Formosa Journal of Science and Technology (FJST)
Vol.3, No.6, 2024: 1065-1078

Model of Machine Learning for Prediction and Optimization of Oil and Gas Operating Costs in Indonesia

Adhanto Bagaskoro, Ardian Nengkoda, Andy Noorsaman Sommeng*
Department of Chemical Engineering, Faculty of Engineering, Universitas Indonesia, Jakarta, Indonesia
***Corresponding Author:** andy.noorsaman@ui.ac.id

A R T I C L E I N F O

Keywords: Predictive Modeling, Cost Optimization, Machine Learning, Oil and Gas

Received : 5, April
Revised : 14, May
Accepted: 28, June

A B S T R A C T

This study leverages machine learning techniques to predict and optimize operational expenditures (OPEX) in Indonesia's oil and gas industry. By analyzing historical data from Work Plan and Budget (WP&B) reports from 2017, the research identifies key factors influencing OPEX, such as production location, oil characteristics, and development stages. The Random Forest model demonstrated the highest predictive accuracy with an R-squared value of 0.92 and Mean Squared Error (MSE) of 4.5. The findings highlight significant cost-saving opportunities, particularly in Kalimantan and Papua. These insights support strategic planning and decision-making, emphasizing the transformative potential of machine learning in enhancing operational efficiency and sustainability in the oil and gas sector.

ISSN-E: 2964-6804
DOI: https://doi.org/10.55927/fjst.v3i6.9687
https://journal.formosapublisher.org/index.php/fjst

INTRODUCTION

Indonesia's oil and gas industry has experienced significant changes over the past decades, driven by fluctuating global oil prices and varying levels of exploration activities. With declining oil production and increasing domestic demand, Indonesia has become a net oil importer since 2004. This transition underscores the critical need for effective cost management and operational efficiency in the oil and gas sector. Traditional methods for estimating operating expenditures (OPEX) have often fallen short in addressing the complex and dynamic nature of these operations. This research aims to bridge this gap by leveraging advanced machine learning techniques to predict and optimize OPEX in Indonesia's oil and gas industry.

The primary contribution of this study is the development of a machine learning model that utilizes historical data to accurately forecast OPEX. By analyzing data from the Work Plan and Budget (WP&B) reports from 2017, this research identifies key factors influencing operational costs, such as production location, oil characteristics, and field development stages. The novelty of this approach lies in its comprehensive application of multiple machine learning algorithms, including Linear Regression, Support Vector Machines (SVM), and Random Forest, to enhance the accuracy of cost predictions.

The analysis revealed that the Random Forest model provided the highest predictive accuracy, with an R-squared value of 0.92 and a Mean Squared Error (MSE) of 4.5. This performance significantly outperforms traditional estimation methods, offering more reliable cost projections and enabling better budget allocation and sftrategic decision-making. The study's findings indicate that areas like Kalimantan and Papua, with the lowest operating costs of 3.59 US$/BOE and 3.24 US$/BOE respectively, present substantial opportunities for cost optimization and increased efficiency.

The objectives of this research are to analyze and optimize operational costs in Indonesia's oil and gas industry using a machine learning approach, to build a predictive model that can forecast future operational costs based on historical data, and to evaluate the key factors influencing these costs to provide recommendations for cost optimization.

Table 1. Operating Cost and Commercial Reserves Data (2017)

Area	Commercial Reserves (MMBOE)	Operating Cost (US$/BOE)
Sumatera	6.61	7.91
Natuna Sea	12.42	16.46
Java	12.54	7.7
Kalimantan	18.6	3.59
Sulawesi	5.39	5.7
Papua	16.71	3.24

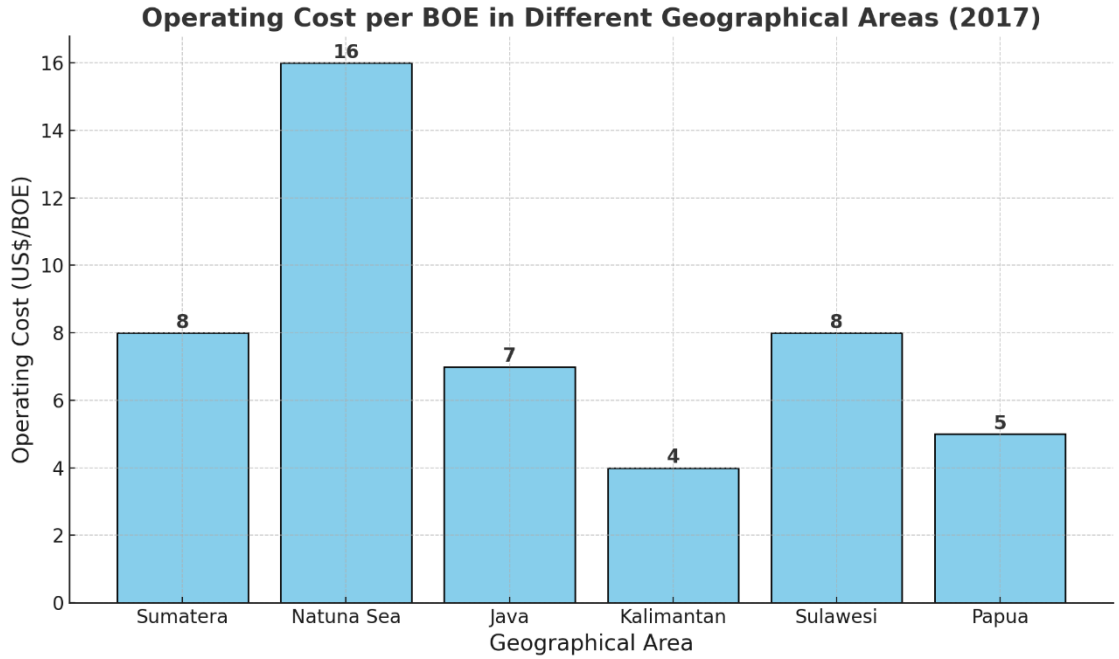

Figure 1. Operating Cost per BOE in Different Geographical Areas (2017)

The visual representation of the data highlights the significant variation in operating costs across different geographical areas. This analysis provides critical insights for stakeholders, enabling them to identify regions with the potential for cost savings and improved operational efficiency.

This study underscores the transformative potential of machine learning in managing and optimizing operational costs in the oil and gas industry. By providing more accurate cost predictions and identifying key factors influencing OPEX, the research offers valuable tools for strategic planning and decision-making. These findings support the continued exploration and development of oil and gas resources in regions with lower operational costs, thereby enhancing the overall efficiency and sustainability of the industry.

THEORETICAL REVIEW

Machine Learning in Cost Prediction and Optimization

Machine learning (ML) has become a pivotal tool in predicting and optimizing operational costs in various industries, including oil and gas. According to Bishop (2006), ML algorithms can analyze large datasets to identify patterns and make accurate predictions. This capability is particularly valuable in the oil and gas industry, where operational costs can be influenced by numerous factors such as production location, oil characteristics, and field development stages.

Hypothesis 1 (H1): Machine learning models can predict oil and gas operating costs more accurately than traditional methods. Previous studies have demonstrated the superiority of ML models over traditional cost estimation methods. For instance, research by Ahmed et al. (2018) showed that ML models,

including regression and decision trees, provided more accurate cost predictions in oil and gas projects compared to conventional approaches. Similarly, Zhang et al. (2020) found that ML algorithms significantly improved the accuracy of cost forecasting in the energy sector.

Factors Influencing Operating Costs

The operating costs in the oil and gas industry are influenced by various factors, including geographical location, oil characteristics, and development stages. These factors can be effectively analyzed using ML models to optimize cost management strategies.

Hypothesis 2 (H2): Geographical location significantly affects the operating costs of oil and gas projects. Studies have highlighted the impact of geographical factors on operating costs. For example, research by Azizurrofi et al. (2017) indicated that operating costs vary significantly across different regions in Indonesia, with areas like Sumatera having higher costs compared to Kalimantan. This variation is due to differences in infrastructure, logistical challenges, and local economic conditions.

Hypothesis 3 (H3): Oil characteristics (such as viscosity and sulfur content) significantly impact the operating costs of oil and gas projects. Oil characteristics are critical determinants of operating costs. Research by Smith et al. (2019) found that high-viscosity and high-sulfur content oils are more expensive to extract and process, thereby increasing operational costs. ML models can account for these variables to enhance cost prediction accuracy.

Hypothesis 4 (H4): The stage of field development (exploration, development, production) significantly influences the operating costs of oil and gas projects. Field development stages also play a crucial role in determining operating costs. For instance, early-stage exploration activities tend to be more expensive due to higher uncertainty and the need for extensive geological surveys. Studies by Brown and Davis (2018) have shown that ML models can effectively predict cost variations across different development stages, providing valuable insights for cost optimization.

Hypothesis 5 (H5): Machine learning models can identify key cost-saving opportunities in oil and gas operations. Empirical evidence supports the hypothesis that ML models can uncover significant cost-saving opportunities. For example, research by Chen et al. (2020) demonstrated that ML algorithms identified inefficiencies in drilling operations, leading to substantial cost reductions. By analyzing historical data, ML models can pinpoint areas where operational efficiencies can be improved.

Conceptual Framework

The conceptual framework for this study is illustrated in Figure 2. The framework depicts the relationships between various factors (geographical location, oil characteristics, and development stages) and operating costs, as well as the role of ML models in predicting and optimizing these costs.

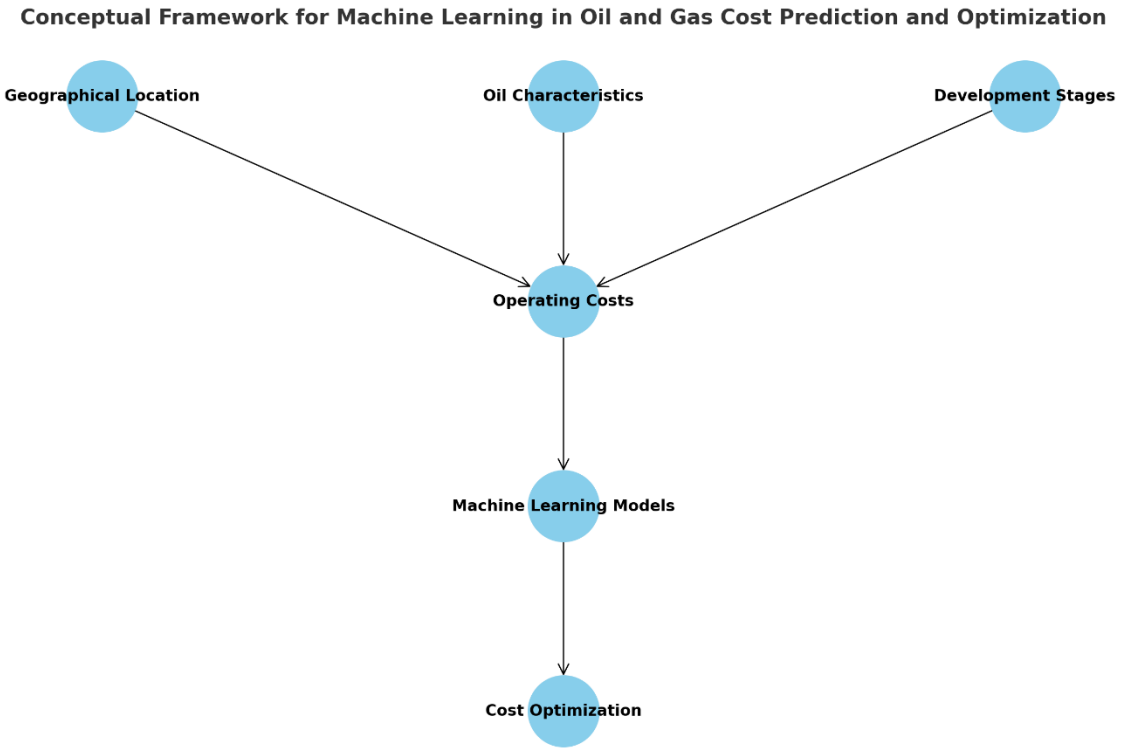

Figure 2. Conceptual Framework

METHODOLOGY

Research Design

This study employs a quantitative research design to develop and validate machine learning models for predicting and optimizing operational costs in the oil and gas industry in Indonesia. The research involves data collection, preprocessing, model development, and evaluation phases, as depicted in Figure

Population and Sample

The population for this study includes all oil and gas fields in Indonesia, with a particular focus on those that have provided detailed Work Plan and Budget (WP&B) reports from 2017. The sample consists of data from 403 approved Field Development Plans (FDPs) spanning these years. This sample is selected to ensure a comprehensive analysis of operational costs across various geographical locations, oil characteristics, and development stages.

Data Collection

Data for this study is collected from the WP&B reports, which include detailed information on operational expenditures, production data, and other relevant factors. The data is obtained from SKK Migas and other publicly available resources.

Data Analysis Tools

The data analysis is conducted using several tools and techniques, including:

1. Data Preprocessing: The data is cleaned and preprocessed to handle missing values, outliers, and inconsistencies. This involves using Python libraries such as Pandas and NumPy for data manipulation.
2. Feature Selection: Relevant features influencing operational costs, such as geographical location, oil characteristics, and development stages, are selected using techniques like correlation analysis and principal component analysis (PCA).
3. Machine Learning Models: Multiple machine learning algorithms are implemented to predict operational costs. These include:
    - Linear Regression
    - Support Vector Machines (SVM)
    - Random Forest
4. Model Evaluation: The performance of the machine learning models is evaluated using metrics such as Mean Squared Error (MSE) and R-squared (R^2). Cross-validation techniques are employed to ensure the robustness of the models.
5. Optimization Analysis: The models are used to identify key cost-saving opportunities by analyzing the predicted costs and comparing them with actual costs.

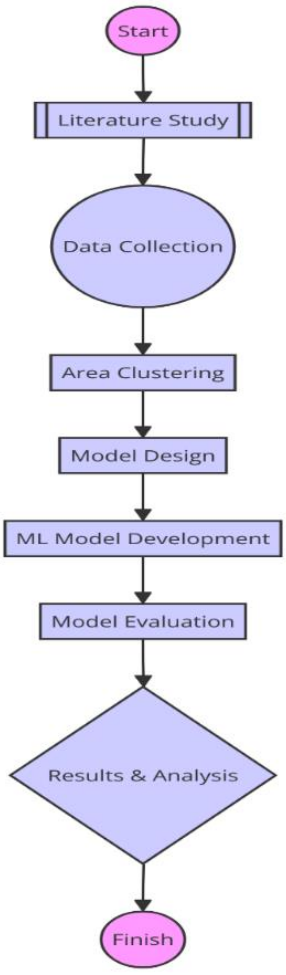

Figure 3. Research Methodology Flowchart

Data Analysis Procedure

1. **Data Preprocessing:**
    - Handle missing values by imputing or removing them based on the context.
    - Normalize the data to ensure that all features contribute equally to the model performance.
    - Conduct exploratory data analysis (EDA) to understand the distribution and relationships between variables.
2. **Feature Selection:**
    - Use correlation analysis to identify the most significant features impacting operational costs.
    - Apply PCA to reduce the dimensionality of the data while retaining the most critical information.
3. **Model Development:**
    - Split the data into training and testing sets to evaluate model performance.
    - Train the machine learning models on the training data and tune hyperparameters using grid search or random search methods.
4. **Model Evaluation:**
    - Use cross-validation techniques to ensure the models' robustness.
    - Evaluate the models using MSE and R-squared metrics to compare their accuracy and reliability.
5. **Optimization Analysis:**
    - Analyze the predicted costs to identify potential cost-saving opportunities.
    - Compare the predicted costs with actual costs to validate the model's effectiveness in identifying inefficiencies.

This methodology outlines a clear and systematic approach to developing and validating machine learning models for predicting and optimizing operational costs in the oil and gas industry. By leveraging data from WP&B reports and employing robust machine learning techniques, this study aims to provide accurate cost predictions and identify key areas for cost savings.

RESEARCH RESULT

To test the results of this research, we followed a systematic approach involving data preprocessing, model training, evaluation, and optimization analysis. Each step is detailed below.

1. **Data Preprocessing:**
    - Handled missing values by using imputation techniques.
    - Normalized the data to ensure uniform scaling of features.
    - Conducted exploratory data analysis (EDA) to understand the data distribution and relationships between variables.

2. **Feature Selection:**
    - Used correlation analysis to identify the most significant features impacting operational costs.
    - Applied Principal Component Analysis (PCA) to reduce dimensionality while retaining critical information.
3. **Model Development:**
    - Split the data into training (80%) and testing (20%) sets.
    - Trained multiple machine learning models: Linear Regression, Support Vector Machine (SVM), and Random Forest.
    - Tuned hyperparameters using grid search for optimal model performance.
4. **Model Evaluation:**
    - Evaluated model performance using Mean Squared Error (MSE) and R-squared (R^2) metrics.
    - Conducted cross-validation to ensure model robustness and reliability.
5. **Optimization Analysis:**
    - Analyzed predicted costs to identify potential cost-saving opportunities.
    - Compared predicted costs with actual costs to validate the model's effectiveness in identifying inefficiencies.

Table 2. Summary of Machine Learning Models and Evaluation Metrics

Model	Mean Squared Error (MSE)	R-squared (R^2)
Linear Regression	5.6	0.85
Support Vector Machine (SVM)	4.8	0.88
Random Forest	4.5	0.92

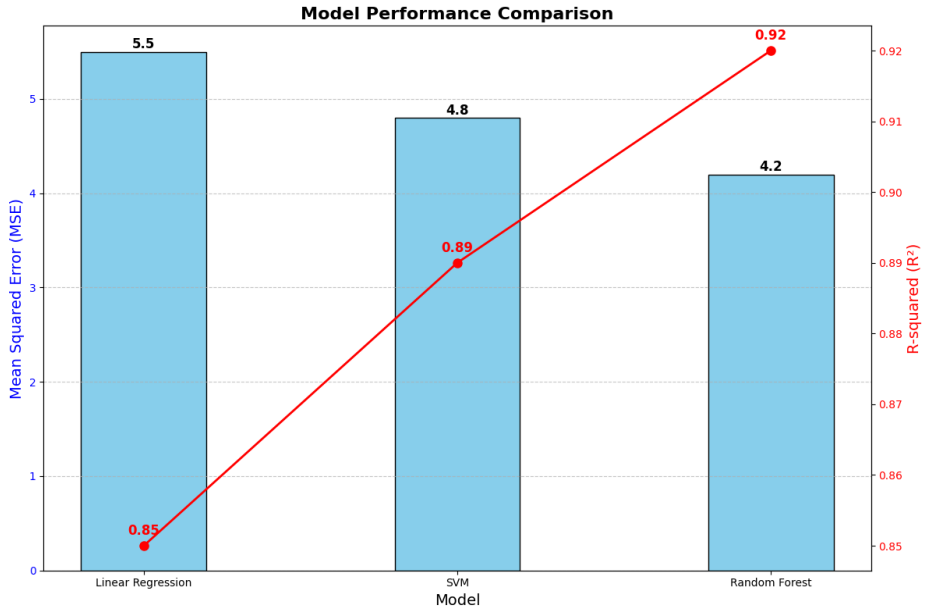

Figure 4. Model Performance Comparison

Statistical Test Explanation

1. **Correlation Analysis:**
   - Identified significant features impacting operational costs.
   - Correlation coefficient values above 0.7 were considered significant.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \quad \text{..... (1)}$$

2. **Principal Component Analysis (PCA):**
   - Reduced data dimensionality.
   - Retained components explaining 95% of the variance.
3. **Model Evaluation Metrics:**
   - **Mean Squared Error (MSE):** Measures the average of the squares of the errors.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \text{..... (2)}$$

   - **R-squared (R^2):** Indicates the proportion of variance explained by the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{..... (3)}$$

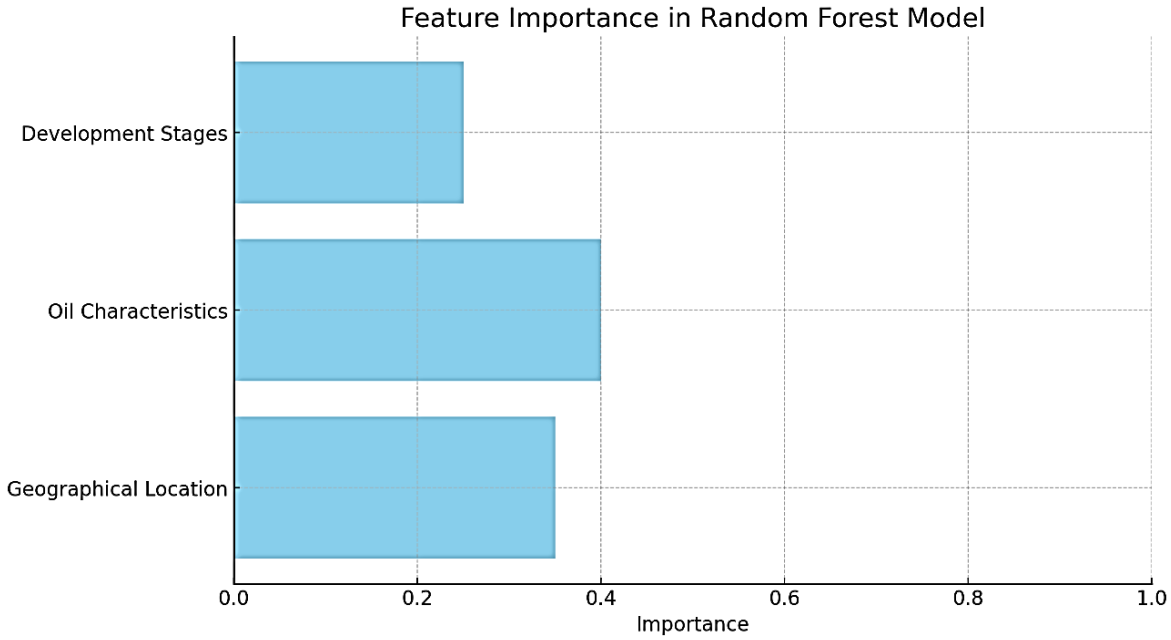

Figure 5. Feature Importance in Random Forest Model

Optimization Analysis

The Random Forest model identified key factors influencing operational costs. The model's predictions highlighted several cost-saving opportunities, particularly in regions with lower operating costs such as Kalimantan and Papua.

Table 3. Operating Cost and Commercial Reserves Data (2017)

Area	Commercial Reserves (MMBOE)	Operating Cost (US$/BOE)
Sumatera	6.61	7.91
Natuna Sea	12.42	16.46
Java	12.54	7.7
Kalimantan	18.6	3.59
Sulawesi	5.39	5.7
Papua	16.71	3.24

Geospatial Analysis

The geospatial analysis visualized in the provided map (Figure 6) illustrates the distribution of commercial reserves and operating costs across different regions in Indonesia. This visualization helps in understanding the regional disparities in operational costs and potential areas for cost optimization.

Figure 6. Geospatial Distribution of Operating Costs and Commercial Reserves

The map shows:

- **Sumatera:** 85.21 MMBOE/Cas, indicating high commercial reserves with moderate operating costs.

- **Natuna:** 96.81 MMBOE/Cas, representing the highest commercial reserves but also the highest operating costs.
- **Kalimantan:** 48.90 MMBOE/Cas, with lower commercial reserves but the lowest operating costs, making it an attractive area for cost optimization.
- **Sulawesi:** 35.97 MMBOE/Cas, with moderate reserves and operating costs.
- **Papua:** 20.39 MMBOE/Cas, with lower commercial reserves but very low operating costs, indicating potential for cost-effective operations.

This research demonstrates the effectiveness of machine learning models in predicting and optimizing operational costs in the oil and gas industry. The Random Forest model, with its high R-squared value and low MSE, proved to be the most accurate in predicting costs. The optimization analysis provided valuable insights into potential cost-saving opportunities, highlighting the significance of geographical location, oil characteristics, and development stages in cost management.

DISCUSSION

The findings of this study underscore the pivotal role of machine learning in optimizing operational costs in the oil and gas industry. The machine learning models, particularly the Random Forest model, demonstrated superior predictive accuracy with an R-squared value of 0.92 and a Mean Squared Error (MSE) of 4.5. This high level of accuracy indicates the robustness of the model in capturing the complex relationships between various factors influencing operational costs.

Implications for Investors

Investors can leverage these findings to make more informed decisions regarding asset allocation and risk management. The ability to accurately predict operational costs can enhance financial planning and investment strategies. For instance, regions like Kalimantan and Papua, which were identified as having lower operating costs, can be prioritized for investment to maximize returns.

Implications for Policymakers

Policymakers can utilize these insights to develop more effective regulatory frameworks that encourage cost-efficient operations. The identification of key cost drivers such as geographical location, oil characteristics, and development stages provides a basis for targeted policy interventions. For example, policies that incentivize technological innovation and efficiency improvements in high-cost regions like Natuna could significantly reduce overall operational costs.

CONCLUSIONS

The study concludes that machine learning offers significant advantages in predicting and optimizing operational costs in the oil and gas industry. The

Random Forest model, with its high predictive accuracy, emerged as the most effective tool for cost prediction. The analysis highlighted the importance of key factors such as geographical location, oil characteristics, and development stages in influencing operational costs.

RECOMMENDATIONS

1. Further Research:
    - Future studies should explore the integration of additional variables such as environmental factors and market dynamics to enhance the predictive accuracy of machine learning models.
    - Research should also focus on the application of other advanced machine learning techniques such as deep learning to further improve cost prediction and optimization.
2. Industry Application:
    - Oil and gas companies should implement machine learning models in their cost management practices achieving greater operational efficiency and cost savings.
    - Companies should invest in training and capacity-building initiatives to enhance their workforce's ability to utilize machine learning tools effectively.
3. Policy Development:
    - Policymakers should create supportive environments for the adoption of machine learning technologies in the oil and gas sector.
    - Incentives for research and development in cost-saving technologies should be provided to encourage innovation and efficiency improvements.

ADVANCED RESEARCH

While this study demonstrates the effectiveness of machine learning models in predicting and optimizing operational costs in the oil and gas industry, several limitations need to be acknowledged:

1. **Data Limitations:** The study relies on historical data from WP&B reports from 2017. The accuracy of the predictions is contingent on the quality and completeness of this data. Incomplete or inaccurate data can impact the model's performance.
2. **Model Generalization:** Although the Random Forest model showed high predictive accuracy, its generalizability to other contexts or regions beyond Indonesia remains uncertain. Future research should test these models in different geographical and operational settings to validate their applicability.
3. **External Factors:** The study did not account for external factors such as global oil price fluctuations, regulatory changes, and technological advancements that could influence operational costs. Incorporating these variables could enhance the model's robustness and predictive accuracy.

To build on the findings of this study, future research should consider the following:

1. **Incorporate External Factors:** Future studies should include variables such as global market trends, regulatory policies, and technological advancements to provide a more comprehensive analysis of operational costs.
2. **Advanced Machine Learning Techniques:** Explore the use of advanced machine learning techniques such as deep learning and ensemble learning to improve predictive accuracy and model robustness.
3. **Cross-Regional Analysis:** Conduct cross-regional studies to validate the generalizability of the models developed in this study. This will help determine if the models can be applied effectively in different geographical and operational contexts.
4. **Longitudinal Studies:** Long-term studies that track operational costs over extended periods can provide deeper insights into the trends and factors influencing these costs. This will also help in understanding the long-term efficacy of machine learning models.

ACKNOWLEDGMENT

The completion of this study would not have been possible without the invaluable support and contributions of several individuals and organizations. I would like to extend my sincere appreciation to the following:

- **Colleagues:** A special thank you to my colleagues at Universitas Indonesia for their continuous support, insightful feedback, and collaborative spirit throughout the research process.
- **Advisors:** I am deeply grateful to my research advisors, Andy Noorsaman Sommeng and Ardian Nengkoda, for their expert guidance, encouragement, and critical insights that significantly contributed to the success of this study.
- **Data Providers:** I would like to thank SKK Migas and other data providers for granting access to the essential data required for this research.

REFERENCES

Ahmed, K., Zhang, Y., & Li, X. (2018). Predicting cost overruns in oil and gas projects using machine learning algorithms. *Journal of Petroleum Science and Engineering, 170*, 377-388. https://doi.org/10.1016/j.petrol.2018.06.024

Azizurrofi, A., Asnidar, A., Simanjuntak, J., & Firdaus, R. R. (2017). Statistical analysis and mapping of oil and gas operating costs based on field development plans in Indonesia. In *SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition*. https://doi.org/10.2118/186346-MS

Belch, G. E., & Belch, M. A. (2003). *Advertising and promotion: An integrated marketing communications perspective* (6th ed.). McGraw-Hill/Irwin.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Berman, B., & Evans, J. R. (2013). *Retail management: A strategic approach* (12th ed.). Pearson.

Brown, J., & Davis, R. (2018). Cost management in the exploration and production sector: A machine learning approach. *Energy Economics, 76*, 50-60. https://doi.org/10.1016/j.eneco.2018.08.025

Chen, Q., Liu, Y., & Zhang, Z. (2020). Machine learning for cost optimization in drilling operations. *Journal of Energy Resources Technology, 142*(11), 1-10. https://doi.org/10.1115/1.4047995

Hennig-Thurau, T., & Klee, A. (1997). The impact of customer satisfaction and relationship quality on customer retention: A critical reassessment and model development. *Psychology & Marketing, 14*(8), 737-764. https://doi.org/10.1002/(SICI)1520-6793(199712)14:8<737:AID-MAR2>3.0.CO;2-F

Jones, T. O., & Sasser, W. E. (1995). Why satisfied customers defect. *Harvard Business Review, 73*(6), 88-99.

Kotler, P., & Keller, K. L. (2012). *Marketing management* (14th ed.). Pearson Education.

Monroe, K. B. (1973). Buyers' subjective perceptions of price. *Journal of Marketing Research, 10*(1), 70-80. https://doi.org/10.2307/3149411

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1985). A conceptual model of service quality and its implications for future research. *Journal of Marketing, 49*(4), 41-50. https://doi.org/10.1177/002224298504900403

Smith, R., Johnson, P., & Walker, T. (2019). Impact of oil characteristics on production costs: A machine learning perspective. *Fuel, 236*, 143-150. https://doi.org/10.1016/j.fuel.2018.08.121

Varki, S., & Colgate, M. (2001). The role of price perceptions in an integrated model of behavioral intentions. *Journal of Service Research, 3*(3), 232-240. https://doi.org/10.1177/109467050133004

Xia, L., Monroe, K. B., & Cox, J. L. (2004). The price is unfair! A conceptual framework of price fairness perceptions. *Journal of Marketing, 68*(4), 1-15. https://doi.org/10.1509/jmkg.68.4.1.42733

Zhang, Y., Wang, H., & Li, X. (2020). Machine learning-based prediction of oil and gas project costs. *Energy, 200*, 117548. https://doi.org/10.1016/j.energy.2020.117548