Çukurova Üniversitesi Mühendislik Fakültesi Dergisi, 39(2), ss. 437-452, Haziran 2024
Cukurova University Journal of the Faculty of Engineering, 39(2), pp. 437-452, June 2024

Using Bigdata for Choosing the Right Forecasting Method, Dataset and Period in a Time Series Analysis

Serap AKCAN[*1] ORCID *0000-0003-2621-9142*
Murat AKCIL[2] ORCID *0000-0003-4963-1826*
Metin ÖZŞAHİN[3] ORCID *0000-0001-9989-526X*

[1]Tarsus University, Faculty of Engineering, Department of Industrial Engineering, Mersin, Türkiye
[2]Süleyman Demirel University, Faculty of Engineering, Department of Industrial Engineering, Isparta, Türkiye
[3]Osmaniye Korkut Ata University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, Osmaniye, Türkiye

Geliş tarihi: 05.01.2024 Kabul tarihi: 27.06.2024

Atıf şekli/ How to cite: AKCAN, S., AKCIL, M., ÖZŞAHİN, M., (2024). Using Bigdata for Choosing the Right Forecasting Method, Dataset and Period in a Time Series Analysis. Cukurova University, Journal of the Faculty of Engineering, 39(2), 437-452.

Abstract

Nowadays especially production companies gathering a huge data due to their daily transactions on the own systems. Production companies should handle this raw data as handling the raw materials too. Today, scientific studies carried out for this purpose are gathered under the title of BigData. The BigData creates many helps to companies' competitive advantages according to their competitors. For this view, the purpose of this study was to determine the best demand forecasts method and forecasting period by using BigData at forest production industry. Using the time series analysis module of the WEKA program, the algorithm and data set providing the most accurate estimate for each of the selected decor papers were determined. As a result, it is thought that this study will provide a route map for about choosing right data period and forecasting method for the forest products.

Keywords: Big data, Data mining, Time series analysis, Demand forecasting, Forest products sector

Zaman Serisi Analizinde Doğru Tahmin Yöntemini, Veri Kümesini ve Dönemi Seçmek İçin Büyük Veriyi Kullanma

Öz

Günümüzde özellikle üretim firmaları kendi sistemleri üzerinde yaptıkları günlük işlemlerden dolayı büyük miktarda veri toplamaktadır. Üretim şirketleri, ham maddeyi ele aldığı gibi bu ham veriyi de ele almalıdır. Günümüzde bu amaçla yapılan bilimsel çalışmalar büyük veri başlığı altında toplanmaktadır. Büyük veri, şirketlerin rakiplerine göre rekabet avantajı sağlamasına birçok katkı sağlamaktadır. Bu doğrultuda bu çalışmanın amacı, orman ürünleri sektöründe büyük veri kullanarak en iyi talep tahmin yöntemini ve tahmin dönemini belirlemektir. Çalışmada, WEKA programının zaman serisi analiz modülü kullanılarak seçilen dekor kağıtlarının her biri için en doğru tahmini sağlayan algoritma ve veri seti belirlenmiştir. Sonuç olarak

[*]Sorumlu yazar (Corresponding Author): Serap AKCAN, *serapakcan@tarsus.edu.tr*

bu çalışmanın orman ürünlerine ilişkin doğru veri periyodu seçimi ve tahmin yöntemi konusunda bir yol haritası oluşturacağı düşünülmektedir.

Anahtar Kelimeler: Büyük veri, Veri madenciliği, Zaman serisi analizi, Talep tahmini, Orman ürünleri sektörü

1. INTRODUCTION

Thanks to technological developments and Industry 4.0, competition among the companies in the production sector has been increasing. Due to this increased competition, companies wanting to maintain their market share produce products of higher quality and lower cost compared to their competitors, and try to shorten delivery times. In this age of online shopping, it is also imperative to stand out in an environment where customers can search for products instantly and find suppliers instantly. Manufacturers want to reduce their product costs as much as possible in order to increase their profitability and make a difference without sacrificing quality, but they have to give their customers timely deadlines. Although it varies by sector, raw material costs are at the top of the product costs in many sectors. Thus, in order to reduce product costs, first the purchase costs of raw materials must be reduced. There are various solutions that manufacturers can use to reduce raw material costs and these solutions can be grouped under three distinct categories. The first of these is improvements in the production process and R&D studies. Manufacturers can examine solutions such as R&D studies, substitute raw material trials, and reduction of waste rates aiming to reduce the raw material usage rates in a way that does not change the quality characteristics of the products. The second category is related to suppliers. Solutions such as increasing the number of companies supplying raw materials, establishing good communication with suppliers, and avoiding purchases with foreign currency payments since the exchange rate is unbalanced in developing countries can be listed under this heading. Thirdly, when determining purchasing strategies, an accurate demand forecast can be made and the determination of raw material order sizes, order frequency, and cost can be presented as a solution. Preparing an accurate budget plan and making the right amount of raw material connections at the right time are among the strategies that can be done at the purchasing stage to reduce raw material costs [1].

In today's world where the competition is so high, it is not only sufficient to lower the prices in order to retain customers, but it is also of great importance to deliver the desired product to the customer on time. Purchasing strategy and demand forecasting are of great importance for companies in order to give accurate deadlines and to comply with the given deadline. The vast majority of companies today benefit from big data. In order for companies to gain competitive advantage, big data needs to be analyzed in a way that will reduce costs and increase customer satisfaction. Machine learning techniques are frequently used in big data analysis.

2. LITERATURE REVIEW

When we evaluate the studies in the literature; It is useful to group and interpret the demand forecasting studies made about i) production and materials, ii) fashion products, and iii) non-fashionable products.

2.1. Studies Related Demand Forecasting of Production

Much research has been carried out on demand forecasting in recent years. Kaes and Azeem (2009) [2] conducted a demand forecasting and supplier selection study at a knitted composite factory which produces fabrics for export. The most suitable model for the selected raw material was scanned by applying different demand forecasting techniques. By examining the results of an analytic hierarchy process (AHP) for demand forecasting and supplier selection, suggestions were made to improve the level of material management and increase profits by reducing waste. Kim et. al. [3] investigated why mass customization is needed in Smart Manufacturing and looked for appropriate demand prediction techniques by comparing the traditional linear analysis method ARIMA time series analysis

with the nonlinear analysis method LSTM neural network model. Arif et. al. [4] examined product demand forecasting in production facilities using machine learning methods. They used KNN, Random Forest, FNN, ANN, and the Holt-Winters model algorithms.

In their study, Gupta and Sihag [5] used the Gaussian process, M5P model, random forest and random tree techniques to predict which materials should be used in which proportions of concrete mixes with the highest concrete strength. As a result of the comparisons, it is seen that the results obtained with the Gaussian process technique are better. Panarese et. al. [6] developed a machine learning-based platform for sales forecasting using a gradient boosting approach. In this study, it is presented that XGBoost regression model is more accurate in predicting future sales in terms of various error metrics, such as MSE, MAE, MAPE and WAPE. Nasseri et. al. [7] applied machine learning in retail demand prediction. In this study, they used over six years of historical demand data from a retail entity. The dataset included daily demand metrics for more than 330 products with 5.2 million records. It is presented in this study that spanning three perishable product categories, reveals that the ETR model outperforms LSTM in metrics including MAPE, MAE, RMSE, and R^2.

2.2. Studies Related Demand Forecasting of Fashionable Products

Aksoy et. al. [8] developed a decision support system for demand forecasting in the clothing industry. Yunishafira [9] conducted a demand forecasting study by using the historical sales data of a store in the clothing industry that buys and sells ready-made products. In the study, time series methods including moving average, simple exponential smoothing, and the holt-winters model were used. While the simple moving average made the most accurate estimates, the results of the study indicated that interpreting the results would better help company managers determine both supply chain and operations management. Ren et. al. [10] conducted an extensive literature review on demand forecasting methods for trendy products and examined how the fashion retailer's future demand forecasting and inventory planning problem was handled in practice through a real-life case study.

2.3. Studies Related Demand Forecasting of Non-fashionable Products

Yadav and Ghosh [11] used MSARIMA and ARMAX forecasting models to forecast monthly demand for farm tractors in India. As a result of their research, it was seen that the ARMAX model made better predictions than the MSARIMA model. Their belief was that accurate monthly forecasts of farm tractors would help manufacturers better manage raw materials, inventory and supply chains.

Huber and Stuckenschmidt [12] presented a daily retail demand forecasting using machine learning methods. Spiliotis et. al. [13] compared statistical and machine learning methods for daily SKU demand forecasting. Panigrahi and Behera [14] focused on a model that can determine the near-optimal structure for artificial neural networks used in time series forecasting in their study using a large number of experimental data sets. As a result of this study, they developed an adaptive DE-based modelling scheme (DEMS) to determine the near-optimal architecture of ANN for a time series. Moroff et. al. [15] presented a study for assessing innovative demand forecasting models. They used the Holt Winters - Triple exponential smoothing (ETS), Seasonal Auto-Regressive Integrated Moving Average Extended (SARIMAX), Extreme Gradient Boosting (XGBoost), Random Forest (RF), Long-term short-term memory (LSTM), and Multilayer Perceptron (MLP) for demand forecasting. Ngo et. al. [16] made demand prediction for the electricity consumption forecasts of buildings, which is one of the important study areas in the literature. In this study, the ANNs, SVR, and M5Rules were applied to predict future building energy consumption. As a result of the study ML models can be proposed as an effective method for forecasting energy consumption in buildings. The ML for an ensemble approach has proved predictive performance in predicting the next 24-h energy consumption.

Pham et. al. [17] used the machine learning algorithms such as random tree (RT), random forest (RF), decision stump, M5P, support vector machine (SVM), locally weighted linear regression (LWLR), and reduce error pruning tree (REP Tree) in their study. Estimation was made using datasets containing data such as groundwater level, average temperature, precipitation and relative humidity for the period 1981-2017 obtained from two wells in the northwest region of Bangladesh. Bagging-RT and Bagging-RF models gave the best results by making the most accurate estimation in the study, where the whole data set was used as the training (1981–2008) and test (2008–2017) dataset.

2.4. Evaluation of the Literature and the Contribution of the Study to the Literature

This study, which we think will make an important contribution to the literature will complete the following deficiencies according to the literature review.

i. As we examined no research related to decor paper in which demand estimations were made using time series analyses could be found.

ii. Forecasting demand using big data is a difficult problem to solve. In the present study, monthly and annual real sales bigdata from 2009-2021 for Melamine-Faced Chipboard (MFC) were obtained and analyzed from a company operating in the forest products sector.

iii. The topic of demand forecasting for fashion products in the clothing industry is a trending research area. But home fashion or forest industry demand forecasting are areas that are rarely studied. Fashionable demand forecasting, especially for forest products, is a needed and important research area for manufacturing companies.

iv. As far as the studies in the literature are examined, in order to decide on the best forecasting method, the importance of choosing the right data set and the right forecast period in demand forecasting in general emerges. For this reason, there is a great need for a guide to help decision makers in this regard. In this study, the demand forecasting problem was examined to help guide the strategic decisions of the company providing the huge sales data. This BigData has divided 3 parts as all data (2009-2019), last 6 years data (2015-2021) and last 3 years data (2019-2021). For these purposes, future demand forecasts for decor papers used in the production of MFC, which changes according to the current trend, were made using the time series analysis module of the WEKA program in an effort to help determine the purchasing strategy of the business.

3. METHODOLOGY

Today, almost all companies use an enterprise resource planning (ERP) software. This type of software is an important structure used in all departments of a company from accounting and purchasing to production and planning and is a tool used to connect all units and processes [18]. Thanks to these types of software systems, records of all work done can be kept. But this unprocessed and raw data is just a chunk of data unless properly analyzed and made sense of. This is where the concepts of data mining and big data come into play. Data mining is defined as making a large number of data interpretable in line with desired targets or categorizing the desired data from big data [19]. The tools used in the development and dissemination of data analysis are important. Some of the most used data analysis tools are open source RapidMiner, WEKA, R Tool, and KNIME [20]. WEKA is a preferred analysis tool in data mining applications due to its ease of use, compatibility with every operating system, algorithms, and the fact that it is open source. In this study, the time series analysis module of WEKA program was used. The flow chart of the methodology used in the study is shown in Figure 1.

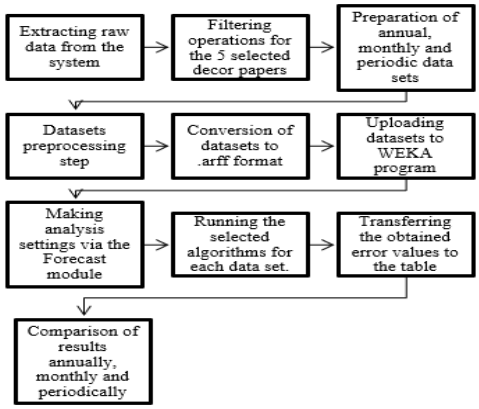

Figure 1. Flow chart of the research methodology

3.1. Time Series Analysis

Algorithms that analyze historical data within a time series enable this data to make predictions for the future by analyzing which trend the data moves on, how seasonal natural conditions are a factor, what effect long-term trends create, and whether there are exceptions in certain periods within the data.

In this study, the following WEKA time series algorithms were used: RandomSupSpace, CVParemeterSelection, MultiScheme, WIHW, InputMappedClasifier, Zeror, RepTree, DecisionStump, LWL, RondomCommitte, RandomForest, M5P, Randomtree, DecisionTable, Bagging, Regression, InputMappedClasifier, Zeror, RepTree, DecisionStump, LWL, RondomCommitte, RandomForest, M5P, Randomtree, DecisionTable, Bagging, Regression, DecisionClasifier, RegressionClasifier, SMO, Regression, and Rules. While running these algorithms, the settings in WEKA and used by default were left constant. In this section, definitions of the SMOreg, LinearRegression, M5Rules, M5P, RandomizableFilteredClasifier, MultilayerPerceptiron, RondomCommitte, RandomForest, and AdditiveRegression algorithms, which are some of the most used algorithms in the literature, are given.

The *SMOreg* (Sequential Minimal Optimization Regression) algorithm uses the support vector machines method [21].

LinearRegression is a method used to examine the numerical relationship between variables. Linear regression analysis is used to explain the relationship between the variables and to create a model that describes this relationship. Simple Linear Regression is also known as Multiple Linear Regression according to whether the variables are dependent or not [22].

The *M5Rules* algorithm divides the data from the whole into parts to create a model tree. Then, the best branches are chosen to create a rule. For the remaining samples, the algorithm continues to work through this branch [23].

The *M5P* algorithm is an improved version of the previously produced M5 algorithm. It creates a regression tree model using experimental data and then a linear regression analysis is performed on each branch within the tree model. First, the data is separated according to the determined features and the standard deviation value is used at the nodes to break up the dataset and find which attribute is the best [23].

Randomizable Filtered Classifier is actually a variant that uses RandomProjection and IBk algorithms with a filter that can bring random selection functionality to the FilteredClassifier algorithm [24].

The *MultilayerPerceptiron* algorithm is used for classification by the back propagation method and learning a multilayer perceptron as the name suggests. One of the most important features is that the created network can also be created manually [25].

The *RondomCommitte* algorithm is used to generate a collection of randomly selected classifiers [26].

The *RandomForest* algorithm is a supervised classification algorithm. As the name suggests, it randomly creates a forest. There is a direct relationship between the number of trees in the algorithm and the result it can achieve. The higher the number of trees, the higher the chance of obtaining accurate results [27].

The *AdditiveRegression* algorithm is used to obtain more accurate results in non-linear values and in cases where linear regression fails [28].

3.2. Performance Statistics

Error tests are essential for measuring the accuracy of the predictions made by the prediction models created by the estimation algorithms and to express them numerically. The main task of these tests is to numerically show the difference between the actual values of the predictions made with the help of models created by prediction algorithms. The smaller this difference, the better the prediction. When using prediction algorithms, the aim is to determine which will make the most accurate prediction. Error measurement techniques are used to determine which algorithm, that is, the model produced by which algorithm, makes the best prediction.

In prediction algorithms, the error is calculated by subtracting the estimated value from the actual value (eq.1). To examine the adequacy of the proposed models, their errors should be investigated. For this purpose, mean absolute error (MAE) (eq. 2), mean square error (MSE) (eq. 3), root mean square error (RMSE) (eq. 4), and mean absolute percentage error (*MAPE*) (eq. 5) are used as performance statistics.

$$E_t = A_t - F_t \quad (1)$$

$$MAE = (\sum_{t=1}^{n} |A_t - F_t|)/\, n \quad (2)$$

$$MSE = (\sum_{t=1}^{n} (A_t - F_t)^2)/n \quad (3)$$

$$RMSE = \sqrt{(\sum_{t=1}^{n} (A_t - F_t)^2)/n} \quad (4)$$

$$MAPE = \frac{\sum_{t=1}^{n} |\frac{A_t - F_t}{A_t}|}{n}\, 100 \quad (5)$$

where, n = number of observations, A_t = actual value at observation t, F_t = estimate value for observation t, E_t = error value at observation t and t = time period. The aim of the present study was to determine the algorithms that make predictions with the least error and the best data set in the time series analysis method. For the 5 selected decor papers, monthly and annual purchase amounts from 2009 to 2021 were analyzed separately and the values realized in 2021 were compared with the estimates made by the algorithms, and the algorithms with the lowest error rate, that is, the most accurate estimation, were determined together with the data sets. Estimations were made with the forecast module of the WEKA analysis tool. As it is predicted that the effect of the pandemic and the raw materials crisis will continue beyond 2021, this year was chosen for the test data. A further aim of this study was to guide the business in issues such as determining order lots, order frequency, and the annual total cost of decor paper, all of which are important aspects of the decor paper purchasing strategy (considering supplier constraints).

4. IMPLEMENTATION

In this study, 5 different decor papers were selected from among the raw materials purchased by a company operating in the forest products sector. Decor paper was chosen as it affects costs the most, is used extensively in production, and is difficult to plan for due to the diversity in selection. To prepare the data set for the 5 selected decor papers, the company's real MFC sales data from the years 2009 to 2021 were taken over the SAP system and transferred to Excel. This study was implemented by using these steps;

i. All data preprocessing was carried out in Excel. The data set consisted of 391,845 rows and 20 columns after noisy data and erroneous entries were removed.
ii. The calculations made on the annual and monthly sales amounts of the 5 decor papers selected in this big data, and the monthly and annual usage amounts for each type of paper were calculated using in Excel using the previously mentioned formulas and new data sets were created.
iii. These new datasets are broken down into 2009-2021, 2015-2021, 2019-2021. Then, in order to process the data in WEKA, the Excel files were converted to. arff using the ExceltoArff application.

iv. In the data set, the Year or Month attribute was formatted as “Date” and Amount as “Numeric”. On the time series data set, the attributes in the "Date" format were taken as independent variables and the "Amount" attribute that occurred over time was evaluated depending on time.

The annual and monthly usage amounts of 5 decor papers selected from the big data were used as test data, while the data from the other years were used as training data.

5. RESULTS AND DISCUSSION

5.1. Computational Results

The data set was analyzed together with the time series prediction algorithms in the forecast module of the 3.9.5 version of the WEKA program. The time series prediction algorithms used in this analysis were the RandomSupSpace, CVParemeterSelection, MultiScheme, WIHW, InputMappedClasifier, Zeror, RepTree, DecisionStump, LWL, RondomCommitte, RandomForest, M5P, Randomtree, DecisionTable, BaggingMappedClassifier, RegressionBy5Discretative, and the Kstar RegressionByRegress MultilayerPerceptiron. The data sets created for 4 different periods for 5 selected decors were run with time series prediction algorithms, and the algorithms with the least MAPE, MAE, RMSE, and MSE error values were determined. In the analysis, the default settings for WEKA's prediction algorithms were utilized. The prediction values of the training data for the test data period were compared with the actual values in the test data and are provided below in tables 1 through 5 along with their demand forecasting performance values. As indicated in Table 1, the MultilayerPerceptron algorithm yielded the best estimate for decor1 on an annual basis with a MAPE error rate of 1.54% using the 2019-2021 dataset. Figure 2 shows the predictive value and the actual value of the MultilayerPerceptiron algorithm. Again, the algorithm that gives the second-best estimation is the DecisionTable algorithm with a MAPE error rate of 1.80% on an annual basis and using the 2009-2021 data set. The best estimate on a monthly basis was the DecisionStump algorithm with the 2019-2021 data set and a MAPE error rate of 34.15%. In Figure 3, the graph showing the actual value and the predicted value of the DecisionStump algorithm is given. As seen in Table 2, the RandomSupSpace algorithm gave the best estimate for decor2 on an annual basis with a MAPE error rate of 11.13% using the 2009-2021 dataset. Figure 4 shows the predictive value and the actual value of the RandomSupSpace algorithm. Again, the algorithm that gives the 2nd best estimation is the Kstar algorithm with a MAPE error rate of 16.79% on an annual basis and using the 2015-2021 data set. The best estimate on a monthly basis was the M5Rules algorithm with the 2019-2021 data set and the MAPE error rate of 24.19%. In Figure 5, the graph showing the actual value and the estimated value of the M5Rules algorithm is given. Table 3 indicates that the RegressionByDiscretization algorithm gave the best estimation for decor3 on an annual basis, with the same result as the 2009-2021 and 2015-2021 datasets, with a MAPE error rate of 1.61%. Figure 6 shows the predictive value and the actual value of the RegressionByDiscretization algorithm. Again, the algorithm that gives the 2nd best estimation is the SMOreg algorithm with a MAPE error rate of 5.68% on an annual basis and using the 2019-2021 data set. The best estimate on a monthly basis was the 2009-2021 data set and the RepTree algorithm with a 29.68% MAPE error rate. In Figure 7, the graph showing the actual value and the estimated value of the RepTree algorithm is given.

Table 1. Performance comparison for Decor1 dataset

Algorithms (Decor1)	MAE	MAPE	RMSE	MSE	Data set	t
MultilayerPerceptiron	215.82	1.54	215.82	46,577.49	2019-2021	Year
DecisionTable	252.00	1.80	252.00	63,504.00	2009-2021	Year
DecisionStump	252.00	1.80	252.00	63,504,00	2015-2021	Year
DecisionStump	441.30	34.15	563.04	317,009.00	2019-2021	Month
M5Rules	471.70	37.73	666.52	444,252.76	2009-2021	Month
DecisionTable	518.20	43.72	739.06	546,207.34	2015-2021	Month

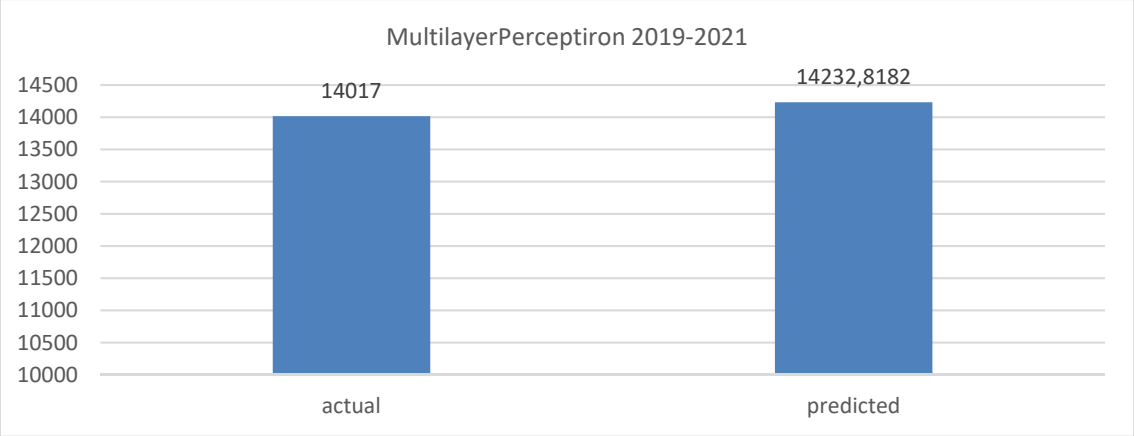

Figure 2. The best predicted value on an annual basis for Decor1

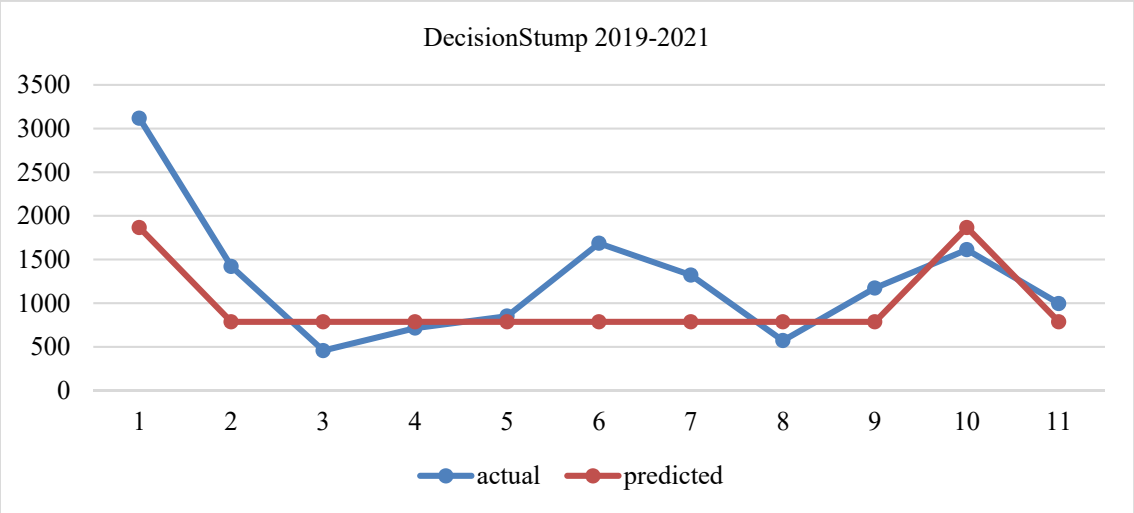

Figure 2. Actual and best forecast values on a monthly basis for Decor1

Table 2. Performance comparison for Decor2 dataset

Algorithms (Decor2)	MAE	MAPE	RMSE	MSE	Date set	t
RandomSupSpace	3,447.74	11.13	3,447.74	11,886,903.03	2009-2021	Year
Kstar	5,202.14	16.79	5,202.14	27,062,294.94	2015-2021	Year
M5Rules	668.50	24.19	851.69	725,376.07	2019-2021	Month
RondomCommitte	850.28	28.07	1,058.70	1,120,846.93	2009-2021	Month
Kstar	9,62900	31.08	9,629.00	92,717,641.00	2019-2021	Year
RepTree	938.94	31.70	1,088.69	1,185,253.43	2015-2021	Month

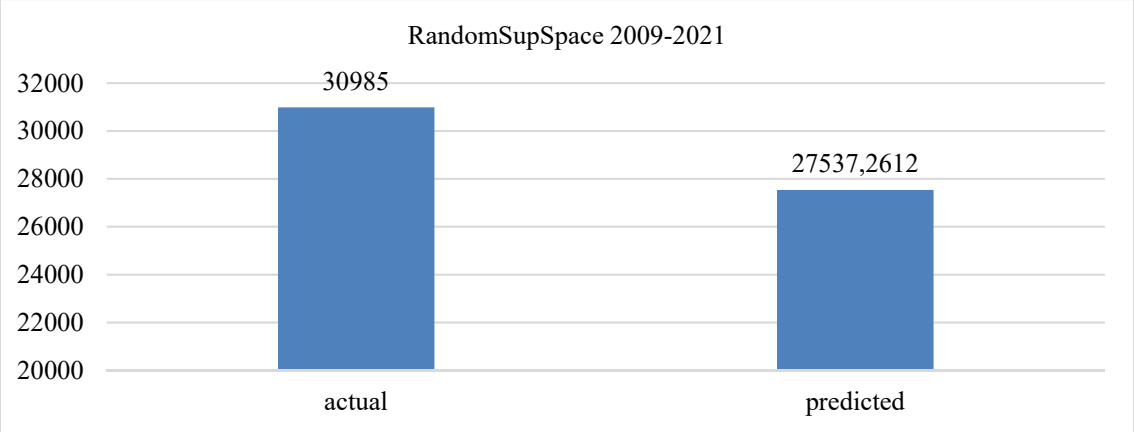

Figure 4. The best predicted value on an annual basis for Decor2

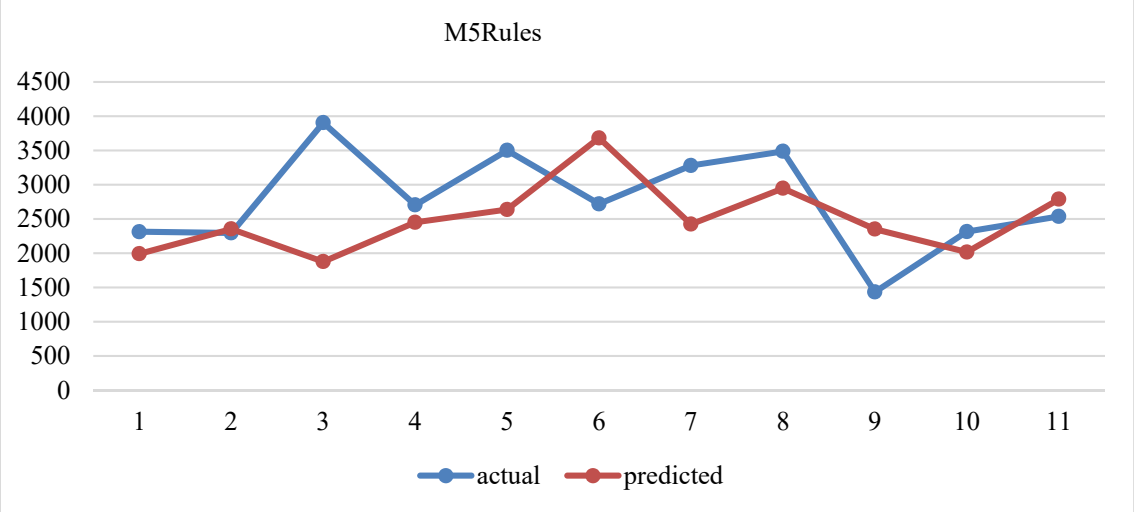

Figure 5. Actual and best forecast values on a monthly basis for Decor2

Table 3. Performance comparison for Decor3 dataset

Algorithms (Decor3)	MAE	MAPE	RMSE	MSE	Data Set	t
RegressionByDiscretization	436.33	1.61	436.33	190,386.78	2009-2021	Year
RegressionByDiscretization	436.33	1.61	436.33	190,386.78	2015-2021	Year
SMOreg	1,54060	5.68	1,540.60	2,373,441.49	2019-2021	Year
RepTree	729.60	29.68	929.04	863,114.35	2009-2021	Month
RegressionByDiscretization	737,33	29.83	938.81	881,358.00	2015-2021	Month
RandomSupSpace	749.01	30.33	951.69	905,722.83	2019-2021	Month

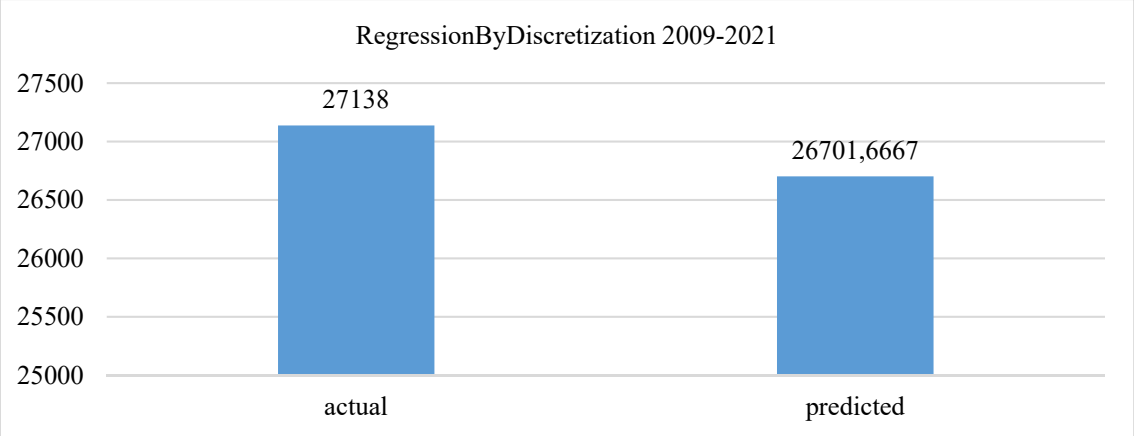

Figure 6. The best predicted value on an annual basis for Decor3

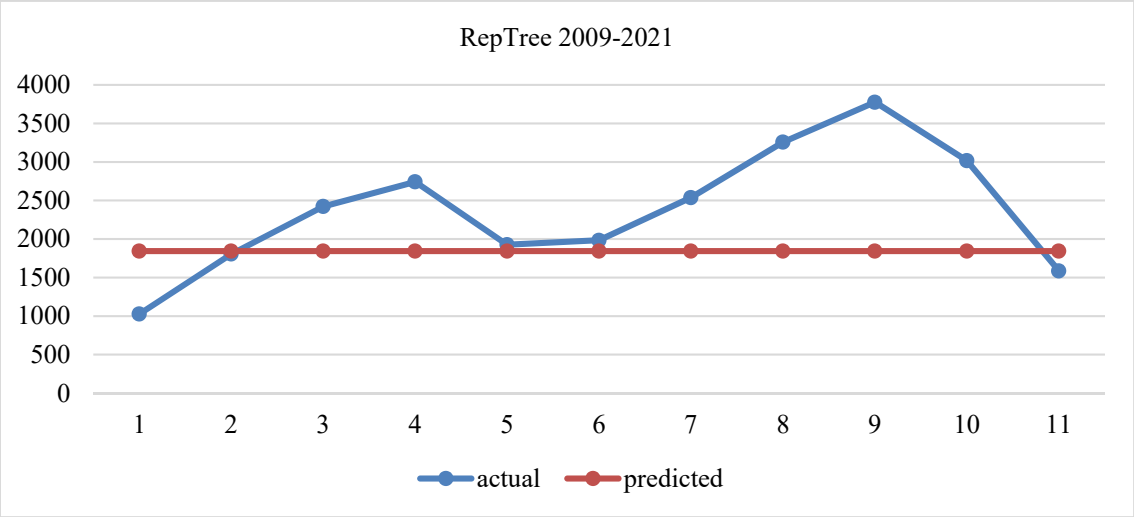

Figure 7. Actual and best forecast values on a monthly basis for Décor3

Table 4. Performance comparison for Decor4 dataset

Algorithms (Decor4)	MAE	MAPE	RMSE	MSE	Data set	t
M5P	4,626.05	4.66	2,286.42	5,227,706.14	2009-2021	Year
RepTree	5,988.33	12.19	5,988.33	35,860,136.11	2015-2021	Year
RandomForest	6,916.00	14.08	6,916.00	47,831,056.00	2019-2021	Year
Bagging	1,418.26	30.85	1,730.82	2,995,754.47	2019-2021	Month
RandomSupSpace	1,384.38	30.98	1,730.65	2,995,150.81	2009-2021	Month
Bagging	1,371.72	31.13	1,731.58	2,998,364.29	2015-2021	Month

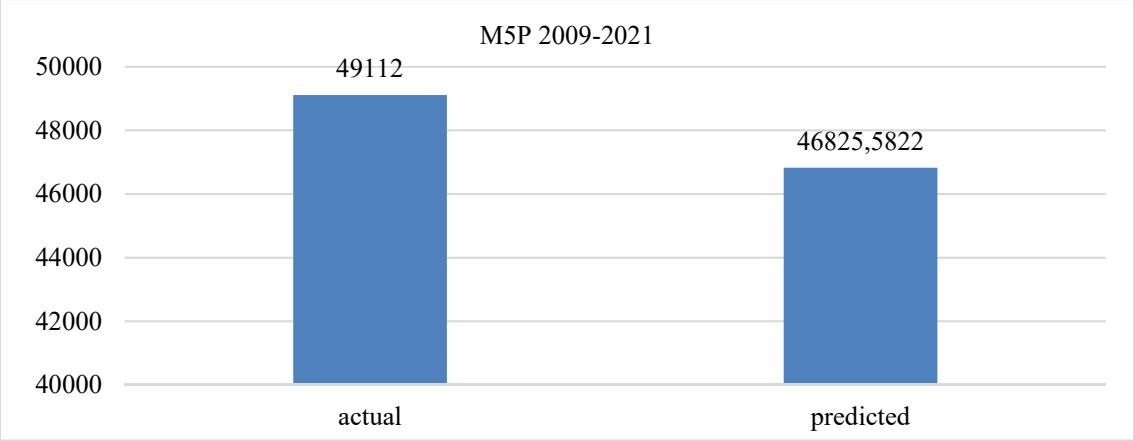

Figure 8. The best predicted value on an annual basis for Decor4

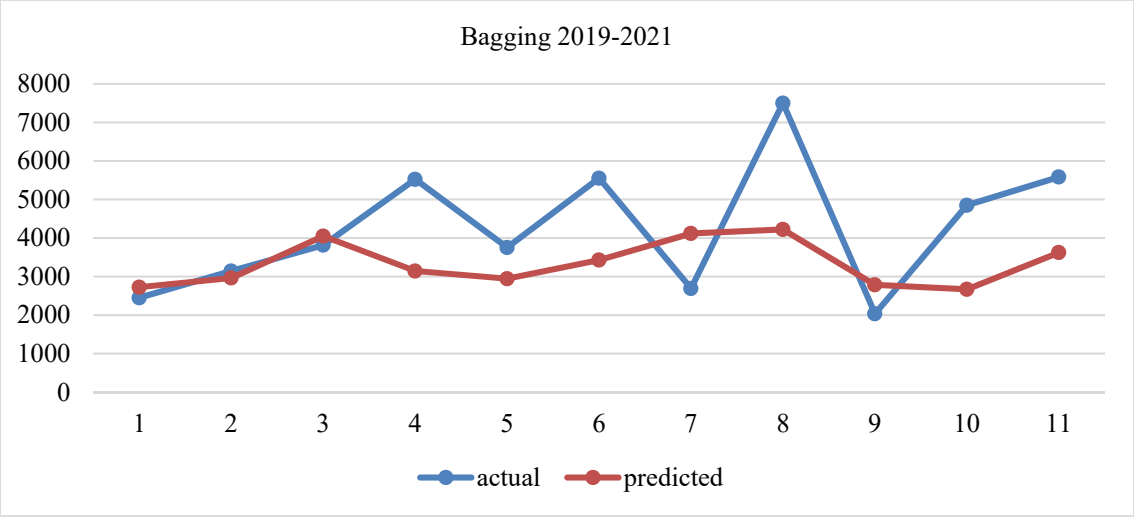

Figure 9. Actual and best forecast values on a monthly basis for Decor4

Table 5. Performance comparison for Decor5 datasets

Algorithms (Decor5)	MAE	MAPE	RMSE	MSE	Data set	T
Kstar	162.00	2.02	162.00	26,244.00	2015-2021	Year
Bagging	316.35	3.95	316.35	100,076.69	2009-2021	Year
SMOreg	3,416.05	42.66	3,416.05	11,669,393.09	2019-2021	Year
Bagging	473.46	51.33	839.88	705,406.72	2019-2021	Month
M5Rules	490.53	56.02	866.33	750,528.29	2015-2021	Month
IBK	608.27	61.92	925.73	856,981.00	2009-2021	Month

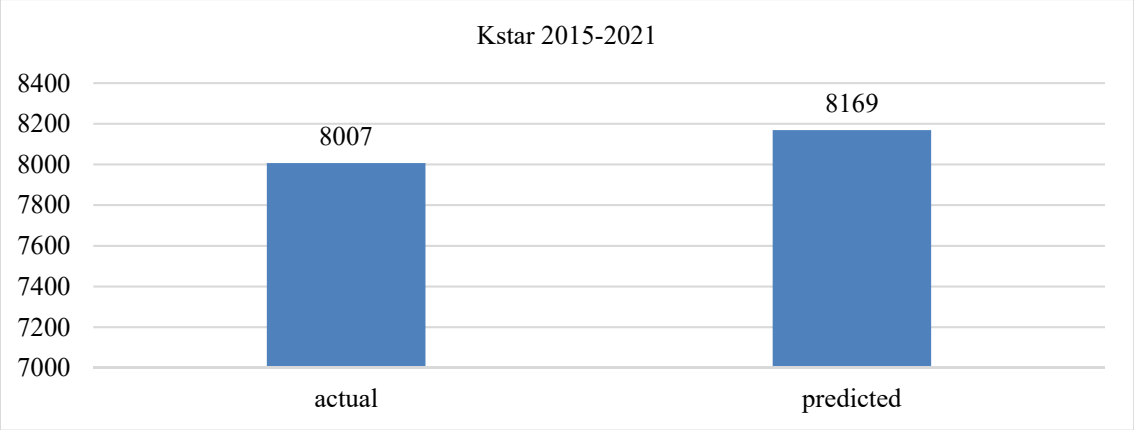

Figure 10. The best predicted value on an annual basis for Decor5

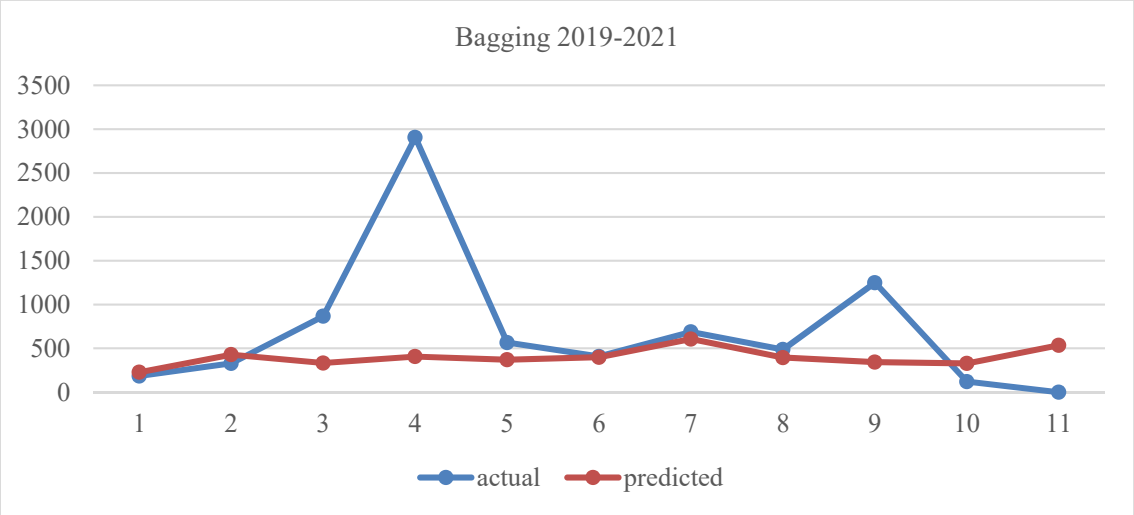

Figure 11. Actual and best forecast values on a monthly basis for Decor5

Table 6. Best Method and Period for Periodic data set according to MAE results

Data Set	**Décor 1**	**Décor 2**	**Décor 3**	**Décor 4**	**Décor 5**
2009-2021 (All Data)	Decision Tree, **Yearly**	Random Committee, **Monthly**	Regression By Discretization, **Yearly**	RandomSup Space, **Monthly**	Bagging, **Yearly**
2015-2021 (Last 6 years)	Decision Tree, **Monthly**	RepTree, **Monthly**	Regression By Discretization, **Yearly**	Bagging, **Monthly**	KStar, **Yearly**
2019-2021 (Last 3 years)	Decision Stump, **Yearly**	M5Rules, **Monthly**	RandomSup Space, **Monthly**	Bagging, **Monthly**	Bagging, **Monthly**

Table 7. Best periodic data set and method for decors according to MAE results

Decor	2009-2021	2015-2021	2019-2021
Décor 1	Multi-Layer Perceptron, **Yearly**		
Décor 2			M5Rules, **Monthly**
Décor 3		Regression By Discretization, **Yearly**	
Décor 4		Bagging, **Monthly**	
Décor 5			Bagging, **Monthly**

Table 4 indicates that the the best estimate for decor4 was the M5P algorithm with a MAPE error rate of 4.66% on an annual basis and using the 2009-2021 data set (Table 4). Figure 8 shows the estimated value and the actual value of the M5P algorithm. Again, the algorithm that gives the 2nd best estimation is RepTree algorithm with a MAPE error rate of 12,19% on an annual basis and using the 2015-2021 data set. On a monthly basis, the Bagging algorithm gave the best estimate with the 2019-2021 data set and a MAPE error rate of 30.85%. In Figure 9, the graph showing the actual value and the estimated value of the Bagging algorithm is given. Table 5 shows that the Kstar algorithm gave the best estimate for decor5 on an annual basis with a MAPE error rate of 2.02% using the 2015-2021 data set (Table 5). Figure 10 shows the predictive value and the actual value of the Kstar algorithm. Again, the algorithm that gives the 2nd best estimation is Bagging algorithm with 3.95% MAPE error rate on an annual basis and using the 2009-2021 data set. On a monthly basis, the Bagging algorithm gave the best estimate with the 2019-2021 data set and the MAPE error rate of 51.33%. Figure 11 shows the actual value and the estimated value of the Bagging algorithm.

5.2. Determining the Best Forecasting Method and Period

Using the data related to the 5 selected decors, the most appropriate estimation method and estimation period for the decor papers are examined here. For this purpose, all results are summarized in Table 6 and Table 7 by looking at the results of the MAE, that is, the average absolute error performance variable. When the data in the table is examined, it is not seen that there is a dominant method in all decor papers. Considering the average absolute error value performance according to the selected data set and period, it is seen that the method differs. On the other hand, when we take the data of the last 3 years as a basis, it is seen that the best results are with the monthly period selection, and in longer periods, the annual and monthly distributions are approximately equally. In general, it is seen that both monthly and annual period selections can be made. In Table 7, the best method and data set, period matching was made on all decor pages, again according to the MAE value. When Table 7 is examined, it is seen that the best performances are realized when data from the last 3 or 6 years are used in all décor papers except Decor 1. In these decors, the weight was also obtained from the monthly estimates. There is no common best practice for all décor papers. This result also makes sense. Demand behavior may differ across all products of a firm, which comes from the nature of demand. On the other hand, in this sense, it is possible to conclude that it is more accurate for decision makers to use the data of the last 3 or 6 years and to make monthly forecasts while using them at forest industry like this study.

Validation of our results according to near literature, we can see that likely results were founded. Yildirim et all [29] were studied about forecasting on production of Non-Wood Forest Products (NWFP) using Turkey import and export value between 1989 and 2011 years. They used MAPE and RMSE values as accuracy of prediction by using ANN models. They founded best results by using yearly import and export values when two hidden ones and one output layer, providing the closest results to the real values. Their MAPE value for test data was reached to 4.66.

Another work was studied by Lin et al. [30]. In their study, they presented forecasting supply and demand of the wooden furniture industry in China. They used ARIMA model as forecasting algorithm. They found the MAPE value as 5.2666 for forecasting accuracy value future among 2018 and 2023 years. An also they took in account yearly data as period.

6. CONCLUSION

Due to various reasons such as energy and shipping container crises and fluctuations in exchange rates, there are constant changes in raw material prices and difficulties in raw material supply. This has made it difficult to prepare an accurate budget at the purchasing stage and to create the right purchasing strategies at the right time. Purchasing strategies made with only experience and foresights from the market are not sufficient in a period when technology is so prominent, and competition is so intense. It is an undeniable reality that companies need to use technology and data science in order to stay one step ahead of their competitors and make a difference. Using ERP Systems like SAP, companies create huge raw data. Today companies and decision makers come across a question, how I can handle this data to produce beneficial results according to competition. At Forest industry like all other industries when it comes to competition, demand comes to mind. Therefore, accurately predicting future demand is closely related to many corporate activities, including purchasing and sales. Scientific studies about this topic are related to nearly Big Data. For this reason, applications where decision makers can see how they will process Big Data will be of great benefit.

For all purposes in this study 5 different decor papers were selected from among the raw materials used in production in order to determine the purchasing strategies of an enterprise operating in the forest products sector. Specific to these decors, time series analyses were carried out using the monthly and annual real sales data of the enterprise from the years 2009 to 2021 together with quantity attributes, and the test data for 2021. Looking at the estimation results, it is seen that different algorithms give the best estimates for different data sets. Considering that the data sets were prepared according to certain time periods, it can be concluded that the estimates vary according to time periods. The importance of not only the algorithms, but also the periodic data sets used are revealed in order to obtain the best estimation results in time series analysis techniques performed on products such as decor paper that change according to current trends and tastes. As a result of the estimation results obtained from this study, the company will be able to make future estimations by considering the best algorithms and data sets for the selected decor papers while making their future purchasing plans and will be able to develop a more efficient purchasing strategy by considering the results that will occur while making its plans. In addition, using the workflow developed in this study, the company will be able to create an integrated strategy by determining the best dataset, period and algorithms for the decor papers that are not covered in the study. Using an artificial intelligence application, it will be possible to automatically run the data on SAP over this flow and display the current best estimates for next month or year.

7. REFERENCES

1. Ferguson, W.C., Hartley, M.F., Turner, G.B., Pierce, E.M., 1996. Purchasing's Role in Corporate Strategic Planning. International Journal of Physical Distribution & Logistics Management, 26(4), 51-62.
2. Kaes, I., Azeem, A., 2009. Demand Forecasting and Supplier Selection for Incoming Material in RMG Industry: A Case Study. International Journal of Business and Management, 4(5), 149-157.
3. Kim, M., Jeong, J., Bae, S., 2019. Demand Forecasting Based on Machine Learning for Mass Customization in Smart Manufacturing. In Proceedings of the 2019 International Conference on Data Mining and Machine Learning, 6-11.
4. Arif, M.A.I., Sany, S.I., Nahin, F.I., Rabby, A.S.A., 2019. Comparison Study: Product Demand Forecasting with Machine Learning for Shop. In 2019 8th International Conference

System Modeling and Advancement in Research Trends (SMART), 171-176.
5. Gupta, S., Sihag, P., 2022. Prediction of the Compressive Strength of Concrete Using Various Predictive Modeling Techniques. Neural Computing and Applications, 34(8), 6535-6545.
6. Panarese, A., Settanni, G., Vitti, V., Galiano, A., 2022. Developing and Preliminary Testing of a Machine Learning-Based Platform for Sales Forecasting Using a Gradient Boosting Approach. Applied Sciences, 12, 11054.
7. Nasseri, M., Falatouri, T., Brandtner, P., Darbanian, F., 2023. Applying Machine Learning in Retail Demand Prediction—A Comparison of Tree-Based Ensembles and Long Short-Term Memory-Based Deep Learning. Applied Sciences, 13, 11112.
8. Aksoy, A., Ozturk, N., Sucky, E., 2012. A Decision Support System for Demand Forecasting in the Clothing Industry. International Journal of Clothing Science and Technology, 24(4), 221-236.
9. Yunishafira, A., 2018. Determining the Appropriate Demand Forecasting Using Time Series Method: Study Case at Garment Industry in Indonesia. KnE Social Sciences, 553-564.
10. Ren, S., Chan, H.L., Siqin, T., 2020. Demand Forecasting in Retail Operations for Fashionable Products: Methods, Practices, and Real Case Study. Annals of Operations Research, 291(1), 761-777.
11. Yadav, A., Ghosh, S., 2019. Forecasting Monthly Farm Tractor Demand for India Using MSARIMA and ARMAX Models. Indian Journal of Agricultural Research, 53(3), 315-320.
12. Huber, J., Stuckenschmidt, H., 2020. Daily Retail Demand Forecasting Using Machine Learning with Emphasis on Calendric Special Days. International Journal of Forecasting, 36(4), 1420-1438.
13. Spiliotis, E., Makridakis, S., Semenoglou, A.A., Assimakopoulos, V., 2020. Comparison of Statistical and Machine Learning Methods for Daily SKU Demand Forecasting. Operational Research, 22, 3037-3061.
14. Panigrahi, S., Behera, H.S., 2020. Time Series Forecasting Using Differential Evolution-Based ANN Modelling Scheme. Arab J Sci Eng, 45, 11129–11146.
15. Moroff, N.U., Kurt, E., Kamphues, J., 2021. Machine Learning and Statistics: A Study for Assessing Innovative Demand Forecasting Models. Procedia Computer Science, 180, 40-49.
16. Ngo, N.T., Pham, A.D., Truong, T.T.H., 2022. An Ensemble Machine Learning Model for Enhancing the Prediction Accuracy of Energy Consumption in Buildings. Arab J Sci Eng, 47, 4105–4117.
17. Pham, Q.B., Kumar, M., DiNunno, F., Elbeltagi, A., Granata, F., Islam, A.R.M., Anh, D.T., 2022. Groundwater Level Prediction Using Machine Learning Algorithms in a Drought-Prone Area. Neural Computing and Applications, 34, 10751-10773.
18. Shi, Z., Wang, G., 2018. Integration of Big-Data ERP and Business Analytics (BA). The Journal of High Technology Management Research, 29(2), 141-150.
19. Han, J., Kamber, M., 2006. Data Mining: Concepts and Techniques, 2nd. Ed. University of Illinois at Urbana Champaign: Morgan Kaufmann, 735.
20. Dwivedi, S., Kasliwal, P., Soni, S., 2016. Comprehensive Study of Data Analytics Tools (Rapidminer, WEKA, R Tool, Knime). In 2016 Symposium on Colossal Data Analysis and Networking (CDAN), 1-8.
21. Shevade, S.K., Keerthi, S.S., Bhattacharyya, C., Murthy, K.R.K., 2020. Improvements to the SMO Algorithm for SVM Regression. IEEE Transactions on Neural Networks, 11(5), 1188-1193.
22. Witten, I.H., Frank, E., 2002. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Acm Sigmod Record, 31(1), 76-77.
23. El-Bendary, N., Elhariri, E., Hazman, M., Saleh, S.M., Hassanien, A.E., 2016. Cultivation-Time Recommender System Based on Climatic Conditions for Newly Reclaimed Lands in Egypt. Procedia Computer Science, 96, 110-119.
24. Asaju, L.A.B., Shola, P.B., Franklin, N., Abiola, H.M., 2017. Intrusion Detection System on a Computer Network Using an Ensemble of

Randomizable Filtered Classifier, K-Nearest Neighbor Algorithm. FUW Trends in Science & Technology Journal, 2(1), 550-553.

25. Pal, S.K., Mitra, S., 1992. Multilayer Perceptron, Fuzzy Sets, Classification. IEEE Transactions on Neural Networks, 3(5), 683-697.

26. Mirmozaffari, M., Alinezhad, A., Gilanpour, A., 2017. Data Mining Classification Algorithms for Heart Disease Prediction. Int'l Journal of Computing, Communications & Inst. Engg, 4(1), 11-15.

27. Lin, W., Wu, Z., Lin, L., Wen, A., Li, J., 2017. An Ensemble Random Forest Algorithm for Insurance Big Data Analysis. IEEE Access, 5, 16568-16575.

28. Pratola, M.T., Chipman, H.A., Gattiker, J.R., Higdon, D.M., McCulloch, R., Rust, W.N., 2014. Parallel Bayesian Additive Regression Trees. Journal of Computational and Graphical Statistics, 23(3), 830-852.

29. Yildirim, I., Ozsahin, S., Okan, O.T., 2014. Prediction of Non-Wood Forest Products Trade Using Artificial Neural Networks. J. Agr. Sci. Tech., 16, 1493-1504.

30. Lin, M., Zang, Z., Cao, Y., 2019. Forecasting Supply and Demand of the Wooden Furniture Industry in China. Forest Products Journal, 69 (3), 228-238.