Failure Prediction in Open-hole Wireline Logging of Oil and Gas Drilling Operation

Maylada Pootisirakorn
Department of Computer Engineering
Chulalongkorn University
Bangkok, Thailand
6070958421@student.chula.ac.th

Prabhas Chongstitvatana
Department of Computer Engineering
Chulalongkorn University
Bangkok, Thailand
Prabhas.C@chula.ac.th

***Abstract*—The failure of open-hole wireline logging leads to an unexpected cost and time that add to drilling operation. The research proposes methods to predict the failure of an open hole wireline logging prior to run the log on actual situation. Three machine learning techniques are used to predict the result of the open-hole wireline logging from drilling process. The success class is the normal well that can run logging to target depth without tool sit down or stuck. Support Vector Machine (SVM), Naive Bayes and Decision Tree are chosen as proposed machine learning techniques for this task. The comparison between each method is discussed. The result of the experiment with the data shows that SVM has the highest accuracy.**

Keywords—machine learning, oil and gas, drilling, open-hole wireline logging, data analysis, support vector machine, naive bayes, decision tree

I. INTRODUCTION

Oil and gas exploration and production is a process based on various kind of data that can describe the unknown surface under the ocean to drill. The operation requires data to make plan and decision. Each of operation requires high cost with operational excellence, however the offshore operation, under the sea level has high pressure, temperature which affects to the operation.

In the past the data from oil and gas field is mostly used for descriptive and diagnostic purpose, however, the focus has been changed to do more on predictive and prescriptive. The machine learning can be applied to get insight from the big amount of data that already kept but have not been used by human. The open hole wireline logging usually run after drilling to target depth with the decision of engineers to run the logging in the next step based on their experience and a standard procedure. After drilling follows the plan, open hole wireline logging will help engineers interpret the real situation from the logs. Log data has been collected from the tools, however, sometimes the tools are stuck in the hole and could not go to the expected target depth. This leads to the loss of rig time and operation time and it could be worst if the tool lost the connection from the wire.

An unexpected cost will be added to get the tools out of the hole prior to resume to normal operation. It was found that the problems encountered cause from several possible issues such as temperature, pressure, directional surveys, formation, fluid density including the circulation. Engineers could not know exactly that the logging will success or fail until the result is found out after sending the log tools.

Machine learning can be used for predicting the failure based on historical data. Three machine learning methods are selected which are applied to the operational data that collected after completed drilling to target depth. Naïve Bayes, Decision Tree and Support Vector Machine (SVM) are techniques in machine learning that can be used to classify data. SVM is a well-known for binary classification based on statistical theory and was first introduced by Boser, Guyon and Vapnik [1] in 1992. SVM concept is to minimize an upper bound of the generalization error by maximizing the margin that separate the hyper planes (Fig. 1). SVM has successfully been applied to several applications including the oil and gas area.

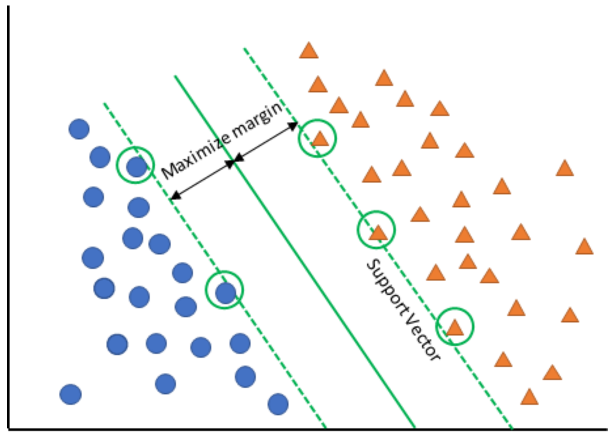

Fig. 1. Support Vector Machine that separate the hyperplane to classify the class with maximum margin

This work focuses on identifying the failure that might happens before starting the open-hole wireline logging job. It uses a well data in Gulf of Thailand. Since data set has a small rate of failure which leads to the imbalance data, the collected data is divided into two groups according to the logging tools. The results are evaluated in terms of precision and recall of both classes. A special emphasize is placed on improving the accuracy of the prediction of failure class.

To choose the best performance which are the most effective, an attribute selection was used to remove the unrelated attributes. The prediction performance of the three machine learning techniques are compared: Naïve Bayes, Decision Tree and Support Vector Machine.

II. RELATED WORKS

There have been increasing research activities related to apply machine learning techniques to predict results or detect errors in the field of oil and gas industry and to get insight and help in making decision. Following work applied the machine learning techniques to their applications (Table I).

The work in [2] uses SVM to select the location of wells. The regularization parameters were determined using grid search to prevent overfitting. The SVM model was trained to rank the locations based on their production capabilities and historical of reservoir data and completion data. The trained model helps the asset team to make data driven decisions.

TABLE I. MACHINE LEARNING TECHNIQUES APPLIED IN OIL AND GAS

Application	ML Technique	Data Set	Year
Selection Infill Location	SVM, K-Means Clustering	Reservoir data, Oil -production rate, and completion data.	2018
Pump Failure Prediction	SVM	Electrical and frequency data from the field	2015
Stuck Pipe Prediction	ANN, SVM	Mud logging and well information	2012
Classification of Well Drilling Operations	SVM	Drilling well information	2006
Risk Level of Lost Circulation	SVM, Random Forest, ANN	Mud logging and well information	2018

The report by [3] presents a data driven approach for failure prediction of the submersible pump system that used in oil and gas industry. The team uses SVM to train a prediction model with the selected features and test it on real world data. The data were collected by sensors based on electrical properties and frequency and other information such as logs of events. They are fed to the machine learning framework to predict the failure of the pump. A timely diagnosis of failure from the model can improve the production performance.

The prediction of stuck pipe in oil and gas industry is done by Islam et al [4]. Their work focuses on using artificial neural networks (ANN) and support vector machine to predict the stuck pipe before it occurs. It is one of the most costly problem. The study classifies stick pipe incidents into two groups as stuck or non-stuck. The SVM can predict stuck pipe occurrences with accuracy over 85%. The report claims that SVM is more convenient than ANN since it needs fewer parameters to be optimized. The model generally works well in the selected area of the operation but may not work in other areas. Previously Siruvuri et al [5] use ANN to predict stick pipe. To combat the erroneous and incomplete data because of the data collection process, the reasonable outputs were generated even the data might have some errors.

SVM has been used to classify petroleum well drilling operations in Adriane et al [6]. Their work presents the development of a system that intends to make better use of the information collected during well drilling operation. The main idea is to amass a great amount of historical data that has not been properly used and hope to extract insight. They use SVM for pattern recognition and develop the automatic classification system that improves the prediction performance. The report presents 6 types of multi-class SVM with the various kernel functions: the gaussian RBF, polynomial and linear functions. The simple linear SVM has a good generalization accuracy with correctness of 92%.

The well problem is one of the most interesting issue that need focus on. Zejun Li et al [7] study three typical machine learning algorithms and analyze drilling data in Iraq to predict the lost circulation issue. They compare three techniques: SVM, ANN and random forest. SVM and random forest have predicted correctly 99% of wells with normal class. However the data is imbalance. Only 55% of the lost circulation samples are correctly classified. The accuracy to identify lost circulation points is not ideal, partly because they occur relatively sparse, and the data becomes imbalance compared to the normal class.

III. OIL AND GAS BACKGROUND

A. Offshore Drilling Operation

Offshore drilling operation [8] refers to the process to drill through soil and rock under the seafloor to create a well which is bored hole that can access to geological reservoirs contained with oil and gas. The development or production wells are drilled to recover oil and gas reserves in the proven economic areas.

The process of drilling oil and gas well involves several steps (Fig. 2):

1. A well is drilled using drill bit and pipe to create a bore hole under the seafloor. The drilling path could not be drilled directly to hit the oil and gas reservoir otherwise it would be blow out or explode before doing the completion and production. It is done by boring a vertical depth with angled to the target reservoir.
2. The circulation process in the hole using mud to circulate and remove the rock cuttings from the hole and maintain the working temperature and pressures of the well.
3. Cementing requires on each section after drill to the planned depth. This is applied to the bore hole to prevent collapse. There are mainly three sections of the well in Gulf of Thailand.

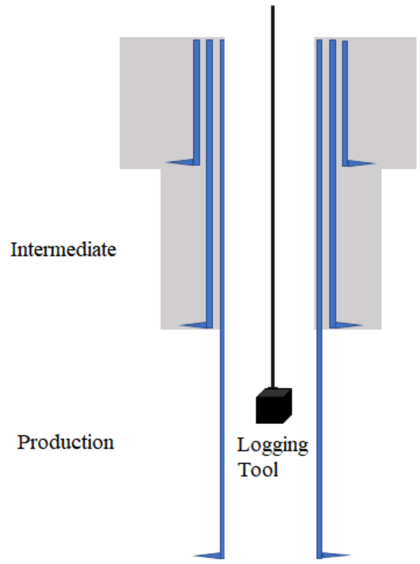

Fig. 2. Wireline logging tool in bored hole or open hole

4. Once the well is drilled to the target depth at production section on bore hole or called open hole before cementing, it usually has the open hole wireline logging or formation test after pulling the drill bit out of the hole. Wireline logging is the process to collect data using the electric instruments to continuously measure the properties of a formation, this data can help making decisions in drilling and production operation.

B. Open-hole Wireline Logging

To drill the well, it is a technological process, however no wells are identical. They vary because of the risk due to the temperature and pressure is increased when drilled to the deeper hole. The information of the subsurface around the hole can be acquired from the electronic logs which represent important source of data to geophysicist and engineers to analyze and explore the rock information and the reservoir target which can produce the oil and gas.

The open hole logging activities [9] are one of a large investment that oil and gas company made to acquire the data. It is important if we can reduce the cost of these activities and ensure that they would not be failed and reduce the non-productive time of the drilling operation. The type of logging [10] to the open hole and objective of the data acquisition are the two main questions that need to be verified prior to start the logging operation. ata gathering would help expert to interpret and making consideration using statistical skill about well integrity and reservoir characteristics.

There are two types of logging in this study:

1. Formation Tester is mainly used for collecting the pressure point for specific depth, the different subtype depends on temperature and the service company.
2. Quad Combo, Triple Combo are used primarily to identify lithology, reservoir porosity and fluid type in formation.

However, it is not a rule for data acquisition of the logging operation of every well. The number can be reduced depends on the hole condition since the logging through casing options still exist. Data that could be get from the open hole are such as assessment of source rock potential, hole volume and shape estimates, sample of lithology, location of hydrocarbon, reservoir capacity assessment, porosity and pressure measurements. Open hole logging operation will be executed after drilled to the production section, there are 6 hours before the operation happen. Actual of parameters after drilled can be used to analyze and make decision prior to run the logging.

IV. RESEARCH EXPIRIMENTS

A. Data Gathering and Preparation Process

To predict the result of open-hole wireline logging, the historical data is used to let the machine learns from the actual situation that occurs. Actual drilling parameters on production section, directional survey, the inclination and dogleg information are captured. Logging tool type and length including number of tight spots in the hole and temperature at the bottom of the hole are also the attributes that are collected. The scope of data in this experiment is based on Gulf of Thailand.

Data from 2014 to 2018 has been gathered from multiple sources and processed prior to be used in this research. There are 1439 records of real cases of historical logging data that had been reviewed by subject matter experts. All of data needs to be grouped into individual wells based on logging tool that being used at that time. The records are labelled with the class of success or failure logging result. The original data is scattered in different database and in the excel spreadsheets. It needs to be collected and compiled prior to preprocessing in the next step. Data from database requires scripting to pull data appropriately from various data formats (Fig. 3).

The success class indicates that there is no stuck of the tool in the hole. The failure class represents the hole with bad conditions that caused the tools stuck in the hole and could not reach to target depth or could not bring the tool out of hole.

Input Data Sources	**Output Label**
Drilling Well Problem Drilling Parameter Logging Tool Information Mud and Fluid information Directional Survey Information	Logging Result Report

Fig. 3. Input data sources and label data for open-hole wireline logging prediction

B. Attribute Selection

In practice, the unrelated attributes in the input data set can confuse the machine learning [11] such as SVM. The attribute selection against the whole data set has been applied. Fig. 4 shows the format of data after preprocessing.

	1	**2**	**3**	**.**	**.**	**.**	**32**	**Class**
Logging#1								Success
Logging#2								Fail
Logging#3								Success

Fig. 4. Data table after processed data with class labeled.

The features selection step is done by the experts which is based on their experience and the major factors that can captured by the attribute. Total of 32 attributes were selected to be used as the input in this experiment and the unrelated attributes were removed.

The following are the list of selected features: Field, Well classification, Rig name, Measure depth, True vertical depth, Subsea true vertical depth, Previous section measure depth, Previous section true vertical depth, Previous section subsea vertical depth, Surface section depth, Intermediate section depth, Tubing depth, Mud weight at intermediate section, Mud weight at tubing section, Wireline logging company, Tool, Well status, Mud weight, Open-hole length, Max degree at tubing, Tubing depth with max degree, Max dogleg at tubing, Tubing depth with max dogleg, Max degree at intermediate, Intermediate depth with max degree, Max dogleg at intermediate, Intermediate depth with max dogleg, Inclination at 7" casing shoe, Type of log, Tool length, Max temperature, Mud logging company, Number of tight spot. Data Separation.

C. Data Separation

Data set is divided into two groups according to logging tool. There are two main type of tools, one is a 'Formation Tester' and another is 'Quad Combo or Triple Combo'. The major equipment and its setting are different between these two types and they are being used for different purpose.

After separation of data into two groups, the well data of each logging type is split 70% for training using cross-validation and 30% of testing (Fig. 5). The number of samples in each group are shown in Table II.

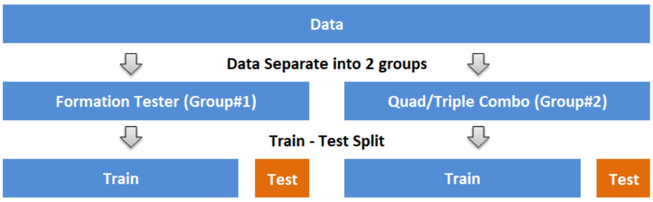

Fig. 5. Data separation process

TABLE II. NUMBER OF SAMPLE OF DATA SET SEPARATED BY GROUP OF TOOL TYPE

Group	Tool Type	Testing Set		Total	Training Set		Total
		Success	*Fail*		*Success*	*Fail*	
#1	Formation Tester	291	71	362	680	167	847
#2	Quad / Triple Combo	49	20	69	116	45	161

D. Training and Validation

The result of training process depends on choosing an efficient method for data partitioning. In practical terms one-third of data is used for testing and the remaining data is used for training. The hold-out method called k-fold cross-validation is an important statistical technique that was applied. In cross-validation we selected 5-fold since the total size of data set is not large. It gives the best estimate of misclassification rate error. In 5-fold cross-validation, the whole data is randomly separated into five equal partitions, each part is held out to be tested and the trained on the remaining four.

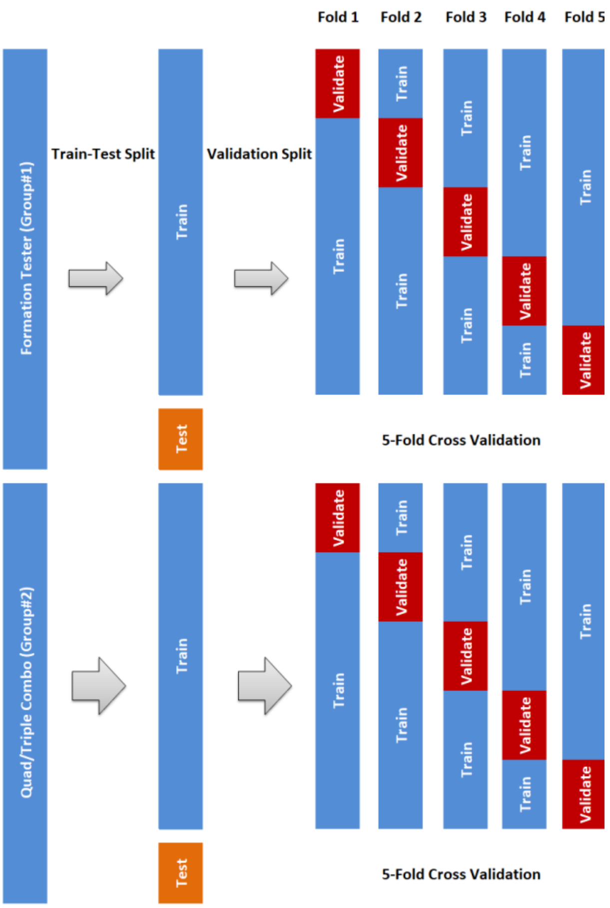

Fig. 6. Cross validation using 5-Fold

The group of data were trained separately with 5-fold cross-validation (Fig. 6). Three machine learning methods are selected for this experiment: SVM, Naïve Bays and Decision Tree.

E. Performance Measurement

Confusion matrix [12] is a summary of the performance of a classification model on a test data set. It is used in evaluation of the classification performance. The number of correct and incorrect predictions are summarized for each class. It gives an insight into the errors from the model and moreover the types of errors are also important besides the classification accuracy.

TP (true positives) referred to the number of correctly predicted success samples. TN (true negatives) referred to the number of correctly predicted fail samples (Fig. 7).

The performance measurement can be used to evaluate the model as below:

- Accuracy measures overall accuracy of the model classification

$$Accuracy = \frac{all\ correct}{all} \quad \frac{TP+TN}{TP+FN+FP+TN} \tag{1}$$

- ROC curve [13] is a plot of values of the False Positive Rate (FPR) versus the True Positive Rate (TPR)

$$True\ Positive\ Rate = \frac{TP}{TP+FN} \tag{2}$$

$$False\ Positive\ Rate = \frac{FP}{FP+TN} \tag{3}$$

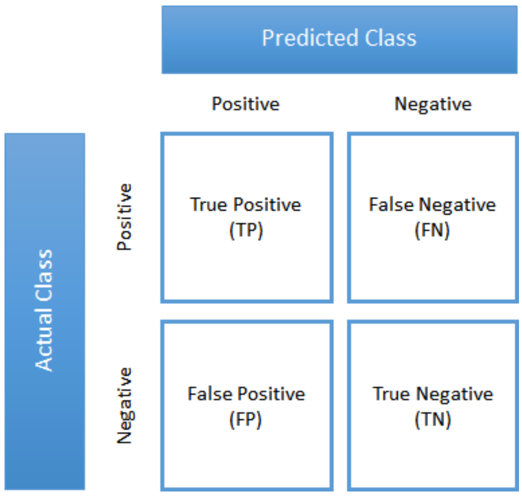

Fig. 7. Confusion matrix as a table summary for a binary classification

In this experiment, the number of failure sample is much less than the number of success sample. This introduces imbalance issue to the training process. We also focus on FP (false positives) which referred to the number of predicted as success, but the actual is failed. The FP should be small since it can cause the additional time and cost if the open-hole wireline logging is failed.

The FN (false negatives) referred to the number of predicted as fail but the actual is success. In this case, it is acceptable since engineers will focus on maintaining the tools and other peripherals to avoid the failure, or perhaps decide not to run the open-hole wireline logging.

V. EXERIMENT RESULT

The results of classification with SVM, Naïve Bayes and Decision Tree are reported in Table III. An accuracy and ROC analysis were calculated as performance measurement. The ROC curves are shown Fig 8, Fig 9 and Fig 10.

TABLE III. CLASSIFICATION RESULT FROM TESTING DATA SET

ML Technique	Group	Test set	Success	Fail	TP	FP	TN	FN
SVM	#1	362	291	71	285	13	58	6
	#2	69	49	20	48	9	11	1
Naïve Bayes	#1	362	291	71	163	12	59	128
	#2	69	49	20	33	2	18	16
Decision Tree	#1	362	291	71	291	71	0	0
	#2	69	49	20	49	3	17	0

The ROC curve gives us a clear picture on the performance measurement. We focus on FP which means to avoid predicting success on actual failure. From the comparison of the ROC curve, we see that the decision tree is not good on prediction for Group 1. It cannot identify the fail case due to the overfit from the training samples. But for the Group 2, the decision tree is the best method.

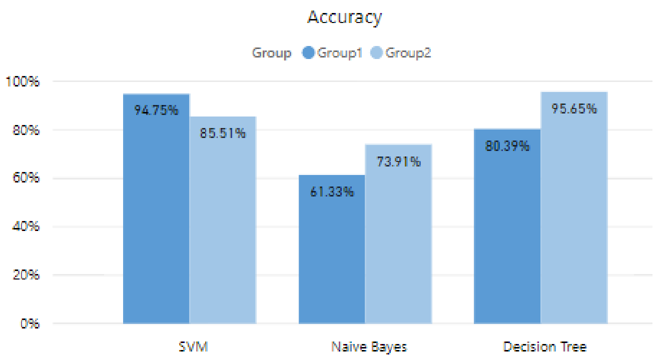

Fig. 8. Accuracy evaluation matric using testing data set

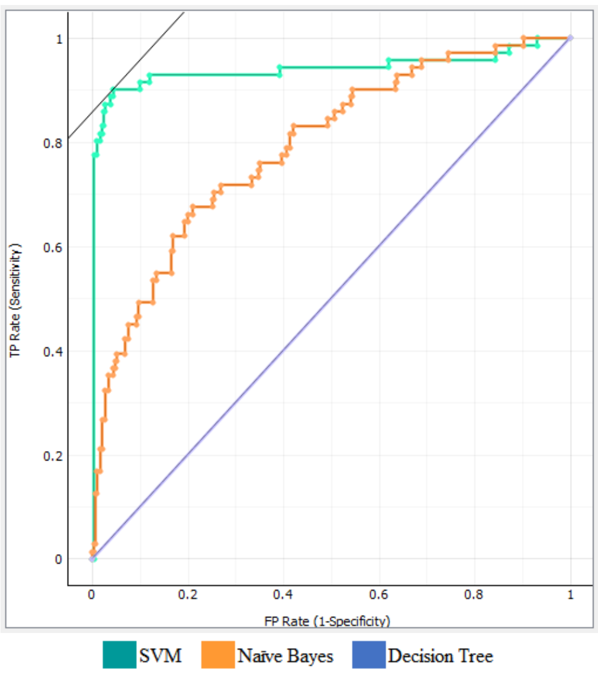

Fig. 9. ROC curve of Formation Tester (Group 1)

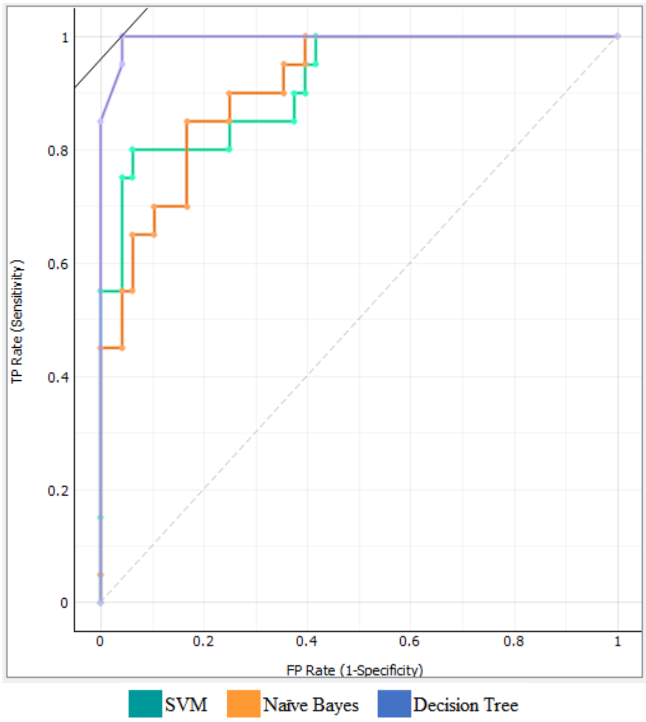

Fig. 10. ROC curve of Quad/Triple Combo (Group 2)

VI. CONCLUSION

A. *Results summary*

From the experiment of Formation tester data set (Group1), SVM has the highest accuracy, precision and recall. Naïve Bayes has least performance for both two data set. For the data set of Quad/Triple combo (Group2), the prediction results were not as good as SVM. The decision tree is more accurate. Nevertheless, the small size of data set decreases statistical power.

TABLE IV. ACCURACY, PRECISION AND RECALL RESULTS

ML Technique	Group	Accuracy	Precision	Recall
SVM	#1	94.75%	95.64%	97.94%
	#2	85.51%	84.21%	97.96%
Naïve Bayes	#1	61.33%	93.14%	56.01%
	#2	73.91%	94.29%	67.35%
Decision Tree	#1	80.39%	80.39%	100.00%
	#2	97.06%	96.08%	100.00%

The accuracy may not be the focus point since we avoid the case that predicts a success, but in fact it is fail. The aim is to reduce FP and FN. SVM gives a good performance for both data set.

From the experiment, the ranking of important features which are captured after running the model is listed below (Fig. 11). It is based on information gain from the expected amount of information.

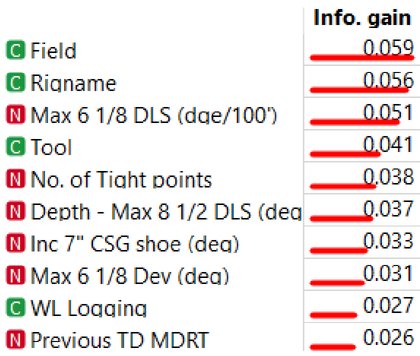

Fig. 11. Top 10 ranking of important features

B. Problem and limitations

The reason that we decide not to re-sampling data in spite of imbalance data is that since the in real world data of the underground subsurface is full of uncertainty we do not want to take risk with the re-sampling data as it may be bias by some re-sampling data. It can lead to misunderstanding. However, the performance from small data set may not be reliable as discussed.

Beside the lack of data, we will need to collect more data records of open-hole wireline logging which will help improving the result from machine learning algorithm.

The missing data and incorrect data required the preparation process which take a lot of efforts to clean up data. This would need knowledge of understanding the data and business process. The good quality data is recommended and the issue of collecting of good data had been in discussion with the experts to let them know and understand about the problem of good quality of raw data.

C. Benefit and contributions

The benefit is the time saving of each well if there is an unexpected non-productive time from open-hole wireline logging stuck (failed). It could save average 2 hours per well from the statistical of 20% failure cases per year (Fig. 12).

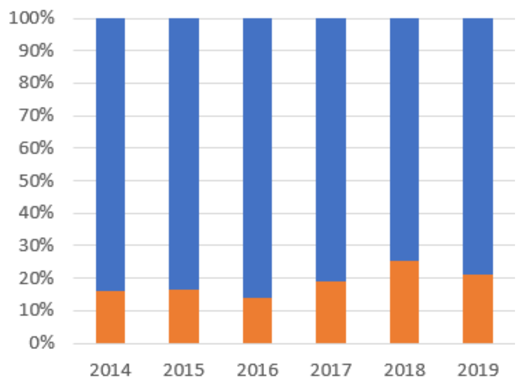

Fig. 12. The percentage of failure by year

ACKNOWLEDGMENT

The work is supported by the oil and gas company in Thailand. The authors would like to thank to the drilling department for providing the dataset of open-hole wireline logging over the Gulf of Thailand.

REFERENCES

[1] Cortes, C. and V. Vapnik, *Support-Vector Networks.* Machine Learning, 1995. **20**(3): p. 273-297.

[2] Wilson, A., *Machine Learning Overcomes Challenges of Selecting Locations for Infill Wells.* Journal of Petroleum Technology, 2018. **70**(10): p. 48-49.

[3] Guo, D., et al., *Data Driven Approach to Failure Prediction for Electrical Submersible Pump Systems*, in *SPE Western Regional Meeting*. 2015, Society

[4] Heinze, L. and I.A. Al-Baiyat, *Implementing Artificial Neural Networks and Support Vector Machines in Stuck Pipe Prediction*, in *SPE Kuwait*

[5] Siruvuri, C., S. Nagarakanti, and R. Samuel, *Stuck Pipe Prediction and Avoidance: A Convolutional Neural Network Approach*, in *IADC/SPE Drilling Conference*. 2006, Society of Petroleum Engineers: Miami, Florida, USA. p. 6.

[6] Serapiao, A.B.S., et al. *Classification of Petroleum Well Drilling Operations Using Support Vector Machine (SVM).* in *2006 International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce (CIMCA'06)*. 2006.

[7] Li, Z., et al., Study on Intelligent Prediction for Risk Level of Lost Circulation While Drilling Based on Machine Learning, in 52nd U.S. Rock Mechanics/Geomechanics Symposium. 2018, American Rock Mechanics Association: Seattle, Washington. p. 8.

[8] J.J. Azar, G.R.S., *Drilling Engineering*. 2007: PennWell Books.

[9] Kleinberg, R., *Well logging overview*. Vol. 13. 2001. 342-343.

[10] Darling, T., *Well logging and Formation Evaluation*. 2005: Gulf Professional Publishing.

[11] Blum, A.L. and P. Langley, *Selection of relevant features and examples in machine learning.* Artificial Intelligence, 1997. **97**(1): p. 245-271.

[12] Ting, K.M., *Confusion Matrix*, in *Encyclopedia of Machine Learning and Data Mining*, C. Sammut and G.I. Webb, Editors. 2017, Springer US: Boston, MA. p. 260-260.

[13] Bradley, A.P., *The use of the area under the ROC curve in the evaluation of machine learning algorithms %J Pattern Recogn.* 1997. **30**(7): p. 1145-1159.