Effective Implementation of Pre-Processing Techniques in Machine Learning for Autism Spectrum Disorder

N. Priya, C. Radhika

Abstract: Autism Spectrum disorder (ASD) is a neurobiological developmental disorder is symbolize by means of the impairment of social interaction, stereotypic behaviours, and communiqué lack. Early deduction of ASD will enhance the fine of lifestyles of the affected person. The objective of the paper is to focus on the application of various Machine Learning strategies applied for the autism dataset for diagnosing ASD. In this study, the effective pre-processing techniques One-hot encoding, Splitting and Scaling are used to standardize the dataset and the Principal Component Analysis (PCA) evaluator method is applied for the best feature selection. This technique is investigated with various Machine learning techniques like Random Forest, SVM, Logistic Regression, KNN, Naive Bayes. Comparatively, the effective Pre-Processing technique with Random Forest model shows the better accuracy of 92% in diagnosing ASD. When with other metrics such as accuracy, precision, recall, F1-score, ROC, error rate.

Keyword: ASD, Machine Learning techniques, PCA.

I. INTRODUCTION

ASD is a neuro developmental ailment that affects someone's studying skills, interplay and verbal exchange. Although the prognosis of autism may be carried out at any age however the signs and symptoms generally appear inside the first 24 months of life and develop thru time[17]. ASD youngsters are identified for their impairment in social interplay skills and deciphering the emotional facial features in other[18]. Autism Spectrum Condition (ASC) is a fixed of neuro development syndromes that affects mind functions [21]. ASD is characterized via repetitive behaviour, verbal and nonverbal communiqué, the want for sameness and irregularity in social interplay [25].

II. LITERATURE REVIEW

Machine Learning techniques play a dynamic role for diagnosing ASD in the analysis of dataset. This section, mainly focus on predicting ASD using different Machine Learning techniques. Fadi Thabtah [1] used DSM-5 tool to identify the ASD with merits and demerits. Padmapriya.S, et al.[2] proposed different pre-processing techniques like Chi-Square, Information Gain, Relevant feature selection and reduced the feature set. Relevant feature selection techniques gave more accuracy when compared to other techniques.

Uma Rani.R et al.[3] proposed the comparison of classification algorithms with statistical models in autism dataset. Arodami Chorianopoulou et al. [5] used with different modalities like audio, text, video and with parent's action of interactions of typically developing(TD) and ASD children. The moderate accuracy occurred only due to engagement on parents behaviour.

Vaishali R. et al.[6] Used ASD data from UCI machine learning repository experimented with binary firefly feature selection wrapper and obtained a better classification reports on minimum feature subsets. Tibaduiza et al.[8] proposed a comparison study between the PCA (Principal Component Analysis) and ICA (Independent Component Analysis) were described. An significant difference between PCA and ICA were related to the quantity of components used in the technique. Kazi Shahrukh Omar et al.[9] proposed a model by merging random forest-ID3, Random forest-CART for predicting the autism traits. The evaluation done with AQ10 dataset and 250 real dataset collected from various persons. The results were compared in terms of different metrics. Sofia Visa et al.[10] created a subset of features based on criteria and yielding better accuracy using CART classification. Mofleh Al Diabat et al.[20] author described a new Ensemble Classification for Autism Screening(ECAS) to predict ASD traits and also reduces biased decisions. Paul Fergus et al.[21] proposed a 3D animation solution developed for Mobile device. It helps ASD people to understand the facial expressions and give awareness to engage real-time situations.

III. THE DATASET

The dataset entitled, "Autistic Spectrum Disorder Screening Data for Toddlers" is an open source dataset from Kaggle Repository. The dataset consists of 1054 observations of 18 features of different variable type. Dataset description is shown in Table1.

It contains categorical variables like Gender, Ethnicity, Jaundice, Family_members_with_ASD, Who is attaining test, Class/ASD Traits, as well as 10 binary variables representing the screening questions (A1 to A10), and 2 numeric variables (Age & Score).

Revised Manuscript Received on March 30, 2020.
* Correspondence Author

Dr. N. Priya*, Associate Professor, Department of Computer Science, SDNBV College for Women, Chennai, Tamil Nadu, India.

C. Radhika, Asst. Professor, Department of Computer Science, SDNBV College for Women, Chennai, Tamil Nadu, India.

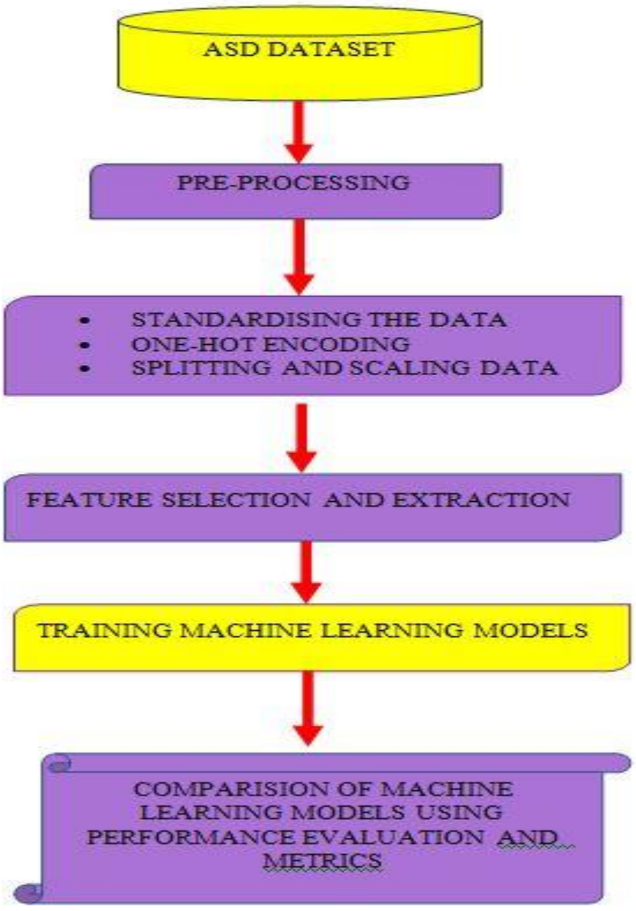
Table1: Dataset description

Feature Name	Type	Description of the Feature
A1: Q1	Binary (0,1)	Does your child look at you call his/her name?
A2: Q2	Binary (0,1)	How easy is it for you to obtain eye contact with your child?
A3: Q3	Binary (0,1)	Does your child point to specify that she/he needs something?
A4:Q4	Binary (0,1)	Does your child point to share curiosity with you?
A5: Q5	Binary (0,1)	Does your child pretend?
A6: Q6	Binary (0,1)	Does your child go behind where you're looking
A7: Q7	Binary (0,1)	When you or someone else in the family is noticeably upset, does your child show signs of counsel to comfort them?
A8: Q8	Binary (0,1)	Your child's first words as:
A9: Q9	Binary (0,1)	Does your child use simple gestures?
A10: Q10	Binary (0,1)	Does your child gaze at nothing with no apparent purpose?
Age	Number	Age in years
Score by Q-chat-10	Number	1-10(Less than or equal 3 No ASD traits, > 3 ASD traits
Gender	String	Boy / Girl
Ethnicity	String	List of Communal ethnicity in text setup
Jaundice	Boolean	Whether case become born with jaundice
Family member with ASD history	Boolean	Whether any immediate member of the family has a PDD
Who is attaining the test	String	Parent, Self, Caregiver, Medical workforce, Clinician, and so forth
Class/ASD Traits	Boolean	ASD ailments or No ASD ailments

IV. METHODOLOGY

The pre-processing of the information is a crucial step in Machine Learning applications. For early stage detection, the proposed system has been implemented with the effective pre-processing methodology for the machine learning models in identifying ASD. The flow chart of the proposed methodology is define in the fig1. The Machine Learning techniques for predictive task that includes Logistic Regression, K-NN, SVM, Naïve Bayes, Random Forest. These methods are compared with different performance metrics and used for better decision making.

Fig1: Framework of Pre-processing.


ASD DATASET
PRE-PROCESSING
STANDARDISING THE DATA
ONE-HOT ENCODING
SPLITTING AND SCALING DATA
FEATURE SELECTION AND EXTRACTION
TRAINING MACHINE LEARNING MODELS
COMPARISION OF MACHINE LEARNING MODELS USING PERFORMANCE EVALUATION AND METRICS

A. Machine Learning Techniques

- LOGISTIC REGRESSION (LR): LR is a statistical linear model used for binary class, taking as input a set of independent attributes. The independent attributes can be discrete or non-stop and are used to predict the possibility of the goal result taking a suitable binary value [18].
- K-NN: K-NN is a kind of non-parametric or instance-based learning in which the undertaking is simplest anticipated domestically and all computation is deferred till classification. K-NN classification, the output is a class club. An item is assessed through a majority vote of its neighbours, with the thing being decided on to the class most common among its k nearest neighbours. If k = 1, then the thing is just assigned to the class of that single adjacent neighbours [22].
- SUPPORT VECTOR MACHINE (SVM): The key objective of SVM model is to maximize the boundary. The boundary is defined as the distance between the unscrambling hyperplane and the training samples that are nearby to this hyperplane, which are the so-called support vectors [4].
- NAIVE BAYES : Naive Bayes is a probabilistic classifier which makes use of Bayes theorem to make the assumption that each one enter attributes are linearly independent after which measure the possibility of every feasible classification given a set of autonomous attributes [23][18].

- RANDOM FOREST : A Random forest is a classifier such as a group of tree-structured classifiers $\{h(x, \Theta_k), k=1, \dots\}$ where the $\{\Theta_k\}$ are independent identically dispensed random vectors and each tree casts a unit vote for the maximum general class at input x [12].

B. DATA PRE-PROCESSING

- Standardising the data: In dataset there are some variables they do not offer any benefit of our analysis.

i) Ethnicity: It will display only the list of common ethnicity.

ii) Who is completed the test: This variable also had only the details of parents, self, medical staff, etc.

ii) Class / ASD traits: It will display only the Class_ASD=yes/ no. The machine learning models would really have the outcome of the target variable. For the purpose of better analysis these three variables are removed.

- ONE-HOT ENCODING: One-hot encoding technique is used to build a new dummy attribute of every distinctive value of the nominal attribute column [4]. Generally it is a method to convert non ordinal categorical variables into numeric data.
- SPLITTING AND SCALING: [4]Splitting a dataset into training sets and test sets and also require to select features is on the same scale for optimal performance. The data will be randomly allocated to training set and one third will allocate to the testing set. In scaling, the different features are transformed into a uniform scale value. There are two general approaches to bringing different attribute of the same scale: Normalization and Standardization. Normalization via min-max scaling is useful in bounded intervals. Standardization via StandardScaler is useful in outliers. This system uses
Standardization for optimal performance. Standardization can be articulated by the subsequent equation:

$$X^{(i)}\text{ std} = \frac{X^{(i)} - \mu_x}{\sigma_x}$$

Where, $X^{(i)}$ - particular sample, μ_x -particular feature column and σ_x - standard deviation.

- DIMENSIONALITY REDUCTION: The dimensionality reduction strategies are classes by means of attributes selection and attributes extraction. Attributes selection is selecting a subset of the novel attributes and Attributes extraction to transform the information onto a brand new attributes subspace [4].

This system uses Sequential Backward Selection (SBS) algorithm to condense the dimensionality of the initial attributes subspace with a minimum. The idea behind the SBS algorithm is sequentially remove attributes from the complete attributes subset until the new attributes subspace contains the desired quantity of attributes[4]. Next for attributes extraction-Principal component analysis(PCA) [8,17] can be used to reduce a complex dataset to a lower dimensionality. Each main aspect is a linear mixture of the innovative variables. All the main additives are orthogonal to every other, so there's no redundant information.

The subset in PCA are define with the aid of the eigenvectors and eigenvalues of the covariance matrix as follows:

$$C_X R = R\Delta$$

Where the eigenvectors of C_X are the columns of R and the eigenvalues are the diagonal terms of Δ[4].

V. RESULT AND DISCUSSION

Evaluation metric parameters for Pre-processing such as accuracy, precision, recall, F1-score, ROC, error rate are compared and the results are listed in Table II and Table III for the Machine Learning Models. The performance evaluation of the classifiers using all the 18 features before feature selection and feature extraction shown in Table-II and the graphical representation of the analysis shown in fig2. Among 18 only 5 important features are selected based on interaction, communication, facial expression and apply the Pre-processing techniques and fed as input to each classifier. The performances of the Classifiers are evaluated using cross validation and confusion matrix. The comparative results indicate that the combination of Random forest with PCA results in a better performance when compare to other classifiers. Random forest outperforms the rest of the classifiers in terms of accuracy (92%), precision, recall, F1-score, ROC, error rate as depicted in Table-III. Figure3 shows the performance evaluation of the classifiers after Dimensionality reduction.

Table -II Predictive Performance of Classifiers before feature selection.

Evaluation Measure	KNN	Logistic Regression	Naive Bayes	Random Forest	SVM
Accuracy	77%	83%	74%	79%	81%
F1-Score	0.79	0.84	0.76	0.8	0.83
Precision	0.78	0.84	0.75	0.83	0.79
Recall	0.79	0.85	0.77	0.78	0.87
ROC	0.76	0.82	0.73	0.76	0.8
Misclassification rate	0.23	0.16	0.26	0.2	0.19

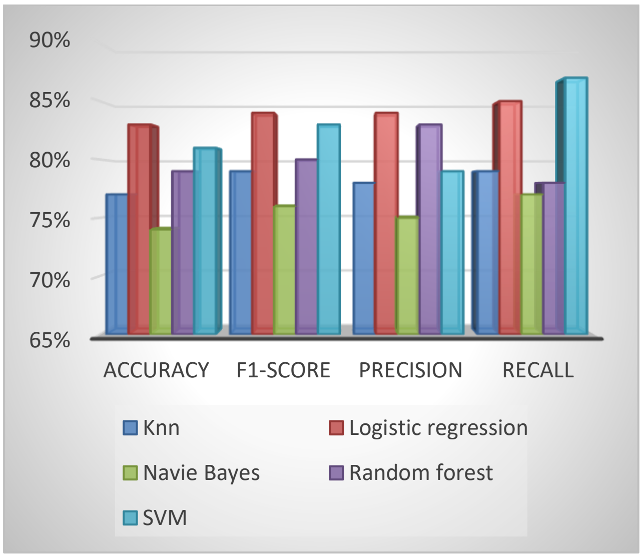


Figure 2: Performance analysis on the classifiers before feature selection

Table-III Predictive performance of classifiers after Feature selection-SBS and Feature extraction-PCA.

Evaluation Measure	KNN	Logistic Regression	Naive Bayes	Random Forest	SVM
Accuracy	71%	85%	74%	92%	84%
F1-Score	0.7	0.86	0.76	0.93	0.85
Precision	0.81	0.86	0.76	0.95	0.85
Recall	0.61	0.87	0.77	0.91	0.86
ROC	0.72	0.85	0.73	0.92	0.84
Misclassifi-cation rate	0.28	0.14	0.25	0.075	0.15

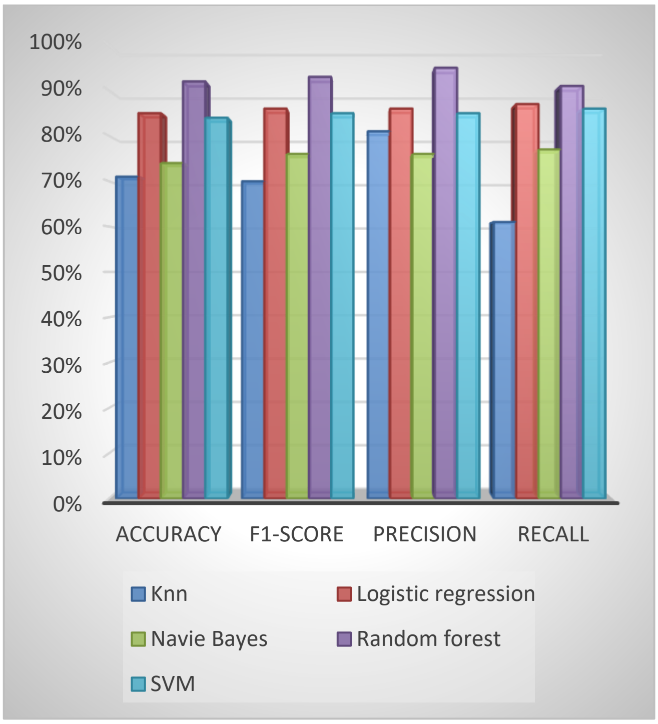


Figure 3: Performance analysis on the classifiers after feature selection

VI. CONCLUSION

The proposed system implementing the effective pre-processing techniques with different metric parameters is used for evaluation of machine learning models to predict ASD. The feature selection and extraction methodology is proposed to with the minimal subset and used to classify the autism. Comparatively, Random forest pooled with PCA gives better accuracy of 92% when compared to other machine learning algorithms. The minimal subset feature to identify the ASD in early stage and improve the life style of the affected people is the future scope.

REFERENCES

1. Fadi Thabtah, "Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment," Nelson Marlborough Institute of Technology 42 Upper Queen Street, Auckland, New Zealand.
2. S. Padmapriya , S. Murugan, "A Novel Feature Selection Method for Pre-Processing the ASD Dataset," International Journal of Pure and Applied Math. 2018, ISSN(p):1311-8080, ISSN(e):1314-3395,Vol 118 No.8 pp.17-25,www.ijpam.eu.
3. R. Uma Rani, R. Suguna, "Exploratory Data Analysis of Autism Data", IOSR Journal of Engineering. ISSN(e):2250-3021,ISSN(p):2278-8719,pp 05-10.
4. Sebastian Raschka, "Python Machine Learning," Sep 2015,ISBN 978-1-78355-513-0, www.packtpub.com
5. Arodami Chorianopoulou, Efthymios Tzinis, Elias Iosif, Asimenia Papoulidi, Christina Papailiou, Alexandros Potamianos, "Engagement detection for children with autism spectrum disorder," March 2017, DOI:10.1109/ICASSP.2017.7953119.
6. R.Vaishali, R.Sasikala, "A machine learning based approach to classify Autism with optimum behaviour sets," International Journal of Engineering & Technology, Aug 2018, DOI: 10.14419/ijet.v713.18.14907 Published.
7. Pream Sudha, "Feature Selection Techniques for the Classification of Leaf Diseases in Turmeric," International Journal of Computer Trends and Technology, Jan 2017,ISSN:2231-2803,Vol 43 No.3, pp.138-142.
8. J.Rodellar, L.E.Mujica, D.A. Tibaduiza, M.Anaya and A.Guemes "Principal Component Analysis vs. Independent Component Analysis for Damage Detection," 6th European workshop on Structural health monitoring-fr.1.D.4.
9. K.S. Omar, P. Mondal, N. S. Khan, M. R.K. Rizvi, M.N. islam, "A Machine Learning Approach to predict Autism Spectrum Disorder," Proc. Int. Conf. Electr., Comput. Commun.Eng,(ECCE), Bangladesh, 2019,pp.1-6.
10. Sofia Visa, Brian Ramsay, Anca Ralescu, Esther van der knap, "Confusion Matrix-based Feature Selection," CEUR Workshop Proc., Vol 710,pp120.127,2011.
11. Fadi Thabtah, Li Zhang and Neda Abdelhamid, "Nba game result prediction using feature analysis and machine learning," Springer-Verlag GmbH Germany,part of Springer Nature, March 2019, Vol 6(1),pp 103-116.
12. Breiman L, "Random Forests, Machine Learning," 45,5-32,(2001).
13. Ertan Mustafa Geldiev, Navden Valkov Nenkov, Mariana Mateeva Petrova, "Exercise of Machine Learning Using Some Python tools and Techniques," CBU International Conference Proceedings, Sep 2018,Vol 6.
14. J. Brownlee, "Machine Learning Mastery with Python understand your Data," Create accurate model sand work projects End-End, pp.123-145.
15. Guido S., Muller A.(2016, Oct), Introduction to Machine Learning with python, A Guide for Scikit-Learn.
16. UCI machine Learning Repository.(2017). Retrieved from Http://Archive.Ics.Uci.Edu/Ml/Index.Php
17. U.Frith and F. Happe, "Autism Spectrum Disorder," Current Biology, Vol.15, No.19, pp.R786-R790,2005.
18. Kayleigh K.Hyde, Marlena N.Novack LaHye, Chelsea Parlett-pelleritil, Raymond Anden Dennis R.Dixon, Erik Linstead, "Application of Supervised Machine Learning in Autism Spectrum Disorder Research: a Review," Review Journal of Autism and Developmental Disorders, Feb 2019, Vol 6(2), pp.128-146.
19. Jollife I.T, "Principal Component Analysis," Springer series in statistics, 2 ed.2002.
20. Mofleh Al Diabat and Najah Al-Shanableh, "Ensemble Learning Model for Screening Autism in Children," International Journal of Computer Science & Information Technology, Vol 11, no 2, Apr 2019.
21. Paul Fergus, Basma Abdulaimma, Chris Carter, Sheena round, "Interactive Mobile Technology for Children with Autism Spectrum Condition," IEEE 11th Consumer Communications and Networking Conference, Jan 2014.
22. Tanvi Sharma, Anand Sharma, Vibhakar Mansotra, "Performance Analysis of Data Mining Classification Techniques on Public Health care Data," International Journal of Innovative Research in Computer and Communication Engineering, June 2016, Vol. 4, pp. 11381-11386.
23. Bishop C.M(2006). "Pattern recognition and machine learning," New York, NY:Springer.

24. S.Baron Cohen, "Autism and Asperger Syndrom the facts," 2008,pp:176.
25. S.Baron-Cohen, "Autism The Empathizing-Systemizing(E-S) Theory," Annals of the New York Academy of Sciences, Vol.1156, No.2009, pp:68-80,2009.
26. Thabtah, F. "Machine Learning in Autism Spectrum Disorder Behavioural Research: A Review and ways forward," Informatics for Health and Social Care Journal, 2018.

AUTHORS PROFILE

Dr. N. Priya, is having more than 15 years of teaching experience. She was presented Ph. D from Bharathiar University Coimbatore. Her research regions consist of datamining, image processing, Neural Networks, Network programming and Fuzzy Logic. Currently she is employed as Associate Professor in the PG Department of Computer Science, SDNBV College for Women.

C. Radhika, is presently employed as Asst. Professor in Department of Computer Science, SDNBV College for Women, Chennai. She has quite 11 years of teaching experience. Her research regions consist of Machine Learning.