PROFESSIONAL TECHNICAL PAPER
ONLINE PRESENTATION
24 - 25 OCTOBER 2020

IATMI

A Breakthrough Approach for Predicting ESP Wells Virtual Flow Rate by Using Supervised Machine Learning Method

Muhammad Irfan*[1], Marda Vidrianto[2], Silvya Dewi Rahmawati[1], and Aulia Ahmad Naufal[3]

[1]Petroleum Engineering, Institut Teknologi Bandung
[2]Artificial Lift Engineering, Pertamina Hulu Energi OSES
[3]Production Engineering, Schlumberger
* Email: muhammadirfan.itb@gmail.com

Abstract. For field that deployed clusters of Electrical Submersible Pump (ESP) wells, implementing a robust surveillance system that could serve as early warning detection, real-time monitoring, and optimizing could give significant advantages. In establishing that system, real-time fluid rate becomes one important aspect but it cannot be obtained yet due to the limited frequency of conventional well test. However, ESP wells which have downhole sensor give valuable benefit to engineers as downhole sensor generates non-stop streaming data that represent ESP condition. Therefore, an idea to convert this non-stop streaming data into real-time fluid rate which could serve as an alternative to the conventional well test arises. An innovation that is discussed in this study is proposed to predict a virtual flow rate by utilizing the collection of data from ESP real-time sensor and wells information which simultaneously train and run a selected machine learning model.

In this study, the dataset collected from formation layers, ESP specification, tubing property, ESP real-time sensor, wellhead pressure, casing pressure, and historical well test data have been cleaned up before it is used to train model and predict the result once it is deployed. Afterwards, feature engineering is conducted to reduce the dimensionally of data. With the value of R-squared as indicator, six regression models comprised of K-Nearest Neighbor (KNN), Support Vector Machine Regression (SVR), Random Forest Regression (RFR), Extreme Gradient Boosting (XGBoost), Linear Regression, and Elastic Net are compared to choose the best predictive model after parameter optimization for each model is applied.

This study used 14,915 data points from 12 mature wells in the Offshore Southeast Sumatera field to train and test the model. The sensitivity study done yielded SVR with the penalty parameters (C) value of 1000 and gamma (γ) value of 0.1 as the best algorithm and parameters for this case. The model reaches 96.05% level of accuracy when it is evaluated with 176 point of historical production test data. This study also shows that the model has succeeded to estimate the value of unknown fluid rate when the wells are not being tested.

The novelty of this paper is associated with the application of new machine learning model that can estimate ESP wells virtual flow rate in Offshore Southeast Sumatera. This study also shows the importance of data preparation, parameter optimization and feature engineering in achieving the proper model for prediction. Post-deployment, the model must be continuously updated its data especially when it is unable to approximate fluid rate properly.

Keyword: ESP, fluid rate, real-time, supervised machine learning, regression

1 Introduction

Electric Submersible Pump (ESP) has become one of artificial lift method that is commonly used in mature wells. Mature wells are the wells which have exceed the peak of production time and lack of energy to lift reservoir fluid to the surface. Therefore, ESP is installed at the certain depth of the well to give additional pressure to keep fluid flowing. Nowadays, ESP has widely known as one of artificial lift with the highest installation and its significant contribution to total oil production in the world. Establishing a robust surveillance system which could serve as early warning detection, real-time monitoring, and optimizing could give significant advantages. Since this system allows engineers to take immediate action when pump performs below desired condition.

Real-time fluid rate becomes one important aspect in ESP surveillance system. However, it cannot be obtained yet as fluid rate from the production flow test or known as well test does not represent real-time condition of ESP as shown in **Figure 1**. As well test consumes 3 – 24 hours for flowing the reservoir fluid from the well to the test separator or the multiphase flow meter and it also has limited frequency especially if the well is located in remote area. Those reasons make fluid rate from conventional well test cannot be used to identify ESP real-time performance.

Nevertheless, ESP wells give valuable benefit to engineers with the non-stop streaming data generated by ESP downhole sensor. ESP downhole sensor captures information such as pump intake pressure, pump discharge pressure, pump intake temperature, motor temperature and electrical current up to hundreds of data in the daily operation. Then, an idea to convert this non-stop streaming data into real-time fluid rate by using supervised machine learning arises. Supervised machine learning is able to create rules based on dataset patterns to get the value of fluid rate using new input data.

In this study, the dataset is collected from ESP real-time sensor data and well information in M Field which is located in Offshore Southeast Sumatera as case study to illustrate the implementation of supervised machine learning method. Supervised machine learning regression which consists of K-Nearest Neighbor, Support Vector Machine Regressor (SVR), Random Forest, Extreme Gradient Boosting (XGBoost), Linear Regression, and Elastic Net is used to predict virtual flow rate in ESP wells. Those regression model are simultaneously trained and compared their accuracy after being evaluated with historical production test data. Afterwards, predictive model is selected based on the highest level of accuracy (average R-squared of test data) and the lowest over-fitting tendency. Afterwards, the chosen model is used to estimate fluid rate value when the well is not being tested.

2 Methodology

2.1 Data Preparation

As shown in **Figure 2**, this study begins with data preparation which consist of gathering data and data preprocessing. In this stage, the available data is collected and chosen based on domain knowledge considered to have impact on fluid rate in ESP wells. After all data required has been gathered, data preprocessing is conducted to clean up the raw data.

2.1.1 Gathering Data

Gathering data becomes the most important aspect in building machine learning model as its quality and quantity will directly affect to the model. In this study, data is obtained from formation layers, ESP specification, tubing property, ESP real-time sensor, wellhead pressure, casing pressure, and historical well test of 12 mature wells in Offshore Southeast Sumatera for past three years, since January 2017 until January 2020. Formation layers gives information about the variation layers opened in each well while ESP specification tells about pump characteristic such as pump type, stage, power, and electrical voltage. Afterwards, tubing property gives information about the size of inner tubing diameter used for each well. Then, wellhead pressure and casing pressure data take into account because they are considered to give significant impact to fluid rate. Meanwhile, ESP real-time sensor provides downhole information such as electrical current, pump intake pressure, pump discharge pressure, pump intake temperature, and motor temperature. However, those sensor data must be converted into daily time first. Thereupon, the previously mentioned data is gathered in one frame and adjusted with the historical well data availability which act as labeled data.

2.1.2 Data Preprocessing

A collection of data from the previous step is known as raw data which may still contains missing value, outlier, error, and imbalance that should be handled first before constructing regression model. Missing value in the dataset can be detected by using missing value plot that shows empty data in the features. There are three ways in handling missing value. First, if one feature contains missing value more than 50%, its feature should be eliminated. Second, if one feature has 10% until 50% missing value, the missing value should be replaced using median or mean by looking at its distribution first. Third, if the missing value has less than 10%, dropping the rows that contain missing value is taken to be an option. As shown in **Figure 3** and **Table 1**, the highest missing value in the dataset is 6.37% so the third option for dropping the rows is selected. Then, the dataset which has been cleaned up its missing value is shown in **Figure 4**. Next, error values from ESP sensor indicated with constant value for long period, higher intake temperature value than motor temperature value, and higher intake pressure value than discharge pressure value are handled by eliminating their rows. On the other side, outlier values which also come from ESP sensor are removed by keeping the value within the following range considered to be the best practice of ESP in M Field.

$$220\ ^{o}\mathrm{F} \leq \text{Intake Temperature} \leq 300\ ^{o}\mathrm{F}$$
$$260\ ^{o}\mathrm{F} \leq \text{Motor Temperature} \leq 400\ ^{o}\mathrm{F}$$
$$100 \text{ psi} \leq \text{Intake Pressure} \leq 2000 \text{ psi}$$
$$1000 \text{ psi} \leq \text{Discharge Pressure} \leq 5000 \text{ psi}$$

Afterwards, ESP type and tubing inner diameter are classified as categorical data as they do not have much variety in the dataset. Next, the last step of data preprocessing is numerical data imbalance checking identified using Multivariate Linear Regression that shows unbalanced coefficients. Multivariate Linear Regression equation is shown below and the table of coefficient can be seen in **Table 2.**

$$Q_f = a.\,(Stage) + b.\,(HP) + c.\,(Volt) + \cdots\ + j.\ (\text{Casing Pressure}) + \text{Intercept}$$

One of the ways in dealing with the imbalanced dataset is by applying standard scale. Although, this technique is only work for normal distribution dataset and it cannot applied properly if there is feature that contains skew. In the dataset, ESP electrical current shows positive skew in the distribution plot while ESP intake pressure shows negative skew as shown in **Figure 5**. After knowing that the dataset has positive and negative skew, Yeo-Johnson transformation is used to normalize those features since this transformation technique can be applied well in both positive skew data and negative skew data.

2.2 Model Development

Next stage of this study is regression model development which consist of K-Nearest Neighbor (KNN), Support Vector Regression (SVR), Random Forest Regression (RFR), XGBoost, Linear Regression, and Elastic Net. Those models are built and tuned their parameters to achieve high level of accuracy which represents the most optimum parameters. The most optimum parameters are reached by using parameter tuning techniques such as Grid Search CV, Randomized Search CV, and Bayes Search CV which are also been validated by using K-Fold Cross Validation.

2.2.1 Dataset Splitting

After the dataset has been cleaned up, the next step to do is dataset splitting. Dataset splitting is the process done to prevent information outside the training dataset enter the model, known as data leakage. In this stage, the rows of clean data are shuffled 101 times first before they are split up into training data and test data. Training data size is set to be 70% while test data size is set to be 30% to ensure that the training data size is neither too high nor too low. Since higher training data size will lead model into over-fitting problem while lower training data size will make model have under-fitting tendency.

2.2.2 Regression Algorithm

K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) begins with storing all the training data so that its distribution pattern can be identified by machine. Afterwards, enumerating distance (N_0) from the test data (x_0) to all neighbor (K) value is carrying out before sorting the K value is done by increasing N_0 from x_0. Then, the output of test data or new data is determined by the majority label of the closest training data. Three matters which must be concerned are the number of K, the length of distance, and the type of weight whether uniform for each neighbor or influenced by distance. The number of K becomes the most important thing in KNN as it should be an odd number in order to prevent indecision. The illustration of KNN is shown in **Figure 6** where there are training data which consists of 5 yellow and 5 purple points. If number of K is 3, new data will be considered as part of Class B. On the other hand, if number of K is 6, new data will belong to Class A.

Support Vector Machine Regressor (SVR)

Support Vector Machine Regression (SVR) separates the label data by creating the widest line (2D) or plane (3D). The widest line or plane is determined by measuring the margin, the distance between closest training data for each class or known as vector. The illustration of line SVR is shown in **Figure 7**. In that figure, the orange circle shows the value of 400 BFPD while the blue star shows the value of 200 BFPD. If the new data is on orange circle area, its value will be 400 BFPD. Meanwhile, if the new data is on the blue star area, its value will be 200 BFPD.

Penalty parameter (C), kernel trick, and Gamma (γ) are the parameters used to adjust line and plane in SVR. Penalty parameter (C) determines the width of margin that directly affects to noise data. As shown in **Figure 8**, higher C value results smaller margin and it makes margin more sensitive to training data while smaller C yields wider margin and it makes margin become more tolerant to noise data. Next, kernel trick plays a role in transforming nonlinear data into other dimension such as 3D when the dataset cannot be separated well in 2D as shown in **Figure 9.** Furthermore, kernel coefficient (γ) in SVR which determines the precision of seeing the data is used to handle inappropriate scaling that yields data cannot be separated properly when it has high standard deviation (σ). However, higher value of kernel coefficient brings higher accuracy but it makes the boundaries fit too smooth that result into over-fitting problem as shown in **Figure 10.** On the other hand, lower value of kernel coefficient has lower accuracy due to under-fitting problem.

Random Forest Regressor

Random Forest Regression uses ensemble method for combining all output values from a number of decision tree and converting them into one final decision. This matter shows that random forest is the further development of decision tree which overcomes the over-fitting problem by performing bootstrap and aggregation called as bagging. As shown in **Figure 11**, random forest starts with building a number of decision tree that run in parallel through dataset which is randomly given a number of rules with replacement or known as bootstrap. Bootstrap reduces correlation and its decision yields lower variance as it will be averaged first. In developing this model, the number of trees (n_estimators), level of features (max_depth),the amount of data split or required in each node (min_samples_split or min_samples_leaf) and the amount of features to be split (max_features) are the common parameters adjusted to find the best model. Higher value of max_depth yields the possibility of over-fitting problem while higher value of n_estimators, min_samples_split, and min_samples_leaf reduce the variance of model.

Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting or known as XGBoost is included as one of ensemble method where next tree boosts current attributes in order to diminish mistake from previous tree. Moreover, XGBoost also has already built with ability to handle missing value and work in large-scale data. XGBoost has three kind of parameters which are general parameters, boosting parameters, and task parameters. General parameters consist of number of thread (n_thread) and tree-based model (booster). Next, learning rate, maximum depth of tree (max_depth), random fraction of observation (colsample_bytree), and regularization term on weight (L1 and L2) belong to a bunch of boosting parameters which are tuned to improve the accuracy.

Learning rate should be maintained below one in order to avoid residual aggressive fitting. Similarly, deeper of maximum depth yields over-fitting as the model will easily fit with its residual as shown in **Figure 12**.

Linear Regression

Linear regression approximates the label data with linear approach between one label data and one or more independent variables. There are two kind of linear regression which are simple linear regression and multiple linear regression. Simple linear regression is linear regression that link two variables assumed have a linear relationship between independent variable and dependent variable as shown in **Figure 13**. On the other side, multiple linear regression has more than one independent variables which give contribution to the label value.

Elastic Net

Elastic net combines lasso and ridge method in applying regularization to the model as shown in **Figure 14**. The combination of both techniques make Elastic net overcome lasso weakness in taking samples in high dimensional data by using the quadratic part of penalty. Quadratic penalty eliminates the limit of selected variables and encourages grouping effect that help the variables to be easily recognized. Moreover, Elastic net has higher number of predictor than the training data so that it helps model in enhancing the level of accuracy.

2.2.3 Parameter Tuning

Grid Search CV

Grid Search CV is one of parameter optimization techniques which tries all possible parameter combinations in accordance with given grid values by user as shown in **Figure 15**. The grid values which are used in this study can be seen in **Table 3**. Despite it has better precision than other optimization techniques, this technique lacks of efficient as it gives try for all possible combinations.

Randomized Search CV

Another optimization technique applied in this study is Randomized Search CV. Unlike previous technique, this technique employs random combination that is set based on the given range of parameter values and number of iteration. This technique brings better computational but sometimes it miss the best parameters combination when the ranges of value are not set properly. The illustration of this technique is seen in **Figure 16** while the given parameter values can be seen in **Table 4**.

Bayes Search CV

The last tuning parameter technique applied in this paper is Bayes Search CV. This technique utilizes last evaluation result to select the next evaluation value until the best parameters combination found. Its computational depends on the number of iteration and range of parameter values. This technique commonly used to tuned parameters of Random Forest and XGBoost. The illustration this technique is shown in **Figure 17** and the given parameter values can be seen in **Table 5**.

K-Fold Cross Validation

After optimizing the model parameters, cross validation is conducted to evaluate model performance. It has purpose to prevent luck in creating model such as getting high level of accuracy model without having validation. In this study, K-Fold Cross Validation is chosen as cross validation method that folds the amount of K and performs the number of K experiment. The number of fold is set to be 5 for every regression algorithm so each model will be validated 5 times. The illustration of K-Fold Cross Validation is shown in **Figure 18** while the application K-Fold Validation in Grid Search CV, Random Search CV, and Bayes Search CV which are respectively shown in **Figure 19**.

2.2.4 Feature Engineering

Next step of this study is feature engineering which aims to enhance the accuracy of model by eliminating some features in the dataset. Feature engineering is divided into two steps which are feature importance and correlation matrix. Feature importance shows feature influence on the labeled data while correlation matrix gives information about the relation between the independent variables. In this study, feature importance can be known using mean score decrease plot which uses the highest level of accuracy model as reference. As shown in **Figure 20**, mean score decrease plot shows that ESP electrical current is the most influential parameter for fluid rate while formation layer A, D, E, and F have low impact to the target variable so they can be eliminated from the dataset.
On the other hand, correlation matrix uses normalized pearson plot as shown in **Figure 21** to show linear relationship between independent variables after being corrected first in order to avoid the outlier such as heteroscedasticity. In this plot, it is shown that intake temperature and discharge pressure have strong correlation (0.8) to motor and wellhead pressure respectively so eliminating one from each features is taken an option to enhance the accuracy of model. Later on, new predictive models are constructed and optimized their parameters to see whether their accuracy are increasing or not. Moreover, this study is also create several models just based on ESP features, tubing pressure, and casing pressure. Subsequently, all predictive models are compared their level of accuracy so the final model which has the highest level of accuracy can be selected.

2.3 Prediction

After finding out the most appropriate model, the next stage of this study is performing virtual flow rate prediction to fill out the missing value of fluid rate from ESP wells in M Field from past three years. Then, the values of unknown flow rate when the wells are not being tested are obtained and it measures the reliability of model.

3 Case Study

M Field is a mature field located in Sunda Basin, Offshore Southeast Sumatera which has been producing since 1982. Nowadays, majority of wells in M Field have equipped themselves with ESP installed in range of depth 4000 - 7000 ft and deeper than 7000 ft. Moreover, ESP wells in M Field have relatively short ESP run lives in the range of 70 up to 365 days so the additional cost for workover and pump replacement for each year are inevitable.

In this study, input data for virtual flow rate prediction is taken from twelve wells in M Field which are MA-04, MA-06, MA-14, MB-07, MB-10, MB-12, MB-13, MB-15, MC-09, MC-10, MC-11, and MC-12. The observation is start from 1 January 2017 until 31 January 2020. Within this range of date, RC4000, RC1000, DN1750, DN1050, and DN460 are the types of ESP installed which result different ESP specifications such as stage, power, and electrical voltage. Next, these wells have been producing in the different layers formation such as Formation A, Formation B, Formation C, Formation D, Formation E, and Formation F. Then, inner tubing diameters in this field vary between 2.441 in and 2.992 in. Furthermore, ESP sensor provides the values of ESP electrical current, intake temperature, motor temperature, intake pressure, and discharge pressure. Moreover, wellhead pressure and casing pressure are entered into the dataset as they have significant effect to fluid rate.

4 Result and Discussion

Six regression algorithms are constructed by using 14,915 data points from 12 mature wells in the Offshore Southeast Sumatera field along with conducting parameter tuning and feature engineering in order to boost the accuracy of models. As shown in **Figure 22**, the final result of this study states that Support Vector Machine Regressor (SVR) built based on ESP features, tubing pressure, and casing pressure appears as the best performance model with the 96.05% level of accuracy and the lowest over-fitting tendency among other model after being evaluated by 176 data points of historical well test as shown in **Table 6**. Moreover, both penalty parameter (C) value of 1000 and gamma (γ) value of 0.1 are chosen as the best parameters for this case.

Afterwards, virtual flow rate prediction is conducted for 12 ESP wells before their accuracy and error which are represented by average R-squared of test data and Root Mean Square Error (RMSE) are evaluated as shown in the **Table 7.** MB-13, MB-10, MC-09, and MA-14 are the top four wells with the highest performance among the rest of the wells by seeing their average R-squared of test data and RMSE. Moreover, those wells have also been validated with great amount of historical well test as shown in **Figure 23**. On the other hand, MC-11, MC-10, MB-15, and MA-06 have good level of accuracy and RMSE but they are lack of historical well test data as shown in **Figure 24**. Furthermore, MA-04, MC-12, MB-07, and MB-12 have lower level of accuracy even though they have more historical well test data than previously mentioned wells as shown in **Figure 25**. However, virtual flow rate prediction cannot be done if there is still error value in the input dataset as it will yield incorrect output as shown in **Figure 26**. Therefore, cleaning data must be conducted first before the data is used for prediction.

5 Conclusion

In conclusion, SVR model which has penalty parameter (C) value of 1000 and gamma (γ) value of 0.1 has been chosen as the best predictive model for estimating virtual flow rate from ESP wells in Offshore Southeast Sumatera with 96.05% level of accuracy. This model is obtained after feature engineering which eliminates some features and leaves only tubing pressure, casing pressure, and several ESP features such as pump type, stage, power, electrical voltage, electrical current, pump intake pressure, and pump intake temperature. As supervised machine learning utilizes the pattern of the dataset, it is very important to prepare data first since the training dataset gives direct impact to the model's performance. Afterwards, optimizing the parameters of model and cross validation are the next step to do to find out the best model parameters. Then, feature engineering can be considered as an option to improve good level of accuracy model into the higher accuracy model by eliminating the features which have no significant effect in the dataset and high correlation among the independent features.

Hereafter, this study shows that virtual flow rate prediction can properly be applied in MB-13, MB-10, MC-09, and MA-14 since the average R-squared of test data in those wells ranges between 84% - 96% and their RMSE ranges between 18 to 45. Moreover, the reliability of virtual flow rate prediction in those wells has already been supported by historical well test data. On the other side, MC-11, MC-10, MB-15, and MA-06 need more production data to have better validation towards virtual flow rate prediction in order to boost the feasibility of model. Meanwhile, fluid rate estimation in MA-04, MC-12, MB-07, and MB-12 can be implemented better if the model updates its data first. Furthermore, it is important to ensure that there is no error or incorrect value in the input data in order to prevent failure in approximating flow rate.

6 Recommendation

The predictive model must be continuously updated its data especially when it is unable to estimate the virtual flow rate correctly. Moreover, it is also needed to update the model when there is an event such as pump replacement, tubing replacement, and other workover activities. Afterwards, it is very suggested to apply deep learning in approximating virtual flow rate value for further development of this study.

7 Acknowledgement

The authors would express the gratitude to PT Pertamina Hulu Energi Offshore Southeast Sumatera for the opportunity to publish the data and share this study.

References

[1] Caicedo, S., Montoya, C. 2012. Estimating Flow Rates Based on ESP Down Hole Sensor Data. Paper presented at the SPE Kuwait International Petroleum Conference and Exhibition held in Kuwait City, Kuwait, 10 – 12 December 2012.

[2] Brown, Kermit E. The Technology of Artificial Lift Methods. Penwell Books, Tulsa Oklahoma, 1984.

[3] Economides, J. Michael., Hill, Daniel A. 1994. Petroleum Production System. Prentice Hall Inc., Englewood Chiffs, New Jersey.

[4] Frank, E., Hall, M. A., Witten, I. H., & Pal, C. J. 2016. Data Mining: Practical Machine Learning Tools and Techniques 4th edition. Morgan Kaufmann.

[5] Guo, Boyun., C.Lyons, William., and Ghalambor, Ali. 2006. Petroleum Production Engineering. Gulf Professional Publishing, Louisiana.

[6] Hastie, T., Zou, H. 2004. Regularization and Variable Selection via the Elastic Net. Stanford University.

[7] Izgec, B., Hasan, A. R., Lin, D., & Kabir, C. S. 2008. Flow Rate Estimation From Wellhead Pressure and Temperature Data. SPE Annual Technical Conference and Exhibition.

[8] Hunter, J., D. 2007. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95.

[9] James, G., Witten, D., Hastie, T., Tibshirani, R. 2013. An Introduction to Statistical Learning with Applications in R. Springer, New York, NY. Available at: https://doi.org/10.1007/978-1-4614-7138-7.

[10] Putra, W.D.K. 2018. J.Cop ML. https://pypi.org/project/jcopml. Accessed on 31 May 2020 at 14.45 WIB.

[11] Putra, W.D.K. 2018. Supervised Learning. https://github.com/WiraDKP/supervised_learning. Accessed on 31 May 2020 at 13.30 WIB.

[12] Putra, W.D.K. 2019. LuWiji. https://pypi.org/project/luwiji. Accessed on 31 May 2020 at 16.30 WIB.

List of Tables

Table 1. Missing Value Percentage

Feature	Missing Value	Missing Value (%)
Layer A	0	0
Layer B	0	0
Layer C	0	0
Layer D	0	0
Layer E	0	0
Layer F	0	0
ESP Type	0	0
Stage	0	0
HP	0	0
Voltage	0	0
Tubing ID	0	0
Average Ampere	0	0
Intake Temperature	**50**	**6.37**
Motor Temperature	**50**	**6.37**
Intake Pressure	**50**	**6.37**
Discharge Pressure	**50**	**6.37**
Tubing Pressure	**2**	**0.25**
Casing Pressure	**2**	**0.25**

Table 2. Coefficient of Multivariate Linear Regression

Name	Coefficient
Intercept	2173.027
Stage	1.016
HP	-3.470
Volt	**0.510**
Average Amps (A)	19.362
Intake Temperature (F)	-15.241
Motor Temperature (F)	3.813
Intake Pressure (psi)	**0.255**
Discharge Pressure (psi)	**0.050**
Tubing Pressure (psi)	**-0.450**
Casing Pressure (psi)	**0.019**

Table 3. Grid Search CV Parameters

KNN	SVR	RF	XGBoost	Linear	Elastic Net

-Number of neighbour : [3, 9, 17] -Weight : [Uniform, Distance] -Distance : [Manhattan, Euclidean]	-Gamma : [0.1, 1, 10] -Penalty parameter : [0.1, 1, 100, 1000]	-Number of tree : [100, 200, 400] -Level of tree : [10, 20] -Max feature : [0.1, 0.3]	-Number of tree : [100, 200, 400] -Level of tree : [5, 10, 20] -Column sample : [0.2, 0.6, 1] -Subsample : [0.2, 0.4, 0.8] -Learning rate : [0.1, 1, 2] -Alpha : [0.1, 10] -Lamda : [0.1, 10]	-Intercept : [True, False]	-Intercept : [True, False] -Alpha : [0.1, 1, 10] -L1 ratio : [0.5, 1]

Table 4. Randomized Search CV Parameters

KNN	**SVR**	**RF**	**XGBoost**	**Linear**	**Elastic Net**
- Number of neighbour : (1 – 40) -Weight : [Uniform, Distance] -Distance : [Manhattan, Euclidean]	-Gamma : (0.001-1000) -Penalty parameter : (0.001-1000)	-Number of tree : (100 – 200) -Level of tree : (20 - 80) -Max feature : (0.1 – 1) -Min samples leaf : (1 – 20)	-Number of tree : (100 - 200) -Level of tree : (1 – 10) -Column sample : (0.1 – 1) -Subsample : (0.3 – 0.8) -Learning rate : (0.01 – 1) -Alpha : (0.001 – 10) -Lamda : (0.001 – 10) -Gamma : (1 – 10)	-Intercept : [True, False]	-Intercept : [True, False] -Alpha : (0.0001–100) -L1 ratio : (0 – 1)

Table 5. Bayes Search CV Parameters

KNN	**SVR**	**RF**	**XGBoost**	**Linear**	**Elastic Net**

- Number of neighbour : (1 – 40) -Weight : [Uniform, Distance] -Distance : [Manhattan, Euclidean]	-Gamma : (0.001-1000) -Penalty parameter : (0.001-1000)	-Number of tree : (100 – 200) -Level of tree : (20 - 80) -Max feature : (0.1 – 1) -Min samples leaf : (1 – 20)	-Number of tree : (100 - 200) -Level of tree : (1 – 10) -Column sample : (0.1 – 1) -Subsample : (0.3 – 0.8) -Learning rate : (0.01 – 1) -Alpha : (0.001 – 10) -Lamda : (0.001 – 10) -Gamma : (1 – 10)	-Intercept : [True, False]	-Intercept : [True, False] -Alpha : (0.0001– 100) -L1 ratio : (0 – 1)

Table 6. Model Performance Comparison

Before FE	**Training R-squared**	**Average Test R-squared**	**Best Optimizer**
KNN	1	0.9323	Bayes Search CV
SVR	0.9794	0.9465	Grid Search CV
RF	0.995	0.9464	Bayes Search CV
XGBoost	0.9991	0.9390	Bayes Search CV
Linear	0.8930	0.8470	Grid Search CV
Elastic Net	0.8913	0.8611	Grid Search CV

After FE	**Training R-squared**	**Average Test R-squared**	**Best Optimizer**
KNN	0.9999	0.9476	Grid Search CV
SVR	0.9752	0.9586	Grid Search CV
RF	0.9940	0.9508	Grid Search CV
XGBoost	0.9929	0.9490	Random Search CV
Linear	0.8872	0.8543	Grid Search CV
Elastic Net	0.8820	0.8684	Random Search CV

ESP Features, Tubing Pressure, and Casing Pressure Only

Before FE	Training R-squared	Average Test R-squared	Best Optimizer
KNN	1	0.9381	Grid Search CV
SVR	0.9774	0.9468	Grid Search CV
RF	0.9868	0.9486	Bayes Search CV
XGBoost	0.9989	0.9401	Bayes Search CV
Linear	0.8290	0.8408	Grid Search CV
Elastic Net	0.8284	0.8411	Grid Search CV

ESP Features, Tubing Pressure, and Casing Pressure Only

After FE	Training R-squared	Average Test R-squared	Best Optimizer
KNN	0.9999	0.9477	Grid Search CV
SVR	**0.973**	**0.9605**	**Grid Search CV**
RF	0.9940	0.9580	Bayes Search CV
XGBoost	0.9993	0.9432	Bayes Search CV
Linear	0.8267	0.8373	Grid Search CV
Elastic Net	0.8261	0.8375	Grid Search CV

Table 7. Prediction Model Performance in ESP wells

Well Name	R-squared	RMSE
MA-04	0.5855	67.99
MA-06	0.820	27.81
MA-14	0.8783	34.92
MB-07	0.5139	62.88
MB-10	0.8420	43.38
MB-12	0.2974	89.45
MB-13	0.8814	32.59

MB-15	0.8588	65.60
MC-09	0.9645	18.17
MC-10	0.8917	26.29
MC-11	0.9222	19.42
MC-12	0.5171	111.37

List of Figures

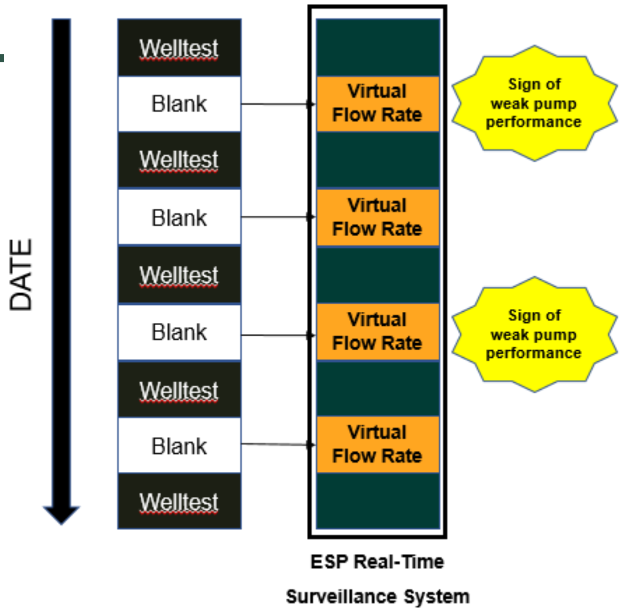


Figure 1. ESP Real-Time Surveillance System.

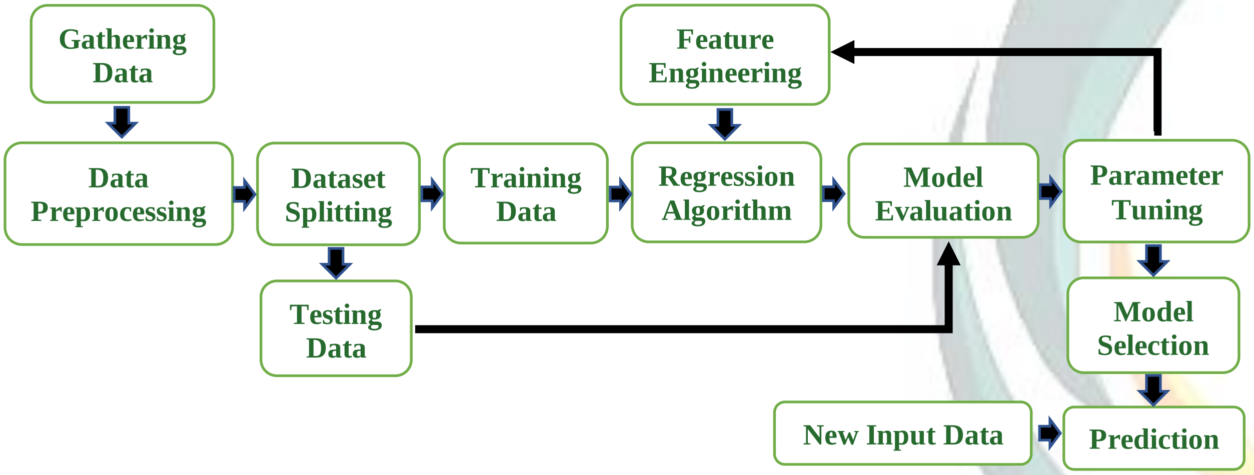


Figure 2. ESP Wells Virtual Flow Rate Prediction Workflow.

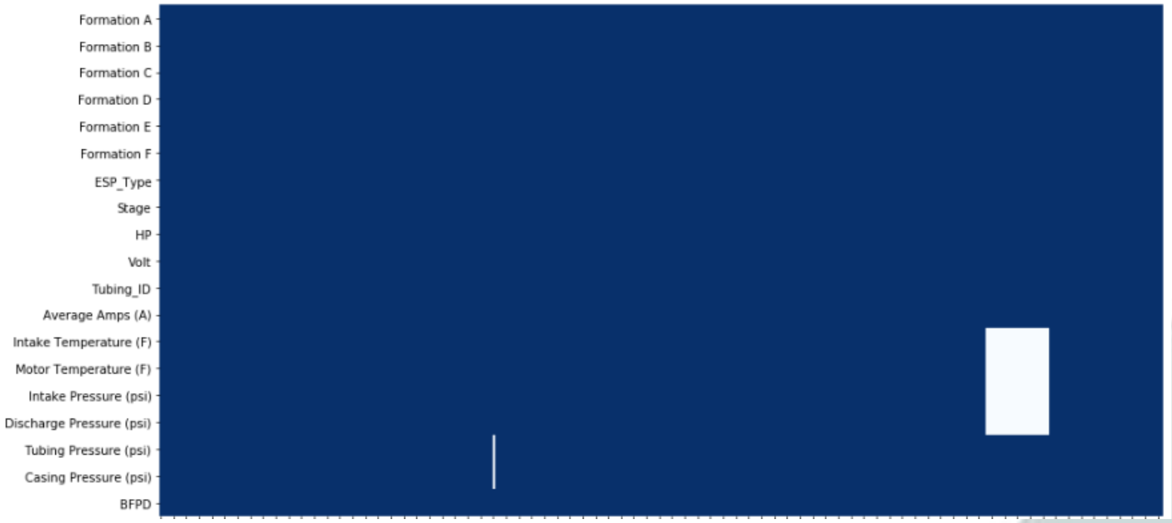


Figure 3. Missing Value Plot of Raw Data

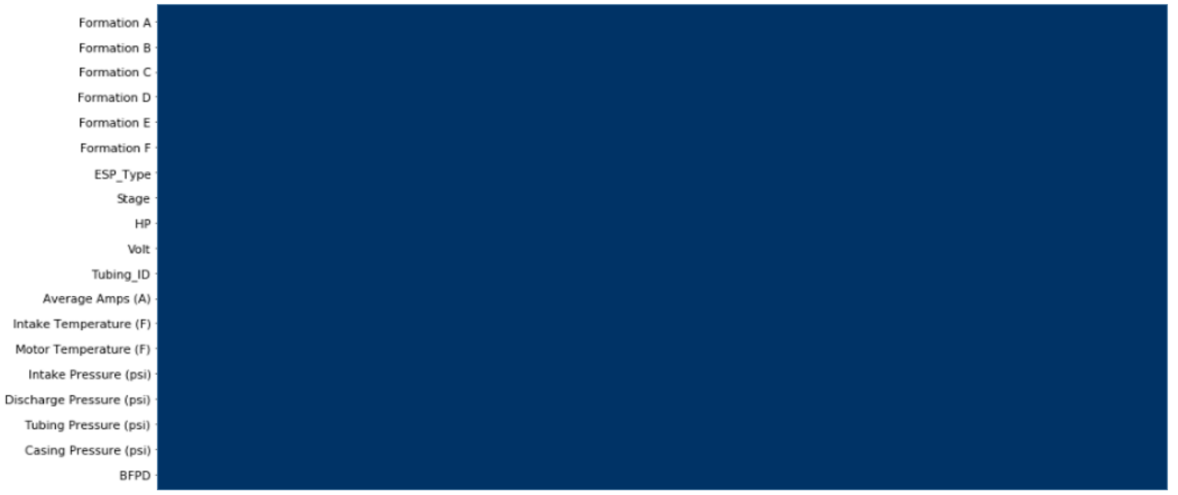


Figure 4. Missing Value Plot of Clean Data

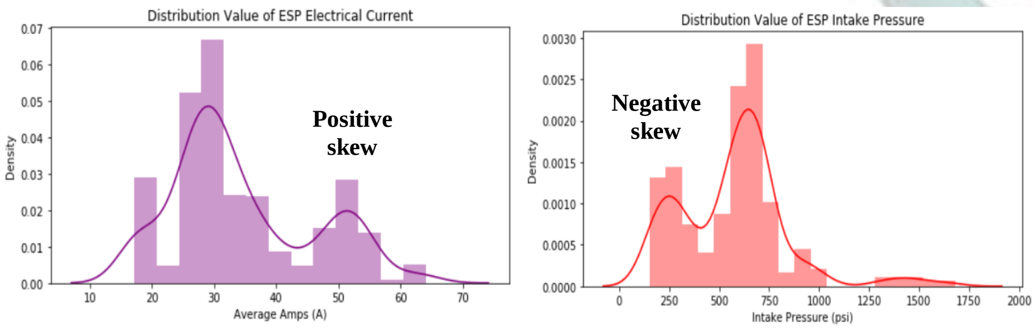


Figure 5. Variation of Data Distribution

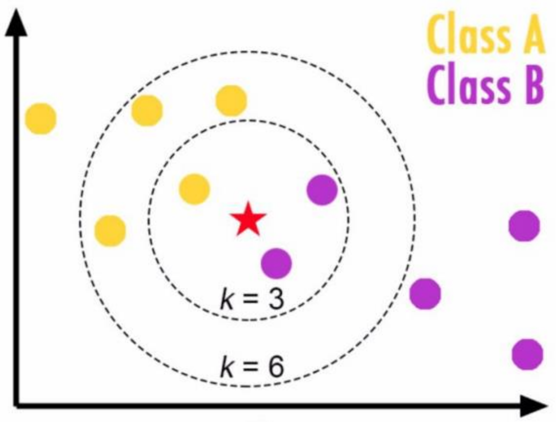


Figure 6. Example of KNN Approach (James, G., 2013).

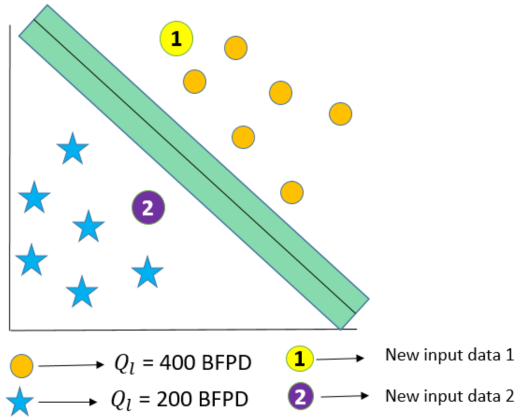


Figure 7. Illustration of line SVR.

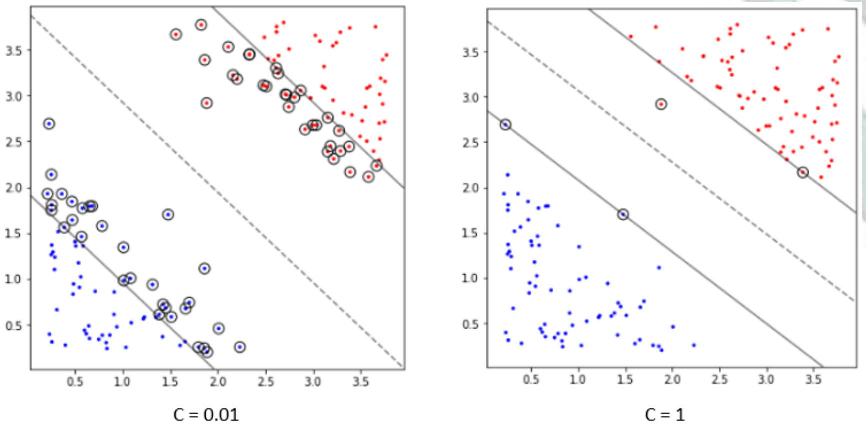
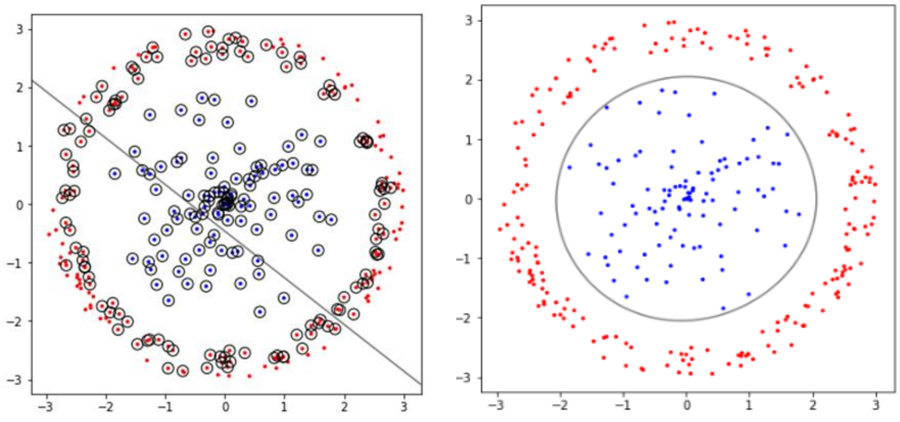


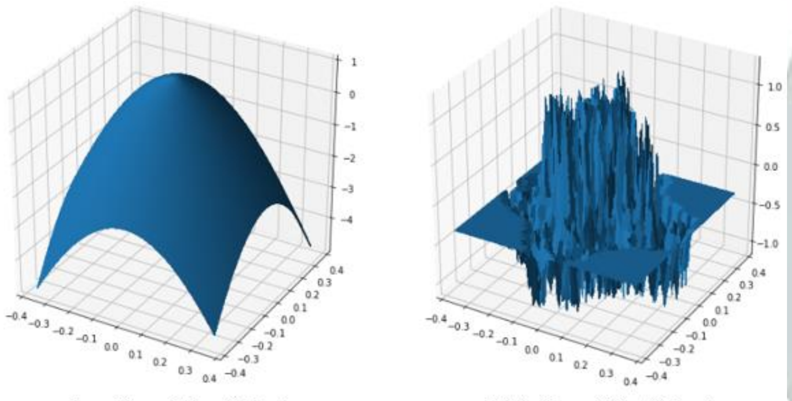
Figure 8. Penalty Parameter (C) in SVR (Putra, W.D.K., 2018).



Before

After Kernel Trick

Figure 9. Kernel Trick in SVR by Using Gaussian (Putra, W.D.K., 2018).



Low Kernel Coefficient

High Kernel Coefficient

Figure 10. Kernel Coefficient (γ) in SVR (Putra, W.D.K., 2018).

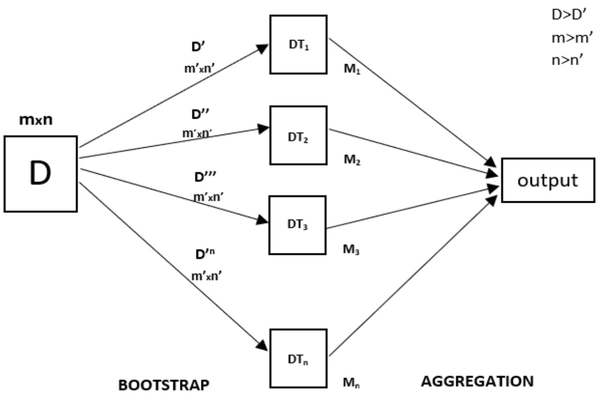


Figure 11. Random Forest Illustration (Dutta., A, 2020).

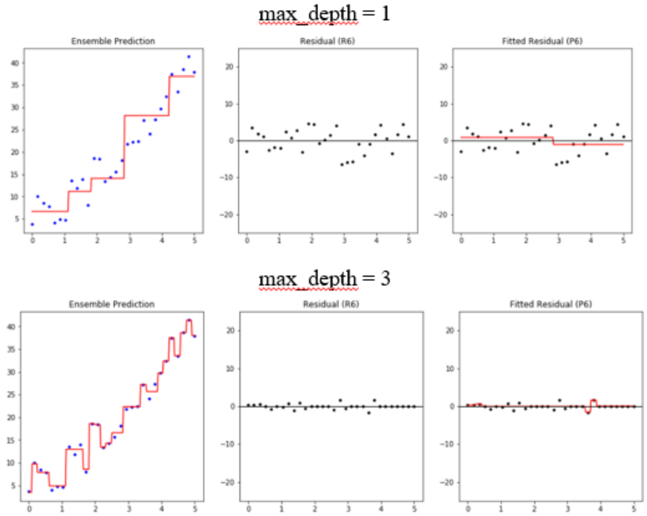


Figure 12. Maximum Depth Comparison (Putra, W.D.K., 2018).

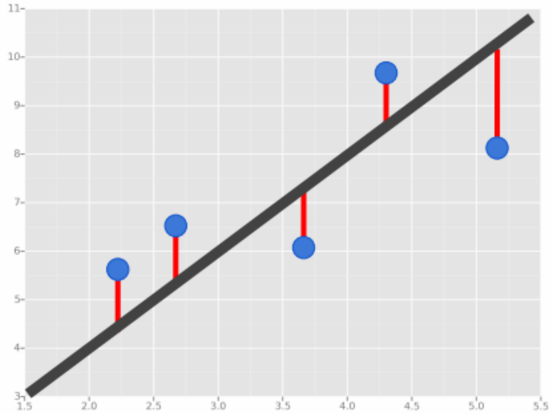


Figure 13. Example of Linear Regression Plot (James, G., 2013).

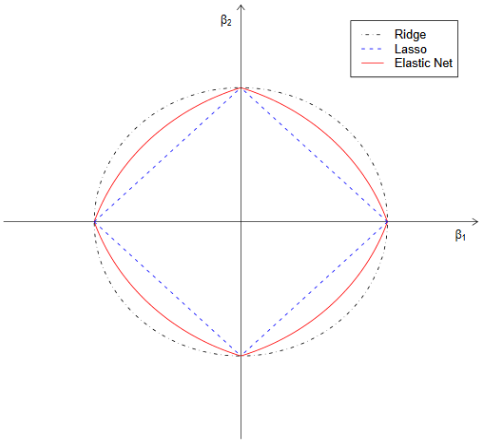
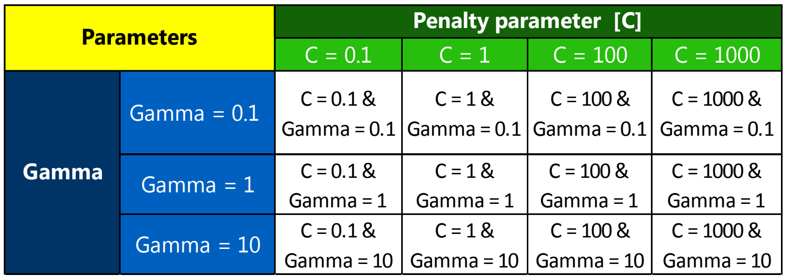


Figure 14. Illustration of Elastic Net (Hastie, T., 2004).



Parameters		Penalty parameter [C]			
		C = 0.1	C = 1	C = 100	C = 1000
Gamma	Gamma = 0.1	C = 0.1 & Gamma = 0.1	C = 1 & Gamma = 0.1	C = 100 & Gamma = 0.1	C = 1000 & Gamma = 0.1
	Gamma = 1	C = 0.1 & Gamma = 1	C = 1 & Gamma = 1	C = 100 & Gamma = 1	C = 1000 & Gamma = 1
	Gamma = 10	C = 0.1 & Gamma = 10	C = 1 & Gamma = 10	C = 100 & Gamma = 10	C = 1000 & Gamma = 10

Figure 15. Example of Grid Search CV.

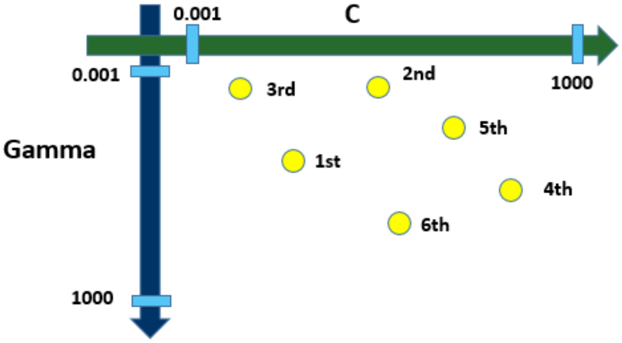


Figure 16. Example of Randomized Search CV.

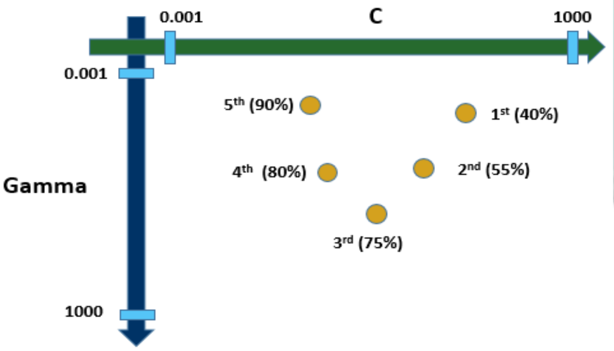


Figure 17. Example of Bayes Search CV.

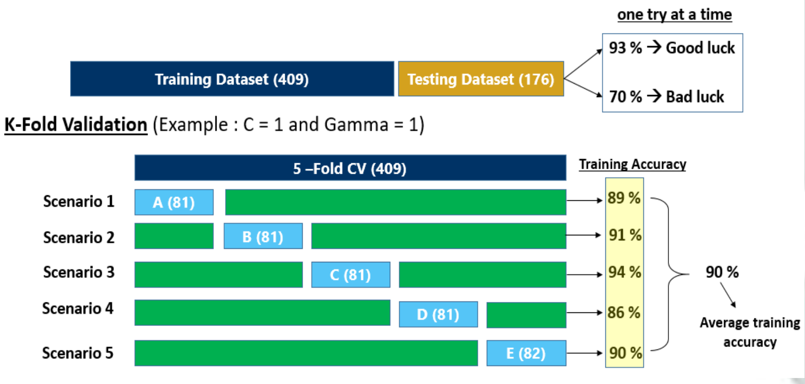
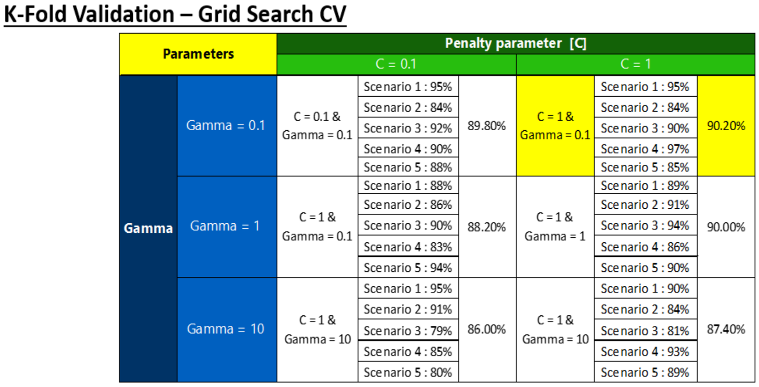


Figure 18. The illustration of K-Fold Cross Validation.



K-Fold Validation – Grid Search CV

Parameters		Penalty parameter [C]					
		C = 0.1			C = 1		
Gamma	Gamma = 0.1	C = 0.1 & Gamma = 0.1	Scenario 1 : 95% Scenario 2 : 84% Scenario 3 : 92% Scenario 4 : 90% Scenario 5 : 88%	89.80%	C = 1 & Gamma = 0.1	Scenario 1 : 95% Scenario 2 : 84% Scenario 3 : 90% Scenario 4 : 97% Scenario 5 : 85%	90.20%
	Gamma = 1	C = 1 & Gamma = 0.1	Scenario 1 : 88% Scenario 2 : 86% Scenario 3 : 90% Scenario 4 : 83% Scenario 5 : 94%	88.20%	C = 1 & Gamma = 1	Scenario 1 : 89% Scenario 2 : 91% Scenario 3 : 94% Scenario 4 : 86% Scenario 5 : 90%	90.00%
	Gamma = 10	C = 1 & Gamma = 10	Scenario 1 : 95% Scenario 2 : 91% Scenario 3 : 79% Scenario 4 : 85% Scenario 5 : 80%	86.00%	C = 1 & Gamma = 10	Scenario 1 : 90% Scenario 2 : 84% Scenario 3 : 81% Scenario 4 : 93% Scenario 5 : 89%	87.40%

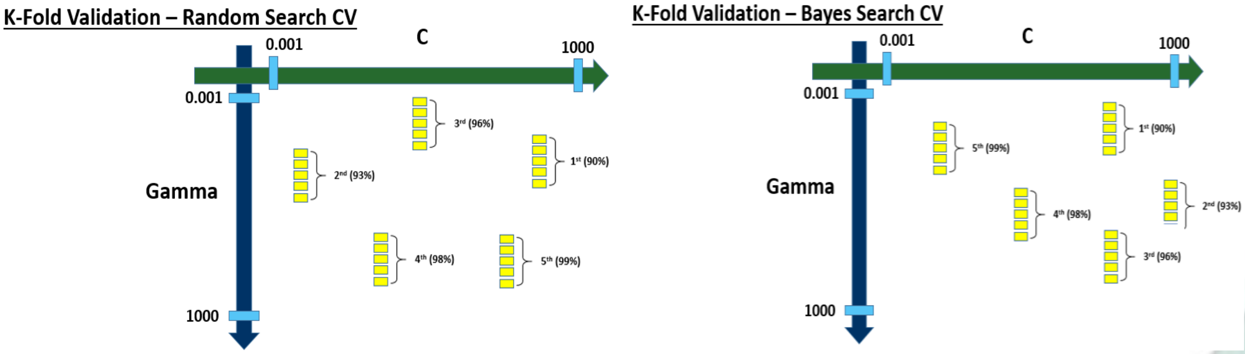


Figure 19. K-Fold Cross Validation in Grid Search CV, Random Search CV, and Bayes Search CV.

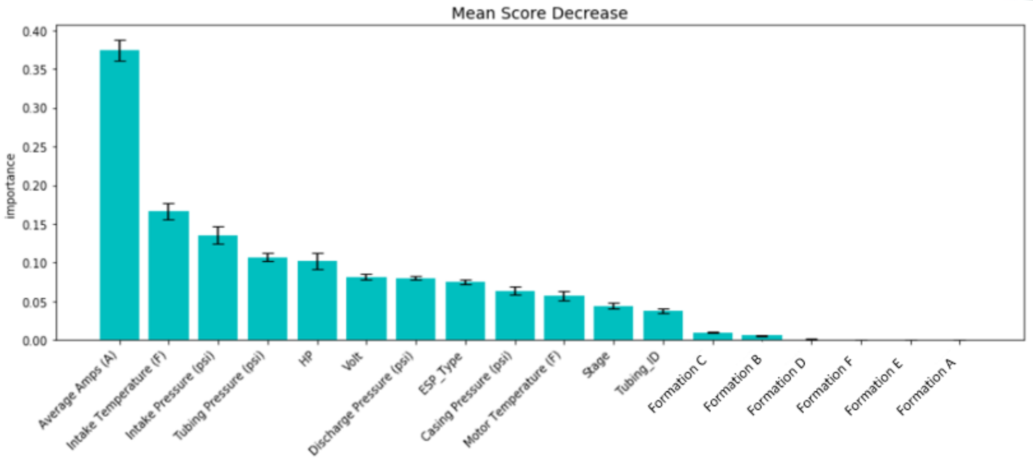
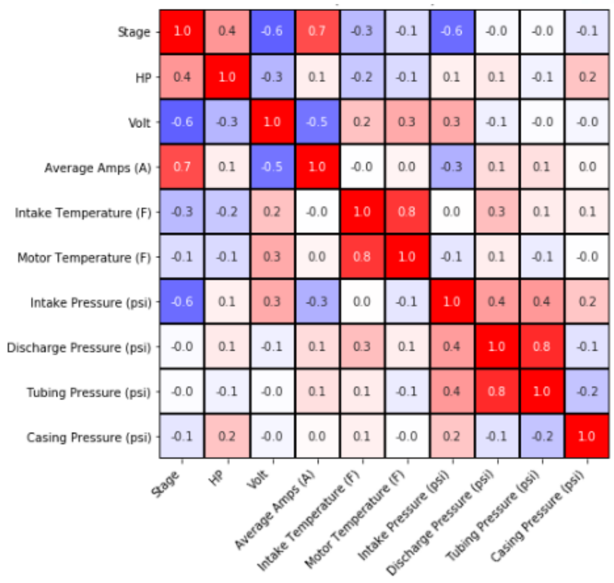


Figure 20. Feature Importance - Mean Score Decrease Plot



	Stage	HP	Volt	Average Amps (A)	Intake Temperature (F)	Motor Temperature (F)	Intake Pressure (psi)	Discharge Pressure (psi)	Tubing Pressure (psi)	Casing Pressure (psi)
Stage	1.0	0.4	-0.6	0.7	-0.3	-0.1	-0.6	-0.0	-0.0	-0.1
HP	0.4	1.0	-0.3	0.1	-0.2	-0.1	0.1	0.1	-0.1	0.2
Volt	-0.6	-0.3	1.0	-0.5	0.2	0.3	0.3	-0.1	-0.0	-0.0
Average Amps (A)	0.7	0.1	-0.5	1.0	-0.0	0.0	-0.3	0.1	0.1	0.0
Intake Temperature (F)	-0.3	-0.2	0.2	-0.0	1.0	0.8	0.0	0.3	0.1	0.1
Motor Temperature (F)	-0.1	-0.1	0.3	0.0	0.8	1.0	-0.1	0.1	-0.1	-0.0
Intake Pressure (psi)	-0.6	0.1	0.3	-0.3	0.0	-0.1	1.0	0.4	0.4	0.2
Discharge Pressure (psi)	-0.0	0.1	-0.1	0.1	0.3	0.1	0.4	1.0	0.8	-0.1
Tubing Pressure (psi)	-0.0	-0.1	-0.0	0.1	0.1	-0.1	0.4	0.8	1.0	-0.2
Casing Pressure (psi)	-0.1	0.2	-0.0	0.0	0.1	-0.0	0.2	-0.1	-0.2	1.0

Figure 21. Normalized Pearson Plot

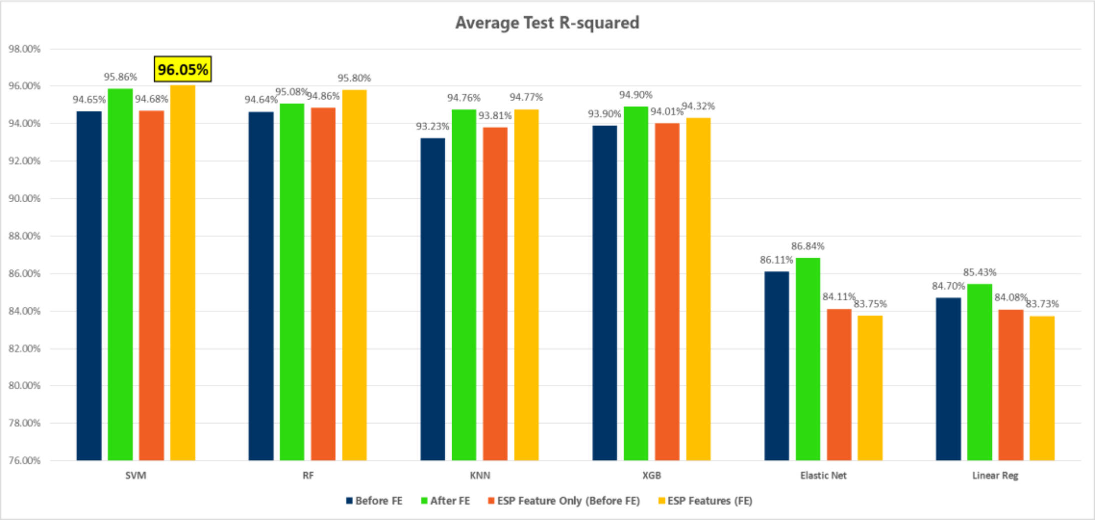


Figure 22. Average R-squared Between Model.

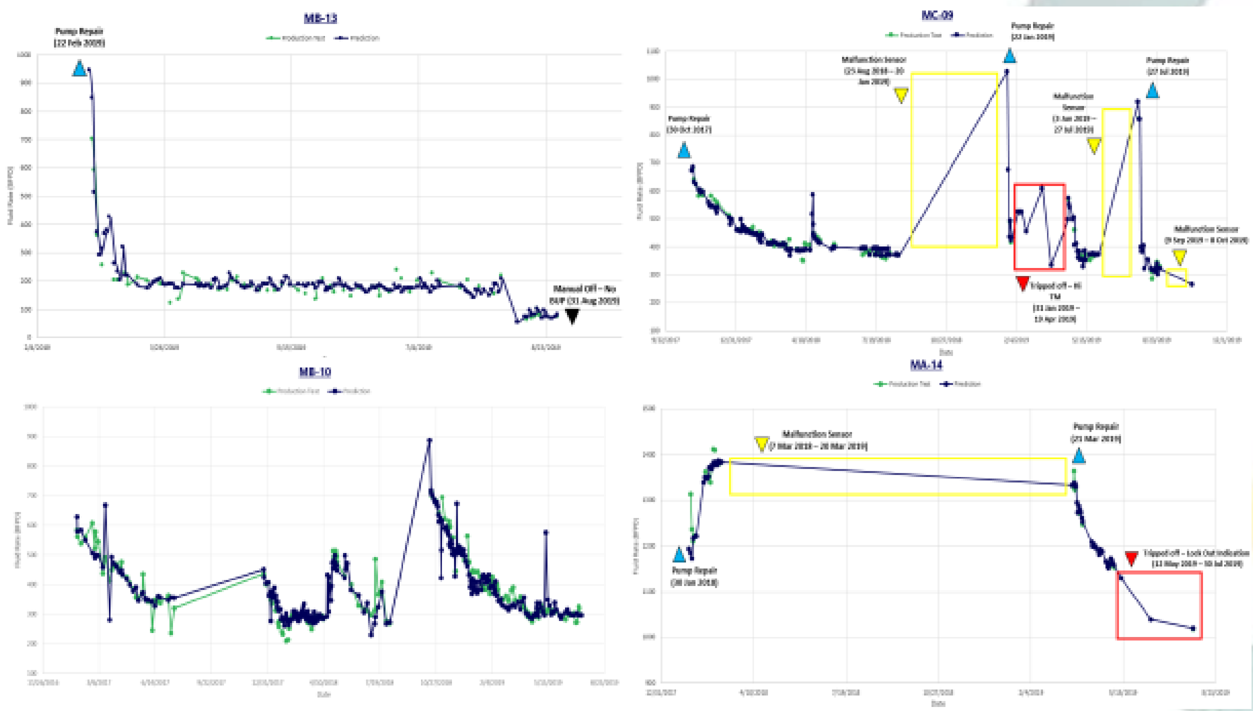


Figure 23. Virtual Flow Rate Prediction in MB - 13, MB - 10, MC - 09, and MA – 14.

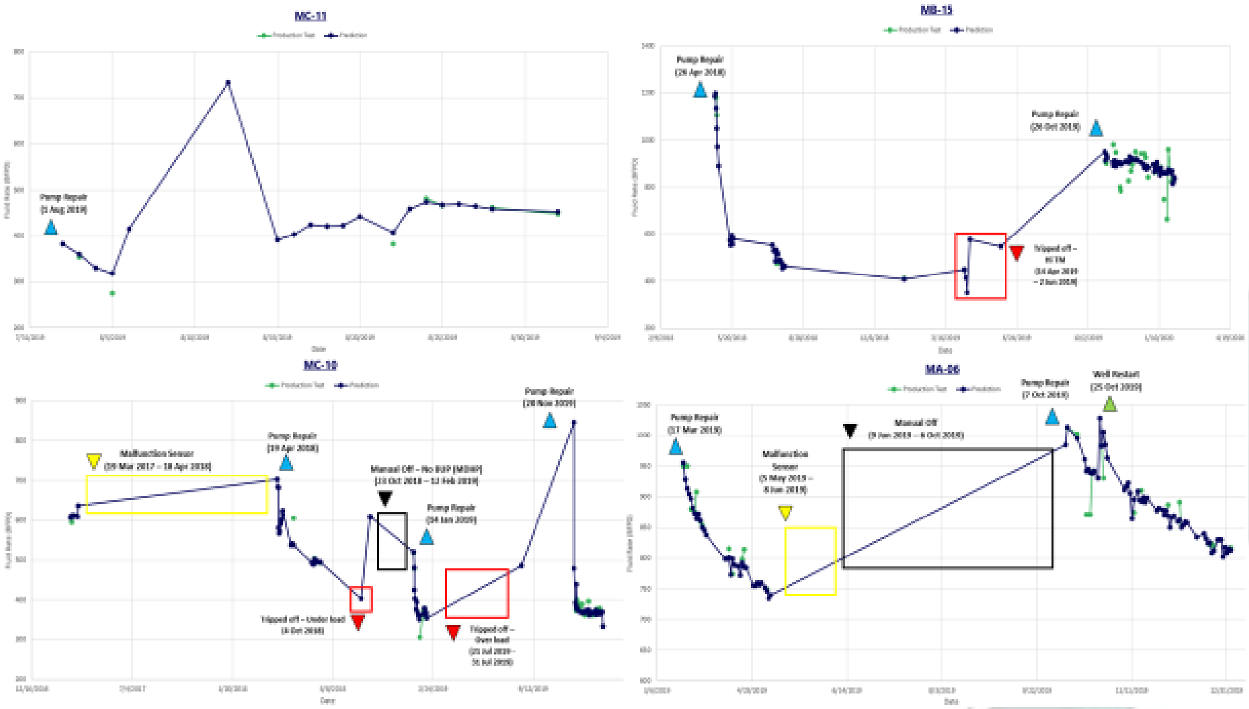


Figure 24. Virtual Flow Rate Prediction in MC - 11, MC - 10, MB - 15, and MA – 06.

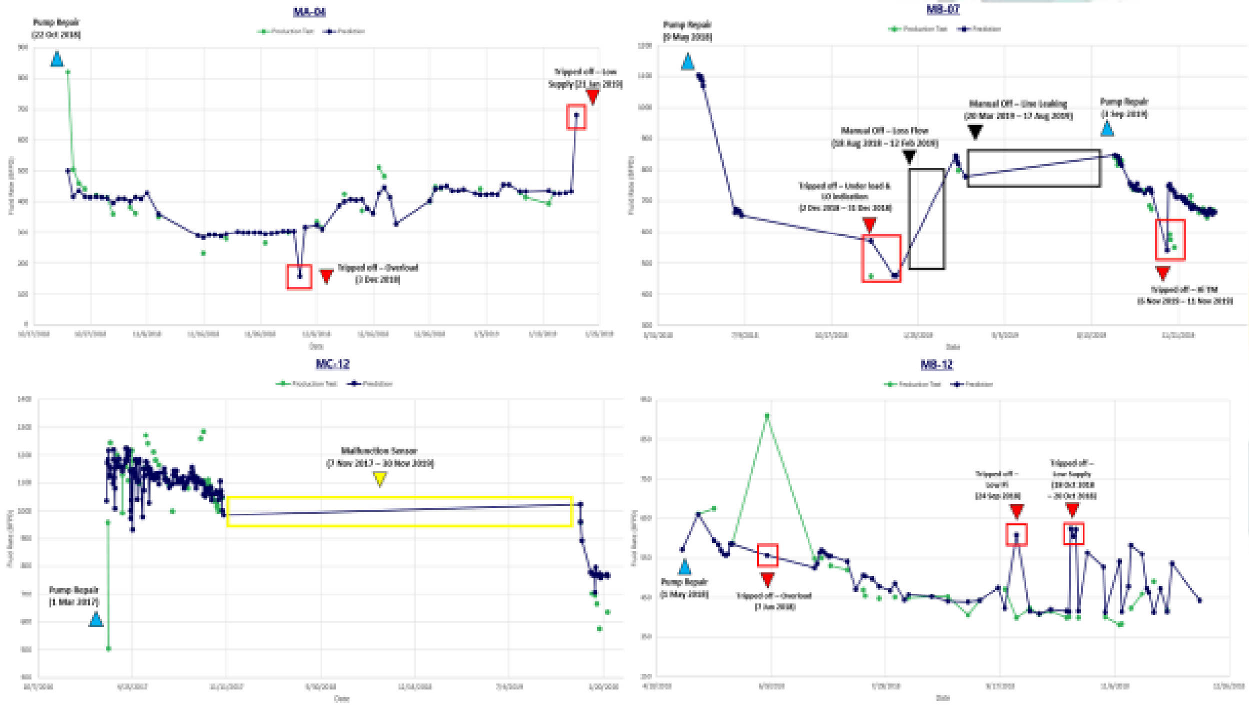


Figure 25. Virtual Flow Rate Prediction in MA - 04, MC - 12, MB - 07, and MB – 12.

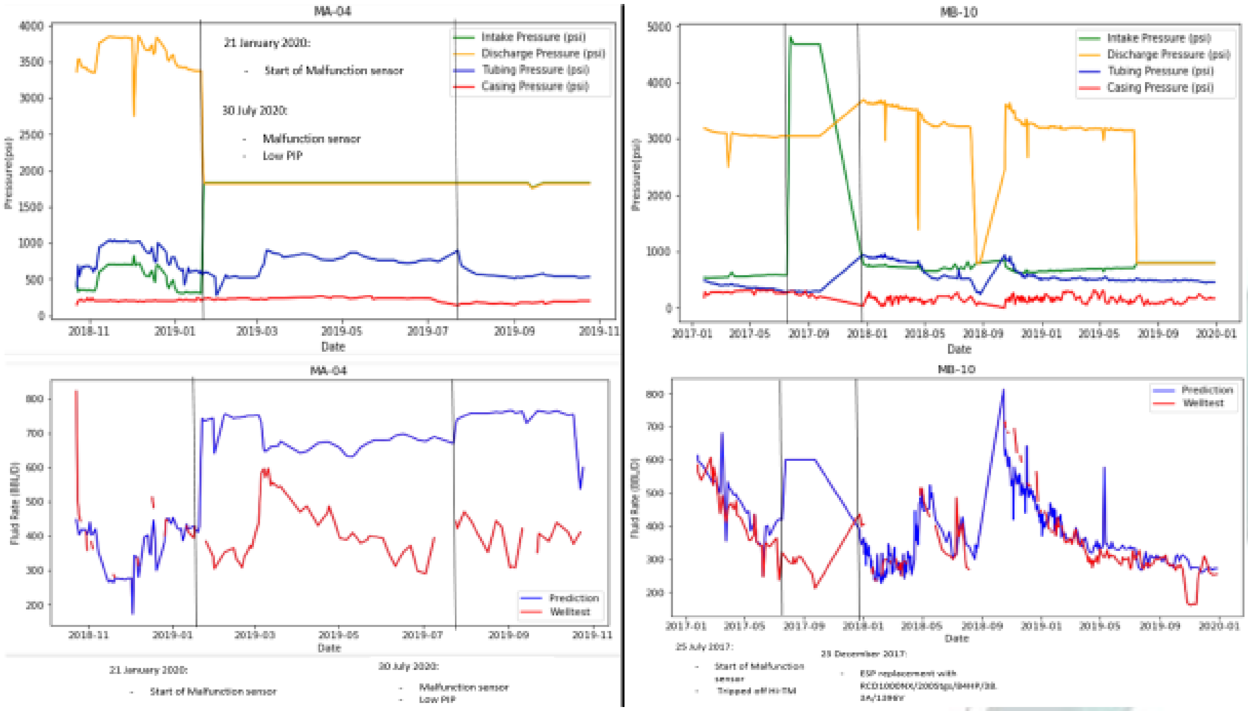


Figure 26. Failed to Predict Virtual Flow Rate due to error value in ESP sensor data