ication_Algorithms_of_the_Autism_Spectrum_Disorder_Diagnosis.pdf

by Armin Lawi

Comparison of Classification Algorithms of the Autism Spectrum Disorder Diagnosis

1st Armin Lawi
Department of Computer Science
Hasanuddin University
Makassar, Indonesia
armin@unhas.ac.id

2nd Firman Aziz
Faculty of Mathematics and Natural Sciences
Pancasakti University
Makassar, Indonesia
firmanaziz88@gmail.com

***Abstract*—ASD sufferers face difficulties in early development compared to normal humans. Various tools, clinical, and non-clinical approaches have been implemented but take a long time to produce a complete diagnosis. the solution by adopting machine learning. This study proposes the application of cross-validation techniques in the Decision Tree method, Linear Discriminant Analysis, Logistic Regression, SVM, and KNN and determines the best k value in each classification method because the shift of datasets when using cross-validation techniques in the classification method is one factor that can cause the estimate to be inaccurate. The results show that the decision tree provides an accuracy of 100% in each of the k values that have been determined previously. 96.9% on Linear Discriminant Analysis with *k = 7, k = 9*, and *k = 10*. 99.7% in Logistic Regression with values of *k = 2* and *k = 3*. 99.9% in Support Vector Machine with values of *k = 9* and *k = 10* and 94.2% for K-Nearest Neighbors with a value of *k = 8*.**

Keywords—autism spectrum disorder, decision tree, linear discriminant analysis, logistic regression, support vector machine, k-nearest neighbor

I. Introduction

The development of the brain that is impaired by limiting social communication and behavior is Autism Spectrum Disorder (ASD) which occurs due to genetic and neurological factors [1]. ASD sufferers face serious difficulties in early development compared to humans in general. These difficulties vary such as difficulty in responding to sensory information (vision, hearing, smell, etc.) and communication difficulties that make it difficult to interact with others [2]. American Psychiatric Association diagnoses ASD clinically by evaluating 3 behavioral domains, namely communication and language, reciprocal social interaction, and limited activity [3].

Various tools, clinical and non-clinical approaches have been applied to diagnose ASD but use manual calculations by determining the score to produce a precise diagnosis that requires concentration from the user and requires an expert doctor for accurate diagnosis and takes a long time to produce a complete diagnosis [4]. Some studies have begun to adopt machine learning methods to diagnose ASD with the aim of efficient diagnosis time, faster access to services and increased diagnosis accuracy [5], [6], [7].

Yun Jiao et. *al* [8] built a diagnostic model for ASD, based on regional thickness measurements extracted from SBM (Surface-based morphometry) and compared it with the diagnostic model based on volumetric morphometry as a dataset, four machine learning techniques were used to diagnose ASD namely support vector machines (SVMs), multilayer perceptrons (MLPs), functional trees (FTs), and logistic model trees (LMTs) [8]. Then the study that Wenbo Liu et al, proposed a new framework based on facial recognition as a dataset has been proposed and used as a classification method for the diagnosis of ASD [9]. Fadi Thabtah's research, developing intelligent diagnostic tools based on machine learning by replacing artificial rules in the ASD screening tool with predictive models [10] and Osman Altay using the dataset to apply it into the Linear Discriminant Analysis and K-Nearest Neighbor method [11] but in research Osman Altay has not used cross-validation techniques to evaluate classifier performance with a tendency to increase accuracy of the classifier. However, cross-validation has the potential to shift datasets, one of the dangerous factors that is often not taken into account and can lead to inaccurate estimates [12].

This study proposes the application of cross-validation techniques in five classification methods, namely Decision Tree, Linear Discriminant Analysis, Logistic Regression, Support Vector Machine, and K-Nearest Neighbors to see the best estimation results and k-fold values in each classification method.

II. Literarture Review

A. Decision Tree

The decision tree algorithm is a prediction model using tree structure or hierarchical structure. The concept of a decision tree is to convert data into decision trees and decision rules. Some steps in making a decision tree with C4.5 algorithm [13] are:

1) Preparing training data, can be taken from historical data that has happened before and has been grouped into certain classes.

2) Determine the root of the tree by calculating the highest gain value of each attribute or based on the lowest entropy index value. Previously the entropy index value was calculated first, with the formula:

$$Entropy(i) = \sum_{j=1}^{m} f(i,j).2f[(i,j)] \quad (1)$$

3) Calculate the gain value using the following formula:

$$gain = -\sum_{i=j}^{p} \frac{ni}{n}.IE(i) \quad (2)$$

4) To calculate the gain ratio we need to know a new term called Split Information:

$$Split\ Information = -\sum_{t=1}^{c} \frac{s_1}{s} log_2 \frac{s_1}{s} \quad (3)$$

5) *Then calculate the gain ratio*

$$Gain\ Ratio\ (S,A) = \frac{Gain(S,A)}{Split\ Information(S,A)} \quad (4)$$

6) *Repeat step 2 until all records are partitioned*

The decision tree partition process will stop when:

a) All tuples in a record in node m get the same class

b) There are no attributes in the records that are partitioned again

c) There are no records in an empty branch.

B. Linear Discriminant Analysis

Basically, Linear Discriminant Analysis operates through the calculation of variance values within and between classes [14], [15]. In Linear Discriminant Analysis, it is necessary to calculate the scatter matrix in class and between classes using equation (5) and equation (6).

$$S_w = \sum_{i=1}^{a} \Pr(C_i)\Sigma_i \quad (5)$$

$$S_b = \sum_{i=1}^{a} \Pr(C_i)(m_i\text{-}m)(m_i\text{-}m)^T \quad (6)$$

To calculate the value of Σ_i equation (7) is used.

$$\Sigma_i = E[(x\text{-}m_i)(x\text{-}m_i)^T] \quad (7)$$

Where x is the sample vector and m_i is the average value in a different class. Next, calculate the discriminative power value using equation (8) [13], [15].

$$j(w) = \frac{\|w^T S_w w\|}{\|w^T S_b w\|} \quad (8)$$

The value of w is an optimal discrimination projection matrix by solving a common eigenvalue problem [14]. calculated as a result of equation (9).

$$S_b w = \lambda_w S_w w \quad (9)$$

The results of the calculation of the linear discriminant function are obtained using equation (10).

$$d(x) = w^T \left(x\text{-}\frac{m}{w}\right) \quad (10)$$

C. Logistic Regression

Logistic regression is a statistical technique that is widely used to solve classification and regression problems by predicting the probability of an event by matching data in the logit function [16]. The classifier function is stated as follows:

$$log\left(\frac{p}{1-p}\right) = \sum_{i=1}^{i=n} \beta^{(i)} * x^{(i)} + e = \beta^T x + e \quad (11)$$

Where $\beta = (\beta^{(1)}, \beta^{(2)}, \dots, \beta^{(n)})$ are the vector coefficients of the hyperplane, the probability in equation (11) can be expressed as equation (12) following.

$$p = \frac{e^{\beta^T x + e}}{1 + e^{\beta^T x + e}} \quad (12)$$

D. Support Vector Machine

Support Vector Machine is used in pattern classification problems that maximize hyperplane to separate data into two classes in the feature space. SVM is very suitable for use in linearly separated data [17]. Figure 1. shows the basic support vector machine [18], [19], [20].

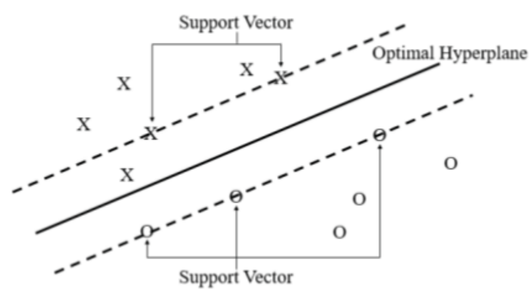

Fig. 1. Base of Support Vector Machines

For grouping data that cannot be separated linearly, the data is transformed to a higher dimension so that it can be separated linearly. This method is called the kernel method, which converts a linear SVM to a non-linear SVM [19], [20]. Moreover, SVM can also be used to classify the data into multiclass [21], [22], [23].

E. K-Nearest Neighbors

The application of the K-Nearest Neighbors algorithm is very easy in the literature study because it is based on the calculation of distances using a numeric dataset [24].

The most commonly used distance calculation method is Euclidean. To calculate the Euclidean use equation (13).

$$Euclidean_{i,j} = \sqrt{\sum_{k=1}^{n} (x_{ik}\text{-}x_{jk})^2} \quad (13)$$

The KNN algorithm consists of 4 steps. In the first step, calculate the distance of the new data to all data. The second step, the distance will be sorted. The third step takes the smallest k value and the last step is to determine the class [11].

F. *Cross Validation*

Cross-validation is a model validation technique to assess how the results of the statistical analysis will generalize independent datasets [26]. This technique is used to predict the model and estimate how accurate a predictive model is when it is run in practice. The use of cross-validation based on the available data is not enough to partition into training and testing data without losing significant modeling or testing capabilities and limiting problems such as overfitting.

Overall, cross-validation is combining the size of the suitability (error prediction) by looking at the average value of each round to get a more accurate prediction of the performance of the prediction model.

G. *Accuracy Performances*

Performance tests are conducted to determine the accuracy of the algorithm. Accuracy is the closer between the prediction data and the actual data or the ratio of the number of data classified correctly [27]. To calculate the level of accuracy using equation (6).

$$accuracy = \frac{TP+TN}{TN+FP+FN+TP} \tag{14}$$

1) *True Positive* (TP): The test is detected having ASD by the model and so it is in the actual..
2) *True Negative* (TN): The test is detected not having ASD and it is not in the actual.
3) *False Positive* (FP): The test is detected not having ASD but it is not in the actual (i.e., having ASD).
4) *False Negative* (FN): The test is detected not having ASD but it is having ASD in the actual.

III. Experimental

A. *Dataset*

The data used in this study were obtained from studies conducted by Tabtah, Fadi [10]. The data set consists of 704 samples aged 17 years and over with 19 different attributes. In the data set, there are 10 questions that are directly related to ASD. All features and questions given in detail are shown in Table I.

The output value of the dataset consists of two classes. 19 different features have been labeled as input values and determine whether they have ASD or not. 0 value indicates 'not having ASD', while 1 indicates 'having ASD'.

TABLE I. Description of The Dataset

Research Variable	Attribute	Description
Response Variable	Class Autistic Spectrum Disorder (ASD)	1 = yes 2 = no
Predictor Variable	Age	Number
	Gender	1 = Male 2 = Female
	Ethnicity	1 = Asian 2 = Black 3 = Hispanic 4 = Latino 5 = Middle Eastern 6 = Pasifika 7 = South Asian 8 = Turkish 9 = White-European 10= Others
	Born with jaundice	1 = yes 2 = no
	Family member with PDD	1 = yes 2 = no
	Who is completing the test	1 = Health care professional 2 = Parent 3 = Relative 4 = Self 5 = Others
	Country of residence	1 = Afghanistan 2 = AmericanSamoa 3 = Angola 4 = Argentina 5 = Armenia 6 = Aruba 7 = Australia 8 = Austria 9 = Azerbaijan 10 = Bahamas 11 = Bangladesh 12 = Belgium 13 = Bolivia 14 = Brazil 15 = Burundi 16 = Canada 17 = Chile 18 = China 19 = Costa Rica 20 = Czech Republic 21 = Cyprus 22 = Ecuador 23 = Egypt 24 = Ethiopia 25 = Finland 26 = France 27 = Germany 28 = Hong Kong 29 = Iceland 30 = India 31 = Indonesia 32 = Iran 33 = Iraq 34 = Ireland 35 = Italy 36 = Japan 37 = Jordan 38 = Kazakhstan 39 = Lebanon 40 = Malaysia 41 = Mexico 42 = Nepal 43 = Netherlands 44 = New Zealand 45 = Nicaragua 46 = Niger 47 = Oman 48 = Pakistan 49 = Philippines 50 = Portugal 51 = Romania 52 = Russia 53 = Saudi Arabia 54 = Serbia 55 = Sierra Leone 56 = South Africa 57 = Spain 58 = Sri Lanka 59 = Sweden 60 = Tonga 61 = Turkey 62 = Ukraine 63 = UAE

Research Variable	Attribute	Description
		64 = United Kingdom 65 = United States 66 = Uruguay 67 = Viet Nam
	Used the screening app before	1 = yes 2 = no
	Scored (positive number of questions)	Number
	It often hears voices that others do not hear.	Number
	It usually focuses on the big picture by going out from small details.	Number
	It can easily follow the conversations of different people in a social group.	Number
	Can easily switch between different activities.	Number
	Does not know how to chat with her peers.	Number
	Good for everyday short chats.	Number
	It difficult for the characters to understand their intentions and feelings when he/she reading a story.	Number
	He/she likes to play games (role plays) that need to be imitated with other children during pre-school education.	Number
	He/she can easily understand what they think and feel by just looking at their faces.	Number
	It is difficult to make new friendships.	Number

B. *K-fold cross- validation*

K-fold cross-validation is used because it can reduce computation time while maintaining the accuracy of estimation accuracy [26]. To improve the accuracy of each algorithm, the data will be divided into 9 k-fold so that it has 9 subsets of data for iteration with $k = 2$, $k = 3$, $k = 4$, ..., $k = 9$ and see the ideal k value in each algorithm.

IV. RESULT

This section presents the results obtained from experiments using the Matlab application. Table II shows the performance of each classification method where the decision tree provides the best performance among other methods with 100% accuracy in all predetermined k values. It is found that the number of iterations does not affect the performance of the decision tree method.

Linear Discriminant Analysis gives the best performance at $k = 7$, $k = 9$, and $k = 10$ with 96.9% accuracy. The lowest accuracy when the value $k = 2$ is 96.3%. The number of iterations greatly affects the performance of this method, the more iterations performed, the higher the accuracy of Linear Discriminant Analysis.

Logistic Regression provides the best accuracy of 99.7% at $k = 2$ and $k = 3$. The value of $k = 4$ to $k = 10$ gives the same accuracy of 99.6%. the more iterations are done, the lower the performance. This is influenced by the missing data in the testing process so that the data cannot be identified in the class 'having ASD' or 'not having ASD'.

Support Vector Machine provides the best accuracy of 99.9% at $k = 9$ and $k = 10$. The small number of iterations results in a performance that is not optimal, whereas a high number of iterations will improve the performance of the Support Vector Machine method. In this research, the use of the kernel on the support vector machine is not optimal because it still uses one kernel function, namely a linear kernel. In the future, the implementation of all kernel functions from support vector machine will be carried out.

K-Nearest Neighbors provides the best accuracy of 94.2% at $k = 8$. The resulting accuracy varies depending on the number of iterations done so it needs to select the right number of iterations to produce the best performance on K-Nearest Neighbors.

TABLE II. RESULT OF THE ALGORITHM FOR ALL K-FOLD VALIDATIONS

Method	Accuracy (%)								
	2 k-fold	*3 k-fold*	*4 k-fold*	*5 k-fold*	*6 k-fold*	*7 k-fold*	*8 k-fold*	*9 k-fold*	*10 k-fold*
Tree	100	100	100	100	100	100	100	100	100
LDA	96.3	96.7	96.7	96.7	96.6	96.9	96.7	96.9	96.9
LR	99.7	99.7	99.6	99.6	99.6	99.6	99.6	99.6	99.6
SVM	97.3	99.3	98.9	99.3	99.4	99.6	99.4	99.9	99.9
KNN	92.8	94	93	93.8	92.9	93.3	94.2	93.9	92.9

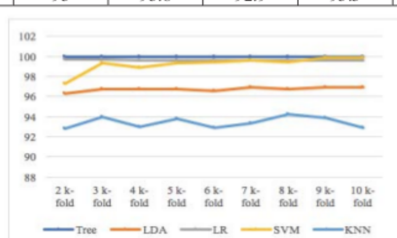

Fig. 2. Performance Algorithm for all *k-fold* validation

V. CONCLUSION

In this study, adult humans aged 17 years and over are used as sample datasets by asking various questions related to ASD to make a diagnosis using the classification method. Five algorithms are used in the classification method, namely Decision Tree, Linear Discriminant Analysis, Logistic Regression, Support Vector Machine, and K-Nearest Neighbors to see the level of accuracy and the application of cross-validation techniques to see the best k-fold values in each classification method. The Decision Tree gives an accuracy of 100% in each of the k values that have been determined previously. Linear Discriminant Analysis gives the best performance at $k = 7$, $k = 9$, and $k = 10$ with 96.9% accuracy. $k = 2$ and $k = 3$ with 99.7% accuracy for Logistic Regression. $k = 9$ and $k = 10$ with 99.9% accuracy for Support Vector Machine and $k = 8$ with an accuracy of 94.2% for K-Nearest Neighbors. All classification methods provide very satisfying performance, but Linear Discriminant Analysis and support vector machine are sensitive to small k values. Logistic regression has classification errors in the testing process, there are data that cannot be identified in class and K-Nearest Neighbors requires the selection of the number of iterations the right one to produce the best performance.

REFERENCES

[1] Bolton, P., Macdonald, H., Pickles, A., Rios, P. A., Goode, S., Crowson, M., ... & Rutter, M. "A case‐control family history study of autism." Journal of child Psychology and Psychiatry 35.5, 1994: 877-900.

[2] Wiggins, L. D., Reynolds, A., Rice, C. E., Moody, E. J., Bernal, P., Blaskey, L., ... & Levy, S. E. "Using standardized diagnostic instruments to classify children with autism in the study to explore early development." Journal of autism and developmental disorders 45.5, 2015: 1271-1280.

[3] American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub, 2013.

[4] Thabtah, Fadi. "Machine learning in autistic spectrum disorder behavioral research: A review and ways forward." Informatics for Health and Social Care, 2018: 1-20.

[5] Mythili, M. S., and AR Mohamed Shanavas. "A study on Autism spectrum disorders using classification techniques." IJSCE, ISSN, 2014: 2231-2307.

[6] Pancerz, K., Derkacz, A., Gomula, J., Ganzha, M., Maciaszek, L., & Paprzycki, M. "Consistency-Based Preprocessing for Classification of Data Coming from Evaluation Sheets of Subjects with ASDs." FedCSIS Position Papers. 2015.

[7] Duda, M., Ma, R., Haber, N., & Wall, D. P. "Use of machine learning for behavioral distinction of autism and ADHD." Translational psychiatry 6.2, 2017: e732.

[8] Jiao, Y., Chen, R., Ke, X., Chu, K., Lu, Z., & Herskovits, E. H. "Predictive models of autism spectrum disorder based on brain regional cortical thickness." Neuroimage 50.2, 2010: 589-599.

[9] Liu, Wenbo, Ming Li, and Li Yi. "Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework." Autism Research 9.8, 2016: 888-898.

[10] 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. ACM, 2017.

[11] Altay, Osman, and Mustafa Ulas. "Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and K-nearest neighbor in children." Digital Forensic and Security (ISDFS), 2018 6th International Symposium on. IEEE, 2018.

[12] Moreno-Torres, Jose G., Jose A. Saez, and Francisco Herrera. "Study on the impact of partition-induced dataset shift on k-fold cross-validation." IEEE Transactions on Neural Networks and Learning Systems 23.8, 2012: 1304-1312.

[13] Gorunescu, Florin. Data Mining: Concepts, models and techniques. Vol. 12. Springer Science & Business Media, 2011.

[14] Zhao, W., Rama Chellappa, and Nagaraj Nandhakumar. "Empirical performance analysis of linear discriminant classifiers." Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on. IEEE, 1998.

[15] Lawi, A., Wungo, S.L., Manjang, A. "Identifying irregularity electricity usage of customer behaviors using logistic regression and linear discriminant analysis." Proceeding of the 3rd International Conference on Science in Information Technology (ICSITech), pp. 552-557. IEEE, 2017.

[16] Agresti, Alan, and Maria Kateri. "Categorical data analysis." International encyclopedia of statistical science. Springer, Berlin, Heidelberg, 2011. 206-208.

[17] Burges, Christopher JC. "A tutorial on support vector machines for pattern recognition." Data mining and knowledge discovery 2.2, 1998: 121-167.

[18] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3, 1995: 273-297.

[19] Lawi, A., Aziz, F., Syarif, S. "Ensemble GradientBoost for increasing classification accuracy of credit scoring." Proceedings of the 4th International Conference on Computer Applications and Information Processing Technology (CAIPT), pp. 1-4. IEEE, 2017.

[20] Lawi, A., Velayaty, A.A., Zainuddin, Z. "On identifying potential direct marketing consumers using adaptive boosted support vector machine." Proceedings of the 4th International Conference on Computer Applications and Information Processing Technology (CAIPT), pp. 1-4. IEEE, 2017.

[21] Jafar, N., Thamrin, S.A., and Lawi, A. "Multiclass classification using Least Squares Support Vector Machine." Proceeding of the International Conference on Computational Intelligence and Cybernetics (CYBERNETICSCOM), , pp. 7-9. IEEE, 2016.

[22] Lawi, A., and Machrizzandi, M. S. "Facial Expression Recognition using Multiclass Ensemble Least-Square Support Vector Machine." Journal of Physics: Conference Series, vol. 979, no. 1, p. 012032. IOP Publishing, 2018.

[23] Lawi, A., and Adhitya, Y. "Classifying Physical Morphology of Cocoa Beans Digital Images using Multiclass Ensemble Least-Squares Support Vector Machine." Journal of Physics: Conference Series, vol. 979, no. 1, p. 012029. IOP Publishing, 2018.

[24] Kramer, Oliver. "K-nearest neighbors." Dimensionality reduction with unsupervised nearest neighbors. Springer, Berlin, Heidelberg, 2013. 13-23.

[25] Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." Ijcai. Vol. 14. No. 2. 1995.

[26] Yadav, Sanjay, and Sanyam Shukla. "Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification." Advanced Computing (IACC), 2016 IEEE 6th International Conference on. IEEE, 2016.

[27] Provost, Foster, and Ron Kohavi. "Guest editors' introduction: On applied research in machine learning." Machine learning 30.2, 1998: 127-132.

ication_Algorithms_of_the_Autism_Spectrum_Disorder_Diagnos..

ORIGINALITY REPORT

20%	14%	15%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University College London Student Paper	3%
2	Submitted to Udayana University Student Paper	1%
3	F Syukmana, E Wahyudi, W Gata, H Wahono, N I Febianto, A Y Kuntoro, R O Nitra, L Effendi, D D Saputra, O R Sulaeman. "Predicting Relegation Clubs in Italian Serie A with Method based C4.5 Decision Tree Algorithm", Journal of Physics: Conference Series, 2020 Publication	1%
4	www.frontiersin.org Internet Source	1%
5	Irawan Dwi Wahyono, Irham Fadlika, Khoirudin Asfani, Hari Putranto, Jehad Hammad, Sunarti. "New Adaptive Intelligence Method for Personalized Adaptive Laboratories", 2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE), 2019	1%

Publication

6	academicpublishingplatforms.com Internet Source	1 %
7	Bo Xing, Wen-Jing Gao. "Innovative Computational Intelligence: A Rough Guide to 134 Clever Algorithms", Springer Science and Business Media LLC, 2014 Publication	1 %
8	Hassan M.E. Azzazy, Robert H. Christenson. "Biochemical Markers of Acute Coronary Syndromes", Wiley, 2006 Publication	1 %
9	hdl.handle.net Internet Source	1 %
10	www.mdpi.com Internet Source	1 %
11	150.214.191.180 Internet Source	1 %
12	worldwidescience.org Internet Source	1 %
13	link.springer.com Internet Source	1 %
14	M A Muslim, A J Herowati, E Sugiharti, B Prasetiyo. "Application of the pessimistic pruning to increase the accuracy of C4.5	1 %

algorithm in diagnosing chronic kidney disease", Journal of Physics: Conference Series, 2018
Publication

15	insights.ovid.com Internet Source	1%
16	Kazi Shahrukh Omar, Prodipta Mondal, Nabila Shahnaz Khan, Md. Rezaul Karim Rizvi, Md Nazrul Islam. "A Machine Learning Approach to Predict Autism Spectrum Disorder", 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019 Publication	1%
17	uvadoc.uva.es Internet Source	<1%
18	Submitted to The Hong Kong Polytechnic University Student Paper	<1%
19	hildarhzamalia.blogspot.com Internet Source	<1%
20	pdfs.semanticscholar.org Internet Source	<1%
21	Osman Altay, Mustafa Ulas, Kursat Esat Alyamac. "Prediction of SFRSCC's T50 Performance Using Regression Tree Algorithm", 2019 1st International Informatics and Software	<1%

Engineering Conference (UBMYK), 2019
Publication

22	pt.scribd.com Internet Source	<1%
23	Siswanto, Abdussomad, Windu Gata, Nia Kusuma Wardhani, Grace Gata, Basuki Hari Prasetvo. "The Feasibility of Credit Using C4.5 Algorithm Based on Particle Swarm Optimization Prediction", 2019 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2019 Publication	<1%
24	Meilin Widyastuti, Agnes Gracella Fepdiani Simanjuntak, Dedy Hartama, Agus Perdana Windarto, Anjar Wanto. "Classification Model C.45 on Determining the Quality of Custumer Service in Bank BTN Pematangsiantar Branch", Journal of Physics: Conference Series, 2019 Publication	<1%
25	Submitted to Universitas Negeri Surabaya The State University of Surabaya Student Paper	<1%
26	tel.archives-ouvertes.fr Internet Source	<1%
27	www.pnas.org Internet Source	<1%

28	Aiguo Wang, Ning An, Guilin Chen, Lian Li, Gil Alterovitz. "Predicting hypertension without measurement: A non-invasive, questionnaire-based approach", Expert Systems with Applications, 2015 Publication	<1%
29	Submitted to University of Johannsburg Student Paper	<1%
30	Submitted to University of Nottingham Student Paper	<1%

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		