

## Using Machine Learning Techniques to Predict Financial Distress in Rural Banks in Indonesia

Maysas Yafi Urrochman<sup>1</sup>. Hasyim Asy'ari<sup>2</sup>. Abdur Ro'uf<sup>3</sup>

<sup>123</sup>Institut Teknologi dan Bisnis Widya Gama Lumajang, Informatika, Institut Teknologi dan Bisnis Widya Gama Lumajang, Jl. Gatot Subroto No. 4, Lumajang, Indonesia

Corresponding Author: Maysas Yafi Urrochman (maysasyafi@gmail.com)

### ARTICLE INFO

Date of entry:  
07 April 2024  
Revision Date:  
25 April 2024  
Date Received:  
30 April 2024

### ABSTRACT

LPS liquidated about 100 people's Rural Banks between 2015 and 2019, indicating that these banks are facing significant issues, particularly financial distress. This study seeks to forecast financial distress through a two-stage classification and regression approach. Researchers used financial report data from Rural Banks in Indonesia from 2015 to 2019, covering a total of 150 banks, with 50 financial ratios from bankrupt banks and 100 from those that remained operational. Data was analyzed for two consecutive years prior to any bankruptcy declarations. The classification targets are categorized into five classes: very healthy, healthy, quite healthy, unhealthy, and distressed. The study results demonstrate that the two-stage classification and regression method can effectively predict the onset of financial distress. This is validated by the classification outcomes using the Decision Tree Algorithm, which achieved an f1-score accuracy of 88%. The evaluation of timing predictions using Random Forest Regression revealed a mean absolute error of 1.2 months and a mean absolute percentage error of 3%. These predictions can assist regulators, bank management, and investors in making better-informed decisions to address financial distress risks in Rural Banks. The superior performance of the Decision Tree Algorithm over Naïve Bayes in classifying financial distress highlights the potential of machine learning techniques in providing robust tools for early warning systems, aiding stakeholders in making informed decisions to mitigate risks.

Keywords: Rural Banks, Decision Tree, Financial Distress, Naïve Bayes, Random Forest Regression



Cite this as: Urrochman, M. Y., Asy'ari, H., & Ro'uf, A. (2024). Using Machine Learning Techniques to Predict Financial Distress in Rural Banks in Indonesia. *Journal of Informatics Development*, 2(2), 36-44. <https://doi.org/10.30741/jid.v2i2.1341>

### INTRODUCTION

The biggest risk faced by Rural Banks is the risk of insolvency or bankruptcy (Wilopo, 2001). Financial distress is a situation where a company experiences a financial crisis and fails to fulfill its obligations to creditors because it does not have the money to continue its business (Sutra, et al., 2019). Financial distress is a risk that financial institutions including Rural Banks in Indonesia

must be aware (Ansar, 2018). This risk can occur due to many factors. such as a decline in property quality poor financial management. changes in economic conditions and so on. Therefore it is important to make efforts to predict the level of financial difficulties in Rural Banks so that they can take appropriate action to reduce this risk. This situation is accompanied by a decrease in profits and fixed assets and usually occurs before bankruptcy. Bankruptcy is a situation where a company cannot fulfill its obligations (Brigham, 2016). There are early indicators before the emergence of financial distress. which can often be identified early by looking closely at the financial reports and then analyzing them using certain methods or methods (Budiwati, et al., 2021). One of them is by looking at financial ratios which can be used as the first indicator of the possibility of financial pressure that could affect company bankruptcy (Muflihah, 2022).

Every year there is data on Rural Banks that are in the process of being liquidated or currently under liquidation. such as the data on the page <https://www.lps.go.id> for the last 5 years. namely around 100 Rural Banks throughout Indonesia (Ratna and Marwati, 2018). This shows that there is a need for a system used by banks. especially Rural Banks. to avoid financial distress which can affect the progress of Rural Banks businesses. However. until now there has been no system used to predict periods of financial stress in banking. especially Rural Banks. as an early warning method for financial health problems that may exist in Rural Banks.

Many studies related to the estimation of the timing of financial stress have used data mining techniques to automatically obtain useful information based on the costs of risk management in companies. In the actual risk assessment process expert knowledge is also considered to have an important function because the expert's predictions are based on his behavior. The two-level classification and regression method can be used to predict the level of financial pressure on Rural Banks in Indonesia using historical financial data. The first category namely the classification category is used to divide Rural Banks into two categories namely financial distress or healthy. The models used at the classification stage can be trained using various machine learning techniques, such as Decision Tree, Support Vector Machine, K-Nearest Neighbors (KNN), Naïve Bayes and others. In this research, Decision Tree and Naïve Bayes were used. After Rural Banks are classified as having the potential to experience financial stress or not stress the next stage is that the Regression Technique is used to predict the maximum period of financial stress for Rural Banks that are determined to be experiencing a serious financial crisis. This regression model can use linear regression techniques, logistic regression, Support Vector Regression (SVR), Random Forest Regression to predict possible periods of financial difficulty for Rural Banks. In this research the Random Forrest Regression method was used. This research was conducted using financial data from Rural Banks in Indonesia for the 2015-2019 period. This research is a continuation of research conducted by Budiwati et al in 2021 in their research entitled Overview of Bankruptcy Predictions for Conventional Banks, Rural Banks in Indonesia. The data used includes financial ratios including APYDAP, NPL, ROE, BOPO, NIM, LDR, IRR. The data collection period is 3 months, 6 months, 9 months, 12 months, 15 months, 18 months, 21 months and 24 months before becoming a debtor (Meyer and Pifer, 1970). Based on previous research, the results are very satisfactory but there is nothing specific regarding the prediction of financial distress in Rural Banks and the use of two-stage machine learning techniques, namely classification and regression. Therefore. this research will discuss specifically and emphasize the prediction of timing of financial distress at Rural Banks in Indonesia by utilizing two-stage machine learning techniques of classification and regression as well as the formation of a bankruptcy prediction model. In summary, this research has the following important contributions:

1. Compile data containing a dataset containing Rural Banks with the label Financial Distress
2. Several prediction models were built to predict Rural Banks that would experience Financial Distress

## METHOD

### A. Method

Studies on financial distress have been carried out by Beaver and Altman, Haldeman, and Narayanan, both of whom are known as pioneers in prediction distress analysis research, therefore their research is often considered basic research for the development of corporate failure research. However, in these two pioneers there are differences in terms of the analytical model used. In Beaver's research using a univariate model where the variables used are single variables (Beaver, 1966), while in Altman et al.'s research, tried to improve Beaver's research by applying multiple linear discriminant analysis (MDA) (Altman, 1968). Meyer and Pifer and Sinkey Jr were pioneers of research into bankruptcy prediction in banking using financial ratios in banks (Meyer and Pifer, 1970).

In 2017, analyzing the application of machine learning to predict bankruptcy using 8 techniques including bagging, boosting, random forest (RF), SVM with two kernels, artificial neural networks, logistic regression and MDA using financial data on American and Canadian companies starting from 1985 to 2013 collected from NYU's Salomon Center database (Kovacova and Kliestikova (2013)). Financial data including liquidity, profitability, productivity, leverage asset turnover, growth of assets, growth in sales, growth of employees, operational margin, changes in return on equity, and changes in price-to-book ratio are used as variables in research to predict bankruptcy. The results show that traditional models (MDA, LR, and ANN) have a lower prediction capacity of 52% and 77% compared to machine learning models which have predictions of 71% and 87%.

In 2017 created a model to predict bank bankruptcies with results showing that all calculated prediction models have achieved a prediction accuracy of around 50% on a dataset of Slovak companies. This is due mainly to the high type I error in the Fulmer (62%), Altman (99%) and Springate (78%) models (Barboza, et al., 2017). On the other hand in the Zmijewski and Taffler model type I and type II errors. Similar errors are around 50%. According to the calculations given, we can assume that we have not found any important differences between the abilities of the tested models to predict future corporate bankruptcies.

In 2018 (Kesumawati, et al., 2018), Research on Machine Learning and Statistical Techniques for Bankruptcy Prediction indicated that the SVM-PSO prediction algorithm achieved the highest accuracy at 95%, with a precision of 94.73% and a specificity of 95.23%, outperforming both traditional statistical analysis and other machine learning algorithms. Apache Mahout, an open-source software, is utilized for data mining and offers robust and scalable implementations of machine learning algorithms suitable for clustering, classification, and heuristic evolutionary algorithms.

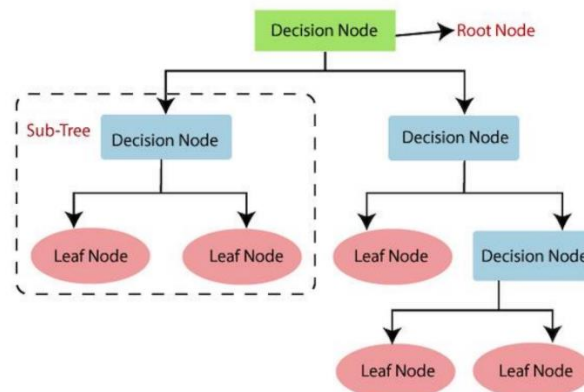
In 2019 (Chang, 2019), who examined company bankruptcy predictions using machine learning models showed research results that there were weaknesses in traditional methods of predicting bankruptcy, so they used the SVM and Random Forrest methods to build prediction models. The result was that when testing the initial model, it had low bankruptcy prediction accuracy due to sample imbalance, so the random forest method was used to increase prediction accuracy and produce an accuracy of more than 70%.

In this research, this researcher used Decision Tree and Naïve Bayes methods for classification. Decision Tree and Naïve Bayes are expected to be able to classify Rural Banks into conditions financial distress or non-financial distress. Where as Random Forest Regression it is hoped that it can predict when financial distress will occur at a Rural Bank.

#### A.1. Decision Tree Algorithm

The decision tree algorithm is a popular method in the fields of machine learning and data mining. This algorithm is used for classification and regression tasks, which models decision making in the form of a tree. This tree structure helps in mapping the relationships between variables in a form that is easy to understand and interpret. The decision tree algorithm works by dividing the dataset into subsets based on certain feature values. This process continues until the subset contains homogeneous data or meets certain stopping criteria. Each internal node in the tree represents a

test on a feature. each branch represents a test result. and each leaf represents a class label or target value.



**Figure 1. Illustration of Decision Trees**

Source : Addapto.com

The root node represents the question or problem to be solved. Then the branch is a decision path, which will later lead to several decisions or internal nodes. Each decision tree can have several internal nodes as alternative answers or decisions. Internal nodes can also have other branch nodes, namely leaf nodes, which will represent the final decision. The following is a simple illustration of the Decision Trees algorithm:

1. Data preparation:
 

Training data must be prepared properly before being entered into the Decision Tree algorithm. This includes selecting relevant features, normalizing data, and dividing data into appropriate classes.
2. Feature Selection
 

The first stage is selecting the features or attributes that will be used to divide the dataset. The selection of these features is usually based on certain criteria that measure how well the features divide the data into different classes.
3. Calculation of Division Criteria
 

There are several criteria that can be used to determine the best features for each division. This criterion measures how well the division is in terms of homogeneity or class irregularity in the resulting subset of data. Some general criteria include:

  - Entropy and Information Gain: Measures the reduction in uncertainty or entropy from division.
  - Gini Impurity: Measures the probability of misclassification if a random data set is classified based on the class distribution in the subset.
  - Gain Ratio: Addresses Information Gain bias towards features with many unique values.
4. Dataset division
 

After the best features are selected based on the division criteria, the dataset is divided into subsets based on the value of the feature. Each subset becomes a branch in the decision tree.
5. Formation of Nodes and Leaves
  - Nodes: Each feature selected to divide the data into subsets will become a node in the tree. This node will have a branch that shows the results of testing on that feature.
  - Leaf: If a data subset is sufficiently homogeneous or meets certain stopping criteria (for example, the amount of data in the subset is too small or there is no significant increase in the splitting criteria), then the node becomes a leaf. This leaf will contain class labels for classification or average values for regression.
6. Repetition of the Process for Each Subset

The process of selecting features, calculating division criteria, and dividing the dataset is repeated for each subset of data resulting from the previous division. This process continues recursively until the termination condition is met.

#### 7. Termination Conditions

There are several conditions that can be used to stop tree formation:

- All data in a subset has the same class label. There are no features left for further division.
- Splitting does not provide a significant improvement in splitting criteria. The amount of data in the subset is too small for further division.

#### 8. Pruning (trimming) trees

Once the tree is established, an additional step that is often taken is pruning. Pruning aims to reduce overfitting by pruning branches of the tree that do not provide a significant improvement in model performance on test data. There are two types of pruning:

- Pre-pruning: Stops tree formation early based on certain criteria.
- Post-pruning: Cutting tree branches after the tree is fully formed.

### A.2. Naïve Bayes algorithm

The Naïve Bayes algorithm is a simple but very effective classification method, especially for text classification tasks such as spam filtering, sentiment analysis, and document recognition (Firdaus and Mukhlis, 2020). This algorithm is based on Bayes' theorem with the assumption of strong (naive) independence between features. The Naive Bayes algorithm works on the principle of probability and uses Bayes' theorem to calculate the probability that an example belongs to a certain class. The formula for Bayes' theorem is as follows:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

Information:

X = Sample data that has an unknown class (label)

C = Hypothesis that X is a data class (label)

P(C) = Probability of hypothesis C

P(X) = Probability of observed sample data (probability C)

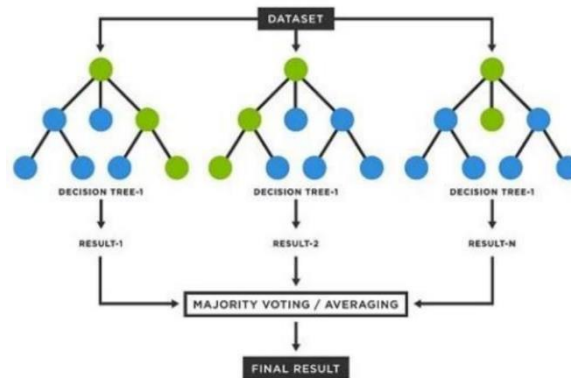
P(X|C) = Probability based on conditions in the hypothesis.

The Follow of the Naïve Bayes method is as follows:

1. Calculate the probability value of a new case from each hypothesis with the class (label) in P(C<sub>i</sub>).
2. Calculate the accumulated probability value of each class P(X|C<sub>i</sub>)
3. Calculate the value of P(X|C<sub>i</sub>) x P(C<sub>i</sub>)
4. Determine the class of the new case.

### A.3. Random Forest Regression Algorithm

The Random Forest method is a method that is similar to the Decision Tree method. This method is one of the most widely used algorithms because of its accuracy, simplicity and flexibility. The fact that it can be used for classification tasks and regression, combined with its non linear nature, makes it highly adaptable to a variety of data and situations.

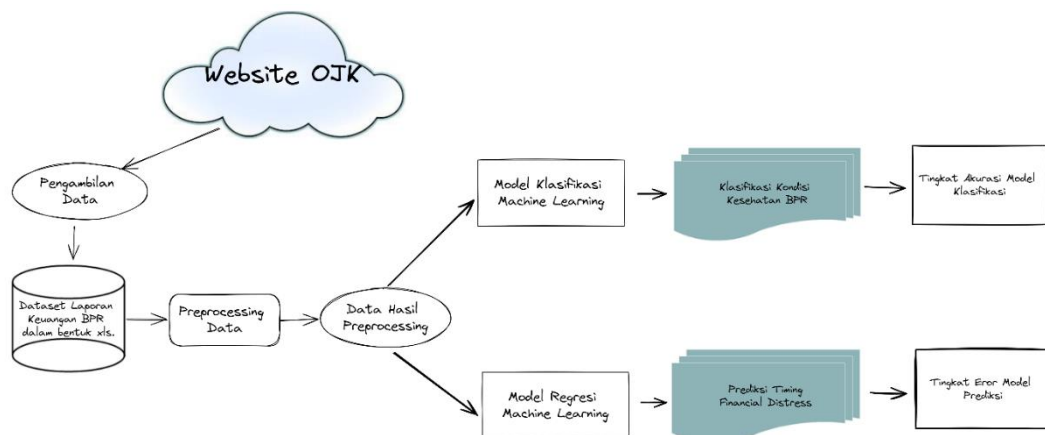


**Figure 2. Illustration of Random Forest Regression**

Source : Lili Zhu (2021)

The image above shows how the Random Forest Regression algorithm works. The available data is divided into different subsets for each Decision Tree created. In each Decision Tree, the selection of variables used to build the decision tree structure is done randomly. After all Decision Trees have been created, the prediction results from all Decision Trees are combined by taking the average of all the resulting prediction values. Random Forest Regression has several advantages, including the ability to handle overfitting problems, being able to deal with complex and unstructured data, and being able to overcome the shortcomings of Decision Trees in cases that are sensitive to data changes. However, Random Forest Regression also has weaknesses, namely high model complexity and long computing time compared to other machine learning algorithms such as Linear Regression or Decision Tree.

The design of the research architecture for predicting the timing of financial distress at people's credit banks in Indonesia utilizing two-stage machine learning techniques classification and regression can be seen in Figure 3.



**Figure 3. Research Architecture**

Source : Researcher (2024)

### B. Datasets

In this research, researchers used data from 150 Rural Bank Financial Ratio Data consisting of 50 Bankrupt Rural Banks financial ratio data and 100 non-bankrupt Rural Banks. Data can be obtained from sources such as financial reports, reports business activities, and Rural Banks publications on the OJK website. Data analysis was carried out 2 years in a row before being

declared bankrupt. By taking financial report data 4 quarters every year. The data required includes financial ratios such as APYDAP. NPL. ROE. BOPO. NIM. LDR. IRR ratios. In the process of collecting the Rural Banks Financial Ratio dataset. researchers used the following criteria:

- Datasets must be available and easily accessible for sustainable use. in this case. financial reports published on the OJK website.
- The dataset must include all the information required for analysis during the observation period.
- The dataset must be consistent in the format and content of the data used so that it is easy to process and use.

**Table 1. Rural Banks Financial Ratio Dataset that has been collected**

NO	STATUS	MP	APYDAP	NPL	ROE	BOPO	NIM	LDR	IRRR
1	BANKRUPTCY	B-3	49,11	40,03	-151,16	238,90	0,28	106,10	109,59
		B-6	49,11	29,57	-126,40	166,77	-10,50	104,81	54,25
		B-9	42,16	35,00	-89,30	136,89	-3,13	78,45	57,69
		B-12	38,37	33,69	-36,75	89,56	-0,28	79,88	95,83
		B-15	35,12	28,84	-45,67	133,68	2,56	85,70	102,67
		B-18	31,16	11,97	-41,23	145,74	3,89	96,00	134,98
		B-21	30,12	20,06	-40,12	87,89	4,89	104,92	167,00
		B-24	11,41	14,61	-38,88	121,24	5,07	91,28	201,89
2	BANKRUPTCY	B-3	52,65	58,68	-210,97	313,16	7,16	179,11	148,91
		B-6	39,78	51,82	-145,85	124,25	5,38	125,83	151,32
		B-9	27,89	37,00	-34,98	110,00	4,32	128,00	167,00
		B-12	11,59	20,98	-4,77	109,32	2,18	145,48	183,58
		B-15	10,45	19,80	-16,78	99,87	4,89	94,32	192,34
		B-18	9,45	18,70	-17,86	98,78	6,98	98,99	202,45
		B-21	8,76	16,70	-20,32	94,56	7,13	101,45	256,34
		B-24	9,31	15,66	-16,03	91,45	4,26	102,32	290,97

Source : OJK

## RESULTS AND DISCUSSION

The Rural Banks financial ratio dataset that has been collected is pre-processed first. Data preprocessing includes removing incomplete data. filling in missing data. performing data transformations such as normalization or standardization of data. and dividing data into training data and test data. In this research. bankruptcy prediction models are categorized into time variations. namely 3 (three) month. 6 (six) month. 9 (nine) month. 12 (twelve) month. 15 (fifteen) month. 18 (eighteen) month prediction models. month. 21 (twenty one) months. 24 (twenty four) months before bankruptcy. This time variation was chosen with the consideration of knowing earlier the bankruptcy signals that might occur and the quarterly time span because the publication of the Rural Credit Bank's financial report position is carried out in quarterly periods. The following is the distribution of data on the financial condition of Rural Banks which can be seen in table 2.

**Table 2. Distribution of Data on Financial Conditions of Rural Banks.**

Category	Count
Very Healthy	569
Healthy	131
Quite Healthy	47
Unwell	176

Distress	127
Total	1050

Source : Data Processing (2024)

After weighting each financial ratio, a value conversion is carried out for each financial ratio. Then an average is carried out to label the financial condition predicate for each quarter. The RURAL BANKS financial ratio dataset from preprocessing can be accessed at <https://doi.org/10.6084/m9.figshare.22558669.v1>

The results of the research show that the level of accuracy in predicting financial distress at Rural Banks in Indonesia using machine learning techniques is shown in Figure 4. During the training period, the weight of each epochs saved and selected the highest validation f1-score and the lowest error. After that, the selected weights are used to predict the test data. Figure 4(b) shows the performance of the Decision Tree which is more accurate than Naïve Bayes in figure 4(a).

Hasil:	precision	recall	f1-score	support		precision	recall	f1-score	support	
CUKUP SEHAT	0.57	0.67	0.62	12		0	0.36	0.33	0.35	12
DISTRESS	0.69	0.71	0.70	35		1	0.81	0.63	0.71	35
KURANG SEHAT	0.75	0.72	0.74	54		2	0.68	0.80	0.74	54
SANGAT SEHAT	0.99	0.96	0.97	172		3	0.95	0.90	0.92	172
SEHAT	0.87	0.95	0.91	42		4	0.65	0.81	0.72	42
accuracy			0.88	315				0.82	315	
macro avg	0.77	0.80	0.79	315		0.69	0.69	0.69	315	
weighted avg	0.88	0.88	0.88	315		0.83	0.82	0.82	315	

(a) Decision Tree f1-score accuracy results

(b) Naïve Bayes f1-score accuracy results

#### Figure 4. Result of Decision Tree and Naïve Bayes

Source : Data Processing (2024)

Figure 4(a) shows the results of evaluating the machine learning classification model on validation data using the Decision Tree Algorithm, obtaining an f1-score accuracy rate of 88%. Meanwhile, Figure 4(b) shows the results of machine learning classification evaluation using the Naïve Bayes algorithm with standard scaler features, getting an accuracy rate of 82%. This shows that machine learning classification using the Decision Tree Algorithm is more effective to apply compared to the Naïve Bayes Algorithm in classifying datasets. In the evaluation results of the Random Forest Regression regression model on validation data using the mean absolute error feature, the error result was 1.2 months and using the mean absolute percentage error feature showed a result of 3%.

### CONCLUSION

Based on the results of research on the application of classification data mining using the Decision Tree Algorithm on the financial ratio dataset of Rural Banks in predicting the timing of financial distress, a result was obtained that the accuracy value for predicting the classification of the health condition of Rural Banks was 88%. while the classification of Rural Bank health conditions using the Algorithm Naïve Bayes shows accuracy results of 82%. This shows that the classification model using Machine Learning techniques with the Decision Tree Algorithm is more effective than using the Naïve Bayes Algorithm.

The results of the prediction of financial distress timing regression using the Random Forest Regression technique show that the results of the model evaluation using the mean absolute error were 1.2 months, while using the mean absolute percentage error the value was 3%. Based on the prediction results, it can be concluded that the model used in this research is quite accurate in predicting financial distress at Rural Banks in Indonesia. The conclusion of this research is that the use of two-stage classification and regression machine learning techniques using machine learning techniques in predicting the timing of financial distress is considered good. This can be seen from the fairly high level of accuracy and small error rate produced by the model. But this needs to be reviewed from the perspective of complexity and number of datasets.



Predicting the timing of financial distress at Rural Banks can help related parties, such as regulators, investors and management of Rural Banks, in taking preventive action or solving financial problems before it is too late. The two-stage classification and regression method used in this research can be applied to Rural Banks and other financial institutions to predict the possibility of financial distress and estimate the time when these financial problems will occur.

Researchers have discussed the use and calculation of the accuracy of financial distress timing predictions at people's credit banks in Indonesia using two-stage machine learning techniques, classification and regression using techniques Machine Learning. In future research, it is hoped that comparisons can be made using other classification and regression methods such as the method K-Nearest Neighbor and ANN as well as more complex amounts of data. So by using many methods you can find out the advantages and disadvantages of each method.

## REFERENCES

- Altman, E. I. (1968). Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), 589–609.
- Ansar, A. (2018). Pengaruh rasio keuangan terhadap financial distress pada perusahaan manufaktur yang listing di BEI periode 2012-2016. STIE Perbanas Surabaya.
- Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83, 405–417.
- Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 71–111.
- Brigham, E. F. (2016). *Financial management: Theory and practice*. Cengage Learning Canada Inc.
- Budiwati, H., Fadah, I., Sukarno, H., & Utami, E. S. (2021). Bankruptcy prediction model conventional bank rural banks in Indonesia. *PalArch's Journal of Archaeology of Egypt/Egyptology*, 18(4), 3790–3802.
- Chang, H. (2019). The application of machine learning models in company bankruptcy prediction. In *Proceedings of the 2019 3rd International Conference on Software and e-Business* (pp. 199–203).
- Firdaus, F., & Mukhlis, A. (2020). Implementasi algoritma Naive Bayes pada data set kualitatif prediksi kebangkrutan. *JURIKOM (Jurnal Riset Komputer)*, 7(1), 15–20.
- Kesumawati, A., & others. (2018). Perbandingan metode Support Vector Machine (SVM) linear, radial basis function (RBF), dan polinomial kernel dalam klasifikasi bidang studi lanjut pilihan alumni UII.
- Kovacova, M., & Kliestikova, J. (2017). Modelling bankruptcy prediction models in Slovak companies. In *SHS Web of Conferences* (p. 1013).
- Meyer, P. A., & Pifer, H. W. (1970). Prediction of bank failures. *The Journal of Finance*, 25(4), 853–868.
- Muflihah, R. (2022). Analisis financial distress Bank Perkreditan Rakyat Syariah (Rural Banks) di Indonesia periode 2019-2020. *La Zhulma: Jurnal Ekonomi dan Bisnis Islam*, 1(1), 17–26.
- Ratna, I., & Marwati, M. (2018). Analisis faktor-faktor yang mempengaruhi kondisi financial distress pada perusahaan yang delisting dari Jakarta Islamic Index tahun 2012-2016. *Jurnal Tabarru' Islamic Banking and Finance*, 1(1), 51–62.
- Sutra, F. M., Mais, R. G., & others. (2019). Faktor-faktor yang mempengaruhi financial distress dengan pendekatan Altman Z-Score pada perusahaan pertambangan yang terdaftar di Bursa Efek Indonesia tahun 2015-2017. *Jurnal Akuntansi dan Manajemen*, 16(01), 34–72.
- Wilopo, R. (2001). Prediksi kebangkrutan bank. *Jurnal Riset Akuntansi Indonesia*, 4(2), 184–198.
- Zhu, L., & Spachos, P. (2021). Support vector machine and YOLO for a mobile food grading system. *Internet of Things*, 13, 100359.