ISSN No:-2456-2165

Implementation of Machine Learning in Analyzing the Effect of Maintenance on the Reliability of Railway Detection Equipment

Fajar Sodik[1], Kusrini[2], Kusnawi[3]
Master of Informatics Engineering
AMIKOM University Yogyakarta
Jl. North Ringroad Yogyakarta, 55283, Indonesia

Abstract:- Maintenance is something that must be done on equipment to maintain its reliability. It is necessary to determine the correct maintenance period to make it more effective and efficient so that reliability is maintained while being efficient in terms of costs incurred. This research aims to determine the best algorithm between polynomial regression and Nadaraya Watson kernel regression to determine the maintenance period for train detection equipment and determine the variables that influence the determination of the maintenance period, which has an impact on equipment reliability. Testing the polynomial regression model produces a mean absolute error of 8.05, a mean squared error of 568.74, and a determination coefficient of 0.999, while the Nadaraya Watson regression model produces a mean absolute error of 3.14, a mean squared error of 19.43, and a determination coefficient of 0.938. Thus, it can be concluded that the Nadaraya Watson Kernel Regression model can be used well to determine the maintenance period for train detection equipment.

Keywords:- *Polynomial Regression; Nadaraya Watson Kernel Regression; Train Detector.*

I. INTRODUCTION

Trains are a mode of mass transportation that has been operating since the Dutch occupation of the country[1]. Until now, trains have been the favorite of users as a means of transportation between cities and provinces and for commuting between regional areas. The large number of people who choose trains as a mode of transportation increases the urgency of ensuring the security, safety, and punctuality of train use. In this case, the train operational facility equipment must be suitable for operation in order to guarantee safe, secure, and timely train travel. However, based on railway news in Indonesia in 2019, there was a signal problem in Karawang, which caused train travel to be late. So equipment is needed that is ready for operation to support the smooth running of trains[2].

Maintenance of equipment is something that must be done periodically with the aim of making the equipment fit for use. In the maintenance manual for equipment, the maintenance period that must be carried out is explained. For Signaling, Telecommunications, and Electrical Equipment in PT Kereta Api Indonesia (Persero), there are regulations regarding monthly maintenance guidelines for train detection equipment.

Inspection and maintenance of signaling, telecommunications, and overhead electrical equipment is carried out to maintain the condition of signaling, telecommunications, and overhead electrical equipment that can function properly and is safe for continuous operation in accordance with the technical requirements of the equipment[3]. However, sometimes, for some reason, the realization of maintenance is earlier or perhaps later than the maintenance schedule. On the other hand, the maintenance period that has been determined does not completely guarantee that the tool will not be disturbed, depending on the value of each parameter.

Much research has been carried out on equipment maintenance, and many use machine learning methods in the hope of increasing the reliability of the equipment. The following are four literature reviews of previous research to serve as a reference for this development, namely :

- One of the machine learning methods used in previous research is Random Forest. In 2020, Nurul Hakim conducted research on data on broken/loose and non-broken/loose rail incidents in the DIVRE III and DIVRE 1V train operation areas in 2017–2019 using 4 algorithms, namely Support Vector Machine (with kernel='rbf'), Decision Tree, Random Forest (with n_estimators = 103 and random_state = 515), and K-Nearest Neighbor (with n_neighbors = 19) [4]. The results of this research state that random forest modeling is the best machine learning method, with the smallest RMSE value of 0.47885622798803357. However, there is still a weakness, namely that the data variations are still incomplete, so it needs to be developed in further research.
- In 2021, Via Ardianto Nugroho, Derry Pramono Adi, Achmad Teguh Wibowo, MY Teguh Sulistyono, and Agustinus Bimo Gumelar, will also conduct research to classify types of maintenance for container cranes using machine learning algorithms [5]. This research aims to determine the performance of mathematical-statistical machine learning genres and tree-based models in classifying the best type of maintenance for container cranes. The models used in machine learning in this

research are Random Forest, SVM, K-Nearest Neighbor, Naïve Bayes, Logistic Regression, J48, and Decision Tree. The results show that J48 shows the best performance with accuracy and ROC-AUC values reaching 99.1%.

- Apart from that, in 2021, Alif Kurnia Utama will also conduct research related to the use of machine learning in river water level monitoring systems [6]. The algorithms used are linear regression, polynomial regression, and neurofuzzy, with results stating that the measurement error produced by neurofuzzy is 3.76%. From these error values, this study concludes that neurofuzzy is the best method for measuring water level.
- Research on the use of machine learning with a linear regression model was carried out by Puteri and Silvanie in 2020 to predict basic food prices with an R-squared result of 0.842, or 84.2%. This percentage value is influenced by independent variables, namely date, commodity, and market, while the remaining 15.8% is influenced by variables not included in the research model [7].

In this research, polynomial regression and Nadaraya-Watson kernel regression models will be used in the hope of obtaining better mean absolute error, mean squared error, and coefficient of determination values so that the best maintenance period can be predicted.

II. RESEARCH METHODS

To conduct research, it is necessary to plan the steps that must be taken so that the research can run smoothly and knowledge findings are obtained.

Fig 1 Flow of Research Methods

In Figure 1, there are 4 steps that must be taken to overcome the problem. The first step taken is to analyze the problem with the aim of finding out the problems of the research. The second step is problem solving, namely by determining the algorithm to be used, collecting the required data, in this case routine maintenance history data and disturbance data, which will later be stored in a database, integrating the data in the database using CSV format, and processing the data into modeling. Polynomial Regression and Nadaraya Watson Regression, as well as testing the modeling results on the mean absolute error, mean squared error, and coefficient of determination parameters The third step, discussion of the results, is analyzing the results of the testing process for each model. The final conclusion is to summarize the modeling results using two algorithms, namely polynomial regression and Nadaraya Watson kernel regression, to determine the best modeling and determine which variables influence the reliability of train detection equipment.

III. RESULTS AND DISCUSSION

➢ *Problem Analysis*

From the background of the problems above, researchers analyzed several problems, including :

- How to predict the best maintenance period to maintain the reliability of train detection equipment at PT Kereta Api Indonesia (Persero)?
- Which is the best model between the Nadaraya-Watson Kernel Regression and Polynomial Regression models in predicting the maintenance period for train detection equipment?

➢ *Solution to Problem*

From the problems that have been explained, it is necessary to predict the best maintenance period by implementing machine learning in analyzing the effect of maintenance on the reliability of train detection equipment. In this research, researchers tried to implement the Nadaraya-Watson Kernel Regression and Polynomial Regression models using Python as an analysis tool. Next, the implementation results will be tested on the Mean Absolute Error, Mean Squared Error, and Coefficient of Determination values.

Regression analysis is a data analysis method that shows the relationship between response variables and one or several predictor variables[8]. For example, if X is the predictor variable and Y is the response variable for n paired observations {x,y}, then the linear relationship between the predictor variable and the response variable can be expressed as follows:

$$Yi = m(X_i) + \varepsilon_i \quad (1)$$

Where ε_i is the residual, which is assumed to be independent with a zero mean and variance σ_i, and m(Xi) is the regression function or regression curve.

After the regression modeling is carried out, testing is then carried out on the regression model used so that we can find out whether the regression model is good or not[9].

The tests carried out were tests on the mean absolute error, mean squared error, and coefficient of determination values. The Mean Squared Error (MSE) measures the discrepancy between the expected value of the model and the observed data value. This discrepancy is then squared to eliminate any negative differences. The average of all the data samples is then calculated by adding the squared differences[10]. One way to assess a forecasting model's accuracy is to use the Mean Absolute Error (MAE). The average absolute error between the forecasted and predicted value and the actual value is displayed by the MAE value[11]. A straightforward metric called coefficient of determination, or R-square, is frequently used to evaluate the accuracy of a regression line equation. An overview of the independent variable's suitability for predicting the dependent variable can be obtained from the R-Square value.

ISSN No:-2456-2165

To be able to make predictions, researchers collected data on the maintenance history and history of train detection equipment disturbances. These two data are processed into a dataset of the best maintenance days before a disruption occurs. Table 1 is the best treatment days dataset, with a total of 55 data points.

The variables used in making predictions are device age, DC VUR Box voltage (VOD1+, VOD1-)(VDC), H(VH) voltage, L(VL) voltage, temperature, and best maintenance day (HPT) with the data type shown in figure 2.

It can be seen that the data type is numerical, so it is sufficient for modeling without data conversion.

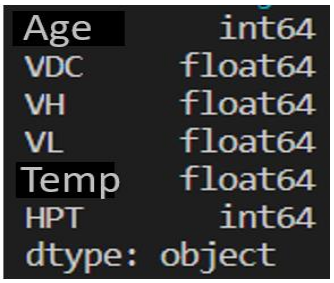

Fig 2 Data Type of Each Variable

Table 1 The Best Maintenance Day Database

Device age	DC VUR Box voltage (VOD1+, VOD1-) (VDC)	H(VH) voltage	L(VL) voltage	Temperature	Best Maintenance Day (HPT)
1715	95,70	5,05	5,05	25,2	68
1745	96,00	4	4	26,4	14
1775	96,00	4	4	26,5	14
1805	96,00	5	4	28	14
…	…	…	…	…	…
3275	92,80	5,07	5,07	28,3	48
3305	94,50	5,07	5,07	25,4	60
3335	93,00	5,07	5,07	25,6	50
3395	95,00	5,07	5,07	27,2	63

Before modeling, correlation testing of the independent variable with the dependent variable is carried out to determine the relationship between the variables. In Figure 3, it can be seen that the variables that influence determining the best maintenance day are DC Power Supply Voltage VOD,1+ VOD1- (0.62), H Voltage (0.75), and L Voltage (0.82).

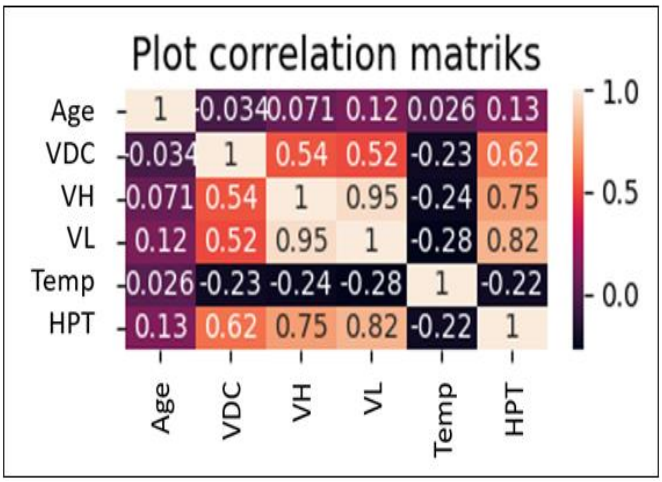

Fig 3 Correlation Matrix between Variables

In figures 5 to 9, it can be seen that the graph of changes in age, DC voltage, H voltage, L voltage, and temperature on the best maintenance day has a pattern. The data distribution is shown in red dots. This pattern can be formulated in the Nadaraya Watson Kernel Regression and Polynomial Regression algorithms.Modeling using polynomial regression is shown on the black line, while modeling using Nadaraya Watson kernel regression is shown on the cyan line.

Descriptive statistical analysis can be shown in Figure 4. It can be seen that the number of samples is 55. The average age of the equipment is 2543 days with a VDC voltage of 93.650182 Vdc, an H voltage of 4.918364 and an L voltage of 4.905636 with a temperature of 27 .032727oC found that the best maintenance days were 55 days. The highest DC voltage was 98 Vdc and the lowest was 5.652009 Vdc, while the best maintenance days were a maximum of 74 days and the lowest was 5 days.

Next, the dataset is separated into training data and testing data with a training: testing data ratio of 80%: 20%, or around 45 data points for training and 10 data points for testing. The training data and testing data were implemented in modeling using the Nadaraya Watson Kernel Regression and Polynomial Regression algorithms.

Fig 4 Descriptive Statistical Analysis

ISSN No:-2456-2165

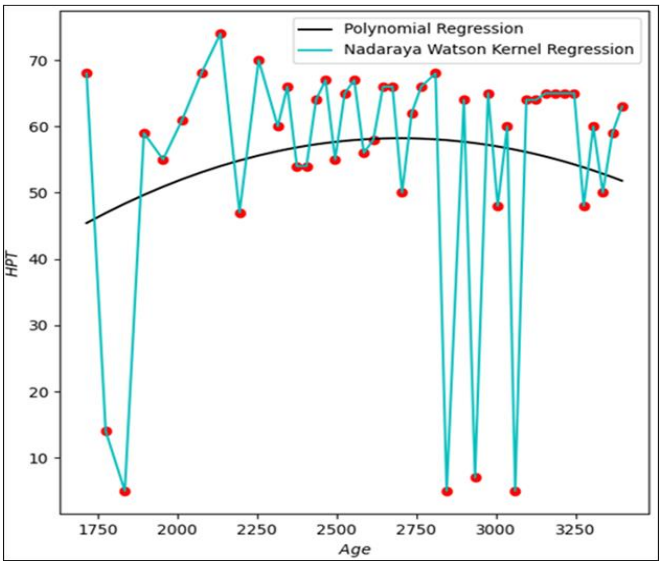

Fig 5 Correlation of Equipment Age to Best Maintenance Day (HPT)

Figure 5 is a correlation graph between the equipment age variable and the best maintenance days. It can be seen that the polynomial regression modeling is in the form of a curved line, but the Nadaraya Watson kernel regression modeling is in the form of a line that passes through each training set.

In figure 6, the correlation between the DC voltage variable and the best treatment day is depicted. It can be seen that the polynomial regression and Nadaraya Watson kernel regression modeling have almost parallel line patterns.

Then in Figure 7, there is the correlation between the voltage variable H and the best treatment day. The line patterns of polynomial regression and Nadaraya Watson kernel regression are opposite, as shown in Figure 8. Figure 7 shows the correlation between the stress variable L and the best treatment day, with the line pattern between polynomial regression and Nadaraya kernel regression modeling being opposite.

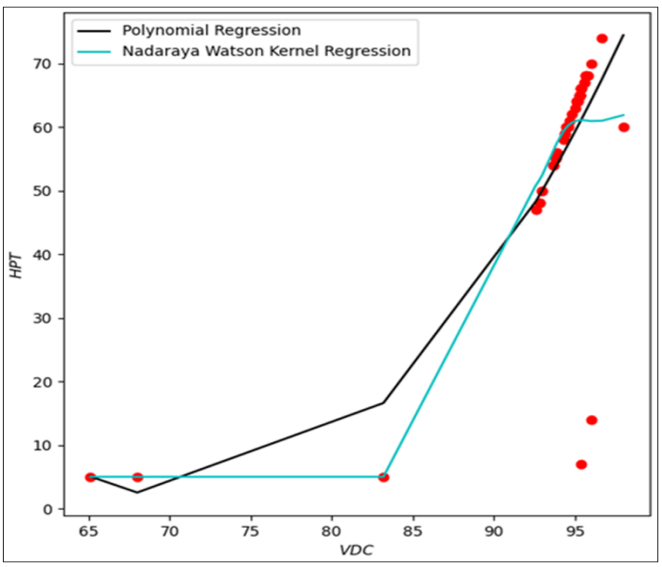

Fig 6 Correlation of DC Voltage (VDC) to best maintenance day (HPT)

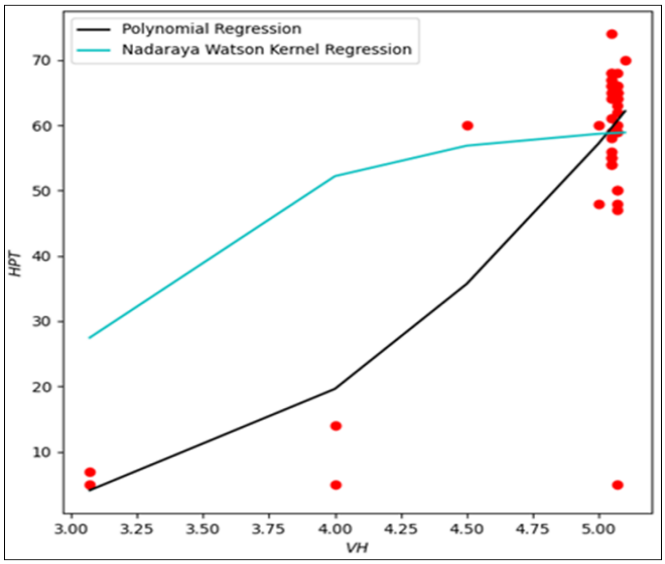

Fig 7 Correlation of H Voltage (VH) To Best Maintenance Day (HPT)

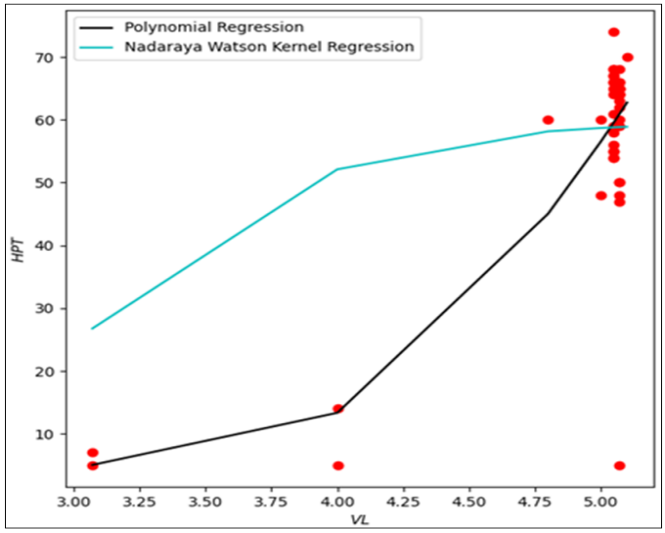

Fig 8 Correlation of L Voltage (VL) to Best Maintenance Day (HPT)

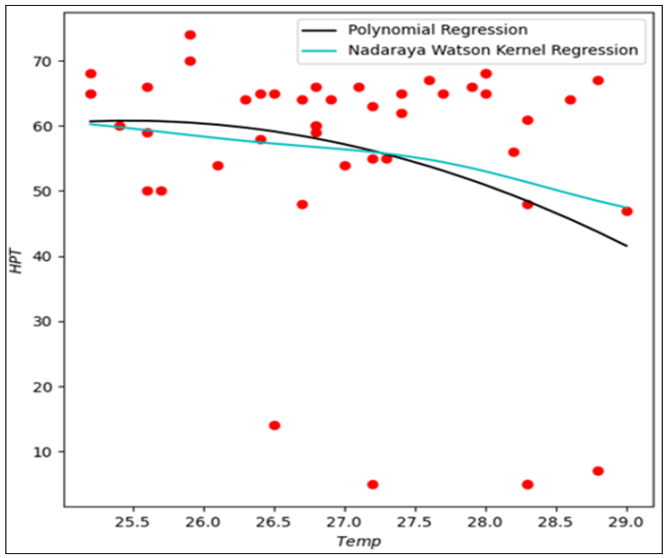

Fig 9 Correlation of Temperature to Best Maintenance Day (HPT)

ISSN No:-2456-2165

Figure 9 is a correlation graph between temperature variables and the best treatment day. The modeling line pattern between polynomial regression and Nadaraya Watson kernel regression is almost parallel.

- *Polynomial Regression*

Polynomial Regression is a method used to see the form of relationship between an independent variable and has a high degree polynomial relationship with the dependent variable with a quadratic function, where the value of the independent variable has a value of power 1, power 2, power n and so on as in the equation (2)[12].

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \quad (2)$$

In this case, y is the best maintenance day, α is a constant number and x is the independent variable.

Next, to determine predictions, modeling is carried out using Python commands from the AstroML library. A Python module called astroML was created inside the Scipy environment for Python and is intended to serve as a source for quick and reliable code for astronomical data analysis[13]. When using modeling, it is necessary to determine in advance the number of degrees that will be used. This number of degrees will determine the modeling ability to make predictions, as shown in the parameters mean absolute error, mean squared error, and coefficient of determination.

	degree	mse_train	rmse_train	r2_train	mse_test	rmse_test	r2_test
0	2	0.104104	0.322652	0.999606	1.814227e+03	42.593743	-2.243616
1	3	0.016033	0.126620	0.999939	2.886926e+06	1699.095786	-5160.472047
2	4	0.002821	0.053114	0.999989	4.324786e+07	6576.310422	-77320.890581
3	5	0.000543	0.023307	0.999998	5.079308e+07	7126.926327	-90810.822794

Fig 10 Results of Polynomial Regression Modeling using Python

Figure 10 is the result of polynomial regression modeling experiments using Python for degrees 2 to 5. If you look at the fine-tuning results in Figure 10, the degree that will be used is degree 2 because a higher degree causes overfitting in the testing set.

Table 2 shows the prediction results using the Polynomial Regression model. From the prediction results, it is necessary to test to ensure that the modeling has a good Mean Absolute Error (equation (3)), Mean Squared Error (equation (4)), and Determination Coefficient (equation (5)).

Table 2 Polynomial Regression Prediction Results

Age	VDC	VH	VL	Temp	HPT Act	HPT Pred
1745	96	4,00	4,00	26,40	14,00	12,32846
1805	96	5,00	4,00	28,00	14,00	89,37861
1865	94,12	5,05	5,05	25,40	57,00	57,34528
1925	95,7	5,05	5,05	26,80	68,00	67,61671
1985	94,6	5,05	5,05	26,30	60,00	60,24727
2045	93,78	5,05	5,05	28,00	55,00	54,95193
2105	95,64	5,05	5,05	27,10	67,00	67,19115
2165	92,6	5,07	5,07	25,80	47,00	48,43485
2225	95,7	5,10	5,10	28,70	68,00	67,58486
2285	96	5,10	5,10	26,90	70,00	70,34392

$$MAE = \sum \frac{|Y' - Y|}{n} \quad (3)$$

$$MSE = \sum \frac{|Y' - Y|^2}{n} \quad (4)$$

Y' = Predicted value
Y = Actual value
n = Number of data

$$R\ square = \frac{SSR}{SST} \quad (5)$$

SSR = square of the difference between the predicted Y value and the average Y value
SST = square of the difference between the actual Y value and the average Y value

By using this formula, test results were obtained between the actual value and the predicted value. Table 3 shows the test results using polynomial regression modeling.

Table 3 Results of Testing the Polynomial Regression Model

No	Parameter	Value
1	MAE	8,045913532
2	MSE	568,744478123832
3	R Square	0,999360600740741

ISSN No:-2456-2165

- *Nadaraya-Watson Kernel Regression*

Nadaraya-Watson Kernel regression was introduced in 1964 by Nadaraya and Watson [14]. This estimator is used to estimate m as a locally weighted average, using the kernel as the weighting function.

Next, to determine predictions, modeling is carried out using Python commands from the astroML library. When using modeling, it is necessary to first determine bandwidth value that will be used. Bandwidth value will determine the modeling's ability to make predictions as shown in the parameters Mean Absolute Error, Mean Squared Error, and Coefficient of Determination. Determination of bandwidth is carried out by prediction experiments with different bandwidth value. Bandwidth value with predicted results that is close to the actual value will be used as the best bandwidth value.

$$m\left(X^{(1)},X^{(2)}\right) = -\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{\phi X_i}K_{h_1}(x^{(1)} - X_i^{(1)})K_{h_2}(x^{(2)} - X_i^{(2)}) \quad (6)$$

Where :

K = kernel functions
h = bandwidth value
X_i = observed value of the predictor variable to - i
y_i = observed value of the response variable to - i
x = random value of variable X or a certain value of variable

Table 4 shows the prediction results using the Nadaraya Watson Kernel Regression model. From the prediction results, it is necessary to test to ensure that the modeling has a good mean absolute error (equation 3), mean squared error (equation 4), and determination coefficient (equation 5). By using this formula, test results were obtained between the actual value and the predicted value. Table 5 shows the test results using Nadaraya Watson kernel regression modeling.

Table 4 Nadaraya Watson Kernel Regression Prediction Results

Age	VDC	VH	VL	Temp	HPT Act	HPT Pred
1745	96	4,00	4,00	26,40	14,00	16,59958
1805	96	5,00	4,00	28,00	14,00	14
1865	94,12	5,05	5,05	25,40	57,00	59
1925	95,7	5,05	5,05	26,80	68,00	57,44211
1985	94,6	5,05	5,05	26,30	60,00	55,68739
2045	93,78	5,05	5,05	28,00	55,00	61,48825
2105	95,64	5,05	5,05	27,10	67,00	69,20183
2165	92,6	5,07	5,07	25,80	47,00	47,13786
2225	95,7	5,10	5,10	28,70	68,00	65,52092
2285	96	5,10	5,10	26,90	70,00	69,41143

Table 1 Nadaraya Watson Kernel Regression Model Testing Results

No	Parameter	Value
1	MAE	3,136566785
2	MSE	19,4282244600571
3	R Square	0,938361042654321

➢ *Discussion of the Results*

From the results of testing the polynomial regression and Nadaraya Watson kernel regression models, a comparison of the mean absolute error, mean squared error, and determination coefficient values is obtained in table 6.

Table 6 Comparison of Test Results of the Polynomial Regression Model with Nadaraya Watson Kernel Regression

No	Parameter	Polynomial regression value	Nadaraya Watson kernel regression value
1	MAE	8,045913532	3,136566785
2	MSE	568,744478123832	19,4282244600571
3	R Square	0,999360600740741	0,938361042654321

Table 6 is a comparison table of the results of testing the polynomial regression model with Nadaraya Watson kernel regression. From this Table, it can be seen that the MAE value in the polynomial regression model is greater than the Nadaraya Watson kernel regression model, and the MSE value in the polynomial regression model is much greater than the Nadaraya Watson kernel regression model. The smaller the MAE and MSE values, the better the quality of the model[15]. The R-Square value is categorized as strong if it is more than 0.67, moderate if it is more than 0.33 but lower than 0.67, and weak if it is more than 0.19 but lower than 0.33[16]. The R square value in the polynomial regression model is greater than the Nadaraya Watson kernel regression model, but the MAE and MSE values are too large, so in this case, the best test results are in the Nadaraya Watson kernel regression model.

ISSN No:-2456-2165

IV. CONCLUSION

➢ *From the results of testing and calculations in the research above, the following conclusions are obtained :*

- The use of the Nadaraya Watson Kernel Regression algorithm has a better coefficient of determination than the polynomial regression algorithm in terms of predicting the best maintenance day, with a value of R squared = 0.938361042654321.
- The variables that really influence the best maintenance day are the DC voltage values on the Power Supply VOD1+ VOD1-, H and L voltages on the UP module. Meanwhile, the temperature and equipment age variables were not significant enough to predict the best maintenance day.
- By paying attention to these variables during maintenance to predict the next best maintenance day, it is hoped that it can increase the reliability of train detection equipment.

SUGGESTION

- In future research, it is recommended to use a dataset from maintenance history data spanning more than 5 years. The more datasets, the better the coefficient of determination value.
- In future research, it is recommended to use a dataset from maintenance history data spanning more than 5 years. The more datasets, the better the coefficient of determination value.
- Principal Component Regression (PCR) could be another alternative in further research.
- In future research, the efficiency of the costs incurred can be calculated based on the results of determining the maintenance period.

REFERENCES

[1]. "PT Kereta Api Indonesia." Accessed: Jul. 30, 2022. [Online]. Available: https://kompaspedia.kompas.id/baca/profil/lembaga/pt-kereta-api-indonesia

[2]. CNN Indonesia, "Gangguan Sinyal di Karawang, Perjalanan Kereta Api Terlambat." [Online]. Available: https://www.cnnindonesia.com/nasional/20191228040004-20-460590/gangguan-sinyal-di-karawang-perjalanan-kereta-api-terlambat

[3]. DIREKTUR PENGELOLAAN PRASARANA, PEDOMAN PEMERIKSAAN DAN PERAWATAN SIGNALLING, TELECOMMUNICATION AND ELECTRICITY. Bandung: PT KERETA API INDONESIA (PERSERO), 2018.

[4]. N. Nurul Hakim, "Implementasi Machine Learning pada Sistem Prediksi Kejadian dan Lokasi Patah Rel Kereta Api di Indonesia," 2020.

[5]. V. A. Nugroho, D. P. Adi, A. T. Wibowo, M. T. Sulistyono, and A. B. Gumelar, "Klasifikasi Jenis Pemeliharaan dan Perawatan Container Crane menggunakan Algoritma Machine Learning," Matics, vol. 13, no. 1, pp. 21–27, 2021, doi: 10.18860/mat.v13i1.11525.

[6]. Y. Alif Kurnia Utama, "Penggunaan Neuro Fuzzy Pada Sistem Monitoring Ketinggian Air Sungai," Jurnal Informatika Kaputama(JIK), vol. 5, no. 1, 2021.

[7]. K. Puteri and A. Silvanie, "Machine Learning Untuk Model Prediksi Harga Sembako Dengan Metode Regresi Linier Berganda," Jurnal Nasional Informatika, vol. 1, no. 2, pp. 82–94, 2020.

[8]. J. A. Saputra, "Pemilihan Bandwidth Pada Estimator Nadaraya Watson dengan Tipe Kernel Gaussian pada Data Time Series," 2016.

[9]. Amalia Yunia Rahmawati, "Landasan teori," no. July, pp. 1–23, 2020.

[10]. H. H. Nuha, "Mean Squared Error (MSE) dan Penggunaannya Ringkasan Penjelasan Referensi," vol. 52, pp. 2021–2022, 2021.

[11]. A. A. Suryanto, "Penerapan Metode Mean Absolute Error (Mea) Dalam Algoritma Regresi Linear Untuk Prediksi Produksi Padi," Saintekbu, vol. 11, no. 1, pp. 78–83, 2019, doi: 10.32764/saintekbu.v11i1.298.

[12]. Gabrielhozana, "Insurance forecast." [Online]. Available: https://github.com/gabrielhozana/insurance_forecast

[13]. J. Vanderplas, "Introduction." [Online]. Available: https://www.astroml.org/user_guide/introduction.html

[14]. F. Lamusu, T. Machmud, and R. Resmawan, "Estimator Nadaraya-Watson dengan Pendekatan Cross Validation dan Generalized Cross Validation untuk Mengestimasi Produksi Jagung," Indonesian Journal of Applied Statistics, vol. 3, no. 2, p. 85, 2021, doi: 10.13057/ijas.v3i2.42125.

[15]. Trivusi, "Perbedaan MAE, MSE, RMSE, dan MAPE pada Data Science." [Online]. Available: https://www.trivusi.web.id/2023/03/perbedaan-mae-mse-rmse-dan-mape.html

[16]. B. University, "MEMAHAMI KOEFISIEN DETERMINASI DALAM REGRESI LINEAR." [Online]. Available: https://accounting.binus.ac.id/2021/08/12/memahami-koefisien-determinasi-dalam-regresi-linear/#:~:text=Menurut Chin (1998)%2C nilai,lebih rendah dari 0%2C33.