Crude oil Assays Classification using Machine Learning

Muhammad Yousuf Jabbar and Luqman Ahmad Shahid

***Abstract*— In this paper, we present classification methods for new and unknown crude oil assays in order to estimate their energy consumption and greenhouse gas emission. Currently University of Calgary has developed a software (PRELIM) to estimate energy consumption and GHG emissions for crude oil assays but its application is limited to only well-known and already available crude oil grades. By using the available data & attributes of existing crude oils in PRELIM software, we have established machine learning algorithms to classify new crudes oils which will further enable using PRELIM software and estimate their GHG emissions**

***Keywords*—Machine learning algorithms; Supervised/ Unsupervised Learning; Crude Oil Assay; Greenhouse Gas emissions; Crude oil Classification**

INTRODUCTION

The petroleum refining industry is the second-largest stationary emitter of greenhouse gases (GHG) in the U.S and third-largest in the world. Annual GHG emissions from a large refinery are comparable to the emissions of a typical (i.e., 500 MW) coal-fired power plant. For U.S. refineries, annual emissions were reported to be close to 180 million tons of CO2eq in 2010, representing nearly 12% of U.S. industrial sector emissions or 3% of the total U.S. GHG emissions.

This industry faces difficult investment decisions due to the shift toward "heavier" crude in the market, both domestic and imported. Each refinery must decide whether and how much they will process heavy crude while considering that processing such crudes requires more energy and results in higher refinery GHG emissions. These major capital investment decisions will impact the carbon footprint of the refining industry for decades to come. Current and future environmental regulations will also affect the decisions faced by this industry.

Hence, it requires more accurate assessments of the emissions intensity upstream of the refinery for each crude. However, the varying quality of these crudes will also have significant implications for refinery GHG emissions. Thus, it require accurate classification of crude oils and assessment of the impact of its quality on refinery emissions to avoid variations in emissions calculation results and avoid potential unintended strict environment regulations.

For this purpose, Petroleum Refinery Life Cycle Inventory Model (PRELIM) was developed which is a mass and energy-based tool for accurate estimation of energy use and GHG emissions associated with processing a variety of crude oils within a range of configurations in a refinery. PRELIM aims to inform decision making and policy analysis by providing a model including data, assumptions and detailed results of energy consumption and GHG emissions. The key input for PRELIM is to select a crude oil assay from the drop-down list of already available 149 assay inventory

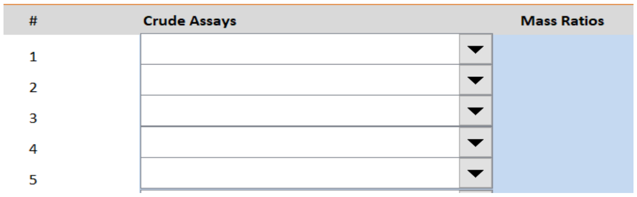
Pick a crude assay

Basrah Medium_COA

Next step is to select options related to the refinery configuration, method of allocation and source for electricity generation. By running the model we get following outputs

- Energy use
- Greenhouse gas emissions
- Mix of crude oil products
- Energy consumption and GHG emissions of each of the final product

The other input option in PRELIM is 'expert inputs' where we need to add manually all the physical characteristics and chemical compositions alongwith crude oil slate/ product details. Also, we can select a crude blend by selecting composition ratio of different assays as shown below.



#	Crude Assays	Mass Ratios
1		
2		
3		
4		
5		

For new discoveries and relatively unknown crudes, its difficult to have all the physical characteristics and chemical compositions needed for expert inputs. Also the crude blending option is not straight forward for these.

For this constraint we have applied machine learning tools for both supervised and unsupervised learning to develop algorithms for classifying new crude assays. Out of the 7 major attributes of crude oil, we have selected the following two main attributes for classification algorithms

- Sulfur content, (wt.,%)
- API gravity

Using the available data of 149 crude assays in PRELIM library, the following machine learning algorithms are applied for training data, clustering and classification

1. K-MEAN
2. KNN (K-Nearest Neighbor)

** Muhammad Yousuf Jabbar is with the Chemical Engineering Department, University of Calgary, Alberta Canada doing his PhD in Energy & Environment (e-mail: muhammadyousuf.jabba@ucalgary.ca).*

Luqman Ahmad Shahid is with the Chemical Engineering Department, University of Calgary, Alberta Canada, doing his Master with specialization in Energy & Environment (e-mail: luqman.shahid@ucalgary.ca).

3. DBSCAN
4. Decision Tree /Classification tree)

Using these algorithms, we will find the closest crude oil assays matching with new unknown grade based on their sulfur contents and API gravity. This will further allow us to select inputs from PRELIM drop down list or creating a blend in expert inputs. The need for having all the physical & chemical characteristics will be eliminated. This leads to calculating energy consumption, environment impacts and product details for new discoveries which are important aspects for economics and environment impact assessment and decision making.

1. Framework and dataset

A. *Crude oil Classification*

The petroleum industry often names crude based on the oil's geographical source, for example, "Cold Lake_Thermal Alberta.ca." However crude quality and refinery configuration are the major factors affecting GHG emissions. Crude quality is defined by physical and chemical properties (e.g., the sulfur content of the crude fractions) that determine the amount and type of processing needed to convert the crude into final products. The technologies used, as well as how they are combined in operation in a refinery, will require different types and amounts of energy inputs and will produce different types and amounts of energy byproducts (e.g., coke) and final products (e.g., gasoline). For example, heavier crudes generally require more energy to process into final products than lighter crudes due to their need for additional conversion processes and their low hydrogen content. Overall Crude oil is classified based on physical characteristics and chemical composition, and these qualities are described with terms such as "sweet," "sour," "light," and "heavy." Based on the classification, it varies in price, usefulness, and environmental impact. Following are the important attributes used to classify crude.

Table 1 Crude oil Attributes for classification

S	Crude oil Attribute	Units
1	Sulphur	wt%
2	Nitrogen	mass ppm
3	API gravity	oAPI
4	Hydrogen	wt%
5	Micro Carbon Residue (MCR)	wt%
6	Characterization Factor	Kw
7	True boiling point (Tb)	[°C]

If we rank above features in terms of economics and environment impacts, following two attributes are considered having the highest impact

1. API gravity
2. Sulfur contents

API Gravity

Crude's classification as either "light" or "heavy" depends on the oil's relative density, based on the American Petroleum Institute (API) Gravity. This reflects how light or heavy a crude oil is compared to water. If an oil's API Gravity is > 10, it is lighter than water and will float on it. If an oil's API Gravity is < 10, it is heavier than water and will sink.

Lighter crude is less expensive to produce accompanied with low GHG emissions. It has a higher percentage of light hydrocarbons that can be recovered with simple refinery process.

Heavy crude can't be produced, transported, and refined by conventional methods because it has high contents of sulfur and several metals, particularly nickel and vanadium. Heavy crude has density approaching, or even exceeding, that of water. Heavy crude oil is also known as "Oilsands" because of its high bitumen content. Overall, it requires complex refining process with high energy consumptions and GHG emissions.

Sulfur Content

Crude oil with low sulfur content is classified as "sweet." and with a higher sulfur content is classified as "sour." Sulfur content is considered an undesirable characteristic for both processing and end-product quality. Therefore, sweet crude is typically more desirable and valuable than sour crude and result in less energy consumption & GHG emissions.

B. *Data set*

The data set was acquired from PRELIM crude oil Assay Library. This comprise of 144 crude assays complete chemical and physical Characteristics alongwith their names and geographical source

C. Features/Attributes selection

For each of the crude assay, following attributes and data is available in the PRELIM library

1 Sulphur (wt.%)
2 Nitrogen mass (ppm)
3 API gravity
4 Hydrogen (wt.,%)
5 Micro Carbon Residue (MCR) (wt.,%)
6 Characterization Factor in (K_w)
7 True boiling point (T_b)

For classification algorithm, one option was to develop model based on all 7 attributes. This strategy makes classification way too complex and there is significant risk of overfitting and low accuracy. Also if we have all the 7 attributes for new crude, we can simulate in PRELIM to calculate GHG emissions and energy usage.

Therefore, to avoid overfitting in our model, we have ranked these attributes based on the relevance and impact on GHG emissions and energy use. As explained above follow two attributes have been selected for our analysis

1. API gravity
2. Sulfur contents

The rest of attributes are not part of our classification algorithms

II Classification algorithms

Following machine learning algorithms developed in Python for classifying crude oil signature.
1. K-MEAN
2. KNN (K-Nearest Neighbor)
3. DBSCAN
4. Linear regression
5. Decision Tree (classification tree)

Training dataset

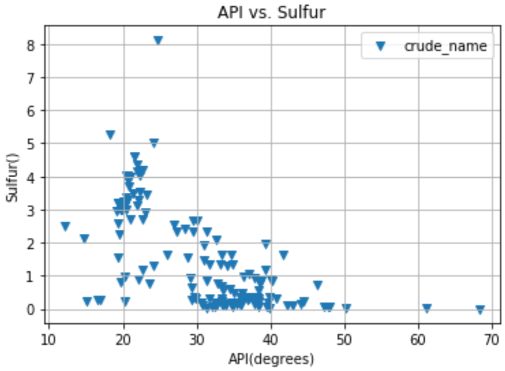


Figure 1 API vs. Sulfur for training data test

There are seven attributes for each crude assay. Based on the trends in 2D plots and most widely used in the industry, crude oil weight as measured by API and crude oil sulfur content are taken as the most important attributes for K-Mean algorithm.

K-Means algorithm

K-means is a numerical, unsupervised learning method. It is simple and very fast, so in many practical applications, the method is proved to be a very effective way that can produce good clustering results. It identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The centroids location is shifted, and it stop creating and optimizing clusters when the centroids have stabilized — there is no change in their values because the clustering has been successful or maximum number of iteration limit has been reached

Based on the K-Means algorithm, there are 81, 52 and 11 oil samples in cluster 0, 1 and 2.

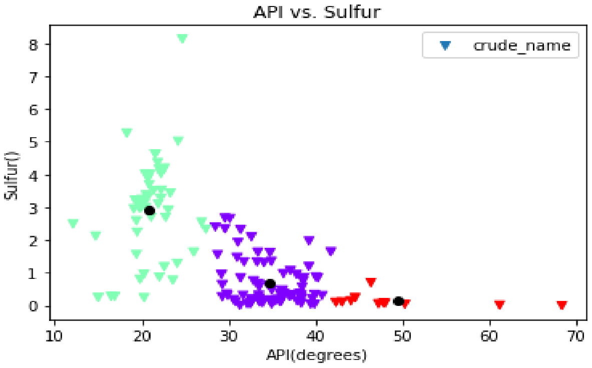


Figure 2 API vs. Sulfur for K-Means algorithm

The results of K-Mean algorithm are shown in Fig. 2 above. Based on the trend three clusters are optimum for 144 data sets.

The centers of the clusters are

Table 1 Cluster centers for K-Mean algorithm

	API, (°)	Sulfur, (wt.%)
Cluster 0	34.75	0.685
Cluster 1	20.88	2.904
Cluster 2	49.44	0.142

The results of K-Mean algorithm are shown in Fig. 2 above. Based on the trend three clusters are optimum for 144 data sets. There are 81 samples in cluster 0, 52 samples in cluster 1 and 11 samples in cluster 2. Cluster 0 with light oil (API range 28-42) have the most oil samples. Cluster 2 is least important with 11 samples and API range > 45.

K-Nearest Neighbor

k-NN performs classification by measuring the similarity of two or more instances. The KNN algorithm assumes that similar data exist in close proximity. In other words, similar things are near to each other. Euclidean distance is the most common metric for measuring the similarity between two instances and is used in this study. The Euclidean distance between points and is defined as follows

$$\text{Euclidean Distance} = \sqrt{\sum (xi - yi)^2}$$

Linear regression

Linear regression is one of the most well-known and well understood algorithms in machine learning. The model assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

Go from $y = \dot{f}(x)$ f: function to approximate
To $\dot{y}$ = Ax
Where A is a matrix : Find the best A

Different techniques can be used to prepare or train the linear regression equation from data, the most common of which is called Least Squares error where we want to find a coefficient of x that minimizes the sum of the squared errors.

$$E(a) = \dot{\sum}_{n=1}^{\infty} (e)^2 = \sum_{n=1}^{\infty} (\acute{y} - yi)^2 f(x) = \sum_{n=1}^{\infty} (a\, xi - yi)^2$$

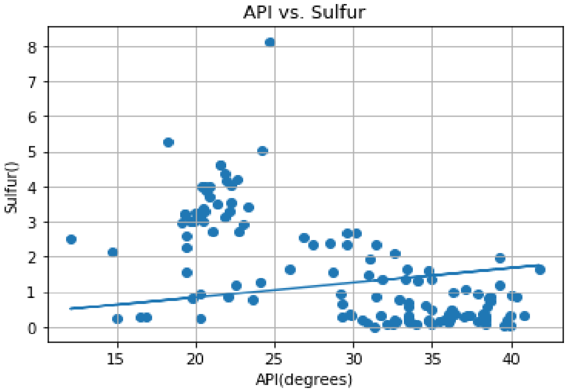


Figure 3 API vs. Sulfur for Linear regression

The linear regression results are shown in Fig. 3 above. The results show a poor fit for estimation of sulfur content given the API content of crude assay. The heavier the crude, the lesser is the gas content and therefore shows poor correlation between the API and sulfur. In the future with more crude oil samples, linear regression could be utilized for each cluster and could be used in complement to KNN method.

DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering method that is used in machine learning to separate clusters of high density from clusters of low density. Given that DBSCAN is a density-based clustering algorithm, it does a great job of seeking areas in the data that have a high density of observations, versus areas of the data that are not very dense with observations. It divide the data into n dimensions and start at a random point and it will count how many other points are nearby. If the point has only a few number of neighbors, we consider it as noise. If it has enough neighbors. We create a cluster containing all these points and or each new point of the cluster, we check if it has (enough) neighbors to add in the cluster. DBSCAN will continue this process until no other data points are nearby, and then it will look to form a second cluster.

The DBSCAN algorithm outputs the noise crude assays and for this data set there were 4 noise points with 2 clusters. The two noise points are well understood with API of higher than 60 which represent condensates used to dilute heavy crude assay. However, the number of clusters are not optimized. Therefore, this data set, KNN algorithm is used for further analysis. For future development with huge data sets, DBSCAN can be utilized to remove the noise data sets.

Decision Tree (Classification tree)

Decision Trees are one of the predictive modeling approaches used in machine learning for achieving high accuracy in many tasks while being highly interpretable. The model starts from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). The Tree models where the target variable can take a discrete set of values are called classification trees and where the target variable can take continuous values (typically real numbers) are called regression trees. In our study, we have used classification trees to identify crude oil grade.

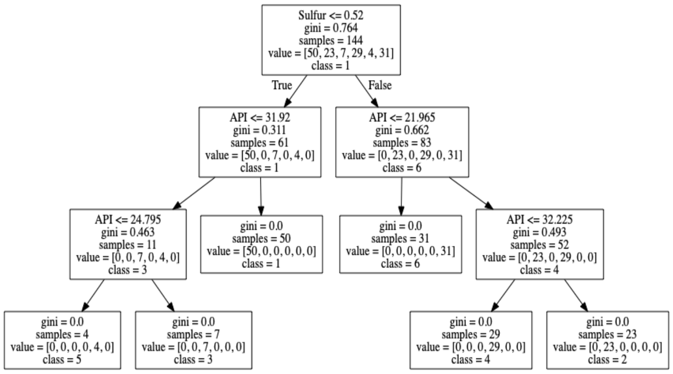


Figure 4 Decision tree for oil assay classification

The decision tree for the crude oil classification is shown in Fig. 4 above. For the decision tree, crude API and sulfur were used to create labels for the whole data set. Based on the criteria established by the Canadian National Energy Board (NEB), each of the available Canadian crudes is designated as conventional (light [>30 API]/ medium [25-30 API]/ heavy [<25 API]), bitumen, or synthetic. For each light, medium and heavy category the crude oil was assign sweet (sulfur weight < 0.5%) and sour (sulfur weight > 0.5%). The six labels created for decision tree output are as follows:

Table 2 Cluster centers for K-Mean algorithm

Crude oil wt.	Sulfur wt., (sweet)	Sulfur wt., (sour)
Light	50 (class 1)	23 (class 2)
Medium	7 (class 3)	29 (class 4)
Heavy	4 (class 5)	31(class 6)

For the training the model all the attributes were used but since the output labels are based on the most important attributes API and sulfur, only these two are shown in the decision tree.

III CRUDE OILS CLASSIFICATION RESULTS

The algorithm can be used to predict crude assay for which crude oil assay is not available in PRELIM model. The two main attributes of input are provided by the operator of the refinery, mainly API and sulfur weight. The algorithm calculates the shortest distance from the center of each cluster and assigns the crude assay to the cluster with the shortest distance. For K-NN algorithm, nearest number of neighbors used are 5 and then based on the nearest distance and similar region (for instance if the test sample is from North America) then more weight is given to the nearest neighbor from the same region. In this way model will output one of the 144 crude assay labels to be used as close proxy for the operator crude oil for modelling purpose in PRELIM tool.

CONCLUSION

In this study, the machine learning algorithms are used to predict crude oil assay signature to be used for GHG estimation in PRELIM model. This approach has considerable advantages such as simplicity, and the presented tool is user friendly. The used data in this model are gathered from the worldwide crude oil assays in PRELIM model. The results of tables and figures that were obtained through modelling K Means. KNN, DBSCAN, decision tree and linear regression. We can conclude that DBSCAN method is the most effective way of crude oil assay characterization. This method can be recommended for characterization of crude oil assays. K Means have also shown good results. Future work is to collect more crude oil assays and therefore have larger training data set. Also, the model will be validated based on the testing data set from clients with crude oil assays not available in PRELIM model. The predicted crude assay by the model will be used by the client to calculate GHG emissions using PRELIM model. The accuracy of the model will be evaluated based on the estimate of GHG from the model to the measured GHG from the client.

APPENDIX

A. Classification Algorithms in python and PDF

B. Crude Oil Assays library from PRELIM

C. Property ranges for each crude oil fraction of the PRELIM's crude assay inventory

D. An example of crude oil assay in PRELIM format

ACKNOWLEDGMENT

This work has been carried out as a part of the final project for Graduate Course ENCH-619 (Data Analytics, Optimization and Control of Energy Systems) at University of Calgary. We would like to thank our instructor Dr. Jean Auriol for his guidance and support.

Abbreviations and Acronyms

GHG	Greenhouse Gas Emissions
PRELIM	Petroleum Refinery Life Cycle Inventory Model
MW	Mega watts
CO2eq	Carbon dioxide equivalents
KNN	K-Nearest Neighbour
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
API	American Petroleum institute
MCR	Micro Carbon Residue
Tb	True boiling point

REFERENCES

- https://www.ucalgary.ca/lcaost/prelim.
- https://inside.mines.edu/~jjechura/Refining/02_Feedstocks_&_Products.pdf.
- EPA. Greenhouse Gas Emissions from Large Facilities; U.S.
- Environmental Protection Agency: Washington, DC, 2012.
- Gale, J., Sources of CO2. In Carbon Dioxide Capture and Storage. Intergovernmental Panel on Climate Change; Metz, B., Davidson, O., de Coninck, H., Loos, M., Meyer, L., Eds.; Cambridge University Press: New York, 2005; pp 75−104.
- J.G. Speight, B. Ozum (Eds.), Petroleum Refining Processes, CRC, 2001.
- R.A. Meyers, Handbook of Petroleum Refining Processes, McGraw-Hill Professional, 2003.
- J. Zieba-Palus, P. Koscielniak, M. Lacki, J. Mol. Struct. 596 (2001) 221–226.
- Bevilacqua, M.; Braglia, M. Environmental efficiency analysis for ENI oil refineries. J. Cleaner Prod. 2002, 10 (1), 85−92.
- Available and Emerging Technologies for Reducing Greenhouse Gas Emissions from the Petroleum Industry; U.S. Environmental Protection Agency: NC, 2010; http://www.epa.gov/nsr/ghgdocs/refineries.pdf. Last accessed September 2012.
- EPA. Technical Support Document for the Petroleum Refining Sector: Proposed Rule for Mandatory Reporting of Greenhouse Gases; 2008.
- EIA. Petroleum and Other Liquids: Imports/Exports & Movements; U.S. Department of Energy, Energy Information Administration,2011.
- Sword, L. Refinery Investments Align with Oil Sands Supplies to 2015. Oil Gas J. 2008, 106 (31), 4.
- Clean Final Regulation Order (Part 1 and 2) to Implement the Low Carbon Fuel Standard; 2009. http://www.arb.ca.gov/regact/2009/ lcfs09/lcfscombofinal.pdf. Last accessed September 2012
- Hu X. DB-HReduction: a data preprocessing algorithm for data mining applications. Applied Mathematics Letters. 2003; 16:889-95.
- Han J, Pei J, Kamber M. Data Mining: Concepts and Techniques. San Francisco: Elsevier; 2011.
- l-Mamory SO, Hasson ST, Hammid MK. Enhancing attribute-oriented induction of data mining. Journal of University of Babylon.
- Han J, Cai Y, Cercone N, Huang Y. Discovery of data evolution regularities in large databases. Journal of Computer and Software Engineering. 1994.
- Zheng A, Casari A. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. Sebastopol: O'Reilly Media;
- Yuan F, Meng Z. H, Zhang H. X and Dong C. R, "A New Algorithm to Get the Initial Centroids," Proc. of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26–29, August 2004.
- Sun Jigui, Liu Jie, Zhao Lianyu, "Clustering algorithms Research",Journal of Software ,Vol 19,No 1, pp.48-61,January 2008.
- Sun Shibao, Qin Keyun," Research on Modified k-means Data Cluster
- Algorithm"I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," Computer Engineering, vol.33, No.13, pp. 200–201, July 2007.
- Merz C and Murphy P, UCI Repository of Machine Learning Databases, Available: ftp://ftp.ics.uci.edu/pub/machine-learning-databases
- Fahim A M,Salem A M,Torkey F A, "An efficient enhanced k-means clustering algorithm" Journal of Zhejiang University Science A, Vol.10, pp:1626-1633,July 2006.
- Machine Learning For Absolute Beginners: A Plain English Introduction (Second Edition)" by Oliver Theobald. ...
- Data Mining: Practical Machine Learning Tools and Techniques" by Ian H. Witten, Eibe Frank, and Mark A. Hall
- Machine Learning in Action" by Peter Harrington
- "Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies" by John D. Kelleher, Brian Mac Namee, and Aoife D'Arcy
- Massaron. Huang Z, "Extensions to the k-means algorithm for clustering large data sets with categorical values," Data Mining and Knowledge Discovery, Vol.2, pp:283–304, 1998.
- K.A.Abdul Nazeer, M.P.Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm", Proceeding of the World Congress on Engineering, vol 1,london, July 2009.