Matematik / *Mathematics* **DOI: 10.21597/jist.703004**

Araştırma Makalesi / *Research Article*

Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 10(4): 2881-2888, 2020
Journal of the Institute of Science and Technology, 10(4): 2881-2888, 2020
ISSN: 2146-0574, eISSN: 2536-4618

On the Financial Situation Analysis with KNN and Naive Bayes Classification Algorithms

Oğuzcan ULUDAĞ[1], Arif GÜRSOY[1*]

ABSTRACT: Classification, a data mining technique, has been applied on the financial parameters used in the Altman Z-Score formulas for a certain number of selected firms in manufacturing industry. The Altman Z-Score is used to estimate a firm's financial difficulties. The Z-Score value shows whether the financial position of the firm is good, moderate or risky. In this study, KNN and Naive Bayes algorithms are used as classification methods. The Z-Score values of all firms are calculated and a certain number of data for all three types are selected and taught to the system as learning data. Algorithms are run on the financial parameters in the Z-Score formulas of companies not taught to the system. Over all data, The KNN and Naive Bayes algorithms achieve 84–88% and 75–86% success, respectively. This study, where data mining techniques are applied on a finance model and successful results are achieved, will contribute to the application of different technologies in many different analysis processes of the financial sector.

Keywords: Data Mining, Classification, Financial Analysis, Altman Z-Score

KNN ve Naive Bayes Sınıflandırma Algoritmaları ile Finansal Durum Analizi Üzerine

ÖZET: Bu çalışmada veri madenciliği tekniklerinden biri olan sınıflandırma, imalat sanayi sektöründe hizmet veren belirli sayıda seçilmiş firmaların Altman Z-Skor formülünde kullanılan finans parametreleri üzerinde uygulanmıştır. Altman Z-Skor değeri bir firmanın finansal zorluklarla karşılaşma durumunun tahminlemesinde kullanılır. Z-Skor değeri, firmanın finansal durumunun iyi, orta veya riskli olup olmadığı hakkında yorum yapar. Bu makalede sınıflandırma yöntemi için KNN ve Naive Bayes algoritmaları kullanılmıştır. Bütün firmaların Z-Skor değerleri hesaplanmış ve her 3 tipten belirli sayıda veri seçilerek öğrenme verisi olarak sisteme öğretilmiştir. Algoritmalar sisteme öğretilmemiş firmaların Z-Skor formülündeki finans parametreleri üzerinde çalıştırılmıştır. Tüm veri üzerinde KNN algoritması yaklaşık %84-88, Naive Bayes algoritması ise %75-86 aralığında başarı ile sonuçlanmıştır. Veri madenciliği tekniklerinin bir finans modeli üzerinde uygulandığı ve başarılı sonuçların elde edildiği bu proje, farklı teknolojilerin finans sektörünün bir çok farklı analiz süreçlerinde uygulanmasına katkı sağlayacaktır.

Anahtar Kelimeler: Veri Madenciliği, Sınıflandırma, Finansal Analiz, Altman Z-Skor

[1]Oğuzcan ULUDAĞ (**Orcid ID**: 0000-0003-0516-0014), Arif GÜRSOY (**Orcid ID:** 0000-0002-0747-9806), Ege University, Faculty of Science, Department of Mathematics, İzmir, Turkey

*Corresponding Author: Arif GÜRSOY, e-mail: arif.gursoy@ege.edu.tr

Geliş tarihi / *Received:* 12-03-2020
Kabul tarihi / *Accepted:* 04-07-2020

INTRODUCTION

Data is the most fundamental part of information. The data forms meaningful or meaningless expressions as it is merged. The main goal is to discover meaningful information from the data warehouses formed by the data. The difficulty of the process of discovering information is directly proportional to the size of the data. The smaller our data warehouse, the easier it is to discover information.

With the development of technology, meaningful expressions, predictions or links can be discovered from large data warehouses. This process is called data mining. Data mining is directly related to the structure of the data. Different data mining techniques can be applied depending on the type or structure of the data being studied. Classification is one of the most commonly used data mining techniques. This method is very important because most existing data models are relational.

Data mining is the method in financial analysis. When data mining methods are applied to financial statements, which are financial indicators of companies, it will be possible to obtain useful information about the financial behavior of companies (Özkan and Boran, 2014). In order to make a decision about the financial situation of a company, the financial statements of that company should be consulted. In these financial statements, briefly, we can see the details of a company's receivables and payables. This data is interpreted with financial statement analysis techniques and enables the firms to comment on their financial situation. There are some financial models that comment on the failure situations of firms using this data. Beaver's 1966 study is an early piece of research on the prediction of financial failure. Beaver used the discriminant analysis method by grouping the financial ratios selected for the prediction of financial failure (Beaver, 1966).

Another approach to the prediction of financial failure is Altman's study, based on multiple discriminant analysis methods. Altman formulated the Z-Scores model by using the financial ratios he had determined (Altman, 1968).

Many different approaches have been applied to financial data. Ohlson used logistic regression analysis in his study on the bankruptcy prediction of firms. He used three different models in his study, consisting of one year before the company's bankruptcy, two years before the bankruptcy, and finally one and two years before (Ohlson, 1980). In 2003, Aktaş used discriminatory, multiple regression and artificial neural network models. This method has been applied to 106 enterprises, one half successful and the other half unsuccessful. The study found that the artificial neural network model is more efficient than the multiple regression model in predicting financial failure (Aktaş, 2003). In their 2016 study on financial failure, Kaygın, Tazegül and Yazarkan used data mining and logistic regression analysis. It has been analyzed that the logistic regression model makes a successful prediction (Kaygın et al., 2016). In 2017, Dewi and Hadri used regression analysis on financial ratios to estimate the financial failures of firms, determining that working capital/total assets, current ratio, stock book value/total liabilities, total debt/total assets, interest and profit before tax/total liabilities ratios are important parameters (Dewi and Hadri, 2017). In the same year, Kürklü and Türk analyzed the financial failure estimate of 166 companies that member BIST by using Altman Z-Score and Springate S-Score models (Kürklü and Türk, 2017). In 2018, Fathi, Saif and Heydari used to data mining models to predict bankruptcy of companies and compared the results with Z Altman model. Their study was shown that data mining model has more power to predict bankruptcy (Fathi et al., 2018).

In this study, classification algorithms, a data mining technique, are applied to the Altman Z-Score finance model. The validity of the process has been measured by comparing the classification results according to the Z-Score values. Thus, a new data mining approach aims to carry out financial situation analysis. There are several important points that increase the value of this study. First, it aims to provide a new perspective on financial analysis methods by using developing technology on today's data. The other, it aims to is a more efficient analysis with use less data.

The aim of this study is to analyze the data status of 156 firms operating in the manufacturing industry using data mining methods. Altman Z-Score values have been calculated using the financial values obtained from the Kamu Aydınlatma Platformu (KAP) between 2013 and 2018. The classification process has been applied to the parameters used in Altman Z-Score calculation. Altman Z-score values for companies in the same class have been compared, and the success of the method has been calculated.

MATERIALS AND METHODS

In this study, the financial values of 156 firms operating in the manufacturing industry between 2013-2018 have been taken from the Kamu Aydınlatma Platformu (KAP). A total of 936 records have been used as dataset. The Z-Score value of each firm record has been calculated through this dataset. A financial status labels have been assigned to each firm according to Z-Score value. KNN and Naive Bayes algorithms have been run on the parameters used in Altman Z-Score formula of each firm regardless of formula coefficients. The classification results have been compared with predetermined labels and the success of the algorithms have been observed.

Data Mining

Data mining is the process of obtaining information that may be useful from a large, noisy, incomplete and fuzzy collection of data that has not been previously discovered (Agrawal et al., 1993). In other words, it is the discovery of potentially useful information from large databases.

Data mining is a process seeking to discover a meaningful relationship between data communities. Data must go through several steps, as shown in Figure 1.

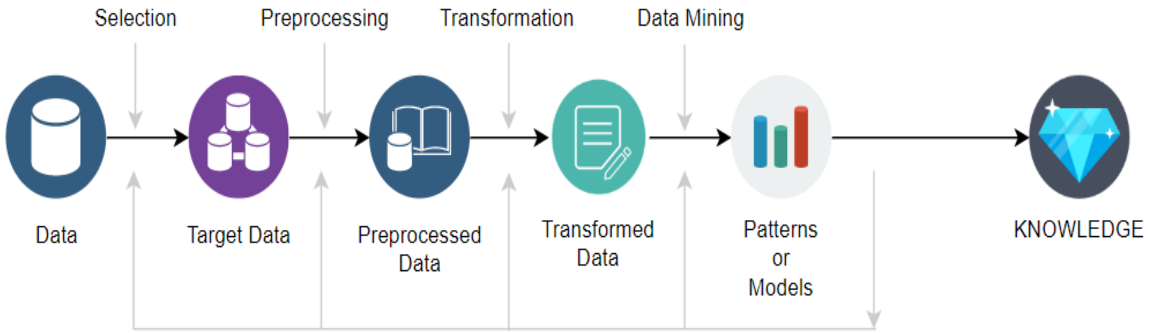

Figure 1. Data mining steps

In Figure 1, data selection is one of the most time-consuming stages. This is where the relevant data is collected and combined. In the preprocessing stage, data that is not suitable for use is cleaned and made available for use in the next steps. In the transformation stage, the cleaned data is converted to the required format for the application of data mining techniques. Data mining techniques are applied on data that is ready for the data mining stage. In the last stage, the results of the applied techniques are analyzed and information is discovered.

Classification is one of the most common methods used in data mining. The purpose of classification is to process the ungrouped data into groups according to similarities. Classification algorithms try to properly accomplish this by learning a classification method from the given training set.

The most common classification methods are as follows:

- Decision Trees
- Artificial Neural Networks
- Bayes Classifiers
- KNN (K-Nearest Neighbors) Methods
- Support Vector Machines
- Genetic Algorithms
- Association-based Classifiers

This study uses KNN and Naive Bayes algorithms. The KNN algorithm, one of the simplest and most important classification methods, memorizes all training data in advance. It finds the *k* data groups closest to the test data to be classified. Within this group, it completes the classification process by assigning a label to the data closest to the test data (Wu et al., 2008). Another important classification method is the Bayes method. A certain amount of learning data is processed into the system, and the algorithm tries to classify the test data by evaluating it using probabilistic operations on the learned data.

KNN (K Nearest Neighborhood) Algorithm

KNN is the most important nonparametric controlled learning algorithm in pattern recognition (Dasarathy, 1991). The algorithm calculates the distance of the data to be classified from the learning data and controls the nearest "*k*" neighbors to it. Based on these nearest neighbors, it then performs the labeling process. There are 3 types of distance functions commonly used for distance calculation operations:

- Euclidean Distance
- Manhattan Distance
- Minkowski Distance

Pseudo Code of the KNN Algorithm:

```
1. Create learning data
2. Determine the number of nearest neighbors (k)
3. For each record in the learning data
     • Calculate distances
     • Choose the nearest k neighbors
     • Label according to the majority of selected classes
4. End
```

Naive Bayes Algorithm

The Naive Bayes algorithm is a probability classification algorithm that calculates the probability set by counting the frequency and combinations of values in a data set (Patil and Sherekar, 2013). The algorithm calculates the probability of each state of the data to be classified, then performs the classification according to the highest calculated probability value.

The Naive Bayes classification process is based on Bayes' theorem:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)} \qquad (1)$$

P(A | B) = Probability of event A when event B occurs
P(A) = Probability of event A
P(B | A) = Probability of event B when event A occurs
P(B) = Probability of event B

Pseudo Code of the Naive Bayes Algorithm:

```
1. Create learning data
2. Calculate the average and standard deviations of the estimated variables
3. Calculate the probability of each state of the data
4. Make the classification according to the highest probability
5. End
```

Altman Z-Score Model

Altman has conducted a study to analyze the bankruptcies of firms. The analysis was conducted on 66 firms in total by selecting 33 companies for two groups, bankrupt and nonbankrupt, operating in the manufacturing industry sector. He has identified 22 significant financial ratios that should be evaluated using financial statements. He has divided these variables into five groups, liquidity, profitability, leverage, solvency and efficiency, and has developed the following formula according to these selected ratios (Altman, 1968):

$$Z = 0.012\, X1 + 0.014\, X2 + 0.033\, X3 + 0.006 X\, 4 + 0.999\, X5 \qquad (2)$$

The parameters used in the formula are as follows:

- X_1 = Working Capital/Total Assets
- X_2 = Retained Earnings/Total Assets
- X_3 = Earnings before Interest and Taxes/Total Assets
- X_4 = Market Value of Equity/Book Value of Total Liabilities
- X_5 = Sales/Total Assets

If the Z-score value

- > 2.99, the company's condition is good
- $1.81 < Z < 2.99$, the company's condition is moderate
- < 1.81, the company's condition is risky

RESULTS AND DISCUSSION

This study has developed a data mining approach for the analysis of the financial situations of firms operating in the manufacturing industry. The ratios obtained from a total of 936 financial data of 156 firms for 2013–2018 have been used as data. The Z-Score value of each firm has been calculated to determine its financial position. Companies with a Z-Score higher than 2.99 have good financial standing, companies with a Z-Score between 1.81 and 2.99 have normal financial standing and companies with a Z-Score below 1.81 have poor financial standing.

From the dataset, 50 financial data have been defined as learning data for every three different groups, with a total of 150 data, financial status good, moderate and risky. The implementation of the KNN and the Naive Bayes algorithms have performed the classification operations on the learned data defined in the system. Algorithms have been run on data unknown to the system. It has been evaluated whether the financial positions of the companies have been correctly classified. The business flow is shown in Figure 2.

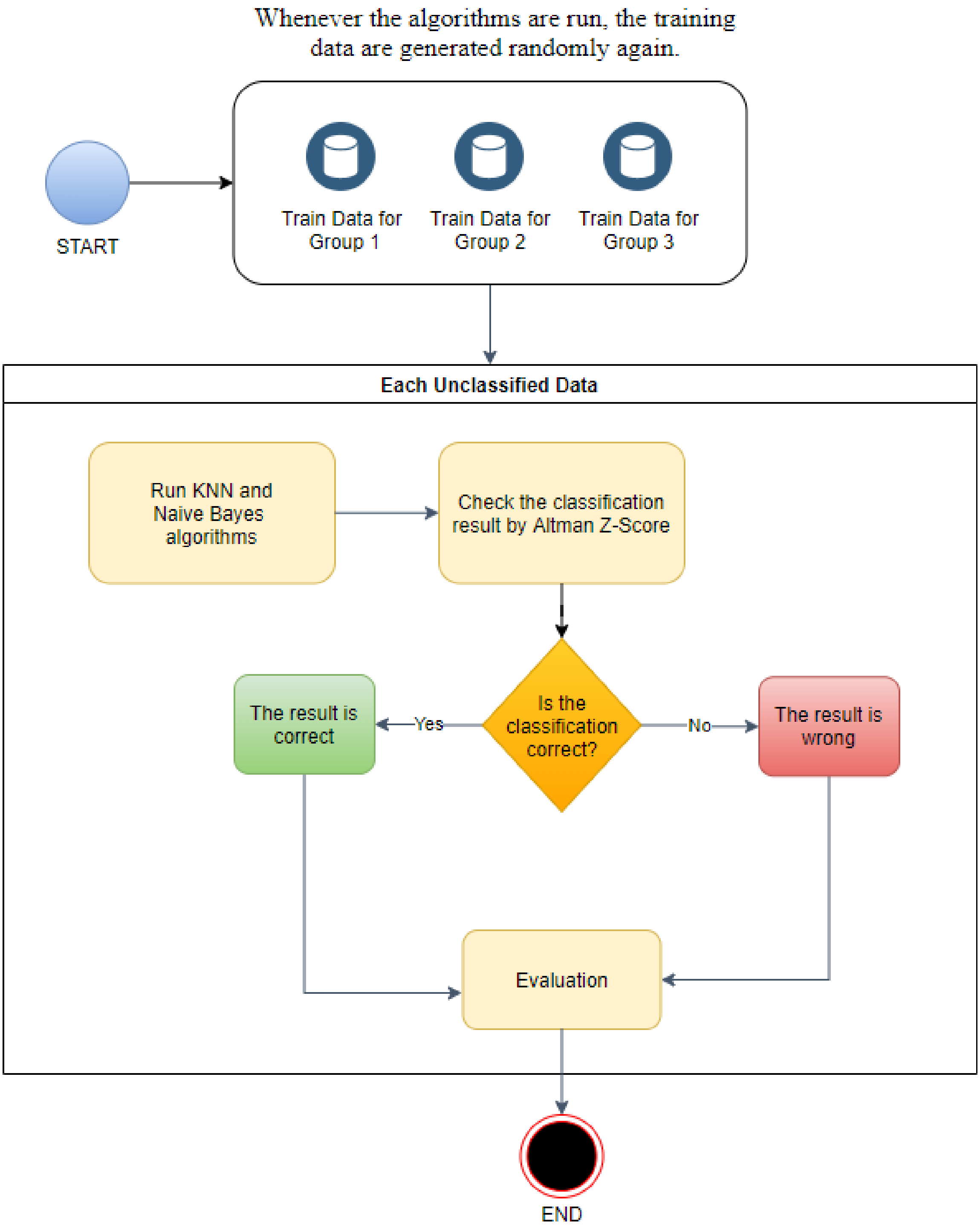
Whenever the algorithms are run, the training data are generated randomly again.
START
Train Data for Group 1
Train Data for Group 2
Train Data for Group 3
Each Unclassified Data
Run KNN and Naive Bayes algorithms
Check the classification result by Altman Z-Score
Is the classification correct?
Yes
No
The result is correct
The result is wrong
Evaluation
END

Figure 2. Business flowchart

These algorithms have been implemented in the C# programming language using Visual Studio IDE and tested on an i7-6700HQ machine with a 2.60 GHz processor and 16GB RAM. Whenever the algorithms have been run, learning data has been generated by selecting from different data. Thus, the consistent success of the algorithm on different learning data has been observed. In the calculation process, the algorithms were run 10 times. The accuracy of the algorithm was determined by the number of companies assigned to the correct class as a result of the classification process. The results are shown in Table 1.

As can be seen from the Figure 3, the algorithms have been compared and we find that learning data directly affects the success of the algorithm. 10 experiments, the accuracy of the KNN algorithm is higher than the Naive Bayes algorithm.

Table 1. The success of KNN and Naive Bayes algorithms

Running Number	KNN Accuracy (%)	KNN Correct Assignment	KNN Running Time (ms)	Naive Bayes Accuracy (%)	Naive Bayes Correct Assignment	Naive Bayes Running Time (ms)
1	87.9	691	11	83.5	657	9
2	86.2	678	15	77.7	611	5
3	84.4	664	14	78.5	617	5
4	86.1	677	16	80.2	631	5
5	85.2	670	14	82.1	646	5
6	85.6	673	12	81.9	644	9
7	84.9	668	10	79.1	622	11
8	84.6	665	15	77.1	606	8
9	85.5	672	18	86.1	655	9
10	84.7	666	12	75.9	597	10

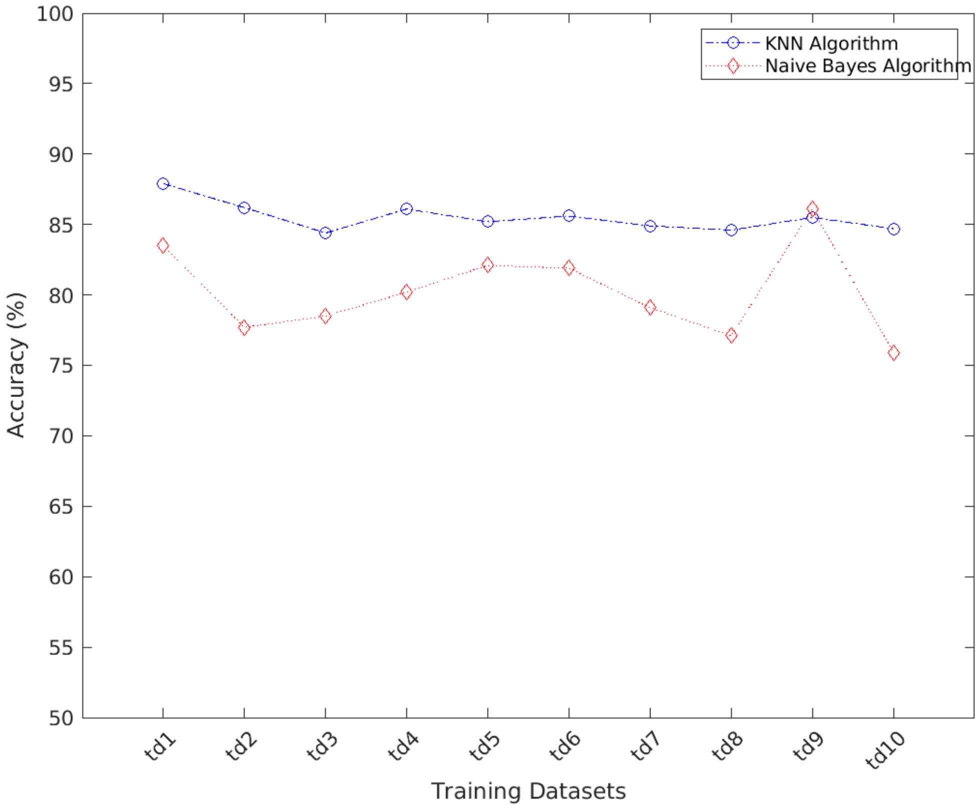

Figure 3. Comparison graph of success of the KNN and Naive Bayes algorithms

CONCLUSION

In this study, a data mining classification technique has been used in the financial situation analysis of firms. KNN and Naive Bayes algorithms have been selected for classification. Fifty data samples from each financial status group have been defined to the system as learning data. Results show that the KNN algorithm completed the classification process with a success of 84–88%, and the Naive Bayes algorithm did so with a success of 75–86%.

Learning data is very important for the success of the classification process. Differences in learning data may change the classification results. In this study, learning data has been randomly selected from each group.

Successful results have been obtained by using data mining classification techniques in the financial situation analysis of firms. This result increases the usefulness of data mining techniques in financial analysis processes. Future studies will aim to increase the accuracies by ensuring the algorithms work more effectively.

ACKNOWLEDGEMENTS

This study is supported by Ege University Scientific Research Projects Directorate with the project number 21759.

REFERENCES

Agrawal R, Imielinski T, Swami A, 1993. Database mining: A performance perspective. IEEE transactions on knowledge and data engineering, 5(6): 914-925.

Altman EI, 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The journal of finance, 23(4): 589-609.

Aktaş R, 2003. Mali Başarısızlığın Öngörülmesi: İstatistiksel Yöntemler ve Yapay Sinir Ağı Karşılaştırılması. Ankara Üniversitesi SBF Dergisi, 58(04).

Beaver WH, 1966. Financial ratios as predictors of failure. Journal of accounting research, 71-111.

Dasarathy BV, 1991. Nearest neighbor (NN) norms: NN pattern classification techniques. IEEE Computer Society Tutorial.

Dewi A, Hadri M, 2017. Financial distress prediction in Indonesia companies: finding an alternative model. Russian Journal of Agricultural and Socio-Economic Sciences, 61(1).

Fathi S, Saif S, Heydari Z, 2018. Predicting bankruptcy of companies using data mining models and comparing the results with Z Altman model. International journal of finance & managerial accounting, 3(10): 33-46.

Kaygın CY, Tazegül A, Yazarkan H, 2016. İşletmelerin Finansal Başarılı ve Başarısız Olma Durumlarının Veri Madenciliği ve Lojistik Regresyon Analizi ile Tahmin Edilebilirliği. Ege Academic Review, 16(1).

Kürklü E, Türk Z, 2017. Financial failure estimate in bist companies with Altman (Z-score) and Springate (S-score) models. Osmaniye Korkut Ata Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 1(1): 1-14.

Ohlson JA, 1980. Financial ratios and the probabilistic prediction of bankruptcy. Journal of accounting research, 109-131.

Patil TR, Sherekar SS, 2013. Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International journal of computer science and applications, 6(2): 256-261.

Özkan M, Boran L, 2014. Veri Madenciliğinin Finansal Kararlarda Kullanımı. Çankırı Karatekin Üniversitesi İİBF Dergisi, 4(1): 59-82.

Wu X, Kumar V, Ross Quinlan J, et al. 2008. Top 10 algorithms in data mining. Knowledge and information systems, 14: 1-37.