All Sciences Proceedings
http://as-proceeding.com/

2nd International Conference on Engineering, Natural and Social Sciences

April 4-6, 2023 : Konya, Turkey

https://www.icensos.com/

Statistical optimization of forecast data from Adaptive Boosting and Support Vector Machine Algorithms

Abdulkadir Atalan*

[1]Industrial Engineering, Gaziantep Islam Science and Technology University, Gaziantep, Turkey

*(abdulkadiratalan@gmail.com)

Abstract **–**This study aimed to calculate the optimum values of estimation data based on adaptive boosting (AB) and support vector machine (SVM) algorithms from machine learning (ML) models with statistical optimization models. Three independent and two dependent variables were used in this study. It was arranged for ML algorithms using 750 data of each variable. The training and testing phases in ML algorithms were set at a rate of 90% and 10%, respectively. The RMSE, MSE, and MAE values, which are the error rates, and the coefficient of Determination R^2 values, were compared to verify the validity of the ML algorithms. The estimation results of the independent variables were analyzed with a nonlinear optimization model. The results obtained were validated with a high degree of desirability and the validity of the optimization model. AB algorithm provided the best performance for y_1 and y_2 dependent variables. The desirability degree of the optimization model of the variables y_1 and y_2 was calculated as 0.945. Based on the AB algorithm, the optimum value of the y_1 and y_2 variables were computed at 6.89 and 0.6169, respectively. The optimum values of the x_1, x_2, and x_3 independent variables for both optimization models were calculated as 3.729, 0.509, and 13.814, respectively. As a result, the desirability values of the optimum values of ML models were calculated, and the validity of the optimum values of the optimum and actual data was verified in this study.

Keywords – Machine Learning; Adaptive Boosting; Support Vector Machine; Statistical Optimization; Variables

I. INTRODUCTION

Machine learning (ML) algorithms are commonly used to analyze large data sets and obtain forecast data [1]. ML algorithms consist of two phases in terms of structure. Most datasets are trained and tested with the rest of the data [2]. Data rates for training and testing phases differ in studies. ML algorithms also have two types of variables; input and output. In the present paper, two different dependent variables were considered besides more than one independent variable.

Researchers widely prefer ML algorithms in many fields, such as healthcare, production, finance, energy, and transportation [3]–[7]. A study used stepwise-multiple linear-regression (SMLR), artificial neural-networks (ANN), support-vector machines (SVM), and gradient-boosting-machine (GBM) models to estimate patients' waiting times in the emergency department (ED) [8]. Another study used various ML algorithms such as neural network (NN), random forest (RF), SVM, elastic net (EN), multivariate adaptive regression splines (MARS), k-nearest neighbor (kNN), GBM, classification and regression tree (CRT), and linear regression (LR) to predict wait times in a radiology facility and postponement times in programmed radiology centers [9]. Li et al. have used the Poisson Lasso-Regression (PLR) model and RF algorithms to predict and classify quality assurance results for volumetric modulated arc therapy plans [10]. One study preferred ML techniques for the relationship between the costs of doctors and nurses from

healthcare resources employed in an emergency department and the cost of patients treated [11]–[14].

A study discussed the accuracy of different ML models and the implications of these models for applicable scenarios to have a clear view of varying ML methods in the field of intelligent transportation systems [15]. Another study compared ML methods to analyze the effects on the production costs of the jet engine portions produced in the production area [16]. One study has utilized ML algorithms to analyze e-scooter delivery vehicles applied in mail or package delivery regarding cost, energy, and environmental factors and to obtain forecast data [17]. Atalan used four different ML algorithms, SVM, NN, and AdaBoost (AB) algorithms, to analyze and predict unit prices of drinking milk [2]. Researchers use ML algorithms frequently to obtain forecast data for energy consumption and cost [18]. Robinson et al. tried to estimate commercial building energy consumption using ML models [19]. Another study used combining multilayer perceptron (MLP), SVM, and CatBoost models to estimate the energy consumption for renewable and non-renewable power bases [20].

ML algorithms are tested with some engineering applications to calculate the prediction data and ensure the accuracy of the obtained data. At the beginning of these applications are statistical and optimization techniques. One study tested the validity with Bayesian optimization and the ML method for processing a sample's Kernel type and hyperparameters evaluated from the Gaussian process (GP) [21]. In another study, discrete-event simulation techniques and ML algorithms were used to predict patient waiting times in the healthcare field [3], [22].

The present study has four main sections. The first part includes the literature review of the study. The second part contains detailed information about the techniques used and the method of the study. The numerical results of the study are discussed in the third part of the paper. General statements about the study are given in the last part of the research.

II. Materials and Method

In this paper, error data were obtained to confirm the validity of the estimation data and ML algorithms to obtain the estimation data of the data belonging to the dependent variables. In addition, a statistical optimization model was developed to obtain the optimum values of the forecast dataset. A visual representation of the method of this study is shown in **Figure 1**.

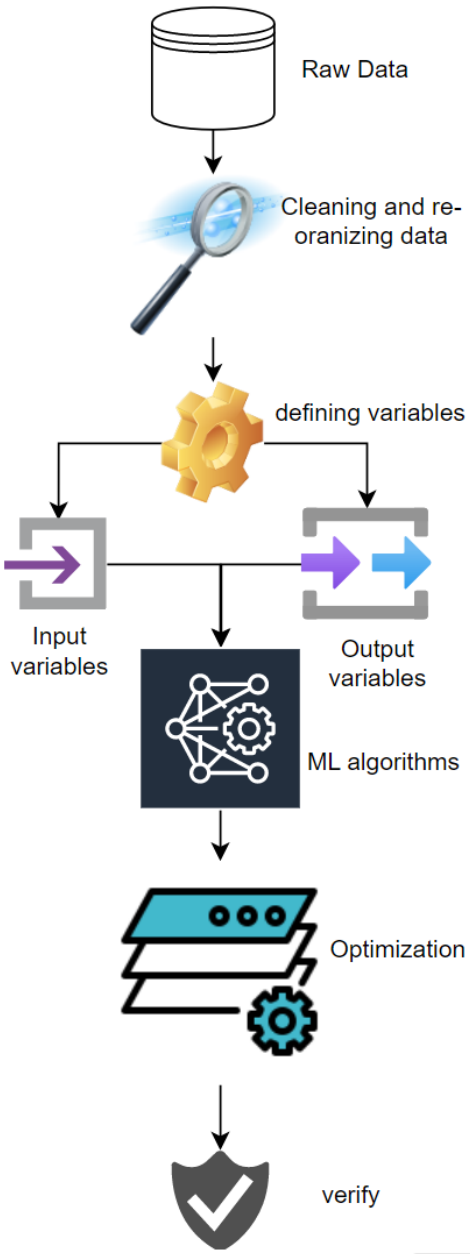

Figure 1: The flowchart of the method for the present study

In this study, three independent and two dependent variables were used. Three thousand seven hundred fifty data were used, provided that there were 750 data belonging to these variables. Descriptive statistics values of these data are shared in Table 1. Generally, descriptive statistical measures consist of the values of sample size, mean, standard deviation, variance, kurtosis, skewness, maximum, and minimum values, etc. [23], [24]

Table 1. The descriptive values of independent and dependent factors

Variable	N	Mean	SE-Mean	St-Dev	Variance	Coef-Var	Min	Max	Skew	Kurt
x_1	750	3.80	0.00	0.10	0.01	2.71	3.47	4.16	-0.01	-0.04
x_2	750	0.80	0.00	0.09	0.01	11.60	0.51	1.13	0.13	0.04
x_3	750	11.87	0.04	1.14	1.30	9.61	8.37	16.19	0.06	-0.08
y_1	750	7.09	0.00	0.08	0.01	1.11	6.18	8.15	1.11	65.02
y_2	750	0.60	0.00	0.02	0.00	3.45	0.30	0.64	-5.25	67.26

*N, the number-of-samples; SE-Mean, standard-error-of-mean; St-Dev, standard-deviation; Coef-Var, coefficient-of-variance; Max, the value of maximum; Min ,the value of minimum; Skew, the value of skewness; Kurt, the value of kurtosis

The correlation values of the input and output factors used for this paper are exposed in Figure 2. Generally, the correlation values between two variables range from -1 to +1. As the correlation values of the factors get closer to the opposite poles, the correlation strength increases [25], [26]. There is a weak, moderate, and strong interaction between the variables regarding correlation values [27].

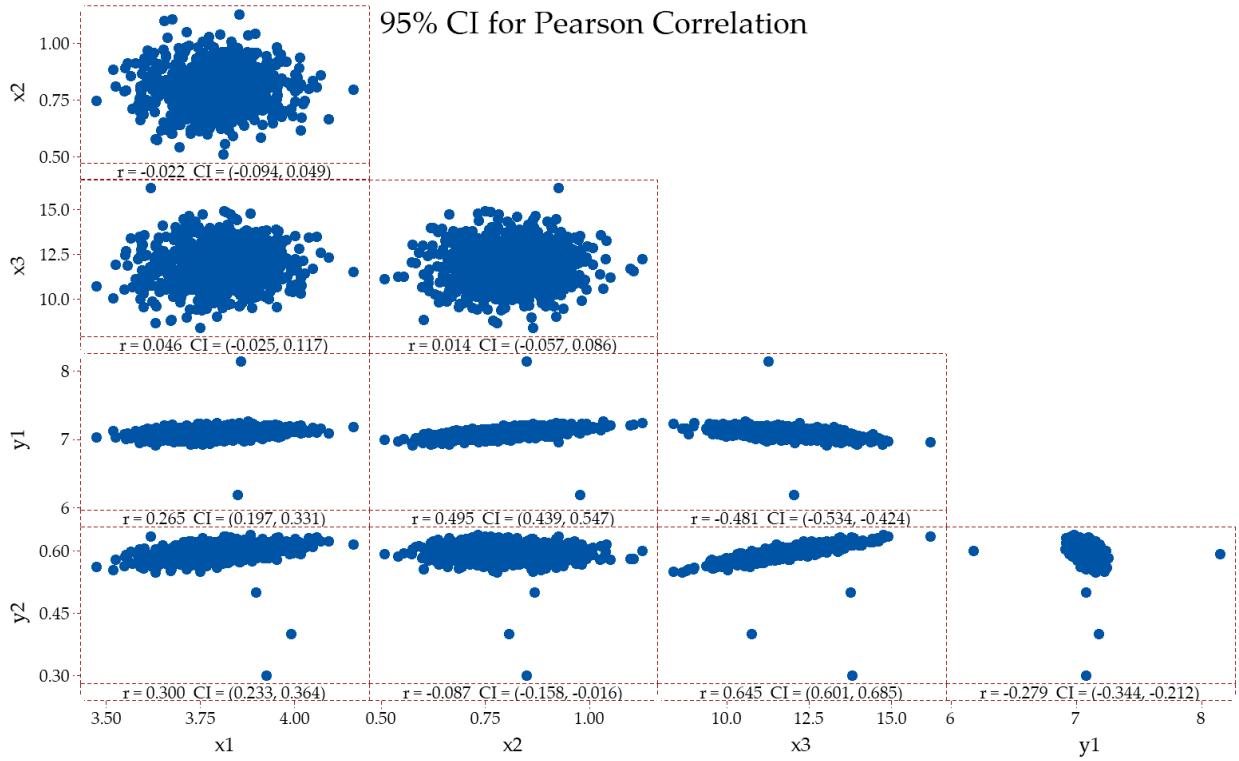

Figure 2: The correlation data between variables

This study preferred two different ML algorithms, adaptive boosting (Adaboost-AB) and support vector machine (SVM). AB algorithm was presented by Freund & Schapire in 1995. The working principle of this algorithm is expressed as creating a set of classifiers and then classifying them as test samples [28]. A support vector machine (SVM) algorithm performs well on the categorical dependent variable. For this reason, this algorithm is usually an ML model that learns by labeling the data contained in the dependent variable. For example, this algorithm is used to predict data of dependent variables with binary or more outcomes such as gender types, pass/fail status, and up/down [29]. The ML algorithms used for this study were run in the Orange 3.14 software computer program, and the prediction data were obtained. The visual of ML models in the Orange 3.14 program is shown in Figure 3.

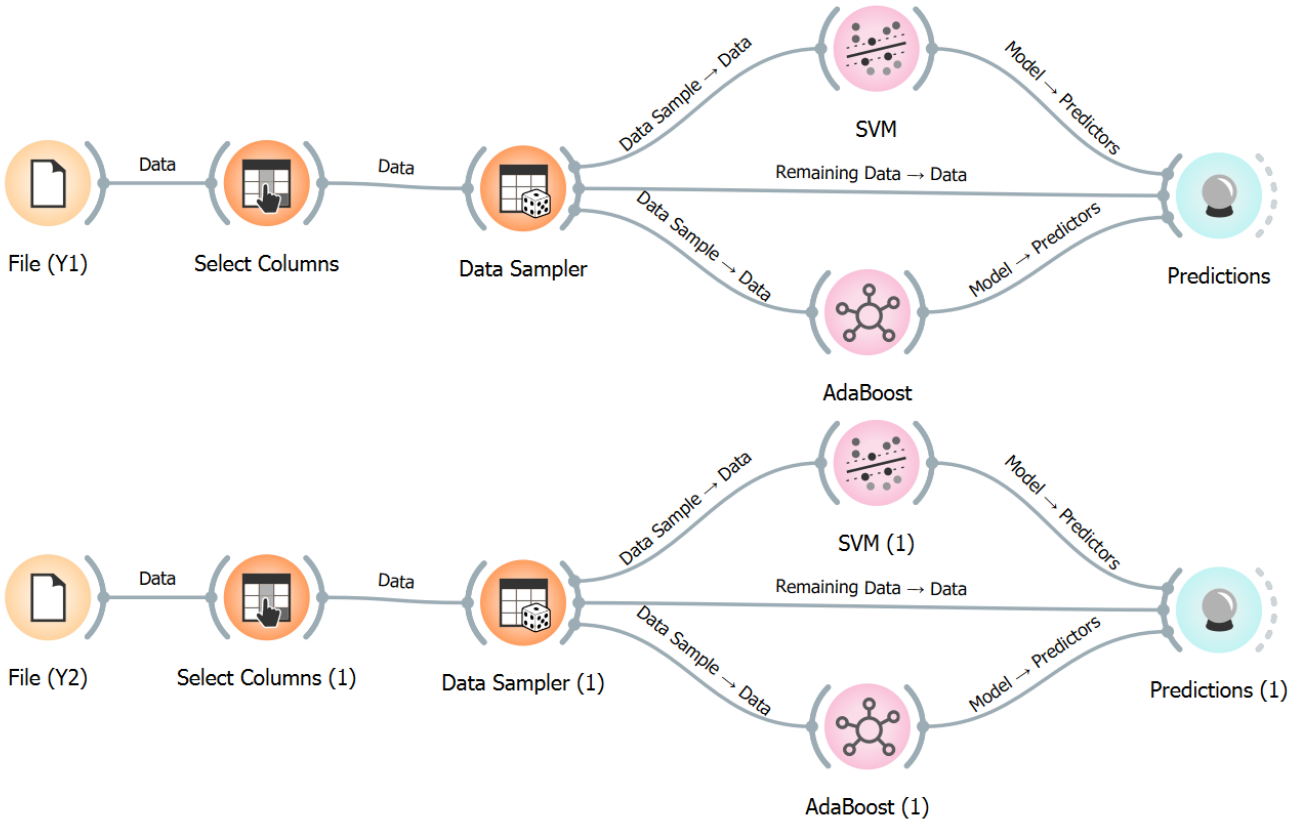

Figure 3. The flowchart of proposed ML models for the present study

RMSE (Root-mean-square deviation), MSE (mean-squared-error), and MAE (Mean Absolute Error) formulas representing error data were used to verify the validity of the results of the preferred ML algorithms for this study. At the same time, the R^2-value, which is the coefficient of Determination, was calculated, and the performance ranking of the ML algorithms was made. The mathematical equivalents of the performance measurement criteria of the algorithms are as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \tilde{y}_i|^2 \quad (1)$$

$$MSE = \sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n}} \quad (3)$$

$$R^2 = \sum_{i=1}^{n}\left[\frac{y_i - \tilde{y}_i}{y_i - \bar{y}_i}\right]^2 \quad (4)$$

Calculation of the optimum values using the values of the estimation data obtained by using the AB and SVM algorithms of the dependent variables was obtained using the following formula:

$$\begin{matrix} max/\min Z \\ l_i \le x_i \le u_i \end{matrix} \quad (5)$$

Since the purpose directions of the two dependent variables in this study differ, this optimization model is expressed as a nonlinear optimization model. The intentions of this model are limited to the lower (l_i) and upper (u_i) limits of the independent variables (x_i).

III. RESULTS

For this study, 3750 data belonging to 3 independent and two dependent variables were analyzed and tested in ML algorithms. To verify the validity of the ML algorithms, the performance measurement values are shown in **Table 2**.

Table 2. Performance measurement values of ML algorithms

Variable	Models	MSE	RMSE	MAE	R^2
y_1	SVM	0.003	0.051	0.044	0.047
	AB	0.000	0.010	0.008	0.961
y_2	SVM	0.006	0.077	0.074	0.561
	AB	0.001	0.038	0.006	0.761

AB algorithm provided the best performance for y_1 and y_2 dependent variables. The RMSE, MSE, MAE, and R^2 values, among the AB algorithm's performance measures for the y_1 variable, were calculated as 0.000, 0.010, 0.008, and 0.961, respectively. Depending on the SVM algorithm, the same variable's RMSE, MSE, MAE, and R^2 values were calculated as 0.003, 0.051, 0.044, and 0.047, respectively.

The RMSE, MSE, MAE, and R^2 values, among the AB algorithm's performance measures for the y_2 variable, were computed as 0.001, 0.038, 0.006, and 0.761, respectively. Depending on the SVM algorithm, the same variable's RMSE, MSE, MAE, and R^2 values were calculated as 0.006, 0.077, 0.074, and 0.561, respectively.

The optimum values of the actual data of the dependent and independent variables are shared in Table 3. In the developed optimization model, while the aim of the y_2 variable is in the maximum direction, the objective of the y_1 variable is set to the minimum. The optimum values of the y_2, y_1, x_1, x_2, and x_3 variables were calculated as 0.634, 6.740, 3.473, 0.509, and 16.193, respectively. The desirability value was calculated as 0.839 to verify the validity of this optimization model. As this value approaches 1, the accuracy of the optimum values obtained increases.

The optimum values of unestimated real data are shown in Figure A1 in the appendix of the study.

Table 3. Optimum values of the real data of the dependent and independent factors

Response	y_2	y_1	x_1	x_2	x_3
Goal	Maximum	Minimum	*	*	*
Lower	0.300	*	*	*	*
Target	0.638	6.177	*	*	*
Upper	*	8.148	*	*	*
Weight	1.000	1.000	*	*	*
Importance	1.000	1.000	*	*	*
SE Fit	0.003	0.012	*	*	*
95% CI	(0.627, 0.641)	(6.717, 6.762)	*	*	*
95% PI	(0.604, 0.664)	(6.636, 6.844)	*	*	*
Optimum	0.634	6.740	3.473	0.509	16.193

* non-applicable

Optimum values were obtained by running the estimation data of ML models in optimization models. The optimum values of the independent factors corresponding to the estimation data of the dependent factors are shown in Table 4. The optimum values of the data estimated by ML models are shown in Figure A2 in the appendix of the study.

Table 4. Optimum values for the predicted data of the dependent variables and the actual data of the independent variables

Response	AB (y_2)	SVM (y_2)	AB (y_1)	SVM (y_1)	Optimum
Goal	Maximum	Maximum	Minimum	Minimum	*
Lower	0.561	0.508			*
Target	0.629	0.533	6.988	7.046	*
Upper			7.195	7.255	*
Weight	1.000	1.000	1.000	1.000	*
Importance	1.000	1.000	1.000	1.000	*
SE Fit	0.001	0.002	0.004	0.017	*
95% CI	(0.615, 0.618)	(0.528, 0.536)	(6.889, 6.904)	(6.951, 7.019)	*
95% PI	(0.614, 0.619)	(0.524, 0.541)	(6.879, 6.914)	(6.910, 7.060)	*
X_{1_1}					3.729
X_{2_1}	*	*	*	*	0.509
X_{3_1}	*	*	*	*	13.814
AB (y_2)	*	*	*	*	0.617
SVM (y_2)	*	*	*	*	0.533
AB (y_1)	*	*	*	*	6.896
SVM (y_1)	*	*	*	*	6.985

* Non-applicable

For the y_1 and y_2 dependent variables, the optimum values of the estimation data obtained in AB and SVM models were computed. The optimum value of the y_1 variable, according to the SVM algorithm, was 6.985. The optimum value of the same variable was computed as 6.896 according to the AB model.

The optimum value of the y_2 variable, according to the SVM model, was computed as 0.533. The optimum value of the same variable was calculated as 0.617 according to the AB algorithm. The desirability degree of the optimization model of the variables y_1 and y_2 was computed as 0.945. The optimum values of the x_1, x_2, and x_3 independent variables for both optimization models were calculated as 3.729, 0.509, and 13.814, respectively.

In this study, a numerical case study was made, and the proposed method was compared in four different ways:

1. Performance measurement data of the estimation data of the AB model,
2. Performance measurement data of the prediction data of the SVM model,
3. Optimum values of actual data,
4. Optimum values according to forecast data,

Thus, the proposed method provides an excellent opportunity for researchers to use in cases where it is difficult to obtain actual data in terms of cost and time. This study has a few limitations. The data used in this study were derived by derivation. The types of arguments in the data set are handled numerically. However, the analysis did not include the independent variable of categorical data. Finally, additional research is required for integer studies, as the decision variables in the created nonlinear mathematical model are not considered integers.

IV. CONCLUSION

This study compared the optimum values of the ML models' prediction and actual data. In this study, 3750 data belonging to three independent and two dependent rudiments were analyzed, and optimum values of these variables were computed. Desirability values were calculated to test the validity of the optimum values between the estimated data and the actual data. The real and predicted data desirability values were calculated as 0.8395 and 0.9450, respectively. It has been determined that the desirability data of the forecast data is higher than the desirability data of the actual data.

In addition, the performance measurement data of the ML algorithms were computed, and the validity of the estimation data of the ML algorithms was

verified. The following points stand out with the method proposed in this study:

1. ML algorithms for forecast data provide robust results,
2. It has been ensured that the statistical optimization method applies to the optimum values of the ML data,
3. The desirability values of the optimum values of ML models were computed, and the validity of the optimum values of the optimum and actual data was verified.

APPENDIX

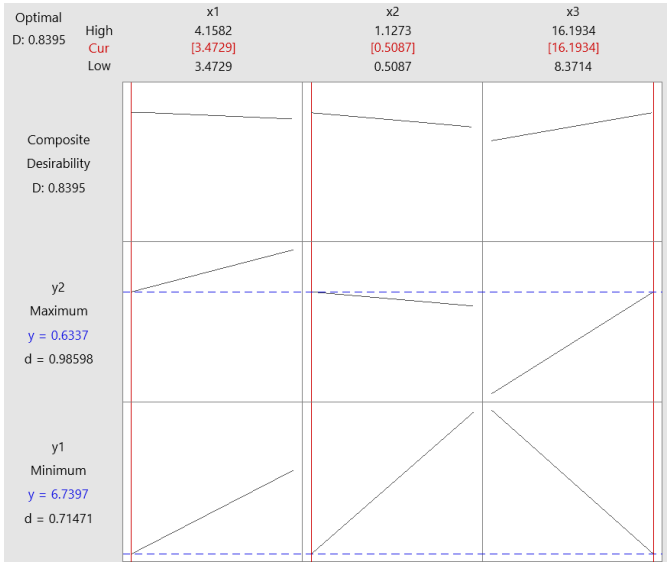

Figure A1. Optimum values of unpredictable real data

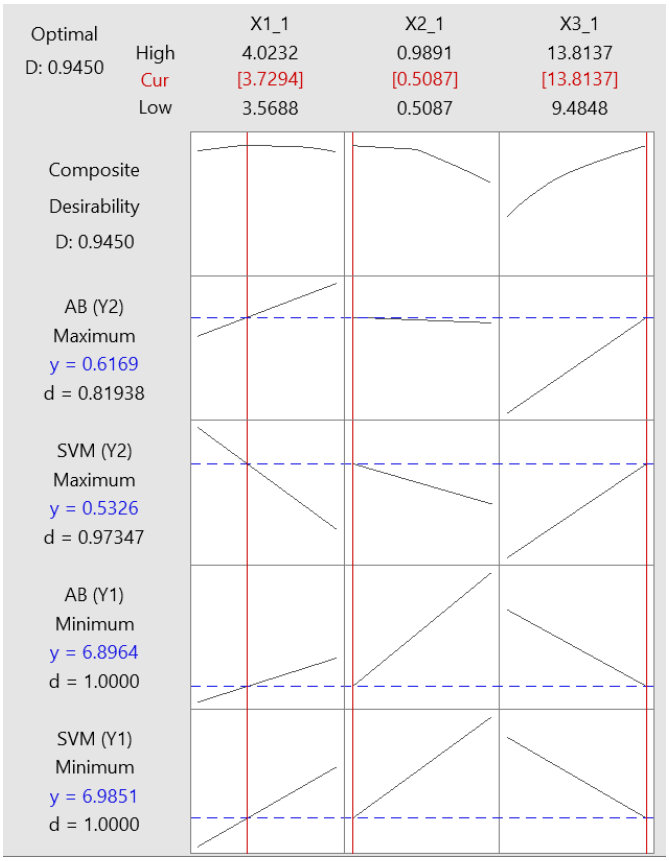

Figure A2. Optimum values of data estimated by ML algorithms

REFERENCES

[1] F. López-Martínez, E. R. Núñez-Valdez, V. García-Díaz, and Z. Bursac, "A Case Study for a Big Data and Machine Learning Platform to Improve Medical Decision Support in Population Health Management," *Algorithms*, vol. 13, no. 4, p. 102, Apr. 2020, doi: 10.3390/a13040102.

[2] A. Atalan, "Forecasting drinking milk price based on economic, social, and environmental factors using machine learning algorithms," *Agribusiness*, vol. 39, no. 1, pp. 214–241, Jan. 2023, doi: 10.1002/agr.21773.

[3] A. Atalan, H. Şahin, and Y. A. Atalan, "Integration of Machine Learning Algorithms and Discrete-Event Simulation for the Cost of Healthcare Resources," *Healthcare*, vol. 10, no. 10, p. 1920, Sep. 2022, doi: 10.3390/healthcare10101920.

[4] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová, "Optimal errors and phase transitions in high-dimensional generalized linear models," *Proc. Natl. Acad. Sci.*, vol. 116, no. 12, pp. 5451–5460, Mar. 2019, doi: 10.1073/pnas.1802705116.

[5] V. Mhasawade, Y. Zhao, and R. Chunara, "Machine learning and algorithmic fairness in public and population health," *Nat. Mach. Intell.*, vol. 3, no. 8, pp. 659–666, Aug. 2021, doi: 10.1038/s42256-021-00373-4.

[6] A. Atalan, "Forecasting for Healthcare Expenditure of Turkey Covering the Years of 2018-2050," *Gümüşhane Üniversitesi Sağlık Bilim. Derg.*, vol. 9, no. 1, pp. 8–16, Apr. 2020, doi: 10.37989/gumussagbil.538111.

[7] Z. Ceylan and A. Atalan, "Estimation of healthcare expenditure per capita of Turkey using artificial intelligence techniques with genetic algorithm-based feature selection," *J. Forecast.*, vol. 40, no. 2, pp. 279–290, Mar. 2021, doi: 10.1002/for.2747.

[8] Y.-H. Kuo *et al.*, "An Integrated Approach of Machine Learning and Systems Thinking for Waiting Time Prediction in an Emergency Department," *Int. J. Med. Inform.*, vol. 139, p. 104143, 2020, doi: https://doi.org/10.1016/j.ijmedinf.2020.104143.

[9] C. Curtis, C. Liu, T. J. Bollerman, and O. S. Pianykh, "Machine Learning for Predicting Patient Wait Times and Appointment Delays," *J. Am. Coll. Radiol.*, vol. 15, no. 9, pp. 1310–1316, 2018, doi: https://doi.org/10.1016/j.jacr.2017.08.021.

[10] J. Li *et al.*, "Machine Learning for Patient-Specific Quality Assurance of VMAT: Prediction and Classification Accuracy," *Int. J. Radiat. Oncol.*, vol. 105, no. 4, pp. 893–902, 2019, doi: https://doi.org/10.1016/j.ijrobp.2019.07.049.

[11] S. Srinivas and A. R. Ravindran, "Optimizing outpatient appointment system using machine learning algorithms and scheduling rules: A prescriptive analytics framework," *Expert Syst. Appl.*, vol. 102, pp. 245–261, 2018, doi: https://doi.org/10.1016/j.eswa.2018.02.022.

[12] A. V. L. N. Sujith, G. S. Sajja, V. Mahalakshmi, S. Nuhmani, and B. Prasanalakshmi, "Systematic review of smart health monitoring using deep learning and Artificial intelligence," *Neurosci. Informatics*, vol. 2, no. 3, p. 100028, 2022, doi: https://doi.org/10.1016/j.neuri.2021.100028.

[13] A. Atalan and C. C. Dönmez, "Optimizing experimental simulation design for the emergency departments," *Brazilian J. Oper. Prod. Manag.*, vol. 17, no. 4, pp. 1–13, 2020, doi: 10.14488/BJOPM.2020.026.

[14] A. Atalan and C. Donmez, "Employment of Emergency Advanced Nurses of Turkey: A Discrete-Event Simulation Application," *Processes*, vol. 7, no. 1, p. 48, Jan. 2019, doi: 10.3390/pr7010048.

[15] A. Boukerche and J. Wang, "Machine Learning-based traffic prediction models for Intelligent Transportation Systems," *Comput. Networks*, vol. 181, p. 107530, 2020, doi: https://doi.org/10.1016/j.comnet.2020.107530.

[16] J.-L. Loyer, E. Henriques, M. Fontul, and S. Wiseall, "Comparison of Machine Learning methods applied to the estimation of manufacturing cost of jet engine components," *Int. J. Prod. Econ.*, vol. 178, pp. 109–119, 2016, doi: https://doi.org/10.1016/j.ijpe.2016.05.006.

[17] H. İnaç, Y. E. Ayözen, A. Atalan, and C. Ç. Dönmez, "Estimation of Postal Service Delivery Time and Energy Cost with E-Scooter by Machine Learning Algorithms," *Appl. Sci.*, vol. 12, no. 23, p. 12266, Nov. 2022, doi: 10.3390/app122312266.

[18] H. Ghoddusi, G. G. Creamer, and N. Rafizadeh, "Machine learning in energy economics and finance: A review," *Energy Econ.*, vol. 81, pp. 709–727, 2019, doi: https://doi.org/10.1016/j.eneco.2019.05.006.

[19] C. Robinson *et al.*, "Machine learning approaches for estimating commercial building energy consumption," *Appl. Energy*, vol. 208, pp. 889–904, 2017, doi: https://doi.org/10.1016/j.apenergy.2017.09.060.

[20] P. W. Khan, Y.-C. Byun, S.-J. Lee, D.-H. Kang, J.-Y. Kang, and H.-S. Park, "Machine Learning-Based Approach to Predict Energy Consumption of Renewable and Nonrenewable Power Sources," *Energies*, vol. 13, no. 18, p. 4870, Sep. 2020, doi: 10.3390/en13184870.

[21] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," *Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, doi: https://doi.org/10.48550/arXiv.1206.2944.

[22] L. Capocchi, J.-F. Santucci, and B. P. Zeigler, "Discrete Event Modeling and Simulation Aspects to Improve Machine Learning Systems," in *2018 4th International Conference on Universal Village (UV)*, Oct. 2018, pp. 1–6, doi: 10.1109/UV.2018.8642161.

[23] A. Atalan and C. Dönmez, "Evaluation of Healthcare Economics of OECD Countries: Multi-Objective Statistical Optimization Model," *Acta Infologica*, vol. 5, no. 1, pp. 197–206, Jul. 2021, doi: 10.26650/acin.836372.

[24] A. Atalan, "Logistic Performance Index of OECD Members," *Akad. Araştırmalar ve Çalışmalar Derg.*, vol. 12, no. 23, pp. 608–619, Nov. 2020, doi: 10.20990/kilisiibfakademik.720604.

[25] A. Atalan, "EFFECT OF HEALTHCARE EXPENDITURE ON THE CORRELATION BETWEEN THE NUMBER OF NURSES AND DOCTORS EMPLOYED," *Int. J. Heal. Manag. Tour.*, vol. 6, no. 2, pp. 515–525, Jul. 2021, doi: 10.31201/ijhmt.949500.

[26] A. Atalan, Z. Çınar, and M. Çınar, "A trendline analysis for healthcare expenditure per capita of OECD members," *Sigma J. Eng. Nat. Sci.*, vol. 10, no. 3, pp. 23–35, 2020.

[27] A. Atalan, "Türkiye Sağlık Ekonomisi için İstatistiksel Çok Amaçlı Optimizasyon Modelinin Uygulanması," *İşletme Ekon. ve Yönetim Araştırmaları Derg.*, vol. 1, no. 1, pp. 34–51, 2018, [Online]. Available:

http://dergipark.gov.tr/download/article-file/414076.

[28] D. D. Margineantu and T. G. Dietterich, “Pruning adaptive boosting,” in *ICML*, 1997, vol. 97, pp. 211–218.

[29] W. S. Noble, “What is a support vector machine?,” *Nat. Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, 2006, doi: 10.1038/nbt1206-1565.