Machine Learning to Examine the Foraging Periods of Bees

Nattakan Thonhual

March 1, 2024

Machine Learning to Examine The Foraging Periods of Bees

Nattakan Thonhual
Department of Information System
Rajamangala University of Technology Suvarnabhumi
Phra Nakhon Si Ayutthaya, Thailand
nattakan.t@rmutsb.ac.th

***Abstract*— This research project on Machine Learning to Examine the Foraging Periods of Bees aims to: 1) Study the number of honey bee foragers and data preprocessing for analysis, 2) Analyze the relationship between the number of honey bee foragers during different time intervals. The research process begins by defining periods from 5:00 a.m. to 4:00 p.m. for counting the number of honey bee foragers, with intervals of 30 minutes. Data preprocessing techniques are applied, and a suitable data schema is designed for data science purposes. In the subsequent steps, the Polynomial Regression algorithm is employed as part of the Machine Learning process to create a model that fits well with data exhibiting polynomial relationships. The relationship between the independent and dependent variables is considered in the form of a polynomial equation with a maximum degree of positivity. This research sets the polynomial degree to 7, yielding results with significant correlation values. The model is capable of efficiently examining the data's variability daily. This is evident from the R-squared and adjusted R-squared values, which approach 1. The insights derived from the analysis are valuable and can be integrated into the next research phase concerning bees.**

Keywords*— *Data Preprocessing, Machine Learning, Polynomial Regression

I. INTRODUCTION

Honey is a product with nutritional content and other advantages, making it a favorite among those who are health-conscious. It can be found in a variety of meals and drinks as well as herbal medicines, cosmetics [1], and other products. Honey, which is fragrant and delicious and is made from the nectar of numerous natural flowers, has unique properties. Sunflowers, Litchi flowers, Longan flowers, wildflowers, and Coral vines are examples of natural flowers that yield sweet nectar and honey.

At the researcher's house, a cluster of lovely Coral vine flowers bloomed in July. They displayed a magnificent fusion of light pink and dark pink hues that daily drew honey bee foragers to collect nectar from their blossoms. The bees consistently followed the same behavioral patterns, such as beginning their work in the morning continuing into the evening, and circling back and forth between the flowers in the same cluster. This observation demonstrated the population of worker bees' diligence and productivity.

These discoveries inspired the research effort to examine the association between the number of honey bee foragers flying to collect nectar at various time intervals [2]. The goal of this investigation is to develop a prediction model for estimating the number of honey bee foragers over a range of periods [3] using AI technology [4], more especially Machine Learning. The information gathered from this study will be useful in future investigations into topics including flower pollination, bee population conservation, and the effects of weather on bee foraging behavior [5].

II. METHODS AND FOUNDATIONS

A. Research Objectives

1) Study the number of honey bee foragers and data preprocessing for analysis.

2) Analyze the relationship between the number of honey bee foragers during different time intervals.

B. Artificial Intelligence and Machine Learning

Artificial Intelligence (AI) technology has been developed to connect with computer systems, enabling them to perform tasks that resemble human brain functions. It integrates various branches of science [4], such as Natural Language Processing, Big Data, and Graph Theory, to enhance and elevate the quality of life by providing convenience in different aspects. A component of Artificial Intelligence (AI) technology, Machine Learning imitates the capacity for learning of the human brain. It features several learning models [6].

1) Supervised Learning: A dataset containing known right answers is used to train a model. Estimating housing values based on geography, for instance.

2) Unsupervised Learning: A learning algorithm that doesn't require right answers and instead learns by seeing hidden patterns or structures in data. Putting similar customers together for marketing efforts is one example.

3) Reinforcement Learning: A model that learns by responding to a changing environment over time. It receives rewards or penalties based on its actions. For example, teaching a computer system to play games, where the model receives scores for correct moves and feedback for incorrect ones [7].

C. Regression Analysis

Regression analysis is the analysis of the relationship between data using mathematical equations. That consists of the dependent variable (y) and one or more independent variables (x). The data that is analyzed will be statistical values that are used to create a model for predicting values using the principles of data science. The primary goal of regression analysis is to estimate the coefficients (a and b's) that minimize the sum of the squared differences between the observed values of the dependent variable and the values predicted by the model. This process helps to create a model that provides the best fit for the data. Regression analysis is a supervised learning model in machine learning with various algorithms for analysis, such as linear regression, ridge regression, polynomial regression, etc. [8].

The regression analysis has an equation of the form:

$$y = a + bx + \epsilon \qquad (1)$$

y is the dependent variable to be predicted

x is the independent variable used for predicting y

a is the y-intercept

b is the slope

ϵ is the error term

D. Linear Regression

The relationship between a dependent variable and one or more independent variables can be studied using linear regression. It is a linear data distribution. through building data-predictive models. For data analysis and predictive modeling, it is one of the most straightforward and widely applied methods in machine learning and statistics [9].

The linear regression has an equation of the form:

$$y = \beta_0 + \beta_1 x + \epsilon \qquad (2)$$

y is the dependent variable to be predicted

x is the independent variable used for predicting y

β_0 is the y-intercept

β_1 is the slope

ϵ is the error term

E. Polynomial Regression

Polynomial regression is a method for creating a linear model for assessing complex interactions between independent variables and dependent variables by utilizing polynomials to modify the model. It is used to analyze the relationship between non-linear data. When high-degree polynomials are used in polynomial regression, the model may become more complex, while polynomial regression aids in better capturing complex interactions between independent and dependent variables. Thus, to prevent both overfitting and underfitting in data analysis, careful model validation and changes are necessary [8].

The polynomial regression has an equation of the form:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n + \epsilon \qquad (3)$$

y is the dependent variable to be predicted

x is the independent variable used for predicting y

β_0 is the y-intercept

n is the degree of the polynomial

F. Data Preprocessing

In the field of Data Science and Data Analytics, data preparation is an essential step to make sure that data is well-prepared for analysis and decision-making at both the business and data analytics levels. This entails a number of procedures, including [10]:

1) Understanding the Data: This entails analyzing and comprehending the current dataset, as well as looking at the data structure to comprehend various aspects of the data, such as data kinds, distribution, and links between various columns.

2) Data Cleaning: The process of verifying, filling up data gaps, and improving data accuracy. This involves handling duplicate data and missing values, as well as data format transformations such translating textual data into numerical values [11].

3) Feature Selection: This step entails choosing significant columns or variables that have an effect on the analysis. To do this, decisions are made to select only the columns that are most helpful in solving issues or responding to prepared queries.

4) Data Splitting: This is the process of splitting a dataset into sections for the purposes of training and testing models as well as assessing the performance of the model.

G. Data Collection

Data collection is the process of gathering information from various sources for analysis, processing, or storage for decision-making. Data can be classified into two types:

1) Quantitative data: information that has been compiled and is displayed quantitatively to represent amounts or measures. It can be categorized into two distinct types of information: continuous data and discrete data.

2) Qualitative data: describes traits, viewpoints, positions, or any other characteristics that cannot be quantified numerically; frequently takes the form of spoken words, tales, descriptions, pictures, or other symbols.

In this research, the data used for analysis is quantitative. The researchers collected data directly from sources, which is within the researcher's home in Phra Nakhon Si Ayutthaya province, Thailand, on July 2023, from 5:00 a.m. to 4:00 p.m. The estimation of the number of honey bee foragers was done by counting three times and taking the average, resulting in primary data [12].

TABLE I. Estimating the Foraging Population of honey bee foragers

Time	**Primary Data on July 1st, 2023**			
	Count 1	***Count 2***	***Count 3***	***Average values***
5:00 a.m.	2	2	2	2
5:30 a.m.	5	5	5	5
6:00 a.m.	12	12	12	12
6:30 a.m.	25	23	24	24
7:00 a.m.	38	39	39	39
7:30 a.m.	45	44	44	44
8:00 a.m.	51	53	50	51
8:30 a.m.	52	51	53	52
9:00 a.m.	53	53	54	53
9:30 a.m.	54	53	52	53
10:00 a.m.	53	55	55	54
10:30 a.m.	55	54	54	54
11:00 a.m.	55	56	54	55
11:30 a.m.	55	55	55	55
12:00 p.m.	53	52	52	52
12:30 p.m.	46	47	47	47
1:00 p.m.	41	40	41	41
1:30 p.m.	32	33	31	32
2:00 p.m.	20	20	21	20
2:30 p.m.	19	20	19	19
3:00 p.m.	3	3	3	3
3:30 p.m.	1	1	1	1
4:00 p.m.	0	0	0	0

The example of counting bees in each time interval by performing three counts and computing the average to two decimal places is shown by the data in Table I. The decimal portion is rounded up if above 0.50 and rounded down otherwise.

III. Methodology

The process involves a sequential set of steps according to the research objectives, as follows:

1) Study the number of honey bee foragers and data preprocessing for analysis.

Thailand's Phra Nakhon Si Ayutthaya province is the area covered by the data collecting for the honey bee foragers population. In July 2023, information will be gathered over a space measuring one meter in width by one meter in length by one meter in depth. Approximately 15 flowering clusters will be present, is shown in Fig. 1.

Fig. 1. The location of storage

The actions involved in creating a dataset are as follows: As a starting point, schedule worker bee collection for the hours of 5:00 a.m. to 4:00 p.m., with one-hour breaks. The second step is to count the number of bees. Save the data in the third step. Define the data's schema in step four. Data cleansing, step five Create data files in the sixth step using the model, is shown in Fig. 2.

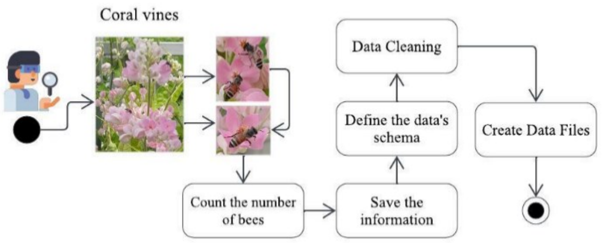

Fig. 2. Model for generating data sets

TABLE II. Number of Honey Bee Foragers (Data set 1)

Time	**July 2023**					
	1st	***2nd***	***4th***	***8th***	***9th***	***15th***
5:00 a.m.	2	3	2	4	4	9
5:30 a.m.	5	8	7	9	11	18
6:00 a.m.	12	14	15	18	22	28
6:30 a.m.	24	22	21	24	30	38
7:00 a.m.	39	41	42	45	34	52
7:30 a.m.	44	45	45	51	48	58
8:00 a.m.	51	52	54	55	52	65
8:30 a.m.	52	52	55	56	54	72
9:00 a.m.	53	54	56	58	59	75
9:30 a.m.	53	55	57	60	62	81
10:00 a.m.	54	55	57	60	64	86
10:30 a.m.	54	56	57	64	71	94
11:00 a.m.	55	57	58	65	72	94
11:30 a.m.	55	54	52	60	55	89
12:00 p.m.	52	52	51	56	47	82
12:30 p.m.	47	45	48	43	43	66
1:00 p.m.	41	40	43	38	42	61
1:30 p.m.	32	18	28	35	40	39
2:00 p.m.	20	24	21	25	14	28
2:30 p.m.	19	15	10	14	8	7
3:00 p.m.	3	5	4	4	4	8
3:30 p.m.	1	3	4	3	4	5
4:00 p.m.	0	0	0	0	0	0

It is evident from the information in Table II that different time periods and numbers of bees were included in the data utilized for analysis. On the 1st, 2nd, 4th, 8th, 9th, and 15th days in July 2023, the data collection starts at 5:00 a.m. and ends at 4:00 p.m.

TABLE III. Number of Honey Bee Foragers (Data set 2)

Time	July 2023					
	16th	*19th*	*22nd*	*23rd*	*28th*	*29th*
5:00 a.m.	10	11	15	16	12	14
5:30 a.m.	24	20	45	28	22	25
6:00 a.m.	33	34	76	53	35	42
6:30 a.m.	41	42	80	59	47	45
7:00 a.m.	55	54	89	64	56	52
7:30 a.m.	76	71	104	81	72	78
8:00 a.m.	87	94	120	95	85	89
8:30 a.m.	92	106	121	110	96	115
9:00 a.m.	106	120	125	115	120	122
9:30 a.m.	115	136	130	124	124	122
10:00 a.m.	120	139	132	125	124	125
10:30 a.m.	123	133	135	126	125	128
11:00 a.m.	125	134	135	127	126	128
11:30 a.m.	114	126	125	120	118	125
12:00 p.m.	98	105	110	112	94	84
12:30 p.m.	77	70	75	86	55	78
1:00 p.m.	65	64	68	45	36	56
1:30 p.m.	41	45	40	30	22	24
2:00 p.m.	25	26	21	15	8	12
2:30 p.m.	12	19	16	9	7	7
3:00 p.m.	16	15	11	6	3	5
3:30 p.m.	11	10	6	2	0	0
4:00 p.m.	0	0	0	0	0	0

It is evident from the information in Table III that different time periods and numbers of bees were included in the data utilized for analysis. On the 16th, 19th, 22nd, 23rd, 28th, and 29th days in July 2023, the data collection starts at 5:00 a.m. and ends at 4:00 p.m.

2) Analyze the relationship between the number of honey bee foragers during different time intervals.

- Algorithm comparison: linear regression vs polynomial regression is shown in Fig. 3 and TABLE IV.

The data used for analysis consists of:

Dependent variable (y) represented by the Number of Honey Bee Foragers.

Independent variable (x) represented by Time.

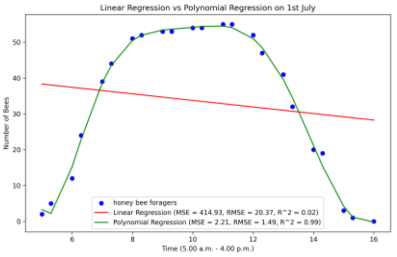

Fig. 3. polynomial regression vs linear regression

TABLE IV. Performance comparison of polynomial regression vs linear regression

Algorithm	MSE	RMSE	R-squared
Linear Regression	414.93	20.37	0.02
Polynomial Regression	2.21	1.49	0.99

Creating a linear regression model for the dataset on July 1st, 2023.

linear_reg.fit(X, y) is used to train a Linear Regression model by providing time data through the variable X and the corresponding number of bees as the target response through the variable y. This process aims to teach the model to learn the linear relationship between X and y by adjusting its parameters.

linear_reg.predict(X) predicts the outcome (y_linear_pred) using the Linear Regression model that has been trained with time data represented by the variable X.

Creating a polynomial regression model for the dataset on July 1st, 2023.

poly_reg.fit_transform(X) transforms the time data represented by variable X into a polynomial format using the **fit_transform** method of PolynomialFeatures.

LinearRegression() Creates a Linear Regression model.

poly_reg_model.fit(X_poly, y) Uses the data of variables (X_poly) and target responses (y) to train a Linear Regression model that utilizes polynomial features.

poly_reg_model.predict(X_poly) Predicts the outcome (y_poly_pred) using the Linear Regression model that has been trained with polynomial features.

As can be seen from Figure 3 and Table IV, the model's predictions of the data indicate that the polynomial regression model is superior to the linear regression model in terms of data analysis and prediction. This is demonstrated by looking at the model's performance using R-squared, RMSE, and MSE, all pointing to a higher prediction accuracy [8].

- Selecting the dataset from the date of 1st and substituting the values into the equation to find the degree of polynomial in representing the relationship between the data is shown in Fig. 4 – 9.

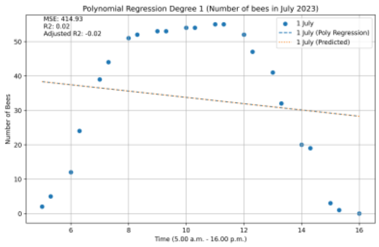

Fig. 4. Polynomial Regression degree=1

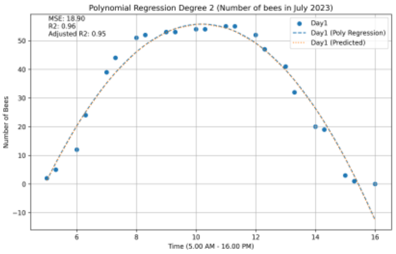

Fig. 5. Polynomial Regression degree=2

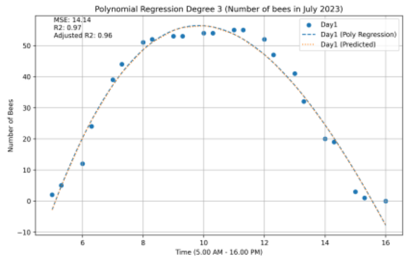

Fig. 6. Polynomial Regression degree=3

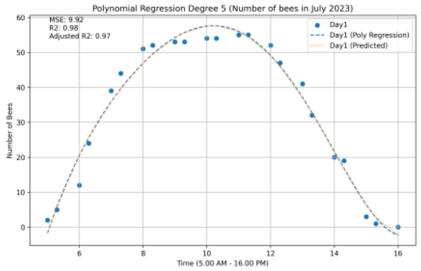

Fig. 7. Polynomial Regression degree=5

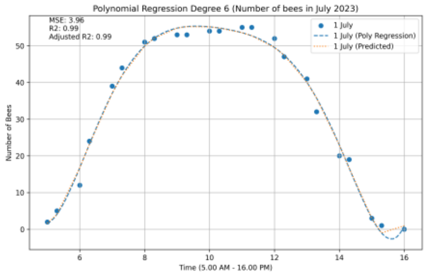

Fig. 8. Polynomial Regression degree=6

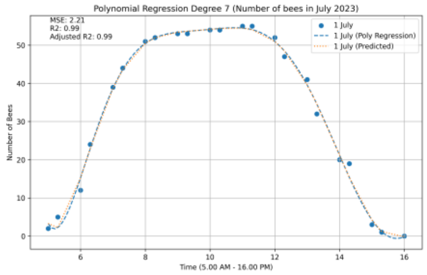

Fig. 9. Polynomial Regression degree=7

TABLE V. A Comparison of Degree Polynomials

Degree	**MSE**	R-squared	Adjusted R-squared
1	414.93	0.02	-0.02
2	18.90	0.96	0.95
3	14.14	0.97	0.96
4	11.32	0.97	0.97
5	9.92	0.98	0.97
6	3.36	0.99	0.99
7	2.21	0.99	0.99

The results of A Comparison of Degree Polynomials in Table V reveal that setting the degree of the polynomial equal to 7 yields efficient and accurate predictions of the data. Therefore, a degree of 7 is selected for predicting the number of honey bee foragers on other days in the dataset.

In the Polynomial Regression equation, the data values are substituted into the equation as follows:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n + \epsilon \qquad (3)$$

From the data on July 1st, 2023

Time: [5.00, 5.30, 6.00, 6.30, 7.00, 7.30, 8.00, 8.30, 9.00, 9.30, 10.00, 10.30, 11.00, 11.30, 12.00, 12.30, 13.00, 13.30, 14.00, 14.30, 15.00, 15.30, 16.00]

Day1: [2, 5, 12, 24, 39, 44, 51, 52, 53, 53, 54, 54, 55, 55, 52, 47, 41, 32, 20, 19, 3, 1, 0]

y = Number of Honey Bee Foragers
x = Time
$n = 7$
$\beta_0, \beta_1, \dots \beta_n$ = coefficients

$$y = 0.00 - 0.06x - 2.03x^2 + 35.91x^3 - 370.86x^4 +$$

$$+370.86x^4 + 2226.15x^5 + 7157.20x^6 + 9477.98x^7 + \epsilon$$

IV. EXPERIMENTAL RESULTS

The following experimental findings were attained via Methodology:

1) Results of the Study on the number of honey bee foragers and data preprocessing for analysis.

Tables II and III present the dataset collected for the analysis of data.

From the data table of honey bee foragers, a Data Schema has been designed to enable the analysis of relationships between the data. The schema includes the following data types:

- Date: Object data type
- Time: Integer data type
- Number of honey bee foragers: Integer data type

With this data preprocessing, the data can be effectively analyzed to understand the relationships between the variables, considering the dates, times, and the corresponding number of honey bee foragers.

2) Results of Analyzing the relationship between the number of honey bee foragers during different time intervals.

The results from setting the degree of the polynomial to 7 enable us to obtain a graph visualization that displays statistical values indicating a strong correlation between the number of honey bee foragers and the time intervals, is shown in Fig. 10 and the accompanying Table VI.

V. CONCLUSION AND DISCUSSION

The following results were reached from the study effort on using Machine Learning to Examine. The Foraging Periods of Bees:

1) The study's findings on the quantity of honey bee foragers and data pretreatment for analysis showed that there is a correlation between the quantity of bees and the passage of time. Every day, the most bees are present between the hours of between 9:00 and 11:30 a.m. The least amount

of bees are present between 3:30 to 4:00 p.m., which is also the time period, is shown in Fig. 10.

2) The Polynomial Regression Algorithm was used to analyze the data after a model was constructed using data splitting for training and testing [13]. The results showed a significant correlation within the data. The R-squared number is astonishingly high, reaching 0.99, and the Adjusted R-squared value is also excellent. The data for Day 1 showed the best stability, with an MSE of 2.21, according to the Mean Squared Error (MSE), which calculates the typical difference between actual and anticipated values. The data on July 1st, 2023, on the other hand, revealed the least stability, with an MSE of 36.98, is shown in Table VI.

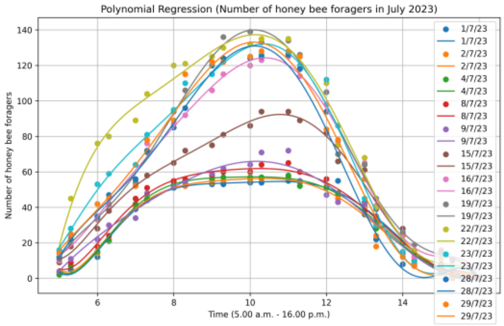

Fig. 10. Polynomial Regression in July 2023

TABLE VI. summary of significant correlations

Day	MSE	R-squared	Adjusted R-squared
July 1st, 2023	2.21	0.99	0.99
July 2nd, 2023	11.27	0.97	0.96
July 4st, 2023	5.23	0.99	0.98
July 8st, 2023	7.02	0.99	0.98
July 9st, 2023	18.73	0.97	0.95
July 15st, 2023	13.41	0.99	0.98
July 16st, 2023	16.83	0.99	0.99
July 19st, 2023	22.25	0.99	0.99
July 22nd, 2023	29.14	0.99	0.98
July 23rd, 2023	26.48	0.99	0.98
July 28st, 2023	27.58	0.99	0.98
July 29st, 2023	36.98	0.98	0.98

The conclusion from Table VI reveals that the model can effectively analyze the data variability for each day. Considering the R-squared and adjusted R-squared values approaching 1, the accuracy of the model is found to be good. Examining the Mean Squared Error (MSE) for predicting accuracy, it is observed that the lowest MSE occurs on the 1st day, indicating the highest prediction accuracy. However, on other days, the MSE values are higher than those in the training dataset. Specifically, on the 29th day, the MSE is relatively higher compared to other days. This elevated MSE may signal overfitting or poor prediction performance on unseen test data. Cross-validation methods may be necessary to assess the model's performance for the given data [14].

ACKNOWLEDGMENT

This research was supported by the Faculty of Business Administration and Information Technology, Rajamangala University of Technology Suvarnabhumi.

REFERENCES

[1] I. Kowalczuk, J. Gębski, D. Stangierska, and A. Szymańska, “Determinants of Honey and Other Bee Products Use for Culinary, Cosmetic, and Medical Purposes,” *Nutrients*, vol. 15, no. 3, pp. 1–17, 2023, doi: 10.3390/nu15030737.

[2] M. D. Rivera, M. Donaldson-Matasci, and A. Dornhaus, “Quitting time: When do honey bee foragers decide to stop foraging on natural resources?,” *Front. Ecol. Evol.*, vol. 3, no. MAY, 2015, doi: 10.3389/fevo.2015.00050.

[3] A. S. Singh and M. C. Takhellambam, “A Method to Study Honey Bee Foraging Regulatory Molecules at Different Times During Foraging,” *Front. Insect Sci.*, vol. 1, no. September, pp. 1–9, 2021, doi: 10.3389/finsc.2021.723297.

[4] C. Collins, D. Dennehy, K. Conboy, and P. Mikalef, “Artificial intelligence in information systems research: A systematic literature review and research agenda,” *Int. J. Inf. Manage.*, vol. 60, no. November 2020, p. 102383, 2021, doi: 10.1016/j.ijinfomgt.2021.102383.

[5] A. S. Alqarni, J. Iqbal, H. S. Raweh, A. M. A. Hassan, and A. A. Owayss, “Beekeeping in the desert: Foraging activities of honey bee during major honeyflow in a hot-arid ecosystem,” *Appl. Sci.*, vol. 11, no. 20, 2021, doi: 10.3390/app11209756.

[6] M. Batta, “Machine Learning Algorithms - A Review,” *Int. J. Sci. Res.*, vol. 18, no. 8, pp. 381–386, 2018, doi: 10.21275/ART20203995.

[7] F. Torres-Cruz *et al.*, “Comparative Analysis of High-Performance Computing Systems and Machine Learning in Enhancing Cyber Infrastructure: A Multiple Regression Analysis Approach,” *Proc. 2nd Int. Conf. Innov. Pract. Technol. Manag. ICIPTM 2022*, vol. 2, pp. 69–73, 2022, doi: 10.1109/ICIPTM54933.2022.9753839.

[8] E. Ostertagová, “Modelling using polynomial regression,” *Procedia Eng.*, vol. 48, pp. 500–506, 2012, doi: 10.1016/j.proeng.2012.09.545.

[9] S. D. Permai and H. Tanty, “Linear regression model using bayesian approach for energy performance of residential building,” *Procedia Comput. Sci.*, vol. 135, pp. 671–677, 2018, doi: 10.1016/j.procs.2018.08.219.

[10] S. B. Kotsiantis and D. Kanellopoulos, “Data preprocessing for supervised leaning,” *Int. J. Comput. Sci.*, vol. 1, no. 2, pp. 1–7, 2006, doi: 10.1080/02331931003692557.

[11] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang, “Data cleaning: Overview and emerging challenges,” *Proc. ACM SIGMOD Int. Conf. Manag. Data*, vol. 26-June-20, pp. 2201–2206, 2016, doi: 10.1145/2882903.2912574.

[12] H. Taherdoost, “Data Collection Methods and Tools for Research; A Step-by-Step Guide to Choose Data Collection Technique for Academic and Business Research Projects,” *Int. J. Acad. Res. Manag.*, vol. 10, no. 1, pp. 10–38, 2021, [Online]. Available: www.elvedit.com

[13] A. Vultureanu-Albisi and C. Badica, “The Model of Regularization Coefficient in Polynomial Regression for Modelling the Spread of COVID-19 in Romania,” *2022 23rd Int. Carpathian Control Conf. ICCC 2022*, pp. 94–100, 2022, doi: 10.1109/ICCC54292.2022.9805938.

[14] Aviral Gupta, Akshay Sharma, and Dr. Amita Goel, “Review of Regression Analysis Models,” *Int. J. Eng. Res.*, vol. V6, no. 08, pp. 58–61, 2017, doi: 10.17577/ijertv6is080060.