Autism Spectrum Disorder Prediction Report

Phuong Thuy Dang
School of Computing Science
Simon Fraser University
8888 University Dr, Burnaby, BC
pdang@sfu.ca

Abstract

Autistic Spectrum Disorder (ASD) refers to a group of developmental disorders that affect the nervous system. Some of the most common ASD symptoms include impairment, challenges in social interaction, and repetitive behaviour that affect communication. ASD has a significant impact on health care not only due to the number of ASD cases raising but also the time involved to diagnose ASD. Moving in line with the rise in machine learning to speed up the time to detect a disease using existing data, the goal is to construct a model that accurately predicts whether an individual has ASD or not in order to provide early intervention for those who has a high chance of having ASD later. We use different models to compare the performance, including Logistic Regression, Support Vector Machines (SVM), Naive Bayes, k-Nearest Neighbours (KNN), Artificial Neural Network (ANN), Convolutional Neural Network (CNN) over 3 datasets: Adult, Children, and Adolescent. For the dataset, all datasets contains of 21 attributes; however, different instances. While adult dataset contains 704 instances, children dataset contains 292 instances, and adolescent dataset contains only 104 instances. After pre-processing the data, applying the model, and evaluating, results strongly suggest the high results for Logistic Regression with the accuracy 98%, 100%, 90%, and neural network with accuracy 100%, 96% and 70% for Adult, Children, Adolescent dataset respectively.

1 Introduction

Autism is a neurodevelopmental disorder with unknown causes, no effective preventive measures, and lifelong disease [1][2]. And because the diagnosis time is too long, the best intervention period is missed [3] and it is easy to be misdiagnosed [4]. A screening method implemented by machine learning technology was developed. The contents of this report are organized as follows: Section 1 presents the contribution for everyone in the group. Section 2 presents the introduction to the Autism Spectrum Disorder problem. Section 3 describes the datasets used in this study. And the different models used in this study are then described in Section 4. Experimental results under different models are presented and discussed in Section V. Finally, the conclusion of this study is drawn in Section VI.

2 Dataset

Three different data types are given, and the information is listed in Table 1.

Table 1: List of ASD datasets

Sr.No.	Dataset Name	Sources	Attribute Type	Number of Attributes	Number of Instances
1	ASD Screening Data for Adult	UCI Machine Learning Repository[5]	Categorical, continuous and binary	21	704
2	ASD Screening Data for Children	UCL Machine Learning Repository[6]	Categorical, continuous and binary	21	292
3	ASD Screening Data for Adolescent	UCL Machine Learning Repository[7]	Categorical continuous and binary	21	104

Each different type of data contains 20 questions, and the information is listed in Table 2.

Table 2: List of Attributes in the dataset

Attribute	Attributes Description
1	Patient age
2	Sex
3	Nationality
4	The patient suffered from Jaundice problem by birth
5	Any family member suffered from pervasive development disorders
6	Who is fulfilment the experiment
7	The country in which the user lives
8	Screening Application used by the user before or not?
9	Screening test type
10-19	Based on the screening method answers of 10 questions
20	Screening Score

3 Proposed Methodology

Figure 1 show the main step of this project.

3.1 Data Pre-processing

Pre-process the data is necessary to make the original data more meaningful and let the machine understand the meaning of the data. In the process of data pre-processing, some useless data are deleted, for example, when there is a "question mark", the sole data is deleted, because reading

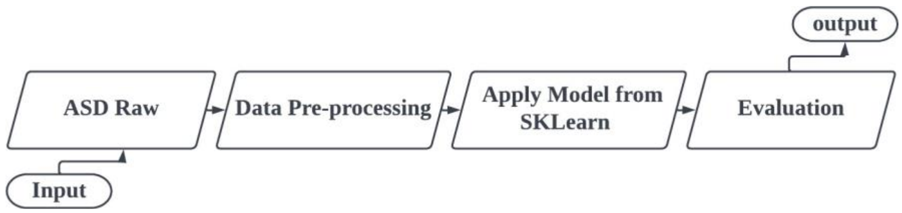

Figure 1: Steps in the proposed ASD detection solution

that data to train the machine does not seem to make any sense. Instead, it may make training more difficult. In addition to deleting useless data, compressed age and result data are also deleted, and the compressed data is less than 1, which makes the data more accurate during training. Then one-hot coding is used to split the country information. Take out all the countries separately, and use ’1’ or ’0’ to represent the country information.

3.2 Model

The data of each part are divided into two parts, the training part and the test part. 80% of the data is used for training and 20% for testing. Figure 2 shows this idea.

Figure 2: Training and testing sets

After split traning and test data part, using 6 different models: Logistic Regression (LR), Support Vector Machine (SVM), Naive Bayes, K-nearest Neighbors Algorithm (KNN), Artificial Neural Network (ANN), Convolutional Neural Network (CNN). First, the training data are used to train various models, and then use the testing data to generate the training results.

4 Performance metrics and result

In order to choose a model that avoid underfitting and overfitting, it is necessary analyze the learning curve. Also, confusion matrix with Specificity, Sensitivity and Accuracy score will be used to measure the effectiveness of each classification model.

4.1 Confusion Matrix

Consider the cases is binary classification, each individual is predicted as having ASD or not having ASD. Thus, every data point will be classified in one of the 4 categories below:

1. True positive (TP): The individual has ASD and are correctly predicted has ASD.
2. True negative (TN): The individual does not have ASD and are correctly predicted does not have ASD.
3. False positive (FP): The individual does not have ASD but are incorrectly predicted as has ASD.
4. False negative (FN): The individual has ASD but are incorrectly predicted as does not has ASD.

Sensitivity refers to the proportion of true positive that are correctly identified. Specificity refers to the proportion of true negatives correctly identified. Accuracy, in other hand, refers to the proportion of true results, either true positive or true negative in a population.

$$Specificity = \frac{TN}{TN + TP}$$

$$Sensitivity = \frac{TN}{TN + FN}$$

$$Accuracy = \frac{TP + TN}{TN + TP + FP + FN}$$

For instance, in the adult dataset, the sensitivity for SVM is 0.95 means when conduct classification task, there are 95% of chance this individual is classified as having ASD. If detecting individuals having ASD is a top priority, a model with high sensitivity can be used, since sensitivity tends to capture all positive outcomes. The Table 3 below are Specificity, Sensitivity, Accuracy score for 3 datasets: Adult, Children and Adolescent.

Table 3: Adult, Children and Adolescent datasets scores (Adult/ Children/ Adolescent)

Model	Specificity	Sensitivity	Accuracy
LR	0.97/ 1.0/ 0.98	1.00/ 1.00/ 1.00	0.98/ 1.00/ 0.90
SVM	0.96/ 0.95/ 0.85	0.95/ 0.96/ 1.00	0.96/ 0.96/ 0.95
Naive Bayes	0.90/ 0.77/ 0.29	0.88/ 0.93/ 1.00	0.89/ 0.86/ 0.75
KNN	0.96/ 0.77/ 0.43	0.93/ 1.00/ 1.00	0.95/ 0.90/ 0.80
ANN	0.98/ 0.95/ 0.42	1.00/ 0.96/ 0.84	0.98/ 0.96/ 0.70
CNN	1.00/ 0.95/ 0.43	1.00/ 0.96/ 0.85	1.00/ 0.96/ 0.70

For the implementation, in Naive Bayes algorithm, MultinomialNB has been used instead of GaussianNB due to the fact that dataset is more likely discrete datasets rather than normal distribution datasets. For SVM, RBF Kernel has been used with 0.1 gamma value. For KNN, the Figure 3 shows the result score after running a loop in range of 1 to 31 and the best k = 7 has been used. In ANN, Adam Optimizer with 0.01 learning rate, binary cross-entropy loss function, 0.2 dropouts, 100 epoch has been used. In CNN, Relu activation Function, Adam Optimizer, binary cross-entropy loss function, 8 filters and 0.5 dropouts with 150 epoch has been used.

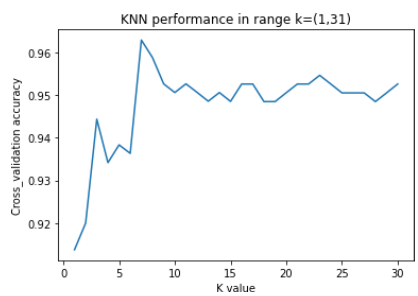

Figure 3: Find best k in range 1 to 31 for KNN

4.2 Learning Curve

Learning curves is a plot of model learning performance over experience or time and is a effective way to detect underfitting, overfitting problem. A learning curve in machine learning is a graph that compare the training score and cross validation score over a varying number of training instance.

Underfitting (or high bias) is the case when both training score and cross-validation score are low. Overfitting (or high variance) is the case when a model performance has large gap between training and validation score due to the model is training too well that affect the ability to deal with unseen data. A good fitting model is considered a model that has high score on both training data and validation score (unseen datas).

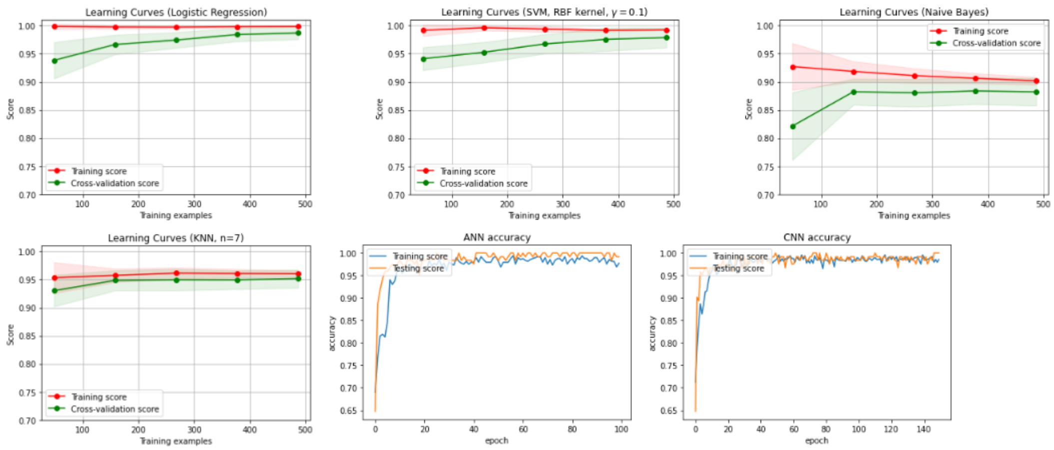

Figure 4: Learning curves for Adult's dataset

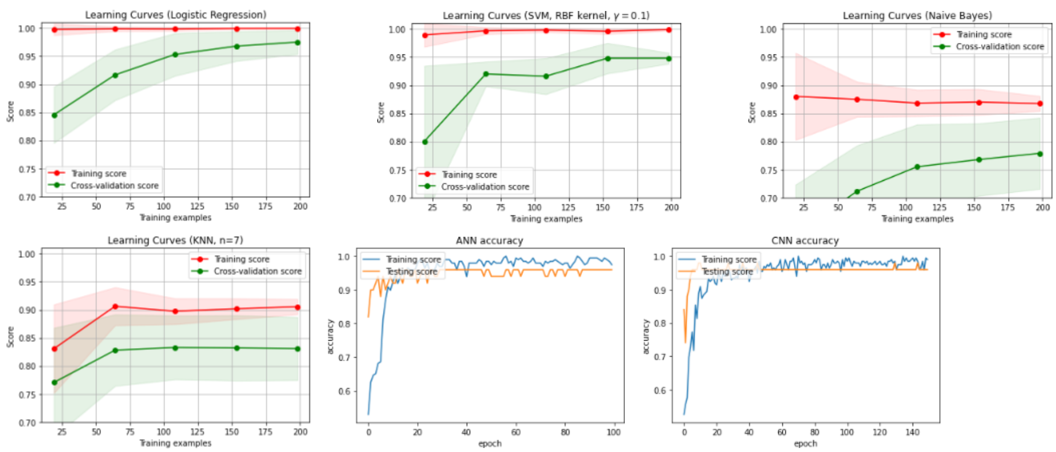

Figure 5: Learning curves for Children's dataset

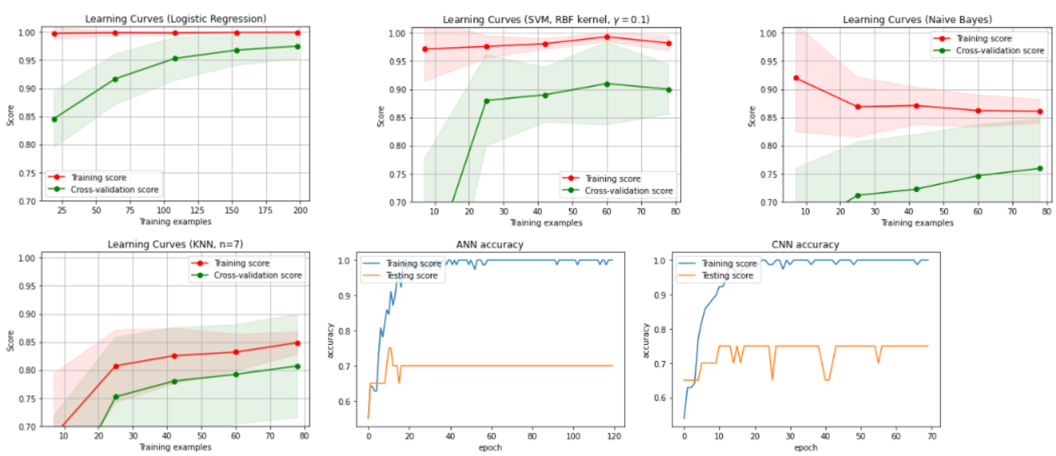

Figure 6: Learning curves for Adolescent's dataset

The learning curve of all models over 3 datasets are shown below.

Logistic Regression and SVM model show a good fitting plot since the training score is still around the maximum overtime and the validation could be increased with more training sample. In other hand, Naive Bayes model has training score and cross-validation score slightly converge at the end, indicate underfitting problem can occur. Based on the learning curve of ANN and CNN for adolescent dataset, overfitting is likely to occur.

The learning curve reflect the performance of all models over 3 datasets. While Logistic Regression and SVM performance results are remain high over 3 datasets, Naive Bayes, KNN and neural network model performance is much lower when dealing with small dataset as the number of instances in children and adolescent dataset is significant smaller than adult dataset (Table 1).

5 Conclusion

In this study, various machine learning techniques were attempted to detect autism spectrum disorders. And use different models and different evaluation indicators to analyze the data performance of three groups of different age groups. In general, based on the results of this study, logistic regression achieved better results overall, while neural networks presented better results when faced with larger data sets. But there is still more to be explored in this study, the most intuitive of which is that compared with the adult and child datasets, the adolescent group has too few data samples, which may affect the results.

References

[1] Tager-Flusberg H. (2010). The origins of social impairments in autism spectrum disorder: studies of infants at risk. Neural networks : the official journal of the International Neural Network Society, 23(8-9), 1072–1076. https://doi.org/10.1016/j.neunet.2010.07.008

[2] American Psychiatric Association (2013). "Autism Spectrum Disorder. 299.00 (F84.0)". Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5). Arlington, VA: American Psychiatric Publishing. pp. 50–59. doi:10.1176/appi.books.9780890425596.

[3] National Center on Birth Defects and Developmental Disabilities (13 March 2020). "Screening and Diagnosis: Autism Spectrum Disorder (ASD)". Centers for Disease Control and Prevention. US.

[4] Blumberg SJ, Zablotsky B, Avila RM, Colpe LJ, Pringle BA, Kogan MD (October 2016). "Diagnosis lost: Differences between children who had and who currently have an autism spectrum disorder diagnosis". Autism. 20 (7): 783–795. doi:10.1177/1362361315607724

[5] Fadi Fayez Thabtah (2017), "Austistic Spectrum Disorder Screening Data for Adult"., https://archive.ics.uci.edu/ml/machine-learning-databases/00426/

[6] Fadi Fayez Thabtah (2017), "Austistic Spectrum Disorder Screening Data for children,"https://archive.ics.uci.edu/ml/machine-learning-databases/00419/ ,2017

[7] Fadi Fayez Thabtah (2017), "Austistic Spectrum Disorder Screening Data for Adolescent",https://archive.ics.uci.edu/ml/machine-learning-databases/00420/.