Aker Y. (2022). Comparison of PCA and RFE-RF algorithm in bankruptcy prediction. Gümüşhane Üniversitesi Sosyal Bilimler Dergisi, 13(3), 1001-1008

# Comparison of PCA and RFE-RF Algorithm in Bankruptcy Prediction

## İşletmelerin İflas Tahmininde PCA ve RFE-RF Algoritmasının Karşılaştırılması

**Yusuf Aker**[1]

## Abstract

*Machine learning prediction models are very important in detecting companies without going into financial distress and have recently become one of the most important research topics in empirical finance. While developing models in this area, data preprocessing steps are applied to make the data ready for analysis. One of these steps is the feature selection method, which can be defined as reducing the size of the financial ratios used as input in the data set. This stage is the process of choosing the best subset of features to be used in the research, or in other words, the selection of the most important features that can represent the data. In this paper, two different feature selection methods, Principal Component Analysis (PCA) and Random Forest - Recursive Feature Elimination (RF-RFE)) are compared. Commercial companies operating in Turkey were used in the experiments. The correct prediction success of the selected features was tested with AdaBoost and Stochastic Gradient Descent model. Our experimental results show that RF-RFE is a more efficient feature selection method compared to PCA.*

***Keywords:*** *Feature Selection, Principal Component Analysis, Random Forest-Recursive Feature Elimination, AdaBoost, Stochastic Gradient Descent*

## Öz

*Makine öğrenmesi tahmin modelleri, şirketlerin finansal sıkıntıya girmeden tespit edilebilmesi açısından çok önemlidir ve son zamanlarda ampirik finansın en önemli araştırma konularından birisi haline gelmiştir. Bu alanda modeller geliştirilirken veriyi analize hazır hale getirmek için veri ön işleme adımları uygulanmaktadır. Bu adımlardan birisi veri setinde girdi olarak kullanılan finansal oranların boyutunun küçültülmesi olarak tanımlanabilen özellik seçimi yöntemidir. Bu aşama araştırmada kullanılacak özelliklerin en iyi alt kümesini seçme veya başka bir deyişle veriyi temsil edebilecek en önemli özelliklerin seçimi sürecidir. Bu çalışmada Temel Bileşenler Analizi (Principal Component Analysis (PCA)) ve Rastgele Orman- Özyinelemeli Özellik Seçimi (Random Forest - Recursive Feature Elimination (RF-RFE)) olmak üzere iki farklı özellik seçim yöntemi karşılaştırılmıştır. Deneylerde Türkiye'de faaliyet gösteren ticari firmalar kullanılmıştır. Seçilen özelliklerin doğru tahmin başarısı AdaBoost ve Stokastik Gradient Descent modeli ile test edilmiştir. Deneysel sonuçlarımız, PCA ile karşılaştırıldığında, RF-RFE'nin daha etkili bir özellik seçim yöntemi olduğunu göstermektedir.*

***Anahtar Kelimeler:*** *Özellik Seçimi, Temel Bileşenler Analizi, Rastgele Orman- Özyinelemeli Özellik seçimi, AdaBoost, Stochastic Gradient Descent*

**Araştırma Makalesi [Research Paper]**

**JEL Codes:** G33, C81, C53

**Submitted:** 24 / 02 / 2022
**Accepted:** 08 / 09 / 2022

[1] Dr., Türkiye Finance Participation Bank, Branch Manager, yusuf_aker@yahoo.com, Samsun, Türkiye, Orcid No: https://orcid.org/0000-0002-6058-068X

## Introduction

In the recent global financial market, there have been radical changes in firm evaluation criteria with the effect of technological developments and big data. Following the developments in information technologies, the cheapening of hardware and the emergence of big data have led to the creation of large databases in many areas and the amount of data stored in these databases to increase exponentially. Many data mining methods have been developed due to the inadequacy of traditional methods in analyzing the said data stacks. Based on various definitions in the literature, data mining can be defined as a multi-stage process that aims to reveal hidden relationships, patterns and information in large data stacks by making use of different tools and technologies. One of the stages of this process is the feature selection process (Budak,2018). Feature selection can be defined as the selection of the subset that can best represent the original dataset. Feature selection is the process of selecting the best k feature among n features in the data set (Forman, 2003). The main purpose here is to shrink the data set by choosing the best separators among the features. In this way, the number of features will be reduced and this will bring many benefits to the analyst. The algorithm will gain speed with the data that is shrinking in size. Noisy and unnecessary data are removed from the data set and the data set becomes better quality and easier to identify. Besides making the visualization of the data easier, less memory is needed to store the data. All these advantages are factors that will increase the success of the model (Ladha and Deepa, 2011).

While making bankruptcy prediction models with machine learning, the size of the data is reduced by the feature selection method in the data preprocessing stage. Thus, it is intended to select a subset of the relevant features for use in model construction. The main purpose of this study is to test which method can choose the best subset of dataset more successfully. Why we chose these algorithms because these two algorithms use different methods while creating the subset. The most important difference was while RF-RFE selects given features without changing them, PCA transforms features into a lower dimension.

## 1. Literature Review

Tsai (2009) examined five feature selection methods (t-test, correlation matrix, stepwise regression, principle component analysis (PCA), factor analysis (FA)) and their prediction performance used in bankruptcy prediction in her study. According to the experimental result of the study, the t-test feature selection method performed better than the others. Liang at al. (2015), a comprehensive study is conducted to examine the effect of performing filter and wrapper based feature selection methods on bankruptcy prediction. Experimental results showed that feature selection does not always improve prediction performance, depending on the chosen techniques. However, at the same time, it is understood from the this study, performing genetic algorithm and logistic regression for feature selection can provide predictive improvements on the credit and bankruptcy datasets. Al-Nafjar (2022) used four different feature selection methods for benchmarking: principal component analysis (PCA), minimum redundancy maximum fitness, recursive feature elimination (RFE) and ReliefF. The study results showed that feature selection improved the performance of all classifiers.

Lu at al. (2007) successfully used the PCA method for face tracking and image acquisition. As a result of the study, they concluded that computers can train hundreds of concepts using pictures (Sun and Li, 2012). According to Parveen at al. (2012), PCA is a linear dimensionality reduction technique widely applied in datasets in all scientific fields. This study aims to contribute to the literature by comparing PCA's success in feature selection for bankruptcy prediction.

Recursive Feature Elimination is widely used in with many classification algorithms to create more efficient classifications such as support vector machines, random forest. RFE was initially implemented with support vector machines. From this model, the data set was trained and the features were ranked, and finally, the lowest ranked features were extracted (Gregorutti at al, 2017; Guyon at al, 2002; Jiang at al, 2004; Svetnik at al, 2004). Chen at al. (2020) examined the important features for the selection of the data. In this study, in which RFE is used as feature selection, the result shows that the RF method has high accuracy in all experiment groups. In the study of Chen et al., in which 1384 features were used, RF model provides 93.31 percent accuracy with six features and 93.36 percent accuracy with four features. This study shows that models that have a low number of features such as 4 or 7 features selected by RFE can have high accuracy. Introduced by Breiman (2001), random forest is a machine learning algorithm used widely. Granitto et al. (2006) used the RF-RFE algorithm to accomplish the feature selection in Proton Transfer Reaction study. And also, according to Voyle at al. (2016), RF-RFE has proven to be more effective compared to other methods and according to them, RF-RFE can use fewer features to achive a higher classification accuracy.

## 2. Feature Selection Methods

### 2.1. Pca

Numerous technologies are used to reduce data size, but one of the most used is PCA (Hasan and Abdulazeez, 2021). PCA is a size reduction process. It has the purpose of destroying the dependency structure between the variables or reducing the size; As it is an analysis technique on its own, it is also used as a data preparation technique for other analyses (Tatlıdil, 1996). With this method, it is aimed to derive new unrelated variables from the variables that are related to each other, in other words, to eliminate the multicollinearity problem. Obtaining the basic components can be explained as follows. Suppose there are initially $p$ variables $X1,....,Xp$. Let the system formed by these variables be represented by $\underline{X}'=( X1,....,Xp)$ the vector variables. In process $E(\underline{X}') = \mu$ *ve* $Cov(\underline{X}') = \Sigma$; $\mu$:*px1* represents the dimensional mass-mean vector and $\Sigma$: *pxp* dimensional mass-mean-variance covariance matrix. The principal components are as shown follow;

$Y1 = a11X1 + a12X2 + \dots + a1pXp$

$Y2 = a21X1 + a22X2 + \dots + a2pXp$

.

.

$Yp = ap1X1 + ap2X2 + \dots + appXp$

*Y1,..,Yp* are principal components, *a11, a21,...,app* are constant numbers and represent principal component loads. Principal component loads are weights that show the variance contribution of principal components on variables (Ozonur at al., 2019).

### *2.2. Rf-Rfe*

We can define Recursive Feature Elimination as the process of selecting estimators backwards (Guyon et al, 2002). In the model created by this method, each predictor is assigned an importance score. After removing the low-significant estimators, the model is rebuilt and the significance scores are recalculated. First of all, the size of the subset and hence the number of subset in relation to it is determined by the analyst. Because RFE is a tuning parameter. With this parameter, the data size is tried to be reduced by selecting the features that can best represent the data. These features will then be used to train the model.

## 3. Methodology and Data

Our data consists of annual financial statements of firms operating in Turkey from 2015 up to 2019. The firms in the study were selected from different sectors such as education, energy, furniture, transportation, mining, automotive, textile, tourism. Firms for which bankruptcy or concordat decisions were made by the commercial courts in the 2018-2019 period were accepted as distressed. The data collected for distressed firms includes annual data two (2016) and three years (2015) before the the judgment date. Our data consists of 166 non-distressed and 219 distressed firms for 2015, and 169 non-distressed and 211 distressed firms for 2016. Data was randomly subsampled as 70% training set and 30% test set. The training set is used to train the prediction model. The model calculates the "prediction" values from the training result. The model is evaluated by comparing it with the test set, which consists of data that is not included in the training set. In this study, 47 initial financial ratios were used and the selection of these ratios is based on previous studies (Mselmi, Lahiani & Hamza, 2017; Aksoy & Boztosun, 2018; Yürük & Ekşi, 2019). These ratios are selected from among the liquidity ratios, financial structure ratios, profitability ratios and turnover ratios.

**Table 1. Methodology of data analysis**

| | | |
|---|---|---|
| *Data pre-processing* | | |
| Process | Obtaining 2015-2016 distressed and non-distressed firms data<br>Missing data detection - Replacing missing values with median<br>Outlier detection - Replacing outliers with Tukey method<br>Data normalization<br>Splitting the data for training and testing (%70 train - %30 test)<br>PCA and RF-RFE Feature Selection | |
| *Analysis methods* | | |
| Methods | AdaBoost | Stochastic Gradient Descent |

The research methodology is planned as follows; obtaining the data set, detecting the missing values in the data set and filling it with the median values, detecting the extreme values in the data set and pulling them to the normal limits according to the Tukey method, separating the data set into 70% training and 30% test set, feature selection with PCA and RF-RFE methods and finding the best subsets of data set and applying each subset to the AdaBoost and Stochastic Gradient Descent (SGD) method. The main purpose is to determine which of the PCA or RF-RFE method has higher accuracy in the established model.

In this study, 47 primary financial ratios, which are the data set, were first reduced in size according to the PCA method. Then, again, among 41 primary financial ratios, the features that can best represent the data set were selected according to the RF-RFE method without changing the data. Finally, these two subsets were applied to the AdaBoost and and stochastic gradient descent model, and the model with the highest accuracy was tried to be found.

**Table 2. Selected initial financial ratios**

| Variable | Meaning | Variable | Meaning |
|---|---|---|---|
| **Liquidity Ratios** | | R25 | Net tangible assets/Total assets |
| R01 | Current ratio | **Profitability Ratios** | |
| R02 | Liquidity ratio | R26 | Net income after tax/Net sales |
| R03 | Cash ratio | R27 | Cost of goods sold/Net sales |
| R04 | Stocks/Current assets | R28 | Gross sales margin/Net sales |
| R05 | Stocks/Total assets | R29 | Operational expenses/Net sales |
| R06 | Stock dependency ratio | R30 | Operating profits/Net sales |
| R07 | Short-term trade receivables/Current ass. | R31 | Operating profits/(Tot.assets-Financial tangible assets) |
| R08 | Short-term trade receivables/Total assets | R32 | Financial expense/Net sales |
| **Financial Structure Ratios** | | R33 | (Fin.expense+Net income before tax)/Total Liabilities |
| R09 | Total foreign assets | R34 | (Fin.expenses+Profit after tax)/Financial Expenses |
| R10 | Debt Ratio | R35 | (Fin. expenses+İncome before tax)/Financial Expenses |
| R11 | Equities/Total foreign assets | R36 | Net profit after tax/Equities |
| R12 | Short term liabilities/Foreign assets | R37 | Profit before tax/Equities |
| R13 | Short term liabilities/Total liabilities | R38 | Net profit after tax/Total assets |
| R14 | Bank loans/Total assets | R39 | (Retained earnings+Reserves)/Total assets |
| R15 | Bank loans/Total foreign assets | **Turnover Rates Ratios** | |
| R16 | Short term bank loans/Short term liabilities | R40 | Equity turnover |
| R17 | Long-term liabilities/Total liabilities | R41 | Working capital turnover |
| R18 | Long-term liabilities/Constant capital | R42 | Net working capital turnover |
| R19 | Current assets/Total assets | R43 | Asset turnover |
| R20 | Fixed assets/Equities | R44 | Accounts receivable turnover |
| R21 | Fixed assets/Total foreign assets | R45 | Stock turnover |
| R22 | Fixed assets/Constant capital | R46 | Fixed asset turnover |
| R23 | Net tangible assets/Equities | R47 | Net tangible asset Turnover |
| R24 | Net tangible assets/Long-term liabilities | | |

Note: This table shows all primary financial ratios used in the analysis..

## 4. Emperical Results

### 4.1. Pca

The accuracy rates of the models in which PCA is used for feature selection are given in the appendix as the confusion matrix.

In t-3, 15 components were used for PCA and the variant ratio was 91.36%. In t-2, 15 components were used for PCA and the variant ratio was 90.18%. As seen in table 3, two years prior to failure, the AdaBoost model correctly classified 45 of the 61 firms that actually failed in the test set. 25 of the 53 non-distress firms were classified correctly and included in the successful firm category. Three years prior to failure, at the AdaBoost model testing set, 49 of the 64 firms that actually failed were correctly classified. 35 of the 52 non-distress firms were classified correctly and included in the successful firm category.

Two years prior to failure, the SGD model correctly classified 39 of the 61 firms that actually failed in the test set. 32 of the 53 non-distress fims were classified correctly and included in the successful firm category. Three years prior to failure, at the SGD model testing set, 49 of the 64 firms that actually failed were correctly classified. 34 of the 52 non-distress firms were classified correctly and included in the successful firm category.

**Table 3. Confusion Matrix of PCA with AGD and AdaBoost.**

| | t-3 / test sample | | | t-2 / test sample | | |
|---|---|---|---|---|---|---|
| | **0** | **1** | **Sum** | **0** | **1** | **Sum** |
| AdaBoost | | | | | | |
| **0** | 77 | 23 | 100 | 74 | 26 | 100 |
| | (49/64) | (15/64) | | (45/61) | (16/61) | |
| **1** | 33 | 67 | 100 | 53 | 47 | 100 |
| | (17/52) | (35/52) | | (28/53) | (25/53) | |
| SGD | | | | | | |
| **0** | 77 | 23 | 100 | 64 | 36 | 100 |
| | (49/64) | (15/64) | | (39/61) | (22/61) | |
| | 35 | 65 | 100 | 41 | 59 | |
| **1** | (18/52) | (34/52) | | (21/53) | (32/53) | 100 |

Note: 0 represents the distressed firms, 1 respresents non-distressed firms.

### 4.2. Rf-Rfe

The accuracy rates of the models in which RF-RFE is used for feature selection are given in the appendix as the confusion matrix.

The data set consists of 380 companies for t-2 and 385 companies for t-3. The data in table 4 were normalized and the extreme values were taken to normal limits according to the Tukey method. The majority of correlation between independet variables are low. For three years prior to failure, the most discriminant financial ratios selected by RF-RFE are stock dependency ratio (R6), equities/total foreign assets (R11), short term bank loans/Short term liabilities (R16), net tangible assets/long-term liabilities (R24), financial expense/net sales (R32), financial expenses+İncome before tax)/financial expenses (R35), profit before tax/equities (R37), net working capital turnover (R42) and fixed asset turnover (R46). For two years prior to failure, the most discriminant financial ratios selected by RF-RFE are R6, R11, R14, R18, R24, R30, R32, R34 AND R47. The independent variables R6, R11, R24, and R32 were the covariates selected in both years.

**Table 4. Summary statistics selected by RFE-RF**

| Variables | | **R6** | **R11** | **R16** | **R24** | **R32** | **R35** | **R37** | **R42** | **R46** |
|---|---|---|---|---|---|---|---|---|---|---|
| t-3 | Count | 385 | 385 | 385 | 385 | 385 | 385 | 385 | 385 | 385 |
| | Mean | 0.47 | 0.46 | 0.22 | 0.58 | 0.19 | 0.54 | 0.51 | 0.49 | 0.22 |
| | Std | 0.20 | 0.17 | 0.27 | 0.18 | 0.26 | 0.19 | 0.25 | 0.26 | 0.25 |
| | Min | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Max | 0.36 | 0.35 | 0.00 | 0.54 | 0.01 | 0.41 | 0.38 | 0.38 | 0.06 |
| | 25% | 0.41 | 0.40 | 0.13 | 0.61 | 0.08 | 0.63 | 0.46 | 0.46 | 0.13 |
| | 50% | 0.55 | 0.51 | 0.42 | 0.61 | 0.29 | 0.63 | 0.62 | 0.62 | 0.31 |
| | 75% | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Variables | | **R6** | **R11** | **R14** | **R18** | **R24** | **R30** | **R32** | **R34** | **R47** |
| t-2 | Count | 380 | 380 | 380 | 380 | 380 | 380 | 380 | 380 | 380 |
| | Mean | 0.47 | 0.41 | 0.27 | 0.47 | 0.57 | 0,45 | 0.18 | 0.55 | 0.46 |
| | Std | 0.20 | 0.17 | 0.25 | 0.14 | 0.21 | 0.21 | 0.24 | 0.22 | 0.18 |
| | Min | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Max | 0.35 | 0.31 | 0.02 | 0.34 | 0.46 | 0.33 | 0.01 | 0.41 | 0.36 |
| | 25% | 0.40 | 0.35 | 0.23 | 0.43 | 0.60 | 0.42 | 0.10 | 0.59 | 0.40 |
| | 50% | 0.56 | 0.46 | 0.45 | 0.59 | 0.62 | 0.55 | 0.26 | 0.63 | 0.52 |
| | 75% | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

As seen in table 5, two years prior to failure, the AdaBoost model correctly classified 55 of the 61 firms that actually failed in the test set. 49 of the 52 non-distress firms were classified correctly and included in the successful firm category. Three years prior to failure, at the AdaBoost model testing set, 60 of the 64 firms that actually failed were correctly classified. 45 of the 52 non-distress firms were classified correctly and included in the successful firm category. Two years prior to failure, the SGD model correctly classified 55 of the 61 firms that actually failed in the test set. 27 of the 53 successful firms were classified correctly and included in the successful firm category. Three years prior to failure, SGD model testing set, 58 of the 64 firms that actually failed were correctly classified. 35 of the 52 successful firms were classified correctly and included in the successful firm category.

**Table 5. Confusion Matrix of RF-RFE with AGD and AdaBoost**

| | | t-3 / test sample | | | t-2 / test sample | | |
|---|---|---|---|---|---|---|---|
| | | **0** | **1** | **Sum** | **0** | **1** | **Sum** |
| | AdaBoost | | | | | | |
| **0** | | 94 | 6 | 100 | 90 | 10 | 100 |
| | | (60/64) | (4/64) | | (55/61) | (6/61) | |
| **1** | | 13 | 87 | 100 | 8 | 92 | 100 |
| | | (7/52) | (45/52) | | (4/53) | (49/52) | |
| | SGD | | | | | | |
| **0** | | 91 | 9 | 100 | 90 | 10 | 100 |
| | | (58/64) | (6/64) | | (55/61) | (6/61) | |
| | | 33 | 67 | 100 | 49 | 51 | |
| **1** | | (17/52) | (35/52) | | (26/53) | (27/53) | 100 |

In machine learning, the performance of the model is often measured with confusion matrice. This matrix shows the results by comparing the estimated values with the actual values. Model results will belong to one of the four evaluations. (Chen at al, 2020):

Accuracy=(TP+TN)/(TP+TN+FP+FN)
Precision=(TP)/(TP+FP)
Recall=(TP)/(TP+FN)
F1 Score=2*[(Precision*Recall)/( Precision+Recall)

where: TP = True positive; FP = False positive; TN = True negative; FN = False negative.

**Table 6. Classification performance**

| | PCA | | | | RF-RFE | | | |
|---|---|---|---|---|---|---|---|---|
| model | AdaBoost | | Sgd | | AdaBoost | | Sgd | |
| time | t-3 | t-2 | t-3 | t-2 | t-3 | t-2 | t-3 | t-2 |
| accuracy | 65 | 61 | 64 | 62 | 91 | 91 | 80 | 72 |
| precision | 69 | 62 | 67 | 65 | 90 | 93 | 77 | 68 |
| recall | 64 | 74 | 67 | 64 | 94 | 90 | 91 | 90 |
| f1-score | 67 | 67 | 67 | 64 | 92 | 92 | 83 | 77 |

As a result of the research, when we compare the accuracy rates of the models in the analysis made with PCA, the AdaBoost model has an accuracy rate of 65% at year t-3 and 61 at year t-2. In the analysis made with SGD, 64% accuracy rate for t-3 year and 62% accuracy rate for t-2 year was achieved. In the research conducted with RF-RFE, the analysis performed with the AdaBoost model had an accuracy rate of 91% in t-3 and t-2 years. In the analysis made with the SGD model, the accuracy rate was also 80% at t-3 and 72% at t-2.

## Conclusions

Focusing on feature selection in bankruptcy forecasting, this work is built on a comparison of two different methods. In both methods, bankruptcy prediction will be made on the data set by reducing the size of the data set. With the first feature called PCA, a new subset is obtained by reducing the size of the dataset. In the second feature, called RF-RFE, a new subset was created from the data set by selecting the most important features without changing the data set. The created subsets were applied to the AdaBoost and SGD model. The empirical results show that the subset selected by RF-RFE achieves higher classification success in both AdaBoost and SGD models. This study focused on the effect of two different feature selections on outcomes in bankruptcy prediction. Apart from the feature selection methods used in the study, there are many methods. Study does not aim to generalize. However, it is aimed to contribute to the studies in this field. The work to be done in the field of bankruptcy forecasting will never end. It is thought that inclusion of more feature selection methods in new studies and their use of larger databases will be beneficial in terms of obtaining more precise and reliable results.

## References

Al-Nafjan A. 2022. Feature selection of EEG signals in neuromarketing. *PeerJ Computer Science* 8:e944 https://doi.org/10.7717/peerj-cs.944

Aksoy, B., & Boztosun, D. (2018). Diskriminant ve Lojistik Regresyon Yöntemleri Kullanlarak Finansal Başarısızlık Tahmini: BİST İmalat Sektörü Örneği. *Finans Politik & Ekonomik Yorumlar Dergisi*, 646, 9–32.

Breiman, L. (2001). Random Forests. *Mach. Learn.* (*45)*, 5–32.

Budak, H. (2018). Feature Selection Methods and a New Approach, Süleyman Demirel University Journal of Natural and Applied Sciences, 22, (Private-10) -1-3. DOI: 10.19113/sdufbed.01653.

Chen, R., Dewi, C., Huang, S., and Caraka, R.E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, volume (7). p,1-7.

Forman, G. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification, Journal of Machine Learning Research, 3, 1289–1305.

Granitto, P.M.; Furlanello, C.; Biasioli, F.; Gasperi. (2006). F. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemom. Intell. Lab. Syst. 83*, 83–90.

Gregorutti B, Michel B, Saint-Pierre P. (2017).Correlation and variable importance in random forests. Stat Comput. 27:659–78.

Guyon, J. Weston, S. Barnhill, and V. Vapnik. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1-3):389–422.

Hasan, B. M. S. and Abdulazeez, A. M., (2021). A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. Journal of Soft Computing and Data Mining Vol. 2 No. 1 pp. 20

Jiang H, Deng Y, Chen HS, Tao L, Sha Q, Chen J, Tsai CJ, Zhang S. (2004). Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. BMC Bioinformatics. 5:81.

Ladha, L., Deepa, T. (2011). Feature Selection Methods And Algorithms, International Journal on Computer Science and Engineering, 3(5), 1787-1797.

Liang, D., Tsai, C. and Wu, H. (2015). The effect of feature selection on financial distress prediction. *Knowledge-Based Systems*. v. 73. p. 289-297. https://doi.org/10.1016/j.knosys.2014.10.010

Lu, Y., Cohen, I.,Zhou, X. Z. and Tian, Q. (2007). Feature selection using principal feature analysis. Proceedings of the 15th ACM international conference on Multimedia p. 301–304 https://doi.org/10.1145/1291233.1291297

Mselmi, N., Lahiani, A. & Hamza, T. (2017). Financial distress prediction: The case of French small and medium-sized firms, *İnternational Review of Financial Analysis,*(50), 67-80.

Ozonur, D., Kılıç, D.,Akdur, H.T.K. and Bayrak, H. (2019). Multi Response Optimization in Food Industry Using Principal Component Analysis and Response Surface Methodology. Erzincan University Journal of Science and Technology. 12(2), 734-744. DOI:10.18185/erzifbed.485762

Parveen, A., Inbarani, H., and SatishKumar, E. (2012). Performance Analysis of Unsupervised Feature Selection Methods. Computing, Communication and Applications (ICCCA), 2012 International Conference. DOI:10.1109/ICCCA.2012.6179181

Sun, J., and Li, H., (2012). Financial distress prediction using support vector machines: Ensemble vs. individual. *Applied Soft Computing*. 12(8). P.2254-2265. https://doi.org/10.1016/j.asoc.2012.03.028

Svetnik V, Liaw A, Tong C, Wang T. (2004). Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. In: Roli F, Kittler J, Windeatt T, editors. Multiple classifier systems. Berlin: Springer.

Tatlıdil, H.(1996). Uygulamalı Çok Değişkenli İstatistiksel Analiz, Ankara: Akademi Matbaası, 1996, 138, 146.

Tsai, C. (2009). Feature selection in bankruptcy prediction. *Knowledge-Based Systems*. 22(2), p. 120-127. https://doi.org/10.1016/j.knosys.2008.08.002

Voyle, N., Keohane, A., Newhouse, S., Lunnon, K., Johnson, C., Soininen, H., Kloszewska, I., Mecocci, P., Tsolaki, M., Vellas, B., et al. (2016). A pathway based classification method for analyzing gene expression for Alzheimer’s disease diagnosis, *Journal of Alzheimer's Disease,* 49, 659–669.

Yürük, M. F., & Ekşi, H. İ. (2019). Yapay Zekâ Yöntemleri İle İşletmelerin Finansal Başarısızlığının Tahmin Edilmesi: BİST İmalat Sektörü Uygulaması. *Mukaddime*, 10(1), 393–422.