Ekonomi, Politika & Finans Araştırmaları Dergisi, 2025, 10(Özel Sayı): 383-402
Journal of Research in Economics, Politics & Finance, 2025, 10(Special Issue): 383-402
Araştırma Makalesi / Research Article, https://doi.org/10.30784/epfad.1813752

THE ROLE OF FINANCIAL MARKETS IN PREDICTING BIST SUSTAINABILITY INDEX PERFORMANCE: NEW EVIDENCE FROM HYBRID MACHINE LEARNING MODELS

BİST Sürdürülebilirlik Endeksi Performansının Tahmininde Finans Piyasalarının Rolü: Hibrid Makine Öğrenmesi Modellerinden Yeni Kanıtlar

Zeynep ÇOLAK*

Keywords:
Sustainable Finance, BIST Sustainability Index, Machine Learning, SHAP, Explainable Artificial Intelligence

JEL Codes:
G19, C53, C58

Abstract

The increasing importance of sustainable finance makes it critical to understand and accurately model the performance dynamics of investment instruments in this area. This study aims to forecast the return of the BIST Sustainability Index using financial market indicators and to explain the underlying dynamics of this forecasting process, thereby understanding the complex structures of financial markets, investor behavior, and information flow. In this study, eleven different machine learning models were compared with a validation strategy suitable for the time series structure, and the most successful candidates were subjected to hyperparameter optimization. In order to overcome the limitations of single models, a sequential hybrid model based on the Residual Fitting approach was developed. According to the results of the study, the two-stage hybrid model, which uses the Voting Regressor as the main predictor and Random Forest as the error corrector, provided the lowest error (RMSE) and the highest R^2 value. The findings indicate that the BIST_100 index is the most critical determinant, while risk aversion indicators such as Gold, USD, and VIX have a negative effect. This evidence has far-reaching implications for understanding the dynamic relationships between the Sustainability Index and macroeconomic variables.

Anahtar Kelimeler:
Sürdürülebilir Finans, BIST Sürdürülebilir Endeksi, Makine Öğrenimi, SHAP, Açıklanabilir Yapay Zeka

JEL Kodları:
G19, C53, C58

Öz

Sürdürülebilir finansmanın artan önemi, bu alandaki yatırım araçlarının performans dinamiklerini anlamayı ve doğru bir şekilde modellemeyi kritik hale getirmektedir. Bu çalışma, finansal piyasa göstergelerini kullanarak BIST Sürdürülebilirlik Endeksi'nin getirisini tahmin etmeyi ve bu tahmin sürecinin altında yatan dinamikleri açıklamayı, böylece finansal piyasaların karmaşık yapılarını, yatırımcı davranışlarını ve bilgi akışını anlamayı amaçlamaktadır. Bu çalışmada, zaman serisi yapısına uygun bir doğrulama stratejisi ile on bir farklı makine öğrenimi modeli karşılaştırılmış ve en başarılı adaylar hiperparametre optimizasyonuna tabi tutulmuştur. Tekil modellerin sınırlamalarını aşmak için, Residual Fitting yaklaşımına dayalı sıralı bir hibrit model geliştirilmiştir. Çalışmanın sonuçlarına göre, ana tahminci olarak Voting Reressor ve hata düzeltici olarak Rastgele Orman kullanan iki aşamalı hibrit model, en düşük hata (RMSE) ve en yüksek R^2 değerini sağlamıştır. Bulgular, BIST_100 endeksinin en kritik belirleyici olduğunu, Altın, USD ve VIX gibi riskten kaçınma göstergelerinin ise olumsuz bir etkiye sahip olduğunu göstermektedir.

* Assist. Prof. Dr., Çanakkale Onsekiz Mart University, Administrative Sciences Department of Business Administration, Department of Numerical Methods, Türkiye, zolak.84@gmail.com

Received Date (Makale Geliş Tarihi): 27.08.2025 Accepted Date (Makale Kabul Tarihi): 17.10.2025

1. Introduction

In countries where sustainability reports are prepared voluntarily, sustainability indices are seen to significantly encourage businesses to prepare and publish sustainability reports. The most important feature of stock exchanges is that they showcase companies in all their aspects and contribute to the formation of transparent and orderly markets by reinforcing the race for excellence among businesses (Kocamış and Yıldırım, 2016). The distinction between sustainable and non-sustainable businesses, the presentation of this distinction to business stakeholders, and the assessment of sustainability performance have led to the development of the BIST Sustainability Index. This initiative was driven by the need for transparent, sustainable businesses in the market, as well as the recognition that the most effective method for achieving this is through performance measurement using indices.

Companies' performance in these areas is explained through Environmental, Social, and Governance (ESG) factors. ESG is described as a non-financial assessment system that considers the environment, society, and corporate governance to support companies' sustainable development and encourages companies to focus on social interests rather than maximizing their own interests (Chen et al., 2023). The performance of sustainable indices is influenced not only by companies' ESG scores but also by macroeconomic and financial indicators (Friede et al., 2015).

The use of Machine Learning models in the financial sector is rapidly expanding. The vast amount of data generated by the sector while performing its financial intermediation functions also provides a favorable working environment for these models (Şahin, 2024). In machine learning, algorithms have been developed that can process large amounts of nonlinear data in modeling frameworks by establishing complex, advanced neural network structures (Seow, 2025). The majority of these models are inherently complex and lack explanations of the decision-making process, causing these models to be termed as 'Black-Box'. (Quinn, 2023). SHAP (SHapley Additive Explanations) is one such method, and it takes the machine learning model out of the black box, allowing for commentary on the model (Lundberg and Lee, 2017). The SHAP method is an XAI method that focuses on identifying the contribution of features to the output, utilizing the mathematical concept known as the Shapley value in game theory. The Shapley value used in this method represents the average marginal contribution of each feature value among all possible values in the feature space. SHAP can be used for both global and local explanations (Bhattacharya, 2022).

The purpose of this study is to predict the returns of the BIST Sustainability Index and its relationship with financial and macroeconomic variables using machine learning models. The role of the variables behind these predictions is then evaluated and interpreted using SHAP analysis. The original contribution of this study to the literature is the development of a sequential hybrid machine learning model based on the Residual Fit approach to predict the returns of the BIST Sustainability Index and interpretation using SHAP analysis.

2. Conceptual Framework

2.1. The Effect of Macroeconomic Variables on Sustainability Indices

Analyzing the financial performance of sustainability indices, it is equally important to demonstrate the interaction of global and local macroeconomic dynamics on these indices as it is to show the impact of ESG factors.

Drimbetas et al. (2010) investigated the effects of macroeconomic factors on the sustainability index in their study, analyzing the relationship between DJSI data, oil prices, 10-year bond prices, exchange rates, and non-agricultural employment data using monthly data from 1999 to 2008. The GARCH model was used in the study. The study concluded that there is a negative relationship between oil prices and exchange rates, as well as the sustainability index. A positive relationship was found between 10-year bonds and the sustainability index, while no relationship was observed between non-agricultural employment and the sustainability index.

In Sharma et al. (2021), the impact of macroeconomic variables on India's sustainability indices was analyzed. The study found that while there was a relationship between the GREENEX index and crude oil prices, interest rates were not related to the index. Kaur and Chaudary (2022) analyzed the relationship between the sustainable stock market index and macroeconomic variables. The result of the study showed that macroeconomic variables have a long-term equilibrium connection with sustainable stock market prices

Özçim (2022) showed that the oil variable did not affect the volatility of the BIST Sustainability Index, while the exchange rate variable increased it, and the interest rate variable decreased it. Kaya (2023) demonstrated that oil-based fuel prices have a more significant impact on the BIST Sustainability Index than prices for natural gas and coal. Kavas (2025) found a positive relationship between the BIST Sustainability Index and the exchange rate.

2.2. Sustainable Finance and Green Bond

Green bonds are one of the instruments that are used to finance environmentally friendly projects. The proceeds from green bonds help businesses raise capital for environmentally friendly projects and contribute to sustainable development for the future. Issuing green bonds involves certain costs, and investors are reluctant to invest in these bonds due to the perceived risk associated with the projects they finance (Bhutta et al., 2022).

Ehlers and Packer (2017) emphasized that conducive market conditions must be in place for the growing green bonds. Both issuers and investors should be satisfied with the returns and safety of such a security. AlGhazali et al. (2025) examined the relationship between sustainability indices, green bond markets, and oil price shocks. The findings indicate that there is a changing connection between all variables over time.

According to the results of Başkaya (2025), a positive long-term relationship was found between the BIST100 Index and the BIST Sustainability Index, while a significant negative relationship was found between the BIST100 Index and the S&P Green Bond Index.

2.3. Machine Learning Applications in Sustainable Finance

In recent years, methods such as Interpretable Machine Learning or Explainable Artificial Intelligence have begun to be used in studies related to sustainability in the financial sector. Zhang and Zhao (2026) developed a prediction model for corporate ESG ratings using an XGBoost algorithm enhanced with SHAP interpretability. Siddique and Karim (2025) employed machine learning, deep learning, and ensemble techniques to assess whether ESG and financial indicators can effectively predict carbon risk. Results demonstrate that advanced AI models significantly outperform traditional regressions by capturing complex, non-linear relationships often overlooked by conventional methods. SHAP analysis further identifies environmental disclosure as the most influential predictor. Çankal and Ever (2025) analyzed the relationship between the financial performance of companies listed on the Borsa Istanbul (BIST) Sustainability Index and their renewable energy consumption using the Explainable Artificial Intelligence method.

A review of the literature reveals that there is a lack of studies examining the impact of sustainability indices on macroeconomic variables and financial indicators for emerging markets. This study aims to fill this gap in the literature.

3. Data Set and Methodology

3.1. Data Set and Variables

The dataset used in this study consists of multivariate time series data covering 2015-2025. The data was obtained from Investing.com. The dependent variable of the study, the BIST Sustainability Index, represents companies listed on the Istanbul Stock Exchange that have high ESG scores and strong corporate sustainability performance. The independent variables include financial and macroeconomic indicators that affect the performance of the sustainability index. These variables are the BIST 100 Index, the main index of the Istanbul Stock Exchange; the Turkey 10-Year Bond Yield, an important tool for national economies and monetary policy; the dollar exchange rate, which expresses the value of the US dollar against the local currency; the S&P 500 Index, which consists of the shares of the 500 largest publicly traded companies in the United States and is weighted by market value; the CBOE Volatility Index, which is based on the fundamental principle that trading volumes and stock option pricing are determined by investors; brent crude oil and gold prices.

3.2. Data Preprocessing and Feature Engineering

By their very nature, financial time series often exhibit non-stationarity. This means that the statistical properties of the series, such as mean and variance, change over time, which contradicts the basic assumptions of many econometric and machine learning models (Tsay, 2010). To improve the performance of the models and to approximate the stationarity assumption, all crude price series were transformed into percentage return series. This transformation was done using first-order differencing between the series. After the transformation, the missing data in the first row (NaN) due to the return calculation were removed from the data set.

3.3. Data Set Partitioning and Validation Strategy

In time series data, there is a temporal dependence between observations. Therefore, data partitioning methods based on random shuffling, such as standard cross-validation, run the risk of "data leakage" (Bergmeir and Benítez, 2012), where the model learns from future information to predict the past. This produces misleadingly high accuracy rates that do not reflect the real-world performance of the model.

To avoid this methodological error, the dataset was split without shuffling (shuffle=False), preserving the temporal order. The first 75% of the dataset is used as training data, and the last 25% is used as test data. This fixed-origin validation strategy simulates a realistic forecasting scenario where the model learns only from past data and is tested on future data that it has never seen before.

3.4. Feature Scaling

The majority of the algorithms evaluated in this study are decision tree-based ensemble models (e.g., Random Forest, XGBoost). These models are insensitive to the scale of features as they partition the feature space parallel to the axes, i.e., they are not affected by monotonic transformations of features (Hastie et al., 2009). Therefore, feature scaling (normalization or standardization) is not a prerequisite for these models. In order to maintain methodological consistency and apply a uniform preprocessing pipeline to all models, no further scaling is performed, even for scale-sensitive models such as SVR and MLP Regressor.

3.5. Control and Evaluation of Excessive Learning

Overfitting occurs when a model loses its ability to generalize to new data by memorizing noise and random fluctuations in the training data, primarily due to the model's high variance (Hastie et al., 2009). In this study, the risk of overlearning was managed and assessed in the following ways:

1. Performance Comparison: The presence and degree of overlearning were quantitatively determined by comparing the performance of each model on the training set with its performance on the test set. A significant difference between the training and test metrics was considered a strong indicator of overlearning.
2. Regularization: Models such as Ridge and Lasso naturally include L2 and L1 regularization mechanisms that penalize coefficients (Tibshirani, 1996). Similarly, algorithms such as Gradient Boosting and XGBoost have regularization parameters that control tree complexity and leaf node values.
3. Ensemble Methods: Bagging-based methods, such as Random Forest, reduce variance by averaging over a large number of models (Breiman, 2001). Meta-aggregation methods such as Voting and Stacking aim to produce more robust and generalizable predictions by combining the biases of different model architectures (Wolpert, 1992). In this study, the default hyperparameters of the models are used, which provide a basic level of regularization.

3.6. Descriptive Statistics

Table 1 presents the basic descriptive statistics of the BIST, CBOE VIX, and S&P indices. The findings reveal that the series have different characteristics in terms of their mean levels, the shape of their distributions, and extreme value characteristics.

Table 1. Descriptive Statistics

Variable	Mean	Median	Mx	Min	Std. Dev.	Skewness	Kurtosis	Jarque-Bera	Probability
BIST	4062.09	1480.44	15440.06	868.47	4503.75	13.286	0.1420	740.71	0.00
CBOE VIX	18.49	16.70	82.69	09.14	07.36	25.558	124.214	18825.10	0.00
SP	1798.25	1789.35	2406.19	1282.51	263.94	0.0315	-0.8658	79.06	0.00

First of all, the average value of the BIST index was 4062.09, with a low of 868.47 and a high of 15,440.06. The high standard deviation of 4503.75 indicates that the index exhibited significant fluctuations throughout the period. The positive skewness coefficient (1.33) indicates that the distribution is skewed to the right, i.e., high values pull most of the observations upwards. The kurtosis value (0.14) is close to normal, indicating that the extreme value density of the distribution does not increase significantly.

CBOE VIX index results show an average value of 18.49 and a highest observation of 82.69. This indicates that the index can occasionally reach very high levels due to market uncertainties. S&P index results show an average value of 1798.25, ranging between 12. Figure 1 shows the evolution of the BIST SE index over the period 2015-2025.

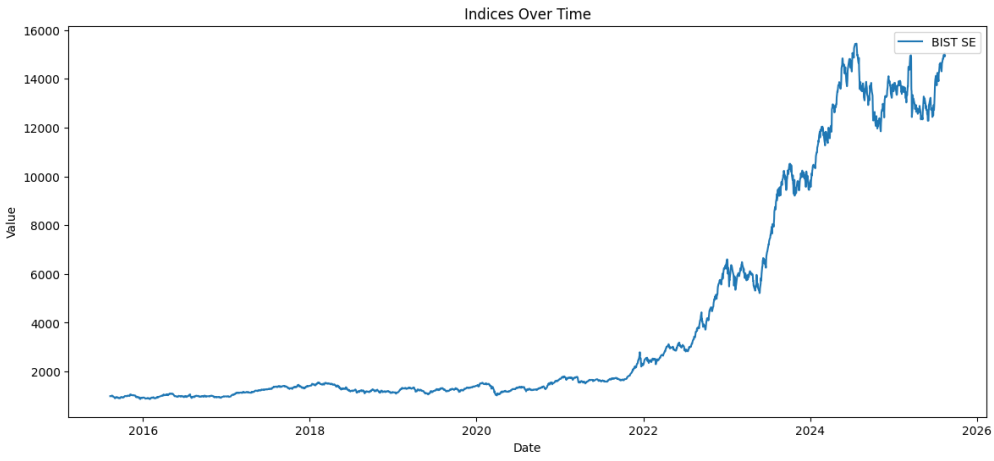

Figure 1. Time Series Dynamics of BIST SE Index

The graph shows that the series remained flat until 2020, then entered a sharp upward trend. Overall, the BIST Sustainability Index demonstrates both growth potential and vulnerability to shocks, in line with the characteristics of emerging financial markets.

4. Methodology

4.1. Voting Regressor

Voting Regressor is a meta-ensemble model that combines the predictions of different machine learning models to produce a final result. The main goal of this approach is to balance the bias or variance that a single model may have by utilizing the collective wisdom of models with different architectures and learning approaches (Dietterich, 2000). Figure 2 shows the schematic architecture of the Voting Regressor model.

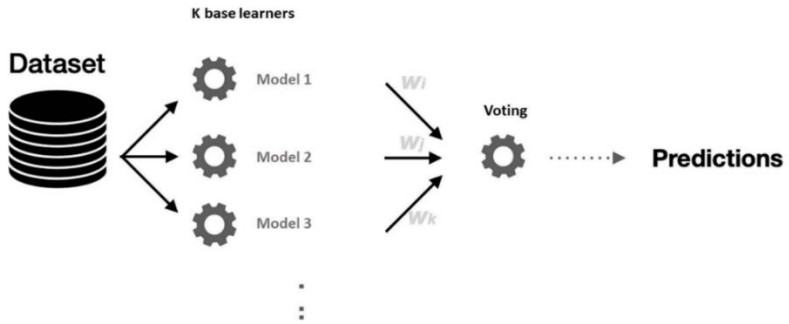

Figure 2. Schematic Architecture of the Voting Regressor Model

As shown schematically in Figure 2, the Voting Regressor architecture starts by presenting the same dataset to multiple independent base learners (K base learners). Each base model (Model 1, Model 2, ...) processes the data according to its own internal learning algorithm and produces an independent prediction (Wi, Wj, ..., Wk). In the final stage, these individual predictions are combined in a voting mechanism. In regression problems, this voting mechanism usually takes a simple or weighted average of all the individual predictions to produce a more robust and usually more accurate final prediction.

In this study, we take advantage of this architecture to integrate the predictions of two structurally different models: (1) a Random Forest optimized to capture nonlinear and complex relationships and (2) a Ridge regression that models more stable and linear relationships through regularization. This diversity allows the models to capture different types of patterns in the data and compensate for each other's weaknesses.

Mathematically, the final estimate ($\hat{y}$) of a Voting Regressor consisting of M base models is expressed as the simple average of the estimate ($h_m(x)$) of each base model, as follows:

$$\hat{y}(x) = \frac{1}{M}\sum_{m=1}^{M} h_m(x) \quad (1)$$

This simple averaging process softens the effect of extreme outlier predictions of a single model and provides a more stable generalization performance (Hastie et al., 2009).

4.2. Random forest

The Random Forest (RF) algorithm has been extremely successful as a general-purpose classification and regression method (Breiman, 2001). RF is able to handle mixed categorical and numerical features, multiple classes, is insensitive to the scale of features, and has been considered as a powerful supervised learner. $h_m(x)$ m. decision and M total number of trees

Mathematically,

$$f(x) = \frac{1}{M}\sum_{m=1}^{M} h_m(x) \tag{2}$$

Since each tree in a random forest is trained independently, the error rate is determined by both the accuracy of individual trees and the correlation between trees. Random feature selection reduces correlation and improves model performance.

4.3. Gradient Boosting

Gradient Boosting is an algorithm developed by Friedman (2001) in which each tree is created sequentially, attempting to correct the errors of the previous tree. This method increases the power of the model by focusing on the errors of weak learners. In each iteration, optimization is performed in the direction of the negative gradient to minimize error in line with the trends of the current model. γ_m the learning rate of the m-th tree and $h_m(x)$ the output of the m-th tree.

The mathematical basis of the gradient boosting method is as follows:

$$f_m(x) = f_{m-1}(x) + \gamma_m h_m(x) \tag{3}$$

Gradient boosting adds new trees modeled according to the negative gradient of the loss function and optimizes the model overall. Mathematically, the optimal γm\gamma_mγm value is determined by the update made in the direction of the gradient at each iteration.

4.4. Multilayer Perceptron (MLP)

MLP was developed by Rosenblatt (1958) as a multi-layer version of the perceptron model. Used as a baseline model in financial forecasting, MLP stands out for its ability to model non-linear relationships (Heaton et al., 2016) and can process time series data with its structure consisting of input, hidden, and output layers. Figure 3 shows the basic architecture of a multilayer artificial neural network (MLP).

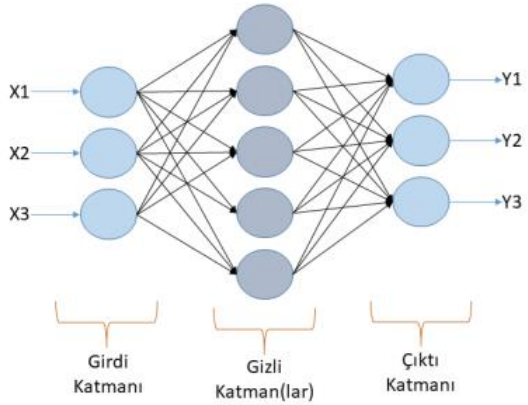

Figure 3. Multilayer Perceptron (MLP)

Haykin (2009) stated that artificial neural networks consist of three basic layers: input layer (X1, X2, X3), hidden layer(s), and output layer (Y1, Y2, Y3) (Haykin, 2009). Rumelhart et al. (1986) demonstrated that hidden layers enhance the network's capacity to learn nonlinear relationships. In this structure, the connections between neurons represent weights, and each layer is fully connected (LeCun et al., 2015). In financial time series, the input layer usually represents historical price and volume data, and the output layer represents the values to be predicted (Heaton et al., 2016).

The flowchart of the methodology of the study is presented in Figure 4.

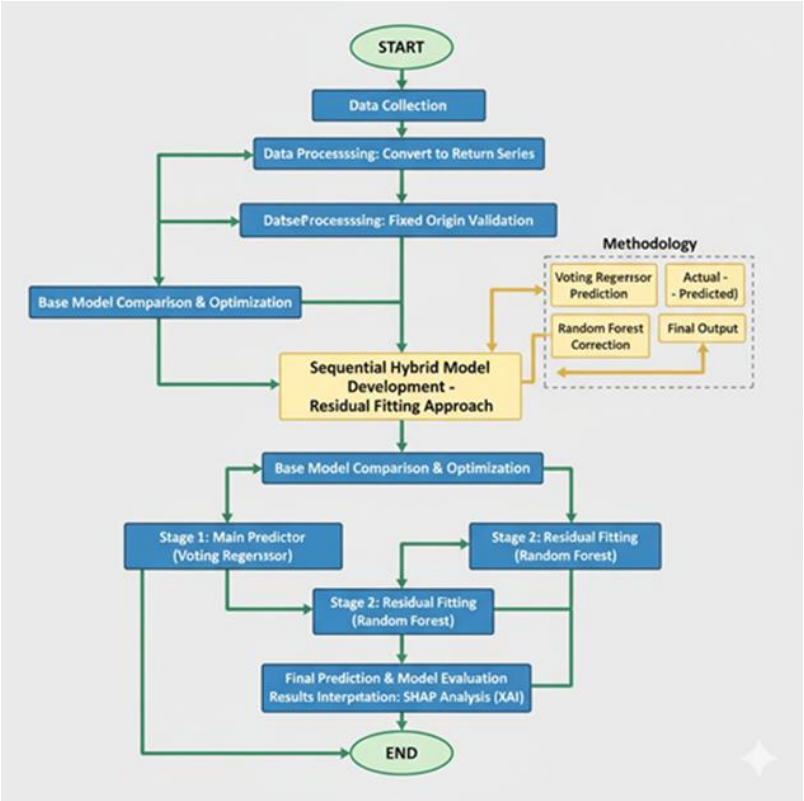

Figure 4. The Flowchart of the Methodology

5. Findings

5.1. Machine Learning Findings

The model results will be tested based on prediction accuracy on test data, the models' generalization capacity, and efficiency. Table 2 presents a comparison of the training and test performance metrics for different machine learning models.

Table 2. Comparison of Training and Testing Performance Metrics of Different Machine Learning Models

Model	Train MAE	Test MAE	Train RMSE	Test RMSE	Train R^2	Test R^2	Time (t)
Voting Regressor	0.6425	0.7522	0.8864	**0.9835**	0.6964	**0.6870**	54.647
Random Forest	0.3941	0.7925	0.5461	10.286	0.8848	0.6577	0.7128
Gradient Boosting	0.8840	0.8048	11.787	10.337	0.4632	0.6542	12.494
MLP Regressor	0.9454	0.8319	12.764	10.597	0.3705	0.6367	54.366
CatBoost	0.5282	0.8211	0.6992	10.870	0.8111	0.6177	37.494
Extra Trees	0.0000	0.8469	0.0000	10.990	10.000	0.6092	0.2539
SVR	10.033	0.7429	14.358	11.012	0.2035	0.6076	0.2807
HistGradientBoosting	0.5687	0.8512	0.7791	11.787	0.7655	0.5505	19.934
LightGBM	0.5656	0.8586	0.7762	11.866	0.7672	0.5444	0.9350
XGBoost	0.1747	0.9163	0.2473	12.367	0.9764	0.5051	0.4320
Stacking Regressor	10.742	0.9985	14.950	13.360	0.1365	0.4225	118.219

The table shows that the Voting Regressor model performed best with an RMSE value of 0.9835. This model was successful because combining the predictions of models based on different algorithms yielded a more stable result. The Extra Trees model achieved a perfect result with zero error in the training data. This indicates that the model has completely memorized the training data. Similarly, the XGBoost and Random Forest models show a significant difference between very high training performance and low test performance. This indicates that the models are prone to overfitting. The analysis also reveals that model complexity does not always result in improved performance. The Stacking Regressor, despite being the most complex and having the longest training time, showed the worst performance.

The results of the hyperparameter optimization process will be presented in two stages: the structural configurations of the models (Table 3) and the effects of these configurations on performance (Table 4).

Table 3. Best Hyperparameter Values for the Optimized Models

Model	Best Parameters
Voting Regressor (Opt)	{'ridge__alpha': 10.0, 'rf__n_estimators': 100, 'rf__max_depth': 5}
MLP Regressor (Opt)	{'learning_rate_init': 0.001, 'hidden_layer_sizes': (50,), 'alpha': 0.01, 'activation': 'tanh'}
Gradient Boosting (Opt)	{'subsample': 0.7, 'n_estimators': 100, 'max_depth': 3, 'learning_rate': 0.01}
Random Forest (Opt)	{'n_estimators': 100, 'min_samples_leaf': 2, 'max_features': 'sqrt', 'max_depth': 5}

Table 4. Final Model Performance Metrics after Hyperparameter Optimization

Model	Train RMSE	Test RMSE	Train R^2	Test R^2	Time (s)
Voting Regressor (Opt)	13.434	**0.9337**	**0.8027**	**0.7179**	25.43
MLP Regressor (Opt)	13.993	11.101	0.6435	0.6013	4.60
Gradient Boosting (Opt)	13.996	12.961	0.4832	0.4565	21.44
Random Forest (Opt)	13.707	13.951	0.4541	0.3703	15.47

Table 3 shows that the search algorithm preferred less complex structures for all models. The selected parameters focused on the more general aspects of the models and the fundamental signal in the data. The performance implications of these structural changes are detailed in Table 4, and the findings are best understood in terms of the bias-variance tradeoff. The most obvious success of the optimization is that it effectively eliminates overlearning by reducing the model variance. In all models, the difference between the Training R^2 and Test R^2 values is almost completely closed compared to before optimization, proving that the models no longer memorize the training data and can consistently generalize the learned knowledge to the test data. However, reducing model complexity to reduce variance has the potential to increase the model's bias, i.e., the error due to the tendency to simplify the underlying structure of the data.

These results have led to different performances among the models. Random Forest and Gradient Boosting models have been oversimplified to prevent overfitting. This causes a loss in their ability to capture meaningful relationships in the data and a decline in test performance. The Voting Regressor has been found to be the model that best achieves this balance. The highest generalization performance was achieved on the test data with an RMSE of 0.9337 and an R^2 of 0.7179.

5.2. Hybrid Modeling Strategy: Sequential Residual Fitting

In the hybrid model, two or more models compensate for each other's weaknesses in sequence. The purpose of the secondary model is to predict errors in the primary model (Aslanargun et al., 2007). The stages of the model used in the study are given as follows:

1. Primary Predictor: The Voting Regressor model was selected as the optimized primary predictor.
2. Residual Corrector: The residual values of the Voting Regressor on the training data are calculated.
3. Final Hybrid Forecast: The final forecast for the test data (Ŷhybrid) is obtained by summing the test forecast of the main model and the error forecast of the error-correcting model for the test data:

$$\hat{Y}_{hybrid} = \hat{Y}_{base}\,(\hat{Y}_{test}) + \hat{Y}_{residual}\,(X_{test}) \quad (4)$$

This approach aims to capture both linear and non-linear patterns more effectively, creating interaction that single models alone cannot achieve. All possible pairwise combinations are tested to determine which model pair produces the strongest empirical interaction. Table 5 shows the role played by the models in the hybrid model architecture.

Table 5. Role of Models in Hybrid Model Architecture

Hybrid Configuration (Main Model -> Error Corrector)	Test RMSE	Test R^2	Test MAE
Voting Regressor -> Random Forest	0.9264	0.7223	0.7299
Voting Regressor -> Gradient Boosting	0.9329	0.7184	0.7305
Gradient Boosting -> Voting Regressor	0.9594	0.7022	0.7623
Voting Regressor -> MLP Regressor	0.9712	0.6948	0.7670
Random Forest -> Voting Regressor	0.9999	0.6765	0.7886
MLP Regressor -> Voting Regressor	10.096	0.6702	0.7459
Gradient Boosting -> MLP Regressor	10.331	0.6547	0.8022
MLP Regressor -> Random Forest	10.384	0.6511	0.7765
MLP Regressor -> Gradient Boosting	10.497	0.6435	0.7815
Random Forest -> MLP Regressor	10.694	0.6300	0.8223
Gradient Boosting -> Random Forest	11.179	0.5956	0.8918
Random Forest -> Gradient Boosting	11.695	0.5574	0.9200

The findings presented in Table 5 reveal the critical impact of the role played by the models in the hybrid model architecture and their interaction on the final performance. According to the results of the analysis, the two-stage hybrid model with Voting Regressor as the main estimator and Random Forest as the error corrector performed the best with a Test RMSE of 0.9264 and a Test R^2 of 0.7223. This result is even better than the performance of the best single optimized model, Voting Regressor (Test RMSE $\approx$ 0.9337), proving the success of the hybridization strategy.

Since Voting Regressor combines models of different natures (linear and tree-based), it is very good at capturing the main trend and more stable patterns in the data. The Random Forest model, which comes in at the second stage and is skilled at capturing flexible, non-linear relationships, improved the overall prediction by effectively modeling these complex and unsystematic residual values (errors) that Voting Regressor misses. A similar interaction is observed in the second-best combination, Voting Regressor -> Gradient Boosting.

Another important finding in the analysis is the tendency for configurations where the Voting Regressor model is used as the main estimator to be at the top of the table. This shows how important it is for the success of the hybrid model that the forecast made in the first stage is stable and has low variance. When more flexible models such as Random Forest or Gradient Boosting are the main estimators, the residuals (errors) they produce are noisier and chaotic, making it more difficult for the second model to model these errors and leading to lower performance.

Finally, more complex hybrid architectures, such as Stacking, were also tried for this pair, which produced the most successful interaction in this study. However, it was observed that these advanced ensemble learning methods do not yield better results than the more intuitive and interpretable Residual Fitting approach. This suggests that, due to the nature of the problem, direct and sequential correction of each other's errors is more effective than an indirect learning process through a more complex meta-model. In the light of these findings, the Voting Regressor -> Random Forest hybrid model was identified as the final model with the highest performance developed in this research.

5.3. Model Interpretability: SHAP Approach

SHAP is a model-agnostic explanation method that draws its theoretical foundations from the Nobel Prize-winning concept of cooperative game theory and Shapley values (Shapley, 1953). SHAP is one of the explainable artificial intelligence approaches and is used as a powerful machine learning interpretation technique that can measure the absolute impact level of each feature on the predicted outcome and also the direction of this impact.).

SHAP creates an explanation model that expresses the prediction of any machine learning model as a simple sum of the values attributed to each feature:

$$f(x) = \emptyset_0 + \sum_{i=1}^{M} \emptyset_i \quad (5)$$

where *f(x)* is the model's final prediction for input x; M is the number of features in the model; φ_0 is the base value, which is the average prediction over the entire data set; and φ_i is the SHAP value, which indicates the impact of the i-th feature on that prediction. A positive value of φ_i indicates that the feature pushes the prediction up from the base value, while a negative value pushes it down.

In this study, we use two basic visualization tools from SHAP to reveal the insights of the best hybrid model (Voting Regressor -> Random Forest):

SHAP Summary Plot: This plot summarizes the impact of each feature on the entire dataset. Each point represents a single prediction for a single feature. The position of the dots on the horizontal axis indicates the SHAP value (impact on the prediction), and the color indicates the value of the feature itself (high or low). In this way, it is globally understandable which features are most important and how the values of these features affect (positively/negatively) the prediction outcome.

SHAP Dependence Plot: This plot shows how the impact of a single feature on the model's output (SHAP value) changes as the value of that feature changes. Furthermore, the color of the dots reflects the value of a second feature that has the strongest interaction with the selected feature, revealing potential interactions between features.

Through these methods, we analyzed not only what the most successful hybrid model predicts, but also which financial indicators influence these predictions, in what direction and to what extent.

6. SHAP Analysis Findings

In this section, the forecasting mechanism of the highest performing Voting Regressor -> Random Forest hybrid model is analyzed using the SHAP (SHapley Additive exPlanations) method. Figure 5 visualizes how important the model assigns to which financial variables and how the values of these variables affect the model's predictions. SHAP Analysis demonstrates that the model goes beyond being merely a black box, linking the nonlinear pattern recognition capabilities provided by machine learning to observable behaviors and market dynamics in financial markets, thereby possessing a meaningful decision-making mechanism. In particular, the relationships captured by the model enrich discussions regarding market inefficiencies, risk-averse behaviors, and the effects of macroeconomic shocks on sustainability indices.

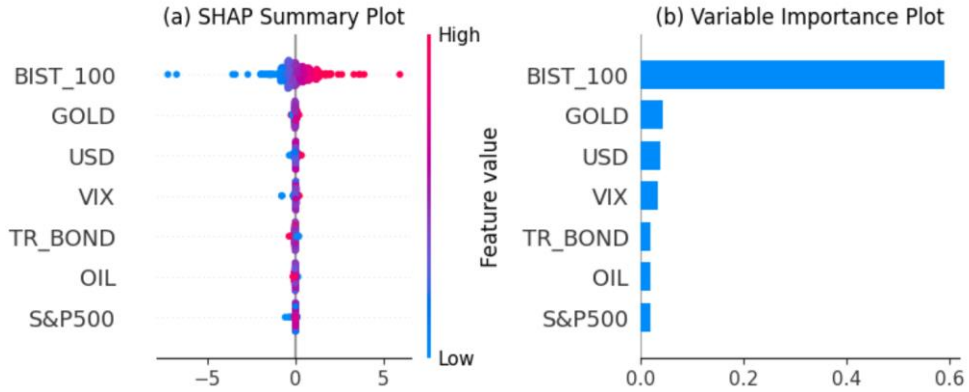

Figure 5. SHAP Summary and Variable Importance Plots for the Best Hybrid Model

The Variable Importance Plot presented in Figure 5(b) shows the average absolute effect that the model attributes to each attribute when making predictions**.** According to this graph, the return of the BIST_100 index stands out as by far the most critical factor among all other variables. This finding is in line with the basic expectation that the performance of the BIST Sustainability Index is strongly influenced by the main index, which reflects the overall market trend. Following BIST_100, variables such as GOLD (Gold), USD (US Dollar), VIX (Volatility Index), and TR_BOND (Turkish 10-Year Bond) constitute a second level of importance. These variables are macroeconomic and financial indicators that generally reflect the perception of risk, uncertainty, and the search for safe havens.

Figure 5(a) explains the dynamics behind this ranking in more detail. This graph shows the effect of the variable value on the model's output. The fundamental relationships analyzed are as follows:

BIST_100: For this variable, high positive returns are associated with positive SHAP values. This indicates that the model has successfully learned a strong and positive correlation between the two indices.

GOLD and USD: The high values of gold and dollar returns are seen to be associated with negative SHAP values. This situation indicates that investors are avoiding risk by exiting stock markets during periods of uncertainty and turning to safe havens such as gold and foreign exchange.

VIX**:** The higher values of the fear index negatively affect the model's predictions. This situation is consistent with financial theories that increased market uncertainty fears put pressure on stock returns.

The SHAP analysis shows that the hybrid model fits statistically. This indicates that it has learned relationships that can be interpreted in terms of the fundamental dynamics of financial markets and economic intuition.

6.1. Analysis of Variable Interactions: SHAP Dependency Graphs

SHAP dependency plots show the effect of variables on the model's predictions. These plots are used to show how these effects interact with other variables. Figure 6 shows these interactions for the four most important variables (BIST_100, GOLD, USD, VIX). In these plots, the horizontal axis shows the value of the feature and the vertical axis shows the influence of that feature on the prediction (SHAP value). The color of the dots represents the value of a second variable, automatically determined by the SHAP library, which has the strongest interaction with the main variable.

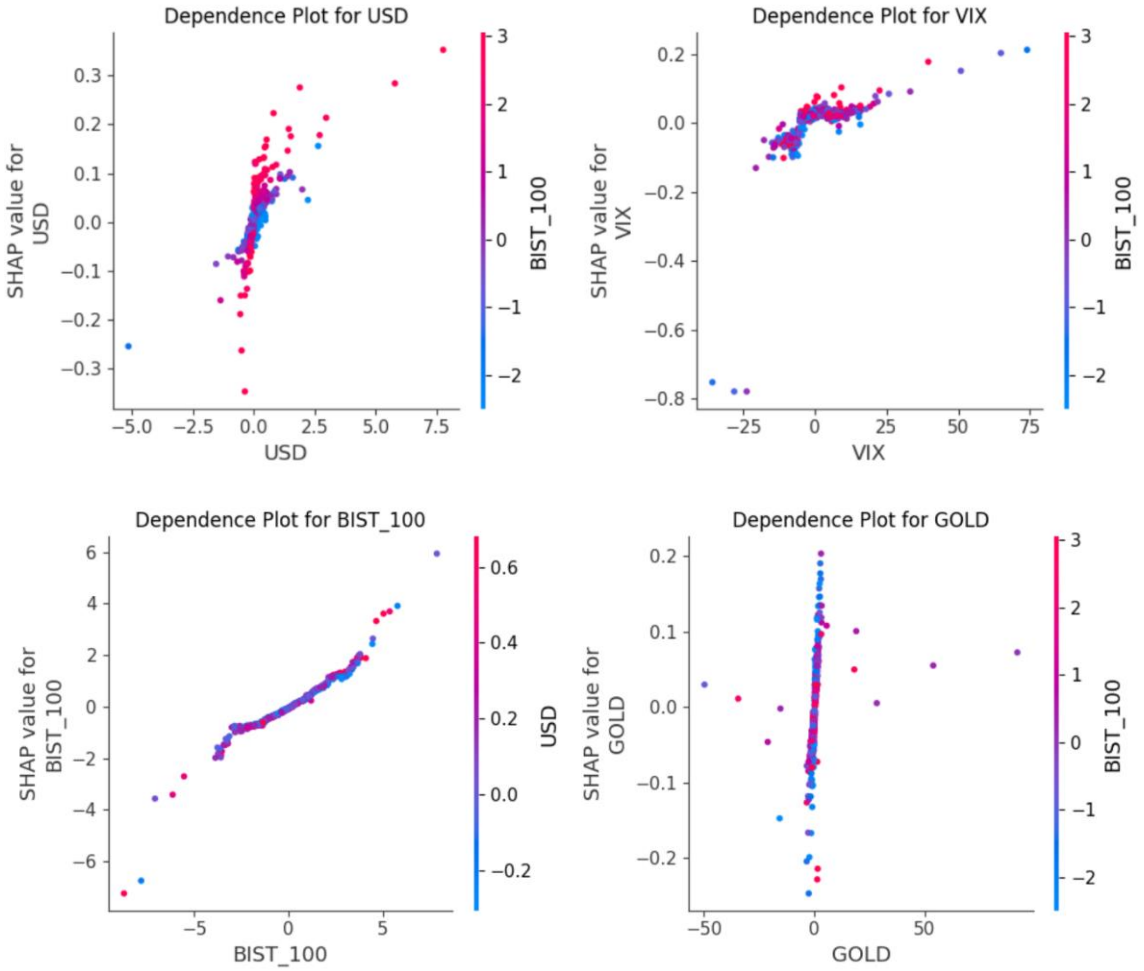

Figure 6. SHAP Dependency and Interaction Plots for the Four Most Important Variables

The graphs in Figure 6 provide important evidence on how successfully the hybrid model learns non-linear and context-sensitive relationships. The BIST_100 graph shows a strong and almost linear positive relationship between the index return and the SHAP value, as expected. According to the SHAP analysis, the factor with the strongest interaction with this variable is USD (Dollar) returns. The color distribution in the chart implies that the positive impact of BIST_100 is more pronounced on days when the USD exchange rate is falling or flat. This suggests that the model has learned that periods when the Turkish Lira appreciates are a more positive signal for the overall market.

The GOLD graph reveals a more complex and noisy relationship structure. Gold returns are concentrated around zero, and SHAP values are also close to zero in this region. However, extreme positive or negative gold returns (dots on the edges of the graph) generally have a negative impact on the model's predictions. This suggests that large price movements in gold are perceived as a signal of uncertainty in the market. According to the color axis, this effect is more pronounced on days when BIST_100 is negative (blue dots), suggesting that the model reinforces the negative relationship between these two variables in risk-off scenarios.

The USD (Dollar) and VIX (Volatility Index) charts exhibit interactions that are highly consistent with financial intuition. Both charts show a positive trend where the SHAP value increases as the value of the variable increases. However, this does not mean that the variables themselves have a positive effect on the Sustainability Index; on the contrary, it should be noted that the SHAP values of these variables are generally below zero.

7. Conclusion

In this study, mean absolute error (MAE), coefficient of determination (R^2), mean square error (MSE), and root mean square error (RMSE) metrics were used to measure the performance of machine learning models. In this study, MAE, R^2, MSE, and RMSE metrics were used to measure the performance of machine learning models, and the SHAP approach was used to evaluate the importance of explanatory variables.

Model performance was evaluated in three stages. First, an initial screening of eleven different models selected the four models that yielded the best results (Voting Regressor, Random Forest, Gradient Boosting, MLP Regressor). In the second stage, hyperparameter optimization on these four models effectively controlled overfitting. In the final stage, a systematic evaluation of hybrid models based on the Residual Fitting technique was performed. The results show that the Voting Regressor -> Random Forest hybrid model has the best performance with the lowest MAE and RMSE and the highest R^2 value.

The empirical findings reveal that the BIST100 index is at the center of market dynamics. The variable importance ranking shows that BIST100 has a much more substantial impact than all other factors. SHAP analysis results support this finding and show that positive returns in the BIST100 have a positive effect on predictions, while negative returns have a negative effect. The variables with the most significant impact after the BIST100 have been gold, the US dollar, and the volatility index. Vardari et al. (2020) found that the BIST Sustainability Index provided returns to the BIST 100 Index. Kaur and Chaudhary (2022) demonstrated a long-term relationship between the sustainability index and macroeconomic variables. Morales et al. (2019) and Shaikh (2022) showed that increases in the VIX index negatively affect various sustainable investment indices. Özçim (2022) revealed that increases in exchange rates increase the volatility of the BIST Sustainability Index. Therefore, the model's learning that increases in risk indicators such as gold price, exchange rate, and VIX have a negative effect on the sustainability index is similar to the risk-averse behavior observed in the literature. The analysis results of the study show that market risk factors have a strong and guiding effect on sustainability indices. It has been concluded that the hybrid machine learning approach can successfully model these complex relationships.

Policy makers play a critical role in making financial and macroeconomic markets more resilient to fluctuations. To this end, concrete incentive mechanisms should be developed to

increase the corporate resilience of sustainability-focused companies against exchange rate and interest rate shocks. Among these incentives, priority should be given to directly applicable policies such as tax breaks for developing the green bond market, easier access to financial instruments for companies to manage foreign exchange risk, or subsidized loans.

In future research, the model can be tested with different algorithms (e.g., LSTM, XGBoost, CatBoost) and expanded data sets to increase the robustness of the findings. Integration of micro-level ESG scores, company reporting, and news/sensitivity data would strengthen the explanatory power of the model.

Declaration of Research and Publication Ethics
This study which does not require ethics committee approval and/or legal/specific permission complies with the research and publication ethics.

Researcher's Contribution Rate Statement
I am a single author of this paper. My contribution is 100%.

Declaration of Researcher's Conflict of Interest
There are no potential conflicts of interest in this study.

References

AlGhazali, A., Mensi, W., Morley, B. and Kang, S.H. (2025). Connectedness and hedging strategies between European sustainability and conventional stock markets. *Journal of Sustainable Finance & Investment*, Advance online publication. https://doi.org/10.1080/20430795.2025.2520523

Aslanargun, A., Mammadov, M., Yazici, B. and Yolacan, S. (2007). Comparison of ARIMA, neural networks and hybrid models in time series: Tourist arrival forecasting. *Journal of Statistical Computation and Simulation*, 77(1), 29-53. https://doi.org/10.1080/10629360600564874

Başkaya, H. (2025). BİST sürdürülebilirlik endeksi ile diğer finansal endeksler arasındaki ilişkinin ve nedenselliğin analizi. *Fiscaoeconomia*, 9(2), 1003-1021. https://doi.org/10.25295/fsecon.1541942

Bergmeir, C. and Benítez, J.M. (2012). On the use of cross-validation for time series prediction. *Information Sciences*, 191, 192-213. https://doi.org/10.1016/j.ins.2011.12.028

Bhattacharya, A. (2022). *Applied machine learning explainability techniques*. Birmingham: Packt Publishing.

Bhutta, U.S., Tariq, A., Farrukh, M., Raza, A. and Iqbal, M.K. (2022). Green bonds for sustainable development: Review of literature on development and impact of green bonds. *Technological Forecasting and Social Change*, 175, 121378. https://doi.org/10.1016/j.techfore.2021.121378

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Broadstock, D.C. and Cheng, L.T. (2019). Time-varying relation between black and green bond price benchmarks: Macroeconomic determinants for the first decade. *Finance Research Letters*, 29, 17-22. https://doi.org/10.1016/j.frl.2019.02.006

Chen, S., Song, Y. and Gao, P. (2023). Environmental, social, and governance (ESG) performance and financial outcomes: Analyzing the impact of ESG on financial performance. *Journal of Environmental Management*, 345, 118829. https://doi.org/10.1016/j.jenvman.2023.118829

Çankal, A. and Ever, D. (2025). The effects of renewable energy consumption on financial performance: An explainable artificial intelligence (XAI)-based research on the BIST sustainability index. *International Journal of Energy Economics and Policy*, 15(4), 204-213. https://doi.org/10.32479/ijeep.19602

Dietterich, T.G. (2000). Ensemble methods in machine learning. In J. Kittler and F. Roli (Eds.), *Multiple classifier systems, first international workshop, MCS* 2000 (pp. 1-15). https://doi.org/10.1007/3-540-45014-9_1

Drimbetas, E., Sariannidis, N., Giannarakis, G. and Litinas, N. (2010). The effects of macroeconomic factor on the sustainability, large-cap and mid-cap Dow Jones indexes. *International Journal of Business Policy and Economics*, 3, 21-36. Retrieved from https://serialsjournals.com/

Ehlers, T. and Packer, F. (2017). Green bond finance and certification. *BIS Quarterly Review*, September, 89-104. Retrieved from https://www.bis.org/

Friede, G., Busch, T. and Bassen, A. (2015). ESG and financial performance: Aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance & Investment*, 5(4), 210-233. https://doi.org/10.1080/20430795.2015.1118917

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. Retrieved from https://www.jstor.org/

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Berlin: Springer.

Haykin, S. (2009). *Neural networks and learning machines, 3/E*. Bangalore: Pearson Education

Heaton, J.B., Polson, N.G. and Witte, J.H. (2016). Deep learning in finance. *arXiv preprint arXiv:1602.06561*. https://doi.org/10.48550/arXiv.1602.06561

Kaur, J. and Chaudhary, R. (2022). Relationship between macroeconomic variables and sustainable stock market index: An empirical analysis. *Journal of Sustainable Finance & Investment*, 1-18. https://doi.org/10.1080/20430795.2022.2073957

Kavas, Ü.Y.B. (2025). Finansal enstrümanların BİST sürdürülebilirlik endeksi üzerindeki dinamik etkilerinin TVP-VAR modeliyle araştırılması. *Mali Çözüm Dergisi*, 35, 1201-1225. Retrieved from https://ismmmo.org.tr/Yayinlar/Mali-Cozum-Dergisi--1

Kaya, M. (2023). BİST sürdürülebilirlik endeksi ile fosil yakıt fiyatları arasındaki ilişkinin analizi. *Abant Sosyal Bilimler Dergisi*, 23(3), 1475-1495. https://doi.org/10.11616/asbi.1327883

Kocamiş, T.U. and Yildirim, G. (2016). Sustainability reporting in Turkey: Analysis of companies in the BIST Sustainability Index. *European Journal of Economics and Business Studies*, 2(3), 41-51. Retrieved from https://revistia.com/ejes

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. https://doi.org/10.1038/nature14539

Lundberg, S.M. and Lee, S.I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 1-10). California: Curran Associates.

Morales, L., Soler-Domínguez, A. and Hanly, J. (2019). The power of ethical investment in the context of political uncertainty. *Journal of Applied Economics*, 22(1), 554-580. https://doi.org/10.1080/15140326.2019.1683264

Özçim, H. (2022). Bist sürdürülebilirlik endeksi ve makroekonomik veriler arasindaki ilişkinin GARCH modelleri çerçevesinde incelenmesi. *Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 50, 115-126. https://doi.org/10.30794/pausbed.1015216

Quinn, B. (2023*). Explaining AI in finance: Past, present, prospects*. New York: Cornell University.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386. https://doi.org/10.1037/h0042519

Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533-536. Retrieved from https://www.nature.com/

Seow, R.Y.C. (2025). Transforming ESG analytics with machine learning: A systematic literature review using TCCM framework. *Corporate Social Responsibility and Environmental Management*, Advance online publication. https://doi.org/10.1002/csr.70089

Shaikh, I. (2022). On the relationship between policy uncertainty and sustainable investing. *Journal of Modelling in Management*, 17(4), 1504-1523. https://doi.org/10.1108/JM2-12-2020-0320

Shapley, L.S. (1953). A value for n-person games. In H.W. Kuhn and A.W. Tucker (Eds.), *Contributions to the theory of games II* (307–317). New Jersey: Princeton University Press.

Sharma, P., Shrivastava, A.K., Rohatgi, S. and Mishra, B.B. (2023). Impact of macroeconomic variables on sustainability indices using ARDL model. *Journal of Sustainable Finance & Investment*, 13(1), 572-588. https://doi.org/10.1080/20430795.2021.1972679

Siddique, M.A. and Karim, S. (2025). Can ESG disclosure predict carbon risk? Evidence from machine and deep learning models. *Finance Research Letters*, 83, 107672. https://doi.org/10.1016/j.frl.2025.107672

Şahin, S. (2024). Finans Sektöründe yapay zeka, makine öğrenmesi ve büyük veri kullanımı: Fırsatlar, zorluklar ve politika yapıcılar için çıkarımlar. *Finans Ekonomi ve Sosyal Araştırmalar Dergisi*, 9(4), 364-381. https://doi.org/10.29106/fesa.1542860

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tsay, R.S. (2010). *Analysis of financial time series* (3rd ed.). USA: Wiley.

Vardari, L., Gashi, R. and Aahmeti, H.G. (2020). The impact of corporate sustainability index on BIST Sustainability Index. *European Journal of Sustainable Development*, 9(2), 375-375. https://doi.org/10.14207/ejsd.2020.v9n2p375

Wolpert, D.H. (1992). Stacked generalization. *Neural Networks*, 5(2), 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1

Zhang, J. and Zhao, Z. (2026). Corporate ESG rating prediction based on XGBoost-SHAP interpretable machine learning model. *Expert Systems with Applications*, 295, 128809. https://doi.org/10.1016/j.eswa.2025.128809