Comparison of data mining algorithms for pressure prediction of crude oil pipeline to identify congeal

Agus Santoso[1*], *F. Danang* Wijaya[1], *Noor* Akhmad Setiawan[1] , and *Joko* Waluyo[2]

[1]Department of Electrical Engineering and Information Engineering, Faculty of Engineering, Universitas Gadjah Mada, Jl. Grafika 2, Yogyakarta, Indonesia

[2]Department of Mechanical and Industrial Engineering, Faculty of Engineering, Universitas Gadjah Mada, Jl. Grafika 2, Yogyakarta, Indonesia

Abstract. Data mining is applied in many areas. In oil and gas industries, data mining may be implemented to support the decision making in their operation to prevent a massive loss. One of serious problems in the petroleum industry is congeal phenomenon, since it leads to block crude oil flow during transport in a pipeline system. In the crude oil pipeline system, pressure online monitoring in the pipeline is usually implemented to control the congeal phenomenon. However, this system is not able to predict the pipeline pressure on the next several days. This research is purposed to compare the pressure prediction of the crude oil pipeline using data mining algorithms based on the real historical data from the petroleum field. To find the best algorithms, it was compared 4 data mining algorithms, i.e. Random Forest, Multilayer Perceptron (MLP), Decision Tree, and Linear Regression. As a result, the Linear Regression shows the best performance among the 4 algorithms with $R^2 = 0.55$ and RMSE = 28.34. This research confirmed that data mining algorithm is a good method to be implemented in petroleum industry to predict the pressure of the crude oil pipeline, even the accuracy of the prediction values should be improved. To have better accuracy, it is necessary to collect more data and find better performance of the data mining algorithm

1 Introduction

Applications of data science and data mining in many fields become very popular recently, one of which is implementation in the petroleum industries [1]. The data mining is able to support the decision making to prevent a massive loss in the operation of the petroleum industries. One of the most common problems in the operation of the crude oil pipeline system is flow congestion, which is caused by the crude oil flow from liquid to pasta phase. In extreme condition, the crude oil flow will lead to the solid phase. This is called congeal flow in the petroleum industry. Congeal is a serious global problem in the petroleum industry, since it leads to significant effects. Congeal phenomenon wastes the operation cost of crude oil exploration in the world and needs to spend millions of USD to solve this problem [2, 3].

To prevent the congeal occurrences, petroleum industries usually implement online monitoring system which watch out and record the pressure of the crude oil pipeline and other parameters [4]. However, such system is designed to record the occurred events and not able to predict the future events as well as the pipeline pressure on the next several days.

The existing studies on congeal prediction mostly focused on the prediction of wax deposition by using static data resulting from controlled experiments [5-8]. Several data mining algorithms indicated relatively good accuracy in predicting congeal flow due to wax deposition [7, 8]. However, the experimental results cannot be directly applied in the operation field of the crude oil pipeline system, because of the dynamic system of the pipeline. Objective of this research is to carry out analytics of the real historical petroleum field data using several algorithms of data mining and to compare the results for finding the best one to predict the pressure of the crude oil pipeline in future.

2 Methodology

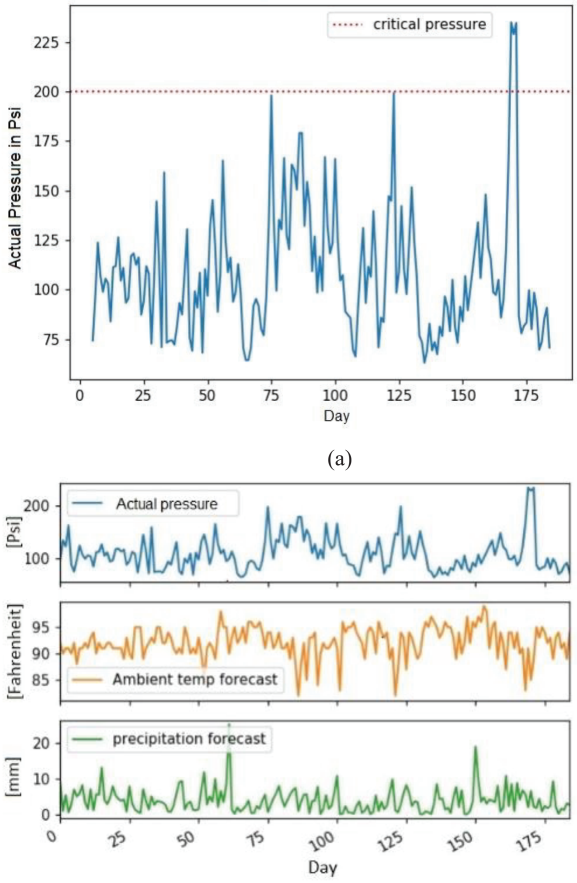
The pressure of crude oil pipeline may be predicted using several algorithms of data mining. In this research, 4 algorithms of data mining were adopted. The first algorithm is Random Forest, which is an ensemble method for regression using a large number of decision trees [9]. The second one, Multilayer Perceptron (MLP), is a type of artificial neural network (ANN) with an individual node named as perceptron organized in a series of layers. Each layer is categorized into three

[*] Corresponding author: agussantosokul@gmail.com

layers of nodes: an input layer, a hidden layer and an output layer [9].

The third algorithm is Decision Tree, that is a tree-like model which can support a decision making based on the conditions [9,10]. The last algorithm, Linear Regression, can be used to make a prediction of dependent attributes based on the several independent variables [10].

Data set employed in this research were a series real historical data acquired from the field of crude oil pipeline of a petroleum company in Indonesia using an online monitoring system. The data set contains the daily actual pressure measured at a certain point in the crude oil pipeline system in Psi. The series length of the daily data is 200 days, as shown in Fig. 1 (a). In this pipeline system, the critical pressure is assumed at 200 Psi. In addition, the daily weather forecast conditions, including ambient temperature and precipitation forecast, were also used as the input data. Figure 1 (b) shows the daily weather forecast conditions in conjunction with the daily actual pressure of the crude oil pipeline system. Example of the data set is shown in Table 1.



(a)

(b)

Fig. 1. (a) The daily actual pressure measured at a certain point in the crude oil pipeline system, (b) The daily weather forecast conditions including ambient temperature and precipitation forecast

Table 1. Example values of the data set.

Day	Act. Press	Amb Temp.	Precipitation
1	97,67	92	8,7
2	134,2	90	0,9
3	122,51	91	5,3
4	162,7	91	0,5
5	87,45	90	2,7
6	74,38	92	6,9
7	96,87	88	5
8	123,75	91	5,8
9	107,04	91	8,2
…	…	…	…

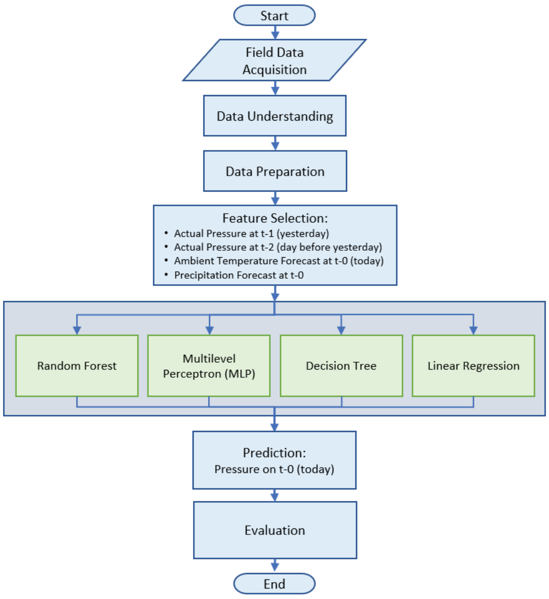

Start
Field Data Acquisition
Data Understanding
Data Preparation
Feature Selection:
• Actual Pressure at t-1 (yesterday)
• Actual Pressure at t-2 (day before yesterday)
• Ambient Temperature Forecast at t-0 (today)
• Precipitation Forecast at t-0
Random Forest
Multilevel Perceptron (MLP)
Decision Tree
Linear Regression
Prediction: Pressure on t-0 (today)
Evaluation
End

Fig. 2. Method for pressure prediction of crude oil pipeline using data mining algorithms.

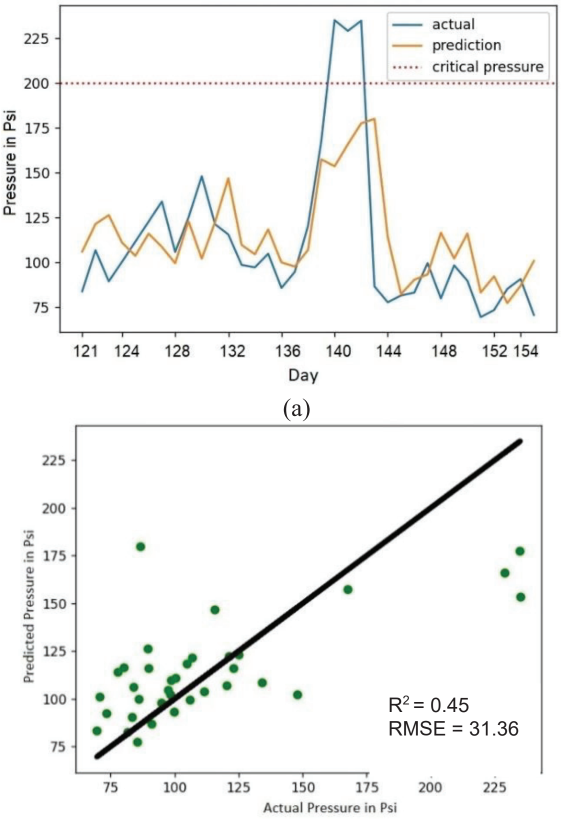
Method employed for the pressure prediction using the data mining algorithms in this research is shown in Fig. 2. The pipeline actual pressure at a day (for example today, t-0) was predicted by the 4 data mining algorithms using 4 input parameters, i.e. the actual historical pressure of 1 day before (yesterday, t-1) in Psi, the actual historical pressure of 2 days before (day before yesterday, t-2) in Psi, the ambient temperature forecast at t-0 in Fahrenheit, and the precipitation forecast at t-0 in mm. Output is the predicted pressure of today (t-0).

The data set was separated in 2 groups. The first group, the series data from day 1 until day 120, was used for the training of the data mining algorithms, while the second group containing the rest series data (from day 121 until day 155) was used for pressure prediction and the blind test evaluation.

Performances of each data mining algorithms employed in this research were measured using the coefficient of determination or R squared (R^2) method, and the root-mean-square deviation (RMSD) or root mean square error (RMSE) method. The coefficient of determination method (R^2) shows whether the predicted pressure is fit with the actual pressure. Values of the coefficient of determination normally ranges from 0 to 1. The value near 1 shows better fitness between the predicted and the actual pressure values. While the RMSE method shows the differences between the predicted and the actual pressure values. The smaller value of the RMSE shows better fitness.

3 Result and discussion

The first algorithm used in the pressure prediction of crude oil pipeline was Random Forest. The algorithm delivers the pressure prediction value for the next day based on the historical data in the data set. Figure 3 (a) shows the predicted pressure of the crude oil pipeline using Random Forest algorithm as a function of time (brown plot) from day 121 until day 155. For comparison, the actual pressure (blue plot) is also shown in the figure. It is confirmed that the values of predicted pressures are going up and down with almost similar trend with those of the actual pressure. Even though there are some error values in the predicted pressure, it can be revealed that data mining algorithm can be used in petroleum industry to predict the crude oil pipeline pressure.

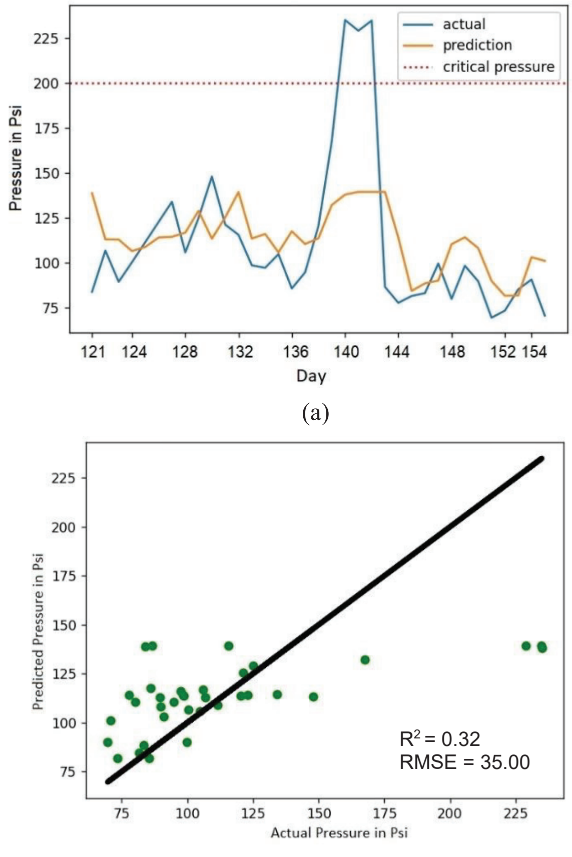


(a)

(b)

Fig. 3. (a) The predicted pressure using Random Forest algorithm and the actual pressure of the crude oil pipeline from day 121 until day 155, (b) Blind test pressure prediction results using Random Forest algorithm.

Figure 3 (b) shows the relationship between the blind test pressure prediction results using Random Forest algorithm and the actual pressure. According to the figure, the predicted values seem accurate at the pressure range below 130 Psi, but rather large deviation for pressure larger than 130 Psi. For blind test evaluation of the accuracy of the pressure prediction, R squared (R^2) and root mean square error (RMSE) tests were adopted in this research. It is attained that the R^2 and RMSE values for the pressure prediction using Random Forest algorithm are 0.45 and 31.36 respectively. The values show that the prediction results still do not accurate enough, therefore the method should be improved.



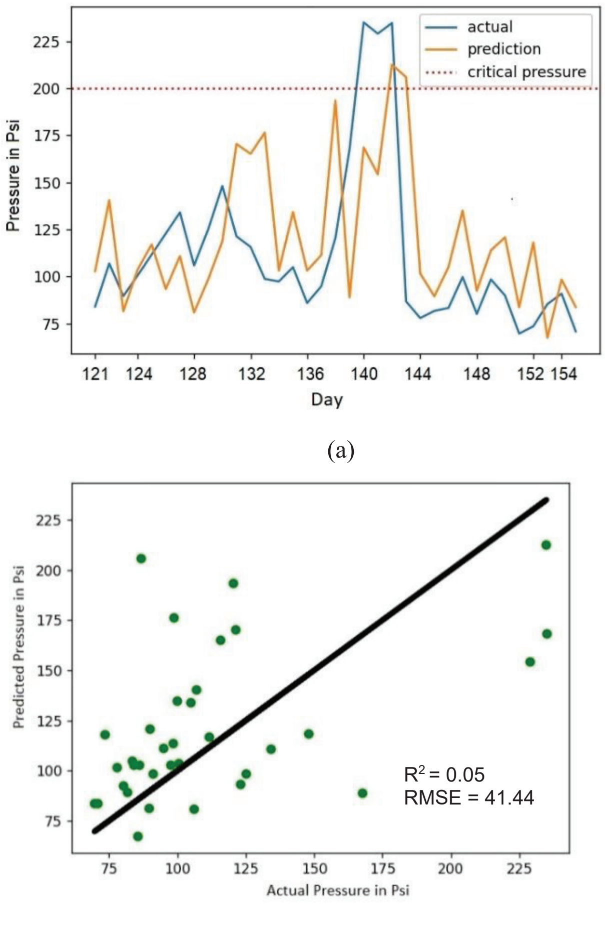
(a)

(b)

Fig. 4. (a) The predicted pressure using Multilayer Perceptron (MLP) algorithm and the actual pressure of the crude oil pipeline from day 121 until day 155, (b) Blind test pressure prediction results using MLP algorithm.

The next algorithm to be used in this research is Multilayer Perceptron (MLP). The brown plot of Fig. 4 (a) shows the pressure prediction values of the crude oil pipeline using MLP algorithm from day 121 until day 155. As the previous figure, the blue plot in the figure shows the actual pressure. The figure shows that the pressure prediction using MLP algorithm seems less accurate than that of using Random Forest algorithm. This conclusion is confirmed by the relationship between the blind test pressure prediction results and the actual pressure as shown in Fig. 4 (b). From this figure, it is known that correlation coefficient R^2 is less than that of using Random Forest algorithm, i.e. 0.32 and the RMSE values is larger than that of using Random Forest algorithm, i.e. 35.00. The larger RMSE means less accurate due to the larger error.

To find the best algorithm to predict the pressure of the crude oil pipeline, this research also employed the Decision Tree and Linier Regression algorithms. The results are shown in the Fig. 5 and Fig. 6. Table 2 shows the pressure prediction result summary of the four algorithms used in this research. Accordingly, the pressure prediction using Decision Tree algorithm indicates better performance than using Random Forest and MLP algorithms, showing by the larger R^2 and the smaller RMSA values. However, the pressure prediction using Linier Regression algorithm shows the best performance comparing to the other 3 algorithms. This tendency agrees with other researches [11].



(a)

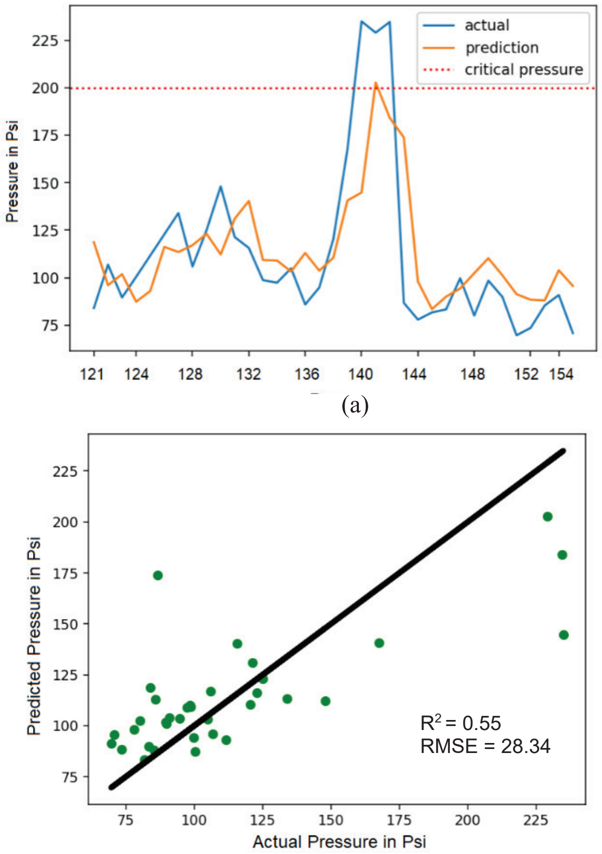
(b)

Fig. 5. (a) The predicted pressure using Decision Tree algorithm and the actual pressure of the crude oil pipeline from day 121 until day 155, (b) Blind test pressure prediction results using Decision Tree algorithm.

Table 2. Summary of the pressure prediction results using data mining with Random Forest, Multilayer Perceptron (MLP), Decision Tree, and Linear Regression prediction models.

Metrics	**Algorithm**			
	Random Forest	**MLP**	**Decision Tree**	**Linier Regression**
R2	0.45	0.32	0.05	0.55
RMSE	31.36	35	41.44	28.34

The 4 data mining algorithms successfully predicted the next day pressure of the crude oil pipeline. However, according to the blind test results, the accuracy should be improved to obtained more confidence predictions. Improvement of the prediction accuracy may be carried out by several methods, such as enrichment of the data transformation, adding the new and more comprehensive data set, and finding the better data mining algorithm and tuning up the parameters.



(a)

(b)

Fig. 6. (a) The predicted pressure using Linear Regression algorithm and the actual pressure of the crude oil pipeline from day 121 until day 155, (b) Blind test pressure prediction results using Linear Regression algorithm.

4 Conclusion

The pressure predictions of crude oil pipeline have been analysed using 4 algorithms of data mining: Random Forest, Multilayer Perceptron (MLP), Decision Tree, and Linear Regression algorithms. The predicted pressures using the 4 algorithms are agree and having similar trend with the actual pressures, even the accuracies which represented by the R^2 are not high, about 0.55 or less. Comparison of the results confirmed that the Linear Regression with $R^2 = 0.55$ and RMSE = 28.34 is the best algorithm among the 4 data mining algorithms adopted in this research for this particular data set. Therefore, it can be concluded that data mining algorithms are useful to predict the crude oil pipeline in the petroleum industry. However, there are still many rooms for continuing this research to improve the results. In future, further recommended works may be focused on the improving the accuracy and reducing the error by data transformation and data set improvement, also finding the better data mining algorithm and tuning up the parameters.

References

1. P. Bangert, Machine Learning and Data Science in the Oil and Gas Industry (Gulf Professional Publishing, 2021)
2. A. K. Manshad, H. Rostami, H. Rezaei, S. M. Hosseini, H. Niknafs and A. H. Mohammadi, Heavy Oil, 161-173 (Nova Science Publishers, New York (2017)
3. Z. Q. Chu, J. Sasanipour, M. H. Saeedi, A. Baghban, and H. Mansoori, J. Petrol. Sci. Tech **35**, 1974-1981 (2017)
4. E.B. Priyanka, C. Maheswari, S. Thangavel, Flow Meas. Instrum **62**, 144-151 (2018)
5. A. Kamari, A. K. Manshad, F. Garagheizi, A. H. Mohammadi, “A Robust Model for Determination of Wax Deposition in Oil Systems”, ACS, I&EC (2013)
6. G. Zhang, G. Liu, J. Petrol. Sci. Eng **70**, 1-9 (2010)
7. T. J. Behbahani, A. A. M. Beigi, Z. Taheri, B. Ghanbari, J. Petroleum **1**, 223-230 (2015)
8. E. O. Obanijesu and E. O. Omidiora, J. Petrol. Sci. Tech **26**:16, 1977–1991 (2008)
9. J. H. Moon,,Y. S. Kim, M. J. Son, and E. J. Hwang, Energies **11**(12), 3283 (2018)
10. Y. S. Kim, Expert Syst. Appl **34**, 1227-1234 (2008)
11. Z. Hu, M. Wu, K. Hu dan J. Liu, J. Petrol. Sci. Tech **33**, 1499-1507 (2015)