Annals of Autism & Developmental Disorders

Review Article
Published: 09 May, 2020

Optimized Machine Learning Classification Approaches for Prediction of Autism Spectrum Disorder

*Devika Varshini G and Chinnaiyan R**

Department of Information Science and Engineering, CMR Institute of Technology, India

Abstract

Autism spectrum disorder is a serious developmental disorder that impairs the ability to communicate and interact. ASD screening is the process of detecting potential autistic traits in individuals using tests conducted by a medical professional, a caregiver, or a parent. These tests often contain large numbers of items to be covered by the user and they generate a score based on scoring functions designed by psychologists and behavioral scientists. In this paper, the effectiveness of various machine learning algorithms and pre- processing techniques for the task of classification for medical datasets that are used for predicting the early autism traits in toddlers and adults is evaluated. Several previous work in this direction use complex pre-processing and machine learning techniques for effective classification. However, this experiment establish that a simple pre-processing steps combined with appropriate encoding of data and different classifier algorithms like logistic regression, KNN and Random Forest give rise to comparable results with the state-of-the-art.

Keywords: Autism; Disorder; Classification; Logic regression; KNN; Random forest

Introduction

There is an increasing prominence of automating processes to curtail the cost and time of any industry. The most significant fields that would benefit from reducing processing time is health care. The speed and efficiency of human health issues diagnostics is vital. In Autism, the huge challenge faced in many health care conditions is the diagnosing time. It takes up to 6 months to firmly diagnose a child with autism due the long process, and a child must see many different specialists to diagnose autism, starting from developmental pediatricians, neurologists, psychiatrists or psychologists. In the current traditional way, the time consumed to finalize the Autism diagnoses is relatively long.

Therefore, Machine Learning methods can make relevant changes to accelerate the process. It is known that early intervention is the key for improving Autistic children. Clearly speeding the diagnosing time is even more crucial in Autism cases. Big data and machine learning technologies can make enormous progress to predict and fasten the complex and time-consuming processes of diagnosis and treatment. A machine learning system can be developed to utilize massive amount of health and medical data available towards predictive modeling and predictive analysis. In this paper, a comparison of several machine learning techniques and models will be tested and analyzed [1-5].

Data is pre-processed to make a prediction based on different categories into which test are classified as Autistic. There are many existing classification algorithms that can be applied. Every classifier is diverse in its way of data accumulation, data filtering, feature extraction and employing these processes towards feeding the model to learn. In this work, the effectiveness of several machine learning algorithms for evaluating the effectiveness of treatment or prediction of outcome of Autism Spectrum Disorder treatment is assessed. The layout of the paper is as follows the next section discusses the previous work done in this direction, while Section 3 details the approaches we have adopted. Experimental evaluation results are detailed in Section 4 while Section 5 concludes the work.

Literature Review

Assessment of Psychological disorders is done by observing the symptoms or features present in a human where the quantitative tests have less involvement during a diagnosis. Hence the clinical expertise is more significant for the differential diagnosis and grading of a disorder which is comparatively challenging than diagnosing a disease. There are classification techniques that have

Correspondence:
Chinnaiyan R, Department of Information Science and Engineering, CMR Institute of Technology, Bengaluru, India, E-mail: vijayachinns@gmail.com

Received Date: *06 Apr 2020*
Accepted Date: *07 May 2020*
Published Date: *09 May 2020*

Citation:
Devika Varshini G, Chinnaiyan R. Optimized Machine Learning Classification Approaches for Prediction of Autism Spectrum Disorder. Ann Autism Dev Disord. 2020; 1(1): 1001.

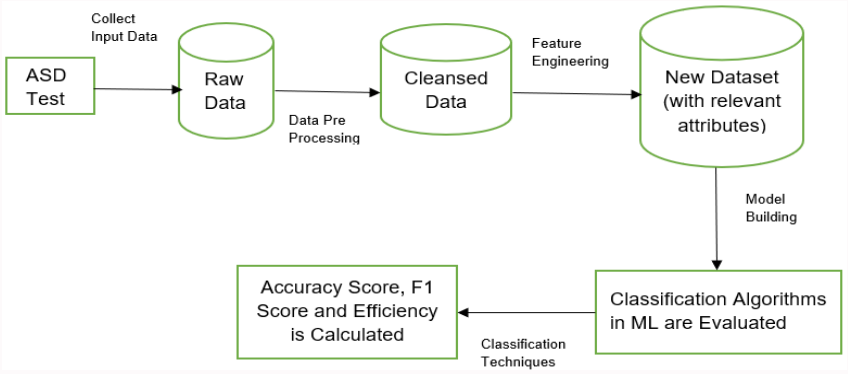

Figure 1: Classification using Machine Learning Techniques.

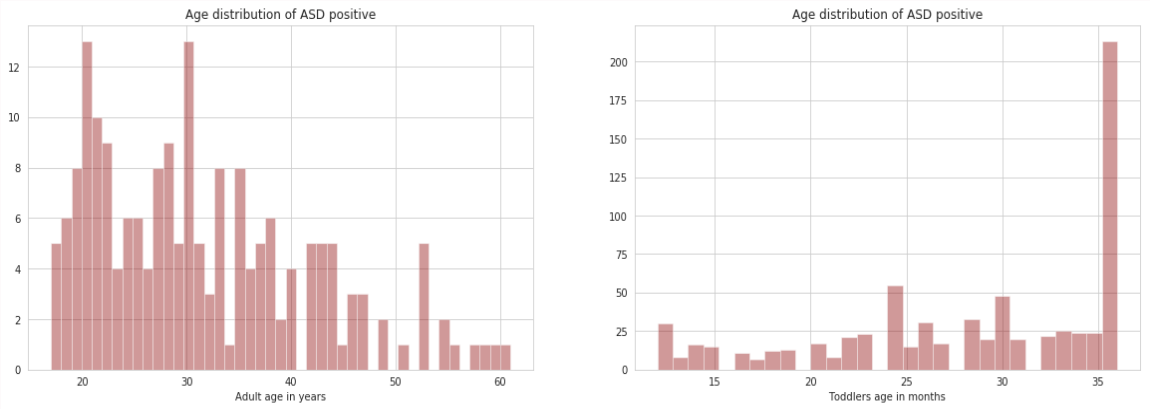

Figure 2: The age distribution of ASD positive.

been applied related to the work and they are discussed as follows.

1. Many types of ensemble learning models such as bagging and boosting. In the bagging learning model, the user must determine the size of the sample used to train the local classifier. This sample is often drawn from the training dataset, randomly based on the size specified by the end user. For example, if the user assigned the sample size to 50%, then new random instances with a size equal to 50% of the training dataset size will be created randomly. Moreover, the parameter of how many bags (classifiers) must be specified by the user. Normally, this parameter will be dynamic and continue to increase until no further improvement on the resulting classifiers is observed [6-8]. However, continuing to increase the number of bags can lead to higher overhead costs in terms of processing time and computing resources.

2. Pratap, Kanimozhiselvi [9], proposed a system for the application of naive Bayes dichotomizer supported with expected risk and discriminant functions in clinical decisions. This thesis investigated the performance of certain soft computing models and observed its applicability in assessment support systems, for a diagnostic confirmation to the clinicians. Frustrations due to misdiagnosis can be avoided to certain extend by the usage of clinical decision support systems, developed with the aid of soft computing techniques.

3. Crippa A, Salvatore C, Perego P, Forti S, Nobile M, Molteni M, Castiglioni I, proposed a system using machine learning to identify children with autism and their abnormal behavioral traits. They have undertaken a proof of concept study to establish whether a simple upper-limb movement could be useful to meticulously classify low-functioning children with Autism Spectrum Disorder (ASD) aged 2-4.

4. Accordingly, researchers in the field try to solve the multiple pass issue with new, enhanced methods and techniques, authors in (Pie, 2001) introduced an efficient approach for frequent rule mining in their "Classification Based on Multiple Class-Association Rules (CMAR)" algorithm for mining large datasets by constructing a class distributed-associated FP-tree. In addition, the authors adopted a CR-tree to preserve the structure of mined association rules and to enhance the storing and the retrieving processes, alongside adopting other rules pruning measures based on correlation rates, confidence as well as database coverage. This is in order to achieve higher accuracy from the classification model when predicting new class labels. CMAR produced better accuracy when compared to C4.5 and CBA models.

5. The existing clinical procedures require long waiting time for diagnosing autism. In an attempt to curtail this and help in early detection of autism amongst children and toddlers, researchers have been attempting to develop new initial screening tools using machine learning. While a few researchers have resorted to acoustic analysis of vocal production to detect autistic traits, others are conducting non-verbal or behavioral analysis using video recordings. A few other researchers follow the traditional approach of a clinical questionnaire that features selection techniques to develop a prediction model using the most effective features.

6. Boosting learning models such as AdaBoost were propagated

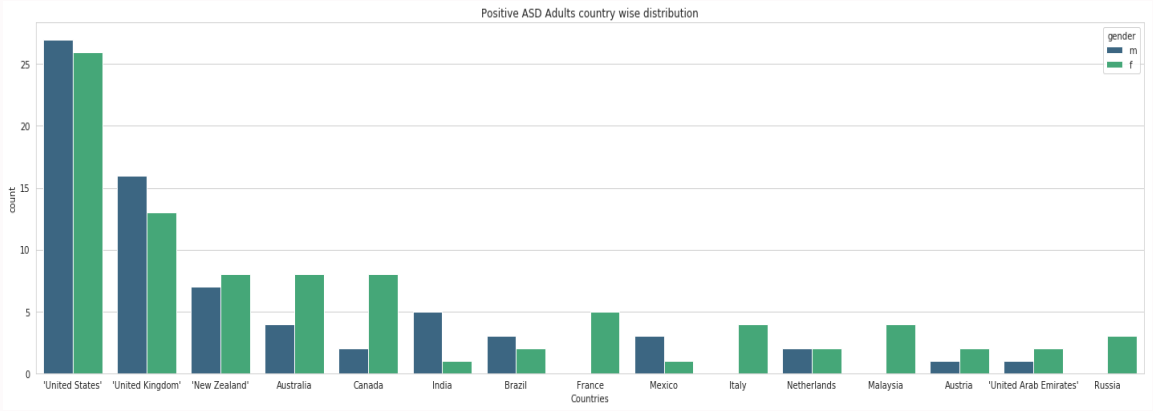

Figure 3: Positive ASD positive Adults based on top 15 countries.

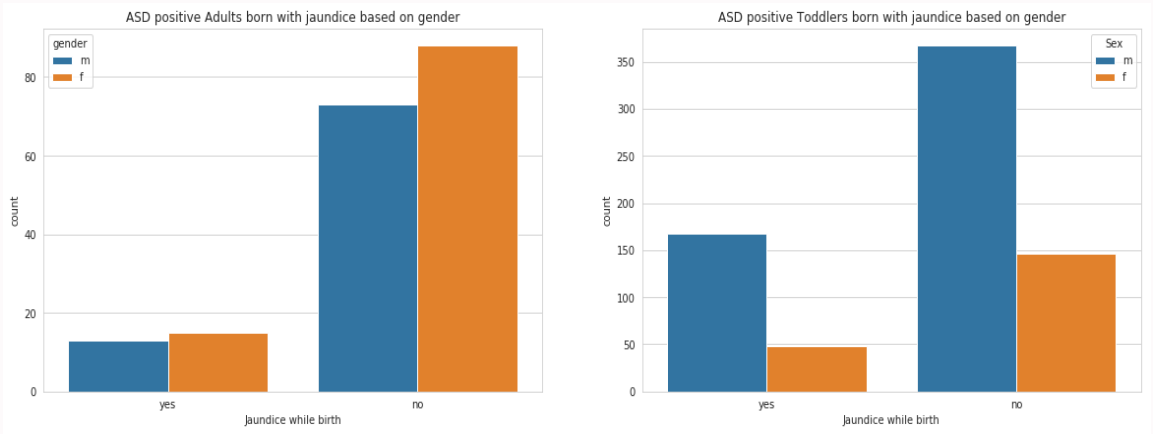

Figure 4: Jaundice born child based on gender.

to overcome the cost associated with the wrong predictions during the class assignment process (classification phase). They work by constructing an initial classifier using a base classification method such as decision trees. Then, the initial classifier will be applied to the training dataset and each wrongly classified training data will be assigned a new weight. Once the training instances have been assigned weights, then a new classifier is derived from the updated training dataset and the process is repeated until no further advancement can be attained in terms of predictive accuracy. Finally, whenever a test data is about to be predicted, then models learned with different data weights are used to decide its class collectively.

7. Bekerom used many ML techniques including naive Bayes, SVM and random forest algorithm to identify ASD traits in children like developmental delay, obesity, less physical activity and compared those results. Wall et al. worked on classifying autism with short screening test and validation and found that AD Tree and the functional tree had performed well with high sensitivity, specificity and accuracy.

Heinsfeld applied deep learning algorithm and neural network to identify ASD patients using large brain imaging dataset from the Autism Imaging Data Exchange (ABIDE I) and achieved a mean classification accuracy of 70% with an accuracy range of 66% to 71%. The SVM classifier achieved mean accuracy of 65%; while the random forest classifier achieved mean accuracy of 63%.

8. ASD screening does not demand a clinical setup and usually includes a set of behavioral questions, i.e. social, repetitive behavior, communication, etc., looking for any symptoms of autism in an individual. Often, the individual's medical professional, caregiver, parent, or teachers answer these questions on behalf of the child during the screening process then a final score is produced to diagnose if the child is potentially exhibiting autistic traits and if he requires any additional medical check-ups. Since the aim of the screening process is to assess any possibility of autistic symptoms based on questions (attributes) and a dependent variable (existence of autistic traits) then we can deal with this problem as a predictive supervised learning task [10-14].

Proposed Methodology

Autistic Spectrum Disorder (ASD) is a neurodevelopment condition correlated with important healthcare costs, which can be reduced by early diagnosis. Unfortunately, waiting time for an ASD diagnosis is tedious and procedures are not cost effective. The economic implications of autism and the increment in the number of ASD cases across the world reveals an urgent need for the development of easily implemented and effective screening methods. The accelerated increase in the number of ASD cases worldwide entails datasets related to the different behavior traits. However, such datasets are uncommon, hence it challenging to perform thorough analyses to enhance the efficiency, sensitivity, specificity and predictive accuracy of the ASD screening process. Presently, only finite autism datasets associated with clinical or screening are feasible

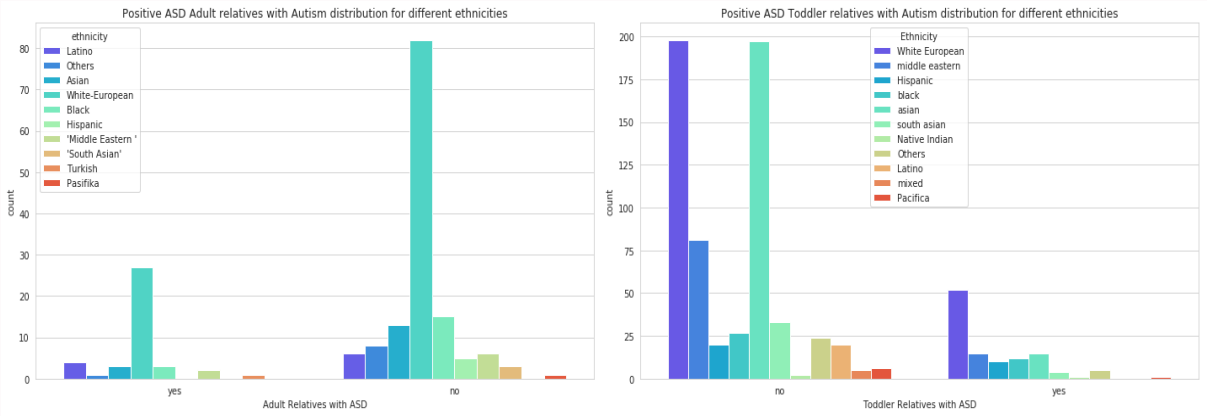

Figure 5: The distribution of autism in family within different ethnicity.

and most of them are genetic in nature.

Proposed algorithm formula

Input Dataset with Categorical, continuous and binary type attributes.

Two datasets - Adults and Toddlers.

A set of classifier models trained on this dataset for model building.

1. Data pre-processing is performed on the two datasets.
2. Graphs are plotted using different instances and attributes.
3. After feature engineering, from adults and toddler's datasets, a new dataset is created with only relevant attributes.
4. Using different classifiers, model building is performed using these relevant attributes.
5. F1 score, recall, precision is calculated and efficiency is estimated.
6. Using F1 score, graphs are plotted to find the most efficient classifier algorithm.

In this paper we have experimented with few of the popular machine learning techniques for performing classification across the datasets:

KNN: abbreviated as KNN, it is an approach to data classification that estimates how likely a data point is to be a member of one group or the other depending on what group the data points nearest to it are in. It is an example of a ”lazy learner” algorithm, meaning that until a query of the data set is performed, it does not build a model using the training set.

Random Forest: It is an ensemble machine learning approach that determines a large number of random decision trees analyzing a set of variables. Each of the decision trees predicts classes that can be aggregated in some manner to arrive at the final prediction.

Logistic regression: It is a technique borrowed from the field of statistics by machine learning. For binary classification problems (problems with two class values) it is a go-to method.

In this paper, the dataset is used is related to autism screening of adults which has 20 features that are utilized for further analysis specifically in determining dominant autistic traits and remodeling

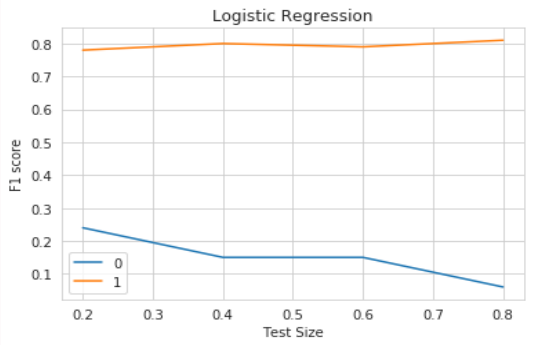

Figure 6: Logic Regression Classifier.

the classification of ASD cases.

In this dataset, ten behavioral features are listed (AQ-10-Adult) along with ten individual's characteristics that have proved to be potent in identifying the ASD cases from the controls in behavior science.

Implementation

Before using the dataset, some data pre-processing was performed. The data type is time-series or multivariate or sequential or univariate or text or domain-theory nominal/categorical, binary and continuous. The attribute type is continuous, binary and categorical. The missing values are checked using heat map for both adults and toddlers datasets. Using these attributes, graphs are plotted with pyplot to check what features or attributes are required for the prediction of ASD traits in both the datasets. Various visualization using graphs is plotted for jaundice born child based on gender, the age distribution of ASD positive, positive ASD Adults based on top 15 countries, the distribution of autism in family within different ethnicity. After the pre-processing method, the features that are irrelevant are neglected and the dataset is formed with the features required for the prediction.

The attributes are changed to age within 24-36, age within 0-12, male, ethnics (10 in total), jaundice born child, ASD genes (whether the family members have a history of ASD), ASD traits. Three different classifiers are used logistic regression, random forest and KNN. The F1 score and precision are evaluated to identify which classifier is more efficient for predicting ASD traits in the given data set.

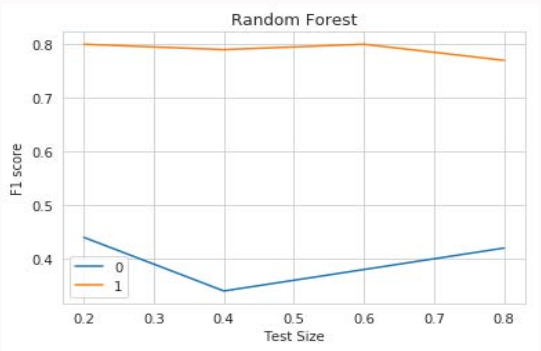

Figure 7: Random Forest Classifier.

Table 1: Autism Spectrum Dataset features.

Feature	Type	Description
Age	numerical	In years
Sex	univariate	Male/female
Ethnicity	multivariate	String list
Born with Jaundice	boolean	Yes/no
Family member with ASD	boolean	Yes/no
Who is completed the test	multivariate	String list
Country of Residence	multivariate	String list
Used the screening app before	boolean	Yes/no
Screening Method Type:		
A1_Score - Answer	integer	0.,1,2,3
A2_Score - Answer	integer	0.,1,2,3
A3_Score - Answer	integer	0.,1,2,3
A4_Score - Answer	integer	0.,1,2,3
A5_Score - Answer	integer	0.,1,2,3
A6_Score - Answer	integer	0.,1,2,3
A7_Score - Answer	integer	0.,1,2,3
A8_Score - Answer	integer	0.,1,2,3
A9_Score - Answer	integer	0.,1,2,3
A10_Score - Answer	integer	0.,1,2,3
Screening Score	integer	

For toddlers, most of them are 36 months. For adults, most of the ASD positive are 20 or 30 years of age. It is clearly noted that in adults as the age increases the number of positive cases decreases whereas in toddlers the number increases along with age.

Hence, for adults, with autism develop different approaches to help them age better. For toddlers, the significant signs of autism reveals around 3 years of age.

Even though the reach of the app affects this distribution, it does quite well describing the report. The most affected countries are Developed countries like Australia, US, Canada, UK. But it is noted that the female population is more distinguishable compared to males, which is quite contrary.

In adults, it is observed that it is almost 6-7 times more and in toddlers, 2-3 times more than of non-jaundice born ASD positive but in the reports it is around 10 times more. A child born with jaundice has a weak link with ASD.

Table 2: Different k values for KNN.

k-Value	precision	F1 Score	Accuracy
1	0.43	0.43	0.5829
3	0.48	0.44	0.6161
5	0.62	0.36	0.6682
7	0.62	0.38	0.6682
9	0.67	0.42	0.6872
11	0.68	0.45	0.6966
13	0.71	0.47	0.706
15	0.74	0.45	0.7061

Table 3: Comparison of Classification Algorithms for ASD Prediction.

Test Size	LR		RF		KNN	
	F1 Score	Precision	F1 Score	Precision	F1 Score	Precision
0.2	0.24	0.73	0.32	0.71	0.47	0.71
0.4	0.15	0.52	0.34	0.52	0.37	0.54
0.6	0.15	0.5	0.38	0.57	0.25	0.38
0.8	0.06	0.62	0.42	0.53	0.29	0.54

Also, according to reports, ASD is more frequent in boys (around 4-5 times) than in girls. But in Adults, a lower ratio is seen, whereas in toddlers it's nearly 4 times more among boys than among girls, which is quite close to actual ratio.

It is observed that in both adults and toddlers, White and European Ethnicities have a higher chance of being ASD positive if they have it in their genes. Black and Asians follow next though with smaller ratios. Nothing can be firmly concluded but it can be said with confidence that there is a genetic link for ASD positive which is backed up by studies.

Experimental Results

Extensive analysis was performed against experimental results for the purpose of assessing the F-measure, recall and precision as statistical measures for three of the well-known AC algorithms early mentioned. Also, varying values for minimum support as well as minimum confidence were used in order to evaluate the reliability of the selected algorithms on autism adult and toddler dataset. The evaluation process presents three well-known statistical measures (F1, Precision, and Recall) are used to reflect the overall performance for all of those algorithms on the Autism Adult UCI Dataset,

Logistic regression

Plotting different F1 score values for different test sizes for 0 and 1 label encoding for label 0, the F1 score decreases as test size increases, whereas, label 1, the F1 score is almost constant.

Random forest

Plotting different F1 score values for different test sizes for 0 and 1 label encoding.

For label 0, the F1 score decreases and gradually increases after a point as test size increases, whereas, label 1, the F1 score is almost constant till test size 0.6 and then gradually decreases.

KNN

Plotting different F1 score values for different test sizes for 0 and 1 label encoding.

For label 0, the F1 score decreases till test size = 0.6 and gradually

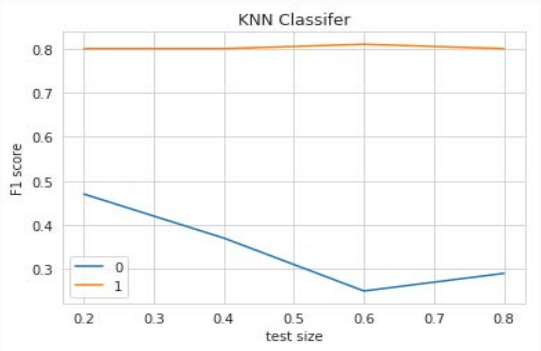

Figure 8: KNN Classifier.

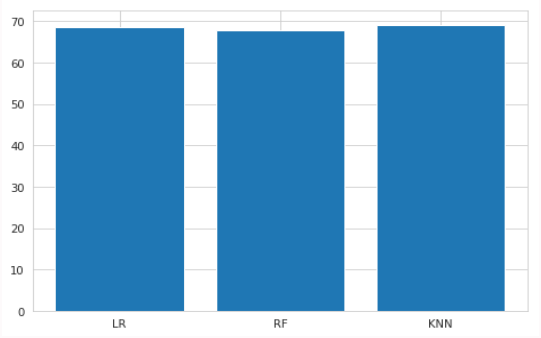

Figure 9: Accuracy score against the Different Classifiers.

increases as test size increases, whereas, label 1, the F1 score is almost constant.

Comparing Algorithms

Accuracy Score is calculated and the accuracy for KNN is 69.2% while accuracy for logistic regression and random forest classifiers are 68.601% and 67.78% respectively.

Out of above three models KNN classifier and random forest classifier performs same overall but much better than logistic regression.

Conclusion

Autism is considered as one of the fastest growing developmental disorder in children, hence the study for its early diagnosis with the support of classification models will certainly contribute to a greater extent, in solving the problem of making a correct assessment. This work focused on the development of some classification models using machine learning algorithms such as random forest algorithm, logistic regression and K nearest neighbor algorithm with two datasets adults and toddler. This study evaluates the performance of machine learning classification techniques which find the efficiency performance of the classification algorithms on these datasets. KNN has higher accuracy score of 69.2% compared to the other two algorithms which is calculated in the experimental results.

References

1. Kosmicki JA, Sochat V, Duda M, Wall DP. Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. Transl Psychiatry. 2015;5(2):e514.
2. Riley M, Karl J, Chris T. A study of early stopping, ensembling, and patch working for cascade correlation neural networks. IAENG Int J Applied Mathematics. 2010;40(4):307-16.
3. Allison C, Baron-Cohen S, Charman T, Wheelwright S, Richler J, Pasco G, et al. The Q-CHAT (quantitative checklist for autism in toddlers): A normally distributed quantitative measure of autistic traits at 18-24 months of age: preliminary report. J Autism Dev Disord. 2008;38(8):1414-25.
4. Duda M, Ma R, Haber N, Wall DP. Use of machine learning for behavioral distinction of autism and ADHD. Transl Psychiatry. 2016;6(2):e732.
5. Bone D, Goodwin MS, Black MP, Lee CC, Audhkhasi K, Narayanan S. Applying machine learning to facilitate autism diagnostics: Pitfalls and promises. J Autism Dev Disord. 2015;45(5):1121-36.
6. Thabtah F. An accessible and efficient autism screening method for behavioural data and predictive analyses. Health Informatics J. 2019;25(4):1739-55.
7. Thabtah F, Kamalov F, Rajab K. A new computational intelligence approach to detect autistic features for autism screening. Int J Med Inform. 2018;117:112-24.
8. Abbas H, Garberson F, Glover E, Wall DP. Machine learning approach for early detection of autism by combining questionnaire and home video screening. J Am Med Inform Assoc. 2018;25(8):1000-7.
9. Pratap A, Kanimozhiselvi CS. Application of Naive Bayes dichotomizer supported with expected risk and discriminant functions in clinical decisions. Case study. In Advanced Computing (ICoAC), 2012 Fourth International Conference on (pp. 1-4). IEEE.
10. van den Bekerom B. Using machine learning for detection of autism spectrum disorder. 2017.
11. Wall DP, Kosmicki J, Deluca TF, Harstad E, Fusaro VA, Use of machine learning to shorten observation-based screening and diagnosis of autism. Translational psychiatry. 2012;2:e100.
12. Ventola P, Kleinman J, Pandey J, Barton M, Allen S, Green J, et al. Agreement among four diagnostic instruments for autism spectrum disorders in toddlers. J Autism Dev Disord. 2006;36(7):839-47.
13. Vllasaliu L, Jensen K, Hoss S, Landenberger M, Menze M, Schutz M, et al. Diagnostic instruments for autism spectrum disorder (ASD). Cochrane Database Systematic Rev. 1-27.
14. Thabtah F. Autism Spectrum Disorder Tools: Machin Learning Adaptation and DSM-5 Fulfilment: An Investigative Study. Proceedings of the 2017 International Conference on Medical and Health Informatics (ICMHI 2017), 2017. p. 1-6. Taichung, Taiwan. ACM.