IMPLEMENTATION OF DATA MINING PREDICTION DELIVERY TIME USING LINEAR REGRESSION ALGORITHM

Tri Wahyudi[1*], Dava Septya Arroufu[2]
Information System Department, STIKOM Cipta Karya Informatika, Jakarta, Indonesia[1,2]
triwahyudi100390@gmail.com[1], davaseptyaarroufu@gmail.com[2]

Received : 15 August 2022, Revised: 02 September 2022, Accepted : 02 September 2022
**Corresponding Author*

ABSTRACT
In the current era of modernization, online shopping has become a habit of the people, and is closely related to freight forwarding services in charge of delivering online shopping items from the seller to the buyer. So that buyers need a fast and safe delivery service to ensure the goods sent on time to their destination. Customer satisfaction is one of the most important factors in the shipping business. However, there are several obstacles that occur in the field that cause delays in the delivery of goods. Therefore, one solution that can be used to overcome this problem is to use data mining technology to predict delivery times. Using 1,000 datasets consisting of 4 Attributes, data processing will be carried out with prediction techniques using the Linear Regression algorithm. By utilizing data when the goods are taken, when the goods are on the way, until they reach the buyer, they can produce forecasts or predictions and produce several analyzes so that in the future there will be no delivery delays. Based on the RMSE (Root Mean Square Error) value which serves to generate the level value the error of the prediction results using this method and in an RMSE value of 0.370 %. It can be concluded that using the Linear Regression algorithm is proven to be accurate in predicting delivery times.

Keywords : *Prediction, Data Mining, Linear Regression, Delivery*

1. Introduction

The use of freight forwarding services has now become a need for every community so that it has increased. With online shopping to various personal and company shipments. Every consumer needs a fast and safe delivery service to ensure that the goods sent arrive on time and safely arrive at their destination(Laia et al., 2018; Öztürk & Başar 2022). There have been many shipping service companies that have emerged and then developed to compete in capturing the market in the modern era as it is today, with a very rapid level of economic and technological development, one of which is PT XYZ which is one of the companies engaged in the delivery of goods by having branch offices spread in every city in Indonesia(Pink & Djohan, 2021; Widiyanto, et al., 2021).

Due to the large number of goods that must be delivered, PT. XYZ has difficulty in achieving the specified travel time until the goods are received by the customer in order to provide customer satisfaction. Efficient delivery will have a positive impact on business development. Goods that reach customers faster will certainly provide space for the company to carry out quality control of the goods sent, so that the company can grow. So we need a Data Mining method that can complete and optimize performance with the aim of increasing customer satisfaction(Hertina, et al., 2021). Based on the discussion of these conditions, the author tries to offer a strategy to deal with delays in delivery of goods, using Rapid Miner application tools and Data Mining Prediction analysis with the concept of the Linear Regression Algorithm(Gilyén, et al., 2022; Artin, et al., 2021; Emioma & Edeki, 2021)..

Systematic Literature Review (SLR) is a systematic literature review method that functions to identify, and interpret findings on a research topic to answer the research question that has been determined(Kumar, et al., 2021; Ahmad & Alsmadi, 2021). This survey methodology is based on PICOC (Population, Intervention, Comparation, Outcomes, Context) as an identification of information needs from previous research sources in the following table(Camargo, et al., 2021):

Table 1 - PICOC review

Implementation Of Data Mining Prediction Delivery Time Using Linear Regression Algorithm	
Population	Linear Regression
Intervention	Predicting Freight Forwarding
Comparison	n/a
Outcomes	Predict Delivery of Goods on Time or Not
Context	Private

2. Research Methods

The algorithm used in this study is the Linear Regression Algorithm (Kohli, et al., 2021; Cosenza, et al., 2021). The shipping data will be processed to get an error value and can be used as a reference in predicting the delivery time of goods. The type of research that researchers do is qualitative research. Qualitative research is research that uses more analysis. The process that is more emphasized in this type of research is the theoretical basis used as a guide so that the research focus is in accordance with the facts on the ground. In addition, the theoretical basis also has a role to provide an overview of the research background and as a material for discussion of research results.

In this study, the process and meaning use more focus on research based on facts in the field. Data analysis in qualitative research is interpreted as an effort to systematically search and organize notes from observations, interviews, and literature studies to increase the researcher's understanding of the case under study and present it as findings.

In general, qualitative research obtains data from interviews and observations. Then the researcher will then analyze the data obtained in detail, and come up with a new theory or concept if the research results contradict the theories and concepts used. In this study, several stages of research will be carried out. In this study, several stages of research will be carried out as shown in Figure 1:

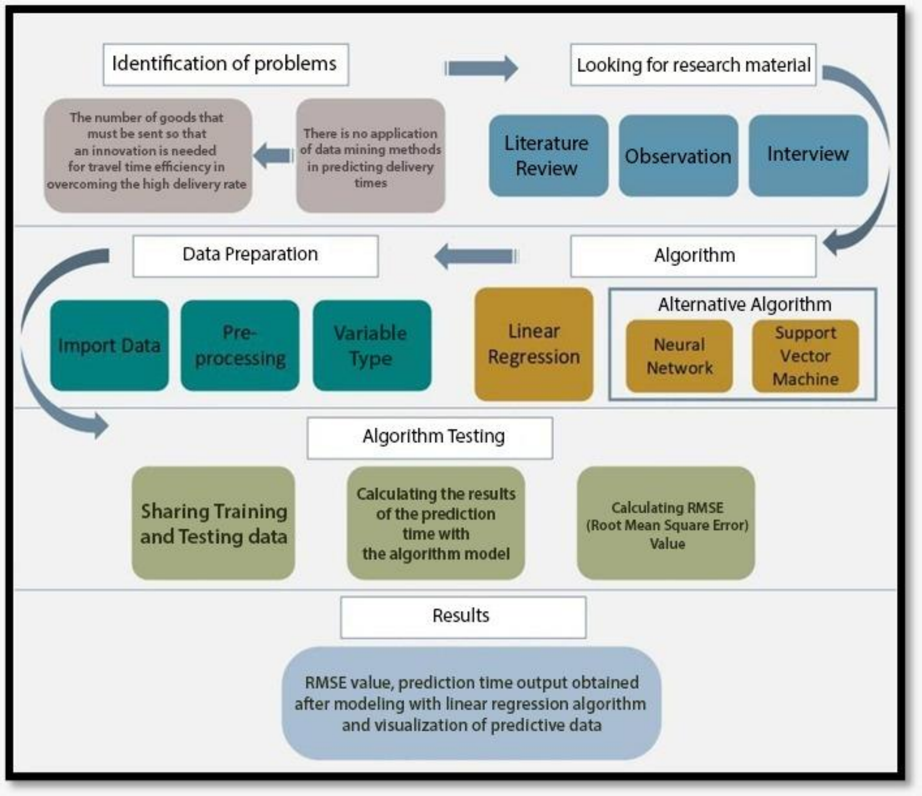

Identification of problems
Looking for research material
The number of goods that must be sent so that an innovation is needed for travel time efficiency in overcoming the high delivery rate
There is no application of data mining methods in predicting delivery times
Literature Review
Observation
Interview
Data Preparation
Algorithm
Import Data
Pre-processing
Variable Type
Linear Regression
Alternative Algorithm
Neural Network
Support Vector Machine
Algorithm Testing
Sharing Training and Testing data
Calculating the results of the prediction time with the algorithm model
Calculating RMSE (Root Mean Square Error) Value
Results
RMSE value, prediction time output obtained after modeling with linear regression algorithm and visualization of predictive data

Fig. 1. Stages of Methodology Application

Data collection is also supported by observations and interviews with related parties in the field as well as library data collection as a reference from previous research.

1) Observation

Observation is a data collection technique that is carried out by systematically recording and observing the PT. XYZ. With this field observation study, the researcher obtained material that is quite accurate and relevant and very helpful in writing this research.

2) Interview

Interview is a method or technique used to obtain data by conducting oral question and answer directly to the Staging Store Leader. The form of information obtained can be in the form of writing, or audio-visual recordings.

3) Literature review

Literature study or literature study is the first step in data collection activities. This can be done by reading scientific journals, recording and processing literature, books, and reports on previous research materials to obtain information and search for data on the internet. Below are some of the results of definitions of important terms obtained from literature studies and related to research:

a. Data Mining

Data Mining is a process to obtain useful information from a large database that needs to be extracted so that it becomes new information and can assist in decision making (Nofitri & Irawati, 2019)

b. Prediction

According to the Big Indonesian Dictionary, prediction is the result of predicting or estimating future values using past data. Prediction shows what will happen in a given situation and is an input for planning and decision-making processes. Understanding Prediction is the same as forecast(Bengnga & Ishak, 2018).

c. Linear Regression

Linear Regression is used to estimate or predict the relationship between two variables in qualitative research. an approach to establish the relationship between one or more dependent variables (simple linear regression) as well as the independent variables (multiple linear regression). Assuming that the relationship between these variables can be approximated by a straight-line equation, the model that approximates the relationship between variables in the data is referred to as linear regression stabilization(Hafizah et al., 2019).

d. Dataset

Dataset or Data Set is a collection of data that is used as the goal of some learning for a particular machine. There are 2 categories of datasets: public and private. (Jukes, 2018). Datasets have attributes that function as factors or parameters that cause classes, labels, and targets to occur. The current trend of data mining research is to test the methods developed by researchers with public datasets, so that research can be comparable, repeatable and verifiable.(Liantoni, 2016)

e. CRISP-DM

CRISP-DM (Cross Industry Standard Process for Data Mining) is a standardization of data mining processing that has been developed where the existing data will pass through each structured and defined phase clearly and efficiently(Hasanah et al., 2021).

f. Rapidminer

Rapidminer is a leading and well-known open-source based data mining application. It includes stand-alone applications for data analysis and as a data mining engine such as for data loading, data transformation, data modeling, and data visualization methods(Baihaqi et al., 2021).

3. Results and Discussions

Based on the stages of applying the methodology, the researchers implemented the Crisp DM design as a test model, and the Linear Regression Algorithm as a data mining method in this research. In this study, the test design carried out using the Cross-Industry Standard Process for Data Mining or CRISP-DM is one of the datamining process models (datamining framework)(Zhang, 2021; Saltz, 2021). The following is an image of the test scheme on CRISP-DM:

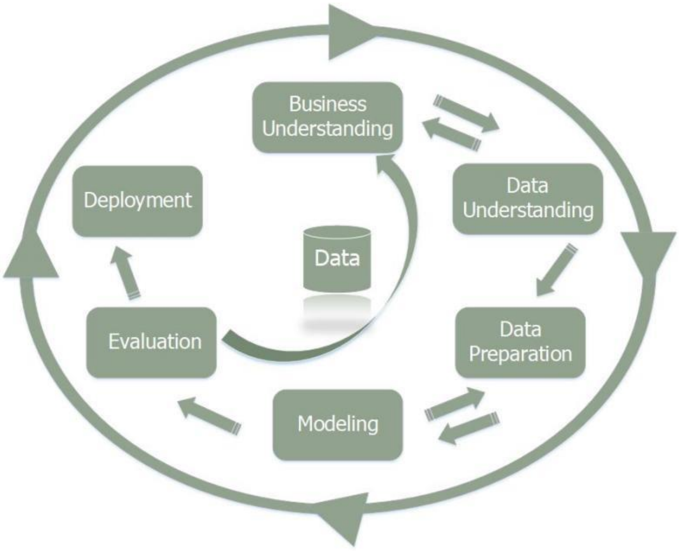


Fig 2. CRISP DM. Test Scheme

In this study using a test model to get good and maximum research results. Before doing data processing and testing the Linear Regression algorithm, the first step that must be done is to implement a data mining test model. The data processed by the CRISP-DM test model will go through several phases in it(Schröer, et al., 2021). There are 6 stages in CRISP-DM, which are described in detail as follows:

3.1 Stages of Business Understanding

Freight forwarding service both in the national and international scope are currently required to have advantages in being competitive, namely by improving the quality of their services. In order for good service to be achieved, it is influenced by the speed of delivery of goods on time and safely when it reaches the recipient. With so many packages, the delivery times are different, so some arrive on time and some don't. Delivery of goods that are not on time can be an important note for shipping companies.

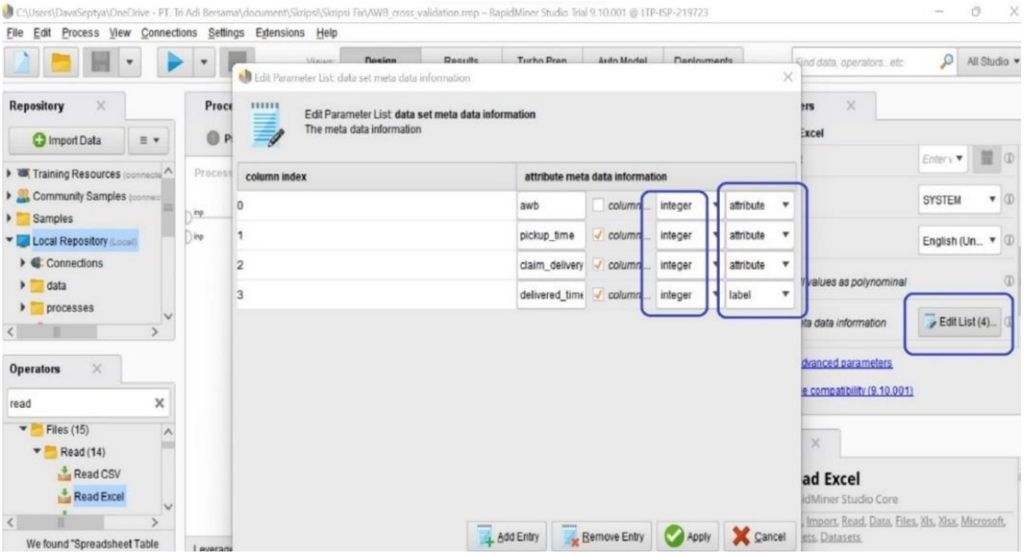
3.2 Stages of Data Understanding

All data that has been collected is 1000 data. The data used in this study is data on delivery of goods at PT.XYZ within a span of 1 day. This data collected from the existing information system in the Management Information System (MIS) division. Not all attributes of the delivery data are used, researchers only use 3 attributes consisting of Pickup Time, Claim Delivery Time, and Delivered Time. The AWB attribute is not used because this attribute is not needed in the research data to predict delivery time, so it only requires 3 attributes. Here is a breakdown of all the attributes:

1. *Pickup Time* or Goods Pickup Time is the time of goods pick-up service so that it uses the date time data type.
2. *Claims Delivery Time* or Time of Claiming Delivery is the time when the courier starts distributing the package of goods and has a date time data type.
3. *Delivered Time* or the time when the goods have arrived at their destination and are received by the recipient using the date time data type.

3.3 Stages of Data Preparation

In this stage, the data is prepared in advance by using pre-processing techniques, namely the raw data is converted into a form that is easier to understand. Using this technique the data is cleaned, the attributes that do not need to be used in the analysis process, such as the AWB attribute, will be removed. Here are the steps in doing pre-processing:



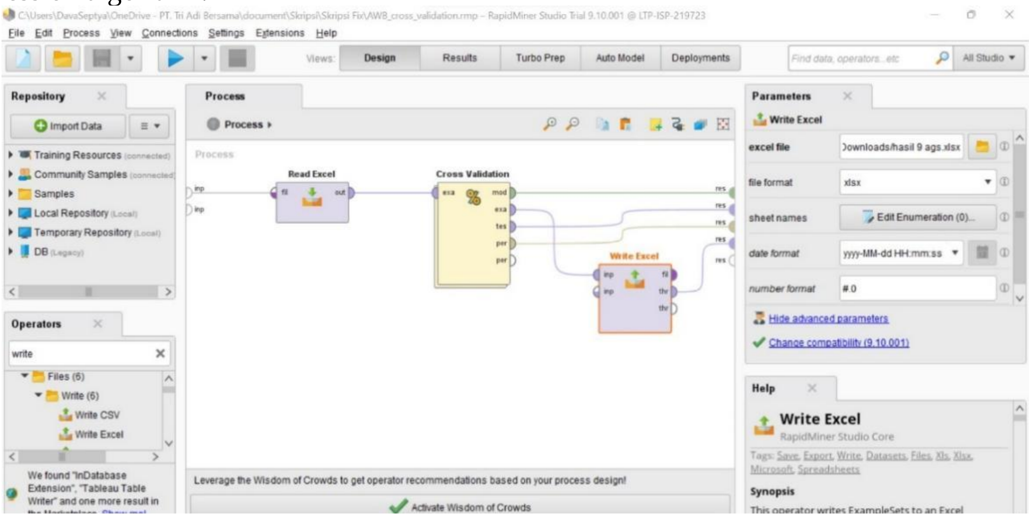
Edit Parameter List: data set meta data information
The meta data information

column index	attribute meta data information			
0	awb	□ column…	integer	attribute
1	pickup_time	☑ column…	integer	attribute
2	claim_delivery	☑ column…	integer	attribute
3	delivered_time	☑ column…	integer	label

Fig 3. Selecting Variable Attributes and Types

3.4 Stages of Modeling

At the modeling stage, are make a prediction model, which is to predict the delivery time of goods. At this stage, you can use read excel to import data sets, cross validation to share training and testing data, and excel writer to export the results of data processing in the cross validation process. The following is a process model that is formed on Rapid Miner using the Linear Regression algorithm:

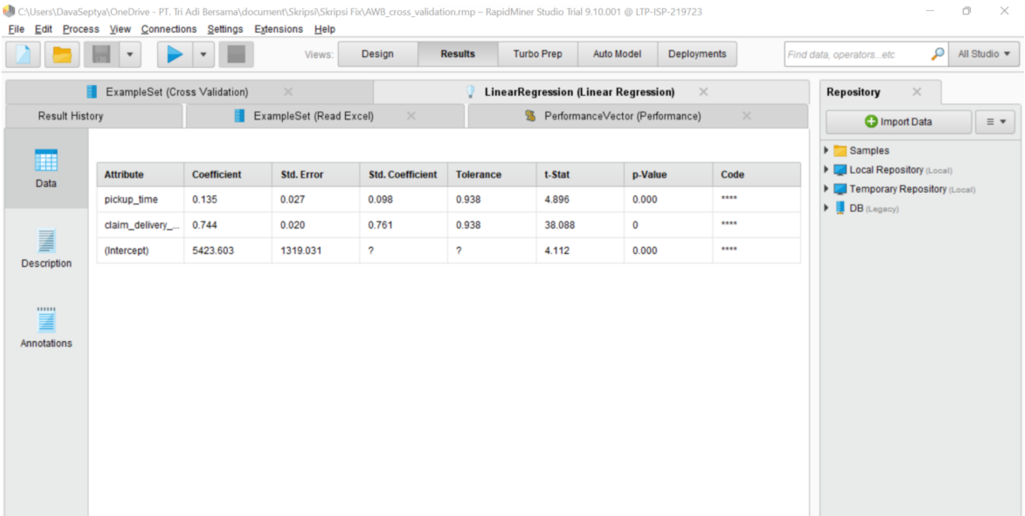


Parameters — Write Excel

excel file	Downloads/hasil 9 ags.xlsx
file format	xlsx
sheet names	Edit Enumeration (0)...
date format	yyyy-MM-dd HH:mm:ss
number format	#.0

Fig 4. Prediction Model Display

Linear regression will generally make predictions based on pre-existing values.

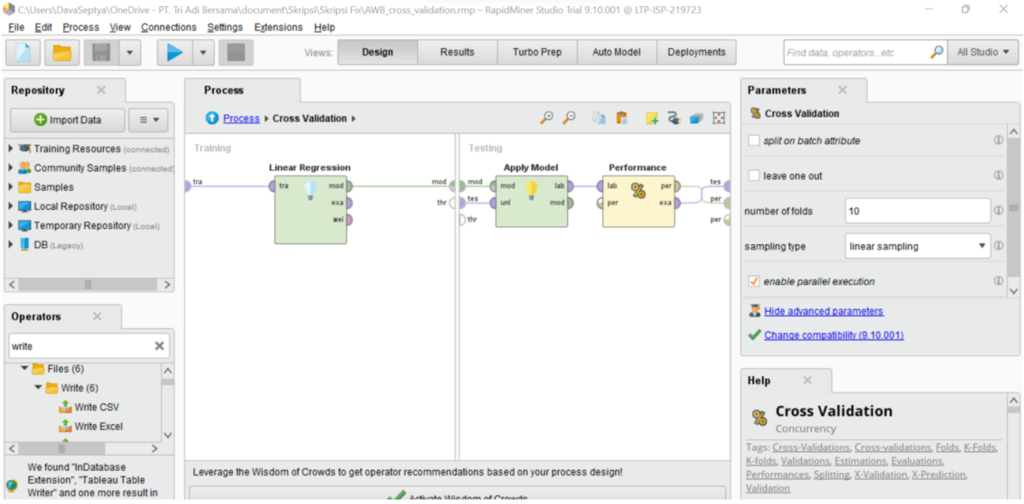

C:\Users\DavaSeptya\OneDrive - PT. Tri Adi Bersama\document\Skripsi\Skripsi Fix\AW8_cross_validation.rmp – RapidMiner Studio Trial 9.10.001 @ LTP-ISP-219723
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
ExampleSet (Cross Validation)
LinearRegression (Linear Regression)
Result History
ExampleSet (Read Excel)
PerformanceVector (Performance)
Data
Description
Annotations
Repository
Import Data
Samples
Local Repository (Local)
Temporary Repository (Local)
DB (Legacy)

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
pickup_time	0.135	0.027	0.098	0.938	4.896	0.000	****
claim_delivery_...	0.744	0.020	0.761	0.938	38.088	0	****
(Intercept)	5423.603	1319.031	?	?	4.112	0.000	****

Fig 5. Results of Linear Regression

3.5 Stages of Evaluation

After modeling, the next step is to evaluate the cross validation process by applying a linear regression algorithm, the way it works is mevaluate the performance of the algorithm where the data is separated into two subsets, namely the data of the learning process (training) and data validation (testing).The following are the results of the evaluation of the model that has been created with this algorithm.


C:\Users\DavaSeptya\OneDrive - PT. Tri Adi Bersama\document\Skripsi\Skripsi Fix\AW8_cross_validation.rmp – RapidMiner Studio Trial 9.10.001 @ LTP-ISP-219723
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Repository
Import Data
Training Resources (connected)
Community Samples (connected)
Samples
Local Repository (Local)
Temporary Repository (Local)
DB (Legacy)
Operators
write
Files (6)
Write (6)
Write CSV
Write Excel
We found "InDatabase Extension", "Tableau Table Writer" and one more result in
Process
Process ▸ Cross Validation ▸
Training
Linear Regression
tra
mod
exa
wei
Testing
Apply Model
mod
unl
lab
mod
Performance
lab
per
per
exa
thr
tes
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Hide advanced parameters
Change compatibility (9.10.001)
Help
Cross Validation
Concurrency
Tags: Cross-Validations, Cross-validations, Folds, K-Folds, K-folds, Validations, Estimations, Evaluations, Performances, Splitting, X-Validation, X-Prediction, Validation

Cross Validation

□ split on batch attribute

□ leave one out

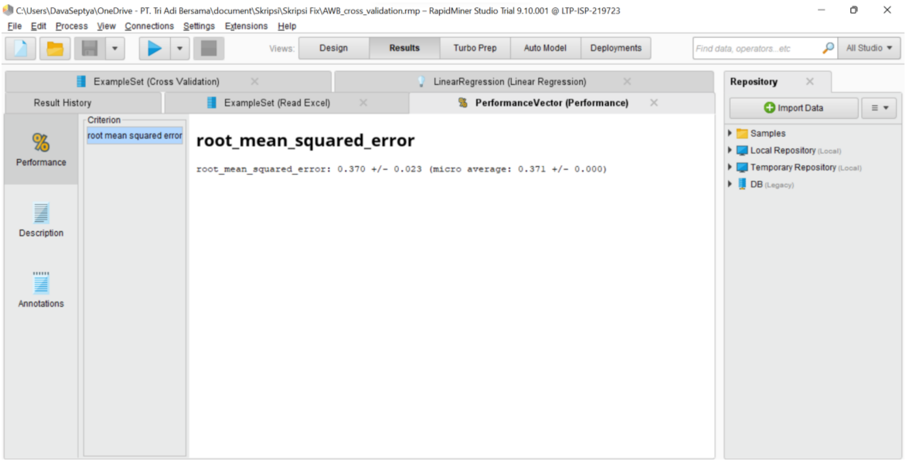
number of folds: 10

sampling type: linear sampling

☑ enable parallel execution

Fig 6. Stages of Cross Validation Evaluation

The performance operator is used to evaluate the performance of a model that provides a list of performance criteria values automatically according to a given task. From the vector in Figure 5, the error value obtained is 0.370. Performance measurement is done by calculating the average error that occurs through the amount of Root Mean Square Error (RMSE). The smaller the value of each of these performance parameters, the closer the predicted value to the actual value.


C:\Users\DavaSeptya\OneDrive - PT. Tri Adi Bersama\document\Skripsi\Skripsi Fix\AWB_cross_validation.rmp – RapidMiner Studio Trial 9.10.001 @ LTP-ISP-219723
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
ExampleSet (Cross Validation)
LinearRegression (Linear Regression)
Result History
ExampleSet (Read Excel)
PerformanceVector (Performance)
Performance
Description
Annotations
Repository
Import Data
Samples
Local Repository (Local)
Temporary Repository (Local)
DB (Legacy)

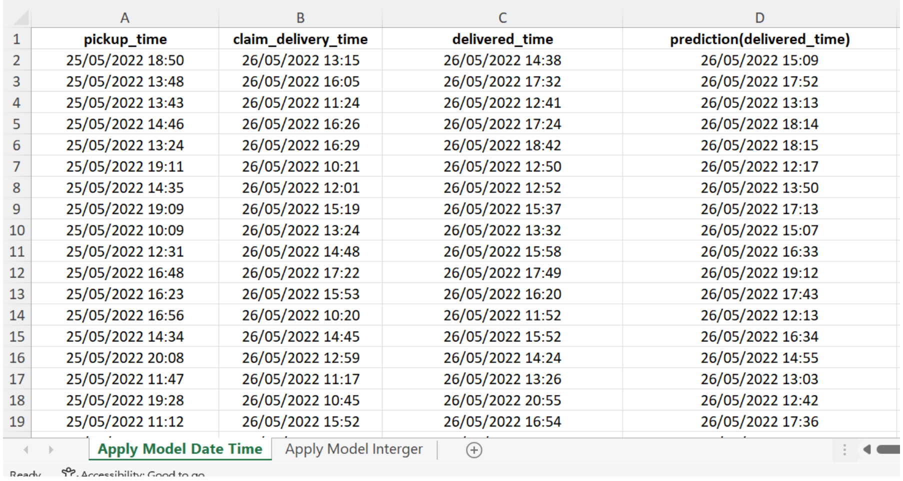
Criterion: root mean squared error

root_mean_squared_error

```
root_mean_squared_error: 0.370 +/- 0.023 (micro average: 0.371 +/- 0.000)
```

Fig 7. RMSE Nilai Value Results

The following is an export of the results of the predicted time that has been converted into excel form. In the data below, the results of prediction delivered time are not much different from the actual data results at delivered time.

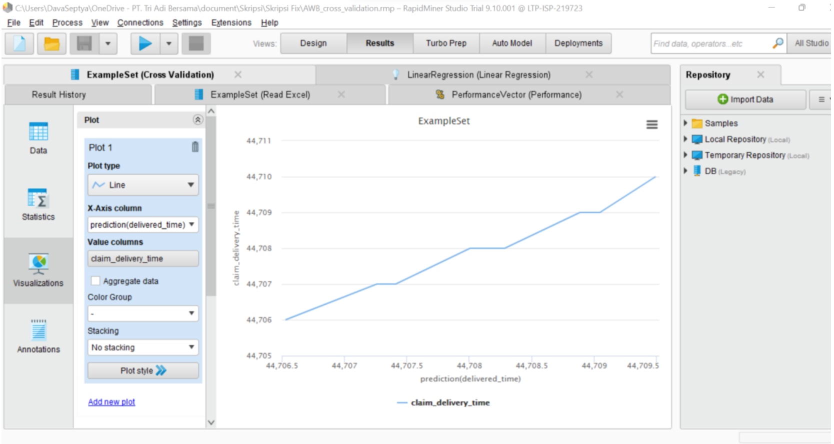


	A	B	C	D
1	pickup_time	claim_delivery_time	delivered_time	prediction(delivered_time)
2	25/05/2022 18:50	26/05/2022 13:15	26/05/2022 14:38	26/05/2022 15:09
3	25/05/2022 13:48	26/05/2022 16:05	26/05/2022 17:32	26/05/2022 17:52
4	25/05/2022 13:43	26/05/2022 11:24	26/05/2022 12:41	26/05/2022 13:13
5	25/05/2022 14:46	26/05/2022 16:26	26/05/2022 17:24	26/05/2022 18:14
6	25/05/2022 13:24	26/05/2022 16:29	26/05/2022 18:42	26/05/2022 18:15
7	25/05/2022 19:11	26/05/2022 10:21	26/05/2022 12:50	26/05/2022 12:17
8	25/05/2022 14:35	26/05/2022 12:01	26/05/2022 12:52	26/05/2022 13:50
9	25/05/2022 19:09	26/05/2022 15:19	26/05/2022 15:37	26/05/2022 17:13
10	25/05/2022 10:09	26/05/2022 13:24	26/05/2022 13:32	26/05/2022 15:07
11	25/05/2022 12:31	26/05/2022 14:48	26/05/2022 15:58	26/05/2022 16:33
12	25/05/2022 16:48	26/05/2022 17:22	26/05/2022 17:49	26/05/2022 19:12
13	25/05/2022 16:23	26/05/2022 15:53	26/05/2022 16:20	26/05/2022 17:43
14	25/05/2022 16:56	26/05/2022 10:20	26/05/2022 11:52	26/05/2022 12:13
15	25/05/2022 14:34	26/05/2022 14:45	26/05/2022 15:52	26/05/2022 16:34
16	25/05/2022 20:08	26/05/2022 12:59	26/05/2022 14:24	26/05/2022 14:55
17	25/05/2022 11:47	26/05/2022 11:17	26/05/2022 13:26	26/05/2022 13:03
18	25/05/2022 19:28	26/05/2022 10:45	26/05/2022 20:55	26/05/2022 12:42
19	25/05/2022 11:12	26/05/2022 15:52	26/05/2022 16:54	26/05/2022 17:36

Fig 8. Prediction Time Results

3.6 Deployment Stages

The Deployment stage is carried out after the evaluation stage. A detailed assessment of the results of a model is carried out with the implementation of the entire model that has been built. In addition, adjustments were made to the model so that it can produce a result that is in accordance with the initial target of this study. So that the delivery time prediction data based on the modeled test and testing data is easy to understand, it is necessary to visualize the data into a dashboard diagram or graphic display.



Plot 1
Plot type: Line
X-Axis column: prediction(delivered_time)
Value columns: claim_delivery_time
□ Aggregate data
Color Group: -
Stacking: No stacking

Fig 9. Display Data Visualization

4. Conclusion

Based on the results of research from the application of data mining to predict the delivery time of goods using the Linear Regression algorithm at PT XYZ, it can be concluded that: The processed data is data obtained in 1 day of delivery with a total of 1000. The processed data produces a coefficient value for pickup time of 0.135 and claim delivery time of 0.744. Factors that affect the prediction results are the attributes of pickup time and claim delivery time, by making delivered time as a label. The results of the summary model that is formed in the delivered time prediction process with the Linear Regression Algorithm produces an RMSE of 0.370, it can be concluded thatThe smaller the RMSE value of each of these performance parameters indicates the closer the predicted value to the actual value.

References

Ahmad, R., & Alsmadi, I. (2021). Machine learning approaches to IoT security: A systematic literature review. *Internet of Things*, 14, 100365.

Artin, J., Valizadeh, A., Ahmadi, M., Kumar, S. A., & Sharifi, A. (2021). Presentation of a novel method for prediction of traffic with climate condition based on ensemble learning of neural architecture search (NAS) and linear regression. *Complexity*, 2021.

Baihaqi, D. I., Handayani, A. N., Pujianto, U., Rahman, A. A., Kurniawan, Y. I., Sulaksono, J., Irawan, R. H., Fahmi, I. N., Iman, Q., Wahyu, A., Agustina, W., Furqon, M. T., Rahayudi, B., Tungadi, E., Thalib, I., Nur, M., Utomo, Y., Firasari, E., Khultsum, U., … Sosial, B. (2021). Klasterisasi Dana Bantuan Pada Program Keluarga Harapan (PKH) Menggunakan Metode K-Means. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *3*(1), 1231–1236. https://doi.org/10.37034/infeb.v3i2.66

Bengnga, A., & Ishak, R. (2018). Prediksi Jumlah Mahasiswa Registrasi Per Semester Menggunakan Linier Regresi Pada Universitas Ichsan Gorontalo. *ILKOM Jurnal Ilmiah, 10*(2), 136–143. https://doi.org/10.33096/ilkom.v10i2.274.136-143

Camargo, C., Gonçalves, J., Conde, M. Á., Rodríguez-Sedano, F. J., Costa, P., & García-Peñalvo, F. J. (2021). Systematic Literature Review of Realistic Simulators Applied in Educational Robotics Context. *Sensors*, *21*(12), 4031.

Cosenza, D. N., Korhonen, L., Maltamo, M., Packalen, P., Strunk, J. L., Næsset, E., ... & Tomé, M. (2021). Comparison of linear regression, k-nearest neighbour and random forest methods in airborne laser-scanning-based prediction of growing stock. *Forestry: An International Journal of Forest Research, 94*(2), 311-323.

Emioma, C. C., & Edeki, S. O. (2021). Stock price prediction using machine learning on least-squares linear regression basis. In *Journal of Physics: Conference Series* (Vol. 1734, No. 1, p. 012058). IOP Publishing.

Hafizah, Tugiono, & Maya, W. R. (2019). Penerapan Data Mining Dalam Memprediksi Jumlah Penumpang Pada CV . Surya Mandiri Sukses Dengan Menggunakan Metode Regresi Linier. *Jurnal Teknologi Informasi Dan Sistem Komputer TGD*, *2*(1), 54–61.

Hasanah, M. A., Soim, S., & Handayani, A. S. (2021). Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir. *Journal of Applied Informatics and Computing*, *5*(2), 103–108. https://doi.org/10.30871/jaic.v5i2.3200

Hertina, H., Nurwahid, M., Haswir, H., Sayuti, H., Darwis, A., Rahman, M., ... & Hamzah, M. L. (2021). Data mining applied about polygamy using sentiment analysis on Twitters in Indonesian perception. *Bulletin of Electrical Engineering and Informatics, 10*(4), 2231-2236.

Gilyén, A., Song, Z., & Tang, E. (2022). An improved quantum-inspired algorithm for linear regression. *Quantum, 6*, 754.

Kohli, S., Godwin, G. T., & Urolagin, S. (2021). Sales prediction using linear and KNN regression. In *Advances in machine learning and computational intelligence* (pp. 321-329).

Springer, Singapore.

Kumar, S., Kar, A. K., & Ilavarasan, P. V. (2021). Applications of text mining in services management: A systematic literature review. *International Journal of Information Management Data Insights, 1*(1), 100008.

Laia, D., Buulolo, E., & Sirait, M. J. F. (2018). Implementasi Data Mining Untuk Memprediksi Pemesanan Driver Go-Jek Online Dengan Menggunakan Metode Naive Bayes (Studi Kasus: Pt. Go-Jek Indonesia). *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer), 2*(1), 434–439. https://doi.org/10.30865/komik.v2i1.972

Liantoni, F. (2016). Klasifikasi Daun Dengan Perbaikan Fitur Citra Menggunakan Metode K-Nearest Neighbor. *Jurnal ULTIMATICS, 7*(2), 98–104. https://doi.org/10.31937/ti.v7i2.356

Nofitri, R., & Irawati, N. (2019). Integrasi Metode Neive Bayes Dan Software Rapidminer Dalam Analisis Hasil Usaha Perusahaan Dagang. *JURTEKSI (Jurnal Teknologi Dan Sistem Informasi), 6*(1), 35–42. https://doi.org/10.33330/jurteksi.v6i1.393

Öztürk, O. B., & Başar, E. (2022). Multiple linear regression analysis and artificial neural networks based decision support system for energy efficiency in shipping. *Ocean Engineering*, 243, 110209.

Pink, M., & Djohan, N. (2021). Effect of ecommerce post-purchase activities on customer retention in Shopee Indonesia. *Enrichment: Journal of Management, 12*(1), 519-526.

Saltz, J. S. (2021, December). CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps. In *2021 IEEE International Conference on Big Data (Big Data)* (pp. 2337-2344). IEEE.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, 181, 526-534.

Widiyanto, P., Endri, E., Sakti, R. F. J., Setiawan, E. B., Manfaluthy, M., Suryaningsih, L., ... & Limakrisna, N. (2021). The relationship between service quality, timeliness of arrival, departure flip ship logistics and people and customer satisfaction: A case in Indonesia. *Academy of Entrepreneurship Journal, 27*(6), 1-12.

Zhang, Y. (2021). Sales Forecasting of Promotion Activities Based on the Cross-Industry Standard Process for Data Mining of E-commerce Promotional Information and Support Vector Regression. *J. Comput*, 32, 212-225.