SHM PUBLISHER
SURYA HIJAU MANFAAT

Journal of Information System Exploration and Research

https://shmpublisher.com/index.php/joiser

p-ISSN 2964-1160 | e-ISSN 2963-6361

Bankruptcy Prediction Using Genetic Algorithm-Support Vector Machine (GA-SVM) Feature Selection and Stacking

Wiena Faqih Abror[1*], Alamsyah[2], Muhammad Aziz[3]
[1,2]Department Department of Computer Science, Universitas Negeri Semarang, Indonesia
[3]Department of Electrical Engineering, Kookmin University, South Korea

DOI: https://doi.org/10.52465/joiser.v1i2.180
Received 08 July 2023; Accepted 18 July 2023; Available online 19 July 2023

Article Info

Keywords:
Bankruptcy;
Genetic algorithm;
Support vector machine;
Stacking

Abstract

Bankruptcy is an impact caused by a company's financial failure. Financial failure in the company must be avoided so as not to cause losses to the company. In the research that was carried out utilizing a data set from the Taiwan Economic Journal as many as 6,819 to be trained using machine learning algorithms using classification techniques. The goal obtained from the research conducted is to obtain a classification technique with the best accuracy results. The method used in this research is preprocessing using the synthetic minority over-sampling technique to handling unbalanced data sets. Then, the results of the balanced data set will be processed using a genetic algorithm-support vector machine feature selection algorithm to reduce the attributes of the data set. Data sets that have experienced reduced attributes will be trained using the stacking method with a single classifier base learner in the form of k-nearest neighbors, naïve bayes, decision trees with classification and regression tree models, gradient boosting decision trees, and light gradient boosting. The meta-learner used in the stacking method is extreme gradient boosting. The results of the accuracy obtained from the research conducted were 99.22%.

1. Introduction

Financial bankruptcy or financial failure can harm global economic circulation. Companies, small traders, traditional markets, and the state can feel the negative impact. Practitioners, investors, government, and academic researchers try to identify the variables that cause bankruptcy [1], [2]. The causes of bankruptcy can be of various types, such as increased raw materials, employee fees, business competition, and managerial incompetence. Causes of bankruptcy can occur in every businessperson. So at every level, there can be internal and external factors for business actors [3]. Managing savings and reducing economic credit risk can improve credit risk assessment [4]. Bankruptcy assessments offer

* ***Corresponding Author:***

Wiena Faqih Abror,
Department of Computer Science,
Universitas Negeri Semarang,
Gunungpati, Sekaran, Semarang, Indonesia.
Email: wf.abror@gmail.com

valuable information to investors, management, shareholders, and governments to make decisions about protecting their finances so that bankruptcy does not occur. Bankruptcy research can provide early warningand detect areas of financial weakness. The prediction of bankruptcy analysis can have benefits such as reducing the cost of credit analysis, financial monitoring, and collection rates [5].

Bankruptcy analysis is similar to the classification model, taken from statistical data provided by the company to map the characteristics and indicators of the causes of bankruptcy. Classification problems can be solved with classification algorithms [6], such as Multivariant Discriminant Analysis (MDA), Logistic Regression (LR), Neural Network (NN), Support Vector Machine (SVM), and Ensemble Method [7]. Theuse of other algorithms that use the characteristics of Neural Networks, such as Recurrent Neural Network(RNN) [8] and Convolutional Neural Network (CNN) [9] have been developed to analyze financial and management topics [7].

High indicator dimensions can be reduced using feature selection to improve predictive performance [10]. Kohavi and John [11] stated that high dimensions can degrade the performance of the resulting prediction accuracy, such as in Decision Trees (DT) and Naive-Bayes (NB), because the attributes in the data are irrelevant to the algorithm. Feature selection can reduce the high dimension obtained from a combination of FRs and CGIs indicators [1].

Feature selection is used according to the type of data object that is owned. Data that has a label is called supervised. The wrapper method is one variation of the supervised feature selection method that can be used to reduce height dimensions [12].

Genetic Algorithm (GA) has natural selection characteristics by applying selection, crossover, and mutationrules [13]. GA characteristic rules can be used to solve nonlinear optimization problems and form classifierdiscriminants. GA can be used as a wrapper in the feature selection step. The wrapper method will iterate over the model used, such as GA, which is limited in iterations based on the desired generation formation [12]. The wrapper method requires data validation using an object model to assess the accuracy of the selected features through GA [14]. The model object used is a Support Vector Machine (SVM). SVM cansolve complicated problems such as high dimensions and nonlinear data [15].

The features subset results and data that have reached the specified generation (stop criteria) will be trained using the ensemble method. The ensemble method will train data in parallel and produce accuracy values [16]. One of the algorithms applied to the ensemble method is stacking. The stacking algorithm can train data based on the machine learning algorithm used. Sequentially, the stacking algorithmwill divide the training into two parts, namely the base learner and the meta-learner [17]. The data will be trained using a predetermined algorithm at the base learner stage. This study uses four machine learning algorithms as base learners, namely a decision tree with the Gini index criteria [18], a gradient-boosting decision tree [19], k-nearest neighbours [20], and a light gradient boosting machine [21]. At the meta-learner stage, Extreme Gradient Boosting (XGBOOST) is used to train the resulting data from the base learner [22], [23]. Zelenkov et al., [24] used an ensemble classifier despite not using a filtering method. Kim and Kang [25] stated that ensemble learning can improve the performance of a single classifier because ensemble learning can improve classifier accuracy.

2. Method

The research approach is carried out through the research design that is made. The research design is madeto explain the research flow comprehensively. In general, the workflow is divided into three stages, namelythe preprocessing stage, the feature selection stage, and the data training stage. The research ends with a validation test using a confusion matrix and k-fold cross-validation. The research workflow is in Figure 1.

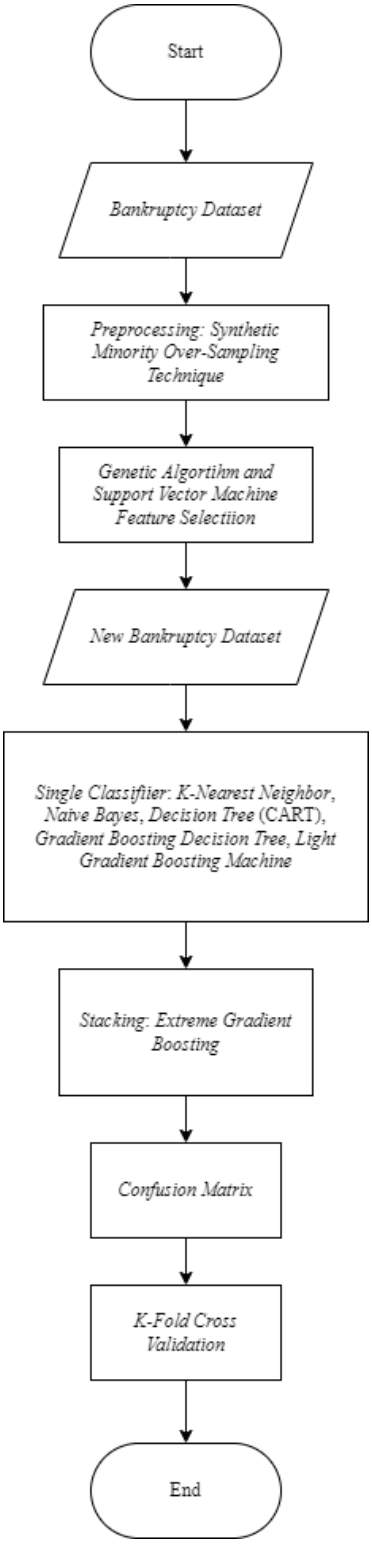

Figure 1. Proposed Method

The general explanation of the research workflow is that the preprocessing stage uses the synthetic minorityover-sampling technique method, which aims to handle unbalanced data sets. The feature selection stage uses the wrapper method with a genetic algorithm as feature selection and uses a support vector machine to test features that have been searched using a genetic algorithm, as well as the data training stage which is divided into two parts using. The first is a single classifier, such as decision trees, naïve bayes, k-nearest neighbours, gradient boosting decision trees, and light gradient boost machines; the second uses stacking to combine a single classifier as a base learner and extreme

gradient boost as a meta-learner. Each part at thedata training stage will be validated using a confusion matrix.

3. Results and Discussion

The problem found in this study is an imbalance of data. The balance of the data in the target Taiwanese Bankruptcy data set is shown in Figure 2, and class 0 shows a distribution of 6,599 data sets. Class 1 shows a spread of 220 data sets. The difference shown on the chart shows the imbalanced data in the Taiwanese Bankruptcy data set. Visualization in the form of a chart of the target data set of Taiwanese Bankruptcy can be seen in Figure 2.

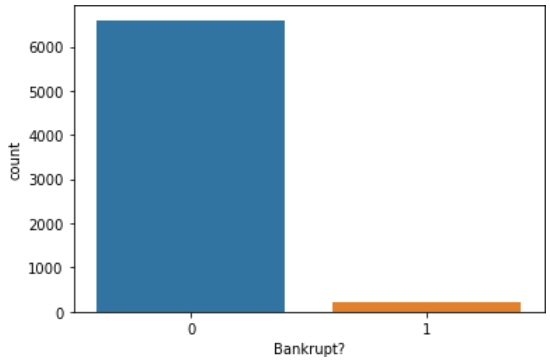

Figure 2. Imbalanced Data set

Visualization of the target data set in chart form is shown in Figure 3. which shows that the distribution ofthe data set has been balanced using the synthetic minority over-sampling technique method.

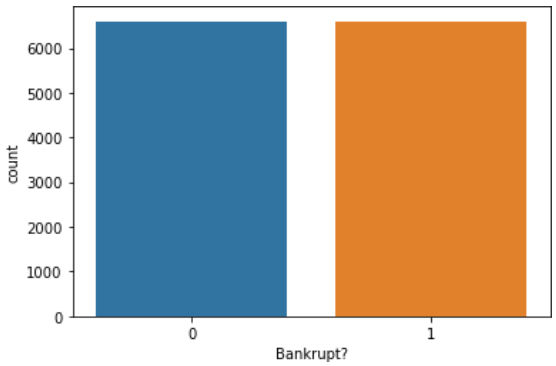

Figure 3. Balanced Data set

The results of balancing the data set in nominal terms are shown in Figure 3. Class 0 shows a distribution of 6,599 data sets and class 1 shows a distribution of 6,599 data sets, which means that the distribution of data sets is balanced.

The Taiwanese Bankruptcy data set is processed using a genetic algorithm-support vector machine featureselection aimed at reducing attributes. The default attribute is 96 along with the target column. The attributes obtained after being processed using a genetic algorithm-support vector machine feature selectionare 44 along with a target column declaring bankruptcy. The new data set that was trained was tested usingcross validation. The accuracy results are in Table 1. Visualization comparison shown in Figure 4.

Table 1. Comparison of Model Performance

Algorithms	Accuracy (%)
K-Nearest Neighbors	96.87
Naïve Bayes	72.34
Decision Tree – CART	95.14
Gradient Boosting Decision Tree	97.13
Light Gradient Boosting Machine	98.74
Stacking – Extreme Gradient Boosting	99.22

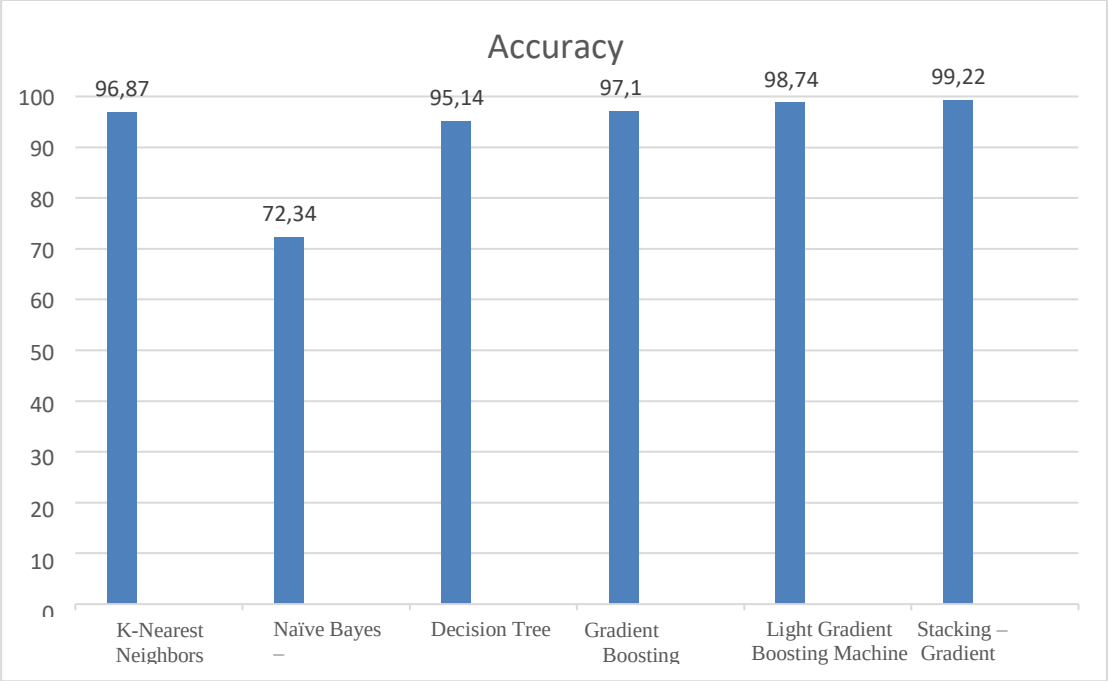

Figure 4. Comparison of Model Performance

4. Conclusion

The research was conducted by combining a single classifier, namely k-nearest neighbors, naïve bayes, decision trees using the classification and regression tree method, gradient boosting decision trees, and lightgradient boosting machines using the stacking method. The single classifier is used as a base learner in thestacking method. The meta-learner used is extreme gradient boosting. The results of the accuracy of the research conducted were 99.22%.

References

[1] D. Liang, C. C. Lu, C. F. Tsai, and G. A. Shih, "Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study," *Eur. J. Oper. Res.*, vol. 252, no. 2, pp. 561–572, 2016, doi: 10.1016/j.ejor.2016.01.012.

[2] M. A. Muslim, Y. Dasril, H. Javed, W. F. Abror, D. A. A. Pertiwi, and T. Mustaqim, "An Ensemble Stacking Algorithm to Improve Model Accuracy in Bankruptcy Prediction," *J. Data Sci. Intell. Syst.*, vol. 1, no. 1, 2023.

[3] A. Kadim and N. Sunardi, "Analisis Altman Z-Score Untuk Memprediksi Kebangkrutan Pada Bank Pemerintah (Bumn) Di Indonesia Tahun 2012-2016 Articles Information Abstract Prodi Manajemen Unpam," *Keuang. dan Investasi)*, vol. 1, no. 3, pp. 142–156, 2018.

[4] D. West, S. Dellana, and J. Qian, "Neural network ensemble strategies for financial decision applications," *Comput. Oper. Res.*, vol. 32, no. 10, pp. 2543–2559, 2005, doi: 10.1016/j.cor.2004.03.017.

[5] S. Lee and W. S. Choi, "A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis," *Expert Syst. Appl.*, vol. 40, no. 8, pp. 2941–2946, 2013, doi: https://doi.org/10.1016/j.eswa.2012.12.009.

[6] P. Pampouktsi, S. Avdimiotis, M. Maragoudakis, M. Avlonitis, P. Hoogar, and G. Ruhago, "Techniques of Applied Machine Learning Being Utilized for the Purpose of Selecting and Placing Human Resources within the Public Sector," *J. Inf. Syst. Explor.*, vol. 01, no. 01, pp. 1–16, 2023.

[7] Y. Qu, P. Quan, M. Lei, and Y. Shi, "Review of bankruptcy prediction using machine learning and deep learning techniques," *Procedia Comput. Sci.*, vol. 162, pp. 895–899, 2019, doi: https://doi.org/10.1016/j.procs.2019.12.065.

[8] E. Chong, C. Han, and F. C. Park, "Deep learning networks for stock market analysis and

prediction: Methodology, data representations, and case studies," *Expert Syst. Appl.*, vol. 83, pp. 187–205, 2017, doi: https://doi.org/10.1016/j.eswa.2017.04.030.
[9] T. Hosaka, "Bankruptcy prediction using imaged financial ratios and convolutional neural networks," *Expert Syst. Appl.*, vol. 117, pp. 287–299, 2019, doi: https://doi.org/10.1016/j.eswa.2018.09.039.
[10] I. Guyon, J. WESTON, S. Barnhill, and V. Vadnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Mach. Learn.*, vol. 46, pp. 389–422, 2002.
[11] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1–2, pp. 273–324, 1997.
[12] N. El Aboudi and L. Benhlima, "Review on wrapper feature selection approaches," in *2016 International Conference on Engineering MIS (ICEMIS)*, 2016, pp. 1–5. doi: 10.1109/ICEMIS.2016.7745366.
[13] A. R. Safitri and M. A. Muslim, "Improved Accuracy of Naive Bayes Classifier for Determination of Customer Churn Uses SMOTE and Genetic Algorithms," *J. Soft Comput. Explor.*, vol. 1, no. 1, pp. 70–75, 2020.
[14] W. C. Lin, Y. H. Lu, and C. F. Tsai, "Feature selection in single and ensemble learning-based bankruptcy prediction models," *Expert Syst.*, vol. 36, no. 1, pp. 1–8, 2019, doi: 10.1111/exsy.12335.
[15] Z. Tao, L. Huiling, W. Wenwen, and Y. Xia, "GA-SVM based feature selection and parameter optimization in hospitalization expense modeling," *Appl. Soft Comput.*, vol. 75, pp. 323–332, 2019, doi: https://doi.org/10.1016/j.asoc.2018.11.001.
[16] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Front. Comput. Sci.*, vol. 14, no. 2, pp. 241–258, 2020, doi: 10.1007/s11704-019-8208-z.
[17] J. Dou *et al.*, "Improved landslide assessment using support vector machine with bagging, boosting, and stacking ensemble machine learning framework in a mountainous watershed, Japan," *Landslides*, vol. 17, no. 3, pp. 641–658, 2020, doi: 10.1007/s10346-019-01286-5.
[18] H. H. Patel and P. Prajapati, "Study and Analysis of Decision Tree Based Classification Algorithms," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 10, pp. 74–78, 2018, doi: 10.26438/ijcse/v6i10.7478.
[19] J. S. Yang, C. Y. Zhao, H. T. Yu, and H. Y. Chen, "Use GBDT to Predict the Stock Market," *Procedia Comput. Sci.*, vol. 174, no. 2019, pp. 161–171, 2020, doi: 10.1016/j.procs.2020.06.071.
[20] W. Xing and Y. Bei, "Medical Health Big Data Classification Based on KNN Classification Algorithm," *IEEE Access*, vol. 8, pp. 28808–28819, 2020, doi: 10.1109/ACCESS.2019.2955754.
[21] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," *Adv. Neural Inf. Process. Syst.*, no. December, pp. 3147–3155, 2017.
[22] M. A. Muslim and Y. Dasril, "Company bankruptcy prediction framework based on the most influential features using XGBoost and stacking ensemble learning," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 6, pp. 5549–5557, 2021, doi: 10.11591/ijece.v11i6.pp5549-5557.
[23] M. A. Muslim *et al.*, "New model combination meta-learner to improve accuracy prediction P2P lending with stacking ensemble learning *," *Intell. Syst. with Appl.*, vol. 18, no. December 2022, p. 200204, 2023, doi: 10.1016/j.iswa.2023.200204.
[24] Y. Zelenkov, E. Fedorova, and D. Chekrizov, "Two-step classification method based on genetic algorithm for bankruptcy forecasting," *Expert Syst. Appl.*, vol. 88, pp. 393–401, 2017, doi: https://doi.org/10.1016/j.eswa.2017.07.025.
[25] M. J. Kim and D. K. Kang, "Classifiers selection in ensembles using genetic algorithms for bankruptcy prediction," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 9308–9314, 2012, doi: 10.1016/j.eswa.2012.02.072.