IT Support Ticket Completion Time Prediction

Mihra Yıldız
ExperTeam R&D Center
Istanbul, Türkiye
mihra.yildiz@experteam.com.tr

Ali Alsaç
Istanbul University - Cerrahpaşa
Istanbul, Türkiye
ali.alsac@ogr.iuc.edu.tr

Taner Ulusinan
ExperTeam R&D Center
Istanbul, Türkiye
taner.ulusinan@experteam.com.tr

Murat Can Ganiz
Marmara University
Istanbul, Türkiye
murat.ganiz@marmara.edu.tr

Mehmet Mutlu Yenisey
Istanbul University - Cerrahpaşa
Istanbul, Türkiye
mehmet.yenisey@iuc.edu.tr

2022 7th International Conference on Computer Science and Engineering (UBMK) | 978-1-6654-7010-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/UBMK55850.2022.9919591

***Abstract*—Prediction of the time that will be spent on IT support tickets is very important for planning and optimization of IT support services that are usually bound with service level agreements. Predicting completion time of a ticket is a difficult problem, which requires substantial experience and technical expertise if done manually by a human. However, it is possible to automate this task using supervised machine learning models given we have a large amount of labeled data. In this study, we employ supervised machine learning algorithms to predict completion time of tickets for IT support. We use a real-world dataset that includes about 17 thousand tickets. We employ data science approaches to preprocess and transform the input and feed to supervised machine learning algorithms for learning models for ticket completion time prediction. More specifically we use Linear Regression, Decision Trees Regression, Random Forest Regression, Support Vector Machines Regression, and Multiple Regression algorithms. For the evaluation of these supervised models, we use several metrics such as MAE, MSE, and MAPE. Our results show varying success levels with different supervised machine learning algorithms for this difficult task. Among the models we train, the Decision Trees and Random Forest Regression show promising results.**

Keywords— IT Support, Prediction, Machine Learning, Supervised Learning, Regression, Data Science

I. INTRODUCTION

Today almost every company or organization extensively use IT systems to support their daily operations. The number and complexity of the IT systems they use increase by the adaptation of new technologies that the businesses require. This makes it difficult to maintain and support these systems and as a result they tend to use remote support at an increasing rate. Especially after the Covid-19 epidemic, remote working and remote technical support became quite popular. Today, although it is not a necessity, remote working continues to increase its popularity due to cost, security, service quality and similar conditions. This is, of course, especially the case for Information Technologies (IT) support. Compared to the other sectors, IT support usually requires skilled personnel and therefore it is expensive. It is important to use advanced technologies in this field to manage and improve IT support systems. A successful IT support management relies on five steps when end-users access a service desk with demands:

1. System creates a ticket and pushes it to a central pool
2. Category/topic of this ticket is determined (automatically or by a manager)
3. Ticket is prioritized and routed to the right agent
4. Ticket gets resolved
5. Customer is notified

Factors determining quality and efficiency in remote IT support services are Service Level Agreement (SLA) constraints and costs. A SLA agreement usually enforces completion time constraints on different types of issues and problems defined in a IT support ticket. If these requirements are not met, if a certain type of ticket is not completed in the stated time window, then the IT support company is subject to certain fees and penalties. Of course, frequent violation of the SLA limits can also cause loss of business. A remote IT support company usually serves many customer companies, each can have different SLAs with different constraints and costs. Maintaining the conditions in SLAs is very important for the commercial success of the remote IT support company.

The successful prediction of the completion times of the IT tickets is very important in order to avoid SLA related fees and costs and retain customers. Furthermore, the output of these prediction models can be used in agent assignment and overall optimization of IT support operations.

In ticket assignment, a wrong expert assignment can cause several problems and may require reassignment. One of the common problems caused by this is the longer than normal ticket completion times. Normal completion time can be

978-1-6654-7010-0/22/$31.00 ©2022 IEEE

achieved by assigning the ticket to an expert that has necessary technical knowledge and experience related to the issue in the ticket and availability. Abnormal completion times can be caused by several reassignments of the ticket until finding the most suitable and available expert. The wrong expert can mean several different things. For example, it can be an expert with insufficient technical knowledge or experience for the ticket or an expert with sufficient qualifications but extremely busy with other higher priority tickets in her queue.

Longer than normal completion time may lead to exceeding SLA, which causes considerable fines and loss of business. Predicting the completion time of a ticket when assigned to an expert is therefore very important for planning purposes.

There are several areas where advanced analytical systems can be used in this domain. One of the popular areas is the category classification of tickets, especially the descriptions of the IT problems that are expressed in text in an email or a web form. Another important area is ticket completion time prediction. In this task, structured data is more commonly used instead of free text problem descriptions. Similarly, supervised machine learning, more specifically numeric prediction or regression algorithms can be used.

Predicting the completion time of the job provides information that supports the decision of assigning the newly opened ticket to the most appropriate agent according to the resolution time. Creating a supervised machine learning model for ticket completion time for these purposes forms the basis of our work. This work is organized as follows; we talk about the related work in the next section. Our approach is described in section III. Following this, we provide our experiment setup and that is followed by results and discussion. We conclude our paper in section VI along with our directions for the future work.

II. Related Work

Machine Learning (ML) is a subfield of Artificial Intelligence (AI). It has strong ties to statistics and mathematical optimization. ML enables the modeling of smarter behaviors by enabling them to learn by themselves rather than being programmed directly. It produces data-based predictions by developing models that discover related or related information in datasets and use these connections to make predictions [1]. Many publications and fields of study in academic fields such as statistics, computer sciences, engineering sciences, numerical methods, economics and business are supported by machine learning methods. In this section, selected ones from supervised learning and prediction studies in various sectors and branches of science mentioned above are presented. The method of Kitchenham [2] was used in the flow system of literature research.

Davies [3] compared four machine learning algorithms, K-Nearest Neighbors (k-NN), Linear Regression, Regression Trees, and SVM Regression, to predict overall operating room usage time. As inputs, 12 variables such as the type of procedure obtained from the records in the operating rooms, the identity of the surgeon and the date of the operation, the age and gender of the patient, the anesthesiologist and the operating site were used. The results show that the Regression SVM and Regression Trees algorithms have the smallest error in predicting the occupancy of the operating room and the prediction model has 80% success.

Master et al [4] aimed to provide solution support to the operating room scheduling problems with the decision support system in their studies where they used Supervised Learning methods. Patient gender (male and female), Patient weight (in kilograms), Patient age (in years), Patient physical condition/score as defined by the American Society of Anesthesiologists (ASA), Primary surgeon ID, Location, Patient class (inpatient and outpatient), a regression model was created using attributes such as the procedure name. This prediction problem is known to be particularly difficult in pediatric hospitals due to the extreme diversity in pediatric patient populations. Two supervised learning approaches were followed. One is a direct prediction of surgical case times using features derived from electronic medical records and hospital operational information, and the other is a classification problem of correct, overestimated, or underestimated to evaluate each prediction made by surgeons. It has been suggested that the prediction model gives as successful results as the experts.

Hosseini et al. [5] were interested in the duration predictions of the processes carried out in the operation rooms in their studies. They designed a two-step model with the dataset selected from hospital records. In the first step, classification was used to group the procedure codes and reduce the number of sub-factors for the procedure code area. In the next step, two separate regression models were developed for the prediction of the duration of the operation processes using classical least square linear regression in which the main factors are included and stepwise regression in which the main factors and second-order interactions are included. Then, the prediction results were compared and the model was evaluated.

Fairley et al. [6] used a combination of ML regressors and integer programming (deterministic optimization) to predict the completion time in the anesthesia department of a hospital. With the proposed Regression Tree algorithm, a prediction model was predicted using records of approximately 18,000 surgeries based on 2016 data.

Bender and Ovtcharova [7] focused on Lead Time Prediction (LTP) in their study where they implemented an AutoML application that can be integrated into the ERP systems of small and medium-sized enterprises.

In a make-to-order system, new orders that come in dynamically must be assigned a delivery date, which requires real-time prediction of order flow time. Alenezi et al. [8]develops a support vector regression model for real-time flow time prediction in multi-source, multi-product systems. The prediction error of the support vector regression model for three different multisource systems of varying complexity is compared with classical time series models and a feedforward neural network. The shared results show that the support vector regression model has lower flow time prediction error.

DeCos Juez et al. [9] stated that the purpose of their article is to predict the production times required for the production of different types of metallic components of aviation engines. This

study not only predicts production times, but also accompanies this prediction with an analysis to identify which factors are most important for on-time lots. Attributes that affect delivery times Batch size (units), CNC machine (minutes), Grinding machine (minutes), Heat treatments (minutes), Horizontal lathe (minutes), Individual serial number (yes/no), Inspection tests time (time) in minutes), Manufacturing forecasted cost (€), Milling machine (minutes), Raw material cost (€), Surface treatments (minutes), Vertical lathe (minutes). Support vector regression was used in this study.

In the article published by Little in 1961, the queuing theory formula known as Little's Law is used in analytical methods in production control systems. Building on Rust's 2008 work, Gyulai et al. [10] compared analytical/classical methods with Machine Learning-based (ML-based) methods in his study on the production control system of a company with Flow Type production method and found that machine learning-based methods achieved more successful results. According to the related study, LT (lead Time) predictions made using Linear Regression, Support Vector Regression, and Decision.

A reliable model for predicting changes in water levels in a river is crucial for better planning to reduce any risk associated with flooding. Ahmed et al. [11], based on the data collected from 1990 to 2019 in Malaysia, linear regression (LR), interaction regression (IR), robust regression (RR), stepwise regression (SR), support Various Machine Learning (ML) algorithms have been developed, including vector regression (SVR), boosted trees ensemble regression (BOOSTER), bagged trees ensemble regression (BAGER), XGBoost, tree regression (TR) and Gaussian process regression (GPR).

Another study comparing prediction systems and deep learning studies and also examining the literature is Zhang et al. [12] Published by. Algorithms in the categories of MLP (Multiple Regression Method), Autoencoder, CNNs, RNNs, RBM, NADE, Neural Attention, Adversary Network, DRL, Hybrid Models are explained and a basis for future studies is presented.

Kadiri and Ravala [13] working on the detection of security threats for cloud service providers, investigated attack types such as Denial of Service and Distributed Denial of Service, Dictionary, Eavesdropping, Password, Phishing, Snooping using Regression Based model and Kernel Based model. In the model, it is claimed that the Kernel-based model works with 87% accuracy.

Liu et al. [14] proposed a hybrid model for the prediction of household energy consumption by using the support vector machine regressors (SVM) together with the Empirical Mode Decomposition method (EMD). Fourier transform was performed on the consumption data, and then it was decomposed with the EMD method and turned into a set for the SRV method. According to the results, the hybrid model EMD-SRV was found to be more successful than the single models.

El-Basyouny and Sayed [30], use an accident, volume and geometric data sample corresponding to 392 arterial segments in British Columbia, Canada in their regression model for accident prediction. The purpose of their paper is to compare two types of regression techniques: traditional negative binomial (TNB) and modified negative binomial (MNB). According to the results they shared, they claim that the MNB method is superior in prediction success.

Different from these studies, we apply machine learning algorithms to a relatively less studied topic of predicting failures in RDBMSs. Furthermore, we apply many machine learning algorithms since the results of algorithms are mostly unpredictable and the best performing algorithms in a particular domain should be selected based on experimentation. This is also known as the no free lunch theorem in the machine learning domain. Additionally, we also consider data preprocessing and feature selection methods as important as the choice of machine learning algorithms. In this regard, we put an additional effort in this.

III. Approach

Our database consists of 16970 IT support tickets from several different customer companies assigned to large numbers or experts. The data in this database is in one of the common normal forms therefore distributed to many tables. So we first denormalize and transform the data into a single table. This preprocessing step also includes feature generation. About 20 features are created at this step. Some of these features are eliminated by manual observation of the domain experts. Most of the available features are categorical. These are also transformed to numeric features using one-hot-encoding for attributes with a small number of values. For attributes with large numbers of values, we use integer encoding if it is meaningful. Missing values are either filled with column average or the instance is deleted depending on the number of missing values in that instance. We also explore several normalization techniques such as z-score normalization. We manually checked for the outliers and eliminated several of them.

Numeric prediction is a supervised machine learning task. Regression algorithms are commonly used in this task. In regression analysis, it is aimed to understand the effect of the change of one or more inputs (independent variable) on the output (dependent variable). The regression problem, which is the subject of this article, aims to predict or model service completion time from historical data. The regression algorithms we use and their classification counterparts are listed in Table 1. We use the most commonly used machine learning library, the scikit-learn in Python programming language. For the preprocessing operations we also use Python libraries such as pandas, numpy, matplotlib, and seaborn along with SQL.

TABLE I. Classification and Regression algorithms

Classification Algorithms	Regression Algorithms
Logistic Regression	Linear Regression (LR) Multiple Linear Regression (MLR)
Support Vector Machines	Support Vector Regression (SVR)
Decision Tree	Decision Tree Regression (CART)
Random Forest	Random Forest Regression (RFR)

Regression is concerned with the prediction of continuous values. Regression is divided into two main categories as Simple Linear and Multiple. In simple linear regression, a straight line is drawn to describe the relationship between two variables (X and Y). In contrast, Multiple regression covers multiple variables and is further divided into linear and nonlinear.

Random forest (RF) method introduced by Leo Breiman [25] is a tree-based ensemble of trees connected to a collection of random variables and is one of the supervised machine learning methods [24]. It is included in our study as it performs well in many different domains.

Another algorithm we use in our model is the Support Vector Machines based regression algorithms. Support vector machines (SVM) is a supervised machine learning method based on the Vapnik-Chervonenkis theory [22],[24]. The adaptation of SVM for regression, which is widely used for classification problems, was proposed by Smola et al., and this method was named Support Vector Regression (SVR) [23].

CART (Classification and regression tree), one of the algorithms based on tree structure [26]. It is an algorithm that creates two child nodes from each node and continues this process until the formation of homogeneous nodes, where the strongest correlation is obtained between the observation variable and the prediction variable.

Evaluation is essential for making comparisons between algorithms. It is necessary to understand the algorithms or combinations of algorithms that are most suitable for our model. The idea of minimizing an error is fundamental in machine learning and is essentially the foundation of all learning algorithms. The regression evaluation criteria and formulas used in our experiments follows [27] including Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Coefficient of Determination R^2, where y_i represents actual value, $\check{y}_i$, predicted value, and $\underline{y}$ mean of real values.

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

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \frac{|y_i - \check{y}_i|}{y_i} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \check{y}_i)^2}{\sum_{i=1}^{n} (y_i - \underline{y})^2} \quad (4)$$

$$\underline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad (5)$$

IV. Experiment Setup

The algorithms we used for prediction are Linear Regression (LR), Decision Tree Regression (DTR), Support Vector Machines Regression (SVR), Random Forest Regression (RFR), Multiple Regression (MRM) from the scikit-learn library of Python programming language. Default parameters which are programmed in the scikit-learn library are used for these algorithms.

Three different experiment plans are created. First experiment plan uses whole data for both training and testing the algorithms. Using the hold-out method in the evaluation of our data set, we divided our data into training, 70% and test data, 30% [28]. In practice, it is quite common to use one-third (1/3) of the available data for testing and the rest for training. It was not considered necessary in our study, but we also state that some of the data to be used for training can be separated as validation data [29]. In the second experiment, the data is split as %70 testing %30 training data by using test_ train_ split function in scikit-learn library. In the third experiment, finally, the confidence interval was set to 95%. This range is especially used when interpreting the OLS (Ordinary Least Squares) table.

In our dataset a ticket represents a task to be completed or a problem to be solved in a particular IT system. Features of a ticket include creation date, description, target completion date, job type, classification, complexity level, priority, customer account (sometimes complemented by a sub account), plus a technician and resolution date assigned later. In our study, we use following features as an input to the machine learning based regression algorithms:

- Issue type: {incident, services request, change request, question, proactive, story, epic, task}
- Issue category: {BI, CST, Custom, DB, EAM, FA, CL, HR, INV, OE, PA, PO, SF, SYSADMIN}
- Priority: {blocker, critical, major, minor, trivial}
- Employee type: {Analyst, DBA, Developer}
- Day of Month: [1, 31]
- Day of Week: [1,7]
- Hour of Day: [0, 23]

Prediction results for completion times were obtained by applying the prediction model with the specified algorithms. The prediction success of the algorithms evaluated at this stage with evaluation metrics and results are provided in the next section.

V. Results and Discussion

All algorithms are run for each experiment setting. You can see the evaluation results for the test on training data settings in Table II. Using the training set as a test set is usually done to see if the problem is learnable. Along with the result of separate - exclusive training set and test set, this also helps us to determine if a certain model is overfitting or not. Indeed, the next table, table III shows the evaluation of the models on a separate test set. Those are the main results that give us an indication of these models' real-world performance. As can be seen in this table, Support Vector Regression (SVR) has the lowest Mean Average Error (MAE) and Mean Average Percentage Error (MAPE). Interestingly, In Table II, where we test the model using the training set, CART model has by far lowest error values but in Table III it performs way worse. We

suspect that CART highly overfits the data. On the other hand, SVR seems quite stable and generalizes well.

TABLE II. EVALUATION OF ALGORITHMS ON TRAINING SET

Algorithms	R^2	MAE	MAPE
LR	0.01	390.95	4.30
SVR	-0.01	281.71	1.08
CART	0.98	61.36	0.39
RFR	0.86	159.86	1.30
PR	0.03	389.06	4.10
MLR	0.01	390.95	4.30

TABLE III. EVALUATION OF ALGORITHMS ON SEPARATE TEST SET

Algorithms	R^2	MAE	MAPE
LR	0.03	389.66	4.31
SVR	-0.02	287.39	1.11
CART	-0.28	353,43	2,34
RFR	0.01	308.36	2.11
PR	0.03	389.66	4.31
MLR	0.03	389.66	4.31

After the data preprocessing was done via Python, the experimental stage was started with the determined algorithms. At this stage, the dataset was run with each algorithm separately, and the results were obtained and the evaluation criteria and the results were arranged to be interpreted.

It is important to look at Ordinary Least Squares (OLS) tables as an analysis based on coefficient data in regression models for evaluating the results [18]. OLS results are available in Python with the statsmodels plugin. It is a result table where prediction success can be checked. It offers important connections for the detection of multicollinearity in multiple regression models [19].

The OLS method aims to minimize the sum of square differences between the observed and predicted values. Using the OLS summary table, the output variables we use in our analyses are Degree of freedom model, Constant term, Coefficient term, p – values, R – squared value, Prob(Omnibus).

Degree of freedom is the number of independent observations on the basis of which the sum of squares is calculated.

The constant term is the intercept of the regression line. From the regression line (Table IV.) the intercept is 139.6268. In regression we omit some independent variables that do not have much impact on the dependent variable, the intercept tells the average value of these omitted variables and noise present in the model.

The coefficient term tells the change in Y for a unit change in X, i.e if issue_type parameter rises by 1 unit then Y rises by 57.5324.

Prob (Omnibus) is supposed to be close to the 1 in order for it to satisfy the OLS assumption. In this case Prob (Omnibus) is 0.000, which implies that the OLS assumption is not satisfied.

In the analyzes which is made according to the OLS table, the R^2 and adjusted R^2 values should be checked. R^2 is a measure of the success of prediction [20]. According to the backward elimination structure, R^2; Adj.R^2 value is determined according to the forward elimination structure [21]. Elimination iterations are made according to the "p>|t|" column in the OLS table. The rows with p value greater than 0.05 are eliminated and the OLS cycle is run again. This iteration is continued until there is no value greater than 0.05 in the p column. The reason why the p value is 0.05 is because the confidence interval was determined as 95%.

The OLS result table of the 1st iteration, it is seen that the COMPNAME, PRIORTY_Blocker, PRIORTY_Critical, PRIORTY_ Minor,EMPLOYEE_TYPE_Analyst, EMPLOYEE_TYPE_DBA properties have no effect on the prediction because the p values are greater than 0.05. These features were removed and the second iteration was run.

The p value of the PRIORTY_ Trivial property is bigger than 0.05 as a result of the elimination applied to the resulting table after the first iteration. At this point, the PRIORTY_ Trivial property is eliminated in the second iteration and a new iteration is started.

At the end of the third iteration, the iteration is stopped because the p values of all the features are less than 0.05.

TABLE IV. OLS OUTPUT VARIABLES

Output	Iteration I	Iteration II	Iteration III
df-Model	10	6	5
Constant	139.62	195.63	197.41
R^2	0,045	0,045	0,045
Adj. R^2	0,044	0,044	0,044

We mentioned that the R^2 values in the OLS table are an important criterion for evaluating the prediction success of the model. Results obtained. It is seen that the Multiple Regression method does not produce a successful prediction for our model.

VI. CONCLUSION AND FUTURE WORK

The completion times prediction of tickets collected through relational database management systems (RDBMS) is important because it affects many critical processes such as personnel planning, cost control, and increasing the level of

customer satisfaction. The benefits of this prediction study are especially high for companies that provide database management and consultancy services to many different companies such as ExperTeam (Uzman Bilişim Danışmanlık A.Ş.). That is why we are conducting extensive experiments to develop a machine learning-based completion time prediction system. Predicting the completion time will definitely help companies control their time, effort and costs, considering the resources spent on each ticket.

Various data science approaches are used to preprocess and convert input data from raw format to format that can be fed into algorithms. After preprocessing, the dataset is evaluated by giving it to machine learning algorithms.

Studies on real-life data may not work with as high success results as in fictional data sets. In this case, it may be necessary to apply data mining methods on the data set. Another solution would be to use feature engineering methods that use data attributes directly. In our future work, it is aimed to increase the prediction success by using feature selection and feature engineering methods.

In the future work we plan to include more information such as the text and image information associated with the ticket and employ several other classification and regression methods.

ACKNOWLEDGMENTS

ExperTeam is the trademark of Uzman Bilişim A.Ş. This work is supported in part by ITEA 3 Call 7 20003 OMD project, TÜBİTAK 1509 grant number 9210017 and TÜBİTAK 2244 grant number 119c056.

REFERENCES

[1] I. Muhammad and Z. Yan," Supervised Machine Learning Approaches: A Survey*,"* Ictact Journal on Soft Computing, vol. 5(3), 2015.

[2] B. Kitchenham, "Procedures for performing systematic reviews," Keele, UK, Keele University, vol. 33*,* pp. 1-26, 2004.

[3] S. I. Davies, Machine learning at the operating room of the future: A comparison of machine learning techniques applied to operating room scheduling. Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science, Doctoral dissertation, 2004.

[4] N. Master, Z. Zhou, D. Miller, D. Scheinker, N. Bambos and P. Glynn, "Improving predictions of pediatric surgical durations with supervised learning," International Journal of Data Science and Analytics, vol. *4*, pp. 35-52, 2017.

[5] N. Hosseini, M. Y. Sir, C. J. Jankowski and K. S.Pasupathy, Surgical duration estimation via data mining and predictive modeling: a case study, In AMIA annual symposium proceedings, Vol. 2015, p. 640, American Medical Informatics Association, 2015.

[6] M. Fairley, D. Scheinker and M. L. Brandeau," Improving the efficiency of the operating room environment with an optimization and machine learning model," Health care management science, vol. *22*, pp. 756-767, 2019.

[7] J. Bender, and J. Ovtcharova," Prototyping Machine-Learning-Supported Lead Time Prediction Using AutoML," vol. 180, pp. 649-655, 2021.

[8] A. Alenezi, S. A. Moses, and T. B. Trafalis, "Real-time prediction of order flowtimes using support vector regression. Computers & Operations Research," vol. 35, pp. 3489-3503, 2018.

[9] F. J. de Cos Juez, P. G. Nieto, J. M. Torres and J. T. Castro, "Analysis of lead times of metallic components in the aerospace industry through a supported vector machine model," Mathematical and computer modelling, vol. 52, pp. 1177-1184, 2010.

[10] D. Gyulai, A. Pfeiffer, G. Nick, V. Gallina, W. Sihn, and L. Monostori, "Lead time prediction in a flow-shop environment with analytical and machine learning approaches," IFAC-PapersOnLine, vol. 51, pp. 1029-1034, 2018.

[11] A. N. Ahmed, A. Yafouz, A. H Birima, O. Kisi, Y. F. Huang, M. Sherif, ... and A. El-Shafie," Water level prediction using various machine learning algorithms: a case study of Durian Tunggal River, Malaysia," Engineering Applications of Computational Fluid Mechanics, vol. 16, pp. 422-440, 2022

[12] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning-based recommender system: A survey and new perspectives," ACM Computing Surveys (CSUR), vol. 52, pp. 1-38, 2019.

[13] P. Kadiri and S. Ravala," Kernel-Based Machine Learning Models to Predict Mitigation Time During Cloud Security Attacks," International Journal of e-Collaboration (IJeC), vol. 17, pp. 75-88, 2021.

[14] D. Liu, Q. Yang and F. Yang," Predicting Building Energy Consumption by Time Series Model Based on Machine Learning and Empirical Mode Decomposition," In 2020 5th IEEE International Conference on Big Data Analytics (ICBDA*),* IEEE, pp. 145-150, May 2020.

[15] R. S. Sutton, and A. G. Barto, "Reinforcement learning: An introduction," MIT press.

[16] P. Michel, K. Baumstarck, A. Loundou, B. Ghattas, P. Auquier and L. Boyer, "Computerized adaptive testing with decision regression trees: an alternative to item response theory for quality-of-life measurement in multiple sclerosis.," Patient preference and adherence, vol. 12, pp. 1043, 2018.

[17] G. D. Hutcheson, "Ordinary least-squares regression. *L. M*outinho and GD Hutcheson, The SAGE dictionary of quantitative management research," pp. 224-228, 2011.

[18] K. Lakshmi, B. Mahaboob, M. Rajaiah, and C. Narayana, "Ordinary least squares estimation of parameters of linear model," J. Math. Comput. Sci., vol. 11, pp. 2015-2030, 2021

[19] N. Shrestha," Detecting multicollinearity in regression analysis," American Journal of Applied Mathematics and Statistics, vol. *8*, pp. 39-42, 2022.

[20] S. G. Fashoto, E. Mbunge, G. Ogunleye, and J. V. den Burg, "Implementation of machine learning for predicting maize crop yields using multiple linear regression and backward elimination," Malaysian Journal of Computing (MJoC), vol. 6, pp.679-697, 2021.

[21] Z. He, L. Li, Z. Huang, and H. Situ, "Quantum-enhanced feature selection with forward selection and backward elimination," Quantum Information Processing, vol. *17*, pp. 1-11, 2018.

[22] V. Vapnik, "The nature of statistical learning theory," Springer science & business media.

[23] A. J. Smola, and B. Schölkopf, "A tutorial on support vector regression," Statistics and computing, vol. 14, pp. 199-222, 2004.

[24] S. Demirezen and M. Çetin, "Rassal Orman Regresyonu Ve Destek Vektör Regresyonu İle Piyasa Takas Fiyatinin Tahmini," Nicel Bilimler Dergisi, vol. 3, pp. 1-15, 2021.

[25] L. Breiman, "Random forests. Machine learning," vol. *45*, pp. 5-32, 2001.

[26] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone,"Classification and regression trees," Routledge, 2017.

[27] S. Uguz and O. Ipek," Prediction of the parameters affecting the performance of compact heat exchangers with an innovative design using machine learning techniques,"Journal of Intelligent Manufacturing, pp. 1-25, 2021.

[28] J. Souza, S. Matwin, and N. Japkowicz, "Evaluating data mining models: a pattern language," In Proceedings of the 9th Conference on Pattern Language of Programs, Illinois, USA, pp. 1-23, September 2002

[29] Ahmed, A. B. E. D., and I. S. Elaraby, "Data mining: A prediction for student's performance using classification method," World Journal of Computer Application and Technology, vol. 2, pp. 43-47, 2014.

[30] K. El-Basyouny and T. Sayed, "Comparison of two negative binomial regression techniques in developing accident prediction models," Transportation Research Record, vol. 1950, pp. 9-16, 2006.