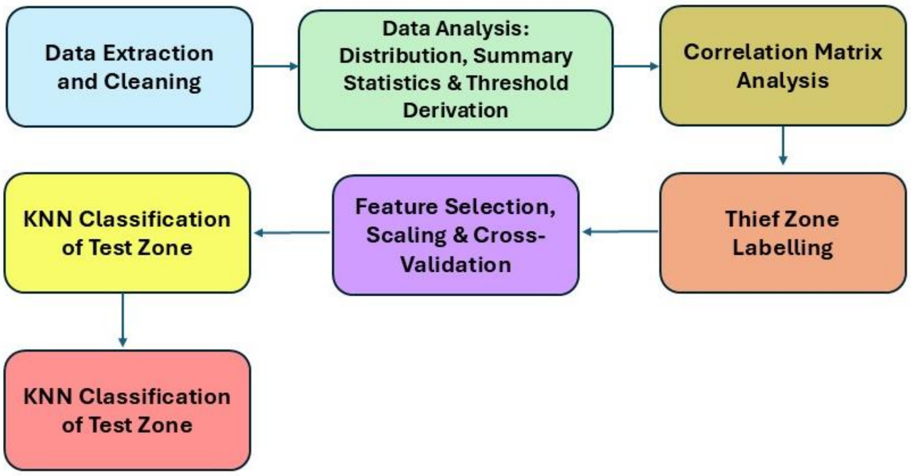
Classification of a Thief Zone using the KNN Algorithm

This task involved using the K-Nearest Neighbours (KNN) algorithm to classify a suspected thief zone in the Lismore field based on petrophysical data. Thief zones are high-permeability intervals that divert injected fluids away from productive zones due to their superior flow capacity as compared with surrounding formations (Liu et al., 2021). The goal was to define a clear classification criteria using geological insight and statistical thresholds, then apply a supervised KNN model to evaluate the test zone.

Workflow Overview


Data Extraction and Cleaning
Data Analysis: Distribution, Summary Statistics & Threshold Derivation
Correlation Matrix Analysis
Thief Zone Labelling
Feature Selection, Scaling & Cross-Validation
KNN Classification of Test Zone
KNN Classification of Test Zone

1. Data Processing and Thief Zone Criteria

The dataset was cleaned by removing non-numeric entries, empty rows/columns, and zones with extensive missing data. The Zechstein zone was excluded due to incomplete petrophysical information.

Histograms and summary statistics (e.g., percentiles, mean, standard deviation) were used to explore the distribution of petrophysical features and identify outliers that might indicate thief zones. Correlation analysis showed PHIF (porosity) was strongly correlated with NTG (r = 0.86) and KLOGH metrics, so it was excluded to reduce redundancy. Similarly, KLOGH_harmonic and KLOGH_geometric were highly correlated (r = 0.95) and offered limited additional value.

Based on this, thief zones were defined using:

- **Primary Indicator – High Permeability:** *KLOGH_arithmetic > 66 mD*, reflecting the 75th percentile. Although the 90th percentile was 139 mD, the 66 mD threshold captured a meaningful contrast in the data distribution, including zones with strong but not extreme permeability.

- **Supporting Indicator – Net-to-Gross Ratio (NTG):** *NTG < 0.75* was selected slightly above the median (0.723) to capture zones with lower NTG that showed thief-like behaviour. These may represent thin, heterogeneous, or high-perm streaks—consistent with trends noted by Liu et al. (2021).

Figure 1. Distribution of Petrophysical Parameter

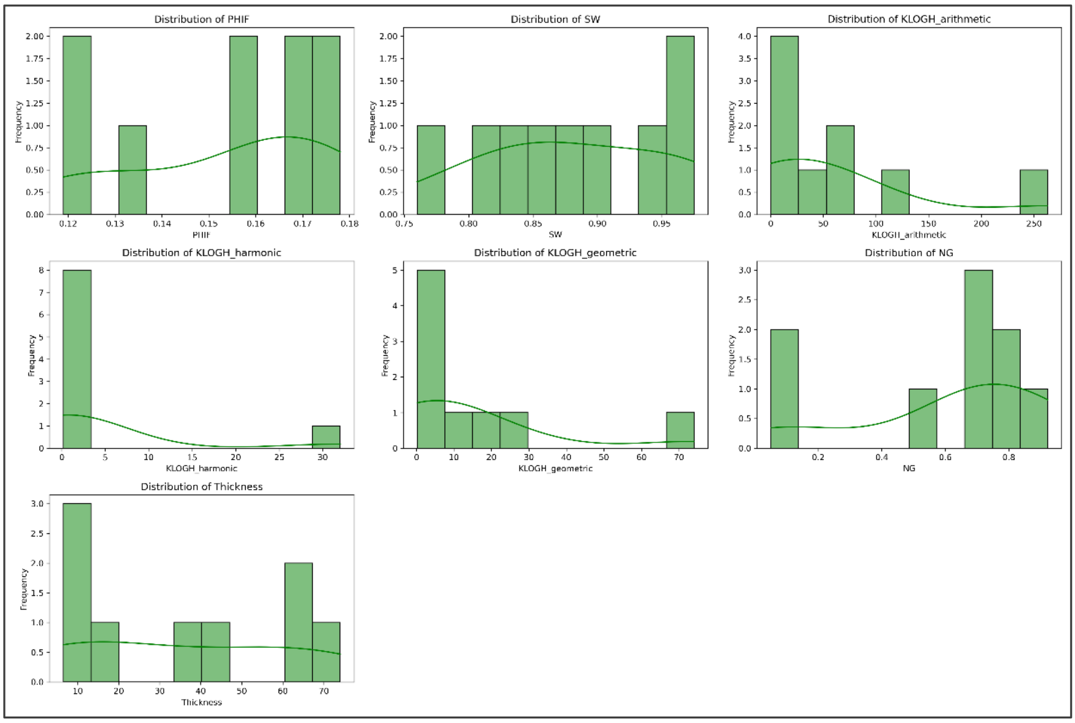
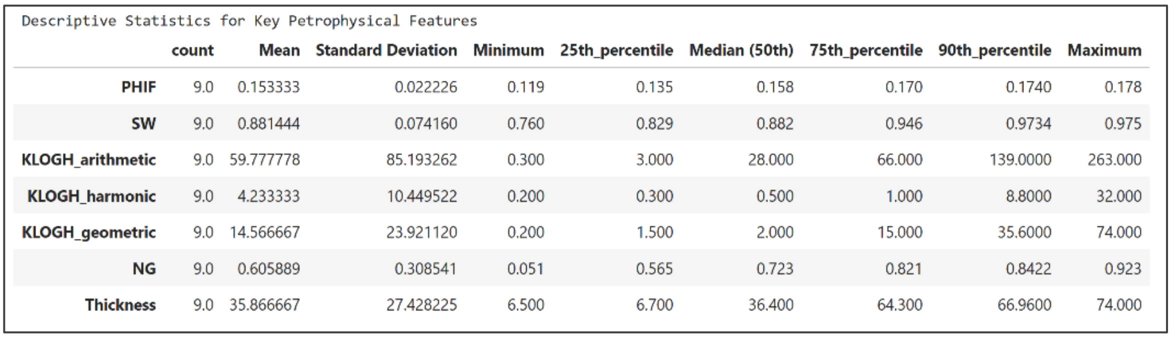


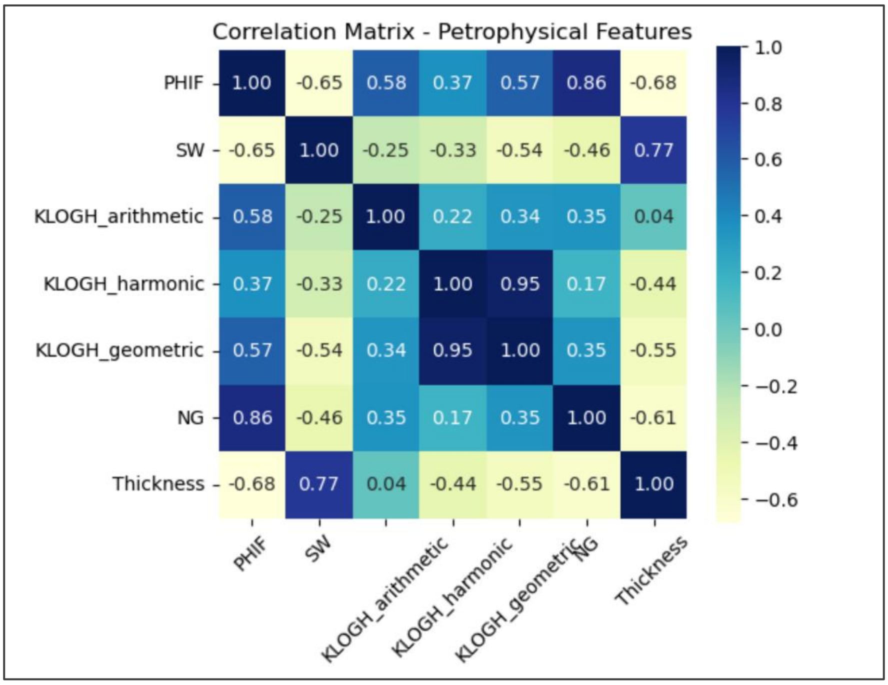
Figure 2: Summary Statistics



Descriptive Statistics for Key Petrophysical Features

	count	Mean	Standard Deviation	Minimum	25th_percentile	Median (50th)	75th_percentile	90th_percentile	Maximum
PHIF	9.0	0.153333	0.022226	0.119	0.135	0.158	0.170	0.1740	0.178
SW	9.0	0.881444	0.074160	0.760	0.829	0.882	0.946	0.9734	0.975
KLOGH_arithmetic	9.0	59.777778	85.193262	0.300	3.000	28.000	66.000	139.0000	263.000
KLOGH_harmonic	9.0	4.233333	10.449522	0.200	0.300	0.500	1.000	8.8000	32.000
KLOGH_geometric	9.0	14.566667	23.921120	0.200	1.500	2.000	15.000	35.6000	74.000
NG	9.0	0.605889	0.308541	0.051	0.565	0.723	0.821	0.8422	0.923
Thickness	9.0	35.866667	27.428225	6.500	6.700	36.400	64.300	66.9600	74.000

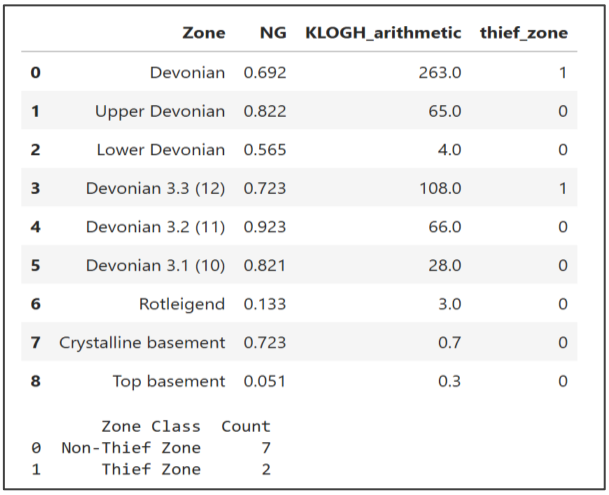
Figure 3: Correlation Heatmap



2. Thief Zone Labelling

Zones were labelled as “thief” (1) or “non-thief” (0) based on the above thresholds. This binary classification formed the target variable (label) for supervised learning. Out of 9 valid zones, 2 were labelled as thief zones.

Figure 4: Visualization of KNN Classification



	Zone	NG	KLOGH_arithmetic	thief_zone
0	Devonian	0.692	263.0	1
1	Upper Devonian	0.822	65.0	0
2	Lower Devonian	0.565	4.0	0
3	Devonian 3.3 (12)	0.723	108.0	1
4	Devonian 3.2 (11)	0.923	66.0	0
5	Devonian 3.1 (10)	0.821	28.0	0
6	Rotleigend	0.133	3.0	0
7	Crystalline basement	0.723	0.7	0
8	Top basement	0.051	0.3	0

```
       Zone Class  Count
0  Non-Thief Zone      7
1      Thief Zone      2
```

3. Feature Selection and Scaling

The eight available petrophysical features (Top, Base, SW, KLOGH_arithmetic, KLOGH_harmonic, KLOGH_geometric NTG, and Thickness) were standardized using **StandardScaler**. This ensured all features contributed equally to distance calculations in the KNN model.

4. Cross-Validation and Selecting Optimal k

Given the small dataset (9 usable zones), we used **Leave-One-Out Cross-Validation (LOOCV)** to evaluate model accuracy, as it maximizes training data while testing each point once (Kohavi, 1995).

To validate the LOOCV results and account for class imbalance (only 2 thief zones), we also used **Stratified K-Fold Cross-Validation**. Both methods returned consistent average accuracies: 77.8% (LOOCV) and 77.5% (Stratified K-Fold).

Although accuracy was stable across all tested values of k, we selected **k = 3** because it helps avoid noise (like with k=1) while still being small enough to reflect local data patterns, and it is a common default in small KNN datasets (Altman, 1992).

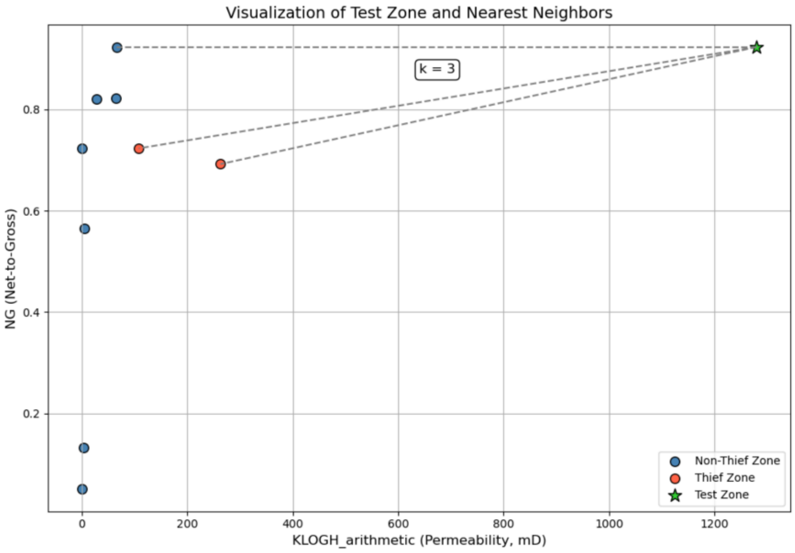
5. KNN Model and Classification

The standardized dataset was used to train a KNN model with k=3, using all eight petrophysical parameters. The test zone, defined by its own petrophysical values, was then classified using this model.

The model predicted the test zone as a **thief zone**, based on the majority vote of its three nearest neighbours (2 thief zones, 1 non-thief).

Notably, the test point had a **high NTG value (0.923)** — above the < 0.75 threshold we used to label thief zones in the training set. This emphasizes that while the NTG threshold helped define known thief zones, it wasn't directly applied to the test point. Instead, the model looked at how similar the test zone was to nearby zones based on all its features, showing that KNN is more flexible than using fixed cutoff rules.

Figure 5: Visualization of KNN Classification



Limitations

While the model performed well, a few limitations remain. Thief zone behaviour may also be influenced by structural factors like faults, fractures, and depositional patterns (Liu et al., 2021), which are not captured in petrophysical logs alone.

The small dataset size also limits how broadly the thresholds and model performance can be applied to other reservoirs. Notwithstanding this method provides a solid foundation for early thief zone identification and can be improved with additional subsurface data such as core samples, structural information, or production history.

Conclusion

By combining geological insight, statistical analysis, and model validation, the KNN approach successfully classified the test zone. The result supported our initial suspicion that the zone may be a thief zone.

Even with a small dataset, using both LOOCV and Stratified K-Fold provided confidence in the model's reliability and justified our choice of $k = 3$.

Overall, the workflow offers a practical tool for early thief zone identification and can guide similar petrophysical analyses in other reservoirs.

References

1. Altman, N. S. (1992). An introduction to kernel and nearest-neighbour non-parametric regression. *The American Statistician*, 46(3).
2. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (pp. 1137–1145).
3. Liu, H., Shi, K., Liu, B., Song, X., Deng, L., Guo, R., Tian, Z., Li, Y., Deng, Y., & Wang, G. (2021). The Characteristics and Origins of Thief Zones in the Cretaceous Limestone Reservoirs of Central and Southern Mesopotamian Basin. *Journal of Petroleum Science and Engineering*, 201, 108395. [Ref]