Logistic regression and SVM have the best performance. Linear Regression, KNN and Naïve Bayes have their shortcomings.

For Logistics regression and SVM, they show the same pattern when assigning feature importance. The most important and only important features are the test scores. It directly connects to the behavioral traits. Other features are not important. The good performance is foreseeable when we draw the t-SNE plot. These data are linear separable in higher dimensions.

For Linear regression, although the ROC and PRC both performs great, the accuracy is not as high. Linear Regression also has high average precision score, which is very good. From t-SNE plot, we can see several outliers, this might be the main reason Logistic Regression outperforms Linear Regression.

For KNN, it also has high average precision, but the true negative rate is low compared to other classifiers. KNN treats features other than result almost the same, which impairs its performance.

For Naïve Bayes, all the metrics are the lowest among the five classifiers. Naïve Bayes has strong assumptions about data distributions, this might be the limitation of the classifier. All the three classifiers show higher feature importance for the features that are not test scores than SVM and Logistic regression. That's probably why they perform less well.

ASD can be thought of as a disease that causes a limitation of communication and social movements as a result of deterioration in brain development. Communication ability and social behavior disorder affect people's whole life. This is why diagnosis of ASD is important for both adults and children.

In this project, we compared several different classifiers to do classification on the ASD data set containing both adult and children patients and control groups. Linear Regression, Logistic Regression, SVM, kNN, Naïve Bayes are the five classifiers used. After training, classifiers' performances were evaluated by several performance metrics including accuracy, ROC, PRC, confusion metrices and F1 score. In addition, feature importance plots are helpful to interpret the results. According to the classifiers' performances, we can see that classification using the proposed data set is successful. Logistic regression and SVM have the best performance with AUC = 1. These classifiers should be implemented to the application where the data set comes from. This would be an effective, low-cost and accessible method towards the diagnosis and screening of ASD.

For future work, although the classifiers have achieved great results, the dataset used in this project is limited because it does not contain large-scale instances with both Autistic and non-Autistic children. An ideal Machine Learning dataset would contain thousands of samples from all diagnostics categories. Another experiment is to use only the scores as features for training. Excluding the demographic information might improve performance of classifiers.

Reference

Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief "red flags" for autism screening: the short autism spectrum quotient and the short quantitative checklist in 1,000 cases and 3,000

controls. *Journal of the American Academy of Child & Adolescent Psychiatry, 51*(2), 202-212. e207.

Association, A. P. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*: American Psychiatric Pub.

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM, 18*(9), 509-517.

Bishop, C. M. (2006). Pattern recognition and machine learning. In: springer.

Comer, R. J. (2007). *Fundamentals of Abnormal Psychology Student Workbook*: Macmillan.

NICE. (2012). Autism spectrum disorder in adults: Diagnosis and management. In: National Institute of Clinical Excellence London.

NIMH. (October 2016). Autism Spectrum Disorder. Retrieved from https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd/index.shtml

Parry, P. (2016, January 28, 2016). The difficulties doctors face in diagnosing autism. Retrieved from https://theconversation.com/the-difficulties-doctors-face-in-diagnosing-autism-53731

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research, 12*(Oct), 2825-2830.

scikit-learn. (2018). Preprocessing data. Retrieved from https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing-scaler

Seber, G. A., & Lee, A. J. (2012). *Linear regression analysis* (Vol. 329): John Wiley & Sons.

Thabtah, F. (2017). *Autism spectrum disorder screening: machine learning adaptation and DSM-5 fulfillment.* Paper presented at the Proceedings of the 1st International Conference on Medical and Health Informatics 2017.

Thabtah, F. F. (2017). Autism Screening Adult Data Set Retrieved from http://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult