Accuracy, Recall, Precision of SVM Kernels in Predicting Autistic Spectrum Disorder In Adults

DidikSetiyadi, Muhammad Dwison Alizah, Yulius Paulus Dharsono, SabarSautomo, Sfenrianto

Abstract: ***Autism is a disorder that is quite difficult to diagnose when the condition of the sufferer is in the adult category. In this era, technology has been able to make predictions including health cases. Autistic Spectrum Disorder (ASD) in adults is felt to be predictable by using machine learning. This study will build a predictor for ASD sufferers. Predictors of machine learning are built using the Support Vector Machine (SVM) algorithm, with the type of kernel used was Gaussian RBF, Polynomial and Sigmoid. From the predictors that are built, the best SVM parameters will be searched based on accuracy. This best parameter is used to build the best new predictor and the results of the prediction are compared in terms of accuracy, recall, and precision. These results can be used to get the best performance when detecting ASD sufferers effectively and efficiently***

Keywords: Autism, Machine Learning, SVM, Kernel, Accuracy, Recall, Precision

I. INTRODUCTION

Autistic Spectrum Disorder (ASD) is a neurodevelopmental disorder that is very common. In general, ASD occurs in children. However, this does not rule out the possibility of this happening to adults. Early diagnosis of ASD is very meaningful for sufferer's life, family, relatives and other people to be more sensitive and have a high understanding of sufferers. Diagnosing ASD in adults is not as easy as ASD in children. Adults with ASD who are not diagnosed at a young age or children are more difficult to meet symptoms.This can make it difficult for medical staff to detect ASD[1].

Technological advances, especially in data mining, are often used to classify an object in the health or medical field. In this study, a classification algorithm will be applied to make predictions of ASD sufferers based on existing data sets. The collection of data used in the study came from data collected by Fadi Fayez Thabtah and published in the machine learning repository from UCI.

The algorithm used in this study is the Kernel Support Vector Machine (SVM). SVM has many types of kernels, but in this study the types of SVM kernels used are Radial Basis Function (RBF), Polynomial, and Sigmoid. The results of the SVM kernel algorithm modeling on the data collection of autistic patients in adults today will be compared in performance. The performance in question is the level of accuracy, recall, and precision of each modeling of the type of kernel that we use.

II. METHODOLOGY

A. Dataset

The data used in this study is "Autistic Spectrum Disorder Screening Data for Adult". This data is a collection of ten behavioral questions and ten characters from individuals with ASD who have proven effective in detecting ASD.

The data consists of 704 cases, 21 features of which the first feature is only indexing of data so that the features to be used in the learning model are only 20 features. ASD data has missing ones on age, ethnicity, and relationship features. The mechanism used to handle missing ones in it is using an average of all the values in the feature. This used to fill in the missing data with a value that does not differ significantly from the other data in the feature. Table I describes the features in the ASD data set.

Table I:Name of the Table that justify the values [7]

Attribute	Type	Description
Age	Number	Age in years
Gender	String	Male or Female
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean (yes or no)	Whether the case was born with jaundice
Family member with PDD	Boolean (yes or no)	Whether any immediate family member has a PDD
Who is completing the test	String	Parent, self, caregiver, medical staff, clinician ,etc.
Country of residence	String	List of countries in text format
Used the screening app before	Boolean (yes or no)	Whether the user has used a screening app
Screening Method Type	Integer (0,1,2,3)	The type of screening methods chosen based on age category (0=toddler, 1=child, 2= adolescent, 3= adult)

Revised Manuscript Received on March 15, 2020.

***** Correspondence Author

DidikSetiyadi,Department of Informatics, Bina Insani University, Indonesia,Email: didiksetiyadi@binainsani.ac.id.

Muhammad Dwison Alizah*, Departement of Computer Science, STMIK Nusa Mandiri, Jakarta, Indonesia. Email: 14002299@nusamandiri.ac.id.

Yulius Paulus Dharsono, Departement of Computer Science, STMIK Nusa Mandiri, Jakarta, Indonesia. Email: 14002320@nusamandiri.ac.id

SabarSautomo, Departement of Computer Science, STMIK Nusa Mandiri, Jakarta, Indonesia. Email: 14002304@nusamandiri.ac.id

Sfenrianto, Information Systems Management Department, BINUS Graduate Program – Master of Information Systems Management, Bina Nusantara University, Jakarta 11480, Indonesia. E-mail:sfenrianto@binus.edu.

Question 1 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 2 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 3 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 4 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 5 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 6 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 7 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 8 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 9 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 10 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Screening Score	Integer	The final score obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner

B. SVMKernel

ASD data were trained and made learning models using the SVM Kernel classification method. SVM is a machine learning method used for classification problems of two different groups / labels [2]. SVM is also able to classify nonlinear functionwith a method called the kernel. This SVM kernel puts input into a higher dimension so that it is able to classify the inputs in a particular kernel. The SVM kernel function consists of linear, nonlinear, polynomial, radians basis functions, and sigmoid [3]. In this study the kernel functions used and will be compared to accuracy, recall and precision are Radial Basis Function (RBF), Polynomial, and Sigmoid.

C. Accuracy, Precision, Recall

Accuracy is the comparison of all correct data to the whole data. Precision is the comparison of correct data to the amount of data query results [4]. Recall is a comparison of the amount of correct data taken against the amount of relevant data [4]. Here is a picture of the confulision matrix and performance. Recall and precision are two of the measurement methods for evaluating performance in the most commonly used information retrieval [5].

D. K-Fold Cross Validation

One of the things that need to be avoided from making a machine learning model is overfitting. This needs to be avoided so that the modeling we build is able to predict new data sets that have not been encountered before by the machine learning model that was built.

There are several techniques that are considered capable of helping to avoid this. Hold-out Validation and K-Fold Cross Validation are some examples of the methods in question. In this research, K-Fold Cross Validation is used to overcome overfitting.

K-Fold Cross Validation is a method used to reduce overfitting of a prediction. K-Fold Cross Validation divides training data and test data as many as K equally with the percentage of training data and test data. The classifier will be built K times according to each part of the fold in each K. The performance of each result of all accuracy is calculated on average [6].

III. IMPLEMENTATION: PERFORMANCE TEST

The implementation of performance tests for the predictor on this project is divided into several steps. These steps are preprocessing data, building classifiers, and testing classifiers and evaluations.Where each stepwill be discussed in the following explanation.

A. Data Preprocessing

At the data preprocessing stage, we separate the ASD data set between the label and its features. From the results of the data separation, we divide the data into 70% of training data and 30% of test data.

Table II: Separation of Train Data and Test Data (Features)

Name	Type	Size
X	object	(704,20)
X_Test	object	(212,20)
X_Train	object	(492,20)

Table III: Separation of Train Data and Test Data (Label)

Name	Type	Size
Y	object	(704,20)
Y_Test	object	(212,20)
Y_Train	object	(492,20)

Features in ASD data are not entirely numeric data. There are several features such as gender, ethnic, country and others which are not data numbers. We change the data on these features using the encoder label. The encoder label is used so that the machine learning model is able to process data from these features.

In the data used there are missing data on age, ethnic and relationship features. Missing data on ASD data will be filled with values from all data from each feature. Missing Data in question is outlined in the following figure. Selection of averages as a result of missing data so that what is provided has a relatively similar scale of values.

B. Building Classifiers

Furthermore, from the data that has been prepared, a classifier of training data is built, the results of which will be used as a model for predicting test data. From the classifier that has been built then used to predict 30% of the test data and see its performance. The classifier used is SVM with the kernel, namely: RBF, Polynomial, and Sigmoid.

At the beginning of making the classifier, it is done with default parameters as the beginning of classifier development and later it will be evaluatedand classifier rebuilt with parameters in accordance with the results of the evaluation. This is done to get more optimal results from the default parameters in the Scikit-Learn module. The default parameters are explained in the following table.

Table IV: SVC Module Default Parameter Value

No	Parameter	Value (Default)	Description
1	C	1.0	-
2	degree	3.0	Kernel = Poly
3	gamma	scale	-
4	coef0	0.0	Kernel = Poly & Sigmoid

C. Validation and Evaluation

The next step is validation and evaluation. In the validation stage, validation is performed on the performance of the built classifier. This validation is used to avoid overfitting. Validation is done by K-fold cross validation. K-fold cross validation is done with K = 20. The results of the performance of the data validation are shown the average performance and standard deviation.

Table V: K-Fold Cross Validation Results

Kernel	Averageperformance	Standarddeviation
RBF	94,81%	0,025
Polynomial	93,88%	0,0408
Sigmoid	90,20%	0,025

Afterwards, an evaluation using grid search is performed to find the most appropriate parameters of each parameter per SVM kernel that we use. The results of the evaluation are the highest accuracy, recall, precision of each kernel and the best parameters used. The following is a table of parameters used to do Grid Search.

Table VI: The parameters used for evaluation using Grid Search

Kernel	C	Gamma
RBF	1 - 10	0.1,0.01,0.02,0.03,0.04,0.05,0.06,0.07,0.08
Polynomial	1 - 10	0.5,0.6,0.7,0.8
Sigmoid	1 - 10	0.1,0.01,0.001,0.0001,0.00001

The results of the evaluation using Grid Search with the parameters above are the best parameters that can be used for optimal results (highest accuracy). In the Polynomial kernel, it produces over fitdata so parameters need to be managed in the kernel. The following is a parameter table that is used to get the optimal value. As stated in Table III.

Table VII: The Best Parameter Results

Kernel	C	Gamma	Degree	Coef0
RBF	9	0,02	-	-
Polynomial	1	0,1	5	1
Sigmoid	8	0,0001	-	-

IV. RESULT AND CONCLUSION

A. Result

The results of the construction of the Machine Learning model for predicting ASD sufferers are explained in the following tables. This model was built using the parameters found to be the most optimal as the results in Table V.

1) Kernel RBF

The following are the results obtained from the Machine Learning model with the 'RBF' kernel. The results are outlined in the Confusion Matrix table and the Classification Report table.

Table VIII: Confusion Matrix in the RBF kernel

		Prediction	
		0	1
RealValue	0	150	5
	1	6	51

Table IX:Classification Report in the RBFkernel

	precision	recall	f1-score	support
0	0,96	0,97	0,96	155
1	0,91	0,89	0,9	57
accuracy			0,95	212
macro avg	0,94	0,93	0,93	212
weighted avg	0,95	0,95	0,95	212

2) Kernel Polynomial

Here are the results obtained from the Machine Learning model with the 'Polynomial' kernel. The results are outlined in the Confusion Matrix table and the Classification Report table.

Table X: Confusion Matrix in the Polynomial kernel

		Prediction	
		0	1
Real Value	0	147	8
	1	10	47

Table XI: Classification Report in the Polynomial kernel

	precision	recall	f1-score	support
0	0,94	0,95	0,94	155
1	0,85	0,82	0,84	57
accuracy			0,92	212
macro avg	0,9	0,89	0,89	212
weighted avg	0,91	0,92	0,91	212

3) Kernel Sigmoid

Here are the results obtained from the Machine Learning model with the 'Sigmoid' kernel. The results are outlined in the Confusion Matrix table and the Classification Report table.

Table XII: Confusion Matrix in the Sigmoid kernel

		Prediction	
		0	1
Real Value	0	152	3
	1	24	33

Table XIII: Classification Report in the Sigmoid kernel

	precision	recall	f1-score	support
0	0,86	0,98	0,92	155
1	0,92	0,58	0,71	57
accuracy			0,87	212
macro avg	0,89	0,78	0,81	212
weighted avg	0,88	0,87	0,86	212

V. CONCLUSION

Building a predictorit's about how well the performance of the predictor is built.From several algorithms used, it was found that the RBF kernel has the best performance, after that the polynomial kernel and sigmoid kernel. This performance can be seen from the results of performance validation using K-Fold Cross Validation as shown in Table 3, the results of the Confusion Matrix and Classification Report of each kernel.

REFERENCES

1. "Data & Statistics on Autism Spectrum Disorder." Centers for Disease Control and Prevention. Ed. Center for Disease Control. Centers for Disease Control and Prevention. Accessed 12 Dec. 2019.
2. Cortes, Corinna; Vapnik, Vladimir N. (1995). "Support-vector networks" (PDF). Machine Learning. 20 (3): 273–297. CiteSeerX 10.1.1.15.9362. doi:10.1007/BF00994018.
3. Talabani, H., &Avci, E. (2018). Performance Comparison of SVM Kernel Types on Child Autism Disease Database. 2018 International Conference on Artificial Intelligence and Data Processing (IDAP).
4. Zhang, P., & Su, W. (2012). Statistical inference on recall, precision and average precision under random selection. 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery.
5. Thabtah, F. (2017). Machine Learning in Autistic Spectrum Disorder Behavioural Research: A Review. To Appear in Informatics for Health and Social Care Journal. December, 2017 (in press).
6. Chicco D, Jurman G (January 2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation".
7. Thabtah, F. (2017). ASDTests. A mobile app for ASD screening. www.asdtests.com [accessed December 20th, 2017].
8. Thabtah, F. (2017). Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment. Proceedings of the 1st International Conference on Medical and Health Informatics 2017, pp.1-6. Taichung City, Taiwan, ACM.
9. Ramtekkar, U. P. (2017). DSM-5 Changes in attention deficit hyperactivity disorder and autism spectrum disorder: Implications for comorbid sleep issues.
10. Doshi-Velez, F., et al. (2014). Comorbidity clusters in autism spectrum disorders: An electronic health record time-series analysis
11. Bressan, P. (2018). Systemisers are better at maths
12. American Psychiatric Association. (2016). Children diagnosed with autism at earlier age more likely to receive evidence-based treatments [Press release].

AUTHORS PROFILE

DidikSetiyadiDepartment of Informatics, Bina Insani University, Indonesia,Email: didiksetiyadi@binainsani.ac.id. With lecturing subject:database system, information system, decission support system and Computer Science

Muhammad Dwison Alizah Departement of Computer Science, STMIK Nusa Mandiri, Jakarta, Indonesia. Email: 14002299@nusamandiri.ac.id

Yulius Paulus Dharsono Departement of Computer Science, STMIK Nusa Mandiri, Jakarta, Indonesia. Email: 14002320@nusamandiri.ac.id

Sabar Sautomo Departement of Computer Science, STMIK Nusa Mandiri, Jakarta, Indonesia. Email: 14002304@nusamandiri.ac.id

Dr. Sfenrianto, S,Kom, M.Kom is a Faculty Member of the Information Systems Management Department, BINUS Graduate Program – Master of Information Systems Management, Bina Nusantara University, Jakarta, Indonesia. (e-mail: sfenrianto@binus.edu). With lecturing subject: Digital Business and E-Commerce Management. Research interest in Digital Business, e-Commerce, business intelligence, E-Learning and Information System.