Journal of Computing and Applied Informatics (JoCAI) Vol. 6, No. 1, 2022 | 1 – 11

DATA SCIENCE
Journal of Computing and Applied Informatics

Predicting Oil and Gas Investment Decision through Machine Learning Approach: Empirical Evidence in Indonesia Oil Exploration Industry

Harry Patria

Newcastle University, School of Computing, United Kingdom

Abstract. Petroleum investment decision remains subject to economic and financial research for decades. Due to capital intensive and higher risk on oil exploration, the investment decision has become more critical than ever before. This study aims to shed some light on this issue by conducting four machine learning algorithms to predict by applying the dataset from 2004 to 2016. This study includes the Decision Tree, Random Forest, Naïve Bayes, and Support Vector Machine. A comparative performance analysis is the illustrated using confusion matrix, Cohen's Kappa value, and the accuracy of each model and Area under the ROC Curve. In this study, a machine learning approach was carried out on the oil exploration data. The findings demonstrate that Naïve Bayes has the most accurate performance for the classification (94.5%), followed by Decision Tree (92.9%), Random Forest (90.9%), and Support Vector Machine (89.6%). In practice, the selected Naïve Bayes model was applied to assess the decision using a new data test. The findings can diminish the subjective blindness and confirmation bias in the investment decision and bring about a reasonable and orderly exploration and development of petroleum reserves.

Keywords: Data Mining, Decision, Exploration, Gas, Machine Learning, Oil

Received 07 November 2021 | Revised 28 January 2022 | Accepted 28 January 2022

1 Introduction

Most of the exploration studies refer to a state-of-the-art developed by Fisher's works. The model estimated exploration equations by applying the success and discovery rate for different American oil fields from 1946 to 1955. Explanatory variables cover economic factors (e.g., oil prices,

*Corresponding author at: Newcastle University, School of Computing, United Kingdom

E-mail address: harry.patria@sbm-itb.ac.id

ISSN: 2580-6769 |e-ISSN: 2580-829X | DOI: https://doi.org/10.32734/jocai.v6.i1-7539
Journal Homepage: https://jocai.usu.ac.id

drilling costs) and geological proxy (e.g., seismic, discovery rate). The most common model is based on aggregate data for regions, continents, groups of countries, and countries [5]-[7].

Over 15 years, the exploration model using disaggregated data was developed by several scholars, e.g., Kolb, Pindyck, Hubbert [5]-[7], [10]. Kolb (1979) examined Fisher's for oil-prone districts, which are slightly more disaggregated settings [4]. Pindyck (1978) originated the interest rate in his drilling model [10]. The role of economic variables paid more attention in that year. Hubbert (1962) addresses the depletion effect on his models, stimulating insightful findings and implications for financial and policy instruments [5]-[7].

The research trends reveal that empirical models become more complex, demanding disaggregated and intertemporal data across the periods. Most of the empirical works in this field are dominated by US and European Countries. Hence, there is a lack of practicability and generalization of earlier findings across other countries. Over the last 20 years, geological and economic perspectives were comprehensively applied to study exploration economics [5]-[7]. Despite the disaggregated studies leading to better insights, most common studies are still developed employing aggregate data for groups of countries, countries, or regions [3], [5]-[7].

In Indonesia, one of the empirical analyses using geological basins conducted by [7] using panel datasets comprising geological and economic variables. The results discussed determinant factors affecting Indonesia's oil exploration in 2004-2013 that opened up a way to uncover the role of geological and economic features [7]. Although earlier research already offered and empirically tested the model, some economic and institutional variables are overlooked, e.g. appraisal spending, national vs. international oil operatorship. It can be argued that appraisal spending becomes a significant part of the early exploration stage. Moreover, institutional features such as management policy and operatorship might play an important role both in exploration and production.

2. Literature Review

Most of the studies of exploration refer to a state-of-the-art developed by Fisher's works [3], [5]-[7]. The model estimated exploration equations by applying the success and discovery rate for different American oil fields over the period 1946-1955. Explanatory variables cover economic factors (e.g., oil prices, drilling costs) and geological proxy (e.g., seismic, discovery rate). The most common model is based on aggregate data for regions, continents, groups of countries, and countries [5]-[7].

Over 15 years, the exploration model using disaggregated data was developed by several scholars e.g., Kolb, Pindyck, Hubbert [5], [6], [10]. Kolb (1979) examined Fisher's for oil-prone districts which are slightly more disaggregated settings. Pindyck (1978) originated the interest rate in his drilling model. The role of economic variables paid more attention in that year. Hubbert (1962)

addresses the depletion effect on his models stimulating insightful findings and implications for economic and policy instruments [5], [6].

The research trends reveal that empirical models become more complex demanding disaggregated and intertemporal data across the periods. Most of the empirical works in this field are dominated by US and European Countries. Hence, there is a lack of practicability and generalization of earlier findings across other countries. Over the last 20 years, geological and economic perspectives were applied to study exploration economics comprehensively [5], [6]. Despite the disaggregated studies leading to better insights, most common studies are still developed employing aggregate data for groups of countries, countries, or regions [3], [7], [9].

In Indonesia, one of the empirical analyses using geological basins conducted by [7] using panel datasets comprising geological and economic variables. The results discussed determinant factors affecting Indonesia's oil exploration in 2004-2013 that open up a way to uncover the role of geological and economic features [7], [9]. Despite that earlier research already offered and empirically tested the model, some economic and institutional variables are overlooked e.g., appraisal spending, national vs. international oil operatorship. It can be argued that appraisal spending becomes a significant part of the early stage of exploration. Moreover, institutional features such as management policy and operatorship might play an important role both in exploration and production.

2 Methodology

This study employs Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, including distinctive stages. First, it commenced by understanding the data in the context of oil exploration in Indonesia (3.1). The data preparation was performed to structure the data in given of oil field with several label data including region, offshore, national oil company and periods (3.1). Afterward, data modeling was conducted to construct a hypothetical regression model with binomial terms of oil wells drilled (3.2). Machine learning models for the classification were proposed and discussed (3.3) before ending with discussion and model evaluation based on ROC and accuracy (4)[14].

2.1 Dataset

The datasets comprise time series for all variables over the period 2004-2016, split between the two regions on the Indonesian geological basins. The dataset is retrieved from the database of Satuan Kerja Khusus (SKK) Migas (Upstream Regulatory for Oil and Gas, Republic of Indonesia). This government institution used to collect and evaluate the data and information for monitoring and optimizing Indonesian upstream oil and gas activities [13].

According to the stylized fact in Figure 1, the exploration in East prone is dominated by emerging fields that vary in many wells drilled by exploration companies. On the other hand, the wells drilled in West prone are relatively few. This descriptive analysis confirms a transition from West to East Exploration, which has a different geological feature. Petroleum exploration of the geological portions of the Indonesia region is quite further from a petroleum industry standpoint [11]. The exploration and production have been intensified on the western basins of Indonesia, which has a less enormous potential of undiscovered fields, most likely made of numerous small to medium size petroleum. On the contrary, the eastern basins are a relatively high-risk frontier which is generally under-explored and under-exploited with half of the basins being undrilled [9], [11].

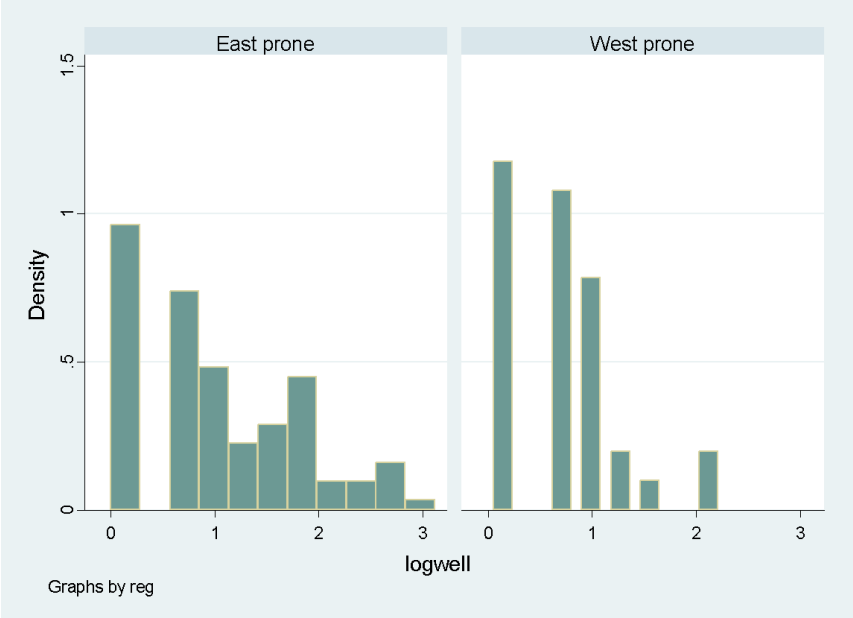

Figure 1 Wells Drilled in East and West Prone (*in Logarithmic*).

Figure 2 describes several variables used in this study, such as wells drilled, discovery size of oil and gas success rate, appraisal spending. The data distribution indicates a difference in drilling activities in these separated geological basins. In line with earlier discussion, the exploration activities in the East region are relatively higher, so do the success rate and volume discovered by exploration companies. This stylized fact confirms the transition from oil to gas exploration.

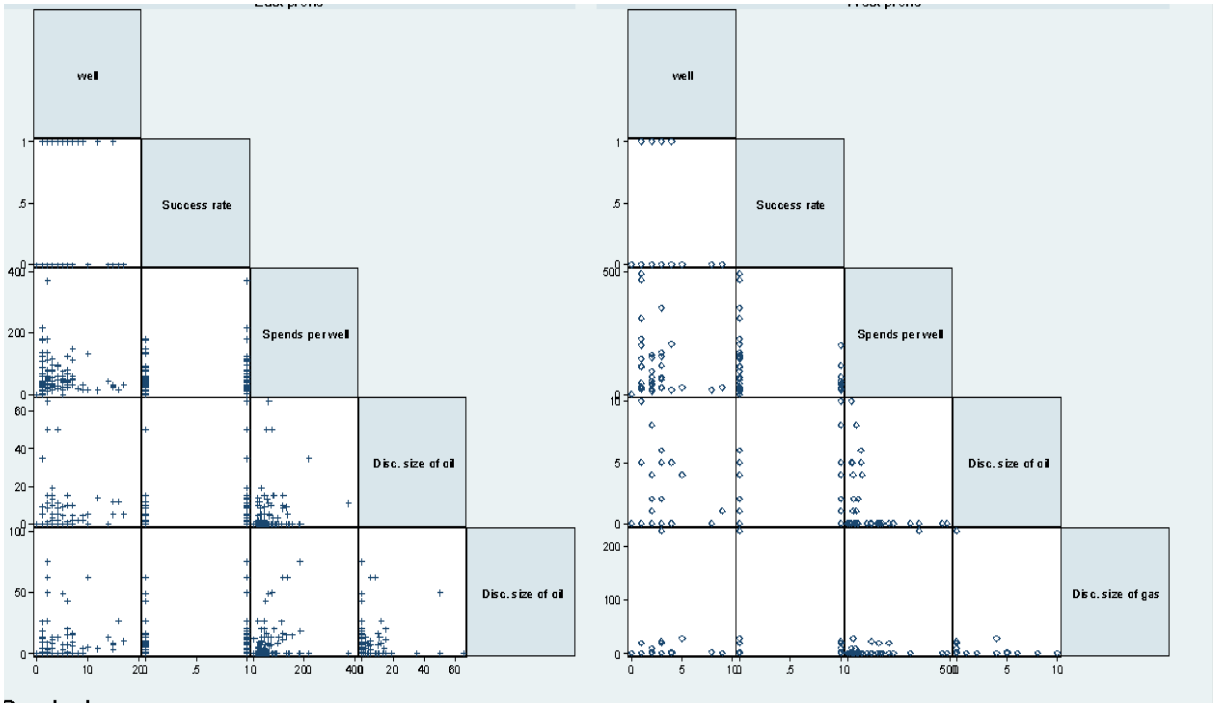

Figure 2 Wells Drilled in East and West Prone

The findings cover spending per well by exploration companies in a specific basin. The East exploration looks more distributed in number. These also spend higher over, in line with the discovery size of oil and gas found in a specific basin. This indication constitutes the key important issue of why the economic factor does matter.

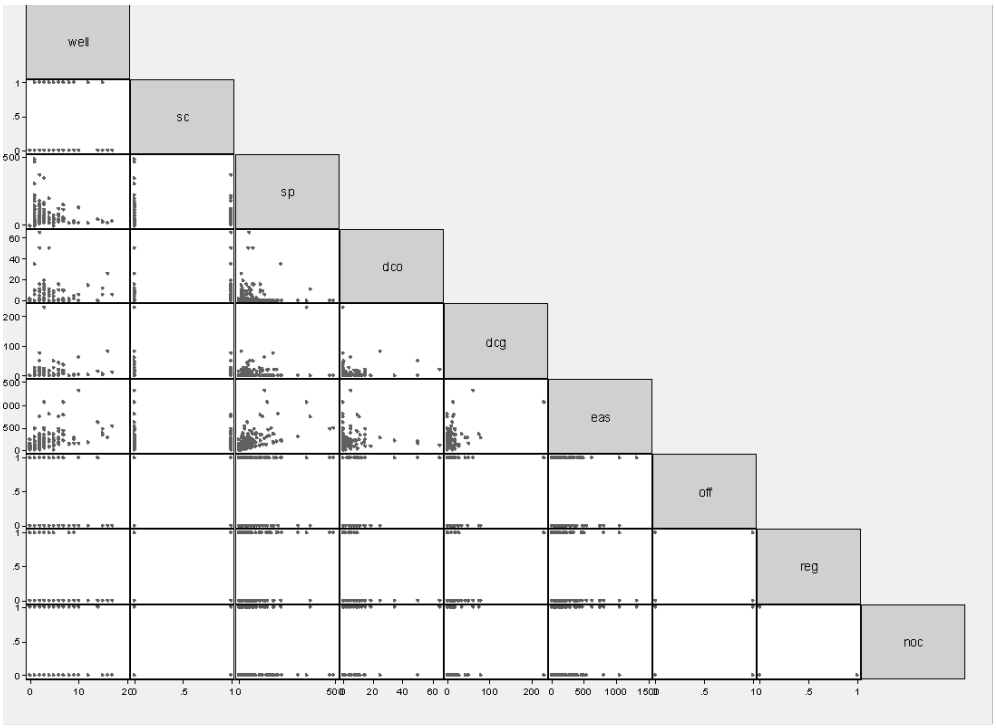

Figure 3 Descriptive Analysis of Overall Datasets

2.2 Method

This section aims to transform the theoretical model into an empirical model for hypotheses development regarding exploring oil and gas wells in Indonesia. Adopting Fisher (1964), a practical model consisting of an average discovery size for oil and gas and success rate. To construct a model, this study covers geological and economic features such as region, location of either drilling on onshore or offshore, operatorship either by the national or international oil company, and appraisal spending.

The multiple regression model is employed to empirically assess the geological and economic predictors that offer a more suitable and compatible causal model. The ceteris paribus is widely applied in multiple regression to maintain the remaining conditions constant. Hence, it allows us to do an interpretation of the parameter of each predictor, with the effect of other predictors statistically eliminated. According to equation (1) and variables discussed above, the equation can be modeled:

$$W_{it} = \alpha_i + \beta_{1i} S_{C_{i,t-1}} + \beta_{2i} D_{O_{i,t-1}} + \beta_{3i} D_{G_{i,t-1}} + \beta_{4i} lnSp_{it-1} + \beta_{5i} Of_{i,t-1} + \beta_{6i} Reg_i + \beta_{7i} No_i + e_{it} \quad (1)$$

W_{it} stands for the number of wells drilled. The geological variables are explained by the success rate $S_{C_{i,t-1}}$ and discovery size of oil and gas respectively $D_{O_{i,t-1}}$ and $D_{g_{i,t-1}}$ in the previous year. The economic variable is then expressed by the exploration cost spent by companies in the previous year $Sp_{i,t-1}$. The β_{ni} are coefficients to be estimated while α_i and e_{it} are intercept and error of the econometric model respectively. Each explanatory variable varies across basins and periods. This model applied lag variables on exploration explaining the behavioral model of companies.

According to Table 1, oil drilling activities had 16.40% of average success rate resulting in 1.51-to-3.23-million-barrel oil or gas equivalent. Gas drilling showed 14.6 of standard deviation which was relatively higher over oil activity. Moreover, offshore and onshore drilling as well as eastern and western prone tends to have a balance ratio with average spends per well of around US$ 27.75 Million and 18.75% local firms doing this exploration activity, dominated by foreign enterprises.

The exploration model used in this study takes into account geological distribution. Each well drilled in geological basins is an integer number. Whereas the discovery size of oil and gas are decimal data. Hence, these different types of data in the model are then estimated by selecting a Poisson regression model. The regression model analyzes the distribution with intensity parameter μ that is determined by explanatory variables.

Table 1 Descriptive Analysis of Drilling Data in Indonesia during 2007 – 2019

Variable	Symbol	Mean	Standard deviation	Minimum	Maximum
Success rate	S_c	0.1640	0.37	0	1
Discovery size of oil	D_o	1.5183	6.00	0	65
Discovery size of gas	D_g	3.2369	14.6	0	248
Spends per well	S_p	27.75	61.0	0	490
Offshore	O_f	0.50	0.50	0	1
Region	R_g	0.4375	0.49	0	1
National oil company	N_o	0.1875	0.39	0	1

3.3 *Machine Learning Model*

The machine learning model for predicting oil drilling was conducted using classification, as part of the supervised machine learning model. The accuracies were compared to select the optimum fit model among four learning algorithms: Decision Tree, Naïve Bayes, Support Vector Machine, and Random Forest learner and predictor nodes on the KNIME analytics platform. The supervised learning was performed using the following features: geological and economic variables.

The dataset was loaded into the model using the Excel File Reader. The processed dataset was applied to the configured learners and the predictors. The results were finally evaluated using the confusion matrix and Receiver Operating Characteristics (ROC) Curve. The workflow annotations were applied to label each stage of data mining. The confusion matrix demonstrates the standard accuracy rate of each model.

The decision trees were applied using Gini Index (GI) splitting criteria when partitioning the branches. The aim is to minimize the Gini Index. The support vector machine (SVM) was applied through a linear classification technique that separates the categories is as far as away from the nearest point.

3 Findings and Discussion

This paper empirically examines the accuracy of each machine learning model to select the fittest model. The ROC curve plotting the true positive rate against the false-positive rates demonstrates the classification performance of the classification model. The larger ROC leads to the better accuracy of the classifiers.

Table 2 ROC and Accuracy of Classification Models

Variable	ROC	Accuracy	Cohen's kappa
Decision Tree	0.929	95%	0.886
Naïve Bayes	0.945	85%	0.647
Random Forest	0.909	95%	0.886
Support Vector Machine	0.896	84.6%	0.655

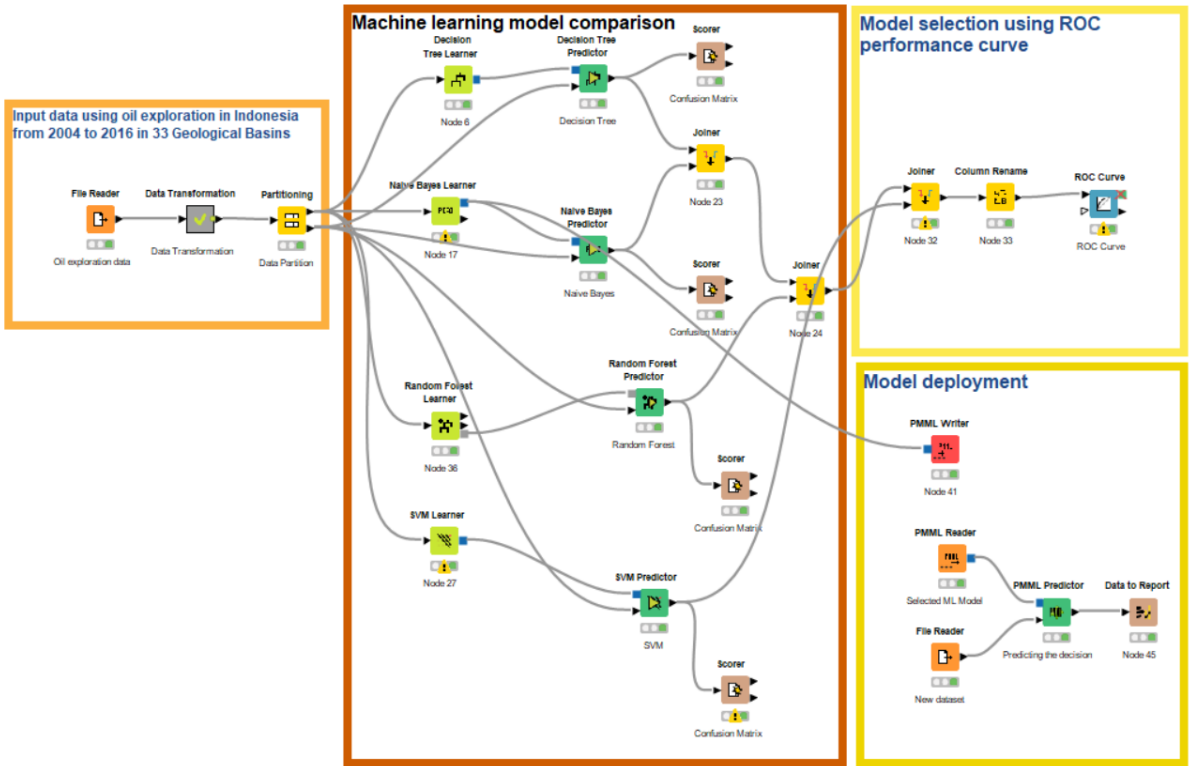

Figure 4 Nodes and workflow of Machine Learning Model Comparison and Selection

The success rate statistically and positively determines the number of wells drilled for all regression models, both fixed and random effects. It can be concluded that the success rate derived from earlier drilling significantly leads to the next drilling effort [2]. The positive result confirms that the higher the success rate the higher chance a company is likely to drill. The result obtains similar results that support earlier pieces of literature on the exploration model [7].

According to ROC found in Table 2 and Figure 5, the findings demonstrate that Naïve Bayes has the most accurate performance based on ROC (94.5%), followed by Decision Tree (92.9%), Random Forest (90.9%), and Support Vector Machine (89.6%). Despite the Decision Tree and Random Forest have higher accuracy over, this study focuses on ROC which is appropriate when the dataset have balance class between drilling and no drilling. These approaches were confirmed by earlier research using classification techniques to predict balance class found in data set [1].

According to Table 1, the dataset has 16.4% of success rate, indicating the drilling class more than 5% of commonly classified as imbalanced dataset. Practically, the selected Naïve Bayes model was applied to assess the decision using a new data test. The findings can diminish the subjective blindness and confirmation bias in the investment decision and bring about a reasonable and orderly exploration and development of oil and gas industry [8].

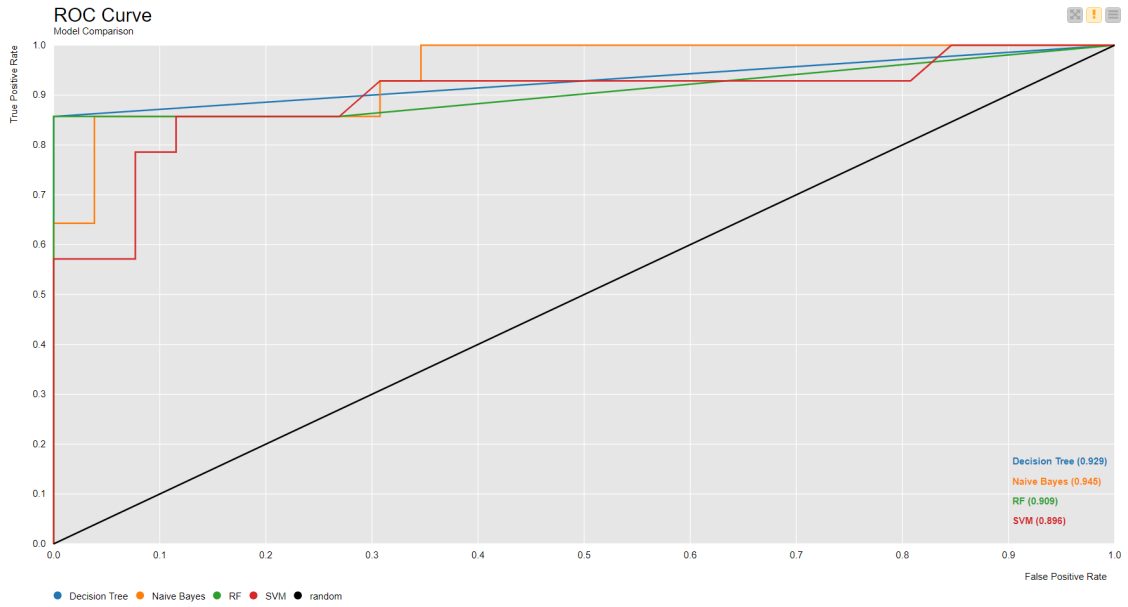

Figure 5 Receiving Operator Characteristics (ROC) Curve of Machine Learning Models

A proxy for the region shows that drilling in eastern areas has increased for years which consists of emerging basins with higher discovery sizes of oil and gas. Besides, these emerging basins are likely to have a higher success rate than mature basins mostly located in western areas. The maturity basins are likely to have less discovery size of oil and gas since these faced a depletion effect naturally. The significance of the geological region of western and eastern Indonesia is in line with earlier studies [7], [11] describing the geological evolution and diversity of sedimentary basins that have been developed and deformed.

This geological regime is the key to understanding the petroleum systems of geological basins to unlock their petroleum potential [11]. Exploration conditions in the geological portions of the Indonesia region are quite different from a petroleum industry standpoint [11]. The exploration and production have been intensified om the western basins of Indonesia. The western relatively has a fewer large potential of undiscovered fields which is most likely made of numerous small to medium size petroleum. On the contrary, the eastern basins are relatively high-risk frontier which is generally under-explored and under-exploited with half of the basins being undrilled [9], [11].

Last , an economic variable such as an expenditure spent by a company for appraisal leads to several wells drilled for all regression models, both fixed and random effects. Appraisal drilling

is a transitioning stage between petroleum exploration and field development. This stage is relatively complicated by which geological characteristics play an important role (Dikkers, 1985). The higher uniqueness and complexity the higher uncertainty that occurs in this early stage of exploration. The result is also in line with a stylized fact in Indonesia's petroleum exploration. Exploration companies selectively invest in proven exploration fields that are likely to have better historical datasets and exploration experience, which in turn, leading to a higher success rate.

4 Conclusion

The purpose of the present paper has been to denote and empirically test a model of exploration behavior of firms on the Indonesian geological basins. The specification of a model offers a way how to identify significant variables and optimize them in the case of investment decisions. The findings have important implications for geological and economic policy: the effect of any change in exogenous variables, like discovery size of oil, success rate, spending for appraisal, and transitioning region, is likely to be fully scrutinized on oil and gas exploration.

From a methodological point of view, these findings highlight the importance of geological and economic factors in the econometric modeling of oil exploration. Subsequently, exploration companies can make a decision and investment plan in oil and gas exploration and development by managing the significant variables ranging from success rate, discovery size of oil, expenditure for appraisal, and region.

In practice, the findings can diminish the subjective blindness and confirmation bias in the investment decision and bring about a reasonable and orderly exploration and development of oil and gas reserves [12]. A decreasing trend is observed for exploration in Western basins across the model, thus suggesting an increased emphasis on the Eastern basins for investment in exploration activities. The results also confirm that there is relatively similar behavior on exploration activity conducted by national oil companies (NOCs) and international oil companies (IOCs). Hence, it can be concluded that there is a weak relation between resource nationalism and exploration intensity.

REFERENCES

[1] Anggraeni, D., Sugiyanto, K., Zam Zam, M. I., & Patria, H. (2022). Stock Price Movement Prediction using Supervised Machine Learning Algorithm: KNIME. Junal Akun Nabelo: Jurnal Akuntansi, Netral, Akuntabel, Objektif. 4(2), 671–681

[2] Boyce, J. R., & Nøstbakken, L. (2011). Exploration and development of U.S. oil and gas fields, 1955-2002. *Journal of Economic Dynamics and Control*, *35*(6), 891–908. https://doi.org/10.1016/j.jedc.2010.12.010

[3] Greiner, A., Semmler, W., & Mette, T. (1989). An Economic Model of Oil Exploration and Extraction. *Computational Economics*, *40*(4), 387–399. https://doi.org/10.1007/s10614-011-9272-0

[4] Kolb, J. A. (1979). An econometric study of discoveries of natural gas and oil reserves in the United States, 1948 to 1970. Ayer Publishing.

[5] Mohn, K., & Misund, B. (2009). Investment and uncertainty in the international oil and gas industry. In *Energy Economics* (Vol. 31). https://doi.org/10.1016/j.eneco.2008.10.001

[6] Mohn, K., & Osmundsen, P. (2008). Exploration economics in a regulated petroleum province: The case of the Norwegian Continental Shelf. *Energy Economics*, *30*(2), 303–320. https://doi.org/10.1016/j.eneco.2006.10.011

[7] Patria, H., & Adrison, V. (2015). Oil Exploration Economics: Empirical Evidence from Indonesian Geological Basins. *Economics and Finance in Indonesia*, *61*(3), 196. https://doi.org/10.7454/efi.v61i3.514

[8] Patria, H. (2020). Determinant factors affeting successful project initiation decision: empirical evidences in Indonesia oil and gas midstream and downstream sector. Disertasi Universitas Indonesia.

[9] Patria, H. (2021). The Role of Success Rate, Discovery, Appraisal Spending, and Transitioning Reion on Exploration Drilling of Oil and Gas in Indonesia in 2004–2015. *Economics and Finance in Indonesia*, *61*(3), 196. https://dx.doi.org/10.47291/efi.v67i2.952

[10] Pindyck, R. S. (1978). The optimal exploration and production of nonrenewable resources. Journal of political economy, 86(5), 841-861.

[11] Satyana, A. H. (2018). Future Petroleum Play Types of Indonesia: Regional Overview. *Proceedings, Indonesian Petroleum Association*, (May 2017). https://doi.org/10.29118/ipa.50.17.554.g

[12] Yuhua, Z., & Dongkun, L. (2009). Investment optimization in oil and gas plays. *Petroleum Exploration and Development*, *36*(4), 535–540. https://doi.org/10.1016/S1876-3804(09)60145-2

[13] Woodmac Research and SKK Migas, Oil Field Exploration Report, 2004 - 2016, unpublished.

[14] Zulfikri, F., Tryanda, D., Syarif, A., & Patria, H. (2021). Predicting Peer to Peer Lending Loan Risk Using Classification Approach. International Journal of Advanced Science Computing and Engineering, 3(2), 94–100. https://doi.org/10.30630/ijasce.3.2.57