ORIGINAL ARTICLE e-ISSN: 2349-0659 p-ISSN; 2350-0964

Exploratory Data Analysis and Decision Tree Modeling for Autism Spectrum Disorder: Machine Learning Approach

R. S. Kamath[1], S. S. Jamsandekar[1], V. L. Badadare[1], R. K. Kamat[2]

Abstract

As the early diagnosis of autism spectrum disorder (ASD) is critical the high accuracy machine learning can be applied to achieve technology based diagnosis. In this context, the present study demonstrates machine learning approach for ASD diagnosis using decision tree (DT) modeling. The dataset employed in the present study comprises two classes of ASD adults with a sample size of 704 instances. The DT model entails a recursive partitioning approach implemented in the "rpart" package of R. The optimum model is derived by tuning parameters such as Min split, Min bucket, Max depth, and complexity. The performance of the model is evaluated in terms of the mean square error estimate of the error rate.

Keywords: Autism spectrum disorder, Decision tree, Exploratory data analysis, Machine learning, Recursive partitioning

Asian Pac. J. Health Sci., (2022); DOI: 10.21276/apjhs.2022.9.4S.31

Introduction

This paper portrays exploratory data analysis (EDA) and decision tree (DT) classification of adults with autism spectrum disorder (ASD). It is a neurological disorder that affects how a person acts and interacts with others.[1] It normally begins in childhood and continues. There are numerous clinical and non-clinical methods for the diagnosis of ASD. Many clinical experts and researchers have conducted studies about the classification of ASD data. Demirhan reported performance evaluation of different machine learning techniques in determining ASD cases.[2] Diabat and Shanableh demonstrated an ensemble learning model for testing autism in children.[3] Yet, another analysis by Praveena *et al*. predicted autism using machine learning techniques.[4] The present study uses ASD screening data of adults to design the machine learning framework for classifying autism data.[5]

EDA

The ASD dataset for the present study retrieved from the UCI machine learning repository is related to autistic spectrum disorder screening data for adults.[6] This dataset consists of 704 instances of ASD adults classified into two classes such as ASD "Yes" and ASD "No." This classification is based on 20 attributes including ten behavioral features and ten individuals characteristics. A basic statistical summary of numerical and categorical attributes is explained in Figure 1. The dataset reveals that the ASD value is "No" for 359 adults and the ASD value is "Yes" for 133 adults. There are 238 female and 254 male respondents with an age range from 17 to 61 years. Figure 2 plots a gender-wise bar chart for both classes

```
      Age            Gender                 Ethnicity    jundice    austim
 Min.   :17.00   f:238   White-European   :160   no :445   no :425
 1st Qu.:21.00   m:254   Asian            : 84   yes: 47   yes: 67
 Median :27.00           'Middle Eastern ': 65
 Mean   :29.05           Black            : 30
 3rd Qu.:35.00           'South Asian'    : 24
 Max.   :64.00           (Other)          : 67
 NA's   :1               NA's             : 62
                contry_of_res used_app_before     result                          relation
 'United States'        : 80   no :485         Min.   : 0.000   'Health care professional':  3
 'New Zealand'          : 59   yes:  7         1st Qu.: 3.000   Others                    :  4
 'United Arab Emirates': 59                    Median : 4.000   Parent                    : 35
 India                  : 55                   Mean   : 4.884   Relative                  : 20
 'United Kingdom'       : 53                   3rd Qu.: 7.000   Self                      :368
 Jordan                 : 28                   Max.   :10.000   NA's                      : 62
 (Other)                :158
 Class.ASD
 NO :359
 YES:133
```

Figure 1: Basic statistical summary of numerical and categorical attributes

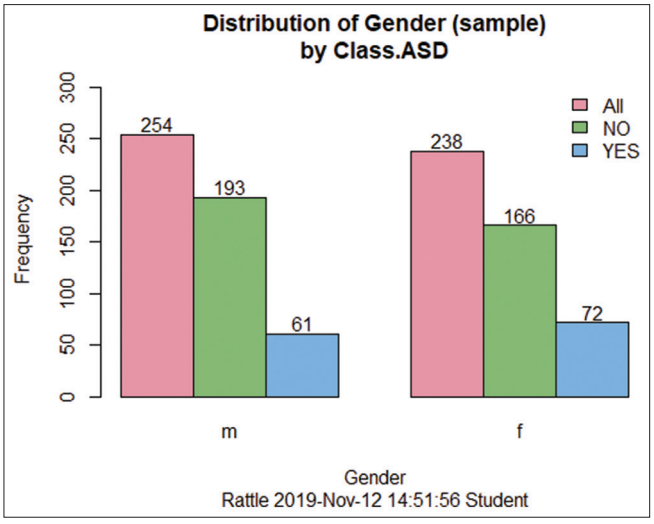

Figure 2: Gender-wise distribution of autism spectrum disorder data

[1]Department of Computer Studies, Chhatrapati Shahu Institute of Business Education and Research, Kolhapur Maharashtra, India

[2]Department of Electronics, Shivaji University, Kolhapur Maharashtra, India

Corresponding Author: V. L. Badadare, Department of Computer Studies, Chhatrapati Shahu Institute of Business Education and Research, Kolhapur 416004. E-mail: vlbadadare@siberindia.edu.in

How to cite this article: Kamath RS, Jamsandekar SS, Badadare VL, Kamat RK. Exploratory Data Analysis and Decision Tree Modeling for Autism Spectrum Disorder: Machine Learning Approach. Asian Pac. J. Health Sci., 2022;9(4S):161-164.

Source of support: Nil.

Conflicts of interest: None.ne

Received: 02/04/2022 **Revised:** 23/05/2022 **Accepted:** 10/06/2022

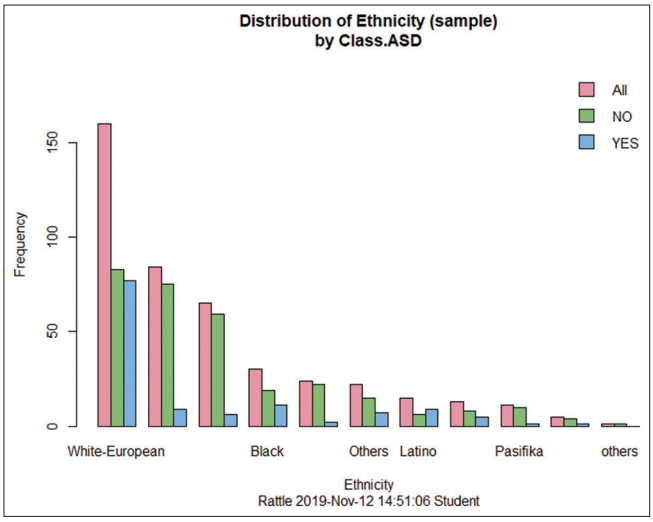

Figure 3: Proportions of each ethnicity

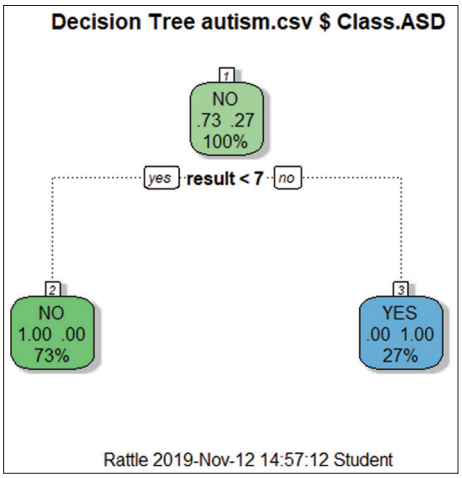

Figure 4: Actual decision tree for the data

of ASD. The bar plot shown in Figure 3 represents the proportions of each ethnicity present in the dataset. The plot reveals that White Europeans account for approximately one-third of the data, followed by Asians and Middle Eastern people.

DT Modeling

DT modeling is a supervised machine learning technique to classify and hence predict to what group a certain data point belongs to be. The DT is designed by splitting the dataset into subsets based on differentiators in input variables. It techniques include the Gini index, Chi-square, information gain, or reduction in variance to identify the most significant variable to get the homogeneous sets of subsets. DT is a visual representation utilized for deciding selection criteria. The DT model entails a recursive partitioning approach by dividing the data into leaf nodes and, hence, leaf nodes represent attributes.

The actual DT is shown in Figure 4. The DT is interpreted as following rules.

Rule number: 2 [Class. ASD = NO cover = 359 (73%) prob = 0.00]
Result <7

Rule number: 3 [Class. ASD = YES cover = 133 (27%) prob = 1.00]
Result ≥7.The dataset includes 20 response variables. It has been found that the following variables do not play a major role in classification modeling

- Age description, since all the observations are adults aged 17 and older
- Used app before, is not a significant feature for our target
- Country of residence, as a factor with more than 60 levels does not play significance importance in modeling
- Result, as explained in Figure 4 always classified as "Yes" for value 7 or more.

Hence, the decision modeling is done with following variables: Below are the remaining variables: Q1, Q2, Q3, Q4, Q5, Q6, Q7, Q8, Q9, Q10, age, gender, ethnicity, jaundice, autism, relation, and class_ASD.The DT model is simulated in an R environment. The model is conceived as a 16-input and single-output arrangement.[7] Figures 5 and 6 show DT derived in the present

```
Summary of the Decision Tree model for Classification (built using 'rpart'):

n= 492

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 492 133 NO (0.729674797 0.270325203)
   2) Q9< 0.5 328  24 NO (0.926829268 0.073170732)
     4) Q5< 0.5 202   1 NO (0.995049505 0.004950495) *
     5) Q5>=0.5 126  23 NO (0.817460317 0.182539683)
      10) Q6< 0.5 92   9 NO (0.902173913 0.097826087) *
      11) Q6>=0.5 34  14 NO (0.588235294 0.411764706)
        22) Ethnicity='South Asian',Asian,Black,Pasifika 11   1 NO (0.909090909 0.090909091) *
        23) Ethnicity='Middle Eastern ',Hispanic,Latino,Others,White-European 22   9 YES
(0.409090909 0.590909091)
          46) Q3< 0.5 10   4 NO (0.600000000 0.400000000) *
          47) Q3>=0.5 12   3 YES (0.250000000 0.750000000) *
   3) Q9>=0.5 164  55 YES (0.335365854 0.664634146)
     6) Q5< 0.5 40   7 NO (0.825000000 0.175000000) *
     7) Q5>=0.5 124  22 YES (0.177419355 0.822580645)
      14) Ethnicity='Middle Eastern ','South Asian',Asian,others,Others 26  13 NO (0.500000000
0.500000000)
        28) Q4< 0.5 9   1 NO (0.888888889 0.111111111) *
        29) Q4>=0.5 17   5 YES (0.294117647 0.705882353) *
      15) Ethnicity=Black,Hispanic,Latino,Pasifika,Turkish,White-European 91   6 YES (0.065934066
0.934065934) *

Classification tree:
rpart(formula = Class.ASD ~ ., data = crs$dataset[crs$train,
    c(crs$input, crs$target)], method = "class", model = TRUE,
    parms = list(split = "information"), control = rpart.control(usesurrogate = 0,
        maxsurrogate = 0))
```

Figure 5: Textual representation of decision tree for autism spectrum disorder data

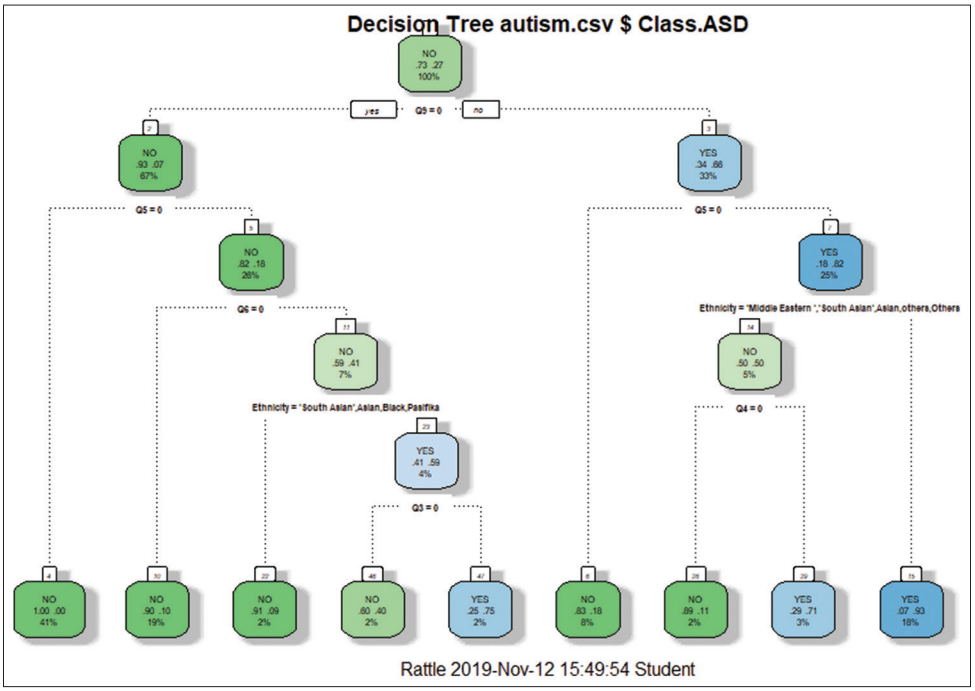

Figure 6: Decision tree for autism spectrum disorder data

```
Tree as rules:

 Rule number: 15 [Class.ASD=YES cover=91 (18%) prob=0.93]
   Q9>=0.5
   Q5>=0.5
   Ethnicity=Black,Hispanic,Latino,Pasifika,Turkish,White-European

 Rule number: 47 [Class.ASD=YES cover=12 (2%) prob=0.75]
   Q9< 0.5
   Q5>=0.5
   Q6>=0.5
   Ethnicity='Middle Eastern ',Hispanic,Latino,Others,White-European
   Q3>=0.5

 Rule number: 29 [Class.ASD=YES cover=17 (3%) prob=0.71]
   Q9>=0.5
   Q5>=0.5
   Ethnicity='Middle Eastern ','South Asian',Asian,others,Others
   Q4>=0.5
```

Figure 7: Sample rules

```
Root node error: 133/492 = 0.27033

n= 492

        CP nsplit rel error  xerror     xstd
1 0.406015      0   1.00000 1.00000 0.074069
2 0.195489      1   0.59398 0.59398 0.061229
3 0.037594      2   0.39850 0.43609 0.053780
4 0.010025      4   0.32331 0.45865 0.054963
5 0.010000      8   0.27820 0.42857 0.053376
```

Figure 8: Summary of decision tree model

```
Error matrix for the Decision Tree model on autism.csv [validate] (counts):

      Predicted
Actual NO YES Error
   NO  74   7   8.6
   YES  3  21  12.5

Error matrix for the Decision Tree model on autism.csv [validate] (proportions):

      Predicted
Actual   NO  YES Error
   NO  70.5  6.7   8.6
   YES  2.9 20.0  12.5

Overall error: 9.5%, Averaged class error: 10.55%
```

Figure 9: Confusion matrix for the validation set

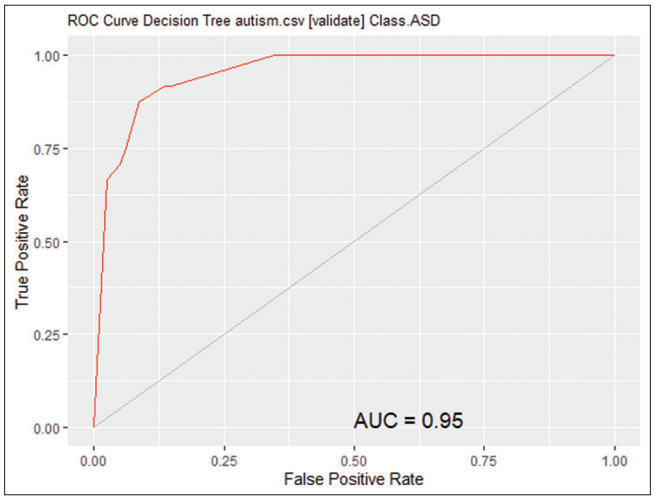

Figure 10: Receiver operating characteristic plot for decision tree model on validation data

investigation represents ASD in adults. That the root node of the DT tests *Q9* value = *0* continues down to the left side of the tree, otherwise right side of the tree. The next test down this left side of the tree is the *Q5* value. Sample rules are explained in Figure 7. Thus, it proceeds and will be able to retrieve the class value for ASD adults. The performance of the model is measured in terms of the mean square error (MSE) between predicted output and actual output. This is found to be 0.27033.

Results and Discussion

Figure 8 summarizes the DT model thus obtained, for classification. This textual view highlights the key interface widgets of DT construction. The variable "CP" stands for complexity parameter, which reveals that as the tree splits into more nodes, the CP is reduced.[8] Thus, derived DT model efficiently classifies validation data with very less errors. Figure 9 shows the error matrix for the DT model on validation data. Receiver operating characteristic (ROC) plots given in Figure 10 reveal the true positive rate against the false-positive rate. Since the area under ROC is 0.95 which is near 1, the model evaluates to be a better one.

Conclusion

This research illustrates the EDA and DT classification of adults with an ASD. The dataset employed in the present study comprises two classes of ASD adults with a sample size of 704 instances. The DT model entails a recursive partitioning approach implemented in the "rpart" package of R. The optimum model is derived by tuning parameters such as Min split, Min bucket, Max depth, and complexity. The performance of the model is evaluated in terms of the MSE estimate of the error rate.

References

1. Rylaarsdam L, Guemez-Gamboa A. Genetic causes and modifiers of autism spectrum disorder. Front Cell Neurosci 2019;13:385.
2. Demirhan A. Performance of machine learning methods in determining the autism spectrum disorder cases. Mugla J Sci Technol 2018;4:79-84.
3. Praveena TL, Muthu Lakshmi NV. Prediction of autism spectrum disorder using supervised machine learning algorithms. Asian J Comput Sci Technol 2019;8:142-5.
4. Al Diabat M, Al-Shanableh N. Ensemble learning model for screening autism in children. Int J Comput Sci Inform Technol 2019;11:45-62.
5. Basu K. Machine Learning Approaches to the Classification Problem for Autism Spectrum Disorder. Available from: https://www.github.com/kbasu2016/Autism-Detection-in-Adults/blob/master/report.pdf [Last accessed on 2019 Nov 15].
6. Autism Screening Adult Data Set. Available from: https://www.archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult. [Last accessed on 2019 Nov 15].
7. Kamath RS, Kamat RK. Modelling fetal morphologic patterns through cardiotocography data: Decision tree based approach. J Pharm Res 2017;12:9-12.
8. Kamath RS, Kamat RK. Modelling of random textured tandem silicon solar cells characteristics: Decision tree approach. J Nano Electron Phys 2016;8:04021.