Final report

Proposal

Autism, or autism spectrum disorder (ASD), refers to a broad range of conditions characterized by challenges with social skills, repetitive behaviors, speech and nonverbal communication. Early diagnosis of and interventions for autism are more likely to have major long-term positive effects on symptoms and later skills[1]. But because of the shortage of medical resources, the waiting list for the ASD diagnosis can be as long as one year. So, a time-efficient and easily accessible ASD screening will be very helpful for the parents whose kids perform autistic behaviors.

In this project, I am going to analyze a dataset developed by Dr. Fadi Fayez Thabtah who uses a mobile app called ASDTests to screen autism in toddlers. Based on the information from the analysis, I am aiming to find a modeling method that can best predict whether a kid is autistic or not. The database will be downloaded form kaggle.com[2]. Although the data set is not a big one, which only have 1045 observations, it is enough to do some simple trainings. And some functions such as cross-validation will be used to compensate the shortages caused by small datasets.

Since the dataset will be used to predict whether a kid can be diagnosed as ASD, the modeling problem is a classification problem. Six main modeling methods that are usually used for classification will be employed to the project, including Decision Tree Classifier (DT), Random Forrest Classifier (RF), Logistic Regression (LR), Gaussian NB (NBG), Multinomial NB (NBM), and K Neighbors Classifier (KNN). Numpy, Pandas and sklearn packages from Python software will be used for the modeling, and the seaborn and matplotlib packages from python will be used for graphing. To get enough background knowledge to finish the project, information from the course lectures[3], scikit-learn user guide[4] and other online resources[5-7] will be used as supports.

The brief schedule of the project will be as follows. First, exploratory data analysis (EDA) will be used to summarize their main characteristics of the dataset. Then, grid search will be used to identify the best parameters for each modeling methods. After figuring out the appropriate parameters, the performance of different modeling methods, as well as the ensembled methods will be compared and the accuracy score will be used for the evaluation. Finally, the method that has the highest accuracy score will be used for the final predicating, and performance of the modeling method in predicating ASD will be evaluated by the accuracy score, confusion matrix and the classification report.

Introduction:

This project will use a dataset related to autism screening of toddlers identify a modeling method that can predict whether a toddler is autistic. The dataset contains influential features to be utilized for further analysis especially in determining autistic traits and improving the classification of ASD cases. The dataset recorded ten behavioral features (Q-Chat-10) plus other individuals characteristics that have proved to be effective in detecting the ASD cases from controls in behavior science. Based on Quantitative Checklist for Autism in Toddlers (Q-CHAT) data provided by the ASD Tests app, this project will try to develop a simple prediction model for toddlers to predict the probability of showing ASD traits so that their parents/guardians can consider taking steps early enough.

This project report includes following information: **1)** The description of the data set; **2)** The experimental methodology of the project; **3)** The results of the project; **4)** Summary and conclusions of the project; **5)** References that have been used to complete the project.

Data set description

Data set "Autism screening data for toddlers" collected data from 1054 toddlers whose ages were less than three years old. The Doctor asked their parents yes/no questions. These questions were behavioral questions, and the number of yeses were added up and have been recorded in the column called "Qchat-10-Score". If this column has more than 3 yes, toddlers will be classified as positive ASD. The ten questions are:

A1: Does your child look at you when you call his/her name?
A2: How easy is it for you to get eye contact with your child?
A3: Does your child point to indicate that s/he wants something?
A4: Does your child point to share an interest with you?
A5: Does your child pretend? e.g. care for dolls, talk on a toy phone?
A6: Does your child follow where you are looking?
A7: If you or someone else in the family is visibly upset, does your child show signs of waning to comfort them? e.g. stroking hair, hugging them)
A8: Would you describe your child's first word as:
A9: Does your child use simple gestures (e.g. wave goodbye)?
A10: Does your child stare at nothing with no apparent purpose?

Besides the questions and Qchat-10-Score features, the data set all includes other features such as "Age_Mons", "Sex", "Ethnicity", "Jaundice", "Family_mem_with_ASD", and "Who completed the test".

A total of 1045 observations and 18 features were included in the dataset.

Experimental methodologies:

1. Modeling methods:

Decision Tree Classifier (DT), Random Forrest Classifier (RF), Logistic Regression (LR), Gaussian NB (NBG), Multinomial NB (NBM), and K Neighbors Classifier (KNN) were used in this project. All of them are widely used to solve the classification problems. DT, RF, LR and KNN were also ensembled and hard/soft voting were used to identify the best prediction model.

2. Hyperparameter tuning

Grid search is a tuning technique that attempts to compute the optimum values of hyperparameters[8]. It is an exhaustive search that is performed on the specific parameter values of a model. Grid search exercise can save us time, effort and resources[8]. In this project, grid search is used to determine the best hyperparameters for each modeling methods.

3. Performance evaluation:

The data set was split into two parts: the training data (80% of the total dataset) and the testing data (20% of the total dataset). All the proposed modeling methods were trained with the training data, and their performances were tested using the testing data, and were evaluated by the accuracy score. After developing the best model, the testing data was used again for the final test, and confusion matrix and the classification report were also used for the evaluation.

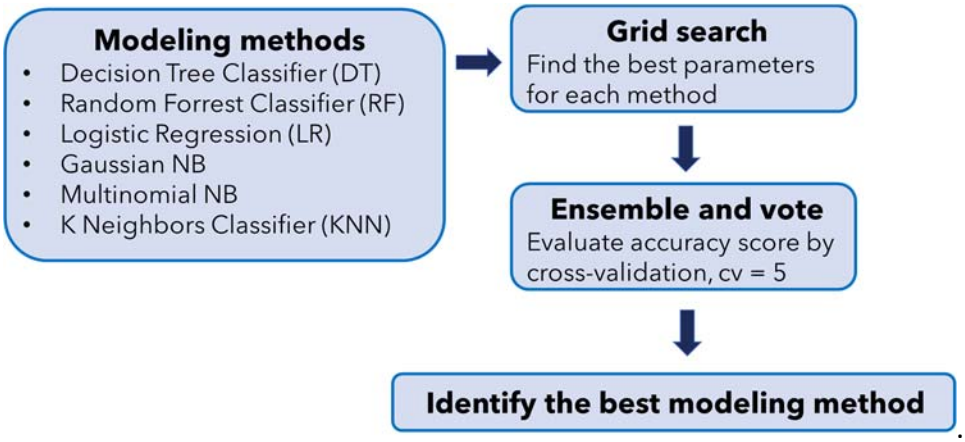
4. Information about the codes

1) *Read and view the basic information of the dataset*: Line 1-10.
2) *Check missing value, value in wrong format and outliers*: Line 13-24. Unique values in each column and their counts were checked, and the outliers were checked via box graphs.
3) *EDA*: Line 27-109. For EDA, how the features of "Age_Mons", "Sex", "Ethnicity", "Jaundice", "Family_mem_with_ASD", and "Who completed the test" will related to the classification were visualized via bar graphs and histogram graphs.
4) *Preprocessing and feature selection*: Line 111-129. Features "Case_No", "A1"-"A10" and "Qchat-10-Score" were removed from the dataset since they were not helpful for the classification. And all the values were converted into numeric values and were standardize. The training / testing data were also split here.
5) *Grid search for hyperparameter tuning*: Line 131-250. The best parameters for each modeling method were determined in this step and were used for all the following modeling steps.

6) *Ensemble and vote:* Line 255-285. DT, RF, LR and KNN were ensembled and both hard and soft voting were performed. Their accuracy scores were evaluated by cross-validation. And their performances were also tested using the testing data.
7) *Determining the most efficient model and performing the final evaluation*: Line 287-308. After comparing the accuracy scores of different modeling methods using the testing data, KNN had the highest accuracy. Confusion matrix and the classification report were made using the KNN method.

5. Modeling pipeline:

The following chart showed the pipeline of the modeling:



Results:

1. Accuracy scores before and after grid search:

After hyperparameter tuning using grid search, the accuracy scores for DT, RF, multinomial NB and KNN were all increased. The performance of LR and Gaussian NB did not change too much after the tuning. Also, Gaussian NB had a very low accuracy score. This might because Gaussian NB performs better in analyzing continuous values, but the values in dataset we analyzed are mainly discrete, so Gaussian NB is not a good model for this analyze. All the best parameters for the methods were listed in the last column of Table 1 and were used for the following modeling.

Table 1: Modeling--- Grid search

Method	Before grid search	After grid search	Best parameters
DT	63.03%	69.67%	criterion: entropy max_depth: 3 min_samples_leaf: 5
RF	69.19%	70.62%	max_depth: None, max_features: sqrt, min_samples_leaf: 4 n_estimators: 100
LR	67.30%	67.30%	C: 0.001 max_iter: 50 penalty: l1 solver: saga
Gaussian NB	34.12%	34.12%	var_smoothing: 1e-09
Multinomial NB	70.62%	71.09%	alpha: 10.0
KNN	69.67%	73.46%	n_neighbors: 19

2. Accuracy score comparison among all the modeling methods:

As shown in Table 2, among all the modeling methods, the KNN method had the highest accuracy score, which is 73.46%. So, the best model to predict whether a toddler is showing ASD traits is KNN.

Table 2: Modeling--- Ensemble and Vote

Method	Accuracy
Ensemble and vote (hard)	72.04%
Ensemble and vote (soft)	70.14%
DT	69.67%
RF	70.62%
LR	67.30%
Gaussian NB	34.12%
Multinomial NB	71.09%
KNN	**73.46%**

Only ensembled DT, RF, LR and KNN

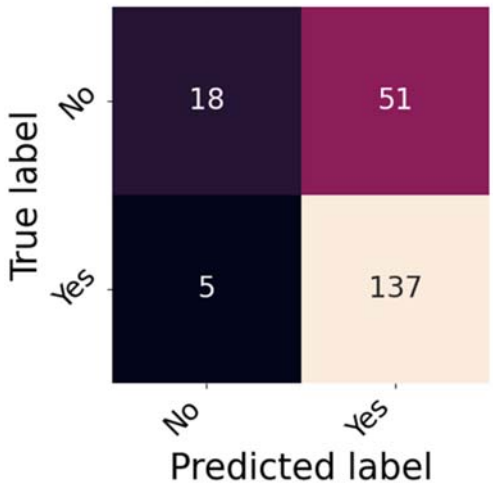
3. Evaluation of the KNN method:

The confusion matrix and classification report were generated after analyzing the testing data with KNN modeling method. The confusion matrix showed this model is good in predicting whether a toddler can be classified as ASD, with 137 correct predictions out of 142 ASD kids. But the method had a very high false positive rate, with 51 misclassified values out of 69 ASD

negative kids. As reported in the classification report, the f1-score in predicting the “Yes” class is 0.83, but is only 0.39 in predicting “No” class, further supporting the information we get from the confusion matrix. Anyway, in predicting a disease, high false positive rate is always better than high false negative rate. So the overall performance of the KNN model is satisficing.

Evaluation of the KNN method

Modeling the dataset using KNN:



	Precision	Recall	f1-score	Support
0 (No)	0.78	0.26	0.39	69
1 (Yes)	0.73	0.96	0.83	142
Accuracy			**0.73**	211
macro avg	0.76	0.61	0.61	211
weighted avg	0.75	0.73	0.69	211

Summary and conclusions:

Among all the modeling methods used in this project, the KNN model is proved to be the best for the classification. Although had a high false positive rate, the model showed a relatively good performance in classifying whether toddlers present ASD-like behaviors. A lot of things can be done to improve the modeling. For example, using sklearn.feature_selection function to do better feature selection might be helpful in increasing the predicting accuracy.

References:

1. https://www.nichd.nih.gov/health/topics/autism/conditioninfo/treatments/early-intervention
2. https://translate.google.com/?hl=zh-CN&sl=en&tl=zh-CN&text=analysis&op=translate
3. https://github.com/amir-jafari
4. https://scikit-learn.org/stable/
5. https://www.youtube.com/watch?v=28xRv-vC9Ys&list=PLreVlKwe2Z0TYh4aCLNw91q9FjRftMSc9&index=2
6. http://rasbt.github.io/mlxtend/user_guide/classifier/EnsembleVoteClassifier/

7. https://coderzcolumn.com/tutorials/machine-learning/scikit-learn-sklearn-naive-bayes
8. https://medium.com/fintechexplained/what-is-grid-search-c01fe886ef0a