INTELLIGENT CLUSTERING AND CLASSIFICATION OF THE GAS-CRUDE OIL SEPARATION PROCESS USING K-MEANS AND RANDOM FOREST: A DATA SCIENCE BASED APPROACH

UDC:665.6:519.68:004.8
Original scientific paper https://doi.org/10.46793/aeletters.2025.10.4.4

Rubén Darío Vega Mejía[1*], Natali Lisbeth Campos Rodríguez[1], Omar José Sánchez Roca[1], Cristhian Ronceros Morales[2]

[1] University of Oriente, School of Engineering and Applied Sciences (EICA), Department of Petroleum Engineering, Maturin, Venezuela
[2] Technological University of Peru, Department of Systems Engineering and Computer Science, Ica, Peru

Abstract:
Applying data science methodology, the study examined the separation of gas and crude oil in the state of the QE2 compressor in Monagas, Venezuela. It began with a descriptive statistical study that found and eliminated 2.22% of anomalous data, revealing a trimodal behavior for crude oil and a bimodal for gas. With skewness and a coefficient of determination (R^2) where 0.7645 for the gas-crude ratio, both variables had a coefficient of variation greater than 20%. The K-means algorithm was used, which found four well-formed clusters. However, the Kruskal-Wallis method could not find statistically significant differences between them, suggesting that the variability is due to different operating rules, crude types or process errors, rather than clearly differentiated groups. Finally, a Random Forest algorithm was developed with one hundred trees. The most significant achieved an accuracy of 0.9929. Despite an initial Gini value of 0.725 (moderate impurity), it was segmented into two branches. The branch with a raw value ≤1.15 Thousands of Barrels of Crude Oil per Day (MBNPD) showed superior performance, with a Gini value of 0.01, indicating near-perfect purity. This shows that this branch classifies with high accuracy.

ARTICLE HISTORY
Received: 25 August 2025
Revised: 13 November 2025
Accepted: 27 November 2025
Published: 15 December 2025

KEYWORDS
Artificial intelligence, Machine learning, Descriptive statistics, Oil production, Natural gas, Gini value

1. INTRODUCTION

When oil is found in subterranean reservoirs under high pressure, natural gas is usually dissolved in the crude oil. During production, when the oil is pumped to the surface, a decrease in pressure is experienced, which causes the release of this gas [1,2]. Consequently, before starting the refining process, it is essential to separate the dissolved gas from the oil [3]. This phenomenon explains why, in the extraction and processing stages, it is often necessary to divide the produced fluid into its main components: oil, gas, water and solids (sediments). For this purpose, specialized equipment is used that applies different separation principles, such as gravity decanting, thermal separation, electrostatic separation, and even combinations of these methods [4,5]. Associated gas is considered a valuable resource in the oil industry, mainly due to its composition, which usually includes liquid hydrocarbons heavier than methane, such as ethane, propane, and butane, increasing its economic value and usefulness compared to dry natural gas [6,7].

In this context, the primary separation of gas and condensate streams, coming from production wells or previous oil separation stages, constitutes a critical operation, as it largely determines the

*CONTACT: Rubén Darío Vega Mejía, e-mail: rvegas@udo.edu.ve

quality of the final products and defines the operating conditions for downstream processes. Efficiency at this stage directly impacts system stability, overall energy efficiency and emission reduction [8,9]. This separation is carried out by high- or low-pressure separator tanks, installed close to the point of extraction, either onshore or on offshore platforms, and designed to receive multiphase hydrocarbon streams [10,11]. The optimal selection of pressure and temperature in surface separators maximizes the yield and quality of the liquids obtained, allowing for efficient separation of the gas, oil/condensate, and water phases [12].

The multiphase flow entering a flow station must be subjected to a separation process, which is characterized by a continuous operation based on a set of interrelated equipment that allows receiving, separating, measuring, temporarily storing and pumping the fluids coming from adjacent wells. This operation requires rigorous control, as well as permanent recording and inspection of each of the operating variables at all stages of the process [13]. The quality of the streams produced and, consequently, the operating conditions of the subsequent stages of the processing system depend to a large extent on the efficiency of this initial stage. A poor separation operation can cause multiple problems during transport, compression and subsequent treatment of the streams, such as the presence of water, gas or sediments out of specification [9,12,14].

Therefore, the separation process is one of the key parts in the natural gas and oil industry [15]. The aforementioned ensures that crude oil represents the main product and its destination is the petrochemical industry, while the separated gas has three main destinations, which are channelling it for commercialization or electricity generation, reinjecting it into reservoirs to obtain more crude oil, or burning it [3].

The studies on crude gas separation initially focused on solving operational problems, such as the work developed by Callaghan et al. [16], who studied severe foam formation issues in first and second stage separators that caused a massive carryover of crude into the pipelines. Technological advancements led to the use of simulators, such as the work of Pan-Echeverría et al. [9], which allowed for the determination of the operational limits of the process through sensitivity to changes in pressure, composition, and temperature, thereby enabling process optimization. Use of artificial intelligence for modeling a biphasic separation system that is part of an initial receiving stage at a crude oil collection station [4]. Additionally, computational fluid dynamics has been applied to verify the efficiency of the separation process [17]. Subsequently, various advanced adsorption and filtration materials have been developed to improve the performance of crude oil and water separation [18]. And finally, in accordance with current times, work is being done to analyze the mechanism of crude-gas separation by a polymer permeable membrane and the key factors affecting the opportunity for gas separation through numerical calculations and simulations [19].

In separation systems specifically, artificial intelligence has supported the identification of two-phase systems, the detection of abnormal conditions, and the discovery of regimes directly from plant data, demonstrating tangible benefits for monitoring and control [4,20-22]. Efforts related to reservoir and production analysis further motivate the integration of machine learning (ML) with domain constraints throughout the hydrocarbon value chain [23].

For unsupervised discovery of operating regimes, K-means remains a simple but powerful benchmark, with well-known behavior, widely evaluated in the literature, and effective for clustering large-volume telemetry into coherent states useful for monitoring and control [24-26]. For supervised tasks, Random Forests (RF) provide high accuracy, robustness to noise, and interpretable variable importance through Gini-based measures, properties that are attractive for the classification and diagnosis of industrial conditions in changing operating environments [27-29]. Previous applications to oil and gas separation processes have shown that combining clustering for regime segmentation with tree-based learners for state discrimination improves both interpretability and operational feasibility [20-22].

Based on the description, in Monagas, Venezuela, there is the QE2 compressor station, which is referred to as a compression gathering station because it increases the pressure of the gas coming from 12 active wells, of which 7 are associated fields and 5 are non-associated fields. The associated gas arrives at the collection station and then proceeds to the flow station, where it is separated from the crude oil. The gas goes to compression, and the crude is stored in tanks for subsequent pumping.

According to the description, the purpose of this research was to analyze the gas-crude separation process at the QE2 compressor station using

machine learning algorithms such as K-means and Random Forest, all based on data science methodology.

2. MATERIALS AND METHODS

This study is observational and retrospective in nature, based on secondary data. Historical records of 1,440 days of production corresponding to the gas-crude separation stage at the QE2 compression station (Monagas, Venezuela) were analyzed. The data comes from PDVSA Gas's operational recording system and was extracted in tabular format. The original operational labels and their engineering units, Thousands of Barrels of Crude Oil per Day (MBNPD) and Millions of Cubic Feet per Day of gas (MMSCFD), were retained. The observation window, sampling frequency, and total number of valid records were reported in a consolidated manner in a quality control table.

Analytical processing was carried out using the Anaconda Navigator program, which facilitated the management of Python packages and environments. To this end, a notebook was created in Jupyter Notebook, which streamlined the preparation and dissemination of the document, which included codes, text, equations, mathematical formulas, and statistics. The methodology applied is shown in Fig. 1, which was based on the data science project process according to O'Neil and Schutt [30].

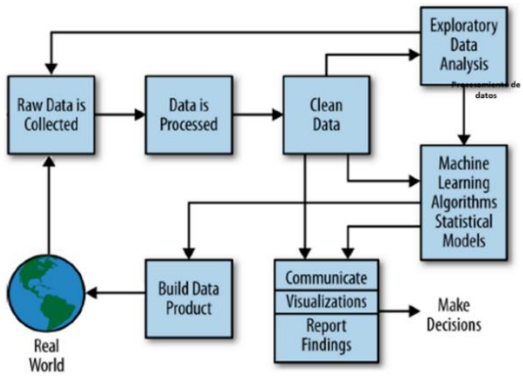


Fig. 1. Data science project process

Based on the phases shown, the activities carried out were:

• Data processing: the necessary libraries, such as numpy and pandas, were initially loaded into the Jupyter Notebook, and a DataFrame (data1) was generated, which contained the information from the "CLUS.csv" file, including information about MBNPD and MMSCFD.

• Data cleaning: the existence of null and missing data was verified, and it was confirmed that the data was numerical. A boxplot and a scatter plot revealed outliers in gas and crude oil.

• Exploratory data analysis: descriptive statistics were applied, and the coefficient of dispersion (R2) of the data was obtained.

• Machine learning, algorithms, static models: K-means algorithms, silhouette index, Kruskal-Wallis method, and Random Forest were applied.

• Communication, visualization, and reporting of findings: this is fulfilled by the writing of this document.

3. RESULTS AND DISCUSSION

In accordance with the established methodology, the necessary libraries were initially imported into the Jupyter Notebook environment, followed by the construction of the corresponding DataFrame. Fig. 2a) presents a visualization of the first and last five records of a set of 1,441 observations, corresponding to the gas (MMSCFD) and crude (MBNPD) production variables. It is important to note that no missing data (Missing values) were detected and that all records were stored as float data, which indicates a numerical representation with decimals. In addition, there was no evidence of non-null values, as shown in Fig. 2b).

	CRUDO (MBND)	GAS (MMPCND)
0	2.60	260.69
1	2.59	255.90
2	2.04	226.69
3	1.85	184.99
4	1.84	183.79
...	...	...
1436	0.81	132.67
1437	0.80	140.59
1438	0.82	144.67
1439	0.76	123.47
1440	0.83	141.64

1441 rows × 2 columns

a)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1441 entries, 0 to 1440
Data columns (total 2 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   CRUDO (MBND)  1441 non-null   float64
 1   GAS (MMPCND)  1441 non-null   float64
dtypes: float64(2)
memory usage: 22.6 KB
```

b)

Fig. 1. a) DataFrame of gas-crude production from the QE-2 compressor plant; b) Non-null data, data types, and missing data of the DataFrame

Fig. 3 shows the boxplot diagrams of the variables analyzed, which show the presence of outliers. In crude oil production, values outside the upper range (whisker) are identified, while in gas production, values below the lower limit are observed. It is worth noting that outliers can significantly affect cluster analysis results, given their sensitivity to non-representative variables. These may correspond to genuine but not very general observations or to limited samples that distort the true structure of the data, leading to unrepresentative clustering [31]. In this study, outlier data represented 2.22% of the total, which is why they were eliminated, leaving a set of 1409 observations for the analysis.

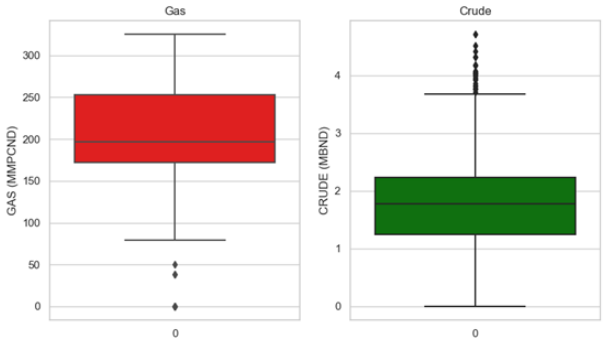
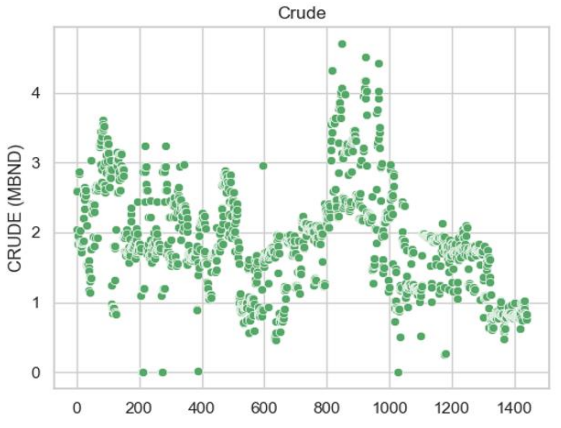


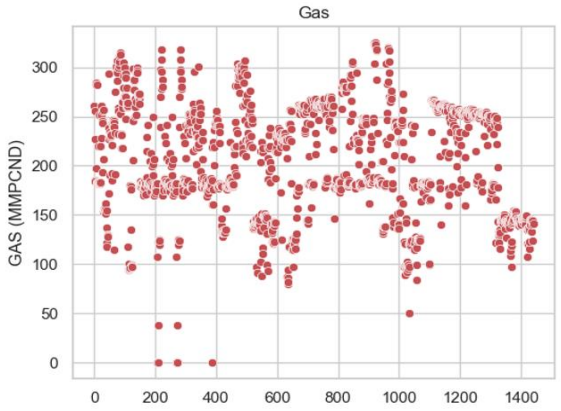
Fig. 3. Boxplot diagram for outlier data

The analysis of the scatter diagram (Fig. 4) for the crude oil and gas production variables shows a high degree of dispersion without a clear trend, attributable to the heterogeneity of the values recorded during the operation. Regarding the gas variable, most of the data are concentrated between 150 and 275 MMSCFD, while for crude oil, the values predominate between 1 and 2.5 MBNPD, although with greater relative dispersion.

This behavior can negatively affect the performance of the production system, since the alternating arrival of large volumes of liquid and gas hinders the efficiency of the separation process and can damage the internal components of the separators. Likewise, flow intermittency compromises the operational efficiency of pumping and compression equipment, increasing the frequency of failures [32]. Fig. 4 below shows the dispersion diagram for crude oil Fig. 4a) and gas Fig. 4b).



a)



b)

Fig. 4. Scatter diagrams: a) crude oil; b) gas

The observed behavior could be attributed to the complex nature of crude-gas mixtures, whose composition varies continuously due to multiple factors. Physical and chemical properties such as density, viscosity and gas-liquid ratio can fluctuate due to the effect of changes in the reservoir, production rates or environmental conditions [2,33,34]. Additionally, the separation process is inherently dynamic and highly sensitive to operating variables such as pressure, temperature, and volumetric flow rate. Variations in these conditions can induce significant instabilities and oscillations in the recorded data [33,35].

In the descriptive statistical analysis of the cleaned data, it was observed that crude oil production ranged from 0.5 to 3.99 MBNPD, with an average of 1.83 MBNPD. Gas production ranged from 83.94 to 321.54 MMSCFD, with an average of 204.96 MMSCFD.

Likewise, Table 1 shows coefficients of variation (CV) of 27.77% for gas and 38.05% for crude oil, which confirms a high variability in both variables. According to Rivas et al. [36], coefficients of variation higher than 20% reflect a high degree of heterogeneity, which justifies the need to analyze

which variables have a more significant influence on the behavior of the system.

In terms of asymmetry, the gas variable shows a slight negative skewness (-0.089), while that of crude oil shows a moderate positive skewness (0.4760). These results are consistent with the degree of dispersion observed: a negative skewness implies a concentration of values towards the upper end (leftward dispersion), and a positive skewness indicates values concentrated at the lower end (rightward dispersion), according to Veliz-Capuñay [37]. Regarding kurtosis, both variables exhibit negative values (K < 0), with the effect more pronounced in the case of gas, indicating a platykurtic distribution. This statistical characteristic suggests a lower concentration of values around the mean and greater dispersion, as pointed out by Contreras et al. [38].

Table 1. Descriptive statistics of plant gas and crude oil production

	Gas (MMSCFD)	Crude (MBNPD)
Minimum	83.94	0.5
Median	197.03	1.79
Mean	204.96	18.32
Kurtosis	-0.962	-0.0115
Skewness	-0.089	0.4760
Standard Deviation	52.811	0.6973
Variance	2789.04	0.4862
CV (%)	27.77	38.06
Maximum	321.54	3.99

The analysis of gas (MMSCFD) and crude oil (MBNPD) production distributions, shown in Fig. 5, reveals a multimodal behavior for both fluids. In the case of gas, a clearly trimodal distribution is observed, while in crude oil, a bimodal distribution with positive skew predominates. This type of behavior suggests that the separation system does not operate under uniform conditions, but is influenced by different operating regimes or sources of variability [39]. In this sense, the presence of several modes could be associated to different factors such as changes in fluid composition as a consequence of variations in pressure, temperature or flow rate of the separator that can generate different modes in the distribution, equipment problems, changes in the separator efficiency that can generate abrupt variations in gas production, stratification of the reservoir due to the fact that in some reservoirs the production can have abrupt changes as a result of the different layers that conform it, due to the fact that the composition of the natural gas varies with time as a consequence of changes in the reservoir or as a result of the influence of external variables that affect the process. Fig. 5 below shows the Production histogram: (a) crude oil; (b) gas.

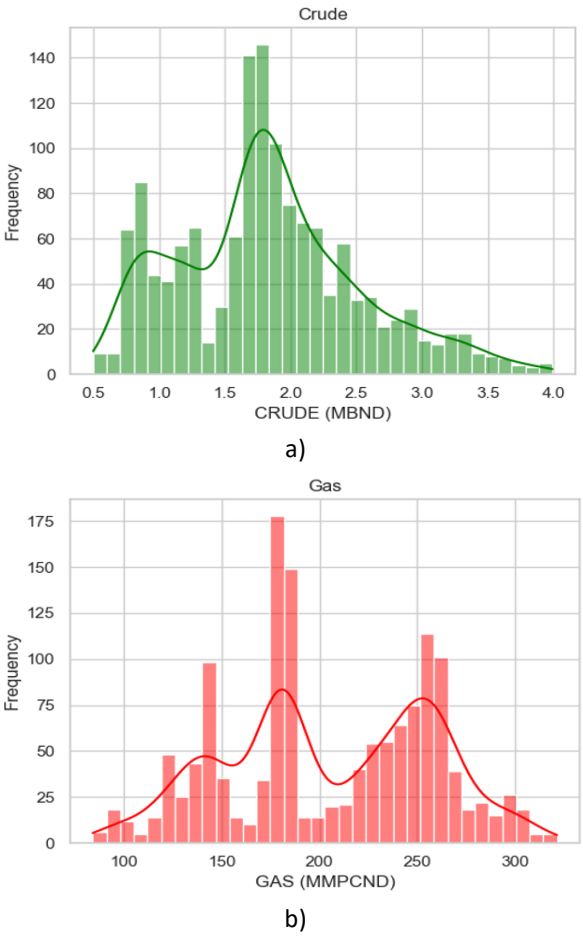


Fig. 5. Production histogram: a) crude oil; b) gas

Also, the bimodal distribution observed in crude oil production suggests the existence of two distinct groups of operating behavior, which could reflect the occurrence of two distinct states in the separation process. This may be due to the presence of free water in the crude, which can generate two different modes in the distribution, changes in crude density due to variations in the crude composition can lead to different densities and, therefore, to different separation efficiencies, problems in the separator that cause failures in the separator internals or sediment accumulations that can generate two different modes in the distribution or the presence of multiphase flow with different phases that lead to generate bimodal readings.

The analysis of the relationship between natural gas and crude oil production, using the scatter plot (Fig. 6), shows a direct correlation between the two

variables. In other words, higher gas volumes are generally associated with higher crude oil volumes, and vice versa. This positive relationship suggests synergistic behavior in the production system, possibly conditioned by common operational characteristics or by the flow regime of the reservoir. The coefficient of determination (R^2 = 0.7645) confirms a strong correlation, implying that approximately 76.45% of the variability of one variable can be explained by the other, while the remaining 23.55% may be due to external factors not considered in the model, such as variations in reservoir pressure, mechanical interference, or differences in separation efficiency [35].

The observed behavior can be attributed to the water, sediments, or salt presence as contaminants in the crude oil. These contaminants can influence the process of gas–liquid separation, which in turn can cause deviations in the behavior of the involved variables. Variations in operating conditions of the separator, such as temperature, flow rate, and pressure, can significantly affect the efficiency of the separation process. Similarly, this process can be influenced by variations in crude oil composition and increases in the gas-oil ratio in some wells.

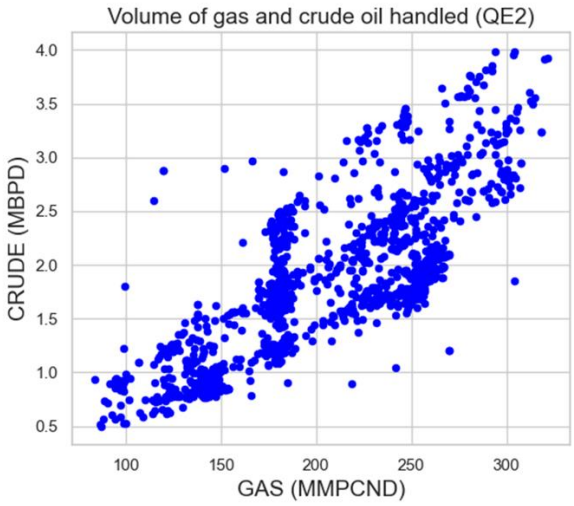


Fig. 6. Dispersion of gas-oil production from the QE-2 plant

To begin the clustering process using the K-Means algorithm, it was necessary to first determine the optimal number of clusters (k). As pointed out by Bishop [40], the choice of the number of clusters can be subjective, since many estimation techniques make implicit assumptions about the structure of the data, which must be met for the model to be valid.

In this study, the elbow method was applied [34], who point out that to calculate the ideal number of clusters k based on the inflection point in the graphical representation, the objective of the algorithm is to reduce the sum of the squared distances in the clusters k. Fig. 7 shows that the elbow is formed from the fourth cluster onwards, indicating that adding more clusters would not significantly improve the quality of the segmentation, so this value was selected as the input value of "k" for the K-Means algorithm.

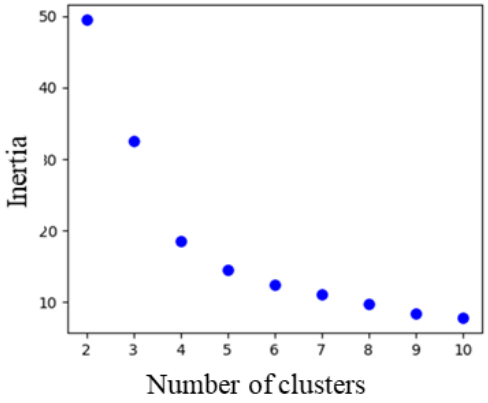
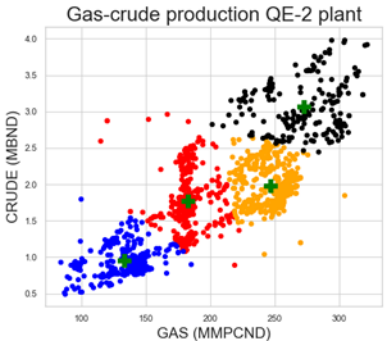


Fig. 7. Elbow method for determining the number of clusters

Fig. 8 shows the four clusters obtained using the K-means algorithm. The red cluster stands out, with points close to each other and around the centroid, indicating high cohesion and a highly stable, predictable separation process from a theoretical point of view. The results obtained are consistent with the findings of Rodriguez et al. [41], who state that the main purpose of cluster analysis is to group objects based on their properties. The resulting clusters of objects exhibit high levels of internal uniformity and external diversity. By forming uniform groups, taxonomies can be described, information simplified, and relationships identified.



Centroids

	Crude	Gas
0	1.977505	247.155098
1	0.955348	133.375633
2	1.771261	183.171599
3	3.066406	272.713333

Fig. 8. Clustered gas-oil relationship and cluster centroids

The dispersion of points across the other clusters could indicate greater variability in the separation process under certain conditions and may provide information about its robustness across conditions. The behavior shown by the clusters leads us to point out that clusters of uniform sizes are observed,

which may mean that there are conditions that are repeated frequently and must be taken into account.

One way to check whether the red cluster is the most cohesive is by measuring the quality of the cluster, for which the Silhouette algorithm was applied. The algorithm evaluates a measure of the similarity of an object (centroid) to its own cluster (cohesion) in comparison with other clusters (separation) [23]. The Silhouette index ranges from -1 to 1, where a high value indicates that the object is well adapted to its own cluster and distant from neighboring clusters [42]. Cluster zero obtained the highest value (Table 2), confirming that it is of the highest quality and therefore exhibits greater data cohesion; however, the other clusters are well formed, as their Silhouette index values tend towards 1.

Table 1. Silhouette coefficient for each cluster

Clúster	Color	Siluetta coefficient
Zero	Red	0.7461
One	Blue	0.6618
Two	Yellow	0.5467
Three	Black	0.5804

To determine whether there were statistically significant differences between the medians of the clusters formed, the nonparametric Kruskal-Wallis method was applied. Fig. 9 shows that there are no statistically significant differences between the data that make up the clusters. Therefore, the centroids are not very far apart, and the behavior shown is a consequence of the presence of different operating regimes, types of crude oil, or even process failures.

```
Estadístico H: 3.0000
Valor-p: 0.3916
No hay evidencia suficiente para rechazar la hipótesis nula.
Los grupos no tienen medias significativamente diferentes.
```

Fig. 9. Kruskal-Wallis Method

Based on the results obtained, it is possible to point out that the formation of clusters in gas-crude production may be related to factors that affect the efficient operating conditions of the wells that converge at the plant. In mature fields, production conditions are not favorable for wells, and in natural gas wells, liquid accumulates at the bottom of the wells, which results in a decrease in gas production due to the generation of a hydrostatic fluid column caused by the accumulation of liquids that can become large enough to stop the well from flowing [43,44].

Likewise, some wells may be experiencing high water cuts and high RGP, mainly due to the release of dissolved gas or the expansion of the gas layer, which is natural behavior in mature reservoirs. This may be causing intermittent flow and, in turn, abrupt variations in production, leading to multiple consequences for fluid separation efficiency.

Finally, in accordance with the proposed methodology, a Randow Forest was applied as an additional approach to provide certainty in decision-making. One hundred trees were obtained, which, according to Prabhu et al. [45] represent the n-estimators or number of decision trees, where a greater number of trees improves performance but can increase the algorithmic complexity of the Randow Forest. Fig. 10 shows the best tree with an accuracy of 0.9929 out of the 100 trees generated by the algorithm. This means that this particular tree achieves an almost perfect classification of the data according to the characteristics used (crude oil and gas).

Within the set of trees generated by the Random Forest model, a tree with an accuracy of 0.9929 was identified, considered the most significant due to its high contribution to the overall accuracy of the model. Breiman [27] established that the importance of a tree within a forest can be evaluated by its impact on overall predictive performance. This tree is used to provide a robust representation of the system's classification rules. It demonstrates the model's ability to segment operational data from the gas-crude separation process using critical variables.

The first decision point is represented as the root node of the tree, which uses the “crude” variable. This highlights its greater discriminatory power compared to the “gas” variable. The model's initial decision is based on dataset segmentation using a cutoff value of 1.165 MBNPD. This value is closer to the centroid of cluster 1 (0.955 MBNPD) than to that of cluster 0 (1.9775 MBNPD), as previously determined by the K-means algorithm. The proposed segmentation indicates that the model prioritizes classifying records linked to small volumes of crude oil, representative of cluster 1. This preference aligns with the interpretation that this group represents the conditions for a high fluid purity.

The impurity of the root node, measured by a Gini index of 0.725, indicates substantial heterogeneity in the data before the first partition. Yet, following the left branch (Crude ≤ 1.15 MBNPD) yields a node with a Gini index of 0.01, indicating almost perfect purity and suggesting that most

observations in that branch belong to a single class. This behavior suggests that the low crude oil production region represents a stable operating state of the system and has structurally consistent characteristics.

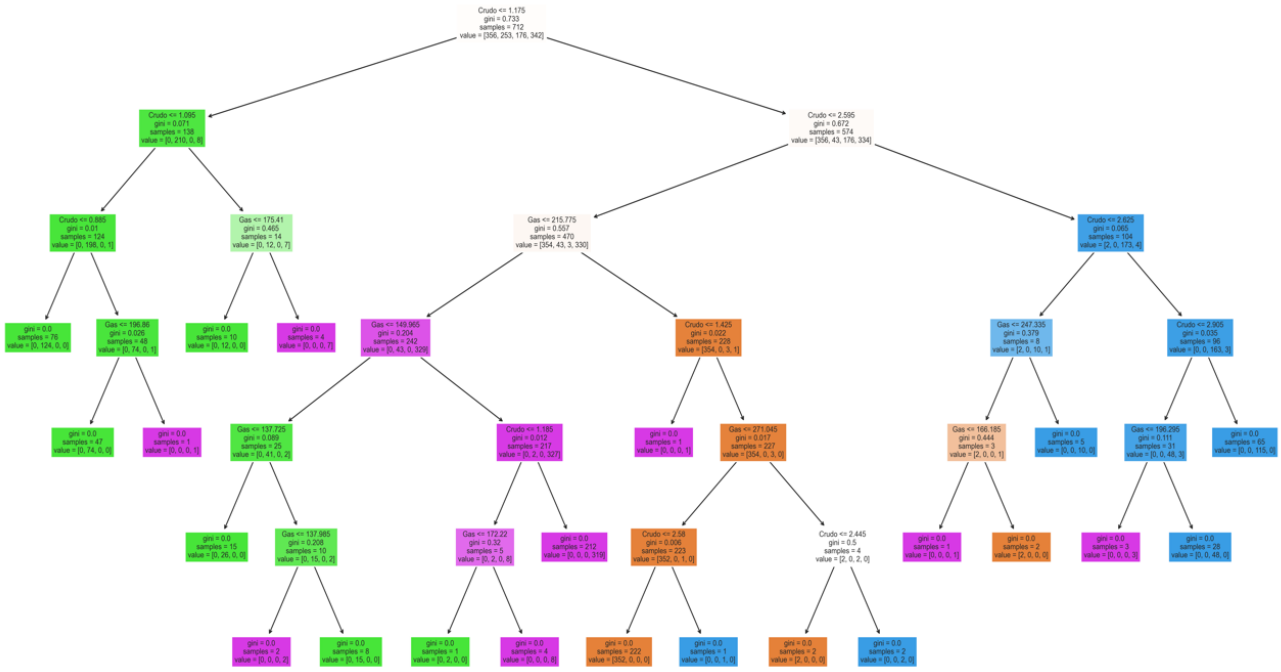


Fig. 10. Tree with the highest accuracy from the Random Forest

On the other hand, when moving towards the right branch (Crude > 1.165 MBNPD), the tree continues with new divisions using the "gas" variable, establishing cuts such as 216.65 MMSCFD, a value that approximates the centroid of cluster 0 (247.15 MMSCFD). However, the Gini value at that node is still 0.664, indicating significant class mixing, which is why the tree branches further to achieve purer divisions.

This segmentation pattern suggests that, although Random Forest is not designed to validate clusters as cluster analysis does, its structure allows for improved separation between classes by progressively reducing impurity in each node. As indicated by Menze et al. [28], the Gini index provides a computationally efficient means to evaluate the effectiveness of successive divisions.

In this context, the variable importance analysis revealed that "crude" contributes 54% of the total relevance of the model, while 'gas' contributes 46%, values extracted from the feature-importance-model attribute. This result supports the selection of "crude" as the initial variable in the most representative tree. The threshold of 1.165 MBNPD appears as an optimal operational separation point, associated with production conditions in which regimes with different stability and efficiency are clearly distinguished.

From a physical-operational point of view, the variable "crude" is interpreted as reflecting characteristics associated with the quality or stability of the fluid, such as water content, condensate concentration, or flow fluctuations, which are more homogeneous in cluster 1 and more variable or disturbed in cluster 0. Thus, the tree architecture reflects not only statistical decisions but also concrete manifestations of the behavior of the separation system, which strengthens the validity of the model as a tool to support operational decision-making.

The present analysis, based on the evaluation of the most representative tree within the Random Forest model (accuracy: 99.29%), is in line with recent studies that highlight the effectiveness of decision tree-based models for diagnosing, classifying, and optimizing complex industrial processes. For example, Li et al. [29] applied Random Forest to detect faults and anomalies in wells and production lines, achieving exceptional accuracies of 96% when they had balanced and well-preprocessed data. Their study coincides with the present analysis in recognizing that the selection of relevant variables (such as "crude oil") has a decisive impact on the architecture of the model and its operational interpretation capacity.

Similarly, Quan et al. [20] employed Random Forest in structural analysis for fault diagnosis in pipelines, enabling the identification of critical variables and improving operational management. They also emphasize that Random Forest, by using the Gini index to assess variable importance, allows the identification of the most influential factors in the occurrence of failures. This facilitates the prioritization of actions and informed decision-making in pipeline system management and maintenance. Similar to the present study, they agree that the value of the decision tree lies not only in its predictive capability but also in its usefulness as a tool for structural system analysis.

Regarding the clustering process, Regarding the clustering process, Yu et al. [21] applied K-means clustering to evaluate the gas content in methane reservoirs in coal beds in order to classify reservoir types. This analysis is directly comparable to the present work, in which the segmentation into four clusters reflected different operating conditions, which were subsequently validated by the Random Forest tree through the proximity of the cut-off points to the centroids.

Finally, Fan et al. [22] emphasize that in complex and highly variable processes such as oil–gas separation, the combination of supervised techniques like Random Forest and unsupervised techniques like K-means enables the construction of hybrid models that leverage the ability of supervised algorithms to identify precise relationships and the capacity of unsupervised methods to uncover hidden patterns. This enhances both interpretability and model performance, a methodological approach that also supports the framework applied in this study.

4. CONCLUSION

During the initial phase of this study's analysis, data quality was verified to ensure there were no null or missing values. A small percentage (2.22%) of outliers were removed to improve the representativeness of the sample. This contributed to the creation of a solid, reliable data set. Preliminary analysis of the variables revealed that gas and crude oil production exhibited noteworthy variability. This was evidenced by coefficients of variation greater than 20%, suggesting significant heterogeneity in operating conditions. The distributions of both variables were multimodal and platykurtic, with negative skew for gas and positive skew for crude oil. These observations suggest the existence of extreme values in the analyzed data. Despite these variations, a strong and direct correlation was found between gas and crude oil (R^2 = 0.7645).

The K-Means algorithm identified four main clusters, with one particularly cohesive cluster representing stable operating conditions. Although the Silhouette index confirmed good cluster segmentation, the Kruskal-Wallis test found no statistically significant differences in their medians, suggesting that the observed variability is due to multiple factors rather than marked structural differences between the groups.

The Random Forest model demonstrated high predictive power, with an accuracy of 99.29%. The “crude” variable was identified as the most important for data classification. This high accuracy validates the robustness of the approach and provides valuable information for decision-making, such as liquid carryover management and separation optimization. The high operational variability found underscores the need to implement continuous monitoring and early warning systems. The development of predictive maintenance mechanisms based on the patterns identified by machine learning models is recommended.

The application of data science and artificial intelligence allowed us to understand the complexity of the gas and crude oil separation process. This methodology not only facilitated the identification of hidden patterns and the segmentation of operating regimes but also laid the foundation for more efficient and proactive management of the resources and equipment involved.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] A.C.C. Rodrigues, Decreasing natural gas flaring in Brazilian oil and gas industry. *Resources Policy*, 77, 2022: 102776. https://doi.org/10.1016/j.resourpol.2022.102776

[2] J.G. Speight, *The Chemistry and Technology of Petroleum*, 5th ed. *CRC Press*, Boca Raton, 2014. https://doi.org/10.1201/b16559

[3] K.K. Orisaremi, F.T.S. Chan, N.S.K. Chung, Potential reductions in global gas flaring for determining the optimal sizing of gas-to-wire (GTW) process: An inverse DEA approach. *Journal of Natural Gas Science and Engineering*,

vol. 93, 2021, 103995. https://doi.org/10.1016/j.jngse.2021.103995

[4] O.E. Gualdrón, L.D. García Mateus, K.D.J. Beleño Sáenz, Identificación de un sistema de separación bifásica en una estación de recolección de crudo a través de técnicas de inteligencia artificial. *Prospectiva*, 2(12), 2014: 18–28. https://doi.org/10.15665/rp.v12i2.285

[5] G. Kooti, B. Dabir, R. Taherdangkoo, C. Butscher, Modelling droplet size distribution in inline electrostatic coalescers for improved crude oil processing. *Scientific Reports*, 13(1), 2023: 20209. https://doi.org/10.1038/s41598-023-46251-4

[6] M. Shahab-Deljoo, B. Medi, M.-K. Kazi, M. Jafari, A techno-economic review of gas flaring in Iran and its human and environmental impacts. *Process Safety and Environmental Protection*, 173, 2023: 642–665. https://doi.org/10.1016/j.psep.2023.03.051

[7] A.H. Al-Rubaye, D.J. Jasim, S.A. Jassam, H.M. Jasim, M. Ameen, F.A. Khoshnaw, Associated Petroleum Gas: Environmental, Utilization, and Economic Rationale. *IOP Conference Series: Earth and Environmental Science*, 1262, 2023: 022026. https://doi.org/10.1088/1755-1315/1262/2/022026

[8] Q. Davarikhah, D. Jafari, M. Esfandyari, Prediction of a wellhead separator efficiency and risk assessment in a gas condensate reservoir. *Chemometrics and Intelligent Laboratory Systems*, 204, 2020: 104084. https://doi.org/10.1016/j.chemolab.2020.104084

[9] G. Pan-Echeverría, T. Gaumer-Araujo, D. Pacho-Carrillo, Simulación y optimización de una planta de separación y estabilización de gas y condensados. *Tecnología, Ciencia, Educación*, 24(1), 2009: 66-75. (In Spanish)

[10] X. Chen, J. Zheng, J. Jiang, H. Peng, Y. Luo, L. Zhang, Numerical Simulation and Experimental Study of a Multistage Multiphase Separation System. *Separations*, 9(12), 2022: 405. https://doi.org/10.3390/separations9120405

[11] A.D. Sarvestani, A.M. Goodarzi, A. Hadipour, Integrated asset management: a case study of technical and economic optimization of surface and well facilities. *Petroleum Science*, 16, 2018: 1221-1236. https://doi.org/10.1007/s12182-019-00356-6

[12] J.A. Massinguil, L.H. Lucas, P. Skalle, Effect of extended heavier hydrocarbon fraction (Cn+) composition on optimum surface separation pressure and temperature. *Journal of Petroleum and Gas Engineering*, 9(5), 2018: 41-55. https://doi.org/10.5897/JPGE2018.0291

[13] C. Ronceros, R. Pomblas, Modelo de Confiabilidad, Disponibilidad y Mantenibilidad Operacional para una Planta Compresora de Gas. *Revista Politécnica*, 51(1), 2023:117–129. (In Spanish) https://doi.org/10.33333/rp.vol51n1.10

[14] T. Jonach, B. Haddadi, C. Jordan, M. Harasek, Dynamic Simulation of a Gas and Oil Separation Plant with Focus on the Water Output Qualit. *Energies*, 16(10), 2023: 4111. https://doi.org/10.3390/en16104111

[15] X. Cao, J. Bian, Supersonic separation technology for natural gas processing: A review. *Chemical Engineering and Processing - Process Intensification*, 136, 2019: 138–151. https://doi.org/10.1016/j.cep.2019.01.007

[16] I.C. Callaghan, C.M. Gould, A.J. Reid, D.H. Seaton, Crude oil foaming problems at the Sullom Voe terminal. *Journal of Petroleum Technology*, 37(12), 1985: 2211–2218. https://doi.org/10.2118/12809-PA

[17] N. Prieto-Jiménez, G. González-Silva, A. Chaves-Guerrero, Revisión del proceso de separación de fases del gas natural a alta presión en la industria Oil&Gas. *Entramado*, 15(1), 2019: 312–329. (In Spanish) https://doi.org/10.18041/1900-3803/entramado.1.5433

[18] J. Yu, C. Cao, Y. Pan, Advances of adsorption and filtration techniques in separating highly viscous crude oil/water mixtures. *Advanced Materials Interfaces*, 8(16), 2021: 2100061. https://doi.org/10.1002/admi.202100061

[19] Y. Jia, C. Shen, Z. Jin, J. Jiang, Numerical Study on Osmotic Equilibrium Timeliness of Oil-Gas Separation Membrane in Online Monitoring System of Transformer Oil Chromatogram. *2023 8th Asia Conference on Power and Electrical Engineering (ACPEE)*, 14-16 April, 2023. Tianjin, China, pp.2113-2117. https://doi.org/10.1109/ACPEE56931.2023.10135698

[20] Q. Quan, D. Li, S. Wang, Research on univariate anomaly diagnosis of gas pipeline measurement data based on Random Forest algorithm. *Journal of Physics: Conference Series*, 2294, 2022: 12004. https://doi.org/10.1088/1742-6596/2294/1/012004

[21] J. Yu, L. Zhu, R. Qin, Z. Zhang, L. Li, T. Huang, Combining K-Means Clustering and Random

Forest to Evaluate the Gas Content of Coalbed Bed Methane Reservoirs. *Geofluids*, 2021(1), 2021: 9321565. https://doi.org/10.1155/2021/9321565

[22] D. Fan, S. Lai, H. Sun, Y. Yang, C. Yang, N. Fan, M. Wang, Review of Machine Learning Methods for Steady State Capacity and Transient Production Forecasting in Oil and Gas Reservoir. *Energies*, 18(4), 2025: 842. https://doi.org/10.3390/en18040842

[23] W.J. Al-Mudhafar, Integrating machine learning and data analytics for geostatistical characterization of clastic reservoirs. *Journal of Petroleum Science and Engineering*, 195, 2020: 107837. https://doi.org/10.1016/j.petrol.2020.107837

[24] P. Nerurkar, A. Shirke, M. Chandane, S. Bhirud. Empirical analysis of data clustering algorithms. *Procedia Computer Science*, 125, 2018: 770–779. https://doi.org/10.1016/j.procs.2017.12.099

[25] J.F. Hair, W.C. Black, B.J. Babin, R.E. Anderson, Multivariate Data Analysis, 7th ed., *Pearson Education*, Harlow, 2014.

[26] L. Marrero, D. Carrizo, L. García-Santander, F. Ulloa-Vásquez, Using K-means algorithm to classify customer profiles with data from smart energy consumption meters: A case study. *Chilean Journal of Engineering*, 29(4), 2021: 778–787. (In Spanish) http://dx.doi.org/10.4067/S0718-33052021000400778

[27] L. Breiman, Random forests. *Machine Learning*, 45(1), 2001: 5–32. https://doi.org/10.1023/A:1010933404324

[28] B.H. Menze, B.M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, F.A. Hamprecht. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, 10, 2009: 213. https://doi.org/10.1186/1471-2105-10-213

[29] W. Li, X. Wang, Q. Sheng, S. Liu, G. Wan, Y. Li, X. Dong, Abnormal Energy Consumption Diagnosis Method of Oilfields Based on Multi-Model Ensemble Learning. *Processes*, 13(5), 2025: 1501. https://doi.org/10.3390/pr13051501

[30] C. O'Neil, R. Schutt, Doing Data Science: Straight Talk from the Frontline. 1st ed., *O'Reilly Media*, Sebastopol, 2014.

[31] S. Mokhatab, W A. Poe, J.Y. Mak, Handbook of Natural Gas Transmission and Processing, 5th ed. *Gulf Professional Publishing,* Cambridge, 2021.

[32] J. Larios-González, T.I. Guerrero-Sarabia, Beneficios de la estabilización y optimización de pozos e instalaciones superficiales con alta RGL: experiencias en un campo marino de aceite pesado. *Ingeniería Petrolera*, 59(1), 2019: 22–35. (In Spanish)

[33] L. Hendraningrat, Complex Fluid Mixtures Characterization of Gas Condensate Reservoir with High CO_2: An Improved Gas Flow Assurance Analysis. *International Petroleum Technology Conference*, February 2025, Kuala Lumpur, Malaysia, IPTC-24854-EA. https://doi.org/10.2523/IPTC-24854-EA

[34] V. Alvarado, E. Manrique, Enhanced Oil Recovery: Field Planning and Development Strategies. *Elsevier Inc.* Kidlington, 2010.

[35] A. Rosiles-Villalobos, L.A. Lugo-Ramírez, M.Á. Clara-Zafra, C.A, Ramírez-Dolores. Statistical analysis of the relationship between work climate and job satisfaction: Case of a government agency in Coatzacoalcos, Mexico. Aposta. *Social Sciences Journal*, (86), 2020: 86–102. (In Spanish)

[36] C.F. Rivas, C. De La Cruz, R. De La Cruz, O. De La Cruz, J. Colivet. Análisis correlacional y contenido de metales pesados en sedimentos superficiales de la avenida Argimiro Gabaldón de la ciudad de Barcelona, Estado Anzoátegui, Venezuela. *Avances en Química*, 7(2), 2012: 111–117. (In Spanish)

[37] C. Veliz-Capuñay, Estadística para la administración y los negocios. *Pearson Educación*, México, 2011. (In Spanish)

[38] J.A. Contreras, G.I. Villalba, E.L. González. Estrategia de cobertura con productos derivados para el mercado energético colombiano. *Estudios Gerenciales*, 30 (130), 2014: 55-64. (In Spanish)

[39] E. Saavedra, Acerca de la moda. *Revista de Educación Matemática*, 36(1), 2021: 75–90. (In Spanish) https://doi.org/10.33044/revem.28231

[40] C.M. Bishop, Pattern Recognition and Machine Learning. *Springer Science + Business Media*, New York ,2006.

[41] M.Z. Rodriguez, C.H. Comin, D. Casanova, O.M. Bruno, D.R. Amancio, L.d.F. Costa, F.A. Rodrigues. *Clustering algorithms: A comparative approach*. *PLoS ONE*, 14(1), e0210236. https://doi.org/10.1371/journal.pone.0210236

[42] L. Orellana, K-means clustering analysis for the ZOO database (Master's thesis), University of Santiago, Santiago, 2020.

[43] N.A. Khairani, E. Sutoyo. Application of k-means clustering algorithm for determination of fire-prone areas utilizing hotspots in West Kalimantan Province. *International Journal of Advances in Data and Information Systems*, 1(1), 2020: 9–16. https://doi.org/10.25008/ijadis.v1i1.13

[44] J.F. Lea, H.V. Nickens. Solving gas-well liquid-loading problems. *Journal of Petroleum Technology*, 56(4), 2004: 30-36. https://doi.org/10.2118/72092-JPT

[45] H. Prabhu, C.M. Ravishankar, A. Ganesan, M. Pandya, H. Bhosale, R. Dhadwal, N.R. Parlikkad, P. Siarry, J.K. Valadi, Enhancing random forest model prediction of gas holdup in internal draft airlift loop contactors with genetic algorithms tuning and interpretability. *Scientific Reports*, 15, 2025: 9325. https://doi.org/10.1038/s41598-025-92728-9

© 2025 by the authors. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).