Failure prediction in open-hole wireline logging of oil and gas drilling operation using support vector machine.

Miss Maylada Pootisirakorn

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2018

การทำนายความผิดปกติของการหยั่งเชิงธรณีของหลุมขุดเจาะน้ำมันและแก๊สโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน

น.ส.เมลดา พุฒิศิรกร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2561

Thesis Title	Failure prediction in open-hole wireline logging of oil and gas drilling operation using support vector machine.
By	Miss Maylada Pootisirakorn
Field of Study	Computer Science
Thesis Advisor	Professor PRABHAS CHONGSTITVATANA, Ph.D.

Accepted by the Faculty of Engineering, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Science

.... Dean of the Faculty of Engineering
(Associate Professor SUPOT TEACHAVORASINSKUN, D.Eng.)

THESIS COMMITTEE

.... Chairman
(Assistant Professor SUKREE SINTHUPINYO, Ph.D.)

.... Thesis Advisor
(Professor PRABHAS CHONGSTITVATANA, Ph.D.)

.... External Examiner
(Associate Professor Worasait Suwannik, Ph.D.)

เมลดา พุฒิศิรกร : การทำนายความผิดปกติของการหยั่งเชิงธรณีของหลุมขุดเจาะน้ำมันและแก๊สโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน. (Failure prediction in open-hole wireline logging of oil and gas drilling operation using support vector machine.) อ.ที่ปรึกษาหลัก : ศ. ดร. ประภาส จงสถิตย์วัฒนา

ความผิดปกติของกระบวนการหยั่งเชิงธรณีของการขุดเจาะน้ำมันและก๊าซธรรมชาตินำไปสู่ความเสียหายทั้งในด้านเวลาที่เพิ่มขึ้นและจำนวนเงินที่ต้องใช้มากขึ้นด้วย การวิจัยนี้ได้นำเสนอวิธีการที่จะสามารถทำนายความผิดปกติของการหยั่งเชิงธรณีของหลุมขุดเจาะก่อนที่จะปฏิบัติการจริง ซัพพอร์ตเวกเตอร์แมชชีนเป็นหนึ่งในการเรียนรู้ของเครื่องที่สามารถใช้จำแนกผลได้ ผลของการหยังเชิงธรณีสามารถจำแนกได้เป็นสองกรณีคือ กรณีที่สำเร็จ และกรณีที่ล้มเหลวซึ่งกรณีที่ล้มเหลวนั้นคือกรณีที่ผลลัพธ์เกิดการติดขัดของอุปกรณ์ในหลุมระหว่างการหยั่งเชิงธรณีซึ่งต้องการหลีกเลี่ยงไม่ให้เกิดขึ้น งานวิจัยนี้ได้มุ่งเน้นไปที่การเรียนรู้ของเครื่องโดยใช้ซัพพอร์ตเวกเตอร์แมชชีนเพื่อทำนายว่าผลลัพธ์ที่ได้จากการหยั่งเชิงธรณีนี้จะสำเร็จหรือมีความผิดปกติเกิดขึ้นได้ ทั้งนี้เพื่อช่วยในการตัดสินใจ รวมทั้งช่วยประหยัดเวลาและค่าใช้จ่ายที่อาจเกิดขึ้นได้อีกด้วย การจำแนกผลลัพธ์จากซัพพอร์ตเวกเตอร์แมชชีนจากการวิจัยนี้จะมีการเปรียบเทียบกับทฤษฎีเบย์อย่างง่ายและต้นไม้ตัดสินใจ

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อนิสิต ....

ปีการศึกษา 2561 ลายมือชื่อ อ.ที่ปรึกษาหลัก ....

6070958421 : MAJOR COMPUTER SCIENCE
KEYWORD: machine learning, oil and gas, drilling operation, open-hole wireline logging, data analysis, support vector machine

Maylada Pootisirakorn : Failure prediction in open-hole wireline logging of oil and gas drilling operation using support vector machine.. Advisor: Prof. PRABHAS CHONGSTITVATANA, Ph.D.

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 classify the result of the open-hole wireline logging from drilling process into two classes, a success class and a failure class which represents a well that might have abnormal conditions which can causes the tool stuck during logging. The success class is the normal well that can run logging to target depth without tool sit down or stuck. Support Vector Machine, Naive Bayes and Decision Tree are chosen as proposed machine learning techniques for this classification.

Field of Study: Computer Science Student's Signature ....
Academic Year: 2018 Advisor's Signature ....

ACKNOWLEDGEMENTS

This thesis would have not been complete without all the supports from colleges, friends and especially my advisor Prof. Prabhas Chongstitvatnan, Ph.D., he is always open whenever I had problem or a question about my research or writing. He gave advise in the right direction and push me through all situations.

Also like to thank the experts Khun Panu Boonwattanopas, Khun Ravis Sampoonachot, Khun Wiwut Saksurakan and Khun Thanachot Udompiyasak who were involved in business process and knowledge for this research project. Without their participation and input, the results could not have been successfully conducted.

I would also like to acknowledge the committees Asst. Prof. Sukree Sinthupinyo, Ph.D from Chulalongkorn University and Assoc. Prof. Worasait Suwannik, Ph.D. from Kasetsart University, I am gratefully indebted for their valuable comments on this thesis.

Nevertheless, I am also thankful to the friends at Chulalongkorn University and senior for helping pass through all the steps, supporting and inspiring each other, Mr. Masayuti Simongyi, Mr. Bunyachai Chantian and 4x100 team, Ms. Patsuda Vittayanonaketavee, Mr. Thiti Kongprasert and Mr. Teeranai Sriparkdee, we had finished the final line. Last but not the least, also thank you to my family that encourage me throughout writing this thesis and studying the Master’s degree.

Maylada Pootisirakorn

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

Chapter I

Introduction

1.1 Statement of the problems

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 as 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 in the past. [Figure 1]

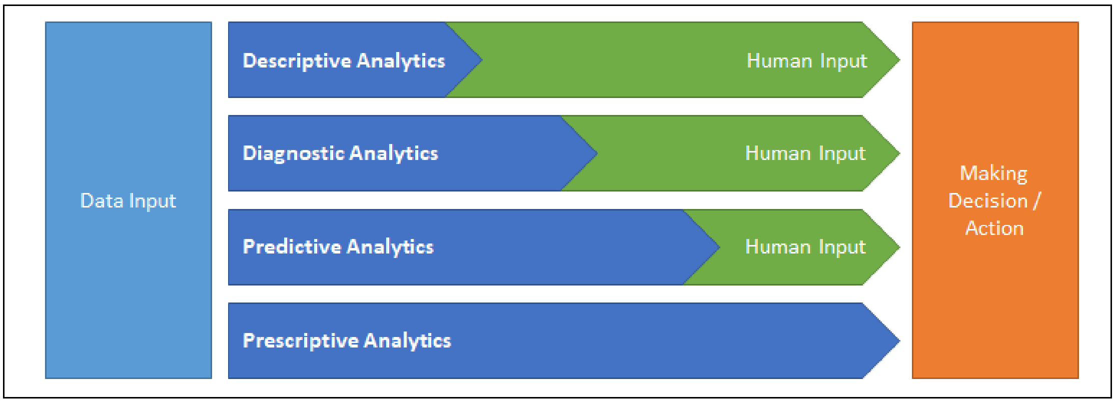


Figure 1 Data analytics focus types

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 follow 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 it may 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 method can be used for predicting this fault based on historical data. The method is applied to the operation data that is collected after completed drilling to target depth. Support Vector Machine (SVM) [1] is a popular technique in machine learning that can be used to classify two-classes data. The data set has a small rate of failure which is lead to the imbalance data set. This research is focus to train SVM to find out the probability of the failure that might happens before running the open-hole wireline logging to save time and cost.

1.2 Objective

Apply the machine learning technique to get the insight of the data which help making decision prior to running open-hole wireline logging of each well. This could help saving time and cost to avoid the failure of the tools.

1.3 Scope of study

1.3.1 This research focuses on the geological data and well logging information of wells in gulf of Thailand.

1.3.2 The input of data must be the well that has been drilled to the target depth.

1.3.3 The output of the model is focused on success or failure of the tools in open-hole condition.

1.3.4 Failure of the tools in the scope means that the tools is stuck or hung up in the hole either it can be fixed and passed through the target depth or needed to cut the tool and used another tool to pull the failed tool out.

1.3.5 The failure will not cover the case of tool failure when it works incorrectly or set up with the wrong adjustment and need to update.

1.4 Expected or anticipated benefit gain

1.4.1 A machine learning technique that learn from historical data and predict the success or failure of the open-hole wireline logging.

1.4.2 The predictive model can be used to improve the decision making.

1.4.3 Save cost and time to pull the tools out if the tools are failed.

1.4.4 Use this framework to apply and find the contributing factors that cause the failure of the open-hole logging.

1.5 Research methodology

1.5.1 Study the workflow and methodology to drill the wells and the open-hole wireline operation.

1.5.2 Study the concept and condition to identify the failure of the open-hole wireline logging

1.5.3 Study on the related works and research about machine learning technique that can be used to predict and classify the success or failure.

1.5.4 Gathering and preparing the data set which include cleaning data, transform data, deal with missing value and convert data into the format that can be used to evaluate the proposed method.

1.5.5 Consult with the subject matter experts to label the class of prepared data set.

1.5.6 Select machine learning technique to apply to the prediction model and design the workflow and experiment prior to build the model.

1.5.7 Run the implemented model then evaluate and validate the experiment results. Modify some of the parameter to optimize the model.

1.5.8 Analyze the result from the experiment and discuss the outcome with the subject matter expert on drilling operation.

1.5.9 Summarize the results of the research and make a report on this research.

1.5.10 Prepare academic articles.

1.5.11 Compile and write dissertation.

Chapter II

Background

2.1 Related theories

2.1.1 Offshore Drilling Operation

Offshore drilling operation [2] 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:

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 out to 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.
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.

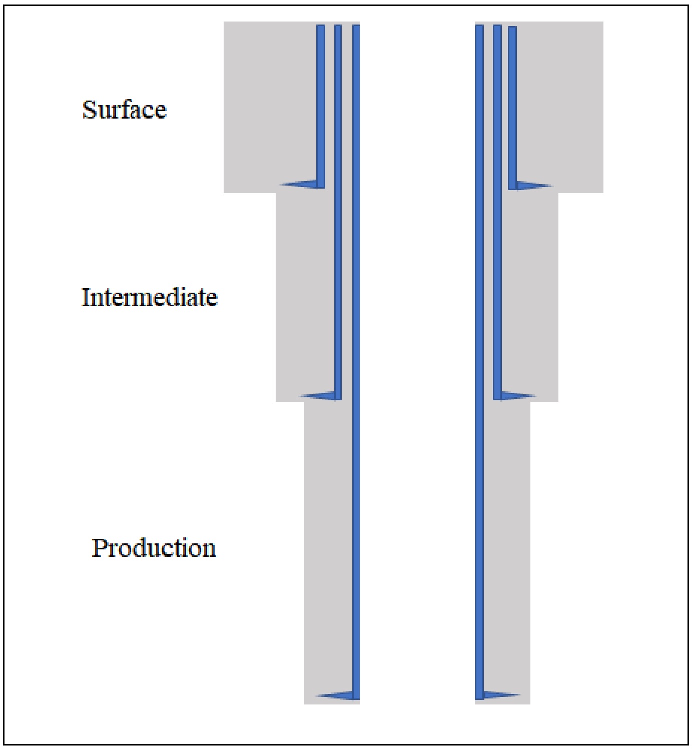


Figure 2 Schematic of the well with cementing on three major sections

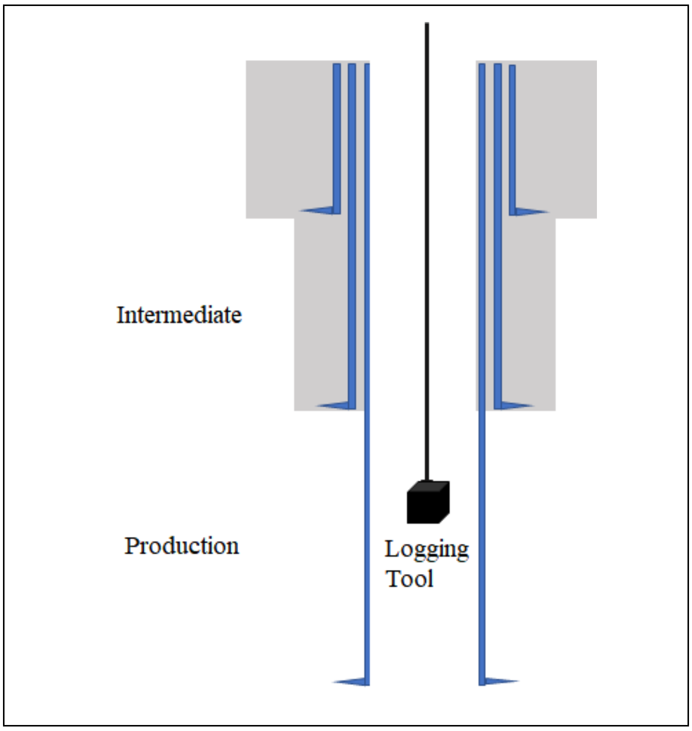


Figure 3 Wireline logging tool in bored hole or open hole

5. After the well has been drilled to the target depth with casing cement, it is ready for completion and production. Completion activity called perforation creates small hole through the casing. The small hole is passed through the production area that connect to the reservoir. This provide a flow path for oil and gas.

6. After the perforation process, the production process continues by maintaining the valves and pump to produce the oil and gas on the platform area. This continue the cycle to perforate to upper target of each well until the reserves is depleted.

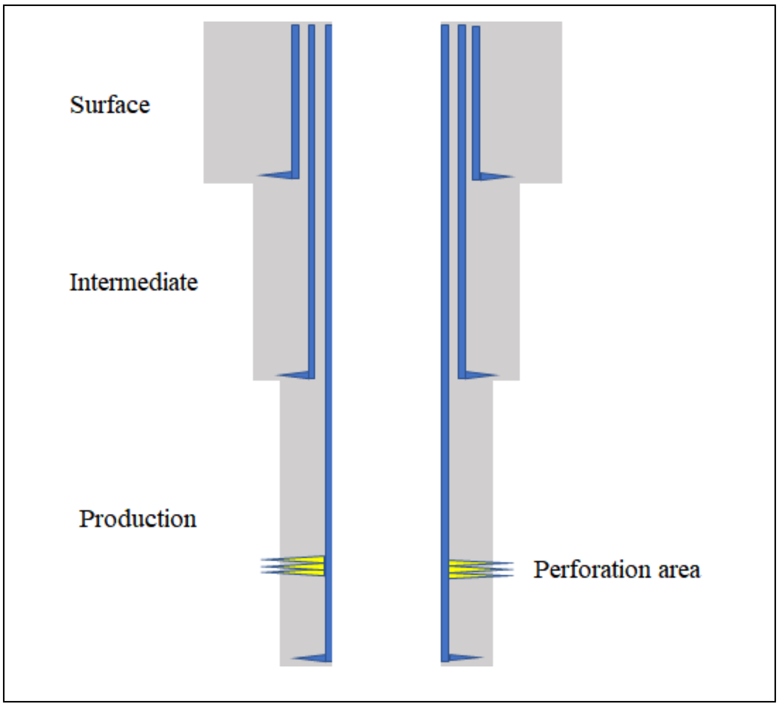


Figure 4 Perforation area after cement and run casing in the production section

2.1.2 Open-hole wireline logging

To drill the well, it is a technological process however no wells are identical with various of 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 activity is 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 this activities and ensure that it would not be failed and reduce the non-productive time of the drilling operation. The type of logging [3, 4] 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 since the data gathering would help expert to interpret and making consideration using statistical skill about well integrity and reservoir characteristics.

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.

Type of logging

1. Formation Tester is mainly used for collecting the pressure point for specific depth, the different subtype depends on temperature and the service company.

    (1) RDT (Reservoir Description Tool)

    (2) HSFT (Hostile Sequential Formation Tester)

    (3) HXPT (High Temperature Xpress Pressure Tool)

    (4) SRFT (Slimhole Repeat Formation Tester)

2. Wireline Logging is used primarily to identify lithology, reservoir porosity and fluid type in formation.

    (1) Quad Combo
    (2) Triple Combo

Quad Combo provides more detail of time-depth and pore pressure which is typically run is delineation or exploration wells to retrieve more information for making decision.

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.

2.1.3 Support Vector Machine

Support Vector Machine (SVM) [5] is a supervised machine learning algorithm which means that the answer must be known and use for training and building a predictive model. There are two main types for SVM, one is Classification, and another is Regression which can handle multiple continuous and categorical variables.

Classification is used for this research. The classification SVM construct the optimal hyperplane that separate data into groups with minimized error function through training process.

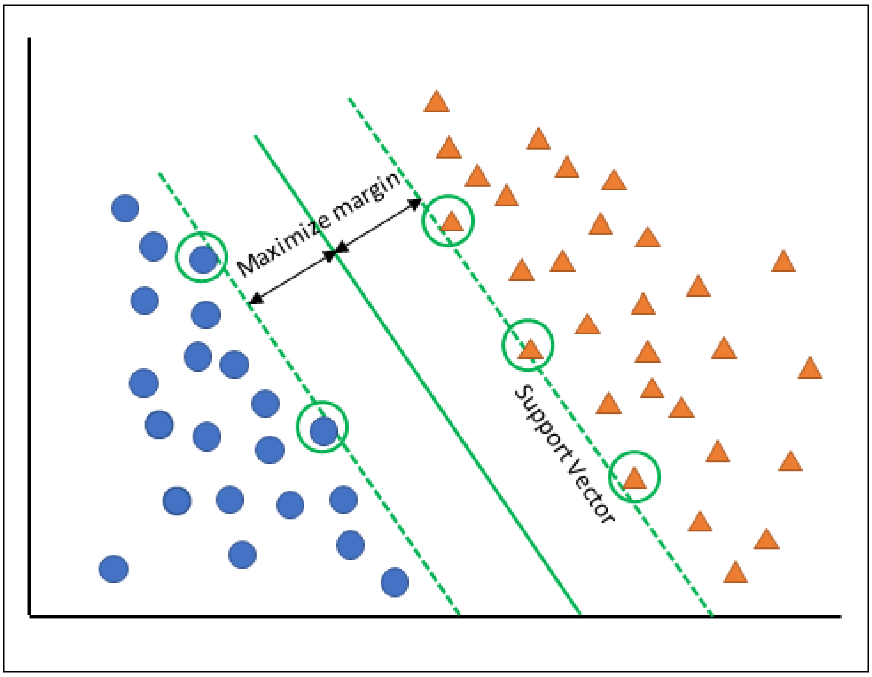


Figure 5 Support Vector Machine that separate the hyperplane to classify the class with maximum margin

2.1.4 Confusion Matrix

Confusion matrix [6] is a technique to summarize and describe the performance of a classification model on a test data set. It is used in evaluation method of the classification problem. The number of correct and incorrect predictions are summarized in table of numbers for each class. It gives an insight into the errors from the model but moreover the types of errors are also important which is not only the classification accuracy.

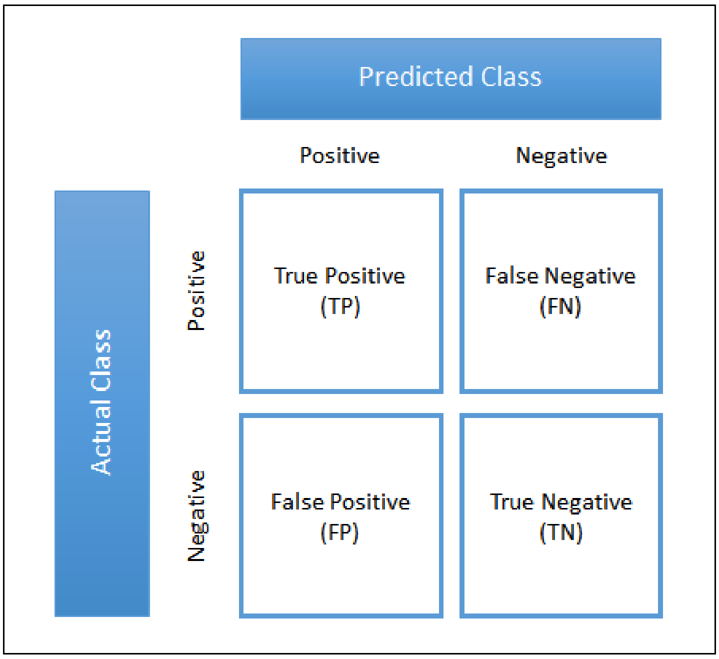


Figure 6 Confusion matrix as a table summary for a binary classification problem

- **Actual Class** is the result as labeled by actual experiment
- **Predicted Class** is predicted results by the test
- **True Positive (TP)** is the number of samples that were predicted as positive and the actual result is labeled as positive.
- **False Positive (FP)** is the number of samples that were predicted as positive, but the actual result is negative. It is also called Type-1 error.
- **True Negative (TN)** is the number of samples that were predicted as negative and the actual result is also negative.
- **False Negative (FN)** is the number of samples that were predicted as negative, but the actual result is positive. It is called Type-2 error.

Moreover, from numbers that are presented in the confusion matrix, there are other performance measurement that can be used to evaluate the model as below:

- **Accuracy** measures overall accuracy of the model classification

$$Accuracy = \frac{\text{all correct}}{\text{all}} = \frac{\text{TP+TN}}{\text{TP+FN+FP+TN}} \qquad (1)$$

- **Precision** is the positive predictive value

$$Precision = \frac{\text{True Positive}}{\text{Predicted Possitive}} = \frac{\text{TP}}{\text{TP+FP}} \qquad (2)$$

- **Recall** is the ratio of correct predictions and the actual positive

$$Recall = \frac{\text{True Positive}}{\text{All Possitive}} = \frac{\text{TP}}{\text{TP+FN}} \qquad (3)$$

Chapter III

Literature Review

3.1 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 to get insight and help in making decision from the data during the past decade.

3.1.1 Machine Learning Overcomes Challenges of Selecting Locations for Infill Wells

The recent research that use machine learning to help in selecting locations for infill well in 2018 [7] applies the support vector machine technique to the data collection and 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 new workflow after has been applied to help the asset team making data driven decisions.

3.1.2 Data Driven Approach to Failure Prediction for Electrical Submersible Pump System

This research in 2015 [8] presents a data driven approach for failure prediction of the pump system that used in oil and gas industry using support vector machine technique to train the selected features and test on real world data. The data collected by sensors based on electrical and frequency data and other information such as logs are feed to the machine learning framework to predict the results. The successful and timely diagnosis of failure from the model can improve the production performance. The paper selected the SVM as it is powerful binary classifier and using the feature extraction and selection in their work.

3.1.3 Implementing Artificial Neural Networks and Support Vector Machines in Stuck Pipe Prediction

The research of stuck pipe prediction in oil and gas industry in 2012 by Islam et al [9] 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 model was designed and constructed by MATLAB built-in function and library. 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% and claim 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 in 2006 Siruvuri et al [10] use ANN to predict stick pipe, the reasonable outputs were accepted even the data might be incomplete or have some errors.

3.1.4 Classification of Petroleum Well Drilling Operations Using SVM

SVM has been used to classify petroleum well drilling operations in 2006 Adriane et al [11] present the development of a system that intends to make better use of the information collected during well drilling operation. The main idea is to use a great amount of data that has not been properly used and it might provide insight. They use SVM for pattern recognition and develop the automatic classification system that can produce performance enhancement. This paper presents the 6 multi-class SVM and tested by the gaussian RBF, polynomial and linear functions using MATLAB. The simple linear SVM has the good generalization accuracy with correctness of 92%.

3.1.5 Study on Intelligent Prediction for Risk Level of Lost Circulation While Drilling Based on Machine Learning

The well problem is one of the most interesting issue that need focus on in 2018 Zejun Li et al [12]. They study three typical machine learning algorithms to analyze drilling data in Iraq to predict the lost circulation issue. They are SVM, ANN and random forest. SVM and random forest have predicted 99% or wells with normal, however the data is imbalance, only 55% of the lost circulation samples are correctly classified. The prediction 45% are incorrect. For ANN in the lost circulation cases 46.6% are correctly classify and 53.4% are incorrect. Compare to overall classification accuracy, the accuracy to identify lost circulation points is not ideal, partly because they are relatively sparse, and imbalance compared to the normal samples.

Table 1 Machine learning techniques application in oil and gas practices

Application	ML Technique	Data Set	Researcher	Year
Selection Infill Location	SVM, K-Means Clustering	Reservoir data, Oil - production rate, and completion data.	Adam Wilson	2018
Pump Failure Prediction	SVM	Electrical and frequency data from the field	Dong Guo et al	2015
Stuck Pipe Prediction	ANN, SVM	Mud logging and well information	Islam B. and Lloyd H.	2012
Classification of Well Drilling Operations	SVM	Drilling well information	Adriane B. et al	2006
Risk Level of Lost Circulation	SVM, Random Forest, ANN	Mud logging and well information	Zejun Li et al	2018

Chapter IV
Research methodology

4.1 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 interested attributes. 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. 1439 records are total number 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 spreadsheet, it needs to be collected and compiled prior to preprocessing in the next step. Data from database requires scripting to pull data appropriately.

The success class indicates that there is no stuck of the tool in the hole and 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.

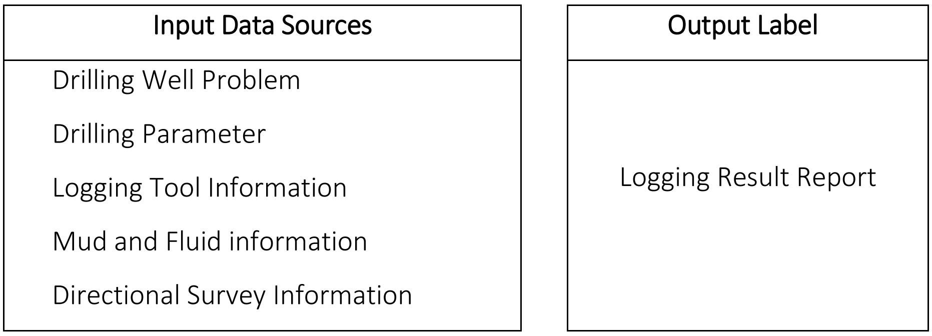


Figure 7 Input data sources and label data for open-hole wireline logging prediction

4.1.1 Data source

- **WellView** is the database that keep records of drilling data and non-productive events that occurs during drilling
- **OpenWorks** is the database that keep records of geophysical data
- **Wireline Tracking Sheet** is the summarize data by wells of logging tool type used for a well and records of tools running which are not kept directly in specific database
- **Engineering Desktop Tools** has one module to plan and calculate for directional well plan parameter

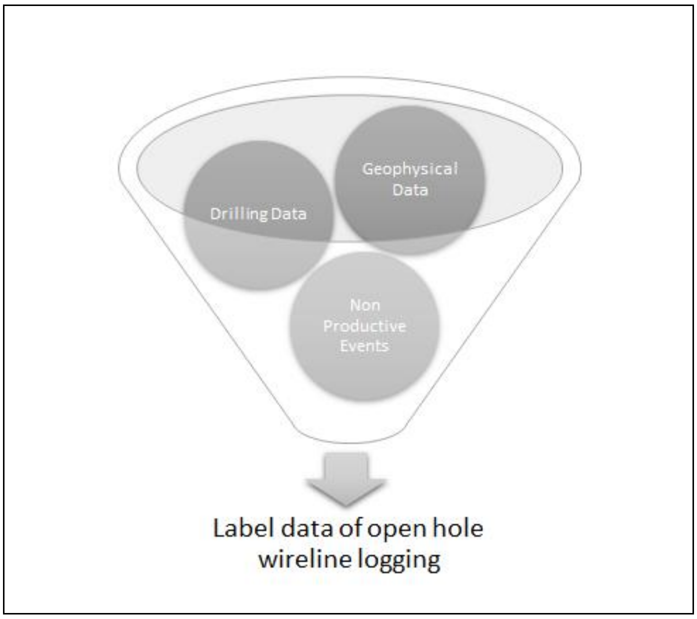


Figure 8 Process of gathering source of data and preparing to label the results for classification training

4.1.2. Clean, Prepare and Manipulate Data

Data from the historical result of open hole log is labelled with the rules of the tool type. In this stage, the record which has missing value will be removed. The calculation to find the representative value for missing value is developed. There are some other steps require to consolidate data into the single format. The format is based on rules given by experts.

- **Wireline Tracking Sheet**
  - Formation Tester (RDT, HXPT, HSFT and others), check that the data has Number of Archive > 0 or not. If number of archives is greater than 0 then fail, otherwise it is success.
  - Wireline Logging (Quad Combo, Triple Combo), check that the Log to TD information is either Yes means success or No means fail.

However, the data need to be cross-check with actual data recorded in the system. The experts help in reviewing the mismatch status of each well.

- **WellView** is a system that has actual activities of logging tool. There are several data cleaning steps being done.
  - 'LOGWL' or 'FISH' are activities with the keyword either 'STUCK', 'SIT DOWN' or 'HUNG UP'. These keywords mean failed. There are some keywords that being misspelled or being in past tense or passive voice in the comment activities. They must be corrected.
  - Extract the depth data that has been record in the comment and compare them with the plan target depth in the system.
  - Check whether there is non-productive time of the unscheduled events and remove them.

4.2 Attribute Selection

In practice, the irrelevant attributes in the input data set can lead to confuse the machine learning [13] such as SVM. The attribute selection against the whole data set has been applied.

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

Figure 9 Data table after processed data with class labeled.

Total of 32 attributes were selected to be used as the input in this experiment and the irrelevant attributes were removed. These attributes are:

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, Casing depth 9 5/8”, Casing depth 7”, Tubing depth, Mud at intermediate section, Mud 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 are selected attributes.

The comparison of the feature selection between expert’s advice and algorithm selection would need to get all data features from data source which required preprocessing step. Due to the limitation of time constraint, the preparation process take time consuming, the features selection step is agreeable to be selected by the experts.

4.3 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 in different purpose.

After separating data into two groups, the well data by each logging type is split 70% for training using cross-validation and 30% of testing.

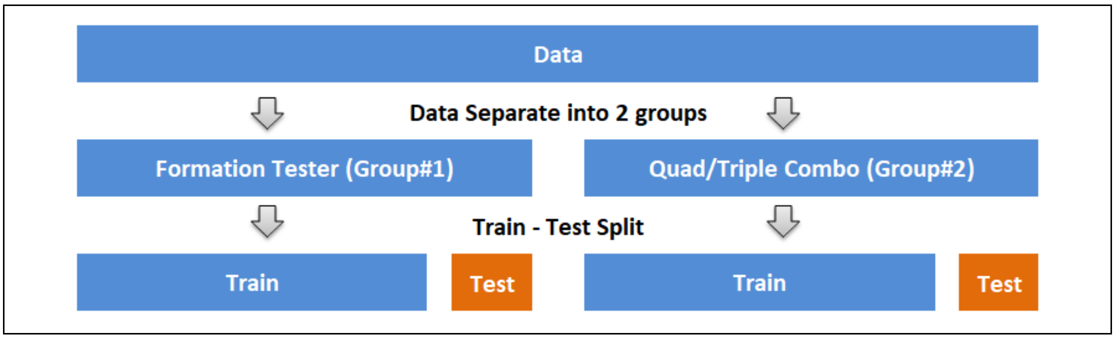

Data
Data Separate into 2 groups
Formation Tester (Group#1)
Quad/Triple Combo (Group#2)
Train - Test Split
Train
Test
Train
Test

Figure 10 Data separation process

4.4 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 set is randomly separated into five equal partitions, each part is held out to be tested and the trained on the remaining four.

Table 2 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

The group of data set were trained separately with 5-fold cross-validation [Figure 11]. Three machine learning are selected to applied on this experiment, SVM, Naïve Bays and Decision Tree.

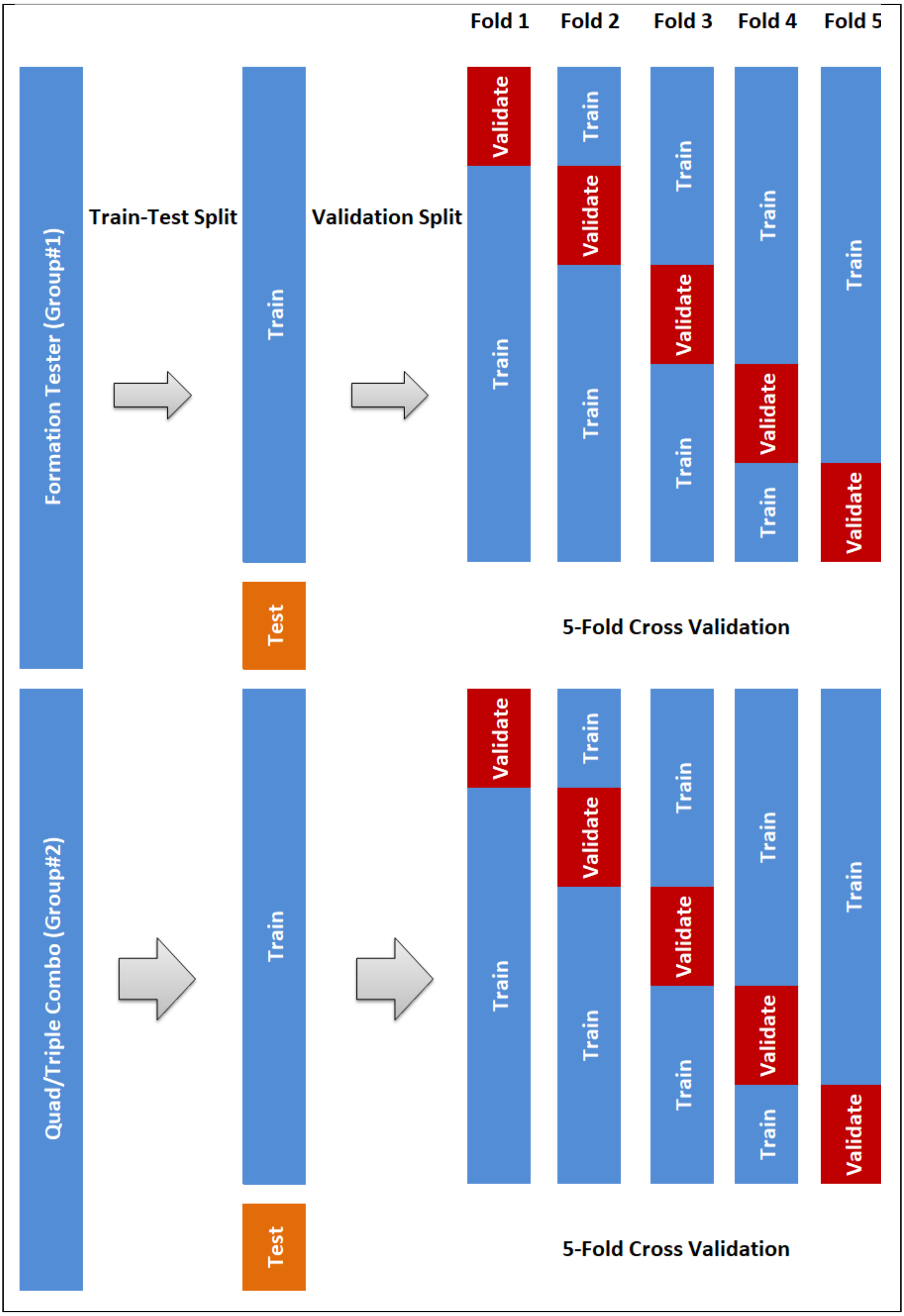


Figure 11 Cross validation using 5-Fold

4.5. Performance Measurement

In addition to accuracy, the performance of the model to classify each class using ROC curve [14] is another one measure. ROC curve is a plot of values of the False Positive Rate (FPR) versus the True Positive Rate (TPR)

$$True\ Positive\ Rate = \frac{TP}{TP+FN} \tag{4}$$

$$False\ Positive\ Rate = \frac{FP}{FP+TN} \tag{5}$$

In this experiment, due to the number of failure sample is far more than the number of success sample, this leads to imbalance issue in the training process, we focus to reduce the FP (false positives) which referred to the number of predicted as success, but the actual is failed. This case is better be avoided 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 failure but the actual is success. In this case, it is acceptable since engineers will focus to maintain the tools and other parameters to avoid the failure, or perhaps decide not to run the open-hole wireline logging.

Chapter V
Results

The result of classification with SVM, Naive Bayes [15] and Decision Tree [16], this experiment are presented. The classification performance was evaluated with test data set. An accuracy and ROC analysis were calculated as performance measurement. The results are shown in Table 3, Figure 11,12,13.

Table 3 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 the clear picture on the performance measurement. We focus on less 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, performance result is in the same trend as accuracy measurement.

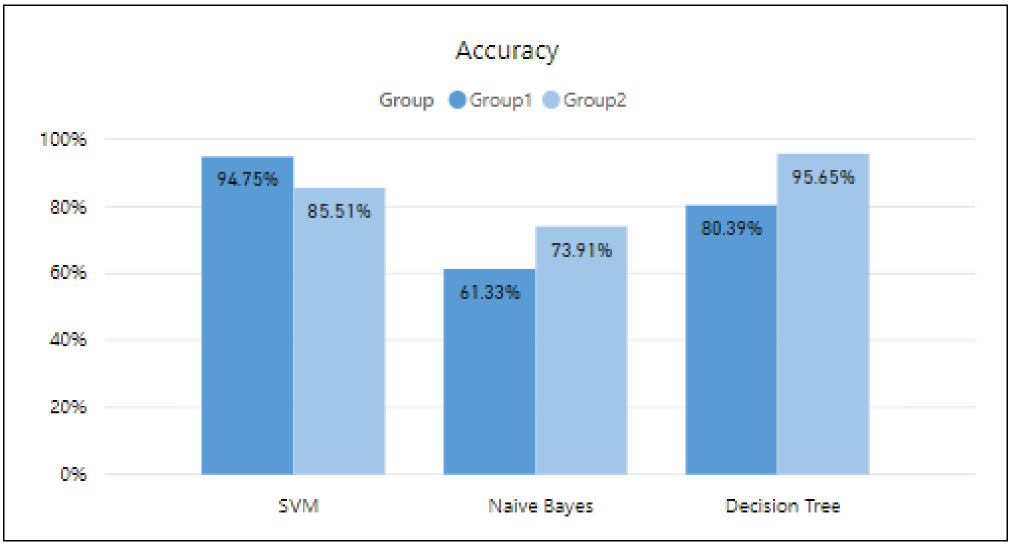


Figure 12 Accuracy evaluation matric using testing data set

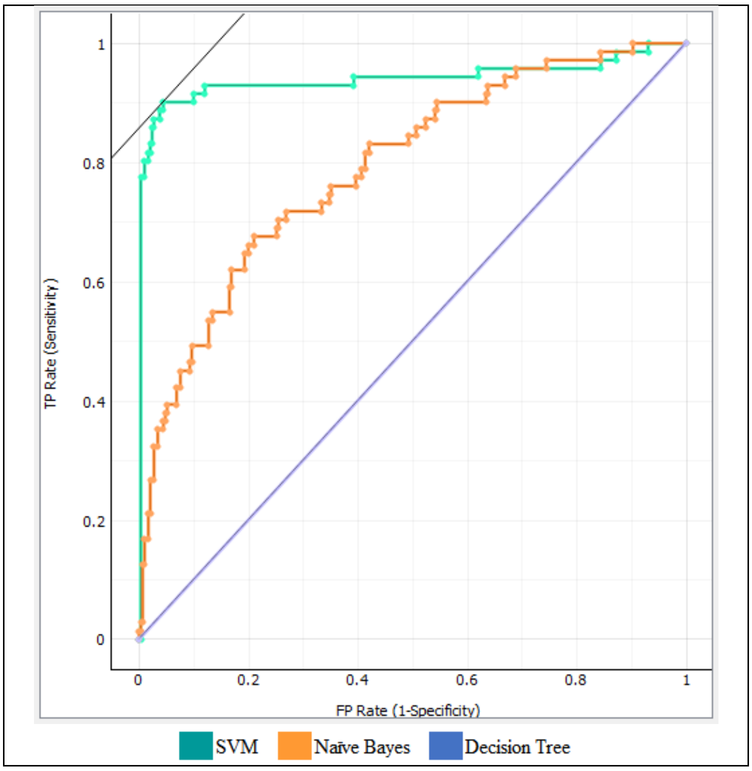


Figure 13 ROC curve of Formation Tester (Group 1)

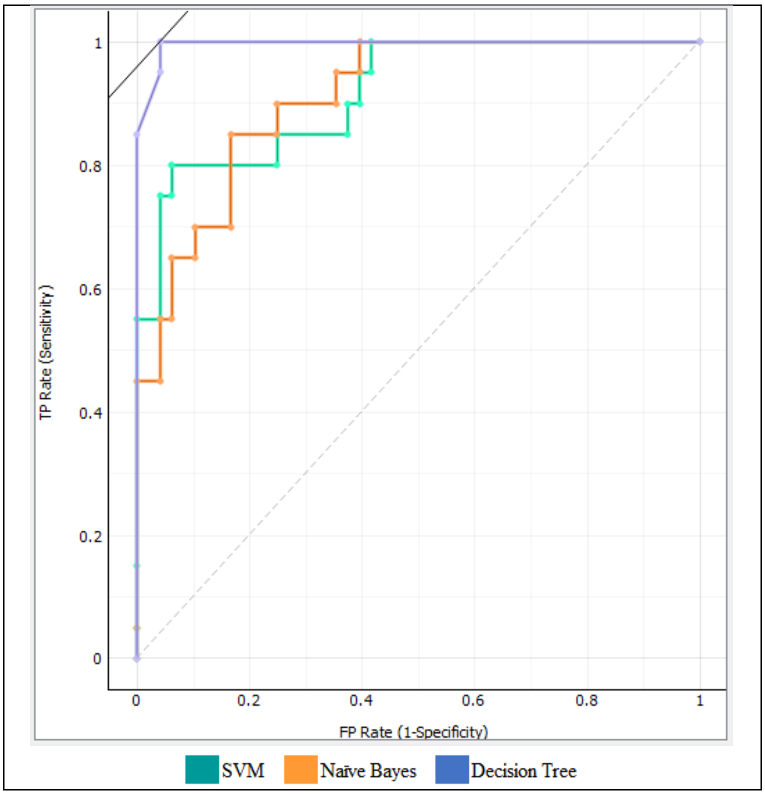
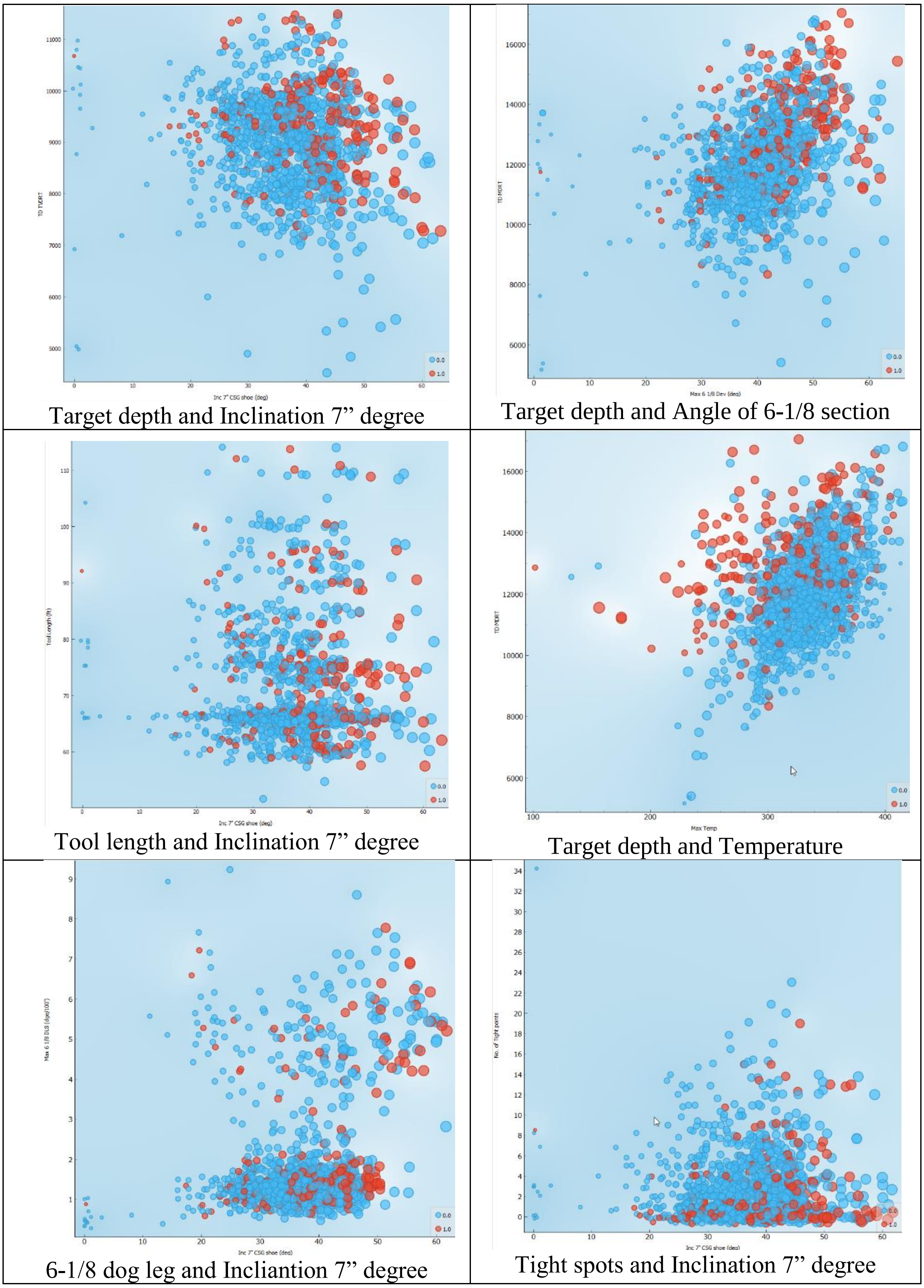


Figure 14 ROC curve of Quad/Triple Combo (Group 2)

The example graphs of correlation between two features show that it could not clearly conclude the relationship between them for the results (failure or success) of open-hole wireline logging.



Target depth and Inclination 7” degree

Target depth and Angle of 6-1/8 section

Tool length and Inclination 7” degree

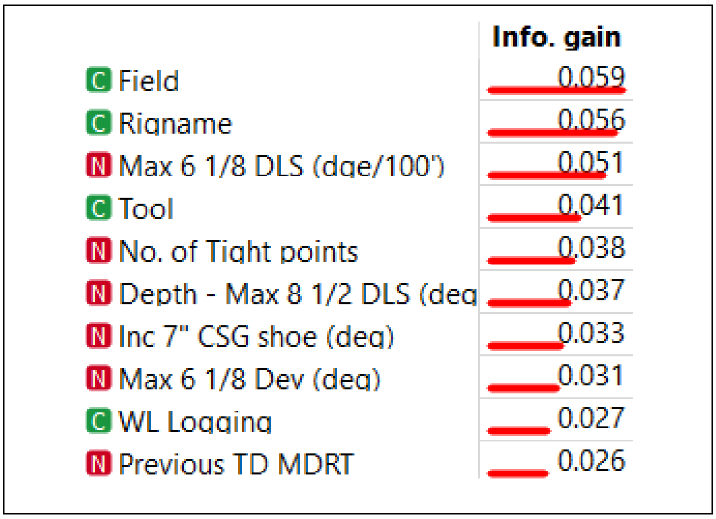
Target depth and Temperature

6-1/8 dog leg and Incliantion 7” degree

Tight spots and Inclination 7” degree

Figure 15 Example graphs of correlation between 2 features

The ranking of important features which are captured after running the model is listed below. It is based on information gain the expected amount of information.



	Info. gain
C Field	0.059
C Rigname	0.056
N Max 6 1/8 DLS (dge/100')	0.051
C Tool	0.041
N No. of Tight points	0.038
N Depth - Max 8 1/2 DLS (deg	0.037
N Inc 7" CSG shoe (deg)	0.033
N Max 6 1/8 Dev (deg)	0.031
C WL Logging	0.027
N Previous TD MDRT	0.026

Figure 16 Top 10 ranking of important features

Chapter VI
Conclusion

In this chapter, the summary of the experiment results, the problem and limitation of this research, future works and suggestions on the failure prediction in open-hole wireline logging of oil and gas drilling operation using support vector machine is described.

6.1 Results summary

The experiment of Formation tester data set (Group 1) SVM has the highest accuracy, precision and recall. Naïve Bayes has least performance for both data set. For the data set of Quad/Triple combo (Group 2), the prediction results were not efficient for SVM, decision tree is more accurate, nevertheless the small size of data set decreases statically power. The reliability of results may not be trusty for small data set.

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

6.2 Problem and limitations

The reason that we decide not to re-sampling since the in real world data under subsurface is full of uncertainty and we do not want to take risk with the 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 samples of quad/triple combo wireline logging which will help improving the result from machine learning algorithm.

The missing data and incorrect data of the preparation process take a lot of efforts to clean up data, this would need knowledge of understanding the data, business process, the good quality data approach had been discussing with the experts to let them know and understand about the problem of raw data also.

6.3 Benefit and contributions

The benefit that can captured 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. The cost could be different per project, but this can helps reduce the cost that occurs if there is less non-productive time.

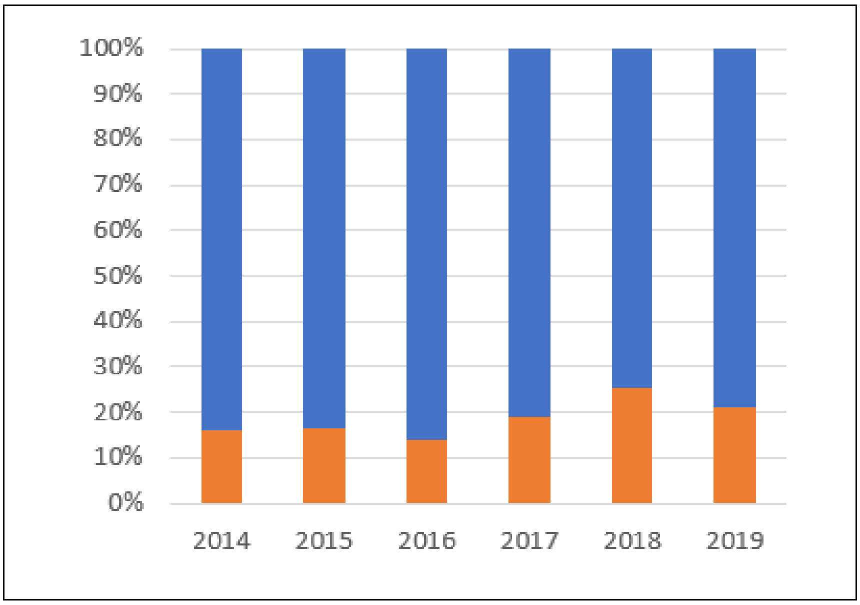


Figure 17 The percentage of failure by year

6.4 Future work and suggestions

From the research experiment, the confidence probability of fail or success result could be developed in the future works. Two approaches based on machine learning technique of support vector machine can be applied to get the probability of the predicted result. The proposed approaches are followed:

1. The concept to calculate distance between the predicted result and the border line that classify two classes of fail or success result. If the distance is low, it means that the probability of the predicted result with low confidence, on the other hand if the distance from the point to classification line is high, this leads to the confidence that the predicted result has high probability to be in the correct class.

2. Another approach is to identify the distance between a prediction result point and a centroid which is a repetitive of the class. If the distance between two point is close to each other, this means the probability of predicted result would be high with high confidence. However, if the distance between point is very far, the predicted result is less confidence.

These two proposed approaches could help describe the confidence level of the predicted result so that the experts can gain more insight to make decision more precisely.

REFERENCES

1. Cortes, C. and V. Vapnik, *Support-Vector Networks.* Machine Learning, 1995. **20**(3): p. 273-297.
2. J.J. Azar, G.R.S., *Drilling Engineering*. 2007: PennWell Books.
3. Kleinberg, R., *Well logging overview*. Vol. 13. 2001. 342-343.
4. Darling, T., *Well logging and Formation Evaluation*. 2005: Gulf Professional Publishing.
5. Tang, Y., et al., *SVMs Modeling for Highly Imbalanced Classification.* IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2009. **39**(1): p. 281-288.
6. 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.
7. Wilson, A., *Machine Learning Overcomes Challenges of Selecting Locations for Infill Wells.* Journal of Petroleum Technology, 2018. **70**(10): p. 48-49.
8. Guo, D., et al., *Data Driven Approach to Failure Prediction for Electrical Submersible Pump Systems*, in *SPE Western Regional Meeting*. 2015, Society of Petroleum Engineers: Garden Grove, California, USA. p. 6.
9. Heinze, L. and I.A. Al-Baiyat, *Implementing Artificial Neural Networks and Support Vector Machines in Stuck Pipe Prediction*, in *SPE Kuwait International Petroleum Conference and Exhibition*. 2012, Society of Petroleum Engineers: Kuwait City, Kuwait. p. 13.
10. 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.
11. 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.
12. 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.
13. Blum, A.L. and P. Langley, *Selection of relevant features and examples in machine learning.* Artificial Intelligence, 1997. **97**(1): p. 245-271.
14. 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.
15. Webb, G.I., *Naïve Bayes*, in *Encyclopedia of Machine Learning*, C. Sammut and G.I. Webb, Editors. 2010, Springer US: Boston, MA. p. 713-714.
16. Quinlan, J.R., *Induction of Decision Trees %J Mach. Learn.* 1986. **1**(1): p. 81-106.

VITA

NAME	Maylada Pootisirakorn
DATE OF BIRTH	5 December 1987
PLACE OF BIRTH	Ratchaburi
INSTITUTIONS ATTENDED	Mahidol University and Chulalongkorn University
HOME ADDRESS	Ngamwongwan Rd. Laksi Bangkok Thailand