

Project Proposal

Data-Driven Classification of Canadian Crude Oils

Ryan Bulger
DTSA-5506
University of Colorado Boulder



Problem Definition

Assessing Current Crude Oil Classifications

Grouping Limitations

Existing classifications of crude oils may not accurately reflect the **true similarities** among different types, leading to potential inefficiencies in management and utilization of these resources.

Importance of Reclassifying Crude Oils

Marketing

Improved classifications can enhance **targeted marketing strategies** effectively.

Pricing

Accurate groupings lead to **fair pricing structures** in the market.

Efficiency

Optimized classifications promote **efficient pipeline utilization** and resource management.

Related Work

Overview of traditional and modern methods

Traditional Approaches

Traditional methods like API gravity and sulfur content often fail to capture the complexities of crude oil. These simplistic classifications may overlook significant variations in crude oil characteristics.

Modern ML Studies

Recent machine learning approaches, such as FTIR spectroscopy combined with clustering techniques, have shown promise. However, they often rely on limited datasets, leading to potential inaccuracies in classifications.

Project Steps

Data Collection



Collect data from **CrudeMonitor** for robust analysis.

EDA



Perform exploratory data analysis to understand trends.

Data Cleaning



Prepare the dataset for modeling.

Clustering



Apply clustering algorithms to identify data-driven groups.

Supervised Comparison



Compare clustering results with supervised methods for validation.

Clustering Algorithms

K-Means

K-Means is a **baseline clustering** method that partitions data into K distinct clusters by minimizing intra-cluster variance, allowing for efficient and straightforward classification of crude oil samples.

Agglomerative

Agglomerative clustering is a **hierarchical approach** that builds clusters by merging similar data points iteratively, resulting in a tree-like structure that captures the relationships among crude oils.

GMM

Gaussian Mixture Models (GMM) utilize a **probabilistic framework** to model data as a mixture of multiple Gaussian distributions, providing flexibility and capturing complex data distributions in crude oil classification.

Model Evaluation Metrics

Internal Metrics

Internal metrics will be used to assess clustering quality using statistical measures.

External Metrics

External metrics will be used to compare cluster assignments to known labels.

Visualization Plots

Visualization plots will be used to illustrate clustering results and distribution patterns.

Supervised Benchmarking

Comparing key machine learning methods

Logistic Regression

Logistic Regression serves as a fundamental method, modeling the **probability of class membership** using a linear approach, effective for binary classification problems in crude oil analysis.

Support Vector Classifier

SVC utilizes hyperplane separation to classify data points, **excelling in high-dimensional spaces**, providing robust classification through kernel trick adaptations tailored for complex crude oil datasets.

Gradient Boosting Classifier

Gradient Boosting Classifier builds an ensemble of decision trees where each model corrects the errors of the previous one, producing an **accurate and flexible classifier** that captures nonlinear relationships in complex crude oil data.

Project Timeline

WEEK 1

**Data
Extraction**

**Data
Cleaning**

EDA

WEEK 2

Modeling

Evaluation

**Final
Presentation**

Potential Challenges

Data Quality

Ensuring high data quality is crucial for accurate clustering results.

Cluster Selection

Choosing the right method for cluster selection impacts the project's effectiveness.

Scaling

Efficient scaling is necessary to handle larger datasets effectively and swiftly.

Computation

Computation resources must be allocated properly to support the algorithms used.