Journal of Student Research Exploration

https://shmpublisher.com/index.php/josre
p-ISSN 2964-8246| e-ISSN 2964-1691

Support vector machine on two-class classification problem to determine an otaku

Farhan Husyen Ramadhan[1], Apri Dwi Lestari[2]
[1, 2] Department of Computer Science, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received April 20, 2024
Revised January 28, 2025
Accepted February 8, 2025

Keywords:

Support vector machine
Classification
Otaku
Anime

ABSTRACT

Machine Learning has become a popular topic among academics and practitioners in recent years. This paper describes the use of SVM for otaku classification problem. The dataset used is a dummy dataset created with a python programme. In this research, SVM will be used as a model. The model aims to predict whether someone is an otaku or not, based on several attributes. The optimal parameters are obtained after several experiments. The parameters consist of kernel='poly', C=0.1, gamma='auto', degree=2, and attribute class_weight=None. The performance obtained by applying the above parameters is 100% accuracy.

1. Introduction

Machine Learning has become a popular topic among academics and practitioners in recent years. Machine Learning (ML) techniques have been rapidly developed and successfully applied in many areas of civil engineering [1], [2]. Machine Learning is one of the methods to find solutions to existing problems. One of them is the classification problem. Classification is the process of grouping samples based on similar attributes by utilising labels as categories [3]. One of the popular machine learning models that can be used for classification problems is SVM.

[1] ***Corresponding Author:***
Farhan Husyen Ramadhan,
Department of Computer Science,
Universitas Negeri Semarang,
Sekaran, Gunung Pati, Kota Semarang, Indonesia.
Email: farhanhr00@students.unnes.ac.id
DOI: https://doi.org/10.52465/josre.v3i1.358

Support Vector Machine (SVM) is one of the machine learning models introduced by Vapink. This model aims to find the best hyperplane that separates two classes in the input space. The accuracy produced by SVM depends on the kernel function and parameters used in the model [4]. SVM can be used for classification, regression, and outlier detection problems. In this article, the SVM model will be used for classification on the dataset. The model is created with the aim of predicting whether an otaku or not.

Otaku is a term that refers to a person who has an interest in anime, manga, and related activities [5]. Otaku is a youth subculture consisting of collectors who are fond of a special lifestyle and obsessed with anime products [6]. Otaku will be used as a label or category that will be determined based on the attributes in the dataset. So that in this classification there are two classes, namely an otaku and someone who is not an otaku.

2. Method

The stages of the experiment start from data collection and dataset creation. The stages can be seen in figure 1. The stages consist of data collection and data set creation. Data preprocessing, splitting data into train data and test data, model building, model training, and finally model evaluation.

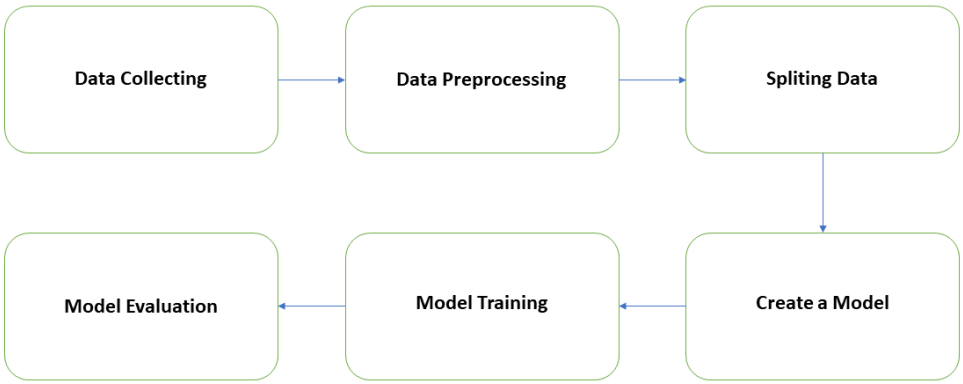


Figure 1. Stages of machine learning modelling

Data Collecting

In this experiment, dummy data is used for the dataset so that the data collecting stage is the stage of creating dummy data. The dummy data is created using python and produces 200 rows of data consisting of 4 columns. One column named 'Otaku' is a label that has values 0 and 1. The value 0 means 'Not Otaku' and the value 1 is 'Otaku' so there are two classes for this classification model. The first five rows of the dataset can be seen in Table 1.

Table 1. Details of otaku dataset

	Jumlah_Anime_yang_ditonton	Baca_Manga	Pergi_ke_Event	Otaku
0	102	1	Null	1
1	179	1	Null	1
2	92	1	1	1
3	14	0	Null	0
4	106	1	1	1

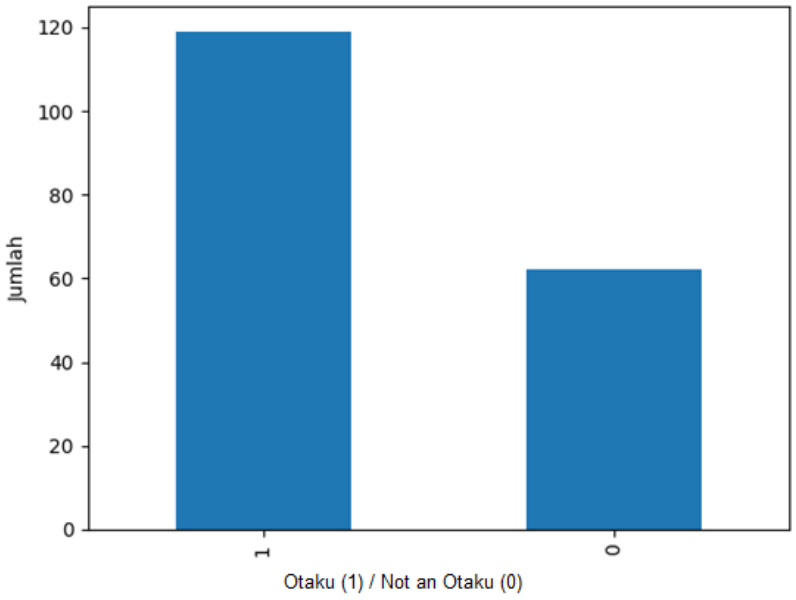
Data Preprocessing

Data preprocessing has become an important technique in today's knowledge discovery scenario, which is dominated by increasingly large data sets [7]. In the data preprocessing stage, one of the things that is done is handling missing values. As can be seen in Table 1, there are several rows that have null values. The null value needs to be removed by deleting the row. The results of handling missing values can be seen in Table 2.

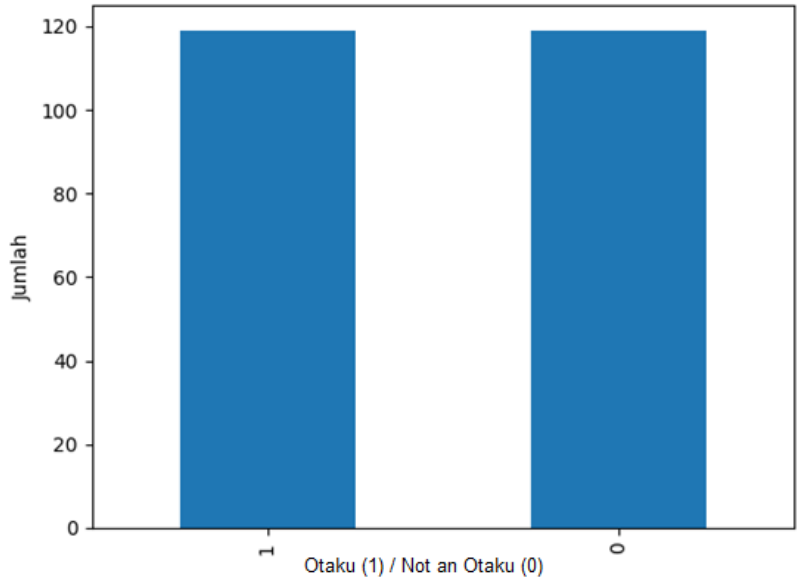
Tabel 2. Otaku dataset details after missing value handling

	Jumlah_Anime_yang_ditonton	Baca_Manga	Pergi_ke_Event	Otaku
2	92	1	1	1
4	106	1	1	1
5	71	0	0	0
6	188	1	1	1
7	20	0	0	0

In addition to missing values, data imbalance also needs to be addressed if the data in the otaku class and the non-otaku class are not balanced. In practical applications, the ratio of small classes to large classes can be as drastic as 1:10, while some fraud detection applications report imbalances of 1:100,000 [8]. Resampling is an effective approach to handling imbalanced data sets that aims to equalise the number of category samples [9]. One way of handling imbalance data can be done by using the oversampling method. Oversampling is a method of handling data imbalance by multiplying data from the minority class so that data from that class can be equivalent to data from the majority class.



(a)



(b)

Figure 2. Comparison of the number of data in (a) imbalance data before oversampling and (b) balanced data after oversampling

Splitting Data

Data splitting is a stage that divides the dataset into two, namely data for training and data for testing. Regarding data splitting, data samples are often divided into two sets of data, including the training set for model training and the testing set for model validation [10]. The division of data for training and testing is important for obtaining training models and test models. Many researchers propose a ratio of

70/30 or 80/20 (training/testing set) to generate datasets in machine learning [11]–[13]. In this research, 80:20 ratio is used for data splitting.

Model Creation, Training, and Evaluation

It is important to determine a suitable model for the problem to get good prediction results. In this research, SVM will be used as a model. SVM works by creating a decision boundary that can separate two classes so it is very suitable in the case of Otaku classification which has two classes.

The parameters used for the first time are the default parameters provided by Sklearn SVC. After the model is trained, an evaluation will be conducted to assess the performance of the model. Further experiments and evaluations will continue with the available parameters until an optimal result is found.

3. Results and Discussion

In this experiment, the trained model will be evaluated for performance. The performance of the model will be evaluated based on the accuracy results of the test. The model will then be retrained with different parameters. Parameters that will be the focus and re-evaluated kernel parameters, C, degree, gamma, and class_weight attributes.

In the first experiment with the default parameters provided by SKlearn SVC, namely, kernel = 'rbf', C = 1.0, degree = 3, gamma = 'scale'. The performance of the model after testing resulted in an accuracy of 77.083%. This result is a pretty good result for the first experiment. Experiments with different parameters continue until getting optimal results.

The optimal parameters were obtained after several experiments. The parameters consist of kernel='poly', C=0.1, gamma='auto', degree=2, and class_weight=None attributes. The performance obtained by applying the above parameters is 100% accuracy. This accuracy is the most optimal result that can be obtained in this classification.

4. Conclusion

The SVM model created successfully predicts someone who is an otaku or not an otaku. With the right parameters, the accuracy of the model gets a score of up to 100%. These results are the most optimal results that can be obtained in this problem. Thus, it is concluded that the SVM model is very suitable for otaku classification problems. In further research it is recommended to use original data to get more correct results.

REFERENCES

[1] H.-B. Ly, T.-T. Le, H.-L. T. Vu, V. Q. Tran, L. M. Le, and B. T. Pham, "Computational Hybrid Machine Learning Based Prediction of Shear Capacity for Steel Fiber Reinforced Concrete Beams," *Sustainability*, vol. 12, no. 7, p. 2709, Mar. 2020, doi: 10.3390/su12072709.

[2] D. Van Dao *et al.*, "A Sensitivity and Robustness Analysis of GPR and ANN for High-Performance Concrete Compressive Strength Prediction Using a Monte Carlo Simulation," *Sustainability*, vol. 12, no. 3, p. 830, Jan. 2020, doi: 10.3390/su12030830.

[3] D. Dalbergio, M. N. Hayati, and Y. N. Nasution, "KLASIFIKASI LAMA STUDI MAHASISWA MENGGUNAKAN METODE ALGORITMA C5.0 PADA STUDI KASUS DATA KELULUSAN MAHASISWA FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM UNIVERSITAS MULAWARMAN TAHUN 2017," *Pros. Semin. Nas. Mat. dan Stat. Vol 1 Pros. Semin. Nas. Mat. dan Stat.*, May 2019, [Online]. Available: https://jurnal.fmipa.unmul.ac.id/index.php/SNMSA/article/view/524

[4] I. M. Parapat, M. T. Furqon, and S. Sutrisno, "Penerapan Metode Support Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 10 SE-, pp. 3163–3169, Feb. 2018, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/2577

[5] S. Irawan, H. Antono, and Y. Windrawanto, "DAMPAK POSITIF OTAKU ANIME TERHADAP PERILAKU MAHASISWA," *J. KONSELING GUSJIGANG*, vol. 8, no. 1, Aug. 2022, doi: 10.24176/jkg.v8i1.7826.

[6] H. Niu, Y. Chiang, and H. Tsai, "An Exploratory Study of the Otaku Adolescent Consumer," *Psychol. Mark.*, vol. 29, no. 10, pp. 712–725, Oct. 2012, doi: 10.1002/mar.20558.

[7] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak, and F. Herrera, "A survey on data preprocessing for data stream mining: Current status and future directions," *Neurocomputing*, vol. 239, pp. 39–57, May 2017, doi: 10.1016/j.neucom.2017.01.078.

[8] A. Aprihartha, "Penyelesaian Masalah Ketidakseimbangan Data Melalui Teknik Oversampling dan Undersampling pada Klasifikasi Siswa Tidak Naik Kelas," *J. Tek. Ibnu SIna*, vol. 9, no. 1, 2024, doi: https://doi.org/10.36352/jt-ibsi.v9i01.807.

[9] S. Guan, X. Zhao, Y. Xue, and H. Pan, "AWGAN: An adaptive weighting GAN approach for oversampling imbalanced datasets," *Inf. Sci. (Ny).*, vol. 663, p. 120311, Mar. 2024, doi: 10.1016/j.ins.2024.120311.

[10] Q. H. Nguyen *et al.*, "Influence of data splitting on performance of machine learning models in prediction of shear strength of soil," *Math. Probl. Eng.*, vol. 2021, pp. 1–15, 2021.

[11] W. Chen *et al.*, "Landslide susceptibility modelling using GIS-based machine learning techniques for Chongren County, Jiangxi Province, China," *Sci. Total Environ.*, vol. 626, pp. 1121–1135, Jun. 2018, doi: 10.1016/j.scitotenv.2018.01.124.

[12] K. Taalab, T. Cheng, and Y. Zhang, "Mapping landslide susceptibility and types using Random Forest," *Big Earth Data*, vol. 2, no. 2, pp. 159–178, Apr. 2018, doi: 10.1080/20964471.2018.1472392.

[13] N. N. Vasu and S.-R. Lee, "A hybrid feature selection algorithm integrating an extreme learning machine for landslide susceptibility modeling of Mt. Woomyeon, South Korea," *Geomorphology*, vol. 263, pp. 50–70, Jun. 2016, doi: 10.1016/j.geomorph.2016.03.023.