Mitigating challenges: Handling mis sing values and imbalanced data in bankruptcy prediction using machine learning

Ednawati Rainarli[1], Amine Sabek[2]

[1]Department of Informatics Engineering, Universitas Komputer Indonesia, Indonesia
[2]Investment bets and sustainable development stakes in border areas, University of Tamanghasset, Algeria

Abstract

The research on financial distress has become essential because the predicted results can serve as an early warning for managers, investors, and banks. Financial ratios calculated in financial reports can serve as indicators to assess the company's condition. One of the approaches used for bankruptcy prediction is employing machine learning methods. Data requirements with balanced classes and the need to process data with complete parameters/features are prerequisites for building an accurate bankruptcy prediction model. In this study, we employed data balancing techniques such as downsampling and filling missing feature values using the average of nearest neighbors in data preprocessing before training the prediction model. From our experiments, we found that by addressing missing values and balancing the data, the F1 score of the prediction model using Random Forest (RF) improved by 30% compared to not addressing missing data and data imbalance. Although our testing used the Polish company dataset, which may have different characteristics from companies in other countries, the proposed strategies can serve as an initial approach for training datasets of other companies using machine learning methods.

Keyword: bankruptcy prediction, financial distress, imbalanced data, machine learning, missing value

Introduction

In a business environment filled with uncertainty and rapid economic changes, predicting company bankruptcies holds excellent relevance. Bankruptcy prediction is a crucial aspect of risk management. The prediction enables companies and stakeholders to identify potential financial risks in advance. The ability to predict bankruptcies allows companies to recognize financial and operational risks that could lead to insolvency. By understanding these risks early on, companies can take proactive measures to manage them effectively. Additionally, investors, creditors, and stakeholders require predictive information to make well-informed investment decisions. Accurate bankruptcy predictions help them identify companies with high potential bankruptcy risks, guiding wise allocation of financial resources.

The company's financial reports provide rich and comprehensive data. They utilized financial statements to build bankruptcy prediction models. There are three approaches to constructing bankruptcy prediction models: using statistical, soft computing, and theoretical approaches. The study conducted by Altman in 1968 represents an early research effort that employed statistical methods to predict corporate bankruptcy. Altman utilized discriminant analysis methods to construct his model. This method measures the differences between two or more groups based on variables that distinguish these groups. In this context, Altman used financial variables, known as financial ratios, to differentiate between companies that are likely to go bankrupt and those that are not.

Although the Altman Z-Score model has proven effective in many cases, there are several limitations to the Z-Score model. Z-Score is a static model that assesses the financial condition at a specific point in time. However, in the dynamic business world, rapid changes can occur, affecting the company's finances. This model does not incorporate market volatility into its calculations. Stock or bond market fluctuations can significantly impact bankruptcy risk assessment. Some companies might have internal information not available to the public, which can affect the accuracy of bankruptcy predictions.

Due to these limitations, researchers have turned to soft computing techniques such as artificial neural networks (Atiya, 2001), fuzzy logic (Rainarli, Aaron, 2015), Support Vector Machine (SVM) (Barboza, Kimura & Altman, 2017; Rainarli, 2019), ensemble methods (Tsai, Hsu & Yen, 2014; Barboza, Kimura & Altman, 2017), and genetic algorithms (Bateni, Asghari, 2020). The advantage of soft computing lies in its ability to handle uncertainty, complexity, and non-linearity in data. It can model complex relationships between various financial and non-financial variables, accounting for market fluctuations and industry dynamics. By employing these techniques, research on bankruptcy prediction can leverage machine learning capabilities to identify complex and non-linear patterns in financial and operational company data. Thus, soft computing techniques offer more flexible and adaptable solutions for the dynamic business environment. The study by Korol (2012) indicates that bankruptcy prediction with a statistical model using the Discriminant Analysis Model resulted in an accuracy of 77.77%, whereas employing soft computing methods such as Neural Network or Fuzzy Logic yielded the same accuracy of 87.03%. This difference represents a 10% improvement when compared to the Discriminant Analysis Model.

While developing predictive models using soft computing approaches can predict bankruptcy, there are challenges in building prediction models, especially with machine learning approaches. Challenges such as missing values, imbalanced data, selecting significant features, and using accurate model evaluation become hurdles in constructing prediction models with machine learning. Therefore, in this study, we evaluate various missing value imputation techniques to observe their impact on prediction model formation, compare the effects of downsampling to address imbalanced data, assess multiple machine learning classification methods in bankruptcy prediction, and employ the F1 score to validate model performance. Understanding the profound need for financial statement-based bankruptcy prediction and addressing challenges related to missing values and imbalanced data using machine learning techniques. The purpose of this research is to construct an accurate bankruptcy prediction model using an imbalanced dataset with missing data. Additionally, this study provides valuable insights for scientific knowledge and business practices relevant to financial risk management. The objectives of this research include analyzing the best methods for handling missing data in bankruptcy prediction cases, examining the impact of dataset balancing on the development of bankruptcy prediction models, and evaluating suitable machine learning classification methods for constructing bankruptcy prediction models.

The structure of this manuscript is as follows, beginning with the problem background on the need for soft computing in building bankruptcy prediction models. We explain the challenges of using machine learning and end up with our proposed solution. Section two discusses the review of related machine learning research and its developments. We outline the framework for bankruptcy prediction in section three, followed by the discussion of results in section four. In conclusion, we summarize the experimental findings and end with a suggestion for further development.

Related work

The research on bankruptcy prediction is valuable as an early warning for managerial, investment, and creditor decision-making. Sun et al. (2014) categorize bankruptcy prediction into two approaches: statistical and artificial intelligence. Various statistical methods used include linear discriminant analysis (LDA) (Altman, 1968; Khan, 2018), multivariate discriminate analysis (MDA) (Lee, Choi, 2013; Mihalovič, 2016), quadratic discriminant analysis (QDA) (Brîndescu-Olariu, Goleţ, 2013), logistic regression (logit) (Mihalovič, 2016; Khan, 2018; Pavlicko, Mazanec, 2022), and factor analysis (FA) (Cultrera, Croquet & Jospin, 2017). The statistical model prediction must fulfill the assumption of independent variables, data distribution following a normal distribution, and equal covariance matrices. If the data fails to meet the requirements, the model generated from statistical approaches becomes biased (Sun et al., 2014). Therefore, developing the bankruptcy model using machine learning approaches became imperative.

Researchers have employed machine learning methods such as Gaussian Process Regression (GPR), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest (RF), and AdaBoost to construct prediction models. Among these methods, Sabek, Horak (2023) used GPR to predict financial distress. The best model was extracted after optimizing the hyperparameters. Remarkably, this fine-tuned model demonstrated outstanding performance. Sabek (2023) conducted an experiment where two distinct ANNs types were pitted against Logistic Regression (LR) to determine if ANNs consistently outperformed regression. Ultimately, his findings led to the conclusion that not all ANNs are superior to regression when it comes to predicting financial health. A study by Barboza, Kimura & Altman (2017) indicated that Random Forest achieved the best performance. Conversely, Danenas, Garsva (2015) optimized SVM to build bankruptcy prediction models. Two challenges arise concerning the construction of prediction models using machine learning approaches: firstly, the issue of imbalanced data (Cleofas-Sánchez et al., 2016), and secondly, the existence of incomplete parameter values in the dataset.

Building prediction models with machine learning requires a substantial amount of data from each class, i.e., the bankrupt and non-bankrupt classes. Bankrupt cases are significantly fewer than non-bankrupt cases. This condition leads to an imbalanced learning process. Learning from imbalanced data tends to be biased because of the model toward recognizing the dominant class. To address the imbalance condition, techniques such as data addition to the minority class (upsampling) and data reduction (downsampling) or their combination are employed. Researchers must limit these processes, as excessive data addition can lead to model overfitting, and reducing data from the significant class can result in misclassification. Additionally, another challenge in utilizing data for training the machine learning model is the need to have complete financial ratio parameters/features in the dataset.

Based on our literature review, this study focuses on data manipulation to overcome data imbalance and missing values. Our testing aims to evaluate the performance achieved when we balance the data and fill in the missing values before training. We employ Support Vector Machine (SVM), Random Forest (RF), and Naïve Bayes (NB) classification methods to determine the optimal approach. There are three main processes for building a bankruptcy prediction model, namely: the data preprocessing stage, the training of the prediction model, and testing the bankruptcy prediction model with unseen data. In this research, we conducted data balancing and data imputation processes as part of the data preprocessing. Our data balancing strategy was employed during the training of the prediction model. After the prediction model was developed, we tested it against new, unseen data and measured its success using precision, recall, and F-measure values.

Proposed Method

To obtain a model capable of predicting bankruptcy, we underwent several processes, as depicted in Figure 1. The initial step involved preprocessing the Polish Company dataset. The dataset consisted of 10,503 companies. The bankrupt companies were analyzed between 2000 and 2012, while the still-operating companies were evaluated from 2007 to 2013. To simplify the process, we combined the evaluations from each year as independent conditions, resulting in a total dataset of 43,405 instances. Each instance provided information on the values of financial ratios, comprising 64 financial ratios

(Tomczak, 2016). The Polish Company dataset is a public dataset. This data has incomplete financial ratio values and an imbalanced distribution of bankrupt and non-bankrupt class propositions. We used The Polish Company dataset to assess the success of the bankruptcy prediction model. Although the company data is from the years 2000-2013, developing a bankruptcy model with this dataset can serve as a baseline if implemented on newer datasets. Additionally, in training prediction models using machine learning methods, the more data involved in model training, the better the model can predict bankruptcy for unseen data.

Table 1 provides an overview of the distribution of data for 64 financial ratios. Each financial ratio contains missing values, resulting in fewer values for each ratio than the total dataset, which amounts to 43,405. The minimum and maximum values for each financial ratio vary significantly. Some ratios, such as X5, X15, X27, X43, X44, X55, X62, have a broad range, while others, like X29, exhibit a narrow range. Therefore, we introduced a data normalization process to ensure consistent data ranges across all financial ratios.

There were three stages in our preprocessing. We began by filling in missing values, removing duplicate data, and performing downsampling. Filling missing data was necessary because out of the 43,405 datasets, only 19,737 instances had complete features. Refrain from discarding incomplete data would result in discarding over 50% of the data. We tested three techniques for filling in missing values. The first technique involved using the median value. For each financial ratio feature, we sorted the values from the smallest to the largest and determined the median value. Then, we used the median to fill in the missing values in the dataset. The second technique utilized the modus value. We used the modus value of each feature to fill in the missing values. The third technique involved using the nearest neighbors' values. Determining values based on nearest neighbors involved selecting the number of neighbors to calculate the missing value and then computing the average value from the nearest neighbors' data. The average value we used to fill in the missing values.

We removed duplicate data to prevent redundant training of instances with similar characteristics. Eliminating duplicate data also helped prevent overfitting during the machine learning model training. Furthermore, the imbalance between bankrupt and non-bankrupt companies posed a challenge during model training. The Polish Company dataset recorded 2,091 bankruptcy cases compared to 41,314 non-bankrupt cases, resulting in a class ratio 1:20. Review findings (Sun et al., 2014) highlighted that balancing data did not always yield optimal performance during testing. However, data balancing was crucial in machine learning to prevent overfitting. Therefore, in our testing, we compared the model's performance using data balancing techniques and without them. We employed downsampling to balance the data.

Table 1: Overview of statistical information from 64 financial ratios of The Polish Company dataset

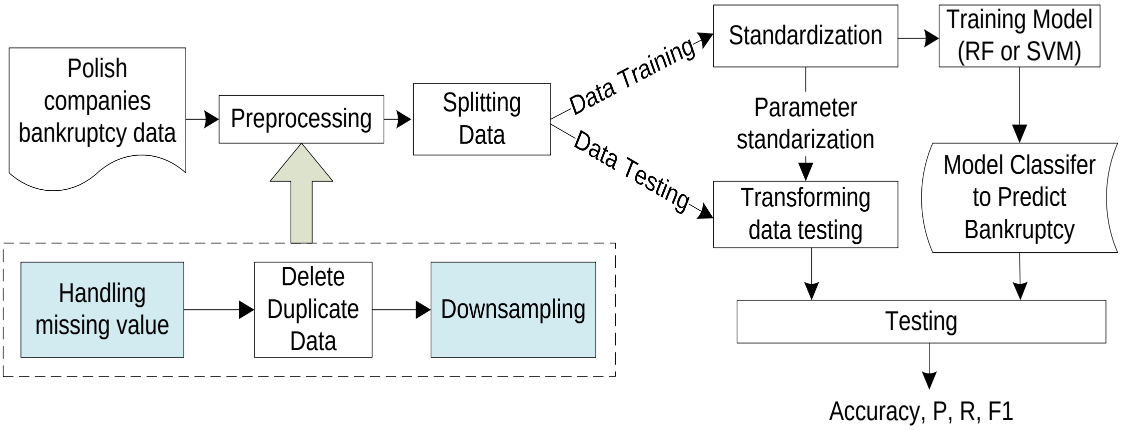
Financial Ratios	count	mean	std	min	median	max
X1 net profit / total assets	43,397	0.04	2.99	-463.89	0.05	94.28
X2 total liabilities / total assets	43,397	0.59	5.84	-430.87	0.47	480.96
X3 working capital / total assets	43,397	0.11	5.44	-479.96	0.20	28.34
X4 current assets / short-term liabilities	43,271	6.31	295.43	-0.40	1.57	53,433.00
X5 [(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365	43,316	-385.35	61,243.03	-11,903,000.00	-1.03	1,250,100.00
X6 retained earnings / total assets	43,397	-0.06	7.20	-508.41	0.00	543.25
X7 EBIT / total assets	43,397	0.09	5.71	-517.48	0.06	649.23
X8 book value of equity / total liabilities	43,311	12.64	505.89	-141.41	1.07	53,432.00
X9 sales / total assets	43,396	2.65	62.93	-3.50	1.20	9,742.30
X10 equity / total assets	43,397	0.63	14.67	-479.91	0.51	1,099.50
X11 (gross profit + extraordinary items + financial expenses) / total assets	43,361	0.13	5.31	-463.89	0.08	681.54
X12 gross profit / short-term liabilities	43,271	1.13	67.59	-6,331.80	0.17	8,259.40
X13 (gross profit + depreciation) / sales	43,278	0.81	86.94	-1,460.60	0.07	13,315.00
X14 (gross profit + interest) / total assets	43,397	0.09	5.71	-517.48	0.06	649.23
X15 (total liabilities * 365) / (gross profit + depreciation)	43,369	1,991.89	96,431.93	-9,632,400.00	846.26	10,236,000.00
X16 (gross profit + depreciation) / total liabilities	43,310	1.41	68.52	-6,331.80	0.25	8,259.40
X17 total assets / total liabilities	43,311	13.80	507.32	-0.41	2.12	53,433.00
X18 gross profit / total assets	43,397	0.10	5.74	-517.48	0.06	649.23
X19 gross profit / sales	43,277	0.16	48.69	-1,578.70	0.04	9,230.50
X20 (inventory * 365) / sales	43,278	243.02	37,545.17	-29.34	35.15	7,809,200.00
X21 sales (n) / sales (n-1)	37,551	3.88	228.67	-1,325.00	1.05	29,907.00
X22 profit on operating activities / total assets	43,397	0.11	5.16	-431.59	0.06	681.54
X23 net profit / sales	43,278	0.14	48.33	-1,578.70	0.03	9,230.50
X24 gross profit (in 3 years) / total assets	42,483	0.27	7.99	-463.89	0.16	831.66
X25 (equity - share capital) / total assets	43,397	0.39	12.89	-500.93	0.38	1,353.30
X26 (net profit + depreciation) / total liabilities	43,310	1.26	66.22	-6,331.80	0.22	8,262.30
X27 profit on operating activities / financial expenses	40,641	1,107.90	35,012.37	-259,010.00	1.08	4,208,800.00

Financial Ratios	count	mean	std	min	median	max
X28 working capital / fixed assets	42,593	6.00	153.47	-3,829.90	0.47	21,701.00
X29 logarithm of total assets	43,397	4.01	0.83	-0.89	4.01	9.70
X30 (total liabilities - cash) / sales	43,278	7.37	814.49	-6,351.70	0.22	152,860.00
X31 (gross profit + interest) / sales	43,278	0.18	48.75	-1,495.60	0.04	9,244.30
X32 (current liabilities * 365) / cost of products sold	43,037	1,162.62	95,593.56	-9,295.60	78.33	17,364,000.00
X33 operating expenses / short-term liabilities	43,271	8.64	118.99	-19.20	4.63	21,944.00
X34 operating expenses / total liabilities	43,311	5.41	120.98	-1,696.00	1.97	21,944.00
X35 profit on sales / total assets	43,397	0.11	4.78	-431.59	0.06	626.92
X36 total sales / total assets	43,397	2.91	62.98	0.00	1.64	9,742.30
X37 (current assets - inventories) / long-term liabilities	24,421	105.09	3,058.43	-525.52	3.10	398,920.00
X38 constant capital / total assets	43,397	0.72	14.75	-479.91	0.61	1,099.50
X39 profit on sales / sales	43,278	-0.29	39.26	-7,522.00	0.04	2,156.50
X40 (current assets - inventory - receivables) / short-term liabilities	43,271	2.15	56.03	-101.27	0.18	8,007.10
X41 total liabilities / ((profit on operating activities + depreciation)* (12/365))	42,651	7.72	1,398.84	-1,234.40	0.09	288,770.00
X42 profit on operating activities / sales	43,278	-0.14	15.99	-1,395.80	0.04	2,156.80
X43 rotation receivables + inventory turnover in days	43,278	1,074.12	147,218.77	-115,870.00	99.40	30,393,000.00
X44 (receivables * 365) / sales	43,278	831.11	110,050.97	-115,870.00	54.77	22,584,000.00
X45 net profit / inventory	41,258	14.83	2,428.24	-256,230.00	0.28	366,030.00
X46 (current assets - inventory) / short-term liabilities	43,270	5.43	295.36	-101.26	1.03	53,433.00
X47 (inventory * 365) / cost of products sold	43,108	357.84	33,146.34	-96.11	38.13	6,084,200.00
X48 EBITDA (profit on operating activities - depreciation) / total assets	43,396	0.03	5.10	-542.56	0.02	623.85
X49 EBITDA (profit on operating activities - depreciation) / sales	43,278	-0.48	45.15	-9,001.00	0.01	178.89
X50 current assets / total liabilities	43,311	5.84	307.38	-0.05	1.22	53,433.00
X51 short-term liabilities / total assets	43,397	0.48	5.44	-0.19	0.34	480.96
X52 (short-term liabilities * 365) / cost of products sold)	43,104	6.48	639.89	-25.47	0.21	88,433.00
X53 equity / fixed assets	42,593	23.77	1,213.80	-3,828.90	1.21	180,440.00
X54 constant capital / fixed assets	42,593	24.65	1,220.88	-3,828.90	1.38	180,440.00
X55 working capital	43,404	7,672.19	70,053.10	-1,805,200.00	1,088.35	6,123,700.00
X56 (sales - cost of products sold) / sales	43,278	-26.22	5,327.86	-1,108,300.00	0.05	293.15

Financial Ratios	count	mean	std	min	median	max
X57 (current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)	43,398	-0.01	13.67	-1,667.30	0.12	552.64
X58 total costs /total sales	43,321	30.03	5,334.45	-198.69	0.95	1,108,300.00
X59 long-term liabilities / equity	43,398	1.33	122.10	-327.97	0.01	23,853.00
X60 sales / inventory	41,253	448.09	32,345.60	-12.44	9.79	4,818,700.00
X61 sales / receivables	43,303	17.03	553.05	-12.66	6.64	108,000.00
X62 (short-term liabilities *365) / sales	43,278	1,502.33	139,266.70	-2,336,500.00	71.33	25,016,000.00
X63 sales / short-term liabilities	43,271	9.34	124.18	-1.54	5.09	23,454.00
X64 sales / fixed assets	42,593	72.79	2,369.34	-10,677.00	4.28	294,770.00

Source: Tomczak, (2016) and own processing for statistical information

Figure 1. Illustration of bankruptcy detection model addressing missing values and overfitting.



Source: Own model

The stages of predictive model training, as depicted in Figure 1, were as follows:

1) Data preprocessing: This included filling in missing values, removing duplicate data, and balancing training data.
2) Splitting data: We split the data into training and testing sets.
3) Data normalization: This involved transforming the feature values of the training data into standardized values.
4) Training: This process included fitting hyperparameters using Stratified Cross Validation. We used accuracy, precision, recall, and F1 score as model evaluation metrics. The model employed RF and SVM to determine the best classifier model for bankruptcy prediction, specifically for the Polish dataset.

Once we established the model, we tested the test data to assess the model's generalization ability in predicting testing data from the Polish dataset.

There are four possible outcomes when classifying companies into bankrupt and non-bankrupt categories, namely:

1) A company correctly predicted as belonging to the bankrupt class. This event is called True Positive (TP).
2) A company correctly predicted as belonging to the non-bankrupt class. This event is referred to as False Positive (FP).
3) A company that should have been classified into the bankrupt class but was predicted to be into the non-bankrupt class. This is known as False Positive (FP).
4) A company that should have been classified into the non-bankrupt class but was predicted to be into the bankrupt class. This is called False Negative (FN).

According to Dalianis (2018), the precision, recall, and F1 score are calculated using equations (1), (2), and (3), respectively.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

Results

We conducted three tests. The first one was to observe the impact of using three techniques for filling in missing data. The second test compared the effect of changing the data proportions in the downsampling process on the bankruptcy prediction model's performance. The third test involved comparing the model's performance based on the classification methods. We carried out these processes step by step. The optimal conditions identified in each test were used for subsequent testing, resulting in the final model tested being the best predictor for bankruptcy on the Polish dataset.

1. Classification results analysis with the filling techniques for missing data

To analyze the influence of adding data through missing data filling, we conducted three events: filling data using median, mode, and nearest neighbors. We compared the measurement results with data without missing values, meaning we only used complete data and deleted incomplete data. After removing duplicate data, we train the prediction model using the RF algorithm. Balancing data was not applied in this test. The purpose was to observe the impact of using strategy to fill the missing value on the prediction model's performance.

Referring to the accuracy values, Table 2 indicates that the bankruptcy prediction model with only complete data achieves higher accuracy than the model with filled missing data. However, when considering the F1 score, the model with missing values differs from that using missing value filling strategies. The reason is that the non-bankrupt data trained is minor in the model without missing values than the model with missing value-filling strategies. This result aligns with the findings of Zahin, Ahmed & Alam (2018), indicating that missing values in data can affect the efficiency of classification models due to the loss of information from those features. Among the three filling methods, the approach using nearest neighbors (NN) proved the most effective in completing bankruptcy prediction data. The model's performance results were reasonable because the NN strategy filled the data locally. Table 2 demonstrates that the F1 score for the bankrupt class improved by 41%.

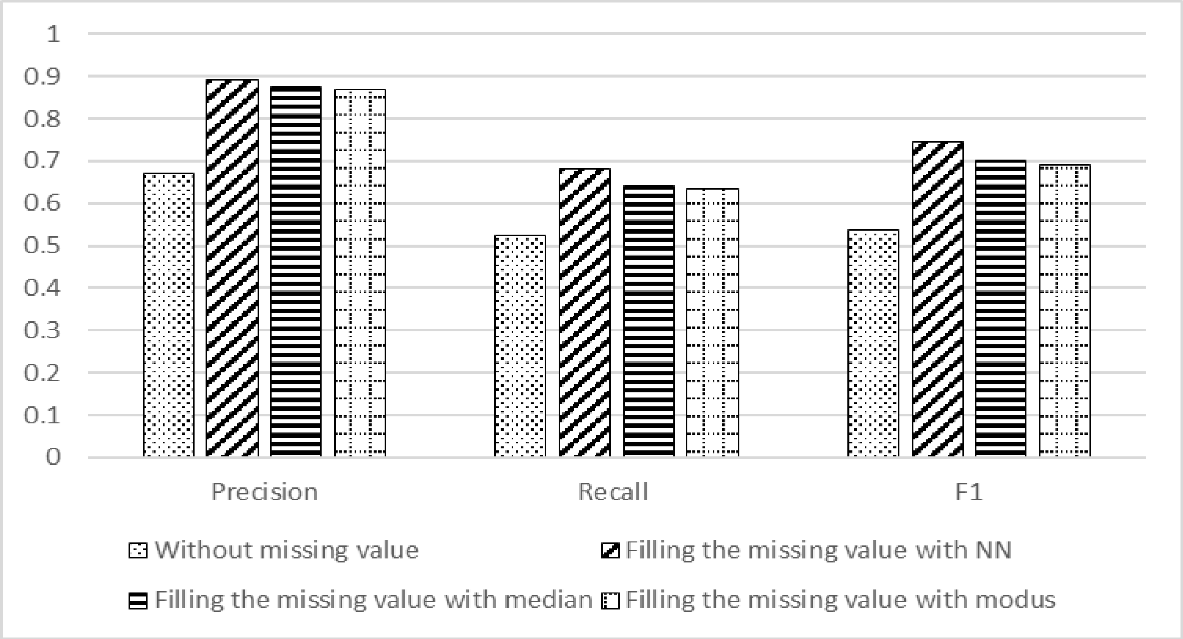
Table 2: Comparison of bankruptcy prediction model performance using RF with missing value strategies

Strategy	Accuracy	F1 score		Average F1
		Bankrupt	Non-bankrupt	
Without missing value	0.9748	0.0859	0.9872	0.5366
Filling the missing values with NN	0.9656	0.5050	0.9821	0.7436
Filling the missing value with median	0.9621	0.4227	0.9804	0.7015
Filling the missing value with modus	0.9611	0.3995	0.9799	0.6897

Source: Own processing

When comparing the precision, recall, and F1 score values from Figure 2, it is evident that filling in missing values enhances precision and recall. The precision, recall, and F1 scores presented in Figure 2 represent the averages of the bankrupt and non-bankrupt classes. The most significant increase in precision occurred for the prediction model utilizing missing data filling with nearest neighbors (NN), with an improvement of 21%. However, the recall values were less substantial than the increase in recall. This phenomenon arises because of the imbalance between bankrupt and non-bankrupt class data. Therefore, in the next experiment, we will evaluate the impact of data balancing on the performance of the bankruptcy prediction model.

Figure 2. Comparison of Precision, Recall, and F1 scores for different filling strategies of missing data



Source: Own processing

2. The impact of downsampling on the performance of the bankruptcy prediction model

Table 3 illustrates the downsampling proportions we employed to balance the data. Given the 2,901 bankrupt data points, we utilized 2,100 data points for the non-bankrupt class. We varied the proportion of non-bankrupt data points. The F1 score for the bankrupt class increased when we reduced the number of non-bankrupt data. We reduced the number of non-bankrupt classes until it was like those of bankrupt classes. Kotsiantis, Kanellopoulos & Pintelas (2006) stated that data imbalance causes classification models to tend to recognize more classes with more significant numbers, in this case, the non-bankrupt class. This result explains why accuracy cannot be used as a reference when fitting a classification model to imbalanced data (Sun et al., 2014).

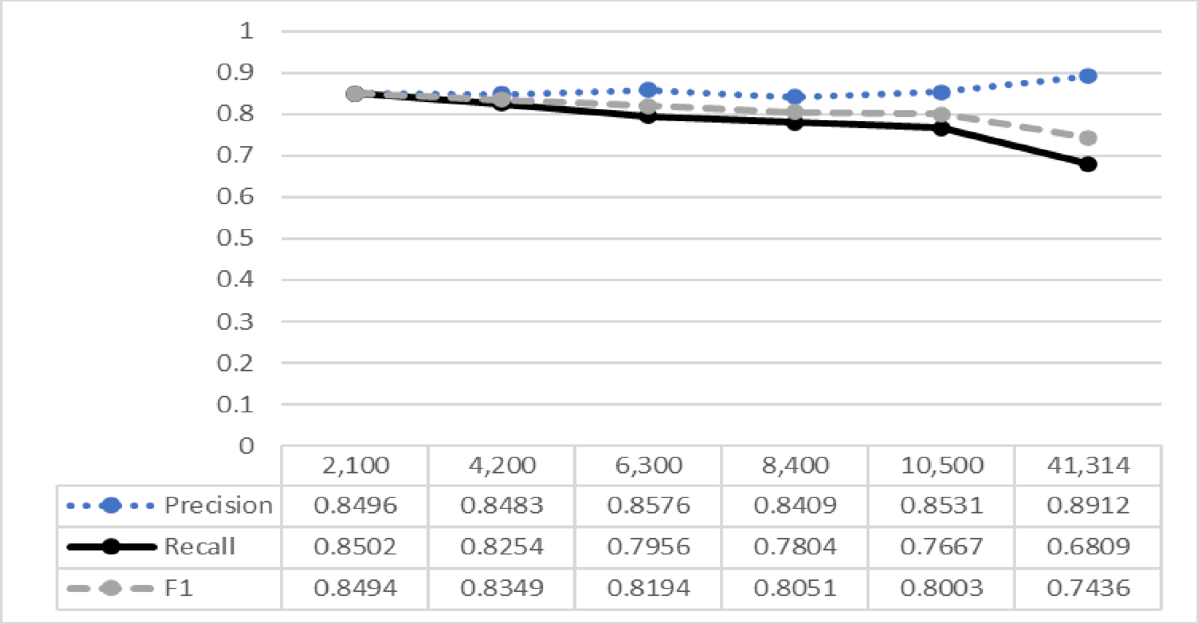
Table 3. Comparison of classification model performance in bankruptcy prediction with down sampling ratios in the RF classification method

Number of data		Accuracy	F1 score		Average F1
Bankrupt	Non-Bankrupt		Bankrupt	Non-bankrupt	
2,091	2,100	0.8494	0.8470	0.8518	0.8494
2,091	4,200	0.8573	0.7741	0.8957	0.8349
2,091	6,300	0.8744	0.7199	0.9190	0.8194
2,091	8,400	0.8878	0.6782	0.9320	0.8051
2,091	10,500	0.9049	0.6559	0.9448	0.8003
2,091	41,314	0.9656	0.5050	0.9821	0.7436

Source: Own processing

Figure 2 shows that by adding non-bankrupt class data, the recall consistently decreases while increasing the precision value. Consequently, when referring to the F1 score, a balanced data condition is the best predictive model, even though its precision could be better than the imbalanced data model. Sun et al. (2014) provides insights on handling imbalanced data. Bankruptcy cases are indeed rare compared to non-bankrupt companies. Therefore, when deciding on using the model, we need to consider the following condition: it is better to predict a bankrupt company as non-bankrupt than vice versa incorrectly. Hence, a high recall value becomes crucial to maintain.

Figure 3. Comparison of precision, recall, and F1 score values with changes in the number of data points in the non-bankrupt class.



	2,100	4,200	6,300	8,400	10,500	41,314
Precision	0.8496	0.8483	0.8576	0.8409	0.8531	0.8912
Recall	0.8502	0.8254	0.7956	0.7804	0.7667	0.6809
F1	0.8494	0.8349	0.8194	0.8051	0.8003	0.7436

Source: Own processing

3. Analysis of classification methods for bankruptcy prediction

We employed three classification methods: Naïve Bayes (NB) as a baseline model, Support Vector Machine (SVM), and Random Forest (RF). In our model fitting process, we utilized the Grid Search method in Python to fine-tune the models based on their hyperparameter values, especially for SVM and RF. For SVM, we tuned parameters such as linear kernel and radial basis function (RBF) kernel, gamma, and C values. In RF, we tuned the number of decision tree estimators and the minimum sample split values to create new trees. Table 4 presents the measurements for these three methods, including comparing their performance without adding missing data.

Among these methods, SVM achieved the highest F1 score in the balanced data group without imputing missing values. SVM performs optimally with smaller, balanced datasets (Danenas and Garsva, 2015). The same trend we observed for the NB method; NB predictive model performance improved when the data was balanced and missing data imputation was not applied. In cases where the data was balanced and handling the missing values, RF emerged as the standout method. The working principle of RF, an ensemble algorithm, explains why RF performs exceptionally well under these conditions (Barboza, Kimura, and Altman, 2017). The study by Barboza, Kimura, and Altman (2017) conducted extensive testing using a large US corporate failure database from 1985 to 2013. They also evaluated predictive models using SVM and found that, for large datasets, RF outperforms SVM. Boosting and bagging algorithms emerged as the most effective choices for bankruptcy prediction.

Table 4. Comparison of classification model performance for RF, SVM, and NB

	Balancing data without missing value				**Missing value and Balancing**			
Classifier	**Accuracy**	**P**	**R**	**F1**	**Accuracy**	**P**	**R**	**F1**
RF	0.7829	0.7835	0.7820	0.7823	0.8494	0.8496	0.8502	0.8494
SVM	0.8217	0.8220	0.8211	0.8213	0.8175	0.8175	0.8180	0.8174
NB	0.6240	0.6783	0.6149	0.5829	0.5219	0.5185	0.5038	0.3892

Source: Own processing.

Table 5 compares precision and recall values for the three classification methods concerning the bankrupt and non-bankrupt classes. SVM with balanced data achieves a balanced recall and precision between the bankrupt and non-bankrupt classes. In contrast, the NB method, despite having a balanced number of bankrupt and non-bankrupt data, fails to detect the non-bankrupt class. This condition is because the NB algorithm requires more non-bankrupt data to recognize the non-bankrupt class. However, adding data with missing values prevents NB from identifying the bankrupt class. We suspect this is due to the non-linear separability of bankrupt data characteristics. Therefore, we cannot use the NB method to build the bankruptcy prediction model. For SVM, the performance tends to decrease when we use data balancing and fill in missing values; however, this decrease is insignificant. The RF method experiences an increase in precision and recall values after balancing data and filling in missing values. The increase in precision and recall occurs for all classes, both bankrupt and non-bankrupt.

Lastly, regarding the limitations of the study, we trained the prediction model using 64 financial ratios as features. If predicting bankruptcy using different financial ratios or introducing additional features beyond the financial ratios, the prediction model needs to be retrained using new training data.

Table 5. Comparison of precision and recall values for bankrupt and non-bankrupt classes in RF, SVM, and NB classification methods.

Classifier	Precision		Recall	
	Bankrupt	Non-bankrupt	Bankrupt	Non-bankrupt
RF with balancing data	0.7770	0.7899	0.8120	0.7520
RF with balancing data (NN)	0.8262	0.8730	0.8688	0.8315
SVM with balancing data	0.8175	0.8264	0.8421	0.8000
SVM with balancing data (NN)	0.7984	0.8365	0.8289	0.8070
NB with balancing data	0.5874	0.7692	0.9098	0.3200
NB with balancing data (NN)	0.5147	0.5223	0.0581	0.9495

Source: Own processing.

Discussion

We first examined the influence of three different techniques for filling in missing data on the bankruptcy prediction model. These techniques included filling with the median, mode, and nearest neighbors. Our results showed that using complete data without any missing values led to higher accuracy in the bankruptcy prediction model. However, when considering the F1 score, which is a better metric for imbalanced data, filling in missing data with the nearest neighbors approach proved to be the most effective. This was because the nearest neighbors strategy filled the data locally, resulting in a substantial 41% improvement in the F1 score for the bankrupt class. These findings align with previous research indicating that missing data can impact the efficiency of classification models due to the loss of valuable information.

To address the issue of data imbalance, we examined the effect of varying the proportions of non-bankrupt data in the downsampling process. We found that reducing the number of non-bankrupt data points improved the F1 score for the bankrupt class, as imbalanced data tends to favor the majority class. The results highlighted the importance of considering metrics other than accuracy when working with imbalanced data, as accuracy alone can be misleading. In this context, the F1 score became a crucial metric for evaluating the predictive model, with a balanced data condition offering the best overall performance.

We compared the performance of three different classification methods, NB, SVM, and RF. We applied grid search to fine-tune these models based on hyperparameters, with a particular focus on SVM and RF. In the absence of missing data and with balanced data, SVM achieved the highest F1 score. The NB method also showed improved performance when data was balanced and missing data imputation was not applied. However, when the data was balanced and missing values were considered, RF emerged as the standout method. The ensemble nature of RF appeared to contribute to its strong performance under these conditions, which is consistent with previous research findings.

Furthermore, when comparing precision and recall values for the three classification methods, SVM demonstrated balanced recall and precision between the bankrupt and non-bankrupt classes when working with balanced data. NB, on the other hand, struggled to detect the non-bankrupt class, likely due to the requirement for more non-bankrupt data to recognize this class properly. While the performance of SVM decreased slightly when working with balanced data and missing values, the RF method exhibited improvements in both precision and recall values for all classes.

Table 6 presents a comparison of performance values for the predictive model using the Polish Company dataset trained with the RF algorithm for predicting bankruptcy. Two parameters, the Area Under Curve (AUC) and F1 values, are employed to assess the performance of the bankruptcy prediction model. The AUC value illustrates the model's ability to differentiate between two classes, in this case, the bankrupt and non-bankrupt classes. Like the F1 value, the maximum of AUC is 1. The higher the AUC or F1 value, the better the generated predictive model. In the study by Dzik-Walczak and Odziemczyk (2021), the same strategy was employed as ours, using stratified Cross Validation for model training. The proposed predictive model achieved an AUC value 9% higher than that of Dzik-Walczak and Odziemczyk (2021). Different results were obtained when comparing with the study by Quynh and Thi Lan Phuong (2020); the AUC value was higher than the predictive model proposed by us. Quynh and Thi Lan Phuong (2020) used more than one RF classifier to achieve these results. When comparing F1 values, the F1 value of our proposed model is only 0.9% lower than that in the study by Quynh and Thi Lan Phuong (2020).

Table 6. Comparison of Area Under Curve (AUC) and F1 values from other studies for the Polish Company dataset using RF.

Research	**AUC**	F1
Dzik-Walczak and Odziemczyk (2021)	0,8342	-
Proposed Method	0,9269	0,8494
Quynh and Thi Lan Phuong (2020)	0,9931	0,8584

Source: Dzik-Walczak and Odziemczyk (2021), own processing and Quynh and Thi Lan Phuong (2020)

Conclusion

We have conducted tests to build a predictive model using two highlighted strategies in machine learning: balancing data through undersampling and handling missing values. Filling in missing values using nearest neighbors and undersampling techniques enables RF to perform optimally compared to SVM and NB. Although the construction of bankruptcy prediction models heavily depends on the characteristics of the dataset used, the choice of the RF method with data balancing and handling missing values strategies become the initial preference in building predictive models for the dataset. These findings can inform future efforts to enhance the accuracy and reliability of bankruptcy prediction models, which can be of significant value to financial institutions and other stakeholders in making informed decisions.

While it would have been advantageous to include data from a different country to facilitate a comparative analysis between nations, thereby corroborating our findings and enhancing the depth of our discussion, regrettably, the majority of available datasets within the databases are associated with unidentified companies, and the specific countries in which these entities operate remain undisclosed. For future developments, we plan to create a hybrid model by applying optimization methods to select relevant features for bankruptcy prediction.

Declaration of competing interests

The author declares no competing financial interests.

References

ALTMAN E.I., 1968. Financial Ratios, Discriminant Analysis And The Prediction Of Corpporate Bankruptcy, *The Journal Of Finance*, **XXIII**(4), pp. 589–609.

ATIYA A.F., 2001. Bankruptcy prediction for credit risk using neural networks: A survey and new results, *IEEE Transactions on Neural Networks*, **12**(4), pp. 929–935. Available at: https://doi.org/10.1109/72.935101.

BARBOZA F., KIMURA H., ALTMAN E., 2017. Machine learning models and bankruptcy prediction, *Expert Systems with Applications*, **83**, pp. 405–417. Available at: https://doi.org/10.1016/j.eswa.2017.04.006.

BATENI L., ASGHARI F., 2020. Bankruptcy prediction using logit and genetic algorithm models: A comparative analysis, *Computational Economics*, **55**(1), pp. 335–348. Available at: https://doi.org/10.1007/s10614-016-9590-3.

BRÎNDESCU-OLARIU D., GOLEȚ I., 2013. Bankruptcy prediction ahead of global recession: Discriminant analysis applied on Romanian companies in Timis Country, *Timisoara Journal of Economics and Business*, **6**(19), pp. 70–94. Available at: https://doi.org/10.1515/9783112597569-toc.

CLEOFAS-SÁNCHEZ L., GARCÍA V., MARQUÉS A.I., SÁNCHEZ J.S., 2016. Financial distress prediction using the hybrid associative memory with translation, *Applied Soft Computing Journal*, **44**, pp. 144–152. Available at: https://doi.org/10.1016/ j.asoc.2016.04.005.

CULTRERA L., CROQUET M., JOSPIN J., 2017. Predicting bankruptcy of Belgian SMEs: A Hybrid approach based on factorial analysis, *International Business Research*, **10**(3), p. 33. Available at: https://doi.org/10.5539/ibr.v10n3p33.

DALIANIS H., 2018. Evaluation Metrics and Evaluation, in Clinical Text Mining. Springer, Cham., pp. 45–53. Available at: https://doi.org/10.1007/978-3-319-78503-5_6.

DANENAS P., GARSVA G., 2015. Selection of Support Vector Machines based classifiers for credit risk domain, *Expert Systems with Applications*, **42**(6), pp. 3194–3204. Available at: https://doi.org/10.1016/j.eswa.2014.12.001.

DZIK-WALCZAK A., ODZIEMCZYK M., 2021. Modelling cross-sectional tabular data using convolutional neural networks: Prediction of corporate bankruptcy in Poland, *Central European Economic Journal,* **8**(55), pp. 352–377. Available at: https://doi.org/ 10.2478/ceej-2021-0024.

KHAN U.E., 2018. Bankruptcy prediction for financial sector of Pakistan: Evaluation of logit and discriminant analysis approaches, *Pakistan Journal of Engineering, Technology & Science*, **6**(2), pp. 210–220. Available at: https://doi.org/10.22555/pjets.v6i2.1966.

KOROL T., 2012. Fuzzy logic in financial management, in *Fuzzy Logic - Emerging Technologies and Applications*. InTech, pp. 259–286. Available at: http://dx.doi.org/10.5772/35574.

KOTSIANTIS S., KANELLOPOULOS D., PINTELAS P., 2006. Handling imbalanced datasets : A review, *GESTS International Transactions on Computer Science and Engineering*, **30**(1), pp. 25–36. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.96.9248 &amp;rep=rep1&amp;type=pdf.

LEE S., CHOI W.S., 2013. A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis, *Expert Systems with Applications*, **40**(8), pp. 2941–2946. Available at: https://doi.org/10.1016/j.eswa.2012. 12.009.

MIHALOVIČ M., 2016. Performance comparison of multiple discriminant analysis and logit models in bankruptcy prediction, *Economics and Sociology*, **9**(4), pp. 101–118. Available at: https://doi.org/10.14254/2071-789X.2016/9-4/6.

PAVLICKO M., MAZANEC J., 2022. Minimalistic logit model as an effective tool for predicting the risk of financial distress in the Visegrad group, Mathematics, **10**(8), pp. 1–22. Available at: https://doi.org/10.3390/math10081302.

QUYNH T.D., THI LAN PHONG T., 2020. Improving the bankruptcy prediction by combining some classification models, in 2020 12th International Conference on Knowledge and Systems Engineering. IEEE, pp. 263–268. Available at: https://doi.org/10.1109/KSE50997.2020. 9287707.

RAINARLI E., 2019. The Comparison of machine learning model to predict bankruptcy: Indonesian stock exchange data, in *IOP Conference Series: Materials Science and Engineering*. Bandung: IOP, pp. 6–12. Available at: https://doi.org/10.1088/1757-899X/662/5/052019.

RAINARLI E., AARON A., 2015. The implementation of fuzzy logic to predict the bankruptcy of company in Indonesia, *International Journal of Business and Administrative Studies*, **1**(4), pp. 147–154. Available at: https://doi.org/10.20469/ ijbas.10003-4.

SABEK A., 2023. Unveiling the diverse efficacy of artificial neural networks and logistic regression: A comparative analysis in predicting financial distress. *Croatian Review of Economic, Business and Social Statistics,* CREBSS), **9**(1), 16-32. Available at: http://doi.org/ 10.2478/crebss-2023-0002.

SABEK A.,, HORAK J., 2023. Gaussian process regression´s hyperparameters optimization to predict financial distress. Retos, Revista de Ciencias Administrativas y EconA3micas, **13**(26), 273-289. Available at: https://doi.org/10.17163/ret.n26. 2023.06.

SUN J., HUI L., QING-HUA H., KAI-YU H., 2014. Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches, *Knowledge-Based Systems*, **57**, pp. 41–56. Available at: https://doi.org/ 10.1016/ j.knosys.2013.12.006.

TOMCZAK S., 2016 Polish companies bankruptcy data, Machine Learning Repository. Available at: https://doi.org/ttps://doi.org/10.24432/C5F600.

TSAI C.F., HSU Y.F., YEN D.C., 2014. A comparative study of classifier ensembles for bankruptcy prediction, *Applied Soft Computing Journal*, 24, pp. 977–984. Available at: https://doi.org/10.1016/j.asoc.2014.08.047.

ZAHIN S.A., AHMED C.F., ALAM T., 2018. An effective method for classification with missing values, *Applied Intelligence*, **48**(10), pp. 3209–3230. Available at: https://doi.org/10.1007/ s10489-018-1139-9.

Contact address of the author(s):
Ednawati Rainarli, Department of Informatics Engineering, Universitas Komputer Indonesia, Jl. Dipatiukur 112 -116 Bandung, Indonesia, ednawati.rainarli@email.unikom.ac.id, ORCID: 0000-0002-5770-1970.

Amine Sabek, Investment bets and sustainable development stakes in border areas, University of Tamanghasset, B.P 10034 Tamanghasset Airport Road, Algeria, sabek.amine@univ-tam.dz, ORCID:0000-0002-6970-4183.