Research on E&P Efficiency Metrics to support SKK Migas Mission utilizing CRISP-DM Methodology

Bobby Suryajaya\textsuperscript{1} Frieda Putri Aryani\textsuperscript{1}
Uma Bala Devarakonda\textsuperscript{1} Alva Erwin\textsuperscript{1}

\textsuperscript{1}Swiss German University, EduTown BSDCity, Tangerang 15339, Banten, Indonesia

Abstract

SKK Migas has mission to supervise and control activities of upstream Oil & Gas. It is believed that by knowing efficiency level of each PSC Contractor, improvement programs can be more effective and directly hit the objective. This paper discusses research proposal on efficiency metrics using data mining techniques and its implementation feasibility as National Indicator.

Keywords: SKK Migas, Indicator, Efficiency, CRISP-DM, Clustering, Classification, Association Rule

1. Background

SKK Migas is The Special Task Force for Upstream Oil and Gas Business Activities Republic of Indonesia, previously known as BPMIGAS. Government as the founder has defined SKK Migas function as its mission, which is to supervise and control the Production Sharing Contracts’ (PSC) implementation through partnerships in order to ensure the effectiveness and efficiency of upstream oil and gas business activities for the greatest welfare of the nation.

In order to successfully perform its mission, SKK Migas will need to have high quality and transparent data that is accurate, qualified, and can be retrieved timely.

On 2011, SKK Migas piloted SOT (Sistem Operasi Terpadu) to allow SKK Migas system to retrieve data from PSC Contractors’ data-source and then process and store the important information to SKK Migas master database that can be summarized and reported automatically to the stakeholders. The SOT pilot selected 5 PSC Contractors that contributes 50% of Indonesia Oil & Gas Production.


This research uses Exploration & Production (E&P) Value Chain as approach framework, which is chain of E&P activities performed by an Oil & Gas producer in order to deliver hydrocarbon where the value increases along the chain (Tordo, S. 2009). Value changes on hydrocarbon can be categorized as follow:

- **Resource Potential** (potential hydrocarbon based on seismic survey).
- **Developable Capacity** (developable hydrocarbon based on exploration activity).
- **Developed Capacity** (maximum well deliverability).
- **Available System Capacity** (maximum production facility throughput).
- **Production** (processed hydrocarbon, ready to sell).
3. Proposed Control Indicators

3.1. Efficiency Metrics

Production data collected from SOT enables SKK Migas to calculate Efficiency Ratios. It is simply a calculation based on input, output, and waste resulted from a production process.

Efficiency ratio is comparison between output of value chain and its input, for example comparing hydrocarbon produced by a facility with hydrocarbon flowed into facility. Normally, efficiency ratio will not exceed 100% since output will never be higher than input, and waste resulted by hydrocarbon processing is called Losses. Other names of losses are deferment or loss production opportunity (LPO).

At least following ratios can be calculated: Asset, Operating, Production, and Field Efficiency.

In theory, the maximum sustainable amount of hydrocarbon that can flow through the well or facility depends on its capacity. Two important components on E&P Value Chain that can be considered as input are:

- Developed Capacity, also known as Maximum Well Deliverability or MWD.
- Available System Capacity or ASC.

Efficiency ratio calculation can use production and sales (lifting) data for the output. Therefore, efficiency ratios can be calculated as the following:

- **Asset Efficiency** = \( \frac{Sales}{MWD} \)  
- **Operating Efficiency** = \( \frac{Sales}{ASC} \)  
- **Production Efficiency** = \( \frac{Production}{MWD} \)  
- **Field Efficiency** = \( \frac{Production}{ASC} \)

ASC can be calculated based on MWD, as built design of facility or field, or simply agreement on throughput or capacity of facility or field. Sometime, ASC can also be used to calculate production if losses are known with following formula:

\[
Production = ASC - Losses
\]

In real implementation, Efficiency Ratios are derived mostly by the Losses instead of Production.

3.2. Efficiency Metrics as Control Indicator

This research proposes the possibility of utilizing Efficiency Ratio as control indicator for SKK Migas to evaluate performance of PSC Contractors. Following is consideration of this research proposal:

- SKK Migas has ability to retrieve data requirement from PSC Contractor through SOT Production Monitoring initiative including MWD, ASC, Production, Sales, and Losses.
- SKK Migas through its PUPO-PPAM support the utilization of efficiency data.
- Efficiency Ratio has been widely benefited Oil & Gas Super Major Companies on evaluating their performance.
• Efficiency Ratio does not require complicated calculation.
• PSC Contractor may be already using Efficiency Ratio although in different form.

To be indicators, efficiency ratio must have criteria to evaluate performance (Paul Stevens, 2008). By using specific data mining technique such as clustering and classification, these criteria can be built.

3.3. Efficiency Metrics Contributor and Co-occurrences

Once efficiency criteria are established, it can be associated with the Losses to determine what losses contribute to specific efficiency criteria. This can be analyzed based on losses occurrences or value using Association Rule Mining technique.

4. Methodology: CRISP-DM

CRISP-DM is a data mining standards drawn up by the three founders of data mining market namely Daimler Chrysler (Daimler-Benz), SPSS (ISL), and NCR. It consists of 6 life-cycle phases (Wirth, R., & Hipp, J.).

4.1. Phase 1 – Business Understanding

In this phase, it takes an understanding of the substance of the data mining activities that will be carried out, from the perspective of business needs. The activities are:

Business Objectives
The business objective is to utilize efficiency ratio as control indicator by SKK Migas to evaluate PSC Contractor performance.

Data Mining Goals
The goal from this research is to construct efficiency criteria and identify the losses that contribute to each of efficiency criteria. In addition it also looks for any co-occurrence of the losses that contribute to specific efficiency criteria.

Project Plan
The project plan is to propose methodology to SKK Migas on carrying out this research based on CRISP-DM life-cycle phases.

4.2. Phase 2 – Data Understanding

This phase tries to study the data to get to know and familiar with the data that will be used and identify data quality, detecting a subset pull off the data to make the initial hypothesis. The activities are:

Data Limitation
Data limitation may include period of data, scope of data, magnitude of data, data complexity, data value, etc. Following might be the limitation:
• Maximum one-year or equal to 365 rows of data.
• The scope can be limited to one PSC contractor.

Data Format
The data may be presented on Microsoft Excel format.

Data Attributes
This research will need at least data attributes as defined on table 1.

4.3. Phase 3 – Data Preparation

This phase includes several activities such as data selection and cleansing. For example, prior to use data mining tools, data must not contain “null” label, especially for losses volume. If there are no losses, then data can be set as “0” (zero) instead of null.

If there are too many data attributes e.g. losses event categories, prioritization can
be done due to hardware or resource limitation. For this purpose, Pareto diagram can be used to filter TOP-20 or TOP-40 losses event categories.

**TABLE I. Data Attributes**

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Description</th>
<th>UOM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>Date of data</td>
<td>-</td>
</tr>
<tr>
<td>Name</td>
<td>Name of Asset, Field, Facility, or PSC Contractors</td>
<td>-</td>
</tr>
<tr>
<td>MWD</td>
<td>Maximum Well Deliverability</td>
<td>BOE*</td>
</tr>
<tr>
<td>ASC</td>
<td>Available System Capacity, maximum sustainable hydro-carbon production</td>
<td>BOE*</td>
</tr>
<tr>
<td>Losses Volume</td>
<td>Waste or Loss Production Opportunity</td>
<td>BOE*</td>
</tr>
<tr>
<td>Production Volume</td>
<td>Hydrocarbon produced from field or facility</td>
<td>BOE*</td>
</tr>
<tr>
<td>Sales Volume</td>
<td>Lifting Volume</td>
<td>BOE*</td>
</tr>
</tbody>
</table>

*BOE = Barrel Oil Equivalent

### 4.4. Phase 4 – Modeling

This phase determines the data mining techniques that will be used, specify tools data mining, data mining algorithms, and determine the parameters to get optimal value.

This research is supported by various data mining techniques visualization such as Clustering, Classification, and Association Rule techniques and Rapid Miner will be used as data-mining tool.

**Clustering**

Clustering algorithms are applied on a given data set iteratively until any meaningful derivations could be found in the result. The clusters thus formed can be non-overlapping collection of data points that are similar to each other (K-Means algorithm). Or they can be a set of nested clusters organized as a hierarchical tree (Agglomerative Hierarchical clustering). (Tan et al., 2006)

Validity of clusters highly depends upon two factors. One, choosing right representative and the other is to identify similarities between various datasets. (V. Ilango, 2011)

Two data points are similar or closer to each other using various means one of which is by calculating Euclidean distance between them. The closer the points are the more are the chances for them to be in one cluster.

**Fig. 3: Clustering Example**

There are many clustering algorithms each with their own strengths and weaknesses. Depending on the nature of data like sparse or dense, fewer dimensions or multi dimensions; one can decide on the algorithm.

**Decision Tree based Classification**

A classification technique is a systematic approach to build classification models from an input data set. The technique proposed is decision tree induction which employs a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data. The model generated by the learning algorithm should both input data well and correctly predict the class labels of records it has never seen.

**Fig. 4: General Approach of Classification Modeling**

In a decision tree, each leaf node is assigned a class label. The non-terminal nodes, which include the root and other internal nodes, contain attribute test conditions to separate records that have different characteristics. (Tan et al., 2006)
**Association Rule**

Association rule mining according to (Kotsiantis & Kanellopoulos, 2006) is used to determine the association rules that meet the minimum support and confidence from a particular database. The process for mining using association rule is divided into two tasks (Agrawal & Srikant, 1994): (a) generate frequent item set, whose objective is to find the entire item sets where their support greater than minimum support and (b) Generate the association rules, whose objective is to extract all the high-confidence rules from the frequent item sets found in previous step.

(Tan et al., 2006) mention about the mathematical representation of the data set will be: Let I = \{i_1, i_2, \ldots, i_n\} be a set of attributes and D = \{d_1, d_2, \ldots, d_k\} be the set of data for these attributes over a period of time from day 1 to day k. Each data contains a subset of I, which is called as item set. An association rule is an implication in the form of X \Rightarrow Y, where X, Y \subset I are sets of items, which are called as item sets. And X \cap Y = \emptyset. X is called antecedent while Y is called consequent, the rule means X implies Y.

There are two important basic measures for association rules (Kotsiantis & Kanellopoulos, 2006): support(s) and confidence(c). Usually thresholds of support and confidence are predefined by users by adjusting the values and applying them in each run of the association rule and look for some meaningful rules generated.

\[
s(\{X \rightarrow Y\}) = \frac{\text{Occurences}(X \cup Y)}{\sum \text{Occurences}(X)} \quad (6)
\]

\[
c(\{X \rightarrow Y\}) = \frac{\text{Occurences}(X \cup Y)}{\sum \text{Occurences}(X)} \quad (7)
\]

Since association rule aims to extract interesting correlations among sets of items data, it is proposed to use the data to determine correlation between losses events and find their association to specific efficiency level or criteria.

**Rapid Miner**

This research proposes to use Rapid Miner, as it is open source and a free data-mining tool, yet equipped with great and latest features for data mining. It comes with various features such as:

- Runs on every major platform and operating system.
- Most intuitive process design.
- Multi-layered data view concept ensures efficient data handling.
- Powerful high-dimensional plotting facilities.
- Access to multiple data sources.

**4.5. Phase 5 – Evaluation**

This phase will interpret the data mining results indicated from previous phases. In this research, it is suggested that SKK Migas use another data (e.g. from different PSC Contractor) to test the validity of model that was built on phase 4.

**4.6. Phase 6 – Deployment**

Deployment is related with how this research can be implemented in both SKK Migas and PSC Contractor environment based on data requirements and people or organization.

**Data Requirements**

The data explained in phase 2 – data understanding should be collected. The collected data must be prepared through pre-processing activity proposed in phase 3 – data preparation. Once the data ready, analysis process can then be done based on proposed method on phase 4 and 5, data modeling and evaluation.

**People & Organization**

There should be coordination with all related department within SKK Migas. The idea is to have same understanding between all related parties on what and how to measure, and what will be achieved after the implementation.
5. Conclusion

Efficiency Metrics are common indicators used by any industry to measure performance of an entity. In Oil & Gas industry, efficiency refers to hydrocarbon process flow that can be measured quantitatively based on E&P Value Chain as explained on section 3.1.

Based on justifications explained on section 3.2 supported by CRISP-DM methodology, SKK Migas has higher successful probability to implement efficiency indicators nationally, considering that all required data is recorded and reported regularly by all PSC Contractors to SKK Migas.

However, it was found that not all recorded and reported data is following SKK Migas standard such as report format and data acquisition point, which may produce different calculation result. This may happen due to different standard adoption by each PSC contractors, where multinational companies normally follow their head quarter’s standard. Therefore, socialization prior to implementation for all PSC contractors is mandatory to ensure same understanding.

In order to ensure successful implementation, initial observation should be conducted to assess all possibilities and find out possible challenges or kickback such as unwillingness of one PSC contractor to provide more data following other PSC contractors.

6. Recommendation for SKK Migas

This research is only one aspect that supports DOE implementation. Furthermore, the research scope can be expanded to following aspects: Governance, Risk Management, System Architecture, Security, etc.

7. Acknowledgment

The authors wish to thank Head of SKK Migas Management Information System (MIS) Division and MIS Data Analysis Team for their full support on this research.

8. References