Online Version Available at: http://bsti.ubd.ac.id/e-jurnal

JOURNAL OF TECH-E

|2581-1916 (Online) | 2598-7585 (Printed) |

Article

Identify Company Insolvency Using Multiple Linear Regression Algorithms

Deny Haryadi[1], Arif Rahman Hakim[2], Dewi Marini Umi Atmaja[3], Amat Basri[4], Risma Adisty Nilasari[5]

[1,] *Institut Teknologi Telkom Jakarta, Information Technology, DKI Jakarta, Indonesia*
[2,3,] *Universitas Medika Suherman, Digital Business, Jawa Barat, Indonesia*
[4] *Universitas Buddhi Dharma, Infiormation Systems, Banten, Indonesia*
[5] *Universitas Duta Bangsa Surakarta, Information Technology, Jawa Tengah, Indonesia*

SUBMISSION TRACK

Received: Januari 24, 2023
Final Revision: February 18, 2023
Available Online: February 25, 2023

KEYWORD

Bankruptcy of a company, Data Mining, Predict, Regresi Linear Algorithm, Root Mean Squared Error

CORRESPONDENCE

E-mail: denyharyadi@ittelkom-jkt.ac.id

A B S T R A C T

Corporate bankruptcy can hurt the company and affect the state of the economy. Therefore, many interested parties want to know the business situation related to the company. These parties include creditors, auditors, shareholders, and management itself who have an interest in knowing the state of the company in the context of bankruptcy. The past financial statements of a company can be used to predict future financial conditions using report analysis techniques. In the risk assessment process, expert knowledge is still seen as an important task, because expert predictions are subjective. This study aims to predict the bankruptcy of the company using influencing factors such as the level of research and development costs, the growth rate of total assets, and the current asset turnover rate. The method used in this research is the prediction method using the Linear Regression Algorithm. Based on the test results show that the variables or attributes used in this study have a significant effect, as evidenced by using a linear regression algorithm to be able to produce a Root Mean Squared Error value: 0.162 +/- 0.000.

INTRODUCTION

The increasingly fierce competition in the economic field has led to changes in economies of scale that are not only national but global. Companies face major challenges, namely how to avoid financial difficulties (bankruptcy), financial distortion, or financial difficulties are the first stage before a company bankruptcy occurs [1]. The emergence of various predictions of company bankruptcy is a system of prevention or anticipation of financial desert [2]. Predicting a company's bankruptcy is the most appropriate step to overcome problems in the company, the purpose is to find out whether the company is in good condition or not [3]. Mistakes in making decisions can affect financial conditions as well as affect shareholders or company owners [4]. The bankruptcy of an enterprise negatively affects

the company and can adversely affect the condition of the economy. So many parties are interested in knowing the business status related to the company. These parties include creditors, auditors, shareholders, and the management of the company itself [5]. Techniques in data mining such as Case-Based Reasoning (CDR), Artificial Neural Networks (ANN), and Decision Tree (DT) in recent years have been widely used as alternative methods to predict a company's bankruptcy [6][7]. The source of the data in this study was taken from the address of https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction website. This study aims to find out how much accuracy the level of accuracy produced by linear regression algorithms. Through this research, it is also expected that a lower average error rate can be known so that it can be used to predict the bankruptcy of a company [8].

Prediction is a systematic process of predicting what will happen in the future and minimizing it [9]. Linear Regression is utilized to determine which bound variables can be predicted through free variables [10]. The usefulness of regression algorithms can be applied to decide whether a class of bonded variables/falls can be done through free variables if the value of the free variable is raised/decreased or vice versa [11]. This research is expected to be a reference in analyzing the bankruptcy of a company. The results of this company data processing are expected to become information and knowledge so that the company can be better, and can find new opportunities or strategic planning in company bankruptcy analysis, besides that it can be used as a means to make decisions in preventing company bankruptcy.

I. LITERATURERES REVIEW

A multiple linear regression algorithm has been applied to predict the price of crude palm oil (CPO), although not with a 100% accuracy value but still within the margin of error in predicting [12]. Other researchers have also applied multiple linear regression algorithms to predict rice production in Bantul Regency, taking into account 3 variables that include harvested land area, rainfall, and pest infestations that can affect rice production. Through validity testing using the MAD method, test results were obtained for rice production predictions of 0.101 so that the predicted results were in the very good category [13]. In multiple linear regression algorithms, independent variables have a partially significant influence on dependent variables. Based on the results of previous research, it is stated that the performance of multiple linear regression models in forecasting the monthly inflation rate of Indonesia produces an accuracy rate with a Mean Absolute Deviation (MAD) value of 0.0380, a Mean Square Error (MSE) of 0.0023, and a Root Mean Square Error (RMSE) value of 0.0481. [14].

II. FRAMEWORK

Before predicting company data that will be tested according to data modeling that will be used to facilitate research and run as desired, the research flow or systematics is made, here is the flow/systematics of this research:

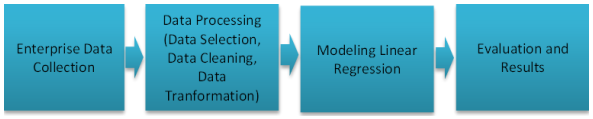

Fig 1. Framework

Data Collection

The source of data in this study was obtained from the internet. To facilitate research so that it runs systematically, stages of research are needed that explain where the source of this research data is carried out, the data sources in this study are taken from the kaggle.com site and literature studies to deepen the study material in this study.

Data Selection

Data Selection is the process of selecting data from datasets before entering the data and information mining stage [15]. The steps at this stage are as follows:

1. The sample data is taken randomly by taking into account the parameters, attributes and the largest amount of data to be used as a dataset, ensuring that the data taken is suitable for use in the modeling process.
2. After the datasets are grouped, it can be known the number of datasets for the next process.
3. Perform a selection of attributes to be used and then analyzed, because in the previous stage there are attributes that are not needed.

Data Cleaning

The data cleaning process is used to delete data by removing missing values, duplicating data and checking for data inconsistencies, and correcting errors in the data. In this data, data cleaning is carried out such as for missing parts, and inconsistent data (incorrect input) [16][17].

Data Transformation

The Data transformation stage is the process of changing the data format before it is processed using an algorithm in a tool or program to be used [15].

Modeling

In this study, prediction techniques in data mining used the Linear Regression algorithm. This algorithm is used to determine the suitability between dependent variables capable of being predicted through independent variables. In this study, the analysis used is multiple regression analysis, this multiple regression analysis aims to predict the state (increase or decrease) of the dependent variable if two or more independent variables become predictors of value manipulation. Multiple regression analysis is performed if the number of free variables is at least 2. The formula for the Simple Linear Regression equation for the values of a and b is established at the following concurrently.

$$b = \frac{n \sum_{i=1}^{n} xiyi - (\sum_{i=1}^{n} xi)(\sum_{i=1}^{n} yi)}{(\sum_{i=1}^{n} xi^2) - (\sum_{i=1}^{n} xi)^2}$$

$$a = \bar{y} - \mathrm{b}\bar{x}$$

so

$$a = \frac{\sum_{i=1}^{n} yi}{\mathrm{n}} - b \frac{\sum_{i=1}^{n} xi}{\mathrm{n}}$$

n is the total amount of data;

I is the value of the bound variable Y to I;

xi is the value of the free variable X to I;

To calculate the linear regression equation is:

$$y = \mathrm{a} + \mathrm{bx}$$

Y is a bound variable;
X is a free variable;
a is a constant;
b is the slope.
To forecast a variable using regression, each variable must be available. The next step is to calculate the regression equation through calculations according to the above formula.

Evaluation and Results

Evaluation of results/performance tests is carried out to find out the calculation results, as well as being a benchmark for how good or not the linear regression algorithm is. The data testing process in this study uses rapid miner software to see whether the data is by the results obtained through these tools and find out how bound variables can be predicted through variables.

III. METHODS

The research instrument used is a quantitative method using data mining tool, namely Rapid Miner software, Rapid Miner is used to help carry out the RMSE calculation process in predicting company bankruptcy data. The data used in this study is company bankruptcy data taken from the https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction with a total of 6819 records has variables or attributes in the form of factors that influence it, namely research and development expense rate, total asset

growth rate, and current asset turnover rate. The data that has been obtained is then carried out in a preprocessing process and then divided into training data and testing data, while the training data used in this study is 90% and the remaining 10% as testing data. The next stage will be calculated on the testing data using a multiple linear regression algorithm to generate RMSE values. The resulting value is used to predict testing data to produce predictable testing data results from multiple linear regression algorithm models.

IV. RESULT

Linear Regression Algorithm Modeling

The algorithm used in this study is linear regression, the performance test used to identify the bankruptcy of the company used is the Root Mean Square Error (RMSE), and the resulting forecasting results can be used to decide whether a company is bankrupt or not. The data source used in this Venetian is sourced from the kaggle.com site. The variables used are the Research and development expense rate, Total Asset Growth Rate, Current Asset Turnover Rate, and Bankrupt variables.

Split Validation

The validation technique used in this study is to partition the data into two parts arbitrarily / randomly, namely, some data is used for training data and the rest as test data. The ratio in the split validation used previously has been determined by the author at 90% of the training data and the remaining 10% for the testing data.

Table 1. Bankruptcy Prediction

Research and development expense rate (X1)	Total Asset Growth Rate (X2)	Current Asset Turnover Rate (X3)	Bankrupt (Y)
441000000	6990000000	9860000000	?
764000000	6700000000	0.000219251	?
357000000	6440000000	0.000108039	?

Data Testing

After the data is collected, the next stage that is carried out is the creation of a model on the rapid miner software. The modeling applied in this rapid miner software is using a linear regression algorithm. The initial step is to enter the company's dataset into the rapid miner software for processing. The dataset used in this study was divided into two parts, namely 90% as training data and the remaining 10% used for testing data, and random using split validation techniques. This technique is a validation technique for randomly partitioning training data and test data.

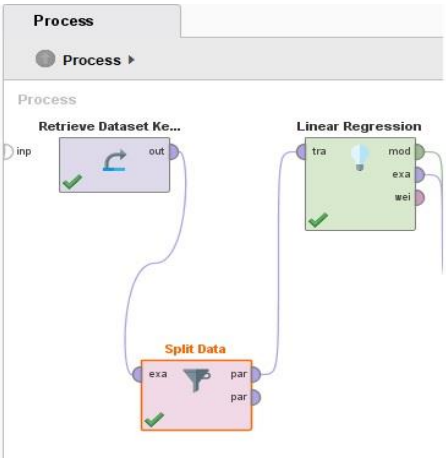

Fig 2. Data Sharing Process with Split Validation in Rapidminer

The next process is to set split validation parameters by dividing training data and testing data in a rapid miner. Next, enter a linear regression algorithm to see the prediction results in the rapid miner. Next, select attributes, namely to find out the prediction results from the rapid miner software, manual calculations, and test results.

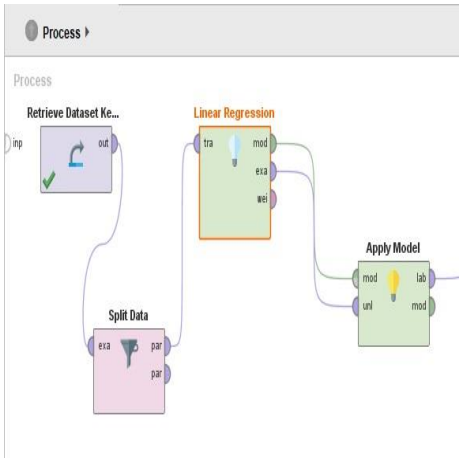

Fig 3. Model Evaluation Process

The next step is to enter training data and test data to generate predictions on the class attributes.

Table 2. Predicted Results

Prediction (Bankrupt)	Research and Development Expense Rate	Total Asset Growth Rate	Current Asset Turnover Rate
0.021	25500000	7280000000	0.002
0.025	730000000	5720000000	0.000
0.034	50900000	6630000000	7290000000
0.045	1900000000	747000000	4760000000
0.025	1210000000	5730000000	0.000
0.038	917000000	6160000000	9230000000
0.051	79800000	830000000	8710000000
0.022	849000000	7030000000	0.000
0.024	119000000	6340000000	475000000
0.023	1040000000	6460000000	0.000
0.023	1190000000	6560000000	0.000
0.026	1240000000	5490000000	0.000
0.033	900000000	6580000000	6370000000
0.023	1440000000	6640000000	0.000
0.024	577000000	5980000000	0.000

After the prediction results are found, the next stage is to measure the accuracy of the prediction results that have previously been made.

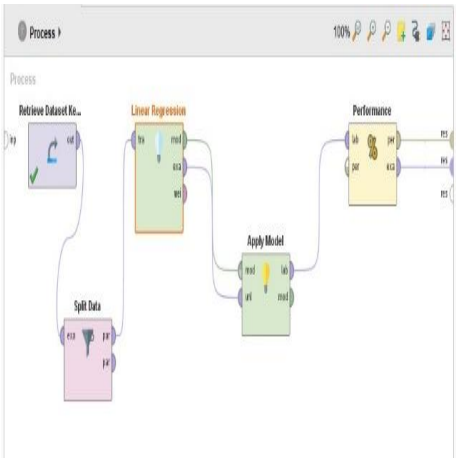

Fig 4. The process of searching for RMSE

To make it easier to read company data, then enter performance tools to find the RMSE value. Here are the results.

root_mean_squared_error

root_mean_squared_error: 0.162 +/- 0.000

Fig 5. RMSE Test Results

The next stage is to implement a linear regression algorithm using rapid miner software. The stages are as follows:

1. By using a rapid miner determine the predicted results of the testing data, as well as the confidence value of the predicted data.
2. Specify the performance value of the result/output that has been obtained to search for RMSE.

By using split validation techniques, namely to partition training data and testing data, as well as the Apply Model feature to apply models to testing and Performance data to display RMSE values.

V. DISCUSSION

After calculations are made to find the values of X_1Y, X_2Y, X_3Y, $X_1X_2X_3$, X_1^2, $X_2^{2,}$ and X_3^2, the result of each of them is for the total value of attribute X_1 is 12263180000, the total value of attribute X_2 is 599287500000, the total value of attribute X_3 is 312791000000, the total value of attribute Y is 8, the total value of X_1Y is 77444000000, the total value of X_2Y is 39884000000, the total value of X_3Y

is 35148000000, the total value of $X_1X_2X_3$ is 327190632200011000000, the total value of X_1^2 is 3167842844400000000, and the total value of X_2^2 is 3969415470370000000000, and the total value of X_3^2 is 24549582330000000000000. From all the total values used to find the value of coefficient a, coefficient b_1, coefficient b_2, and coefficient b_3 the value of a=0.20, value b_1=-1.24, value b_2=-2.13, and value b_3= 6.85 are obtained.

The values of the coefficients a, b_1, b_2, and b_3 will be used for the equation of the double linear regression algorithm. From the coefficient values above, then by implementing the model, we can apply the equation of the Linear Regression algorithm with the model Y = 0,20. (-1,24.X_1). (-2,13.X_2).(6,85.X_3), where variable X_1 is the Research and development expense rate, variable X_2 is the Total Asset Growth Rate variable and X_3 is the Current Asset Turnover Rate.

Using the multiple regression equation above, the next step is to implement predictions on the test data. Overall, the application of linear regression equations is applied to predict three data sets such as predefined test data. The results of manual calculations are compared with the results of the Rapid Miner software calculation process. There was no significant difference between the two results. So there is no big difference. Therefore, we can conclude that calculations and manual results are available. It is processed by the Rapid Miner software and shows similar results.

Table 3. Bankruptcy Prediction Comparison

Research and development expense rate (X1)	Total Asset Growth Rate (X2)	Current Asset Turnover Rate (X3)	Bankrupt (Y) Manual	Bankrupt (Y) Rapidminer
441000000	6990000000	9860000000	0.11465132	0.033924707
764000000	6700000000	0.000219251	0.049215229	0.021457922
357000000	6440000000	0.000108039	0.059812686	0.022160371

Meanwhile, the comparison of the value of the observation Y variable with the predicted variable Y value in the rapid miner software can be seen through the following graph.

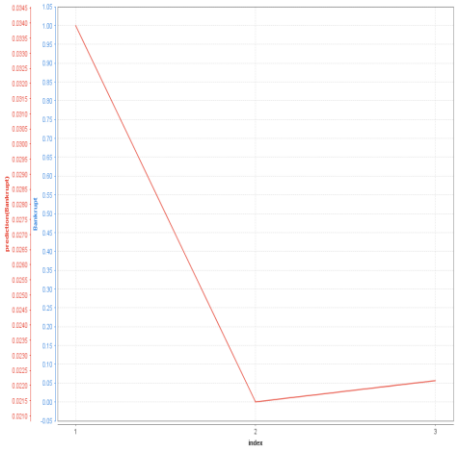

Fig 6. Comparison between Values (Y) Observation and (Y) Prediction

Based on the test results, the variables Research and development expense rate, Total asset growth rate, and Current asset turnover rate have a significant effect as evidenced by using a linear regression algorithm capable of generating the Root Mean Squared Error value: 0.162 /- 0.000. This is because there is a correlation between bound variables and free variables. This testing process is carried out to identify the bankruptcy of the company using a linear regression algorithm.

VI. CONCLUSION

Based on the results of research that has been done, then the following conclusions are obtained:

The use of linear regression algorithms in company bankruptcy data consisting of variables Research and development expense rate, Total asset growth rate, and Current asset turnover rate can be used to identify company bankruptcies. The test results using a linear regression algorithm yielded an RMSE value of 0.162 /- 0.000. This proves the existence of a functional relationship (toleration) between bound variables and free variables.

REFERENCES

[1] P. Yoga, R. B. D. Putra, and E. S. Budi, "Prediksi Kebangkrutan Perusahaan Menggunakan Artificial Neural Network," *J. Ris. Komput.*, vol. 5, no. 5, pp. 503–510, 2018.

[2] D. Oktafia and D. D. L. C. Pardede, "Perbandingan Kinerja Algoritma Decission Tree Dan Naive Bayes Dalam Memprediksi Kebangkrutan," *repository.gunadarma.ac.id*, 2008.

[3] A. S. Malaka and Hartojo, "Model Prediksi Kepailitan Bank Umum Di Indonesia Menggunakan Algoritma Backpropagation," *J. Ilmu Manaj.*, vol. 2, no. 4, pp. 1714–1724, 2014.

[4] L. Handayani and Fitriandini, "Prediksi Kebangkrutan Perusahaan Menggunakan Support Vector Machine (SVM)," *SITEKIN J. Sains, Teknol. dan Ind.*, vol. 11, no. 1, pp. 31–35, 2013, [Online]. Available: http://iaesjournal.com/online/index.php/IJECE

[5] H. Amalia, A. Fitria Lestari, and A. Puspita, "Penerapan Metode Svm Berbasis Pso Untuk Penentuan Kebangkrutan Perusahaan," *J. Techno Nusa Mandiri*, vol. 14, no. 2, pp. 131–136, 2017, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Qualitativ

[6] I. Arieshanti and Y. Purwananto, "Model Prediksi Kebangkrutan Berbasis Neural Network Dan Particle Swarm Optimization," *JUTI J. Ilm. Teknol. Inf.*, vol. 9, no. 1, pp. 29–34, 2011, doi: 10.12962/j24068535.v9i1.a65.

[7] H. Saleh, "Prediksi Kebangkrutan Perusahaan Menggunakan Algoritma C4.5 Berbasis Forward Selection," *Ilk. J. Ilm.*, vol. 9, no. 2, pp. 173–180, 2017, doi: 10.33096/ilkom.v9i2.97.173-180.

[8] H. Eni, "PREDIKSI KEBANGKRUTAN BANK DENGAN MENGGUNAKAN ANALISIS DISKRIMINAN (Studi Kasus pada Bank yang Terdaftar di Bursa Efek Indonesia) Eni Handayani," *J. Ilm. Mat.*, vol. 1, no. 6, pp. 8–13, 2017, [Online]. Available: www.idx.co.id

[9] F. Firdaus and A. Mukhlis, "Implementasi Algoritma Naive Bayes Pada Data Set Kualitatif Prediksi Kebangkrutan," *JURIKOM (Jurnal Ris. Komputer)*, vol. 7, no. 1, pp. 15–20, 2020, doi: 10.30865/jurikom.v7i1.1757.

[10] A. Setiadi, "Data Mining Untuk Prediksi Kebangkrutan Perusahaan Berdasarkan Data Kualitatif," *Snit 2014*, vol. 1, no. 1, pp. 414–423, 2014, [Online]. Available: http://seminar.bsi.ac.id/snit/index.php/snit-2014/article/view/252

[11] D. Haryadi, A. R. Hakim, D. M. U. Atmaja, and S. N. Yutia, "Implementation of Support Vector Regression for Polkadot Cryptocurrency Price Prediction," *Int. J. Informatics Vis.*, vol. 6, no. 1–2, pp. 201–207, 2022, doi: 10.30630/joiv.6.1-2.945.

[12] A. Fitri Boy, "Implementasi Data Mining Dalam Memprediksi Harga Crude Palm Oil (CPO) Pasar Domestik Menggunakan Algoritma Regresi Linier Berganda (Studi Kasus Dinas Perkebunan Provinsi Sumatera Utara)," *J. Sci. Soc. Res.*, vol. 3, no. 2, pp. 78–85, 2020, [Online]. Available: http://jurnal.goretanpena.com/index.php/JSSR

[13] E. Triyanto, H. Sismoro, and A. D. Laksito, "Implementasi Algoritma Regresi Linear Berganda Untuk Memprediksi Produksi Padi Di Kabupaten Bantul," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 4, no. 2, pp. 73–86, 2019, doi: 10.36341/rabit.v4i2.666.

[14] Amrin, "Data Mining Dengan Regresi Linier Berganda Untuk Peramalan Tingkat Inflasi," *J. Techno Nusa Mandiri*, vol. XIII, no. 1, pp. 74–79, 2018.

[15] D. M. U. Atmaja and R. Mandala, "Analisa Judul Skripsi untuk Menentukan Peminatan Mahasiswa Menggunakan Vector Space Model dan Metode K-Nearest Neighbor," *IT Soc.*, vol. 4, no. 2, pp. 1–6, 2020, doi: 10.33021/itfs.v4i2.1182.

[16] D. Haryadi and R. Mandala, "Prediksi Harga Minyak Kelapa Sawit Dalam Investasi Dengan Membandingkan Algoritma Naïve Bayes, Support Vector Machine dan K-Nearest Neighbor," *IT Soc.*, vol. 4, no. 1, pp. 28–38, 2019, doi: 10.33021/itfs.v4i1.1181.

[17] D. Haryadi, D. Marini Umi Atmaja, A. Rahman Hakim, and N. Suwaryo, "Identifikasi Tingkat Resiko Penyakit Stroke Menggunakan Algoritma Regresi Linear Berganda," *SNTEM*, vol. 1, no. 1, pp. 1198–1207, 2021.

BIOGRAPHY

Deny Haryadi, Graduated from the Informatics Engineering study program at Pelita Bangsa University in 2017 (S1), and graduated from the Informatics Engineering Masters at the President University in 2020. Currently working as a lecturer at Institut Teknologi Telkom Jakarta

Arif Rahman Hakim, Graduated from the Informatics Engineering study program at Pelita Bangsa University in 2017 (S1) and graduated from the Informatics Engineering Masters at the President University in 2020. Currently working as a lecturer at Medika Suherman University.

Dewi Marini Umi Atmaja, Graduated from Informatics Study Program, at Jenderal Achmad Yani University in 2018 (S1), and graduated from the Informatics Engineering Masters at the President University in 2020. Currently working as a lecturer at Medika Suherman University.

Amat Basri, Graduated from the Information Technology Study Program (S1) in April 2002, continued his Master in Information Systems in January 2015, and graduated in September 2016. Currently, I am a lecturer in the information systems program at Buddhi Dharma University.

Risma Adisty Nilasari, Majoring in Informatics Engineering, Faculty of Computer Science, University of Duta Bangsa Surakarta.