Weight of Evidence and Information Value on Support Vector Machine Classifier

M Dika Saputra[1](✉), Zahroatul Fitria[1], Bagus Sartono[2], Evi Ramadhani[3], and Alfian Futuhul Hadi[1]

[1] Department of Postgraduate Mathematics, University of Jember, Jember, Indonesia
m.dikasaputra963@gmail.com, afhadi@unej.ac.id
[2] Department of Statistics and Data Science , IPB University, Bogor, Indonesia
bagusco@apps.ipb.ac.id
[3] Department of Statistics, Syiah Kuala University, Aceh, Indonesia
evi.ramadhani@unsyiah.ac.id

Abstract. In building a classification model, variables containing low predictive information are sometimes used. This can increase the bias on classification. Weight of Evidence (WoE) and Information Value (IV) provide a good theoretical foundation to explore, filtering, and transforming variables in binary classification. The value of IV can help measure the predictive power possessed by a variable in separating binary classes. This research implements this framework to screen 24 predictor variables that will be used in the svm classification model to improve the evaluation of the food insecure household classification model. We use the National Socioeconomic Survey by the Indonesian Central Bureau of Statistics in 2020 for West Java Province and 2021 for East Java Province to produce a classification model. The results of this study showed that WOE was able to improve the model evaluation value from the AUC value of 0.81 to 0.83 for West Java Province and the AUC value of 0.58 to 0.66 for East Java Province.

Keywords: weight of evidence · information value · feature selection · classification · machine learning

1 Introduction

Big data has recently gained a lot of interest among data scientists. In addition to data that has a large size, another feature is the diverse form of data and high speed. This causes classical analysis such as linear regression, etc. to be unable to solve big data problems properly. Machine learning is one method that has been widely used in analyzing big data [1]. Big data problems can be solved with models found in machine learning. One of the important big data problems that needs immediate attention is classification [2]. Classification is a process to find a model that can distinguish between data classes, with the aim that the model obtained is useful for predicting unknown classes of observed objects. Support vector machine (svm) is a classification model in machine learning. Research conducted by Phangtriastu, et al. (2017) compared directed machine learning models between neural networks and svm [4]. Based on this research, the accuracy value of the svm model is better than the neural network model.

I. H. Agustin (Ed.): ICONNSMAL 2022, AISR 177, pp. 113–124, 2023.
https://doi.org/10.2991/978-94-6463-174-6_11

Basically, the svm model is very strong in data classification problems.. Despite having a good theoretical basis and high classification accuracy, svm is usually not suitable for classification on big data, because the complexity of svm training is highly dependent on the dataset size. In addition, in the analysis of directed machine learning classification, low prediction information on observation features is of particular concern. This can increase bias in classification, because low prediction information in features is still included. The big data problem in this research can be overcome one of them by weighting category data using the WoE and IV methods on prediction features [6]. WoE and IV are derived from the same logistic regression technique. These terms serve as benchmarks for sorting out variables in risk modeling such as default probability.

Related to the description that has been explained, in this research we will use food insecurity data as a simulation. The food insecurity data used comes from the national socioeconomic survey of East Java and West Java provinces. Each data amounted to 24792 sample households for East Java province and 24769 sample households for West Java province. Observation features in the food insecurity data for both East Java province and West Java province amounted to 24 features of food insecurity characteristics. The objectives of this research are (i) WoE and IV methods are able to improve the evaluation model of big data classification, (ii) provide information about the predictive power of each feature.

2 Method

2.1 Machine Learning

Big data refers to datasets that usually include both many observations and many variables, making the use of traditional statistical methods difficult [7]. Thus, a better model than the classical statistical analysis model is needed. In general, big data has characteristics such as data volume, variation, and velocity. Big data are also often less structured than traditionally collected data [8]. The development of computing devices, models and data storage caused big data to grow rapidly.

Machine learning is a subfield of artificial intelligence that has various approaches [9] by giving advantages of computers to know tasks without being explicitly programmed. Machine learning algorithms are mathematical model mapping methods used to learn the underlying patterns embedded in data [3]. Machine learning has two categories including unsupervised and supervised approaches. In a supervised approach, a dataset of samples that have data classes is used by the learning algorithm to learn patterns in the explanatory features. The trained model is then applied to make predictions on new data [10].

2.2 Support Vector Machine

In processing huge visual information in big data such as ImageNet, it is necessary to use powerful methods to address the problem. In computer science, machine learning is designed to achieve the following goals: Machine learning has become a popular field of research in recent years [16]. Svm is one of the most popular linear machine

learning classifiers with some attractive properties [17]. Svm is a supervised machine learning model. Based on SVM optimization theory using linear function hypothesis in high dimensional features.. The level of accuracy in the svm model is highly dependent on the kernel functions and parameters used [5]. Linear and non-linear SVMs are svm models when viewed from their characteristics. Linear SVM uses a soft margin on the hyperplane and is linearly separated. While non-linear SVM implements data in a high-dimensional space against the kernel function.

The goal of svm is to find the best hyperplane by separating two classes with maximum margin and can be seen in Fig. 1. Below illustrates the separation of classes, namely class + 1 and class -1, this separation uses the svm model. $x_i \in \mathbb{R}^D$ is the notation of the data. While each class is denoted $y_i \in \{+1, -1\}$ for $i = 1, 2, \ldots, n$ and i is the number of data.

svm model using kernel function k, can be seen in the following equation:

$$f(x) = \sum_{i=1}^{n} a_i k(x_i, x) + b$$

where the following function is minimized using the coefficients a_i and b:

$$\sum_{i,j=1}^{n} a_i a_j k\left(x_i, x_j\right) + C \sum_{i=1}^{n} \varsigma_i$$

Subject to

$$y_i f(x_i) \geq 1 - \varsigma_i$$

where ς_i measures the misclassification rate of x_i and Cost is the misclassification penalty parameter. The function $f(x)$ maps the training data vectors x to a higher dimensional space. Based on the function $f(x)$, svm determines a linear hyperplane that separates the training samples by maximizing the margin in the higher dimensional space.

The kernel function is used to modify the SVM. This is to solve non-linear problems. In non-linear svm, the data $\overrightarrow{x}$ is mapped by the function $\phi(\overrightarrow{x})$ into a higher dimensional

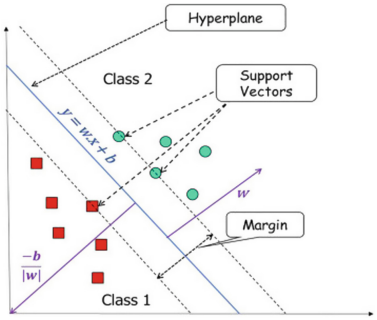

Fig. 1. Classification of svm model (Source: Cognitive Data Science in Sustainable Computing) [18]

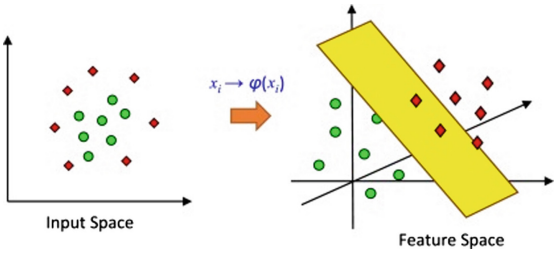

Fig. 2. Non-linear svm on higher dimensional vector spaces (Source: Handbook of Neural Computation) [19]

vector space. In the vector space, a hyperplane is used to separate two classes. An illustration can be seen in Fig. 2.

In non-linear svm, the performance is measured with four types of kernels: polynomial, linear, sigmoid, and linear [17].

2.3 Weight of Evidence and Information Value

Based on Information Theory conceived in the later 1940s and initially developed for scorecard development, WoE and IV have been gaining increasing attention in recent years for such uses as segmentation and variable reduction [11]. This method of analysis is usually simple and comparatively consumes less time [12]. WoE works by recoding variable values into discrete categories and assigning a unique WoE value to each category with the aim of generating the largest difference between the recoded ones. An important assumption here is that the dependent variable must be binary to indicate the occurrence and non-occurrence of an event. In the example of food insecurity analysis where households are neither food insecure (good) nor food insecure (bad), the WoE for each household segment is calculated as follows.

$$WoE = \left[ln\left(\frac{\%bad_i}{\%good_i} \right) \right] \times 100$$

While WoE analyzes the predictive ability of a variable in relation to its targeted outcome, IV assesses the overall predictive ability of the variables that have been used. IV can be used to compare the predictive ability among competing variables. The following is the calculation of IV.

$$IV = \sum_{i=1}^{n} \left((\%bad_i - \%good_i) \times \left(\frac{\%bad_i}{\%good_i} \right) \right)$$

2.4 Model Evaluation

Classification models are expected to produce correct classification of all data, but it cannot be denied that the performance of a model can provide accurate results. Model evaluation can be done by calculating the confusion matrix. The confusion matrix is a

Table 1. Commonly Used Kernels in SVM

Kernel Function	**Definition**
Linear	$K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i . \vec{x}_j)$
Polynomial	$K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i . \vec{x}_j + 1)^D$
Gaussian	$K(\vec{x}_i, \vec{x}_j) = \exp\left(-\frac{\|\vec{x}_i - \vec{x}_j\|}{2\sigma^2}\right)$
Sigmoid	$K(\vec{x}_i, \vec{x}_j) = tanh(\alpha \vec{x}_i . \vec{x}_j + \beta)$

Table 2. The Confusion Matrix

	Actual Positive (AP)	***Actual Negative (AN)***
Predicted Positive (PP)	True Positive (TP)	False Positive (FP)
Predicted Negative (PN)	False Negative (FN)	True Negative (TN)

cross-tabulation between the response feature data included in the prediction class and the actual [14], as shown in Table 2 and (Table 1).

Based on Table 2., the accuracy, sensitivity values can be obtained as follows:

Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

Sensitivity

$$Sensitivity = \frac{TP}{TP + FN}$$

Model performance measurement can also be done by using the area under the curve (auc). The auc value is obtained by calculating the area under the roc curve from 0 to 1 [15]. The greater the auc value, the better the performance of the classification model.

3 Design and Experiment

3.1 Big Data Source

This research uses food insecurity data of East Java province and West Java province as simulation. The establishment of food insecurity data is taken from the susenas of East Java province which includes 24792 sample households and 24679 sample households for West Java province. The food security module used in this survey is the food insecurity experience scale (FIES). The FIES serves to measure the level of food insecurity of households or individuals. This measurement uses yes or no answers to 8 questions about the respondent. The level of food insecurity in this study consists of not food insecure (y = 0) and food insecure (y = 1) [20]. The food insecurity predictor features used in this research are listed in Table 3. and (Table 4.).

Table 3. Predictor Features

Features Name	Scale
Head of household education	Nominal
Vulnerable Household Head	Ordinal
Number of savin	Ordinal
Land assets	Nominal
Floor Size	Ordinal
Roof Types	Nominal
Floor Types	Nominal
Wall Types	Nominal
Transferee	Nominal
Proper Drinking Water	Nominal
Grantee of Health Insurance Local Program	Nominal
Grantee of Non Cash Social Assistance	Ordinal
Grantee of Hopeful Family Program	Ordinal
Internet Access	Nominal
Grantee of Scholarship Social Program	Nominal
Number of Illiterate	Ordinal
Grantee of Health Insurance National Program	Nominal
Sick but not Outpatient	Nominal
Grantee of Social Assistance From Local Government	Ordinal
Proper sanitation	Nominal
Cooking Fuel	Nominal
Drinking Water Source	Nominal
Grantee of Prosperous Family Program	Ordinal
Electricity	Nominal

3.2 Procedure Analysis

In general, the research process includes collecting data, preprocessing data, transforming data and selecting (WoE and IV), building classification models, evaluating models, and comparing the results of evaluation models between the transformed WoE and IV data and the original dataset. This process is described in a flowchart, as shown in Fig. 3.

The analysis phase began with data collection. Data was collected from the Central Statistics Agency (BPS) regarding the National Socio-Economic Survey. In the preprocessing stage, data were prepared by aggregating individual data to the household level. Next, missing observation values, “no answer” codes or “don’t know” codes for the 8 susenas FIES questions were removed.. Assigning food insecurity classes from “not food insecure” to “food insecure”. Furthermore, the finished data is divided into two:

Table 4. WOE in East Java and Barat Java Provinces

Head of household education	**WOE**	
	East Java	**West Java**
Elementary school	0,02	0,47
Didn't finish elementary school	0,64	0,19
Junior High School	-0,22	-0,21
Senior High School	-0,29	-0,51
College	-1,29	-2,32
Cooking Fuel		
Firewood	0,20	0,47
Liquefied Petroleum Gas 3 kg	0,03	0,06
No cooking	-1,34	-1,00
Kerosene	-0,37	-1,02
Other	-1,64	-1,54
Liquefied Petroleum Gas 12kg/5,5kg/Bluegaz	-2,69	-3,12
Number of savings		
0–1	0,24	0,24
2–10	-0,83	-0,73
Floor Area		
4–30	0,83	0,37
31–48	0,46	0,35
49–52	0,18	0,30
53–60	0,08	0,01
61–77	-0,04	-0,06
78–100	-0,46	-0,35
101–882	-0,73	-0,66

70% training data and 30% test data. Calculate the WoE and IV values on the training data. Transformed data is perform by using WoE values on both training and test data. Transformed training data and original training data are modeled with svm models and get the best model with optimal hyperparameters. Hyperparameter tuning is performed with the following parameters: Cost (misclassification penalty parameter); kernel (linear, poly, rbf, and sigmoid); and gamma. In addition, to obtain balanced results, grid search cross validation is performed by randomly dividing the data cluster into ten parts for 10-fold cross validation. The best model will be selected through model evaluation values, accuracy, sensitivity, and AUC obtained from the ROC curve.

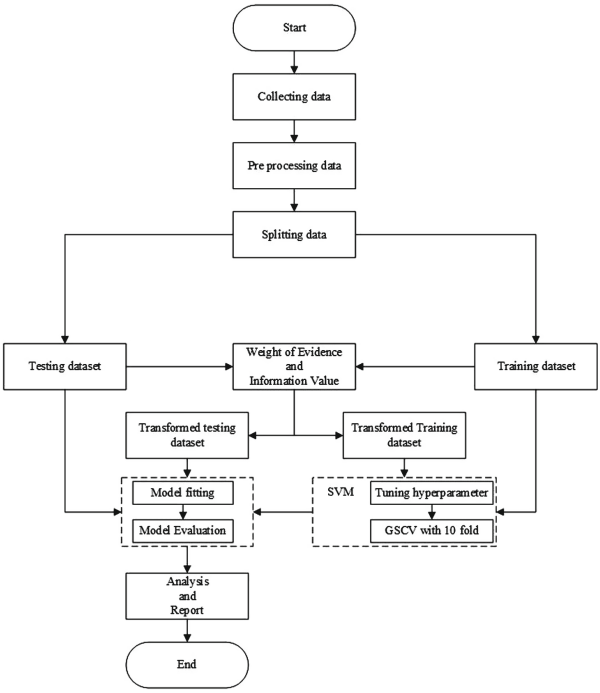

Fig. 3. Research Framework.

4 Results and Discussion

WoE and IV analysis were performed on the training data. The results of the WoE method on data from East Java and West Java provinces have similarities, such as minimum and maximum values. The positive value in WoE shows the chance of features in predicting food insecurity events. The greater the WoE value, the greater the feature predicts food insecurity in each province. The WoE and IV values are shown in the 2 tables below.

The selection of variables in the prediction model uses the technique of information value. This technique aims to rank variables based on their level of importance. Siddiqi

Table 5. In East Java and Barat Java Provinces

Variable	Information Value	
	East Java	West Java
Household head Education	0,21	0,38
Cooking Fuel	0,20	0,26
Number of Family Members Having Saving Account	0,20	0,17
Floor Area	0,20	0,13

Table 6. Rules Related to Information Value

Information Value	**Variable Predictiveness**
< 0.02	Unpredictive
0.02 *to* 0.1	Weak
0.1 *to* 0.3	Medium
0.3 *to* 0.5	Strong
> 0.5	suspicious

suggests a rule for evaluating 4 as shown in Table 5. [13]. The higher the IV, the more important the variable is in the prediction model (Table 6.).

Based on the analysis of WoE and IV, classification of food insecurity events is carried out with the transformed data and the original data using the svm model. Variable selection is done by looking at the strength of predictive information value. Determination of the best hyperparameter of the svm model is done simultaneously with grid search cross validation 10 times. Some hyperparameters (HP) used are kernel function, C parameter, degree for polynomial kernel, and gamma for radial. The parameters used can be seen in Table 7..

The best hyperparameters obtained from several simulations are presented in Table 8..

Simulations used for East Java and West Java provinces are using original data, original data with feature selection, transformed data, and transformed data with feature selection.

Table 7. Hyperparameter SVM'S

HP	SVM 1	SVM 2	SVM 3	SVM 4
Kernel	Linear	Rbf	Sigmoid	Polynomial
Cost	0.5 to 1500	0.5 to 1500	0.5 to 1500	0.5 to 1500
Gamma	-	auto, scale	-	-
Degree	-	-	-	2 and 3

Table 8. Best Hyperparameter

hyperparameter	East Java Provinces	West Java Provinces
Kernel	Rbf	Rbf
Cost	200	500
Gamma	auto	auto

Table 9. Model Evaluation for Classification

Province	Simulations	Acc	Sensitivity	Auc
East Java	Original Data	0.48	0.67	0.58
West Java	WoE Tranformed	0.61	0.68	0.66
	Original Data	0.74	0.77	0.81
	WoE Tranformed	0.75	0.79	0.83

Experimental results show that, the Weight of Evidence (WoE) method selects features based on the value of predictive information that can improve the evaluation value of the model used. This can be seen through the comparison of model evaluation values presented in Table 9.. The combination of WoE and IV methods with the SVM model presents a confusion matrix accuracy of 0.61, sensitivity of 0.68 for the province of East Java and accuracy of 0.75, sensitivity of 0.79 for the province of West Java. In addition, the optimal AUC value obtained is 0.66 for East Java province and 0.83 for West Java province.

5 Conclusions and Future Works

In this paper, the combination method of WoE and IV with svm model is able to increase the evaluation value of the model used. The auc value obtained is 0.66 for East Java province and 0.88 for West Java province. The selection features used are taken from the information value of the predictive power of each feature. Features used in the classification of food insecurity events range from 0.1 to 0.5 in information value.

Acknowledgment. This research was funded by Indonesia's Ministry of Education, Culture, Research, and Technology, with grantee contract No. 3603/IT3.L1/PT.01.03/P/B/2022. The authors also thank Mr. Sudarko, PhD. For his kindness in helping and facilitating us to conduct the python Cuda code on the RTX 3090 GPU engine of the https://pchembl.id project, Computational Chemistry Laboratory, University of Jember.

References

1. A. Asyiva, B. Susetyo, B. Sartono, A. F. Hadi, and E. Ramadhani, "Interpretable machine learning to characterize food insecurity in Aceh and West Java provinces", *The Proceeding of the Fifth ICCGANT*, 2021.
2. S, Suthaharan, "Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning", *Integrated Series in Information Systems*, Vol. 36, New York : Spinger, 2015.
3. K. Palanichamy, "Integrative Omic Analysis of Neuroblastoma", *Computational Epigenetics and Diseases*, vol. 9, pp. 311-326, 2019.

4. M. R. Phangtriastu, J. Harefa and D. F. Tanoto, "Comparison Between Neural Network and Support Vector Machine in Optical Character Recognition", *Procedia Computer Science*,Vol. 116, pp. 351-357,2017
5. B.C. Kristina, A.F. Hadi, A. Riski, A. Kamsyakawuni, and D. Anggraeni, "The visualization and classification method of support vector machine in lymphoma cancer", *Journal of Physics: Conference Series*, 2020.
6. M. Collett, "Photosensitisation diseases of animals: Classification and a weight of evidence approach to primary causes",*Toxicon: X*, Vol. 3, 100012,2019.
7. C. Snijders, U. Matzat, U.D. Reips, ""Big Data" : Big Gaps of Knowledge in the Field of Internet Science", *International Journal of Internet Science*, Vol. 7, pp. 1-5, 2012.
8. N. Dedić and C. Stanier, "Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery", *International Conference on Enterprise Resource Planning Systems*, Page 114-122, 2017.
9. I.H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions", *SN Computer Science*, Vol. 2, page 160, 2021.
10. J.D. Morgenstern, L.C. Rosella, A.P. Costa, R.J. de Souza, and L.N. Anderson, "Perspective: Big Data and Machine Learning Could Help Advance Nutritional Epidemiology", *Advances in Nutrition*, Vol. 12, **3**, pp. 621-631, 2021.
11. A.Z. Lin, "Variable Reduction in SAS by Using Information Value and Weight of Evidence", *proceeding in SUGI Conference*, 2015.
12. A.H. Alsabhan, K. Singh, A.Sharma, S. Alam, D.D. Pandey, S.A.S. Rahman, A. Khursheed, and F.M. Munshi, "Landslide susceptibility assessment in the Himalayan range based along Kasauli – Parwanoo road corridor using weight of evidence, information value, and frequency ratio", *Journal of King Saud University - Science*, Vol. 34, 2,2022.
13. A.Z. Lin, "Expanding the Use of Weight of Evidence and Information Value to Continuous Dependent Variables for Variable Reduction and Scorecard Development", *proceeding in SUGI Conference*, 2014.
14. M. Kuhn and K. Johnson, "Applied predictive modeling", *Springer*, New York, pp. 247–273, 2013
15. A.C.Muller and S.Guido,"Introduction to Machine Learning with Python", California: *O'Rielly Media*, 2016
16. S. Liu, J. McGree, Z. Ge, and Y. Xie,"4 - Computer vision in big data applications", *Computational and Statistical Methods for Analysing Big Data with Applications*, *Academic Press*, Pages 57–85, 2016.
17. S.K. Mohapatra and M.N. Mohanty, "Chapter 7 - Big data classification with IoT-based application for e-health care", *Cognitive Data Science in Sustainable Computing*, Cognitive Big Data Intelligence with a Metaheuristic Approach, Academic Press, Pages 147-172, 2022.
18. A. Rani, N. Kumar, J. Kumar, J. Kumar, and N.K. Sinha,"Chapter 6 - Machine learning for soil moisture assessment", *Cognitive Data Science in Sustainable Computing*, Deep Learning for Sustainable Agriculture, Academic Press, Pages 143–168, 2022.
19. R. Gholami and N. Fakhari, "Chapter 27 - Support Vector Machine: Principles, Parameters, and Applications", Handbook of Neural Computation, Academic Press, Pages 515-535, 2017.
20. H. Dharmawan, B. Sartono, A. Kurnia, A. F. Hadi and E. Ramadhani, "A study of machine learning algorithms to measure the feature importance in class-imbalance data of food insecurity cases in Indonesia", *Commun. Math. Biol. Neurosci*, 2022.