JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 6. NO. 1 AUGUST 2020
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v6i1.1390

A PREDICTION MODEL OF COMPANY HEALTH USING BAGGING CLASSIFIER

Green Arther Sandag

Informatics Study Program
Universitas Klabat
http://www.unklab.ac.id/
greensandag@unklab.ac.id

Abstract—In business, have many competitions between companies occur to obtain as many profits as possible, Financial Distress is a financial decline that occurs in companies, reflecting the health of the company before bankruptcy started. Therefore, to avoid bankruptcy, it requires a method or tool with high accuracy in identifying company health. This research uses a bagging classifier, which is one type of Ensemble Learning algorithm. To predict financial difficulties, the authors use the bagging classifier algorithm with 0.13% more accurate results than previous studies using the XGBoost algorithm.

Keywords: Financial Distress, Bagging Classifier, Xgboost.

Abstrak—*Dalam persaingan bisnis layaknya sebuah kompetisi antar perusahaan agar perusahaan tersebut mendapatkan keuntungan, Financial Distress adalah penurunan kondisi keuangan yang terjadi pada perusahaan, mencerminkan kesehatan perusahaan sebelum terjadinya kebangkrutan. Oleh karena itu untuk menghindari kebangkrutan terjadi dibutuhkan metode atau tools yang memiliki keakuratan yang tinggi mengidentifikasi kesehatan perusahaan. Penelitian ini menggunakan bagging classifier yang merupakan salah satu jenis algoritma Ensemble Learning. Untuk memprediksi financial distress penulis menggunakan algoritma bagging classifier dengan hasil 0.13% lebih akurat dibandingkan dengan penelitian sebelumnya yang menggunakan algoritma..*

Kata Kunci*: Financial Distress, Bagging Classifier, Xgboost.*

INTRODUCTION

Companies compete when it comes to business. They compete with each other to obtain as many profits as possible. There a lot of companies are competing though they are running the same sector of business. A more reliable and experienced company has a higher chance of getting more profit, as we can see from its strategy, product, asset, and other elements [1].

There is a chance for another company to compete with big companies. On the other side of the competition, it could harm the other party, the losing company. The company which has lost the competition will only get a smaller profit from the big company that was their competitor. The lost company has the potential to go bankrupt. It is known as Financial Distress. Financial distress is a condition of Financial because it is unable to meet its financial obligations, which is happened to a company that can lead to a bankruptcy [2].

We live in a modern era where technology and economy have developed rapidly, so people need to learn more about technology and economy data [3]. Finance Distress could be a good indicator for identifying the company condition, which it was known before, that a Financial Distress is needed to survive in whatever the situation is, especially when the company is in an unbearable condition. Financial distress can predict the things that are going to happen through company data [4].

According to that problem, the research has created a model to predict the Financial Distress. The data used on this research consists of information about the company name, a periodical of the year of data taking, the number of financial distress, and 83 columns representing the financial and non-financial factors of a company. The column's name and its contents are numbers instead of characters. Meanwhile, the user data are in italic, although there are no missing data.

There is more than one algorithm that can be used to analyze this data, and in this case, we used a bagging classifier, which is one of Ensemble Learning algorithms [5]. Bagging Classifier compiled many presumption values and turned them into one value. This method is specifically used to analyze data without understanding the

details of the data context that is going to be used. Its random ability characterizes this algorithm, and its minimum name is biased [6].

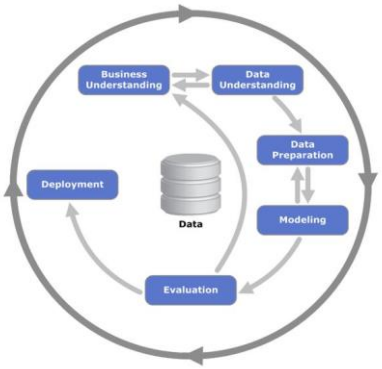
According to previous research by Huang in predicting the level of Financial Distress in Taiwan company from the 2010-2016 data using four algorithms which are Support Vector Machine (SVM), Hybrid Associative Memory with Translation (HACT), Hybrid GA-fuzzy Clustering and Extreme Gradient Boosting (XGBoost), and Deep Belief Network (DBN), which is found that XGBoost algorithm is one of the Ensemble Learning Algorithms that its accuracy level has reached to 90.6% [7].

The purpose of this research is to predict the company's health and the tendency to bankrupt based on financial distress and obtain an excellent performance using the bagging classifier method. The benefit of this research is to help the company to know its bankrupt potential in advance.

MATERIALS AND METHODS

This research uses the Cross-Industry Standard Process for Data Mining (CRISP-DM) as the research method. CRISP-DM is a standard process for creating a model using a general approach used by data mining experts. There are 6 phases or steps in CRISP-DM; each of them holds a different task: *business understanding, data understanding, data preparation, modeling, evaluation,* and *deployment*. Picture 1 is showing the CRISP-DM steps.

According to Vorhies, whom one of CRISP-DM method creators. He stated that all of the data science projects were started with business understanding, then data collecting, data processing, and data mining algorithm implementation. CRISP-DM provides firm guidelines for data science activity today. Therefore, the researcher chose to use this CRISP-DM method.



Picture 1. Cross-Industry Standard Process for Data Mining Steps [8]

Business Understanding

The first step of the CRISP-DM process is to understand the purpose that will be reached within the business perspective. The purpose of this step is to see the essential factors that will impact the project's final result. Hence the researcher determined to create a prediction model for the company health based on the financial distress factor.

Data Understanding

This stage required us to get and to understand the data. The data that will be used for this research is a Financial Distress Prediction dataset from Kaggle. Financial Distress Dataset has 3672 rows and 86 variables, which can be seen in Table 1.

Table 1. Dataset Parameter

Parameter	Details	Value type
COMPANY	Representing the name of the company as a sample	*Integer*
TIME	Serving all of the yearly period in data sample taking	*Integer*
FINANCIAL DISTRESS	Representing the number of company health in a specific period, if it is higher than – 0.5, the company is in health; on the contrary, it is in an unstable condition.	*Polynomial*
X1-X79	Representing the financial and non-financial factors in the company	*Polynomial*
X80	Serving the industrial sector of the company work	*Integer*
X81-X83	Representing the financial and non-financial factors of the company	*Polynomial*

Data Preparation

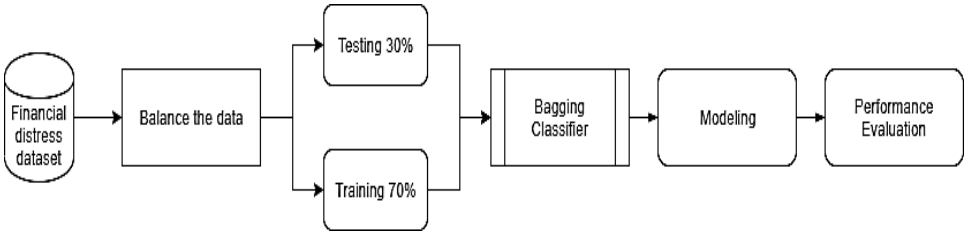
This step is a step to decide which data will be used for analyzing and evaluating the data quality. This step is cleaning the data and integrating data. There are many variables in missing value. Therefore the researcher has to clean the uncompleted data.

Modeling

This step is creating the prediction model using data mining algorithms such as bagging classifier, decision tree, naïve Bayes, k-nearest

VOL. 6. NO. 1 AUGUST 2020
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v6i1.1390

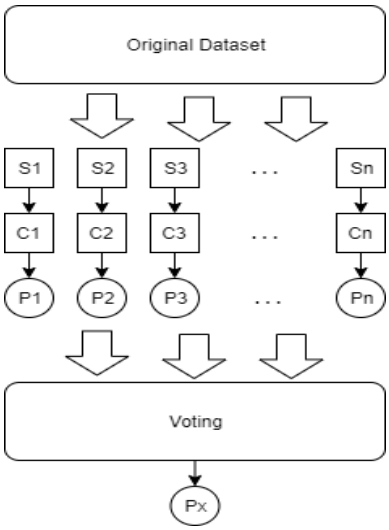
neighbor, etc. Picture 2 is showing the financial prediction process. The tool that is going to be used is a phyton script to process the dataset. The first stage is obtaining the Financial distress dataset from Kaggle. After the data was balanced, we can manipulate the data to approach its deviation standard and reach the maximum value. The dataset is divided into 70% of training data and 30% of testing data [9] the same as the ratio from the previous research [7]. The next stage is to analyze the dataset and creating its model using the Bagging Classifier algorithm. The result of the modeling will be represented into graphics, then going through the performance evaluation.



Picture 2. The architecture to predict Financial Distress

Bagging Classifier Algorithm is an algorithm that uses many kinds of sample data from datasets to divide them into some data training and data tests. Bagging Classifier Algorithm resulted in some presumptions or probability values and voting them to obtain one real value [7]. Picture 3 shows us the procedure of the bagging classifier, starting from getting a few samples from the original dataset (S1- Sn) then implementing the random forest algorithm (C1-Cn) inside the classifier whence this is the foundation of bagging classifier algorithm. After obtaining the presumption or probability values (P1-Pn), they need to be voted (P1-Pn). Then the result of the most votes will become a general probability or presumption value (Px).

After we have successfully created the model using a bagging algorithm, the next step is to measure it.



Picture 3. Bagging Classifier *Evaluation* Procedure

This performance measurement step is to observe the performance of each algorithm in creating the model. The evaluation step is to assess how far this model has fulfilled its purpose from the first step. Confusion Matrix is a bagging classifier algorithm that is entirely useful to measure the performance of each algorithm. This algorithm is a measuring tool to identify an algorithm's performance in doing data testing [8]. The measurement values are true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN). True-positive (TP) and true-negative (TN) are the profit values that are reached through the right prediction, and false-negative (FN) and false-positive (FP) are the fault or error values which are reached through a projection that has errors. [8].

F-Score identifies the accuracy by examining the precision and recall to determine the accuracy accurately and reliable. To do the evaluation using F-Score, a β value configuration must be executed first to get the specific result. If β= 1, then the data is already balanced, but if β> 1, the evaluation emphasized its precision, and on contrary, the assessment emphasized its recall. To examine this model on the dataset, the researcher is also using 5-fold cross-validation. It is to test the training dataset and divide it into a 5-fold dataset with the same size then examining them.

Performance measurement is useful for testing the model created from classifiers like Recall, Precision, Accuracy, Root Means Square error (RMSE), and F-Score. The recall is a process of classification in a positive data collection, classified correctly as positive data. Precision is a process of classification in a positive classified data

collection, which is resulted in positive also. Accuracy is a data classification resolve [10]. RMSE helps us to reveal the errors of a model [11]. F-Score identifies accuracy by examining the precision and recall to determine accuracy and reliability accurately [12]. These are the formulas in searching of Recall, Precision, Accuracy, RMSE, and F-Score:

Recall Formula:

$$Recall = \frac{TP}{TP+FN}, \quad \text{.....(1)}$$

Precision Formula:

$$Precision = \frac{TP}{TP+FN}, \quad \text{.....(2)}$$

Accuracy Formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \quad \text{.....(3)}$$

RMSE Formula:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}i - yi)^2}{n}}, \quad \text{.....(4)}$$

F-Score Formula:

$$\text{F-Score} = \frac{(\beta^2 + 1) * precision * recall}{\beta^2 * precision + recall}, \quad \text{.....(5)}$$

Description:

TP : *true positive*
TN : *true negative*
P : Number of Total Positive data
FP : *false positive*
N : Number of Total Negative data
FN : *false negative*
$(\hat{y}i - yi)$: Difference between predicted and actual values
n : Total value
β : F-Score configuration value

Deployment

Deployment is the last step of CRISP-DM. This step will draw the evaluation results of every tested classifier when creating the prediction model of company health and determining a strategy to implement the established model.

RESULTS AND DISCUSSION

This is the part of analyzed and modeling results in the financial distress dataset using the Bagging Classifier algorithm.

The Comparison between XGBoost Algorithm and Bagging Classifier Using 5-Fold Cross-Validation Dataset

Table 2 shows the performance between two algorithms, which are the XGBoost that is used on the previous research and Bagging Classifier, which is this research object algorithm that generates a close result. However, Bagging Classifier works higher than XGBoost. We can see from the accuracy value which got 0.13% higher than F-Score, except the recall value, which reached 0.09% lower. Therefore, Bagging classifier is the algorithm with higher performance in predicting the company health with 96.01% accuracy, 98.3% recall, 97.56% precision, 0.19 RMSE, and 97.93% F-Score

Table 2. The Performance Comparison between XGBoost and Bagging Classifier (BC) on 5-Fold Cross-Validation Dataset

Algorithm	Acc (%)	Rec (%)	Prec (%)	RMSE	f-Score
XGBoost	95.73	98.39	97.20	0.206	97.80
Bagging Classifier	96.01	98.30	97.56	0.199	97.93

The Comparison Between XGBoost Algorithm and Bagging Classifier using Independent Dataset

After completing the testing phase using a 5-fold dataset, the researcher executes an independent dataset using XGBoost and Bagging Classifier Algorithms, where the results can be seen in table 3. Bagging classifier algorithm performs the best performance among other algorithms in its kind, such as XGBoost with 95.01 % accuracy, 95.8% recall, 96.2% precision, and 0.184 RMSE and 94.83% F-Score. Bagging classifier algorithm can predict the company health using financial distress dataset with an increase of 2 % from the XGBoost algorithm.

Table 3. The Performance Comparison between XGBoost and Bagging Classifier (BC) on Independent Dataset

Algorithm	Acc (%)	Rec (%)	Prec (%)	RMSE	f-Score
XGBoost	92.43	93.25	92.20	0.231	93.8
Bagging Classifier	95.01	95.80	96.20	0.184	94.83

Bagging Classifier Algorithm Performance in Predicting Company Health Using 5-fold Cross-Validation Dataset

The best algorithm performance in creating a prediction model of company health is using bagging classifiers with 97.01% accuracy, 96.2 % recall, 97.36% precision, and 0.183 RMSE and 97.03% F-Score that can be seen in table 4.

Accredited Rank 3 (Sinta 3) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti
No. 28/E/KPT/2019, September 26, 2019. Published by PPPM STMIK Nusa Mandiri

VOL. 6. NO. 1 AUGUST 2020
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v6i1.1390

Table 4. Independent Test Performance Comparison

Algorithm	Acc (%)	Rec (%)	Prec (%)	RMSE	f-Score
Bagging Classifier	97.01	96.2	97.36	0.183	97.03
Support Vector Machine	94.45	93.54	97.50	0.341	94.93
Logistic	93.25	90.52	96.38	0.422	95.04
Decision Tree	94.50	95.16	98.10	0.225	95.89

Bagging Classifier Algorithm Performance in Predicting Company Health Using Independent Dataset

Table 5 shows the results of performance evaluation from the classification algorithm without using a sampling method but an independent dataset. Based on the results, Bagging Classifier got a higher accuracy, recall, RMSE, F-Score than other algorithms with accuracy in 96.01%, 98.30% of recall, 0.199 of RMSE, and 97.93% of F-Score but the precision is the second lower in 97.56%.

Table 5. Independent Test Performance Comparison

Algorithm	Acc (%)	Rec (%)	Preci (%)	RMSE	f-Score
Bagging Classifier	96.01	98.30	97.56	0.199	97.93
Support Vector Machine	90.65	91.04	99.17	0.305	94.93
Logistic	90.83	91.23	99.18	0.302	95.04
Decision Tree	94.10	96.32	97.52	0.242	96.91

CONCLUSION

After evaluating the five algorithms, the best performing algorithm among the other algorithms, XGBoost, Support Vector, Logistic, and Decision tree, go to the Bagging Classifier algorithm with 0.13% - 54.36% of higher accuracy in predicting the company health.

RECOMMENDATION

This model can be used as a foundation for implementing the other Ensemble Learning algorithm. For further research, it is recommended to add more sampling methods to increase data quality.

REFERENCE

[1] R. K. Brahmana, "Identifying Financial Distress Condition in Indonesia Manufacture Industry." Birmingham Business School, University of Birmingham, United Kingdom, pp. 1–19, 2007.

[2] H. D. Piatt and M. B. Piatt, "Predicting corporate financial distress: Reflections on choice-based sample bias," *J. Econ. Financ.*, vol. 26, no. 2, pp. 184–199, 2002.

[3] D. Arditi and T. Pulket, "Predicting the Outcome of Construction Litigation Using an Integrated Artificial Intelligence Model," *J. Comput. Civ. Eng.*, vol. 24, no. 1, pp. 73–80, Jan. 2010.

[4] M. S. Gameel and K. El-Geziry, "Predicting Financial Distress: Multi Scenarios Modeling Using Neural Network," *Int. J. Econ. Financ.*, vol. 8, no. Oct, 2016.

[5] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Woźniak, "Ensemble learning for data stream analysis: A survey," *Inf. Fusion*, vol. 37, pp. 132–156, Sep. 2017.

[6] A. Wibowo and A. Purwarianti, "PENERAPAN BAGGING UNTUK MEMPERBAIKI HASIL PREDIKSI NASABAH PERUSAHAAN ASURANSI X," Kota Batam, 2011.

[7] Y. P. Huang and M. F. Yen, "A new perspective of performance comparison among machine learning algorithms for financial distress prediction," *Appl. Soft Comput. J.*, vol. 83, p. 105663, Oct. 2019.

[8] C. Sammut and G. Webb, *Encyclopedia of Machine Learning and Data Mining*, 2nd ed. Boston: Springer US, 2017.

[9] A. Abdelaziz, A. S. Salama, A. M. Riad, and A. N. Mahmoud, "A Machine Learning Model for Predicting of Chronic Kidney Disease Based Internet of Things and Cloud Computing in Smart Cities," in *Security in Smart Cities: Models, Applications, and Challenges* , Springer, Cham, 2019, pp. 93–114.

[10] M. Bramer, *Principles of Data Mining*, Fourth. London: Springer London, 2020.

[11] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the

literature," *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, Jun. 2014.

[12] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation," in *AAAI Workshop - Technical Report*, 2006, vol. WS-06-06, pp. 1015–1021.

Accredited Rank 3 (Sinta 3) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti
No. 28/E/KPT/2019, September 26, 2019. Published by PPPM STMIK Nusa Mandiri