Data-Driven Classification of Canadian Crude Oils

Ryan Bulger

University of Colorado Boulder

DTSA 5506

ABSTRACT

This project investigates whether the current classification of crude oil grades accurately reflects their underlying chemical and physical similarities. Using unsupervised machine learning techniques, the analysis explores whether natural groupings emerge within the data that differ from existing defined categories. The objective is to determine if data-driven clustering can uncover more meaningful relationships among crude grades, potentially revealing a simpler and more representative grouping structure that better captures the true variation in crude oil characteristics.

1. INTRODUCTION

This study addresses the problem of whether hidden patterns within Canadian crude oil quality data can reveal more natural groupings of crude oils than those defined by current classification systems. Accurate grouping is important because the existing categories, often based only on density and sulfur content, may not fully capture the chemical and physical complexity of each crude. As a result, misclassification can distort pricing, complicate blending and transport decisions, and reduce overall efficiency in the supply chain.

To address these limitations, this project applies unsupervised machine learning techniques to cluster crude oils based on a comprehensive range of laboratory-measured properties. By uncovering data-driven groupings that reflect intrinsic relationships between crudes, the analysis provides a more objective framework for understanding quality variation. The findings can support more accurate market valuation, optimize pipeline and blending operations, and improve refinery yield predictability within the Canadian energy sector.

The analysis proceeds through several structured stages: data summary to understand the dataset's structure, exploratory data analysis (EDA) to identify patterns and correlations, and data cleaning to handle missing or inconsistent values. Three clustering algorithms, K-Means, Agglomerative Clustering, and Gaussian Mixture Models (GMM), are then applied and evaluated using multiple performance metrics and visualized through PCA and t-SNE. The unsupervised results are later compared with supervised learning models, Multinomial Logistic Regression, Support Vector Classifier, and Gradient Boosting, to assess alignment with true class labels. The project concludes with final discussion and takeaways, highlighting key insights and implications for applying machine learning to crude quality analysis in the energy industry.

2. RELATED WORK

Over the past two decades, several approaches have been developed to classify crude oil grades based on their physical and chemical properties. Traditional methods rely on API gravity and sulfur content to distinguish between light, medium, and heavy crudes, as well as sweet and sour categories [1]. While effective for general use, these classifications overlook finer compositional differences. Chemical engineering studies have expanded on this by incorporating indices such as the Watson characterization factor and correlation index to better capture variations in paraffinic and aromatic content [2].

More recently, researchers have introduced spectroscopic and chemometric techniques to automate and improve accuracy in crude classification. Studies using Fourier Transform Infrared (FTIR) spectroscopy, Near-Infrared (NIR) spectroscopy, and Nuclear Magnetic Resonance (NMR) have demonstrated that supervised machine learning models such as Support Vector Machines and Partial Least Squares Discriminant Analysis (PLS-DA) can predict crude oil grades and key properties with high accuracy [3][4]. These methods provide faster, non-destructive analysis compared to conventional lab assays.

Unsupervised statistical methods have also been used to discover natural groupings within crude oil datasets. Techniques like K-Means, hierarchical clustering, and self-organizing maps have revealed that many crude oils can be clustered into fewer groups than their assigned categories, often aligning with geographic origin or refining characteristics [5]. This supports the idea that data-driven clustering can yield more meaningful categorizations than fixed classification systems.

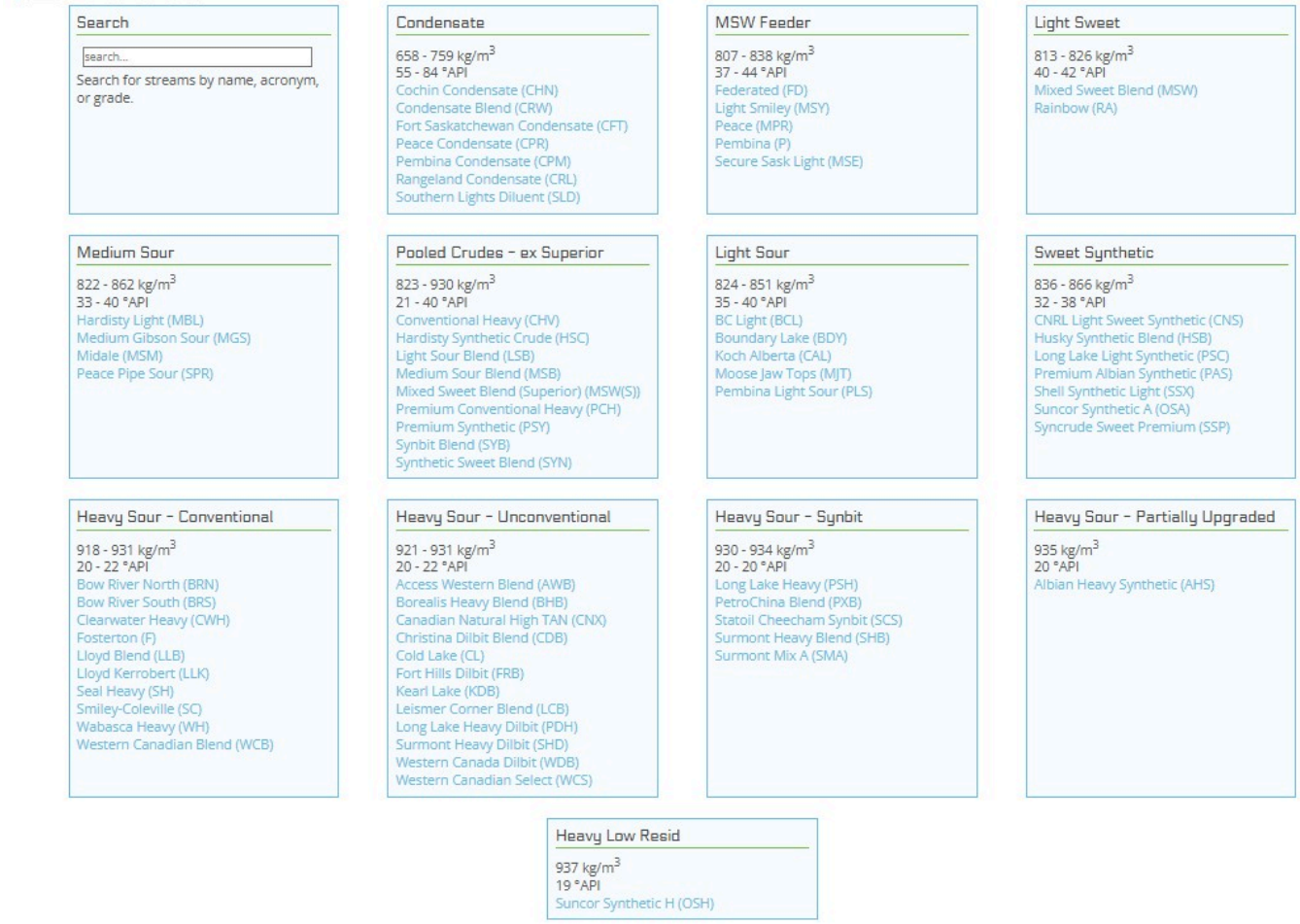
The proposed study builds upon this prior work by combining multiple clustering techniques, K-Means, Agglomerative Clustering, and Gaussian Mixture Models, and comparing them with supervised learning models. Unlike previous studies that focused on narrow datasets or specific analytical tools, this project integrates a broad range of crude quality features to evaluate whether unsupervised methods can uncover more representative groupings, offering both technical and commercial insights into crude oil classification.

3. METHODOLOGY

3.1 Data Source

The dataset for this project will be derived from CrudeMonitor.ca, an established platform managed by Crude Quality Inc., which collects and publishes detailed crude oil assay data from pipelines and refineries across Western Canada [6]. Each assay provides extensive information on the chemical and physical characteristics of crude oils, including density, sulfur content, metals (nickel, vanadium), distillation yields, and hydrocarbon compositions. As shown in Figure 1, CrudeMonitor categorizes these crude oils into 12 defined groups based on quality and refining characteristics. The platform serves as the industry-standard source for Canadian crude oil quality data, used extensively by producers, refiners, marketers, research institutions, and government agencies. Because the data is both publicly accessible and quality-assured, CrudeMonitor is an ideal and trusted source for developing a transparent and reproducible dataset for this project.

Figure 1
CrudeMonitor Crude Oil Grades



To construct the working dataset, I will develop a Python function that interacts directly with the CrudeMonitor API. This function will loop through each of the 12 defined crude groups, extract the associated crude assay data, and convert it into structured dataframes. Once all groups have been collected, the dataframes will be concatenated into a single unified dataset for analysis. This automated process will ensure consistency across all groups and simplify future data updates if new assays or crude types are introduced.

3.2 Data Summary

The first stage of analysis will involve a systematic review of the dataset's structure and composition. I will determine the overall shape of the dataset (rows and columns) and inspect each column's datatype to verify data integrity. A concise summary describing each column, such as physical property, unit of measurement, and expected value range, will be documented. This summary will serve as the foundation for identifying how the dataset should be preprocessed, encoded, and modeled in subsequent stages.

3.3 Exploratory Data Analysis

The exploratory analysis phase will begin with organizing the dataset into logical column groups, numeric, categorical, and datetime, to facilitate consistent handling across the EDA process. A bar chart will be created to visualize the count of missing values per column. Identifying missing data early is essential since proper imputation prevents data loss and ensures that clustering models are not biased by incomplete information.

Next, bar charts will be generated to display the frequency of each category within the categorical columns, including the target crude group. Visualizing these distributions will help identify whether the dataset is balanced across all categories. If any classes are underrepresented, that imbalance will be noted for consideration later when comparing unsupervised and supervised model performance.

To assess how numerical features vary across groups, histograms will be plotted for all numeric columns, color-coded by crude group. These visualizations will help reveal whether certain features (e.g., sulfur or metal content) naturally separate some groups, potentially hinting at clustering tendencies even before formal modeling. A box plot will also be used to identify numeric

columns that may contain extreme outliers. Finally, a correlation heatmap will be constructed to evaluate inter-feature relationships. While correlations are generally less critical for clustering than for supervised learning, understanding them can still provide valuable insight into redundant or highly related variables that might influence distance-based clustering algorithms.

3.4 Pre-Processing

The data cleaning process will prepare the dataset for modeling. The dataset will be split into features (X) and the target (y) variable, where the target represents the 12 crude groups. The target column will be encoded into numeric values, a requirement for external clustering metrics and supervised learning comparisons. Any columns deemed irrelevant, redundant, or containing excessive missing data will be dropped to maintain dataset reliability.

All columns will then be converted to appropriate datatypes, numeric, categorical, or datetime. New columns will be created from datetime fields to extract month and year, which may help identify temporal patterns in assay sampling. Categorical values will be cleaned by standardizing names and correcting any misspellings to ensure consistent encoding. Numeric outliers will either be winsorized or replaced with missing values to be imputed later, reducing the influence of extreme values.

For imputation, numeric columns will be filled using a K-Nearest Neighbors (KNN) imputation method with ten neighbors, which accounts for relationships among features when estimating missing values. Categorical columns will be imputed using their most frequent category. After imputation, all numeric values will be standardized so that each feature contributes equally to clustering distance calculations. This step is important because many clustering algorithms, such as K-Means and GMM, are sensitive to feature scaling. Finally, categorical features will be encoded into dummy variables to transform them into numeric form, ensuring full compatibility with all planned modeling techniques.

3.5 K-means Clustering

The first model I will be using is the K-Means clustering algorithm, one of the most widely used unsupervised learning methods for identifying patterns within multivariate data. K-Means partitions the dataset into a predefined number of clusters by minimizing the sum of squared distances between data points and their respective cluster centroids. The algorithm begins by assigning random centroids, then iteratively reassigns data points and recalculates centroids until convergence is reached and cluster assignments stabilize. The result is a compact grouping of data points where each point is closer to its own cluster centroid than to any other.

K-Means is computationally efficient and performs well on large datasets with continuous numeric features, making it a strong initial choice for clustering high-dimensional crude assay data. However, the algorithm assumes spherical clusters of similar variance and is sensitive to feature scaling and the initial choice of cluster centroids. It may also struggle with overlapping clusters or when the true underlying structure is not well separated. Despite these limitations, its simplicity, speed, and interpretability make it a practical starting point for exploring the underlying structure in crude oil quality data.

I am choosing to include K-Means in this project because it will provide a clear and interpretable baseline for understanding the primary cluster formations within the crude dataset. Since the goal of the analysis is to determine whether natural groupings exist that differ from current classifications, K-Means offers a straightforward way to observe initial cluster boundaries and compare them against known crude oil groups. Its performance will also serve as a benchmark for evaluating more complex models.

3.6 Agglomerative Clustering

The second model used will be Agglomerative Clustering, a type of hierarchical clustering that builds nested groupings by progressively merging the most similar data points or clusters. The process begins with each observation as its own cluster and iteratively combines pairs based on a chosen linkage criterion, such as average, complete, or Ward's linkage, until all data points are merged into a single hierarchy. The resulting structure is often visualized as a dendrogram, which provides insight into how clusters form and merge at different levels of similarity.

Agglomerative Clustering has the advantage of not requiring a predefined number of clusters, allowing flexibility to explore hierarchical relationships within the dataset. It also performs well on data where the natural clusters vary in shape or size. However, it can be computationally expensive for large datasets and may be sensitive to noise and outliers. Additionally, once clusters are merged, the algorithm cannot reverse previous steps, making it somewhat less flexible than iterative partitioning methods like K-Means.

This model is being included in the project because it can reveal hierarchical relationships that K-Means may overlook. In the context of crude oil quality data, hierarchical clustering can help visualize how different grades relate to one another and whether certain groups form subclusters with shared characteristics. This can provide a richer understanding of how crude oils group together beyond simple distance-based partitioning.

3.7 Gaussian Mixture Model Clustering

The third model applied will be the Gaussian Mixture Model (GMM), a probabilistic clustering approach that assumes data points are generated from a mixture of multiple Gaussian distributions, each representing a cluster. Unlike K-Means, which assigns points to a single cluster, GMM uses a soft clustering approach where each data point has a probability of belonging to each cluster. The model uses the Expectation-Maximization (EM) algorithm to iteratively estimate the parameters of these distributions, means, covariances, and mixing weights, until the likelihood of the data under the model is maximized.

GMM is more flexible than K-Means because it can capture clusters of different shapes, sizes, and orientations due to its use of covariance matrices. It also provides probabilistic cluster memberships, which can be valuable for interpreting ambiguous cases where samples exhibit characteristics of multiple groups. However, GMM can be sensitive to initialization, prone to overfitting if too many components are chosen, and may require more computational effort than simpler models.

I am selecting GMM for this project because its probabilistic framework aligns well with the goal of exploring nuanced relationships among crude oil grades. Many crude samples exhibit overlapping physical and chemical properties, and GMM's ability to model such uncertainty provides a more realistic representation of these boundaries. This model will complement the results from K-Means and Agglomerative Clustering by offering a probabilistic perspective on how crude grades may overlap or transition

between one another, thereby helping assess whether the existing classification system adequately reflects the natural structure of the data.

4. MODEL EVALUATION

4.1 Clustering Scoring Metrics

The performance of the clustering models will be evaluated using a combination of internal and external validation metrics. Internal metrics assess clustering quality using only the model's output, without reference to true labels, while external metrics require the known target labels to measure how closely the model's cluster assignments align with the actual classes. By employing both types, the evaluation will capture both the intrinsic structure of the data and the extent to which that structure reflects the existing classification system.

Three internal metrics will be applied to each clustering result. The Silhouette Score measures how well each sample fits within its assigned cluster compared to neighboring clusters, with higher values indicating better-defined separation. The Davies–Bouldin Index (DBI) calculates the average ratio of intra-cluster similarity to inter-cluster separation, where lower scores represent better clustering. The Calinski–Harabasz Index (CHI) assesses the ratio of between-cluster dispersion to within-cluster dispersion, with higher scores indicating more distinct cluster boundaries. Together, these internal metrics provide a robust assessment of cluster compactness and separation.

External validation will use three additional metrics that compare predicted clusters to the true crude group labels. The Adjusted Rand Index (ARI) measures the similarity between two clusterings while correcting for chance alignment, where higher values suggest closer agreement. The Normalized Mutual Information (NMI) quantifies how much information is shared between the predicted and true labels, with values closer to 1 indicating greater mutual dependence. Lastly, the Fowlkes–Mallows Index (FMI) calculates the geometric mean of precision and recall between predicted and actual labels, balancing false positives and false negatives. These external metrics provide insight into how accurately the unsupervised models capture known classifications.

4.2 Clustering Plots

To visually interpret the clustering performance, a set of three plots will be created for every model and cluster size combination. The first is a Silhouette Plot, which displays the silhouette coefficient for each sample within its assigned cluster. The average silhouette score is also plotted as a reference line. When clusters are well-defined, the silhouette coefficients will be high and relatively consistent in width across groups, indicating that points are strongly associated with their clusters and well-separated from others. In contrast, negative or unevenly distributed silhouettes suggest overlapping or poorly defined clusters. The ideal cluster configuration is typically identified where the mean silhouette score is maximized, and cluster widths are approximately uniform.

The second visualization will be a Principal Component Analysis (PCA) scatter plot. PCA reduces the multidimensional dataset into two principal components that capture the highest variance in the data, allowing the clusters to be plotted in two dimensions. While PCA does not always preserve nonlinear relationships, it provides a straightforward and interpretable way to observe how the algorithm has partitioned the data and whether distinct boundaries exist between clusters.

The third visualization will use t-distributed Stochastic Neighbor Embedding (t-SNE), a nonlinear dimensionality reduction technique designed to preserve local relationships between data points. Unlike PCA, t-SNE focuses on maintaining relative distances between nearby points, which makes it particularly effective for visualizing high-dimensional data such as crude assay features. The resulting 2D plots often reveal more nuanced groupings and substructures that may not be visible in PCA space.

Finally, to directly compare model predictions to known labels, three 12×12 heatmaps, one for each clustering model, will be created for a cluster size of 12. Each heatmap will display the relationship between predicted clusters and actual crude groups, illustrating how well each model aligns with existing classifications. This provides an intuitive summary of model performance and misclassification patterns.

4.3 Comparison to Supervised Learning

To further evaluate the clustering models, their performance will be compared to supervised machine learning algorithms trained on the same feature matrix and true target labels. The three supervised models selected are the Multinomial Logistic Classifier, Support Vector Classifier, and Gradient Boosting Classifier. Each of these algorithms represents a different approach to classification, logistic regression for linear separability, SVC for margin-based classification, and gradient boosting for nonlinear decision boundaries.

After training, the supervised models will be used to predict crude oil groups, and the results will be visualized through three additional 12×12 heatmaps, one for each model. These will mirror the layout of the unsupervised heatmaps, allowing for direct comparison between how the supervised and unsupervised approaches classify samples relative to the true labels.

This comparative evaluation will help determine how closely the data-driven, unsupervised clusters reflect the structure captured by supervised learning methods. If the unsupervised models achieve similar grouping performance without label guidance, it would suggest that the existing classification system may not fully represent the natural structure of the data, supporting the hypothesis that fewer or alternative groupings of crude oils may be more appropriate.

5. DISCUSSION

This project will be developed over the course of two weeks, primarily during evenings and weekends. The first week will focus on data extraction, cleaning, and exploratory analysis, while the second week will concentrate on model development, evaluation, and visualization.

Several potential challenges may arise during this process. First, data completeness and consistency could pose issues, as some crude assays may contain missing or irregularly reported values that require careful imputation. Second, feature scaling and high dimensionality may affect clustering performance, making it necessary to fine-tune preprocessing steps and algorithm parameters. Third, determining the most meaningful number of clusters may be difficult since internal and external metrics do not always converge on the same optimal value. Finally, computational efficiency could become a constraint when generating and evaluating multiple model–cluster size combinations, particularly for algorithms like GMM or t-SNE that are resource intensive. Despite these challenges, a structured and iterative approach will

help ensure that the analysis progresses efficiently toward achieving the project objectives.

6. CONCLUSION

This project proposes an unsupervised machine learning approach to explore whether the existing classification of Canadian crude oil grades accurately represents their underlying chemical and physical relationships. By using clustering algorithms such as K-Means, Agglomerative Clustering, and Gaussian Mixture Models, the analysis aims to uncover natural groupings within the data that may suggest a simpler or more representative structure than the current system. The project will also compare these unsupervised results with supervised classification models to evaluate alignment with known labels. Through careful data preparation, visualization, and performance evaluation, this study seeks to contribute a data-driven perspective to crude oil classification, one that could enhance understanding of crude quality relationships and potentially inform future decisions in refining, transportation, and market optimization.

7. REFERENCES

[1] Sancho, C., López, M., & Acha, E. (2021). *Clustering of crude oils using key quality parameters*. Energy & Fuels, 35(7), 5671–5678.

[2] Al-Dahhan, M., & Mahmood, A. (2019). *Characterization of Basra crude oils using Watson and correlation indices*. Journal of Petroleum Research and Studies, 28(1), 1–10.

[3] Mohammadi, A., Kamal, M. R., & Shahraki, F. (2020). *Rapid classification of crude oils using FTIR and chemometric methods*. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 239, 118488.

[4] Moro, M., dos Santos, L. M., & Sousa, J. (2021). *Prediction of crude oil properties using NIR and machine learning models*. Fuel, 289, 119918.

[5] Olea, R. A., Houseknecht, D. W., & Garrity, C. P. (2021). *Cluster analysis of crude oil geochemical data for source characterization*. Marine and Petroleum Geology, 133, 105292.

[6] Crude Quality Inc. (n.d.). CrudeMonitor.ca. Retrieved October 19, 2025, from https://www.crudemonitor.ca/