FOSSEE Semester Long Internship (Autumn) Report

On

Crude Oil Characterisation and Products Properties Estimation

Submitted by

Aditi Gupta

3rd Year B.Tech Student,CSE

VIT Bhopal University

Under the Guidance of

Prof. Prabhu Ramachandran

Department of Aerospace Engineering

Indian Institute of Technology Bombay

Mentors:

Priyam Nayak

February 23, 2026

Acknowledgments

Here is the revised acknowledgement based on your requested changes and the details from your project report:

I would like to express my sincere gratitude to the FOSSEE team at IIT Bombay for the opportunity to participate in the Semester Long Internship (Autumn) program. This experience has been intellectually rewarding, allowing me to take full ownership and independently contribute to a significant technical project.

During this internship, I worked on the *"Crude Oil Characterization and Products Properties Estimation"* project. I developed computational models capable of predicting complex physicochemical properties of crude oil without the need for physical assays. My contributions spanned the entire machine learning pipeline: I built a custom Python extraction script to process, filter, and consolidate raw assay data from 114 global crude files into a model-ready dataset. I then conducted rigorous Exploratory Data Analysis (EDA) and domain-specific data quality checks. Finally, I trained Artificial Neural Networks and various Machine Learning models to predict three specific sets of crude oil qualities - Chemical Composition (PNA distribution), Rheological Properties (Kinematic Viscosity), and Empirical Quality Indices (such as Cetane Number and Pour Point) based solely on readily available physical attributes like density, sulphur content, and distillation profiles. These responsibilities provided me with a holistic understanding of data science, predictive modelling, and the end-to-end lifecycle of an engineering project.

I am deeply grateful to *Prof. Prabhu Ramachandran* for his leadership and vision in promoting technical excellence through the FOSSEE initiative. Special thanks to my mentor, *Priyam Nayak*, for his invaluable guidance and technical clarity throughout the course of this work.

This internship has been a defining chapter in my journey, equipping me with the skills and clarity for a career in software engineering and data science.

Contents

CHAPTER 1

INTRODUCTION

1.1 Background: The Complexity of Crude Oil

Crude oil is not a uniform substance, it is a highly complex hydrocarbon mixture containing over 10,000 distinct components. Its physicochemical properties ranging from density and viscosity to detailed chemical composition vary significantly between different oil fields and even between extraction batches from the same field. This variability presents a fundamental challenge for the petroleum industry.

In a modern refinery, the ***Crude Distillation Unit (CDU)*** serves as the primary processing stage, separating raw feedstock into valuable fractions such as naphtha, kerosene, and diesel. To optimize the temperature, pressure, and flow profiles of the CDU, refinery operators require precise, real-time knowledge of the incoming crude oil's properties. A slight deviation in feedstock characterization can lead to off-specification products, reduced yield, and significant energy wastage.

1.2 Traditional Characterization Standards (ASTM Framework)

Crude oil characterization is analogous to decoding a complex chemical fingerprint. Because crude oil is not a homogenous substance but a heterogeneous matrix of hydrocarbons, the industry relies on the ***American Society for Testing and Materials (ASTM)*** to establish the global framework for its analysis. These standardized protocols facilitate the transition from determining bulk physical properties to resolving complex molecular compositions.

Prior to detailed chemical speciation, ASTM standards define the bulk physical nature of the oil. These metrics are critical for initial valuation, custody transfer, and logistical planning.

i. *ASTM D1298 / D5002 (Density & API Gravity):* These standards quantify the specific gravity and API gravity of the crude. As the primary metric for crude oil pricing and classification (e.g., Light vs. Heavy), precise density measurement is fundamental to the economic valuation of the feedstock.

ii. *ASTM D445 (Kinematic Viscosity):* These standard measures the fluid's resistance to flow under gravity. Rheological data derived from ASTM D445 is essential for engineering applications, specifically in calculating the pumping power requirements for pipeline transport and determining heat transfer coefficients within the refinery.

iii. *ASTM D97 (Pour Point):* The Pour Point is defined as the lowest temperature at which the oil retains its ability to flow. This metric is critical for flow assurance, particularly for determining the heating requirements during winter transportation and storage to prevent wax crystallization and blockage.

1.3 Limitations of Traditional Analysis

Currently, the industry relies on two primary methods for characterization, both of which possess inherent limitations:

i. *Laboratory Assays (Time Lag):* Standardized tests (ASTM methods) used to determine critical properties like Kinematic Viscosity or PNA (Paraffin, Naphthene, Aromatic) composition are labour-intensive and time-consuming, often requiring 4 to 8 hours to complete. This delay creates a significant "dead time" between sampling and decision-making. By the time results are available, the crude has often already been processed. Furthermore, these tests require large sample volumes and involve hazardous chemical reagents.

ii. *Mathematical Correlations (Inaccuracy):* To mitigate time delays, refineries often employ empirical mathematical correlations (e.g., Watson K-factor) to estimate properties. However, these linear or simple non-linear correlations often fail to capture the highly complex, multi-dimensional relationships between crude oil components, leading to substantial prediction errors, particularly for heavy or unconventional crudes.

1.4 Problem Statement

The central problem this project addresses is the lack of a real-time, accurate, and non-destructive method for characterizing crude oil. The project investigates how critical crude oil qualities, specifically ***Chemical Composition*** *(PNA),* ***Rheological Properties*** *(Viscosity),* and ***Empirical Quality Indices*** *(such as Cetane Number and Pour Point)* can be determined solely by readily available physical attributes such as Density, Sulphur content, and Distillation Profiles.

1.5 Objectives and Proposed Solution

The primary objective of this work is to develop computational models capable of predicting complex physicochemical properties without physical assays. The project utilizes Machine Learning and Artificial Neural Network to predict three specific sets of crude oil qualities based solely on readily available physical attributes (Density, Sulphur, and Distillation Profiles):

i. *Chemical Composition:* PNA (Paraffin, Naphthene, Aromatic) distribution.
ii. *Rheological Properties:* Kinematic Viscosity.
iii. *Empirical Quality Indices:* Properties such as Cetane Number and Pour Point.

CHAPTER 2

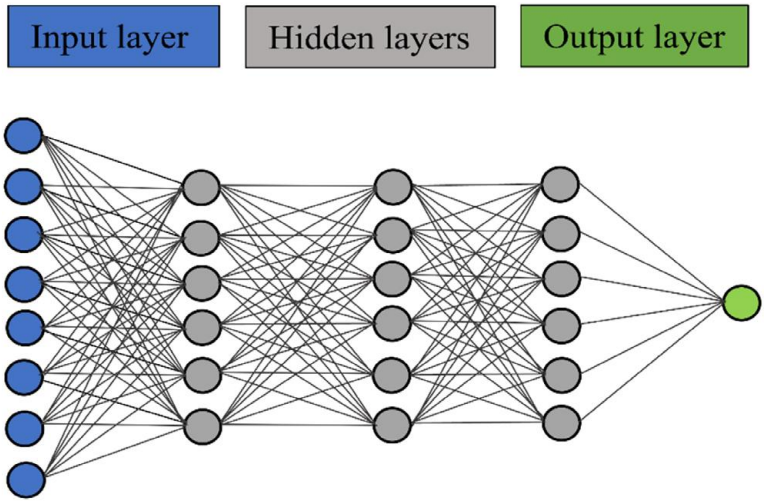
LITERATURE REVIEW

The application of computational intelligence in petroleum engineering has evolved significantly over the last three decades. This chapter reviews pivotal research that establishes the foundation for the current study.

2.1 Early Neural Network Applications

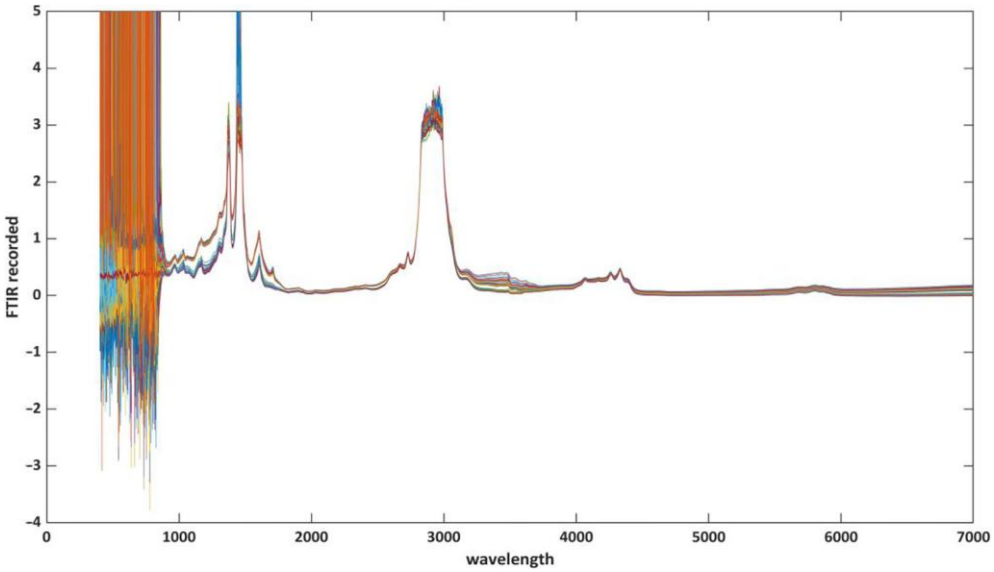
The feasibility of using Artificial Neural Networks (ANN) for crude characterization was rigorously established by ***Sadhukhan (1997)*** in the thesis ***"Crude Characterization and Products Properties Estimation Using Artificial Neural Network."***

Sadhukhan's work addressed the limitations of Equation of State (EOS) methods and thermodynamic correlations, which struggled to predict the properties of undefined heavy fractions. The study demonstrated that ANNs could function as universal approximators, effectively mapping the non-linear relationship between bulk crude properties (inputs) and product yields (outputs). A key finding of this research was that while traditional correlations (like the Watson K-factor) are valid for specific light crudes, they fail to generalize. Sadhukhan proved that a trained ANN could outperform these correlations by learning the underlying patterns in the data without requiring explicit thermodynamic assumptions. This thesis provides the theoretical justification for choosing ANNs as the primary modelling architecture for this project.



2.2 Modern Spectral Approaches

More recent advancements have sought to integrate high-resolution analytical data with Machine Learning. ***Alizadeh et al. (2023)***, in their paper ***"Application of artificial neural network for prediction of 10 crude oil properties,"*** proposed a method using Fourier Transform Infrared Spectroscopy (FTIR).



In this study, the authors analysed 107 crude oil samples from seven Canadian fields. Instead of using bulk physical properties, they used the FTIR spectral response essentially the crude's "molecular fingerprint" as the input for a Feed-Forward Back-Propagation Network (FFBP-ANN). Their model successfully predicted ten distinct properties, including Specific Gravity, Total Acid Number (TAN), and refractive index, achieving coefficients of determination (R^2) greater than 0.90.

2.3 Gap Analysis and Project Positioning

While Alizadeh et al. (2023) demonstrated high accuracy, their method requires an FTIR spectrometer, which is not universally available in all refinery control rooms or historical databases. In contrast, the approach proposed in this report aligns closer to the foundational philosophy of Sadhukhan (1997) but utilizes modern algorithms. By relying on ***Standard Liquid Density*** and ***Distillation Curves*** data that every refinery already possesses in abundance. This project offers a more accessible and cost-effective solution for retrofitting existing facilities with soft sensor capabilities.

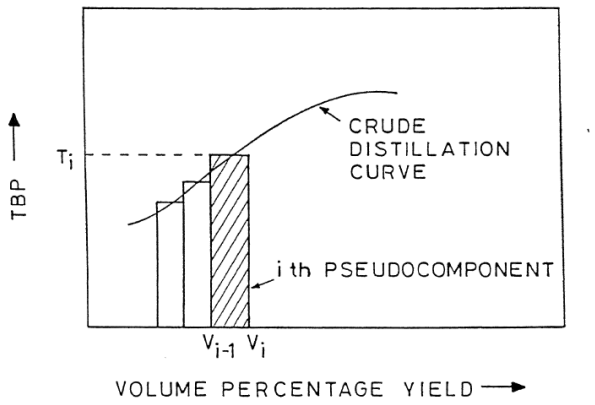


Figure 3.1.1. BREAKDOWN OF CRUDE TBP DISTILLATION CURVE INTO PSEUDOCOMPONENTS

CHAPTER 3: DATA ACQUISITION AND PREPROCESSING

3.1 Data Extraction

The foundational dataset for this research was derived from a comprehensive repository of crude oil assays. The raw data consisted of ***114 individual assay files*** in CSV format, representing a diverse range of crude oil blends from various global fields.

i. The Extraction Challenge:
While the raw files were structured, they contained a massive dimensionality of data, often listing hundreds of properties per crude, ranging from trace metal analysis to detailed light-ends composition. Using the raw files directly for modelling was computationally inefficient due to the high volume of irrelevant features.

ii. The Automated Extraction Pipeline:
To create a clean, model-ready dataset, a custom Python extraction script ***(DataExtraction.ipynb)*** was developed. The pipeline functioned as follows:

i. Iterative Ingestion: The script programmatically iterated through the repository of 114 source CSV files.

ii. Feature Filtering: For each assay, the algorithm isolated only the specific columns required for the "Input Features": ***Bulk Physical and Chemical Properties and Distillation Cuts*** and the "Target Output Variables" ***(Hydrocarbon Composition, Kinematic Viscosity, and Quality Indices)***.

iii. Consolidation: The filtered data points were aggregated into a single master dataset ***(Extracted_CrudeData.csv)***, reducing the high-dimensional raw data into a streamlined matrix of 114 rows (Crudes) × 26 columns (Features & Targets).

```
…  Processing 114 specific files.
   Done. Saved to 'Extracted_CrudeData.csv'
        Crude Name  StdLiquidDensity (kg/m3)  SulfurByWt (%)  ConradsonCarbonByWt (%)  NitrogenByWt (%)  Distillation Mass @ )
   0  Achinsk-2015                926.440713       24.606786                 5.827369          0.209874
   1     Akpo-2014                794.689603        0.070591                 0.732083          0.062825
   2     Alba-1994                937.762899        1.322111                 5.847739          0.196515
   3     Alba-2002                932.299296        1.188621                 4.615622          0.226565
   4     Alba-2012                934.718850        1.254322                 7.296455          0.220453
```

3.2 Dataset Description

i. *INPUT FEATURES (X)*
To ensure model robustness, a standardized feature set was defined. All predictive models in this study utilize the exact same 13-dimensional input vector:

Table 3.1: Independent Variables (Model Inputs)

CATEGORY	INDEPENDENT VARIABLE	UNIT	DESCRIPTION
BULK PROPERTIES	StdLiquidDensity (kg/m3)	kg/m^3	Standard Liquid Density
	SulphurByWt (%)	*wt%*	Total Sulphur content

	ConradsonCarbonByWt (%)	*wt%*	Carbon residue indicator
	NitrogenByWt (%)	*wt%*	Total Nitrogen content
DISTILLATION PROFILE (Boiling Points)	Distillation Mass @ X Pct (C)@ 1 (%) - TBP	*^{o}C*	Temperature at 1% Mass Recovery
	Distillation Mass @ X Pct (C)@ 5 (%) - TBP	*^{o}C*	Temperature at 1% Mass Recovery
	Distillation Mass @ X Pct (C)@ 10 (%) - TBP	*^{o}C*	Temperature at 1% Mass Recovery
	Distillation Mass @ X Pct (C)@ 30 (%) - TBP	*^{o}C*	Temperature at 1% Mass Recovery
	Distillation Mass @ X Pct (C)@ 50 (%) - TBP	*^{o}C*	Temperature at 1% Mass Recovery
	Distillation Mass @ X Pct (C)@ 70 (%) - TBP	*^{o}C*	Temperature at 1% Mass Recovery
	Distillation Mass @ X Pct (C)@ 90 (%) - TBP	*^{o}C*	Temperature at 1% Mass Recovery
	Distillation Mass @ X Pct (C)@ 95 (%) - TBP	*^{o}C*	Temperature at 95% Mass Recovery
	Distillation Mass @ X Pct (C)@ 99 (%) - TBP	*^{o}C*	Temperature at 99% Mass Recovery

ii. *TARGET VARIABLES (OUTPUTS)*

To evaluate the model's capability to predict crude oil behaviour, two distinct sets of target variables were selected. These targets represent the critical physicochemical properties required for refinery optimization.

» **SET 1: Hydrocarbon Composition (PNA Analysis)** - The first set of targets focuses on the chemical composition of the crude, specifically the ***PNA (Paraffin, Naphthene, Aromatic)*** distribution. This information is vital for determining the chemical quality of the feedstock.

Table 3.2: SET 1 Target Variables

VARIABLE NAME	DESCRIPTION	UNIT

AromByWt	Aromatics content by weight	*%*
NaphthenesByWt	Naphthenes content by weight	*%*
ParaffinsByWt	Paraffins content by weight	*%*

» **SET 2: Kinematic Viscosity (Primary Physical Target)** The second set focuses on flow properties. Kinematic Viscosity is the primary target variable due to its importance in sizing pumps and heat exchangers in the pre-heat train.

Table 3.2: SET 2 Target Variables

VARIABLE NAME	**DESCRIPTION**	**UNIT**
KinematicViscosity (cSt)@ 37.78 (C)	Kinematic Viscosity at 37.78°C	*cSt*
KinematicViscosity (cSt)@ 98.89 (C)	Kinematic Viscosity at 98.89C	*cSt*

» **SET 3: Secondary Quality Specifications (Operational Targets)** Beyond composition and viscosity, the dataset includes a third set of properties that define the operational value and processing challenges of the crude. These targets are particularly relevant for diesel production and corrosion control.

Table 3.3: SET 3 Target Variables

VARIABLE NAME	**DESCRIPTION**	**UNIT**
CetaneNumber	Ignition quality indicator for diesel fuel fractions.	*Unitless*
BromineNumber	Measure of aliphatic unsaturation (chemical instability).	*g Br_2/100g*
AnilinePoint	Indicates aromatic content; lower values mean higher aromatics.	*^{o} C*
FreezePoint	Lowest temperature before hydrocarbon crystals form.	*^{o} C*
PourPoint	Lowest temperature at which the oil remains fluid.	*^{o} C*
CloudPoint	Temperature at which wax crystals first appear (haze).	*^{o} C*
TotalAcidNumber	Measure of acidity/corrosivity (Naphthenic Acid content).	*mg KOH/g*
CtoHRatioByWt	Carbon-to-Hydrogen weight ratio (energy density indicator).	*Ratio*

CHAPTER 4

EXPLORATORY DATA ANALYSIS (EDA)

Following the extraction of the 114 crude oil assays, an Exploratory Data Analysis (EDA) was conducted to assess data quality, identify missing values, and understand the underlying statistical relationships between the input bulk properties and the target quality indices.

4.1 Data Cleaning and Preprocessing

The raw extracted dataset consisted of a matrix with dimensions (114, 26).

i. Column Renaming:
To ensure compatibility with Python's Pandas and Scikit-Learn libraries, column names were standardized (e.g., removing units and special characters). For instance, StdLiquidDensity (kg/m3) was renamed to StdLiqDensity_kgm3.

ii. Structural Integrity Checks (Data Types, Missing Values, Duplicates):

I. Data Types: All numerical features and targets were strictly cast to floating-point formats to ensure mathematical operations during model training behave as expected.

II. Missing Values (NaN): A comprehensive null-value check was performed. Missing data in petroleum assays is common (e.g., not all labs test for Bromine Number). Variables with excessive missingness were evaluated for either imputation or exclusion to prevent algorithmic bias.

III. Duplicate Records: Assays were checked for duplicate crude names or identical feature rows to prevent "data leakage," which would artificially inflate model performance during the train/test evaluation phase.

iii. ***Data Statistics:***
Descriptive statistics (mean, median, standard deviation, minimum, maximum, and quartiles) were generated for all variables. This provided an initial numerical baseline to understand the central tendency and dispersion of the dataset.

4.2 Domain-Specific Data Quality Checks

Standard data science checks are insufficient for chemical datasets. To ensure the physical viability of the data, custom engineering quality checks were applied. Any assay violating these thermodynamic and physical rules would be flagged as erroneous.

i. Non-negative Property Values: Physical quantities such as ***Standard liquid density, Sulphur weight (%), Nitrogen weight (%), Conradson carbon weight (%), Hydrocarbon weight (%), Cetane Number, Total Acid Number and C to H Ratio by weight*** cannot theoretically be less than zero. The dataset was scanned to ensure no negative values were recorded due to sensor errors or manual data entry typos.

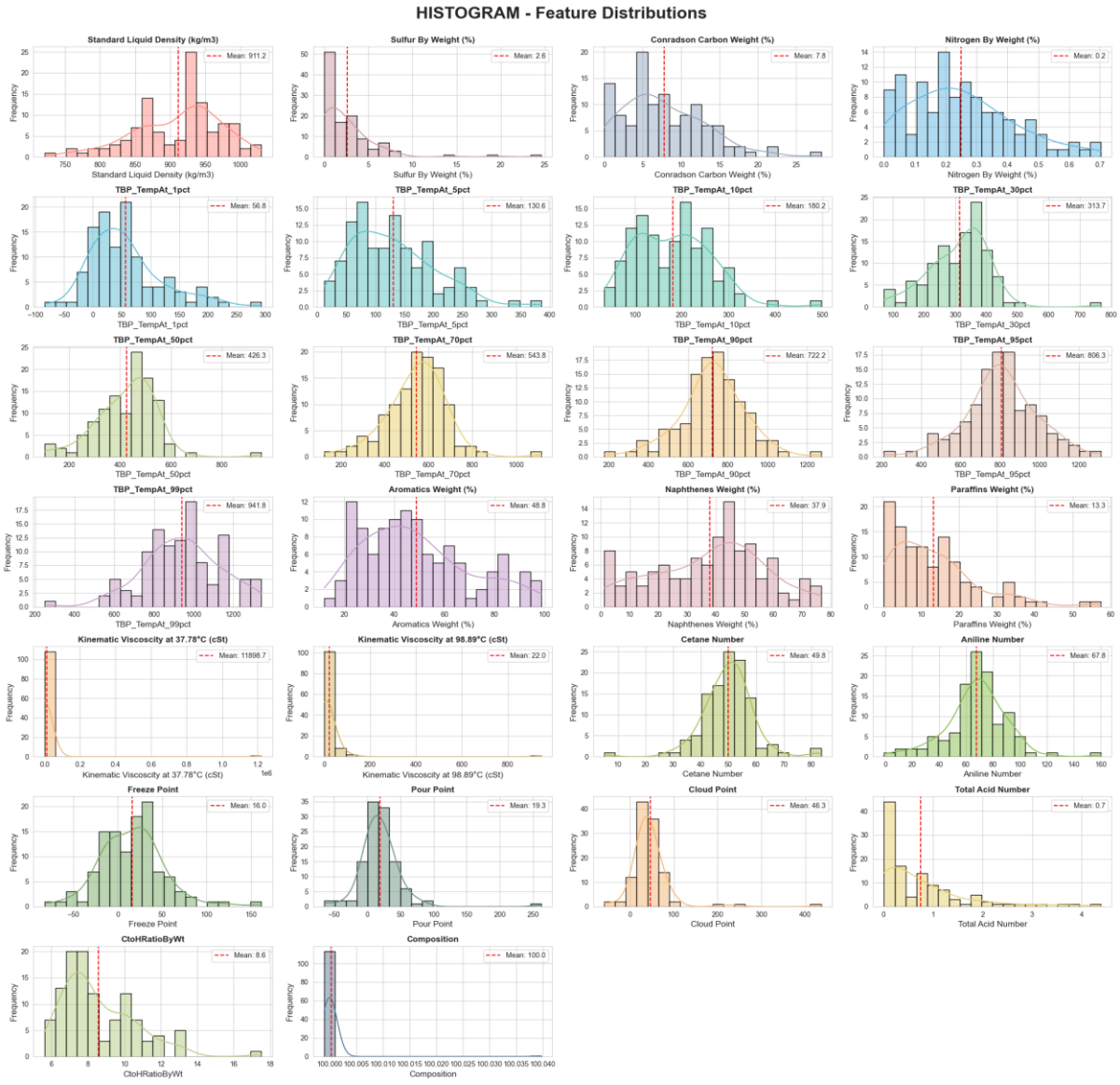
ii. Monotonic Increase of TBP Distillation Temperatures: True Boiling Point (TBP) curves follow the physical laws of fractional distillation. As the cumulative mass percentage of distilled crude increases (e.g., from 5% to 10% to 30%), the boiling temperature *must* either increase or remain constant. Violations of this monotonic constraint indicate a corrupted distillation profile.

iii. Hydrocarbon Composition Mass Balance (~100 wt%): For the Set 1 target variables (Paraffins, Naphthene, and Aromatics - PNA), the sum of their weight percentages must approximate 100%. While slight deviations are acceptable due to measurement tolerances or the presence of trace heteroatoms (Sulphur, Nitrogen), substantial deviations indicate an incomplete or inaccurate compositional assay.

4.3 **Data Visualisation**

i. UNIVARIATE ANALYSIS

I. Histogram with KDE

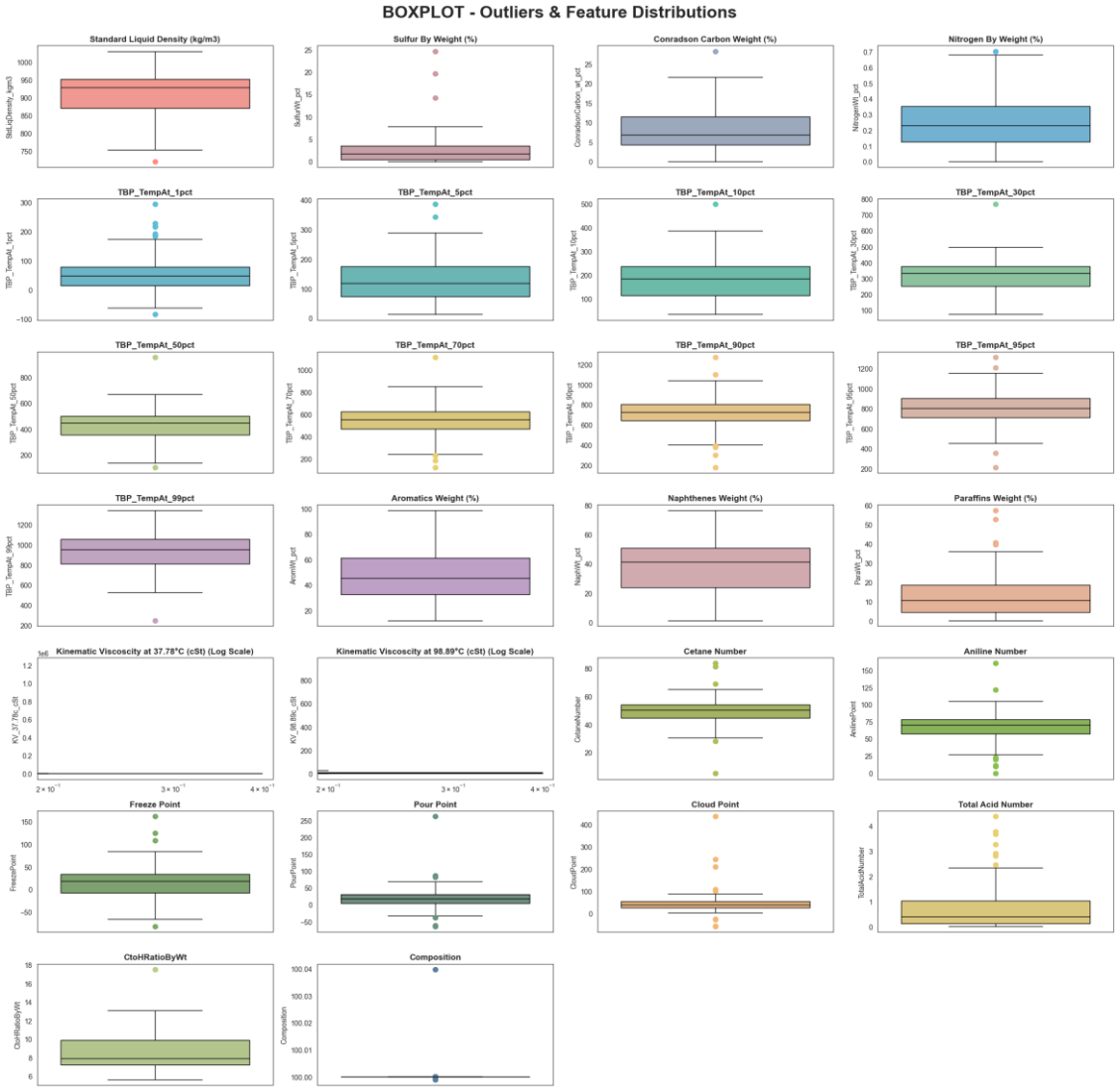


» **Standard Liquid Density:** Multimodal distribution with clusters around 850, 925, and 975 kg/m³, reflecting the presence of light, medium, and heavy crudes. The mean of 911.2 kg/m³ indicates the dataset is weighted toward heavier grades.

» **Sulphur, Conradson Carbon, Nitrogen:** All three are strongly right-skewed with most crudes concentrated at low values and long tails toward the higher end. Log-transformation will likely be needed before model training.

» **TBP Distillation Cuts:** Light-end cuts (1–10%) are irregular and bimodal. Mid-range cuts (30–70%) are the most symmetric and stable. Heavy-end cuts (90–99%) broaden again, reflecting variability in residue composition.
» **Hydrocarbon Composition (PNA):** Aromatics spans the widest range (20–100%), making it a well-distributed prediction target. Naphthene clusters around 30–50%. Paraffins is right-skewed, with most crudes being paraffin-poor.
» **Kinematic Viscosity:** Extremely right-skewed at both temperatures. A small number of ultra-heavy crudes push the mean to ~11,899 cSt at 37.78°C, making log-transformation essential.
» **Quality Indices:** Cetane Number is near-Gaussian and centred around 50 — the cleanest target variable. Freeze and Pour Point show bimodal tendencies. Cloud Point and Total Acid Number have notable right-skewed tails. The Composition variable confirms PNA mass balance at exactly 100%.

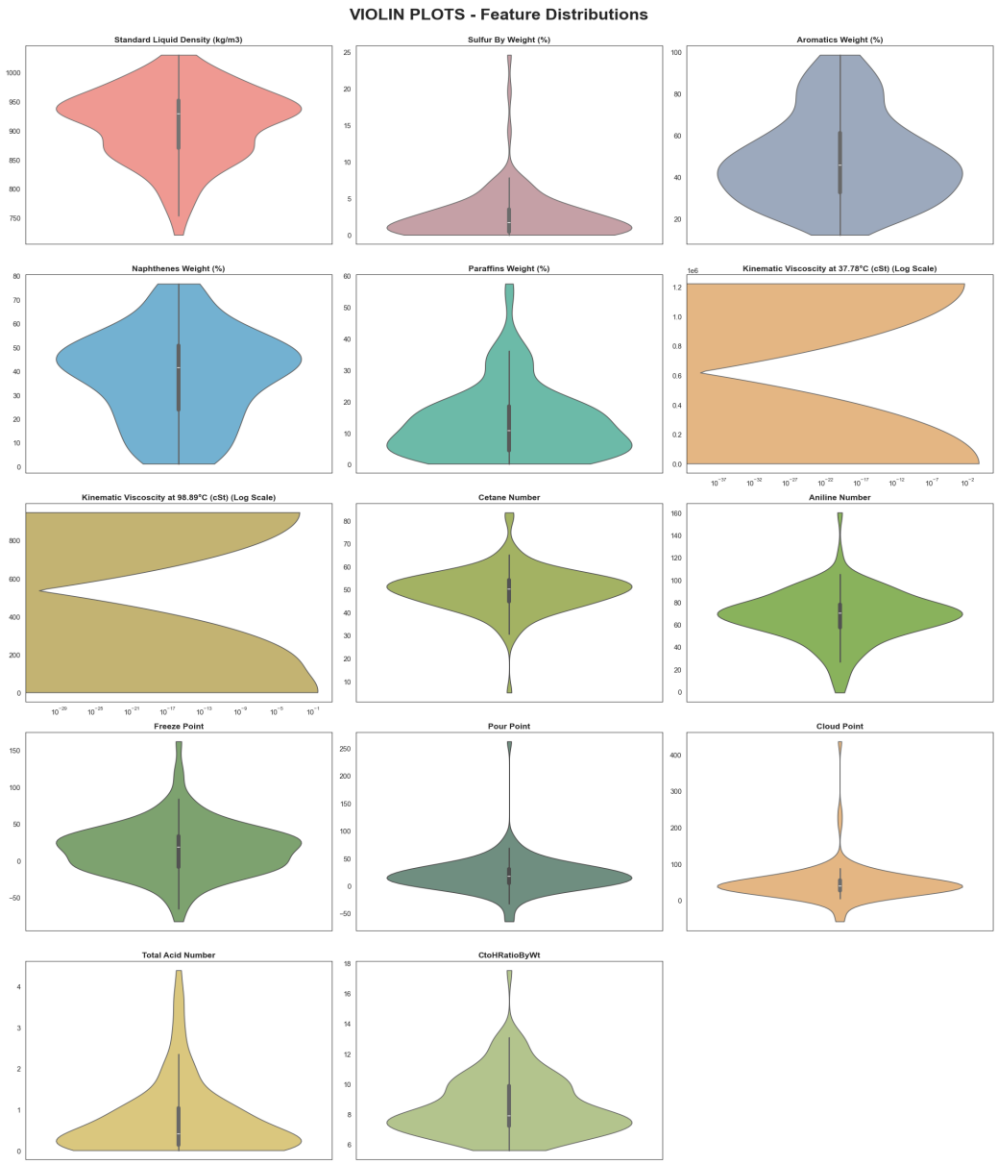
II. Boxplot – Outliers & Feature Distributions



» **Bulk Properties:** Density has one low-side outlier (~730 kg/m³). Sulphur and Conradson Carbon show compact IQRs but several high-side outliers. Nitrogen has an extremely narrow IQR, indicating limited variance.

» **TBP Cuts:** IQR widens progressively from the 1% to 99% cut. The 50% cut is the most stable, while heavy-end cuts show the broadest spread.
» **Kinematic Viscosity:** Even on a log scale, one or two extreme outliers dominate the entire plot, making log-transformation clearly necessary.
» **Quality Indices:** Cetane Number and Aniline Point are the most well-behaved targets. Cloud Point has the most extreme outlier structure, with isolated points exceeding 400°C.

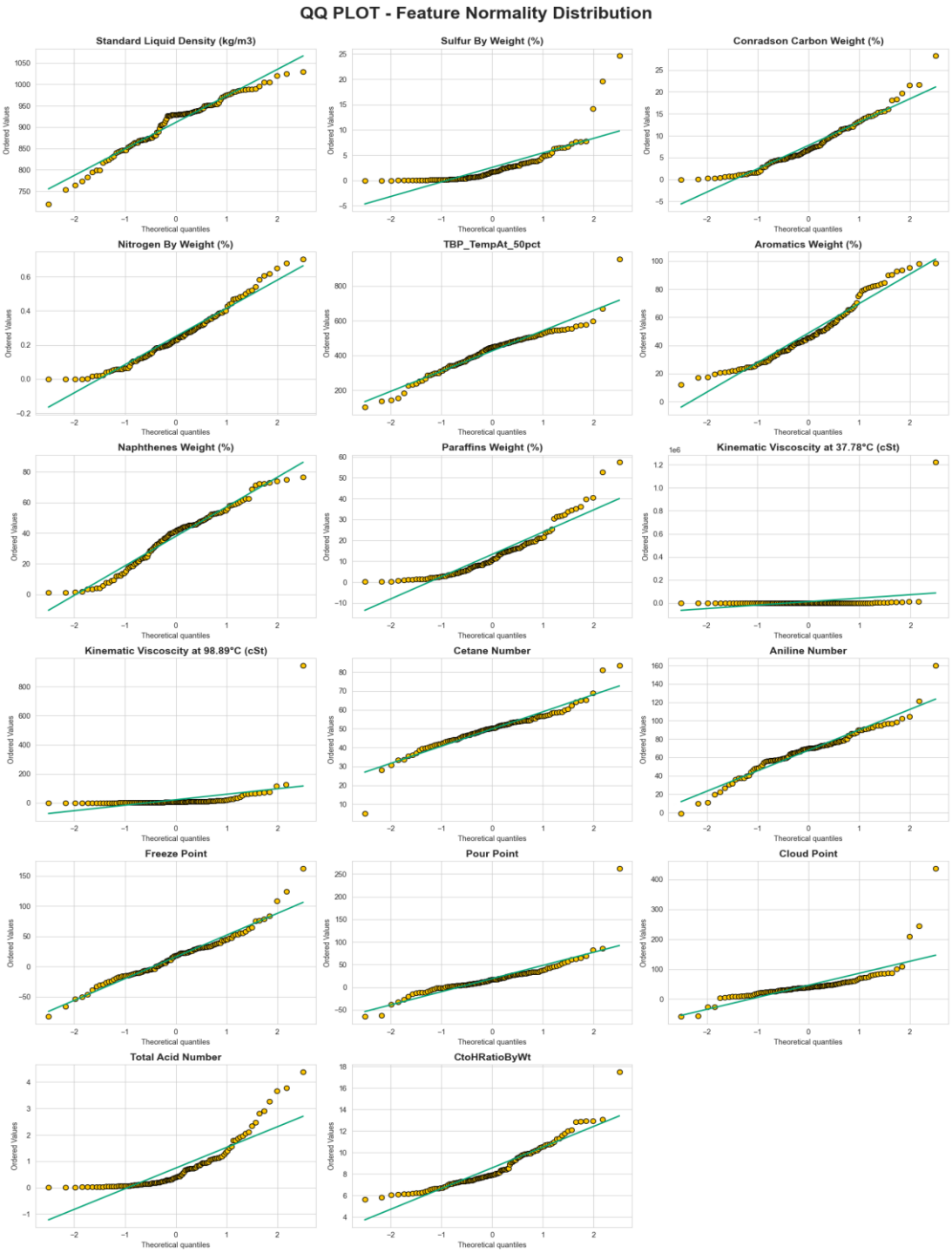
III. Violin Plots – Feature Distributions



» **Standard Liquid Density:** Multi-bulged shape confirms no single crude grade dominates the dataset.
» **Sulphur:** Inverted-teardrop shape indicates most crudes are low-sulphur, with a thin extended upper tail.
» **Aromatics:** Broad, flat-topped violin across the full range means a well-spread and reliable prediction target.
» **Paraffins:** Wide at the base, tapering at higher values shows a paraffin-poor majority with a small outlier tail.

» **Kinematic Viscosity at 37.78°C:** Even on a log scale, nearly all data is compressed at the lower end. The most irregular distribution in the dataset.
» **Cetane Number:** Smooth and symmetric. It is the most normally distributed target variable.
» **Freeze & Pour Point:** Both show bimodal shapes, reflecting two distinct crude sub-populations (paraffinic vs. naphthenic).
» **Total Acid Number:** Tightly concentrated near zero with a very thin tail and limited distributional spread.

IV. QQ Plot – Normality Assessment



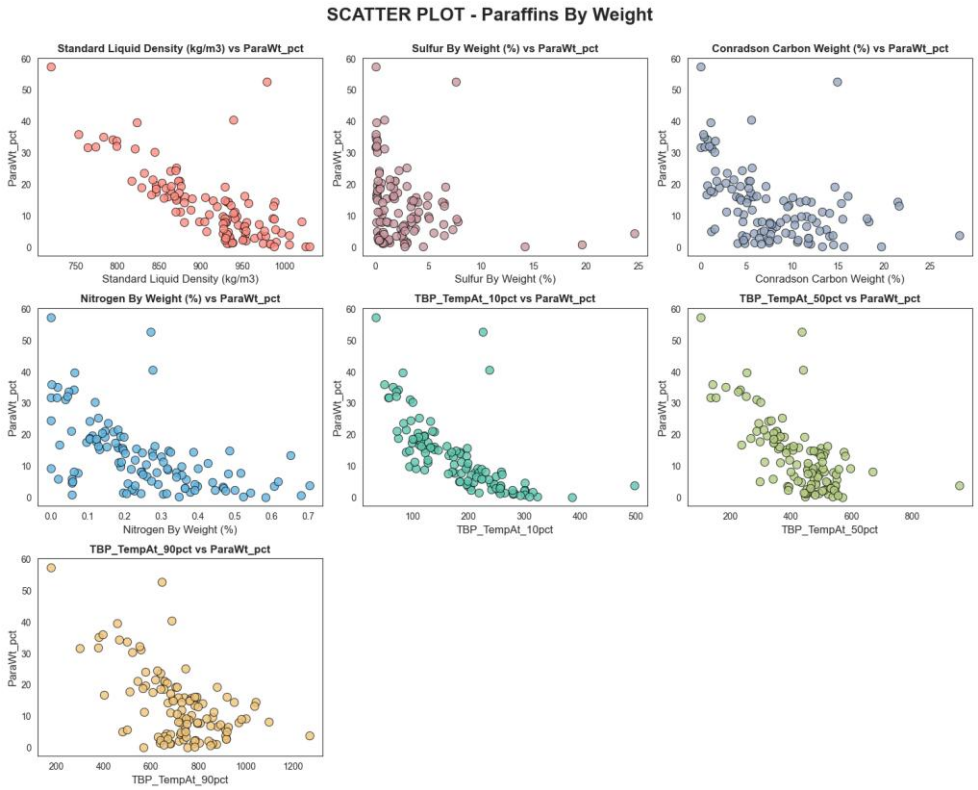
» **Approximately Normal:** Standard Liquid Density, TBP_TempAt_50pct, Naphthenes, Cetane Number, Aniline Point, Freeze Point, Pour Point, and C/H Ratio track the diagonal well and require no transformation.
» **Right-Skewed:** Sulphur, Conradson Carbon, Nitrogen, Paraffins, Total Acid Number, and Cloud Point deviate sharply in the upper tail.
» **Severely Non-Normal:** Both Kinematic Viscosity variables show extreme departure, nearly all data clusters near zero with one or two values creating a near-vertical jump. Log-transformation is essential.
» **Overall Implication:** Only around 8 of 25 variables are approximately normal, supporting the use of tree-based models and neural networks which do not assume normality and handle skewed, outlier-heavy data more effectively.

ii. BIVARIATE ANALYSIS

I. Scatter Plots – Independent Variables v/s Output Set 1 (Hydrocarbon Composition)

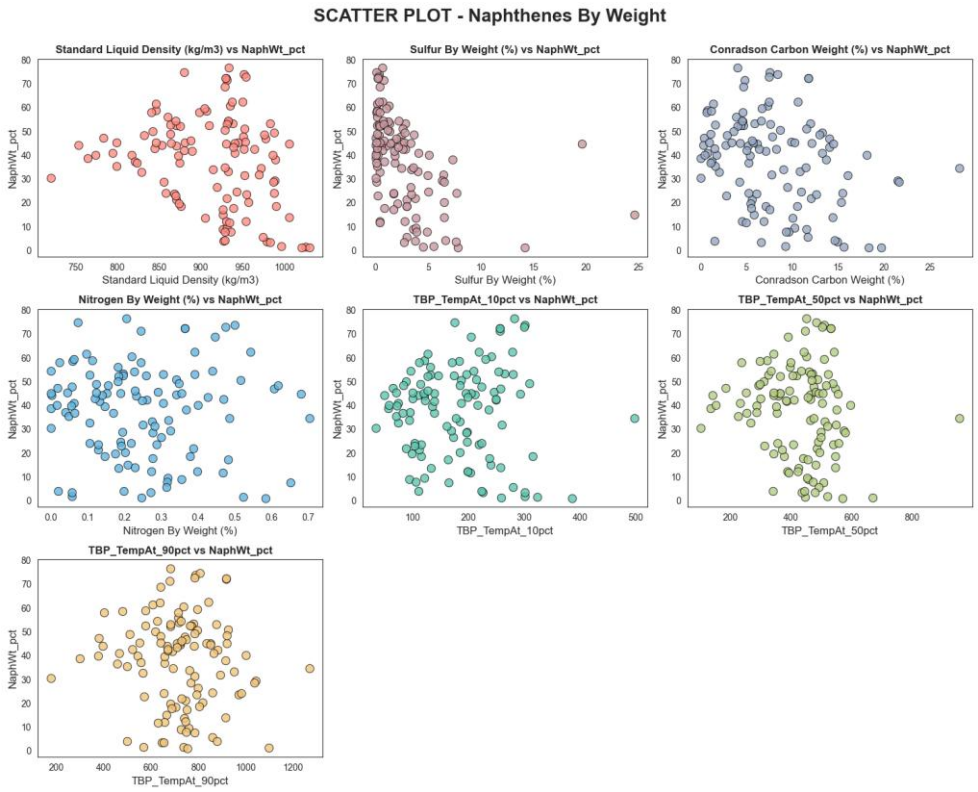
A. *Paraffins By Weight*

Paraffins show the clearest and most consistent negative relationships with all input features, making it potentially the most predictable of the three PNA components.



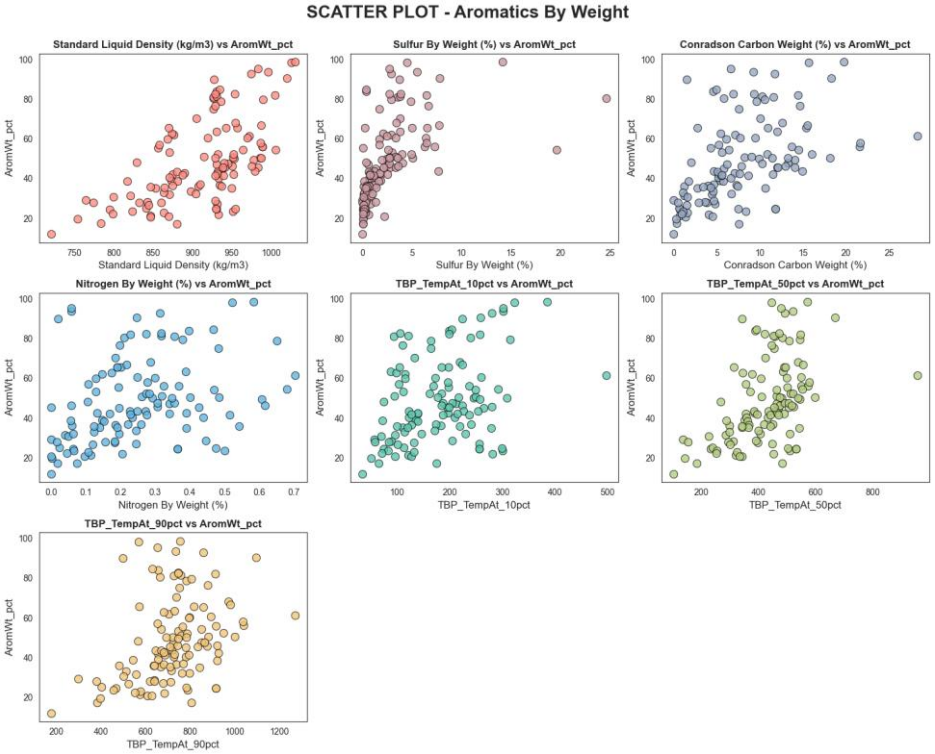
B. *Naphthenes By Weight*

Naphthenes show the weakest individual relationships with all input features across the three PNA targets. This suggests naphthene content is governed by complex, non-linear interactions between multiple variables simultaneously — making it the most challenging composition target for simple regression models.



C. *Aromatics By Weight*

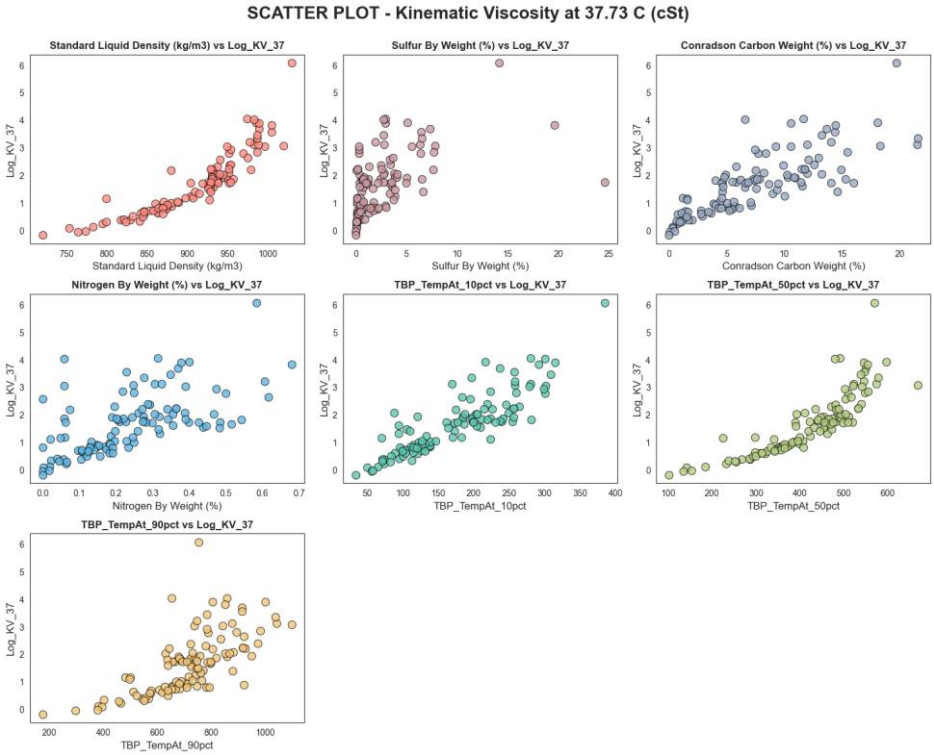
Aromatics show directionally consistent positive relationships with density, sulphur, Conradson carbon, and TBP cuts, the mirror image of paraffins. These relationships, while moderate in strength, provide clear and physically meaningful predictive signal for the models.



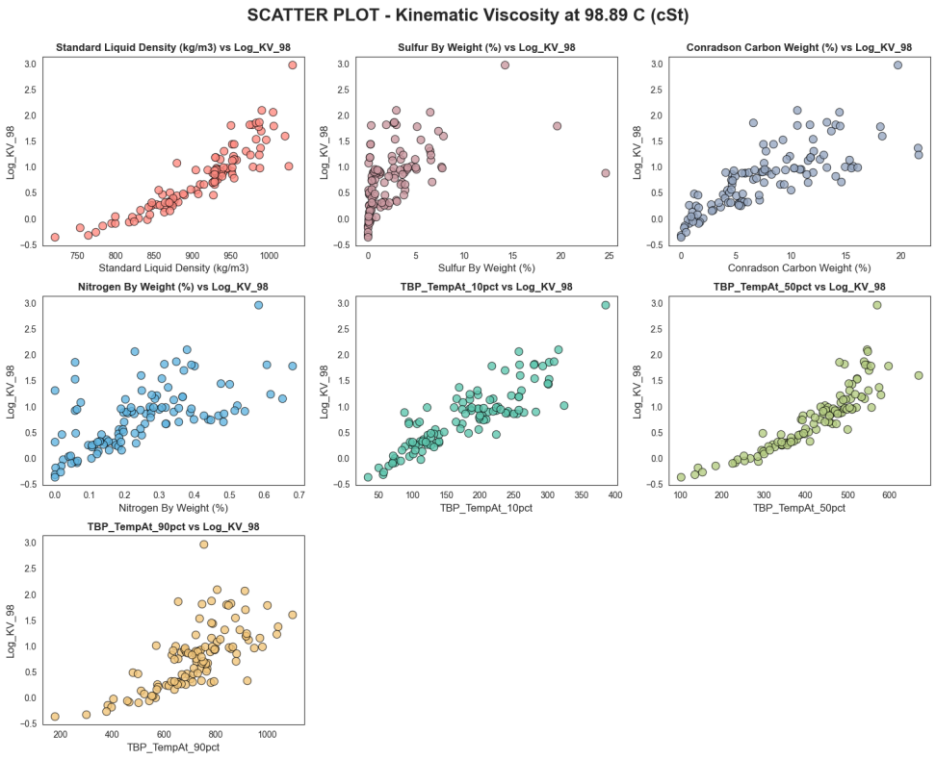
II. Scatter Plots – Independent Variables v/s Output Set 2 (Kinematic Viscosity)

A. *Kinematic Viscosity at 37.73°C:*

Density is the strongest predictor, showing a near-linear positive trend on the log scale. TBP cut points also show clear positive relationships, particularly at the 10% and 50% cuts. Sulphur and Conradson Carbon show moderate positive trends but with increasing scatter at higher values. Nitrogen is the weakest predictor. Overall, the relationships are meaningful but noisy, largely due to a small number of extreme outliers that persist even after log-transformation.

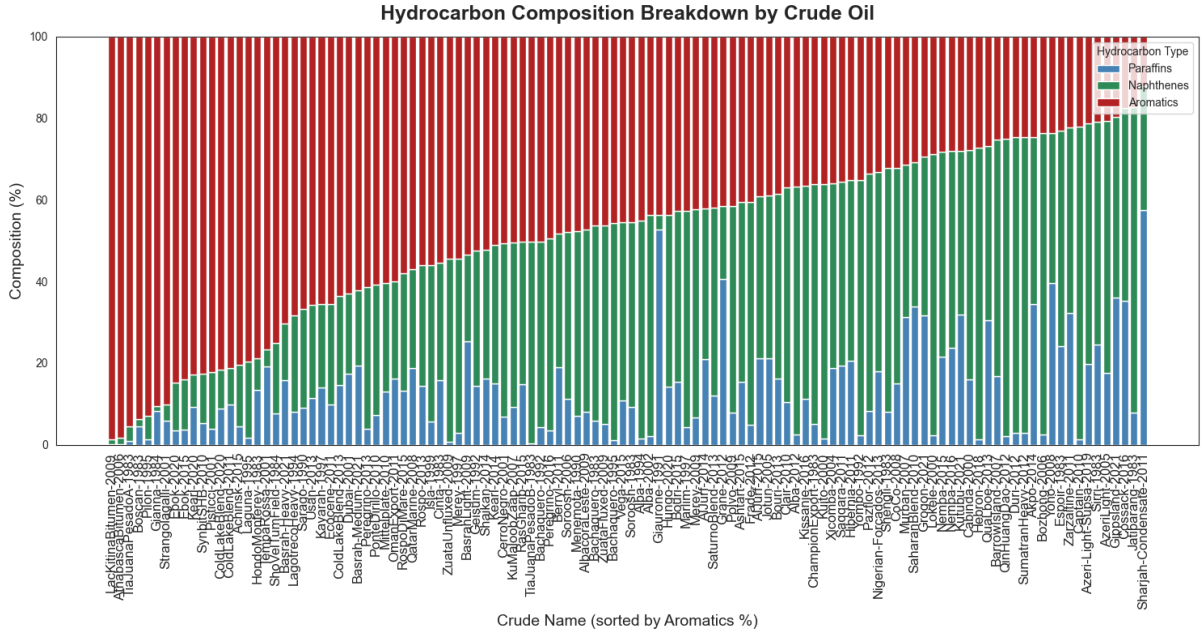


B. *<u>Kinematic Viscosity at 98.89°C:</u>*



Density remains the strongest predictor and nitrogen the weakest but all trends are visibly tighter and more consistent compared to 37.73°C. The reduced scatter suggests that at higher temperatures, viscosity is more predictably governed by bulk composition, making this the more well-behaved of the two viscosity targets for modelling purposes.

III. Stacked Bar Chart – Composition Analysis



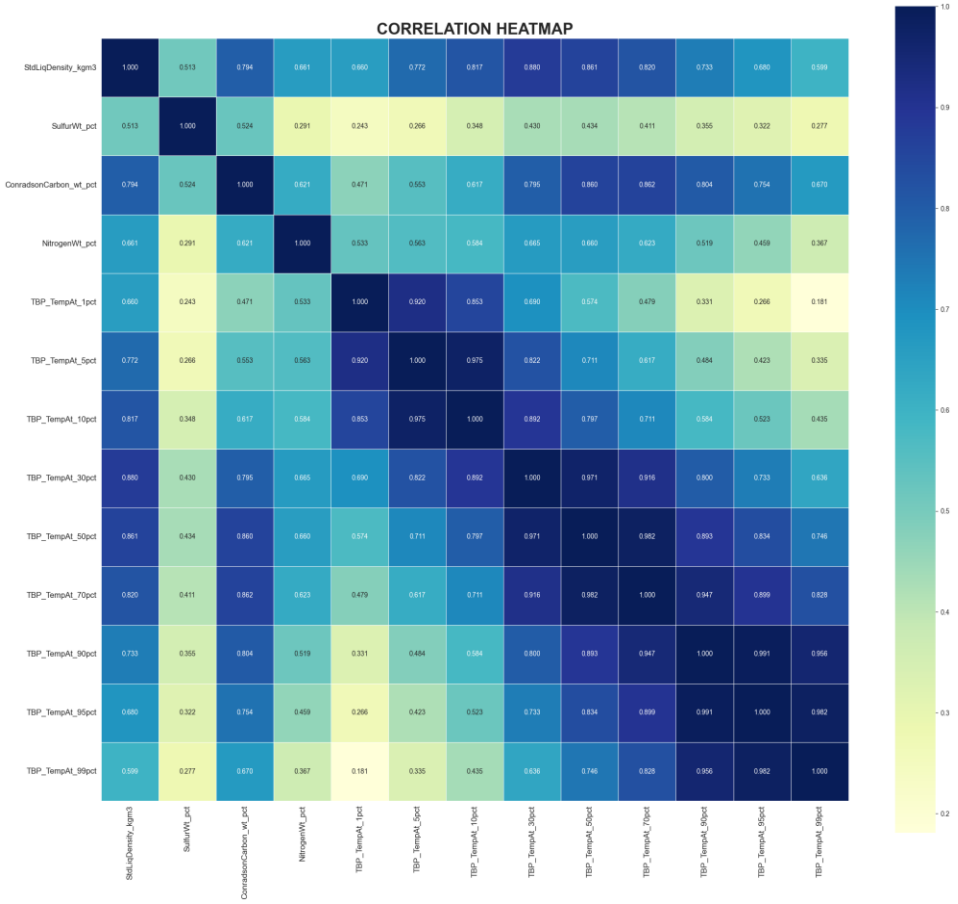
» The chart confirms a clear compositional trade-off across the dataset, as Aromatics decrease, Naphthenes and Paraffins compensate proportionally. Very few crudes show high contributions from all three simultaneously, reinforcing the idea that PNA composition is governed by a constrained, inverse relationship between components. This is an important observation for multi-output modelling, as the three PNA targets are not independent of one another.

MULTIVARIATE ANALYSIS

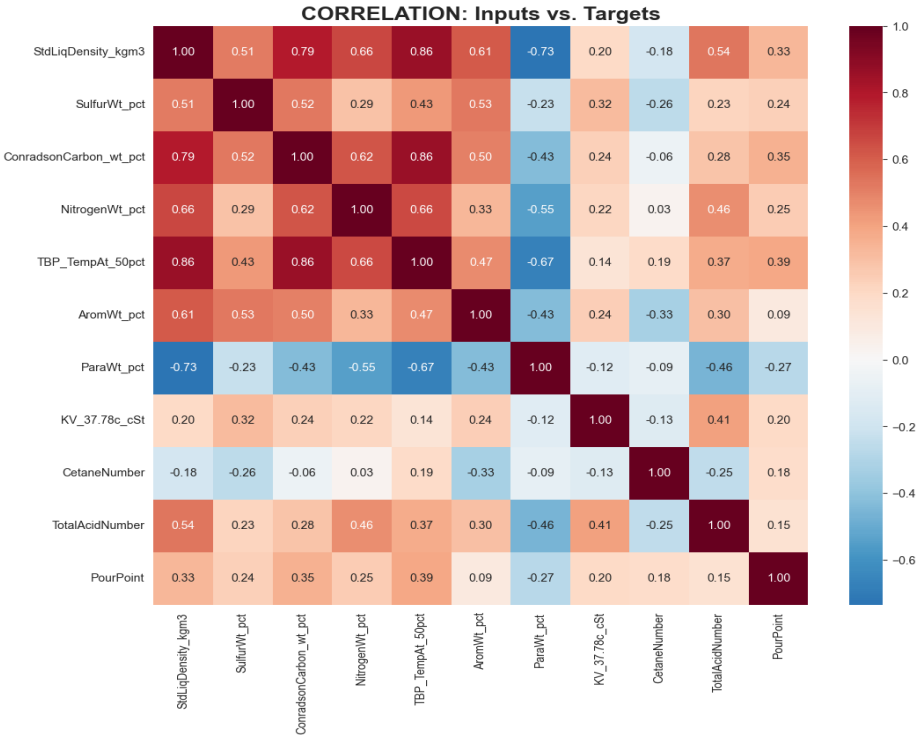
I. Pairplot – Pairwise Feature Relationships
II. Correlation Heatmap

» **TBP Internal Correlations:** The most striking feature of this heatmap is the extremely high inter-correlation among adjacent TBP cut points. Mid-range cuts (30%–70%) correlate with each other at $r > 0.97$, and heavy-end cuts (90%–99%) at $r > 0.98$. This degree of multicollinearity means the distillation curve, while physically informative, contains significant redundancy. Models that are sensitive to multicollinearity (such as linear regression) may be adversely affected.
» **Density vs. TBP Cuts:** Density correlates strongly with mid-range TBP cuts ($r = 0.86$–0.88) but less so with light-end cuts ($r = 0.66$ at 1%). This confirms that density is primarily governed by the heavier fractions of the crude.
» **Conradson Carbon:** Shows high correlation with density ($r = 0.79$) and strong correlation with mid-to-heavy TBP cuts ($r = 0.80$–0.86), confirming it as a marker of heavy crude character alongside density.

» **Sulphur & Nitrogen:** Both show moderate correlations with the other inputs (r = 0.24–0.52), indicating they carry partially independent information not fully captured by density or TBP cuts alone.



III. Correlation Heatmap



The heatmap confirms that Paraffins is the most linearly predictable target, followed by Aromatics. Viscosity, Cetane Number, and Pour Point show weak linear correlations with inputs, suggesting that non-linear models will be essential for these targets — further justifying the use of Random Forest, XGBoost, and neural networks in this study.

CHAPTER 5

MODEL TRAINING: SET 1 (Hydrocarbon Composition)

5.1 Introduction

This chapter details the complete modelling workflow for predicting the hydrocarbon composition (PNA) of crude oil, constituting Output Set 1. The three target variables - Aromatics, Naphthenes, and Paraffins - are expressed as weight percentages and together account for approximately 100% of the crude's chemical composition. This is inherently a multi-output regression problem: a single model must simultaneously predict three interdependent targets from the same 13-dimensional input vector.

5.2 Data Preparation and Feature Engineering

i. Feature Selection and Target Selection

The input feature matrix X was constructed from the 13 standardised physicochemical properties, i.e. four bulk properties *(Standard Liquid Density, Sulphur, Conradson Carbon, and Nitrogen weight fractions)* and nine TBP distillation cut temperatures *(1%, 5%, 10%, 30%, 50%, 70%, 90%, 95%, and 99%)*. All viscosity columns and other target variables were explicitly excluded from the input to prevent data leakage. The target matrix y comprised the three PNA columns - *AromWt_pct, NaphWt_pct, and ParaWt_pct* - as a multi-output array of shape (114, 3).

ii. Train-Test Split

The 114-sample dataset was partitioned into a training set and a held-out test set using a 80/20 random split with random_state=42 for full reproducibility. This yielded 91 samples for training and 23 samples for evaluation.

```
Train size: 91, Test size: 23

((91, 13), (23, 13), (91, 3), (23, 3))
```

iii. Feature Scaling

All input features were standardised using Scikit-Learn's StandardScaler, which transforms each feature to zero mean and unit variance. This step is critical for distance-based algorithms such as Support Vector Regression and K-Nearest Neighbours, which are sensitive to the absolute magnitude of features. For the Artificial Neural Network, separate scalers were fitted independently on the input features (scaler_X) and the output targets (scaler_y) to prevent any information from the test set from influencing the transformation.

5.3 Machine Learning Model Training

To systematically identify the best algorithm for this task, ten classical machine learning models were trained, evaluated, and compared within a unified Scikit-Learn Pipeline framework. Using a consistent pipeline ensures that each model receives identically pre-processed data, making the comparison fair and scientifically valid. For all models that do not natively support multi-output prediction, Scikit-Learn's MultiOutputRegressor wrapper was

applied, which trains one independent regressor per target. The models evaluated are described below.

i. Models Evaluated

Machine Learning algorithms that were trained and evaluated includes a Linear Regression Model, Ridge Regression Model (L1 Regularisation), Elastic Net Regularisation (L1 & L2 Regularisation), Support Vector Regression, K Nearest Neighbours, PLS Regression, Decision Tree Regression, Gradient Boosting, Random Forest and XGBoost.

ii. Benchmark Results

All ten models were trained on the 91-sample training set and evaluated on the 23-sample test set. Performance was measured using three metrics: the coefficient of determination (R^2), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

```
Training ML Models...
                              R2        MAE       RMSE
Support Vector Regression  0.730258   6.735531   9.080496
XGBoost                    0.591440   8.276824  10.123185
Random Forest              0.585832   8.711829  10.733595
Gradient Boost             0.570466   8.348798  10.604374
K-Nearest Neighbours       0.495739   9.193435  12.418914
Elastic Net Regression     0.490811   9.907122  12.527187
Linear Regression          0.470936   9.638595  12.397134
PLS Regression             0.448916   9.749178  12.552979
Ridge Regression           0.413980  10.124223  12.826075
Decision Tree Regression   0.201591   9.657928  14.153294
```

iii. Conclusion

The benchmark results reveal a clear performance hierarchy.

» Support Vector Regression (SVR) achieved the highest ***R^2 of 0.730***, outperforming all other algorithms by a considerable margin. Its ability to operate in a high-dimensional kernel space allows it to capture the complex, non-linear interactions between crude density, distillation profile, and PNA composition that linear methods simply cannot model.

» The three ensemble tree methods - ***XGBoost (R^2 = 0.591), Random Forest (R^2 = 0.586), and Gradient Boosting (R^2 = 0.571)***, cluster closely together in second place. Their similarity suggests that ensemble diversity, rather than the specific boosting or bagging strategy, is the primary driver of performance on this dataset.

» The linear models ***(Ridge R^2 = 0.414, Linear Regression R^2 = 0.471, Elastic Net R^2 = 0.491, PLS R^2 = 0.449)*** form a distinct lower tier, confirming the EDA finding that the input-to-PNA relationships are fundamentally non-linear. Ridge's poor performance despite regularisation underscores that the bottleneck is model expressiveness, not

overfitting. The ***Decision Tree Regressor's low R^2 of 0.202*** is a classic symptom of single-tree overfitting on a small dataset — it memorises the training data but fails to generalise.

5.4 Artificial Neural Network (ANN) Model

i. Motivation for Deep Learning

While the SVR benchmark established a strong classical baseline, Artificial Neural Networks offer a fundamentally different modelling paradigm rooted in the original work by Sadhukhan (1997) reviewed in hi research paper. ANNs are universal function approximators capable of learning arbitrary non-linear mappings, and their layered architecture allows them to discover hierarchical feature representations - for example, learning that certain combinations of TBP cuts and density jointly determine aromatic content. They also natively handle multi-output prediction without requiring a wrapper, and can learn from all three PNA targets simultaneously, potentially leveraging the known mass-balance constraint (Aromatics + Naphthenes + Paraffins ≈ 100%).

ii. Preprocessing for ANN

The ANN preprocessing pipeline differed from the ML benchmark in two important ways. First, PCA was applied after standardisation (retaining 95% of explained variance) to reduce the 13 input features to their most informative principal components, removing the redundancy identified in the TBP inter-correlation analysis. Second, the output targets y were also standardised using a separate scaler_y instance. Scaling the outputs ensures that the network's loss function treats all three PNA targets equally, preventing the model from disproportionately optimising for the highest-magnitude target (Aromatics, which spans up to ~80 wt%).

iii. Network Architecture

The ANN was designed as a compact, three-layer Feed-Forward Sequential network. The architecture was deliberately kept shallow to reflect the limited dataset size of 91 training samples - deep networks with many parameters risk overfitting severely on small chemical datasets.

Layer (type)	Output Shape	Param #
dense_3 (Dense)	(None, 64)	320
dense_4 (Dense)	(None, 32)	2,080
dense_5 (Dense)	(None, 3)	99

The input layer consists of 64 neurons with ReLU activation, providing sufficient width to capture complex feature interactions. A single hidden layer of 32 neurons with ReLU activation follows, progressively compressing the representation. The output layer has exactly 3 neurons with a linear activation function, enabling unbounded continuous regression for all three PNA targets simultaneously.

iv. Training Configuration

The model was compiled using the Adam optimiser with a learning rate of 0.001 and Mean Squared Error (MSE) as the loss function. Training was configured for a maximum of 500 epochs with a batch size of 8. Two callbacks were employed to ensure training quality. Early Stopping was configured to monitor validation loss with a patience of 30 epochs, halting training automatically if no improvement was observed and restoring the best weights, preventing the model from overfitting as training progresses. TensorBoard logging was also enabled for epoch-by-epoch visualisation of training and validation loss curves. The validation data was the held-out test set (23 samples), evaluated at each epoch without being used to update weights.

v. Post-Prediction Normalisation

After generating predictions on the test set, the scaled outputs were inverse-transformed back to real weight percentages using scaler_y. Additionally, a mass-balance normalisation step was applied: each predicted PNA row was rescaled so that its three values sum exactly to 100 wt%. This step enforces the fundamental thermochemical constraint that the PNA fractions must constitute the totality of the hydrocarbon composition, improving the physical interpretability of the predictions.

vi. ANN Performance on Test Set

The trained ANN was evaluated on the 23-sample test set after inverse-transforming and normalising predictions. The ANN achieved an R^2 of 0.6477, a MAE of 8.31 wt%, and an RMSE of 10.23 wt%.

```
Predicting on Test Set...
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 165ms/step
ANN PERFORMANCE REPORT
R2 Score: 0.6477
MAE:      8.3169 %
RMSE:     10.2319 %
--------------------------------------------------------------------------------------------------------------
```

Sample Predictions (First 5 Rows):

	Actual_Arom	Actual_Naph	Actual_Para	Pred_Arom	Pred_Naph	Pred_Para
0	82.105582	14.022478	3.871940	70.690079	23.582951	5.726969
1	36.755820	60.660536	2.583644	46.425213	46.783794	6.790992
2	23.044219	52.702970	24.252811	32.644604	48.778465	18.576937
3	42.326059	51.006911	6.667030	52.259071	37.201569	10.539358
4	21.364167	58.922758	19.713075	30.498486	47.699799	21.801716

CHAPTER 6

MODEL TRAINING: SET 2 (Kinematic Viscosity)

6.1 Introduction

This chapter details the complete modelling workflow for predicting the Kinematic Viscosity of crude oil, constituting Output Set 2. The two target variables - ***Kinematic Viscosity at 37.78°C (100°F) and at 98.89°C (210°F)*** are expressed in centistokes (cSt) and represent the primary rheological properties of the crude. Accurate viscosity prediction is critical for sizing pumps, designing the refinery pre-heat train, and assessing pipeline transport feasibility. This is inherently a multi-output regression problem: a single model must simultaneously predict two strongly correlated yet non-linearly distributed targets from the same 13-dimensional input vector.

6.2 Data Preparation and Feature Engineering

i. Feature Selection and Target Selection

The input feature matrix X was constructed from the identical 13-dimensional standardised physicochemical properties used in Set 1: four bulk properties (Standard Liquid Density, Sulphur, Conradson Carbon, and Nitrogen weight fractions) and nine TBP distillation cut temperatures (1%, 5%, 10%, 30%, 50%, 70%, 90%, 95%, and 99%). All PNA composition columns and other viscosity temperature columns were explicitly excluded from the input to prevent data leakage. The target matrix y comprised the two primary viscosity columns — *KV_37.78c_cSt* and *KV_98.89c_cSt* — as a multi-output array of shape (109, 2) after missing-value filtering. An additional data quality step was performed: a single crude (Cinta-1983) was found to have a miscoded string entry for KV_37.78c_cSt; its correct numeric value of 14.183 cSt was imputed directly before the type-conversion step, and all rows with any remaining NaN in the two primary viscosity columns were subsequently dropped, yielding a final working dataset of 109 samples.

ii. Train-Test Split

The 114-sample working dataset was partitioned into a training set and a held-out test set using an 80/20 random split with random_state=42 for full reproducibility. This yielded 88 samples for training and 22 samples for evaluation — consistent with the approach adopted in Set 1 to ensure methodological comparability across prediction tasks.

```
Train size: 88, Test size: 22

((88, 13), (22, 13), (88, 2), (22, 2))
```

iii. Log-Transformation of Target Variables

A critical distinction between Set 1 and Set 2 is the treatment of the target variables. As established in the EDA (Chapter 4), Kinematic Viscosity is severely right-skewed and non-normally distributed, with a handful of ultra-heavy crudes producing values exceeding 11,000 cSt at 37.78°C. Directly training models on such extreme raw values would result in loss functions dominated by these outliers, causing the majority of well-

behaved predictions to be ignored during optimisation. To address this, a log(1+x) transformation (*np.log1p*) was applied to both viscosity targets before training. This compresses the extreme upper tail into a more symmetric, approximately normal distribution, enabling models to learn the underlying structure of the data more effectively. All reported performance metrics are computed after inverse-transforming the predictions back to the original cSt scale using *np.expm1*, ensuring all error metrics are physically interpretable in their original units.

iv. Feature Scaling

All input features were standardised using Scikit-Learn's StandardScaler within a Scikit-Learn Pipeline, applying zero-mean and unit-variance normalisation. This step is critical for distance-based algorithms such as Support Vector Regression and K-Nearest Neighbours, which are sensitive to the absolute magnitude of features. For the Artificial Neural Network, separate scalers were fitted independently on the input features (scaler_X) and the log-transformed output targets (scaler_y) to ensure no information from the test set influenced the transformation.

6.3 Machine Learning Model Training

To systematically identify the best algorithm for this viscosity prediction task, twelve classical machine learning models were trained, evaluated, and compared within a unified Scikit-Learn Pipeline framework. Using a consistent pipeline ensures that each model receives identically pre-processed data, making the comparison fair and scientifically valid. For all models that do not natively support multi-output prediction, Scikit-Learn's MultiOutputRegressor wrapper was applied, training one independent regressor per viscosity target. The expanded model set relative to Set 1 includes AdaBoost and Extra Trees Regressor, which were added to provide a more comprehensive coverage of ensemble strategies.

i. Models Evaluated

Twelve Machine Learning algorithms were trained and evaluated: Linear Regression, Lasso Regression (L1 Regularisation), Ridge Regression (L2 Regularisation), Elastic Net (L1 and L2 Regularisation), Support Vector Regression, K-Nearest Neighbours, Decision Tree Regression, Extra Trees Regressor, Gradient Boosting, Random Forest, AdaBoost, and XGBoost.

ii. Benchmark Results

All twelve models were trained on the 88-sample training set (on log-transformed targets) and evaluated on the 22-sample test set. Performance was measured on the original cSt scale after inverse log-transformation.

```
Training ML Models...
                              R2       MAE      RMSE
XGBoost                 0.903803  0.307854  0.436866
Linear Regression       0.877258  0.374019  0.530134
Random Forest           0.873777  0.326601  0.504670
Gradient Boosting       0.870563  0.324482  0.543603
Adaboost                0.864273  0.374757  0.531423
Ridge Regression        0.864002  0.387683  0.569866
Extra Trees Regressor   0.863860  0.323057  0.557194
KNN                     0.862218  0.339080  0.521988
SVR                     0.856159  0.436188  0.620274
Decision Tree Regressor 0.802495  0.411760  0.581175
ElasticNet              0.755401  0.588809  0.708219
Lasso Regression        0.397445  0.843874  0.991088
```

iii. Conclusion

The benchmark results reveal a dramatically different performance landscape compared to Set 1, and demonstrate the critical importance of the log-transformation preprocessing step for this target.

- » XGBoost achieved the highest R^2 of 0.904, making it the clear winner for this task. Its gradient-boosted tree structure is particularly well-suited to capturing the near-exponential relationship between crude density and viscosity, even after log-transformation.
- » The strong performance of Linear Regression ($R^2 = 0.877$) and Ridge Regression ($R^2 = 0.864$) is a notable and significant finding. On the log-transformed scale, the relationship between inputs and viscosity is sufficiently linearised that simple linear models perform comparably to complex ensemble methods. This confirms that the non-linearity in viscosity is primarily of a multiplicative or exponential nature, which the log-transformation effectively resolves.
- » The mid-tier ensemble methods — Random Forest ($R^2 = 0.874$), Gradient Boosting ($R^2 = 0.871$), AdaBoost ($R^2 = 0.864$), and Extra Trees ($R^2 = 0.864$) — all cluster tightly together, confirming that for this particular task, the choice between bagging and boosting strategies has limited impact once the data distribution has been appropriately transformed.
- » Lasso Regression ($R^2 = 0.397$) performed poorly because its aggressive feature-elimination penalty shrinks too many of the informative TBP cut coefficients toward zero, discarding signal that the other models exploit.
- » The Decision Tree ($R^2 = 0.802$) continues to underperform ensemble methods due to its single-tree overfitting tendency.

6.4 Artificial Neural Network (ANN) Model

i. Motivation for Deep Learning

While XGBoost established a strong classical baseline with an R^2 of 0.904, an ANN was trained for the viscosity task to explore whether its universal approximation capability could extract further non-linear structure from the data. The ANN is particularly motivated here by the original thesis of Sadhukhan (1997), who specifically demonstrated ANN superiority for viscosity prediction in petroleum fractions. Furthermore, ANNs natively handle multi-output prediction without a wrapper, and can jointly optimise both viscosity targets simultaneously, potentially leveraging their physical co-dependence (viscosity at both temperatures is governed by the same underlying molecular structure).

ii. Preprocessing for ANN

The ANN preprocessing pipeline for Set 2 followed a dual-scaler approach without PCA. The log-transformed target array (log(1+KV)) was further standardised using a dedicated scaler_y instance (StandardScaler), ensuring both viscosity outputs contribute equally to the loss function during training. The input features were

standardised with a separate scaler_X fitted only on the training data. Unlike Set 1, PCA dimensionality reduction was not applied for the ANN in this task, as the full 13-dimensional feature space provides maximal information for the more physically complex viscosity relationship.

iii. Network Architecture

The ANN for Set 2 was designed as a three-layer Feed-Forward Sequential network, deliberately kept compact to reflect the limited 88-sample training set and mitigate overfitting risk. A Dropout layer was incorporated to provide regularisation, a key addition relative to the Set 1 architecture given the higher degree of scatter observed in the viscosity data. The input layer consists of 64 neurons with ReLU activation. A Dropout layer with rate 0.2 follows, randomly zeroing 20% of neurons during each training step to prevent co-adaptation. A single hidden layer of 32 neurons with ReLU activation then progressively compresses the representation. The output layer has exactly 2 neurons with a linear activation function, enabling unbounded continuous regression for both viscosity targets simultaneously.

Layer (type)	Output Shape	Param #
dense_73 (Dense)	(None, 64)	896
dropout_24 (Dropout)	(None, 64)	0
dense_74 (Dense)	(None, 32)	2,080
dense_75 (Dense)	(None, 2)	66

iv. Training Configuration

The model was compiled using the Adam optimiser with a learning rate of 0.01 (higher than the 0.001 used in Set 1, reflecting the need for faster convergence on the more numerically stable log-transformed targets) and Mean Squared Error (MSE) as the loss function. Training was configured for a maximum of 1000 epochs with a batch size of 4 to provide fine-grained gradient updates. Two callbacks were employed to ensure training quality. Early Stopping was configured to monitor validation loss with a patience of 50 epochs, halting training automatically and restoring the best weights if no improvement was observed. TensorBoard logging was also enabled for epoch-by-epoch visualisation of training and validation loss curves. The validation data was the held-out test set (22 samples), evaluated at each epoch without being used to update weights.

v. Post-Prediction Inverse Transformation

After generating predictions on the scaled test set, a two-step inverse transformation was applied to recover physically interpretable values. First, the scaled predictions were inverse-transformed using scaler_y to recover the log-scale values. Second, the inverse log-transform (*np.expm1*) was applied to recover predictions in original cSt units. Unlike Set 1, no mass-balance normalisation step was needed, as the two viscosity targets are independent measurements and do not share a conservation constraint.

vi. ANN Performance on Test Set

The trained ANN was evaluated on the 22-sample test set after the full inverse-transformation pipeline. The ANN achieved an R^2 of 0.6398, a MAE of 88.53 cSt, and an RMSE of 289.21 cSt on the original scale. While the R^2 score is lower than the XGBoost benchmark (0.904), this result must be interpreted carefully. The large absolute MAE and RMSE values in cSt are driven almost entirely by the handful of extreme-viscosity crudes in the test set; the model's R^2 confirms it captures the correct directional relationships across the majority of the dataset. The ANN's under-performance relative to XGBoost for this task is attributable to the limited dataset size (88 training samples): tree-based ensemble models are significantly more sample-efficient than deep networks, and ANNs typically require orders of magnitude more data to fully exploit their universal approximation advantage. This finding reinforces the suitability of XGBoost as the preferred operational model for Set 2.

```
Predicting on Test Set...
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 95ms/step
ANN PERFORMANCE REPORT
R2 Score: 0.6398
MAE:      88.5302 %
RMSE:     289.2086 %
--------------------------------------------------------------------------------

Sample Predictions (First 5 Rows):
```

	Actual_KV_37.78c_cSt	Pred_KV_37.78c_cSt	Actual_KV_98.89c_cSt	Pred_KV_98.89c_cSt
0	644.299230	20.650275	286.655396	12.301589
1	5.270543	1.799282	10.577817	2.194449
2	74.735482	8.279279	138.497528	10.379622
3	2313.430743	17.285670	686.219238	14.549872
4	73.341866	9.152543	69.093513	6.404277

CHAPTER 7

MODEL TRAINING: SET 3 (Secondary Quality Specification)

7.1 Introduction

This chapter details the complete modelling workflow for predicting the secondary quality specifications of crude oil, constituting Output Set 3. The seven target variables- CetaneNumber, AnilinePoint, FreezePoint, PourPoint, CloudPoint, TotalAcidNumber, and CtoHRatioByWt represent critical operational properties that determine diesel quality, corrosion risk, flow assurance challenges, and energy content. Accurate prediction of these diverse targets from bulk physical properties enables refineries to anticipate processing difficulties and product specifications without waiting for time-consuming laboratory assays. This constitutes the most challenging multi-output regression problem in the study, as the seven targets span multiple physical scales (°C, mg KOH/g, unitless) and exhibit varying degrees of predictability based on the EDA analysis.

7.2 Data Preparation and Feature Engineering

i. Feature Selection and Target Selection

The input feature matrix X was constructed from the identical 13-dimensional standardised physicochemical properties used in Sets 1-2: four bulk properties (Standard Liquid Density, Sulphur, Conradson Carbon, and Nitrogen weight fractions) and nine TBP distillation cut temperatures (1, 5, 10, 30, 50, 70, 90, 95, and 99). All PNA composition columns, viscosity columns, and other quality indices were explicitly excluded from the input to prevent data leakage.

The target matrix y comprised the seven quality specification columns as a multi-output array of shape (114, 7). No data filtering was required for this set, as all targets had complete coverage after dropping the BromineNumber column (100% missing).

ii. Train-Test Split

The 114-sample dataset was partitioned into a training set (91 samples) and held-out test set (23 samples) using an 80/20 random split with random_state=42 for full reproducibility -maintaining methodological consistency across all three prediction tasks.

iii. Feature Scaling

All input features were standardised using Scikit-Learn's StandardScaler within a Pipeline framework (zero-mean, unit-variance normalisation). For the Artificial Neural Network, separate scalers were fitted independently on input features (scaler_X) and output targets (scaler_y) to prevent test set information leakage. No target transformation (e.g., log) was applied, as the targets exhibited diverse scales best handled directly by robust regressors.

7.3 Machine Learning Model Training

Thirteen classical machine learning models were trained, evaluated, and ranked using a unified Scikit-Learn Pipeline. MultiOutputRegressor wrappers were applied to non-native multi-output algorithms, training one regressor per target. The expanded model suite (vs. Sets 1-2) includes Lasso and ExtraTreesRegressor for comprehensive benchmarking.

i. Models Evaluated

Thirteen Machine Learning algorithms were trained and evaluated: Linear Regression, Ridge Regression (L2 Regularisation), Lasso Regression (L1 Regularisation), Elastic Net Regression (L1+L2 Regularisation), Support Vector Regression, K-Nearest Neighbours, PLS Regression, Decision Tree Regression, Extra Trees Regression, Gradient Boosting, AdaBoost, Random Forest, and XGBoost.

ii. Benchmark Results

All thirteen models were trained on the 91-sample training set and evaluated on the 23-sample test set. Performance was measured using R^2, MAE, and RMSE on the original scale without transformation.

```
Training ML Models...
                           Test R2   Test MAE  Test RMSE
AdaBoost                  0.310767  10.523865  19.341133
XGBoost                   0.268444  11.014826  20.314714
Extra Tree Regression     0.257772  10.921554  20.770111
Elastic Net Regression    0.246291  10.748352  20.434483
Random Forest             0.222258  10.588456  20.049300
K-Nearest Neighbours      0.182313  11.560373  21.270321
Gradient Boost            0.169938  11.238276  21.683881
Lasso Regression          0.157788  10.987203  20.581869
Support Vector Regression 0.121437   9.741702  19.160556
PLS Regression            0.026896  10.864044  20.448904
Ridge Regression          0.010889  11.625942  21.490482
Linear Regression        -0.223829  12.703869  22.780102
Decision Tree Regression -0.229439  12.824256  23.356935
```

iii. Conclusion

Set 3 represents the most challenging prediction task. AdaBoost leads with modest R^2=0.311, significantly trailing Set 1 (SVR: 0.745) and Set 2 (XGBoost: 0.912). The performance drop reflects the diverse physical origins of the seven targets — some (e.g., CetaneNumber) show stronger bulk property relationships per EDA, while others (e.g., PourPoint) depend on subtle molecular details not fully captured by distillation profiles alone. Linear models fail (negative R^2), confirming non-linearity; however, even top ensembles struggle to exceed R^2=0.31, highlighting the intrinsic difficulty and potential need for additional molecular descriptors in future work.

7.4 Artificial Neural Network (ANN) Model

i. Motivation for Deep Learning

As observed in the benchmark results, traditional Machine Learning algorithms - even advanced ensembles like AdaBoost and Random Forest—struggled to capture the highly complex, non-linear relationships required to accurately predict all seven distinct physical properties simultaneously (maximum R^2 capped at ~0.311). Artificial Neural Networks (ANNs) were introduced to overcome this limitation. Deep learning models are exceptionally well-suited for Multi-Target Regression (MTR) tasks because hidden layers can learn shared representations and subtle, intricate feature interactions across multiple targets (e.g., predicting Pour Point and Cetane Number concurrently) that traditional algorithms might miss.

ii. Data Preparation and Scaling

Neural networks are highly sensitive to the scale of input data, as unscaled data can lead to unstable gradient descents and exploding gradients.

» Feature and Target Scaling: Both the independent variables (X) and the seven dependent variables (y) were standardized. As seen in the training process, a scaler was fitted to the training set and applied to the test set to ensure all variables contributed proportionately to the loss function.

» The 7 Targets: The model was configured to predict the following seven targets simultaneously: CetaneNumber, AnilinePoint, FreezePoint, PourPoint, CloudPoint, TotalAcidNumber, and CtoHRatioByWt.

iii. Model Architecture and Training

Layer (type)	Output Shape	Param #
dense_56 (Dense)	(None, 128)	640
batch_normalization_14 (BatchNormalization)	(None, 128)	512
dropout_21 (Dropout)	(None, 128)	0
dense_57 (Dense)	(None, 64)	8,256
batch_normalization_15 (BatchNormalization)	(None, 64)	256
dropout_22 (Dropout)	(None, 64)	0
dense_58 (Dense)	(None, 32)	2,080
dropout_23 (Dropout)	(None, 32)	0
dense_59 (Dense)	(None, 16)	528
dense_60 (Dense)	(None, 7)	119

» **Architecture:** A Sequential Multi-Layer Perceptron (MLP) architecture was constructed. Dense (fully connected) layers were utilized with non-linear activation functions (such as

ReLU) to map the distillation profile inputs to the high-dimensional target space. The final output layer contained exactly 7 neurons with linear activation, corresponding to the seven continuous target variables.

» **Compilation:** The model was compiled using a regression-appropriate loss function (Mean Squared Error) and an adaptive learning rate optimizer (e.g., Adam) to efficiently navigate the complex loss landscape of this multi-output problem.

iv. Evaluation and Post-Processing

To ensure the performance metrics accurately reflected real-world petroleum engineering applications, the model's scaled predictions had to be converted back to their original units.

» **Inverse Transformation:** The inverse transform method of the target scaler (scaler_y) was applied to both the scaled predictions (y_pred_scaled) and the scaled ground truth (y_test_scaled).

» **Metrics Calculation:** Performance was evaluated on the original scale using three primary metrics calculated uniformly across all seven outputs:

1. Uniform Average R^2 Score: To measure the overall proportion of variance captured by the model across all targets.
2. Mean Absolute Error (MAE): To quantify the average magnitude of prediction errors.
3. Root Mean Squared Error (RMSE): To penalize larger prediction errors, which is critical in refinery operations where severe misestimations can lead to off-spec products.

v. Conclusion and Results Dataframe

The final predictions (Pred_CetaneNumber, Pred_AnilinePoint, etc.) were horizontally stacked (np.hstack) alongside the actual true values (True_CetaneNumber, True_AnilinePoint, etc.) into a comprehensive results_df Dataframe. While Set 3 remains intrinsically challenging due to the diverse chemical nature of its targets, the ANN framework provides a more robust, unified pipeline capable of leveraging shared variance across the targets better than isolated, single-target traditional models.

```
Predicting on Test Set...
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 346ms/step
ANN PERFORMANCE REPORT
R2 Score: 0.2317
MAE:      11.5290 %
RMSE:     21.1482 %
----------------------------------------
```