ABSTRAK

Kebutuhan energi global akan minyak dan gas bumi terus meningkat, sedangkan cadangan yang dapat dieksploitasi semakin berkurang. Identifikasi zona potensial hidrokarbon secara tradisional dilakukan dengan menganalisis data *well log* yang memerlukan waktu lama dan ketelitian tinggi. Tantangan utama dalam klasifikasi otomatis menggunakan *machine learning* adalah distribusi kelas yang sangat tidak seimbang (*imbalanced*), di mana zona *Oil* dan Gas jauh lebih sedikit dibandingkan zona *NonReservoir* dan *NoHC*. Oleh karena itu, diperlukan metode untuk menangani ketimpangan distribusi tersebut sekaligus mengidentifikasi fitur-fitur pada *well log* yang paling berpengaruh dalam proses klasifikasi.

Penelitian ini menerapkan model *XGBoost* dengan metode *Weighted-XGBoost* dan optimasi *hyperparameter* menggunakan *GridSearchCV* untuk mengklasifikasikan zona potensial hidrokarbon ke dalam empat kelas yaitu *NonReservoir*, *NoHC*, *Oil*, dan Gas pada data *well log* sumur Poolowanna 1 yang diperoleh dari *Australian Government Geoscience Australia*. Data tersebut terdiri dari 20.082 baris dan 20 kolom. Metode *Weighted-XGBoost* diterapkan dengan menggunakan fungsi *compute_sample_weight* untuk menentukan bobot invers proporsional pada setiap kelas, sedangkan *GridSearchCV* dengan *Stratified 5-Fold Cross Validation* digunakan untuk mencari kombinasi *hyperparameter* terbaik dari 216 kombinasi yang diuji. Selain itu, analisis *feature importance* dilakukan untuk mengidentifikasi fitur *well log* yang paling berpengaruh terhadap hasil klasifikasi.

Penerapan metode *Weighted-XGBoost* dan optimasi *hyperparameter* menggunakan *GridSearchCV* menghasilkan model dengan nilai *precision macro* sebesar 99.85% pada *test set* yang terdiri dari 4.017 sampel, dengan *precision* per kelas *NonReservoir* 99.8%, *NoHC* 99.6%, *Oil* 100%, dan Gas 100%. Kombinasi *hyperparameter* terbaik yang diperoleh adalah *n_estimators* = 200, *learning_rate* = 0,05, *max_depth* = 6, *subsample* = 0,8, dan *colsample_bytree* = 1,0. Analisis *feature importance* mengidentifikasi SP (*Spontaneous Potential*), RHOB (*Bulk Density*), dan *Vshale* (*Shale Volume*) sebagai tiga fitur terpenting dari *well log* yang memberikan kontribusi sebesar 71% dari total *gain* model, sedangkan fitur NPHI, DRHO, dan RMED dieliminasi karena nilai *feature importance*-nya di bawah ambang batas yang ditentukan.

Kata Kunci: *XGBoost*, Klasifikasi Hidrokarbon, *Well Log*, *Imbalanced Dataset*, *Feature Importance*

ABSTRACT

The global demand for oil and natural gas continues to rise while exploitable reserves are increasingly depleted. Conventional identification of potential hydrocarbon zones is conducted through well log data analysis, which requires considerable time and high precision. The primary challenge in automated classification using machine learning is the highly imbalanced class distribution, where Oil and Gas zones are far less frequent than NonReservoir and NoHC zones, necessitating specialized methods to handle this imbalance while simultaneously identifying the most significant well log features in the classification process.

This study implements an XGBoost model with the Weighted-XGBoost method and hyperparameter optimization using GridSearchCV to classify potential hydrocarbon zones into four classes, namely NonReservoir, NoHC, Oil, and Gas, using well log data from the Poolowanna 1 well obtained from the Australian Government Geoscience Australia, consisting of 20,082 rows and 20 columns. The Weighted-XGBoost method was applied through the compute_sample_weight function to assign inversely proportional weights to each class, while GridSearchCV with Stratified 5-Fold Cross Validation was employed to identify the best hyperparameter combination from 216 tested combinations. Furthermore, feature importance analysis was conducted to identify the most significant well log features contributing to the classification results.

The application of the Weighted-XGBoost method and hyperparameter optimization using GridSearchCV produced a model achieving a macro precision of 99.85% on a test set of 4,017 samples, with per-class precision scores of 99.8% for NonReservoir, 99.6% for NoHC, and 100% for both Oil and Gas. The best hyperparameter combination obtained was n_estimators = 200, learning_rate = 0.05, max_depth = 6, subsample = 0.8, and colsample_bytree = 1.0. Feature importance analysis identified SP (Spontaneous Potential), RHOB (Bulk Density), and Vshale (Shale Volume) as the three most significant well log features with a collective contribution of 71% of the total model gain, while features NPHI, DRHO, and RMED were eliminated as their feature importance values fell below the established threshold.

***Keywords**: XGBoost, Hydrocarbon Classification, Well Log, Imbalanced Dataset, Feature Importance*