A Deep learning of Autism Spectrum Disorder using Naïve Bayes, IBk and J48 classifiers

S. Gomathi

Abstract**: *Deciding the right classification algorithm to classify and predict the disease is more important in the health care field. The eminence of prediction depends on the accuracy of the dataset and the machine learning method used to classify the dataset. Predicting autism behaviors through laboratory or image tests is very time consuming and expensive. With the advancement of machine learning (ML), autism can be predicted in the early stage. The main objective of the paper is to analyze the three classifiers Naïve Bayes, J48 and IBk (k-NN). An Autism Spectrum Disorder (ASD) diagnosis dataset with 21 attributes is obtained from the UCI machine learning repository. The attributes have experimented with the three classifiers using WEKA tool. 10-fold cross validation is used in all three classifiers. In the analysis, J48 shows the best accuracy compared with the other two classifiers. The architecture diagram is shown to depict the flow of the analysis. The Confusion matrix with other relevant results and figures are shown.***

Index Terms**: *Autism, Machine learning, Weka, J48, IBk, k-NN, classifier, Naïve Bayes.***

I. INTRODUCTION

Autism is a childhood, neurodevelopmental disorder which affects a person's interaction, communication and learning skills which has become more predominant among younger generations in the recent decade. Clinical examination method conducted conferring to the DSM-V (Diagnostic and Statistical Manual of Mental Disorders) standards for disorder classification [1]. The DSM standards are devised by the US Mental health professionals based on the positive diagnostic knowledges and contributions. These measures are generally implemented in behavioral analytics for classification of ASD from non-ASD. Questionnaire-based and interview oriented clinical examinations are also followed for behavior classification addition to DSM-V standards. ADI-R (Autism Diagnostic Interview-Revised) and ADOS (Autism Diagnostic Observation Schedule) are some common behavior tests. The diagnosis of autism can be done at any age, its symptoms generally appear in the first two years of life and develop through time [2]. These clinical experiments are practiced by certified professionals in laboratory conditions. Autism patients face different types of challenges such as difficulties with concentration, learning disabilities, mental health problems such as anxiety, depression, etc, motor difficulties, sensory problems, and many others. Consolidated scores decide the severity [3] of autism in the patients. According to the Centre for disease control and prevention, there is a sustainable growth in the number of children diagnosed with Autism disorder and 1 among 68 Children under the age of 8 in the United States of America is diagnosed with autism and According to WHO [5], about 1 out of every 160 children has ASD. Data mining plays an important role to classify and predict the disease in the early stage [9].

II. PROBLEM SPECIFICATION

Practitioners used to make decisions based on their experience and physical analysis which may lead to wrong predictions and medications. Meanwhile, any machine learning algorithms cannot be used at a random but a proper analysis is necessary to choose the best algorithm for the prediction of the disease.

III. OBJECTIVE OF THE PAPER

The objective of this work is to propose an analysis of three algorithms for autism prediction using ML techniques and to find the best classifier that could effectively predict autism traits of any age.

IV. ARCHITECTURE DIAGRAM

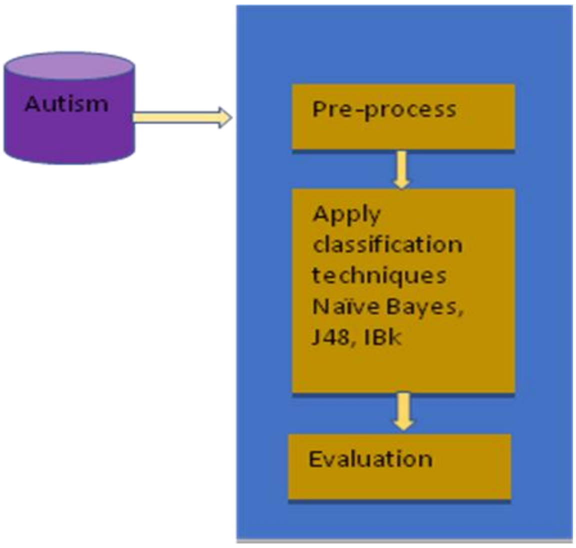


Fig 1. Architecture diagram for the proposed work

Revised Manuscript Received on 30 July 2019.
* Correspondence Author

S. Gomathi*, Assistant Professor, Department of Computer Science (PG), PSGR Krishnammal College for Women, Coimbatore, India.

The Fig 1 shows the architecture diagram of the proposed work. The raw data set is obtained from the UCI repository [10] and other sources. The data set has many missing values, random values which has to be preprocessed before analyzing the result.

The missing values may lead to the wrong analysis and can misclassify or wrongly predict the outcome. The preprocessing is done with the Weka tool [4] itself. Then the three algorithms are evaluated to see the performance and other major metrics to check the best algorithm to predict the disease.

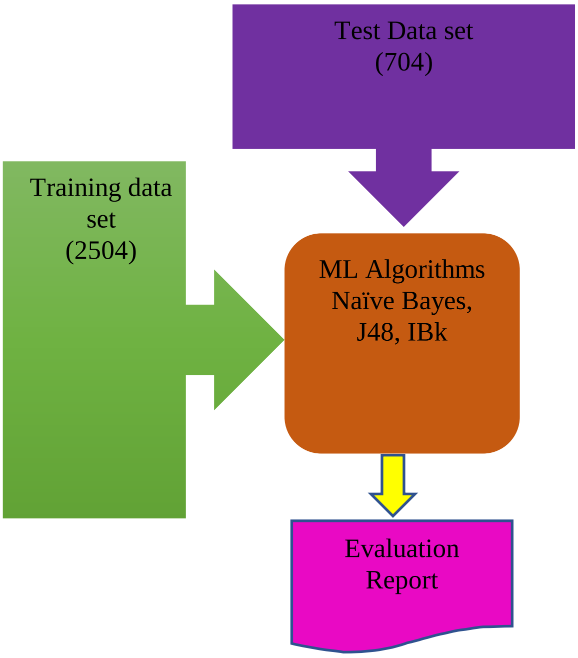


Fig 2. Training and test set

Fig 2 shows that 2504 records have been given as a training data and 704 data is used to test the outcome. The evaluation report is used to analyze the algorithms. Autism can be effectively analyzed by the machine learning algorithms [8].

V. ALGORITHMS

A. Naïve Bayes

Let $C_1C_2, \ldots, C_m$ be m possible classes. Let p $X_1X_2, ..., X_p$ be a set of p predictor values of a record, then the probability that the record belongs to class C_i is:

$$P(C_i \mid X_1,...,X_p) = \frac{P(X_1,...,X_p \mid C_i)P(C_i)}{P(X_1,...,X_p \mid C_1)P(C_1) + \cdots + P(X_1,...,X_p \mid C_m)P(C_m)} \quad (1)$$

Where $P(C_i)$ is called the prior probability and $P(C_i \mid X_1, \ldots, X_p)$ is called the posterior probability. Naïve Bayes is primarily used for situations where all attributes are categorical (numeric attribute values are typically grouped into intervals). Table 1 shows the confusion matrix of the Naïve Bayes algorithm.

Table 1. Confusion Matrix of Naïve Bayes classifier

N=704	Predicted No	Predicted Yes
Actual No	496	19
Actual Yes	2	187

B. k-NN (IBk)

The k-NN classifier is named as IBk in Weka tool. k-NN is a popular, non-parametric method used for regression and classification [6]. The algorithm is

Step 1: Input: an integer value k.

Step 2: To classify a new record, find the nearest k neighboring records in the training set, based on a distance measure which is the normalized Euclidean distance.

Step 3: For a classification problem, classify the record as a member of the majority class of the k nearest neighbors. For a numeric prediction problem, take the average value of the target attribute of the k nearest neighbors as the predicted value.

Distance Calculation

In a p-dimensional space, the Euclidean distance between two records, a= $(a_1, a_2,\ldots,a_p)$ and b=$(b_1,b_2,\ldots,b_p)$, is defined as:

$$d(\mathbf{a},\mathbf{b}) = \sqrt{(a_1-b_1)^2 + (a_2-b_2)^2 + \cdots + (a_p-b_p)^2} \quad (2)$$

It is not necessary to perform square root operation if the purpose is to compare distance.

The Euclidean distance is typically calculated based on normalized values. The Euclidean distance measure implicitly assumes data are numeric. When applied to the categorical data, the difference between two categorical values is defined as zero, if they are the same, and one otherwise. Table 2 shows the confusion matrix of IBk classifier.

Table 2. Confusion Matrix of IBk classifier

N=704	Predicted No	Predicted Yes
Actual No	494	21
Actual Yes	15	174

C. J48

J48 is an extension of ID3. The added features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. An open source Java implementation of the C4.5 algorithm is J48 in WEKA. The WEKA tool provides a number of options like tree pruning. Most of the classification algorithms perform recursively until every single leaf is pure to make sure the classification [9] of the data to be as perfect as possible.

The J48 algorithm generates the rules from which specific uniqueness of that data is produced. The objective is increasingly generalization of a decision tree until it gains a balance of accuracy and flexibility. The best attribute is found on the base of the recent selection criterion and that attribute selected for branching.

$$Entropy(\vec{y}) = -\sum_{j=1}^{n} \frac{|y_i|}{|\vec{y}|}\log\left(\frac{|y_i|}{|\vec{y}|}\right) \quad (3)$$

$$Gain(\vec{y}, j) = Entropy(\vec{y} - Entropy(j|\vec{y})) \quad (4)$$

The outliers are significant to the result. Some instances are existing in all data sets which are not well-defined and vary from the supplementary instances on its neighborhood [7].

Retrieval Number: B2090078219/19©BEIESP
DOI: 10.35940/ijrte.B2090.078219
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The classification is achieved on the instances of the training set and finally, the tree is formed.

Table 3. Confusion Matrix of J48 classifier

N=704	Predicted No	Predicted Yes
Actual No	515	0
Actual Yes	0	189

Table 3 shows the confusion matrix for J48 Algorithm. The pruning is performed for reducing classification errors which are being formed by specialization in the training set. Pruning is achieved for the generalization of the tree.

VI. RESULTS AND DISCUSSIONS

Table 4. Evaluation Results

Evaluation	Classifier		
	Naïve Bayes	J48	IBk
Error Rate	3.07%	0%	5.11%
Percent_Correct	97.27%	100 %	95.23 %
Percent_incorrect	2.73%	0%	4.77%
Entropy_gain	53.54%	59.08 %	27.62 %
Kappa Statistics	0.93	1.00	0.88
Mean Absolute error	0.03	0.00	0.05

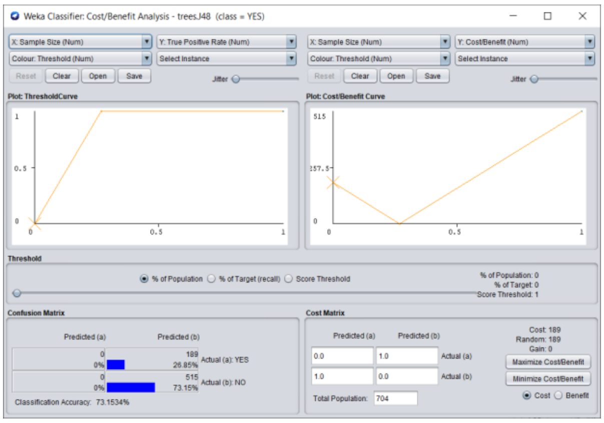
Table 4 shows the Evaluation results for the three classifiers. The error rate is calculated by

Error Rate = Incorrectly classified instances / Total number of instances

The other evaluations like Percent_correct, percent_incorrect, entropy_gain, kappa statistics, mean absolute error are calculated through Weka experimenter. From the results we can easily understand that J48 outperforms compared with other algorithms. The Naïve Bayes performed well compared with IBk algorithm,

Decision tree rules

(result >= 7) => Class/ASD=YES (189.0/0.0)
 => Class/ASD=NO (515.0/0.0)


Weka Classifier: Cost/Benefit Analysis - trees.J48 (class = YES)

X: Sample Size (Num) | Y: True Positive Rate (Num) | Colour: Threshold (Num) | Select Instance

X: Sample Size (Num) | Y: Cost/Benefit (Num) | Colour: Threshold (Num) | Select Instance

Threshold: ● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 0
% of Target: 0
Score Threshold: 1

Confusion Matrix

	Predicted (a)	Predicted (b)	
	0 0%	189 26.85%	Actual (a): YES
	0 0%	515 73.15%	Actual (b): NO

Classification Accuracy: 73.1534%

Cost Matrix

Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Total Population: 704

Cost: 189
Random: 189
Gain: 0

● Cost ○ Benefit

Fig 3. Cost benefit analysis of J48 classifier.

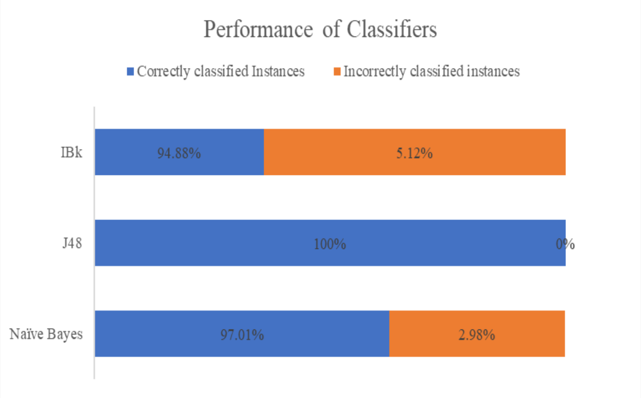
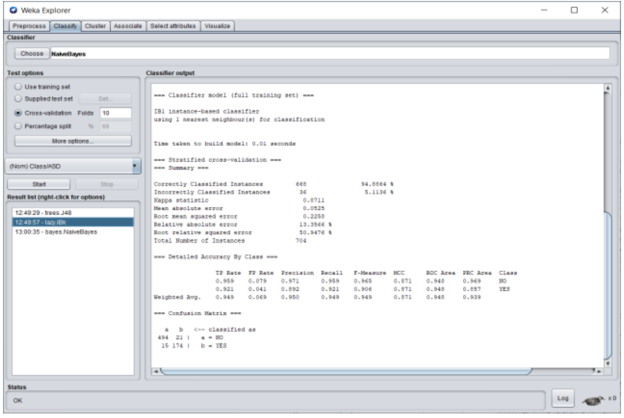


Fig 4. Correctly and incorrectly classified instance


Weka Explorer — Classifier: NaiveBayes; Test options: ● Cross-validation Folds 10; (Nom) Class/ASD

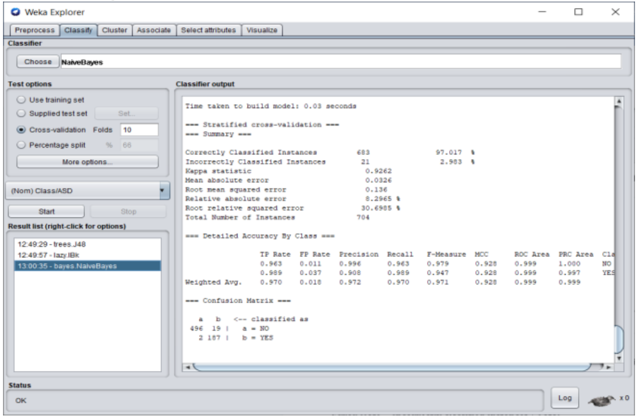
Result list (right-click for options):
12:49:29 - trees.J48
12:49:57 - lazy.IBk
13:00:35 - bayes.NaiveBayes

Status: OK

Classifier output

```
=== Classifier model (full training set) ===

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         668               94.8864 %
Incorrectly Classified Instances        36                5.1136 %
Kappa statistic                          0.8711
Mean absolute error                      0.0525
Root mean squared error                  0.2258
Relative absolute error                 13.3566 %
Root relative squared error             50.9476 %
Total Number of Instances              704

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.959    0.079    0.971      0.959    0.965      0.871    0.948     0.969     NO
                 0.921    0.041    0.892      0.921    0.906      0.871    0.948     0.857     YES
Weighted Avg.    0.949    0.069    0.950      0.949    0.949      0.871    0.948     0.939

=== Confusion Matrix ===

   a   b   <-- classified as
 494  21 |   a = NO
  15 174 |   b = YES
```

Fig 5. Summary of Naïve Bayes classifier


Weka Explorer — Classifier: NaiveBayes; Test options: ● Cross-validation Folds 10; (Nom) Class/ASD

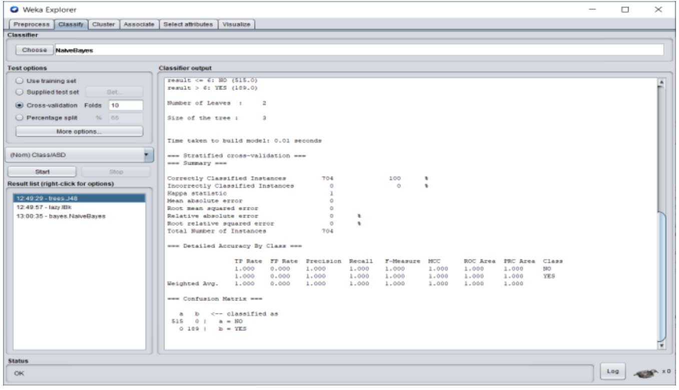
Result list (right-click for options):
12:49:29 - trees.J48
12:49:57 - lazy.IBk
13:00:35 - bayes.NaiveBayes

Status: OK

Classifier output

```
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         683               97.017  %
Incorrectly Classified Instances        21                2.983  %
Kappa statistic                          0.9262
Mean absolute error                      0.0326
Root mean squared error                  0.136
Relative absolute error                  8.2965 %
Root relative squared error             30.6985 %
Total Number of Instances              704

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0.963    0.011    0.996      0.963    0.979      0.928    0.999     1.000     NO
                 0.989    0.037    0.908      0.989    0.947      0.928    0.999     0.997     YES
Weighted Avg.    0.970    0.018    0.972      0.970    0.971      0.928    0.999     0.999

=== Confusion Matrix ===

   a   b   <-- classified as
 496  19 |   a = NO
   2 187 |   b = YES
```

Fig 6. Summary of IBk classifier


Weka Explorer — Classifier: NaiveBayes; Test options: ● Cross-validation Folds 10; (Nom) Class/ASD

Result list (right-click for options):
12:49:29 - trees.J48
12:49:57 - lazy.IBk
13:00:35 - bayes.NaiveBayes

Status: OK

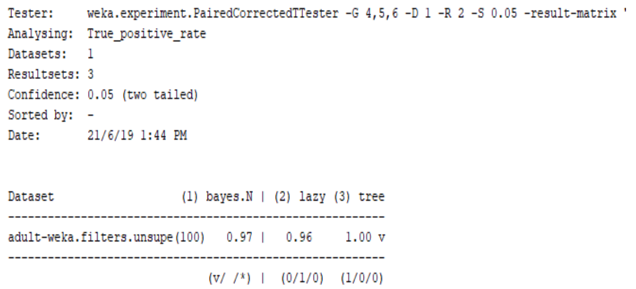
Classifier output

```
result <= 6: NO (515.0)
result > 6: YES (189.0)

Number of Leaves  :     2

Size of the tree :      3


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         704              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances              704

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     NO
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     YES
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

   a   b   <-- classified as
 515   0 |   a = NO
   0 189 |   b = YES
```

Fig 7. Summary of J48 classifier

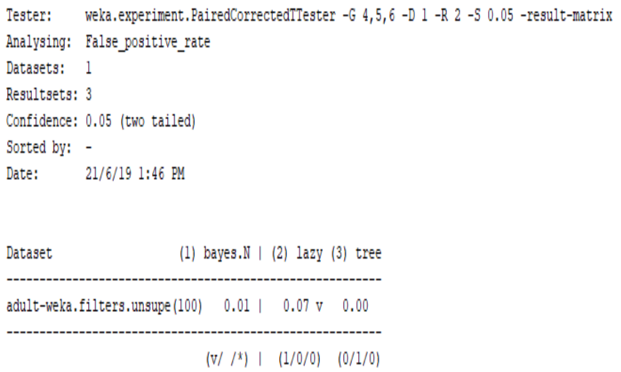
Retrieval Number: B2090078219/19©BEIESP
DOI: 10.35940/ijrte.B2090.078219
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication



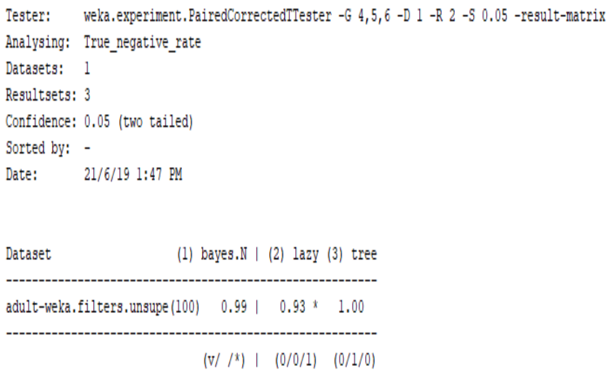
```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix '
Analysing:  True_positive_rate
Datasets:   1
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       21/6/19 1:44 PM


Dataset                   (1) bayes.N | (2) lazy (3) tree
---------------------------------------------------------
adult-weka.filters.unsupe(100)   0.97 |   0.96     1.00 v
---------------------------------------------------------
                          (v/ /*) |  (0/1/0)   (1/0/0)
```

Fig 8. True positive calculation of the three classifiers



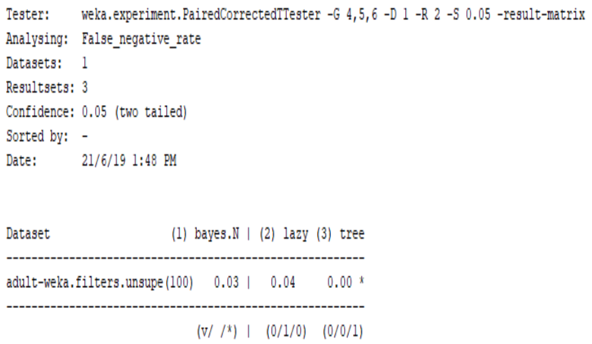
```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix
Analysing:  False_positive_rate
Datasets:   1
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       21/6/19 1:46 PM


Dataset                   (1) bayes.N | (2) lazy (3) tree
---------------------------------------------------------
adult-weka.filters.unsupe(100)   0.01 |   0.07 v   0.00
---------------------------------------------------------
                          (v/ /*) |  (1/0/0)   (0/1/0)
```

Fig 9. False positive calculation of the three classifiers



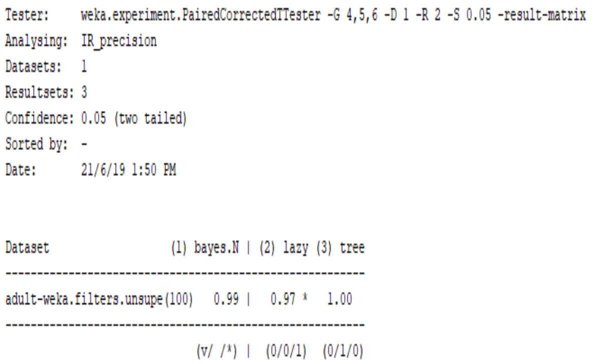
```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix
Analysing:  True_negative_rate
Datasets:   1
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       21/6/19 1:47 PM


Dataset                   (1) bayes.N | (2) lazy (3) tree
---------------------------------------------------------
adult-weka.filters.unsupe(100)   0.99 |   0.93 *   1.00
---------------------------------------------------------
                          (v/ /*) |  (0/0/1)   (0/1/0)
```

Fig 10. True negative calculation of the three classifiers



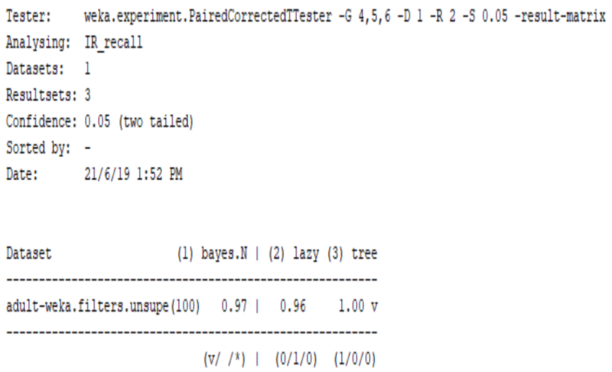
```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix
Analysing:  False_negative_rate
Datasets:   1
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       21/6/19 1:48 PM


Dataset                   (1) bayes.N | (2) lazy (3) tree
---------------------------------------------------------
adult-weka.filters.unsupe(100)   0.03 |   0.04     0.00 *
---------------------------------------------------------
                          (v/ /*) |  (0/1/0)   (0/0/1)
```

Fig 11. False negative calculation of the three classifiers



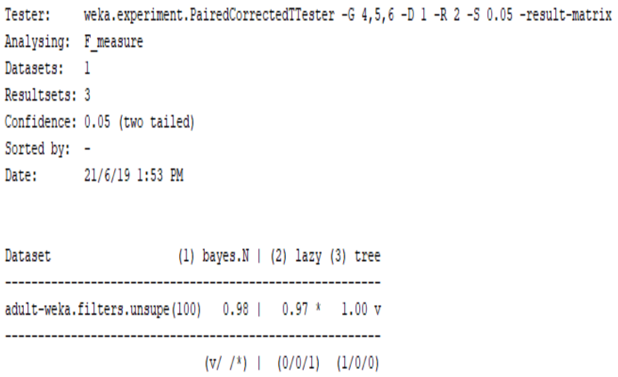
```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix
Analysing:  IR_precision
Datasets:   1
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       21/6/19 1:50 PM


Dataset                   (1) bayes.N | (2) lazy (3) tree
---------------------------------------------------------
adult-weka.filters.unsupe(100)   0.99 |   0.97 *   1.00
---------------------------------------------------------
                          (v/ /*) |  (0/0/1)   (0/1/0)
```

Fig 12. IR_precision calculation of the three classifiers



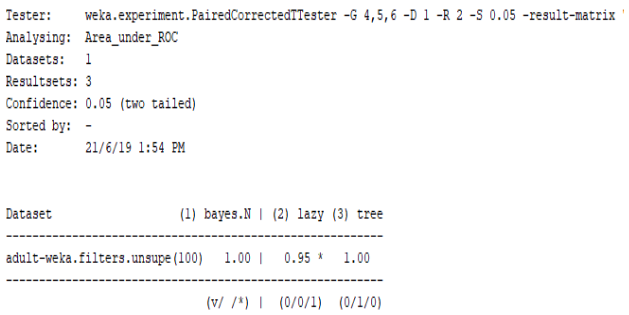
```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix
Analysing:  IR_recall
Datasets:   1
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       21/6/19 1:52 PM


Dataset                   (1) bayes.N | (2) lazy (3) tree
---------------------------------------------------------
adult-weka.filters.unsupe(100)   0.97 |   0.96     1.00 v
---------------------------------------------------------
                          (v/ /*) |  (0/1/0)   (1/0/0)
```

Fig 13. IR_recall calculation of the three classifiers



```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix
Analysing:  F_measure
Datasets:   1
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       21/6/19 1:53 PM


Dataset                   (1) bayes.N | (2) lazy (3) tree
---------------------------------------------------------
adult-weka.filters.unsupe(100)   0.98 |   0.97 *   1.00 v
---------------------------------------------------------
                          (v/ /*) |  (0/0/1)   (1/0/0)
```

Fig 14. F_measure calculation of the three classifiers



```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix '
Analysing:  Area_under_ROC
Datasets:   1
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       21/6/19 1:54 PM


Dataset                   (1) bayes.N | (2) lazy (3) tree
---------------------------------------------------------
adult-weka.filters.unsupe(100)   1.00 |   0.95 *   1.00
---------------------------------------------------------
                          (v/ /*) |  (0/0/1)   (0/1/0)
```

Fig 15. Area_under_ROC calculation of the three classifiers

Figure 3 to Figure 15 are the analysis which has been done through Weka.

VII. CONCLUSION

This paper aimed to analyze the three classifiers to find the outperforming classifier to predict ASD using the data set obtained from UCI repository. The hypothesis of the paper is to find the novelty of the machine learning models which are trained with minimum behavior sets are proficient of better performance or not. The dataset is processed using WEKA tool. The tool shows that J48 predicted the disease with 0% error rate. The future work will be to utilize this result and to develop a software or mobile application to predict the ASD in advance.

REFERENCES

1. J. Baio, “Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years-Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014,” vol. 67, 2018
2. A. P. Association and others, Diagnostic and statistical manual of mental disorders (DSM-5®), American Psychiatric Pub, 2013.
3. F. Thabtah, “Machine learning in autistic spectrum disorder behavioral research: A review and ways forward,” Informatics Heal. Soc. Care, vol. 0, no. 0, pp. 1–20, 2018.
4. G. Holmes, A. Donkin, and I. H. Witten, “Weka: A machine learning workbench,” in Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on 1994, pp. 357–361.
5. WHO, Autism spectrum disorders, 2017 [Accessed August 22, 2018]? [Online]. Available: http://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders
6. D. Bone, S. L. Bishop, M. P. Black, M. S. Goodwin, C. Lord, and S. S. Narayanan, “Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion,” Journal of Child Psychology and Psychiatry, vol. 57, 2016.
7. B. van den Bekerom, “Using machine learning for detection of autism spectrum disorder,” 2017.
8. W. Liu, M. Li, and L. Yi, “Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework,” Autism Research, vol. 9, no. 8, pp. 888–898, 2016.
9. Gomathi, S., and V. Narayani, “Early prediction of systemic lupus erythematosus using hybrid K-Means J48 decision tree algorithm”, International Journal of Engineering & Technology 7.1.3 (2017): 28-32
10. https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult.

AUTHORS PROFILE

S. Gomathi is working at PSGR Krishnammal College for Women. She has 7 years of teaching experience. Her area of interest are Big data, data mining, machine learning and ETL. She published 5 papers in SCOPUS indexed journals and 10 papers in International journals. She submitted her Ph.D. thesis in Computer Science at Bharathiar University. She has working experience in the latest technologies like Apache Hive, Apache Pig, Hadoop, Apache Flume, Solr and the data analytics tools like Weka, Python, and Rattle. She has good experience in the ETL and reporting tools like Informatica, Talend, Pentaho, MicroStrategy SSIS, and SSRS.