KAMMA, S.P., BANO, S., NIHARIKA, G.L., CHILUKURI, G.S. and GHANTA, D. 2022. Cost-effective and efficient detection of autism from screening test data using light gradient boosting machine. In Raj, J.S., Palanisamy, R., Perikos, I. and Shi, Y. (eds.) *Proceedings of the 4th International conference on intelligent sustainable systems (ICISS 2021)*, 26-27 February 2021, Tirunelveli, India. Lecture notes in networks and systems, 213. Singapore: Springer [online], pages 777-789. Available from: https://doi.org/10.1007/978-981-16-2422-3_61

# Cost-effective and efficient detection of autism from screening test data using light gradient boosting machine.

KAMMA, S.P., BANO, S., NIHARIKA, G.L., CHILUKURI, G.S. and GHANTA, D.

2022

*This is the accepted manuscript version of the above paper, which is distributed under the Springer Nature AM terms of use: https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms*

OpenAIR @RGU

This document was downloaded from https://openair.rgu.ac.uk

SEE TERMS OF USE IN BOX ABOVE

DISTRIBUTED UNDER LICENCE

# Cost-Effective and Efficient Detection of Autism from Screening Test Data using Light Gradient Boosting Machine

Sai Pavan Kamma[1], Shahana Bano[2], Gorsa Lakshmi Niharika[3], Guru Sai Chilukuri[4], Deepika Ghanta[5]

Department of Computer Science & Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Andhra Pradesh, India

[1] saipavank2000@gmail.com
[2] shahanabano@icloud.com
[3] niharikagorsa2000@gmail.com
[4] gurusai21102000@gmail.com
[5] deepikaghanta15@gmail.com

**Abstract:** Autism spectrum disorder (ASD) is a developmental disorder that effects the brain. Autism constrains a person's ability to interact and communicate with others. The cause of autism, in general, is unknown though genetics does play a role in the manifestation of the condition. In the absence of clear identifiable biomarkers, Shortcomings of the available prognostic approaches creates a need for a new technique that is speedy, cost-efficient, and provides an error-free diagnosis. The system should also be able to adapt to the varying characteristics of subjects with ASD. The amelioration Machine Learning brings to automated medical diagnosis has inspired us to come up with a solution. An adept screening and diagnostic test for patients exhibiting known autistic symptoms is a well-compiled, specific, and approved questionnaire, which facilitates an easy and cheap diagnosis. Autistic Spectrum Disorder Screening Test data is collected from one such questionnaire. We used a combination of three publicly available datasets containing records related to ASD in children, adolescents, and adults. There are a total of 1100 instances along with 21 attributes. The proposed study uses a Light Gradient Boost (LGB) based model for classification, along with Random Search for hyper-parameter optimization, which yielded a high accuracy of 95.82%.

**Keywords:** Autism, Autism Spectrum Disorder (ASD), Random Search, Light Gradient Boosting Machine (LGBM), pre-processing.

## 1. Introduction

The human brain is one of the largest and most complex organs in the body, contains roughly one hundred billion neurons. The brain is the centre for cognitive

activities. The brain provides us with the ability to react to stimuli, have emotions and stores incidents as memories. Autism spectrum disorder is a child-development disability which causes lack of social skills, speech difficulties, and inability to grasp non-verbal cues. Autism was initially elucidated by Leo Kanner [1] in 1943 as an innate inability to create normal, biologically determined, emotional contact with others Malfunctioning of mirror neurons and genetics are usually presented as reasons but, there is no one known cause. ASD can be hereditary [2], but the way the progeny gets the disorder is not known. People with allele alterations associated with Autism have a higher risk factor. A person suffering from Autism Spectrum Disorder may not have good people skills [1].
Fragile X Syndrome and Rett syndrome are associated as the genetic causes of autism. The general symptoms of autism may include the following [4]:

- avoiding eye contact or no awareness when others are speaking
- difficulty expressing needs
- no facial expressions or unusual facial expressions
- repeating words or phrases
- difficulty expressing needs
- hypersensitivity to sound, smell, taste, sight, or touch
- delayed speech and language skills, or lack of gesturing

ASD is found to be more common in boys as compared to girls [3]. Although, on a brighter aspect many children diagnosed with ASD live an independent, productive life. The symptoms of majority of the children assuage with age. Speech and psychological therapy can help with behavioural, depression issues. In some cases, people might require continual support throughout their life. Also, autistic individuals can work and live individually [2]. The word spectrum is used to illustrate a broad range of developmental delays and symptom severity. ASD includes people who have a few mild autistic traits to those who need help with day-to-day functioning. The differences between one type and another type can be subtle and difficult to determine. A diagnosis on the spectrum means you can turn attention to assessing individual needs [4].

The diagnostic test for autism is expensive and sometimes involves extensive equipment such as MRI. Various other techniques involve speech tests and observation schedules. An easier way we can quickly diagnose is through a screening tool that is built based on the eclectic phenotype and other behavioural symptoms of autism. The questionnaire consists of ten open-ended questions to which the test-taker (either parents or health care employees) can agree or disagree (example: "I often notice small sounds when others do not") based on the traits exhibited by the patient. The questionnaire also consists of general information like age,

gender, and ethnicity. Two other important questions are whether any immediate family members have ASD and If the patient was born with jaundice, which drastically increases the chances of autism.

## 2. Related Work

Thabtah [5] has developed a quick and available mobile application for screening of ASD called ASDTests. The application design aims to screen patients of all age groups. The users are given a questionnaire of ten questions and their answers are used for the diagnosis. This work uses data collected from existing applications and machine learning (ML) techniques like Naïve Bayes classifier and Logistic Regression classifier, with the later performing better in all parameters like accuracy, sensitivity, etc. across all ages. The data collected from the app has been made available publicly for academic purposes.

Suman Raja and Sarfaraz Masood [6] have worked on a study to classily ASD using three datasets with 20 common features acquired from UCI's ML Repository. Authors of the work have implemented ML and deep learning algorithms, compared accuracy of them across all the datasets. The study reported that the CNN based model produced the highest accuracy of 99.53%.

Fadi Thabtah [7] has proposed a screening model for autism making use of DSM-5 and Machine Learning Adaptation, the work uses datasets collected by the author. In the paper, the researcher highlights the advantages of using a DSM-5 which helps to overcome the problems with consistency of the existing tools.

M. S. Mythili, et al [8] studied autism using classification techniques. The objective of this paper was diagnosis of autism and the severity of it. They implemented Support Vector Machine (SVM), Neural Network and Fuzzy logic techniques using WEKA workbench tools to examine student's behaviour and social interaction.

The goal of our research study is to devise a model that can handle surging data from frequent surveys conducted by health departments, parents, and hospitals on subjects who could have ASD. The data also requires a considerable amount of pre-processing and may also contain outliers, implementation regression techniques might not always produce better results. The pre-processing also includes encoding the character features whereas our proposed model negates the requirement.

## 3. Methodology

### 3.1 Importing the Dataset:

The data utilized has been sourced from the Machine Learning Repository of University of California, Irvine (UC Irvine) . Autistic Spectrum Disorder Screening Data of adult subjects [9] which consists of 704 instances with 21 attributes. ASD Screening records of kids [10] has 292 instances and 21 attributes. ASD Screening records of teenagers [11] includes 104 instances with the same number of attributes as the above datasets. We used a combination of above datasets where the records are randomly shuffled to prevent possibility of overfitting, the cumulated dataset consists of 1104 records. All the datasets mentioned above have the following 19 attributes common among them, which we have used for our model.

Table 1: Overview of attributes used for the study

| Attribute Id | Attributes Description | Data Type |
|---|---|---|
| 1 | Patient age | int32 |
| 2 | Gender | String (m: Male, f: Female) |
| 3 | Ethnicity | String (White-European, Latino, Asian, Middle Eastern, etc.) |
| 4 | Whether the patient suffered from Jaundice at the time of birth | String (Yes, No) |
| 5 | If, any immediate family member suffered from | String (Yes, No) |

| | | |
|---|---|---|
| | autism | |
| 6 | Country of residence | String (Austria, Ireland, Jordan, etc..) |
| 7 | Test attempted Individual | String (Mother/Father, Hospital Personnel, Relative, etc.) |
| 8 | Answers to the screening test (10 questions) | String (Yes, No) |
| 9-18 | Screening Score | int32 |
| 19 | Class/ASD | String (Yes, No) |

### 3.2 Pre-processing the Data:

The process of modifying and cleaning the raw data into feasible format for training the model is called data pre-processing. Eclectic pre-processing techniques are used to handle missing and inconsistent data. The data contains multiple columns of categorical features like ethnicity which have encoded. However, the use of Light GBM eliminates the necessity because it automatically encodes the categorical attributes. The null and missing values were imputed to increase the feasibility of the record. Dataset Standardization has been done using Standard Scalar from Scikit-Learn library.

### 3.3 Training and Testing of the Proposed Model:

The dataset is split into a couple of parts in ratio of 80:20, one segment is for training the model and the another is for testing data to check the model performance. Further twenty percent of the training records are used as validation data to avoid overfitting.

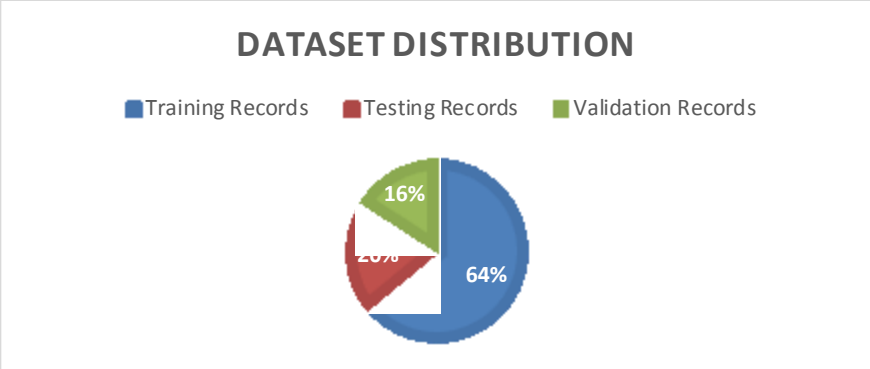


Fig.1: Dataset Distribution

### 3.4 Random Search Optimization:

Hyperparameter tuning is method which is used to find the most suitable parameters of the model architecture from the available space. In simple terms, is nothing but searching for the right hyperparameters to find high precision and accuracy. Random search is a technique where random combinations of the hyperparameters are used to find the right combination for the model built. In each search pattern, random combination of parameters is considered in every iteration. Then it checks if these are the right fit for the model and the data involved, if so, the search ends otherwise it continues to another random set of parameters. Random search is a group of statistical optimization procedures that lacks the requirement of the gradient of the problem to be optimized, and Random Search can hence be used on functions that are not continuous or differentiable.

The advantage of hyperparameter tuning is that by finding the ideal combination of hyperparameters for the model we can control the performance of the model thereby reducing the loss and improving accuracy. Random search unearths the best possible values of the parameters by efficiently looking at a larger, less outcome-oriented search area in comparison to other techniques. The aim of a ML algorithm *ml* is to identify a mathematical relation (function) $f(n)$ that limits loss $l(n; f(n))$ over the trials n from a ground truth gaussian distribution $t_n$ [12]. A ML algorithm *ml* is a relation that matches a data point $p^{(train\ data)}$(a finite array of experiments from $t_n$) to a function $f(n)$ [12]. Still, the ML algorithm itself have some features named hyper-parameters $\lambda_i$, and the final algorithm for the model is available strictly after choosing $\lambda_i$, which can be denoted $ml_\lambda$, and $f(n) = ml_\lambda(p^{(train\ data)})$ for

the training records $p^{(train\ data)}$ [12]. Random search has the benefit of finding the optimal parameters because of the randomness of the process.

PROCEDURE random search():

Find $x^*$ which optimizes the function f

Initialize $x, f_0$

Repeat

Generate z, a randomly approximated point over S;

if f(x+z)>max(f(x)) then x:=x+z;

$x^* \leftarrow x+z$;

RETURN($x^*$)

END random search;

### 3.5 Light Gradient Boosting Machine:

Light GBM is an ensemble learning technique which trains a Gradient Boosted Decision Tree (GBDT), by default. LGBM works for classification and regression tasks. Microsoft is credited with the first stable relaease of LGBM. LightGBM is a descent based boosting technique that employs tree-oriented learning techniques. It is developed to be distributed and efficacious as compared to other boosting algorithms. However, in GBDT, contains no innate set of weights, and thus the sampling methods recommended for other boosting algorithms cannot be automatically used, here the adavantage of gradient sampling is observed.

The framework is fast,light and designed for distributed training. It also provides the facility of handling missing values and categorical feature support. LGBM handles categorical labels by using the data of column. It does not transform to one-hot encoding rather LGBM uses a unique technique to find the division value of categorical labels, which is quicker. Light GBM makes use of leaf-wise tree growth technique, in comparsion with depth-wise growth, this helps in faster converging [14].

The following are the advantages [14]:

- Accelerated training and superior efficacy.
- Lesser memory required for operation.
- Better performance.
- Capable of handling large-scale data.
- Enables Graphics Processing Unit(GPU) training

GBDT uses decision trees to learn a function from the input space X s to the gradient space *G*[15]. Suppose that we have a training set with n instances {x1, · · , xn}, where each $x_i$ is a vector with dimension s in space $X^s$ . In each iteration of gradient boosting, the negative gradients of the loss function with respect to the output of the model are denoted as $\{g_1, \cdots, g_n\}$. The decision tree model splits each node at the most informative feature (with the largest information gain) [13]. For GBDT, the information gain is measured by the variance after splitting [13].

Hyperparameters which can be tuned for leaf-wise (LightGBM uses leaf-wise growth) Tree are :

- learning_rate
- num_leaves : The value should be less than $2^{max\_depth}$
- max_bin : Controls the limit of bins that features will be bucketed into.

Light GBM Algorithm:

For several boosting rounds M and a differentiable loss function L:

Let $F_0(x)=arg_{\Upsilon} \min \sum_{i-1}^{n} L(y_i,\Upsilon)$

For m = 1 to M:

1. Calculate the *pseudo* residuals

$$r_{im}=\frac{\left(-\partial L(yi,Fm-1(xi))\right)}{(\partial Fm-1(xi))}$$

2. Fit decision tree $h_m(x)$ to $r_{im}$
3. Compute the step multiplier $\Upsilon_m$ for each leaf of $h_m(x)$
4. Let $F_m(x)=F_{m-1}(x)+\lambda_m \Upsilon_m h_m(x)$, where $\lambda_m$ is the learning rate for iteration *m*.

### 3.6 Working Approach:

This is a classic binary class classification problem, the classes are (ASD and non-ASD). The dataset we pooled (ASD data of children, adolescents, and adults) contains 21 attributes that include test takers demographics and answers of the screening test which includes 10 questions (named A1 to A10) that test takers answered. In each of these 10 questions, the test takers were given a statement with which they had to agree or disagree. Dataset used for training consists of 1104 rows and 21 columns after pre-processing and removing less correlated features, the number of columns(features) was reduced to 19.

Pre-processing was done on this modified dataset, imputing missing values and scaling. Random search was used for hyper parameter tuning. The random search algorithm randomly searches the sample space (all possible parameter values of the Light GBM algorithm) in correspondence to the goal (find the best parameters for the data).Random search was implemented using the Sci-kit learn python library. The best estimator of the random search is returned by the random search algorithm; these are the most appropriate hyperparameters pertaining to this task.

Light GBM is a gradient boosting ensemble method which focusses on accuracy of results and supports GPU Learning. The Light GBM also executed using Sci-kit learn. A GPU enabled version of LGBM Classifier was used which can also multiple categorical features thereby eliminating the need for encoding the values. For training of the classifier 64% of the data(earlier mentioned split)was used. The testing was done using the test set and performance metrics accuracy, precision and f1-score were used to evaluate the model. The efficiency of training the model determines the implementation result.

## 4. Flowchart

The flow we represented here in flowchart is started with an algorithm of importing all the packages required for building the model. These include NumPy, Pandas , Matplotlib, SkLearn, Seaborn, etc. Then the data is imported using the pandas and pre-processing is done by imputing the null values. Afterwards the data is split into train set and test set using SkLearn in 80:20 ratio, validation set is formed from the train set. Then Random search is used for finding hyperparameters of Light GBM suitable for the problem. The best set is returned for the algorithm and the training is done using the train set. Evaluation measures are used to assess the performance of the model.

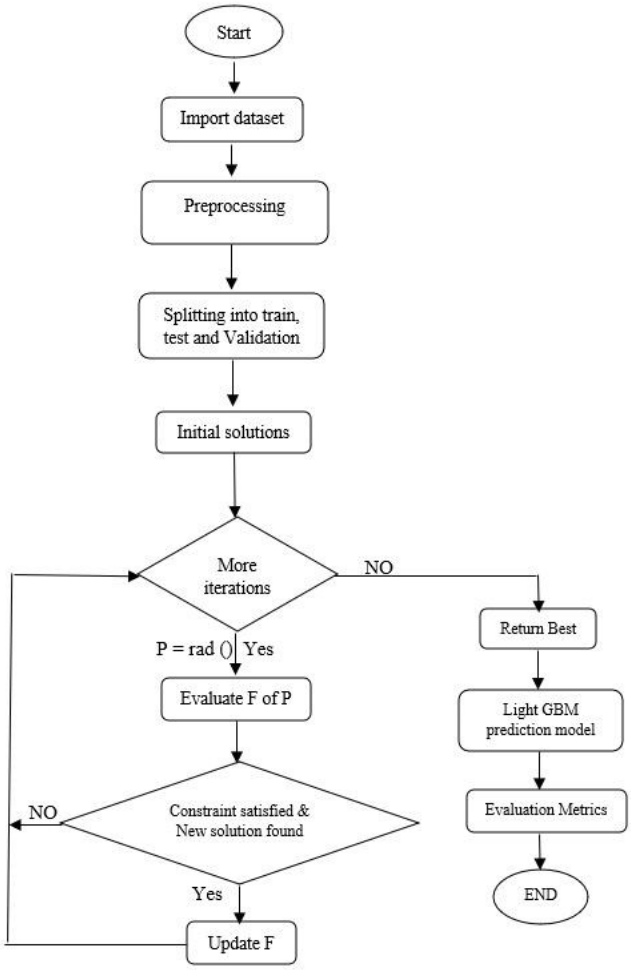


Fig 2: schematic of proposed work

## 5. Results

| | A1_Score | A2_Score | A3_Score | A4_Score | A5_Score | A6_Score | A7_Score | A8_Score | A9_Score | A10_Score | age | gender | ethnicity | jundice | austim | contry_of_res | result | relation | Class/ASD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 26 | 0 | White-European | 0 | 0 | United States | 6 | Self | 0 |
| **1** | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 24 | 1 | Latino | 0 | 1 | Brazil | 5 | Self | 0 |
| **2** | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 27 | 1 | Latino | 1 | 1 | Spain | 8 | Parent | 1 |
| **3** | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 35 | 0 | White-European | 0 | 1 | United States | 6 | Self | 0 |
| **4** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 40 | 0 | White-European | 0 | 0 | Egypt | 2 | Self | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **699** | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 25 | 0 | White-European | 0 | 0 | Russia | 7 | Self | 1 |
| **700** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 34 | 1 | Hispanic | 0 | 0 | Mexico | 3 | Parent | 0 |

Fig 3: Snapshot of a Dataset

The image shows a brief overview of the Autistic Spectrum Disorder Screening Data for Adult [9].

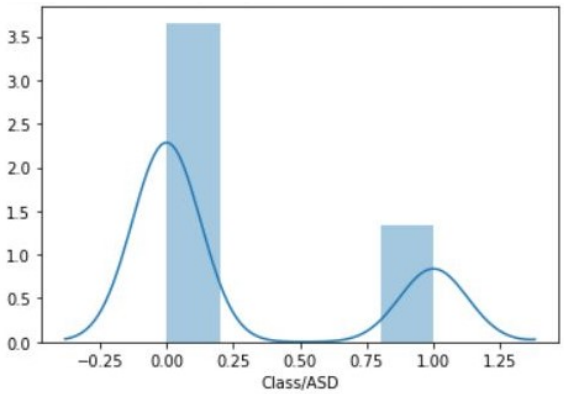


Fig 4: Distplot of Class/ASD

The image shows a displot(shows a histogram with a line on it) which is plotted using Seaborn library distplot plots a univariate distribution of observations.

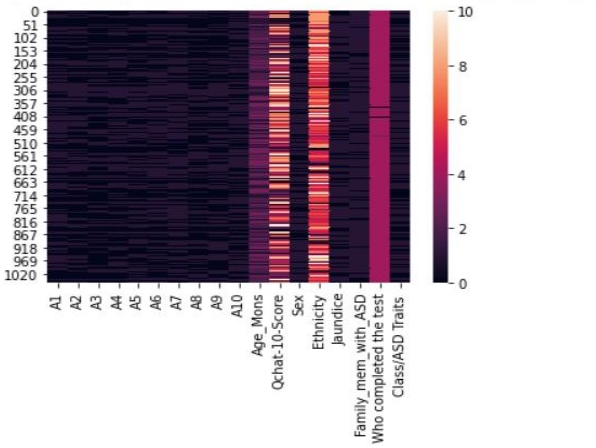


Fig 5: HeatMap visulaization of the dataset.

The above image is heatmap of the dataset. Heatmap is powerful way to visualize relationships between variables in high dimensional space.

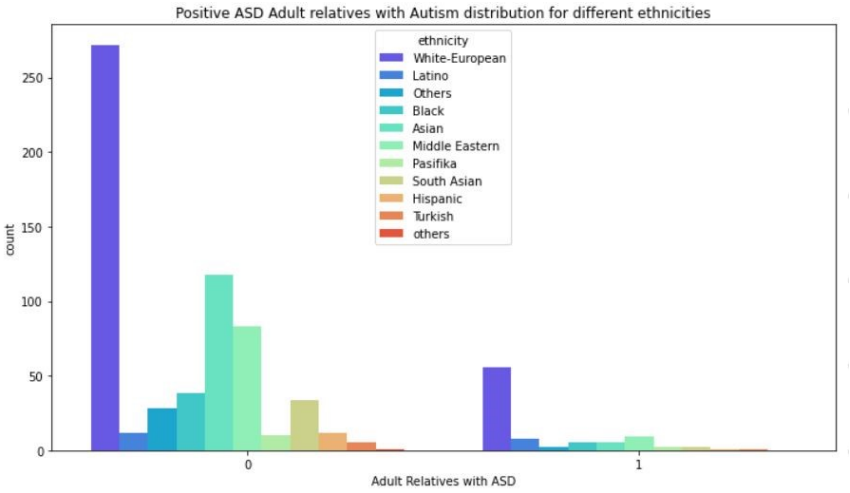


Fig 6: Positive ASD Adult relatives vs different ethnicities

Bar Graph visualization of Ethnicity of the subject and Immediate adult relatives of the subject with ASD.

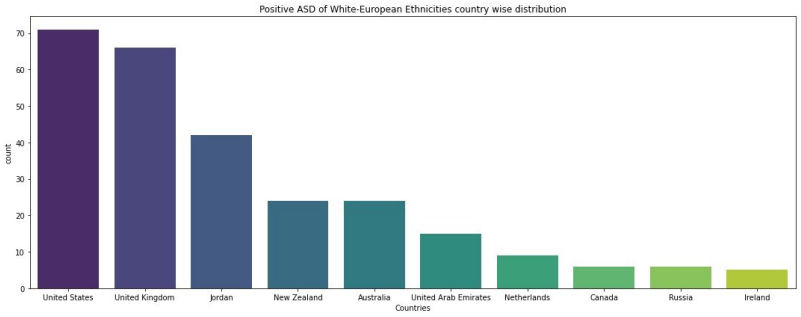


Fig 7: Positive ASD in subjects belonging to white enthinicty vs Country of residence

Bar Graph visualization of subjects of white ethnic group and their Native Country

| | precision | recall | f1-score |
|---|---|---|---|
| 0 | 0.94 | 1.00 | 0.97 |
| 1 | 1.00 | 0.81 | 0.90 |
| accuracy | | | 0.95 |
| macro avg | 0.97 | 0.91 | 0.93 |
| weighted avg | 0.96 | 0.95 | 0.95 |

Fig 8: Classification report of the model

The image show the classification report of the classification model it shows well-known metrics used to calculate the effectiveness of the classification : precision, recall and f1-score for each class. The metrics are computed by using true positives (TP), true negatives (TN), false positive (FP), false negatives (FN).

The result was assessed in metrics of accuracy, precision and recall by using classification report. Classification report automatically presents the precision, recall and f1-score of the experimentation with the help of confusion matrix. The formulae for the above mentioned metrics can be seen in the below given equations. The best fit hyperparameters(learning_rate=0.05,num_leaves=11, n_jobs=3) returned by random search were used for the LGBM Classifier which was trained using 300 estimators. The overall accuracy of the model is 95.82%. The average precision,recall and f1-score for both the classes (ASD and non-ASD) is 97%,91% and 94% respectively.

Table 2: Elements of a Confusion Matrix

| Original data values | Prognostic data values | |
|---|---|---|
| | TP | FP |
| | FN | TN |

$$\text{Accuracy} = \frac{TP+TN}{(TN+TP+FP+FN)} \qquad \text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN} \qquad \text{f-1 Score} = 2\times \frac{Precision*Recall}{Precision+Recall}$$

## 6. Conclusion

Around One in fifty-four children are diagnosed with autism according to the reports from the Centre for Disease Control's Autism and Developmental Disabilities Monitoring (ADDM) Network [3]. This work focuses on providing a scalable solution for efficient autism screening. In this study, we used machine learning techniques to classify ASD data from a screening test as to whether the subject is autistic or not (ASD or non-ASD). For this, we implemented a Light GBM model with the algorithm hyperparameters tunned using Random Search algorithm. The rationale for using LGBM is it handles huge datasets with relative ease thereby improving the accessibility and efficiency for real-world use. A comparison with some existing studies [6] shows that our model outperforms even while using three fewer features and most importantly required less pre-processing. This work could be further used to create an automatic screening tool using a mobile application for several other ASD screening tests with subjects from different age groups.

## 7. References

[1] Eisenberg, L. (1981). Leo Kanner, MD 1894--1981. The American Journal of Psychiatry, 138(8), 1122-1125.

[2] https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Autism-Spectrum-Disorder-Fact-Sheet

[3] Baio, Jon, et al. "Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2014." MMWR Surveillance Summaries 67.6 (2018): 1.

[4] https://www.healthline.com/health/types-of-autism#autism-symptoms

[5] Thabtah, Fadi. "An accessible and efficient autism screening method for behavioural data and predictive analyses." Health informatics journal 25.4 (2019): 1739-1755.

[6] Raj, Suman, and Sarfaraz Masood. "Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques." Procedia Computer Science 167 (2020): 994-1004.

[7] Thabtah, Fadi. "Autism spectrum disorder screening: machine learning adaptation and DSM-5 fulfillment." Proceedings of the 1st International Conference on Medical and health Informatics 2017. 2017.

[8] Mythili, M. S., and AR Mohamed Shanavas. "A study on Autism spectrum disorders using classification techniques." International Journal of Soft Computing and Engineering 4.5 (2014): 88-91.

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

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

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

[12] Bergstra, James, and Yoshua Bengio. "Random search for hyper-parameter optimization." The Journal of Machine Learning Research 13.1 (2012): 281-305.
[13] Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems. 2017.
[14] https://lightgbm.readthedocs.io/en/latest/index.html
[15] Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." Annals of statistics (2001): 1189-1232.

View publication stats