MACHINE LEARNING AND OPTIMIZATION OF OIL EXTRACTION RATE (OER) IN PALM OIL MILLS

Developed by: Amir Ashraf Izham

Date: 13 January 2025

INTRODUCTION

Malaysia is the second-largest producer of palm oil in the world, supplying about one-third of the global demand. Palm oil is a key part of the country's economy, contributing 2.4% to its GDP. The industry has also led to the growth of related sectors, such as palm oil mills, machinery manufacturers, and oleochemical factories.

Despite its importance, the industry faces challenges in improving efficiency. For example, between 1988 and 1994, the national average Oil Extraction Rate (OER) declined by 1.24%, with a 1.36% drop recorded in Peninsular Malaysia. This highlights the need for better methods to improve productivity.

Artificial Intelligence (AI) and Machine Learning (ML) offer exciting opportunities to address these challenges. By using these technologies, the industry can optimize processes, increase the OER, and improve overall performance. This document explores how AI and ML can help transform Malaysia's palm oil sector for a more efficient and sustainable future.

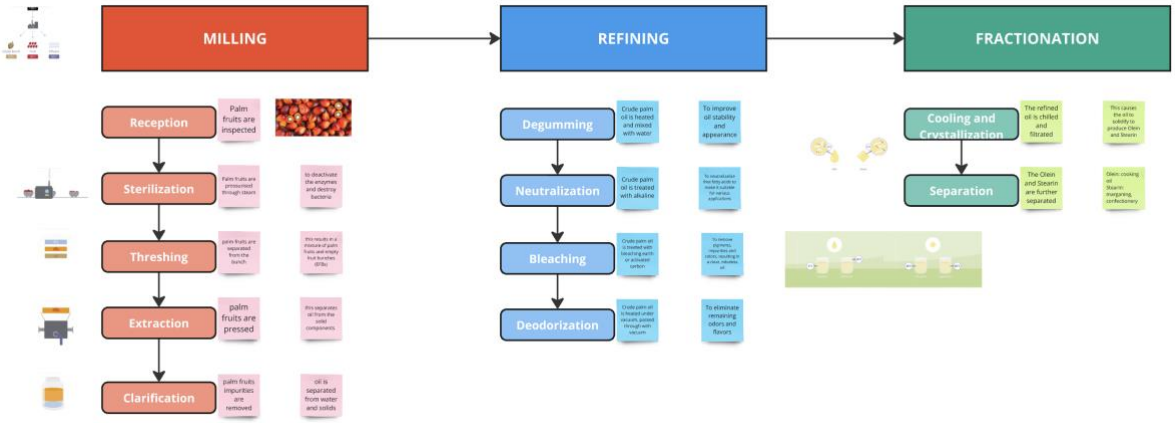


Figure 1: A diagram of how palm oil is processed

Before implementing our AI and ML systems, it is important to understand the process of palm oil production. The production of palm oil involves three main steps:

1. Milling

The milling process extracts crude palm oil (CPO) from fresh fruit bunches. The key stages in this process include:

- **Reception**: Receiving and inspecting fresh fruit bunches.
- **Sterilization**: Cooking the fruit bunches with steam to loosen the fruit and deactivate enzymes.
- **Threshing**: Separating the palm fruits from the bunches.
- **Extraction**: Pressing the fruits to extract the crude palm oil.
- **Clarification**: Removing impurities from the extracted oil to produce clear CPO.

2. Refining

Refining improves the quality of crude palm oil to make it suitable for consumption or industrial use. The refining process includes:

- **Degumming**: Heating and mixing CPO with water to improve oil stability
- **Neutralization**: Treating the CPO with alkaline to reduce acidity
- **Bleaching**: Treating the CPO with bleaching earth to remove color pigments and impurities.
- **Deodorization**: Heating the CPO under pressure to eliminate remaining odors.

3. Fractionation

Fractionation separates palm oil into different components based on melting points. This step involves:

- **Cooling and Crystallization**: Cooling the oil to form crystals.
- **Separation**: Separating the liquid oil (olein) from the solid fat (stearin).

By understanding these steps, we can identify areas where AI and ML can enhance efficiency, optimize processes, and improve overall productivity in palm oil production. For this project, our primary focus will be on the Milling section, as it aligns with the data provided by the client.

Variable	Remarks	Target/Threshold
Oil Extraction Rate (OER) %	The output to be forecasted	Acceptable threshold: 19% and above
Crop Freshness Score	When the crops processed are fresher, the OER% improves.	Acceptable threshold : 280 and above
Ripe %	A higher ripe% leads to higher OER%.	Acceptable threshold : 90% and above
Long Stalk %	Higher long stalks% mean lesser oil extracted.	Acceptable threshold : 5% and below
Rat Damage %	The more damage to the fruits caused by rat bites, the lower the OER.	Acceptable threshold: 5% and below
Loose Fruits %	More loose fruits mean more oil can be extracted because they have higher oil content to weight ratio.	Acceptable threshold: 8% and above
Rainfall (mm)	Heavy rainfall reduces OER%. 25mm/day is considered heavy rainfall.	Optimal rainfall level is 150mm/month (~5mm/day)
Age Profile (years)	Higher age profile causes lower OER%.	N/A
Total Oil Loss (OL)	The lower the OL%, the higher the OER%.	Acceptable threshold: 1.40 and below
Downtime %	The higher the downtime %, the lower the OER%.	Acceptable threshold: 6% and below
FFB Processed MT	Fresh Fruit Bunch - Higher FFB processed, higher utilisation rate, higher OER	N/A
Seed Type	Seed A + Seed B + Other Seeds = 100%	N/A
Topography	Coastal % + Inland % = 100%	N/A

Figure 2: Plant Milling Data

Using the data provided, we will develop developing a machine learning and optimization model to predict and increase the **Oil Extraction Rate (OER)**. The model will leverage the variables outlined in the table to identify key factors influencing the OER and provide actionable insights for improvement

SECTION 1: DESCRIPTIVE ANALYTICS

Before building the machine learning models, we will conduct a thorough analysis of the provided data to uncover key insights and patterns. This step is crucial to ensure the model accurately represents the palm oil mill's operations and effectively addresses the problem at hand.

i. How are the numerical features correlated to the label?

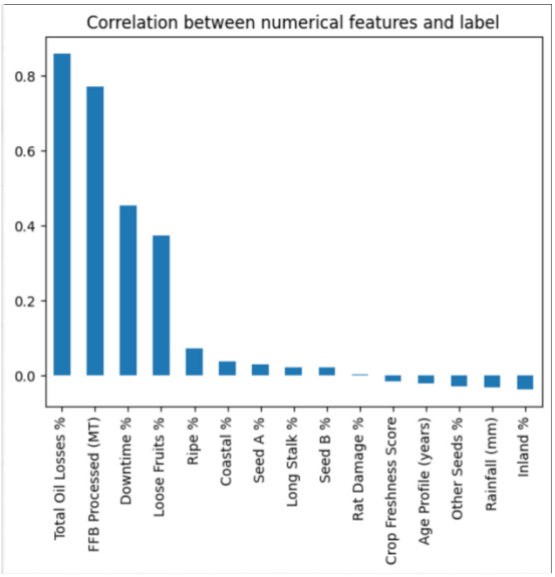


Figure 3: Global Correlation

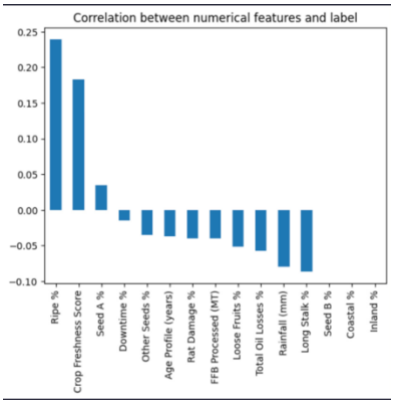


Figure 4: Correlation at Mill Z022

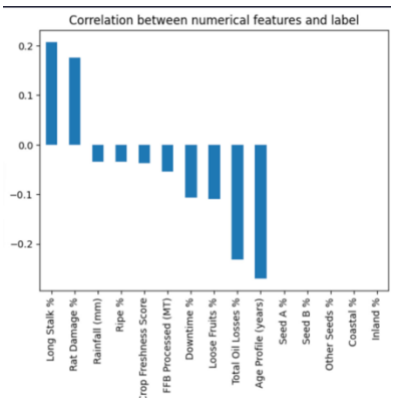


Figure 5: Correlation at Mill Z017

To identify the features most strongly correlated with the label, we calculate **Pearson's correlation coefficients**. Values closer to +1 indicate a strong positive relationship, while values closer to -1 signify a strong negative relationship. From the diagram above, the top three features most correlated with the Oil Extraction Rate (OER) are:

- Total Oil Losses (%)
- FFB Processed (MT)
- Downtime (%)

However, the correlation mentioned above is based on data aggregated across all palm oil mills. When we analyze the correlation within individual mills, the relationships between features and the Oil Extraction Rate (OER) differ significantly. In Mill Z022, the top three features correlated with OER are: Ripe %, Crop Freshness Score, Seed A%. In Mill Z017, the top three features are: Long Stalk %, Rat Damage %, Rainfall (mm). This factors influencing OER differ across mills might be due to varying operational and environmental conditions. **This insight suggests that developing individual models tailored to each mill is more effective than using a single global model** since the factors influencing the OER vary significantly between mills.

ii. How are the numerical features correlated with one another?

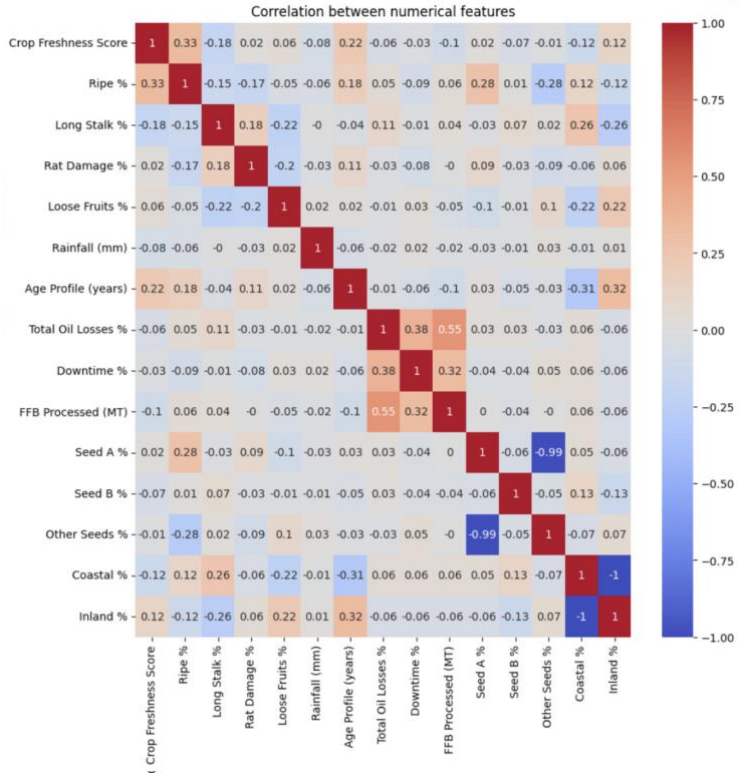


Figure 6: Heatmap of correlation between Features

Similarly, we use **Pearson's correlation coefficients,** visualized through a heatmap, to determine how the numerical features are correlated with one another. This helps us identify multicollinearity, which could weaken prediction performance. **Multicollinearity** occurs when two or more features are highly correlated, meaning they provide redundant information. This can negatively impact the performance of our machine learning model by making it harder to isolate the contribution of individual features to the predictions.

From the chart, we observe that perfect correlations exist between:

- Seed A% and Other Seeds % (-0.99): *Seed A is inversely related to Other Seeds*
- Coastal % and Inland % (-1): *An increase in one proportion automatically decreases the other*

These correlations are expected, as the features share a linear relationship based on the following equations:

Relationship 1:

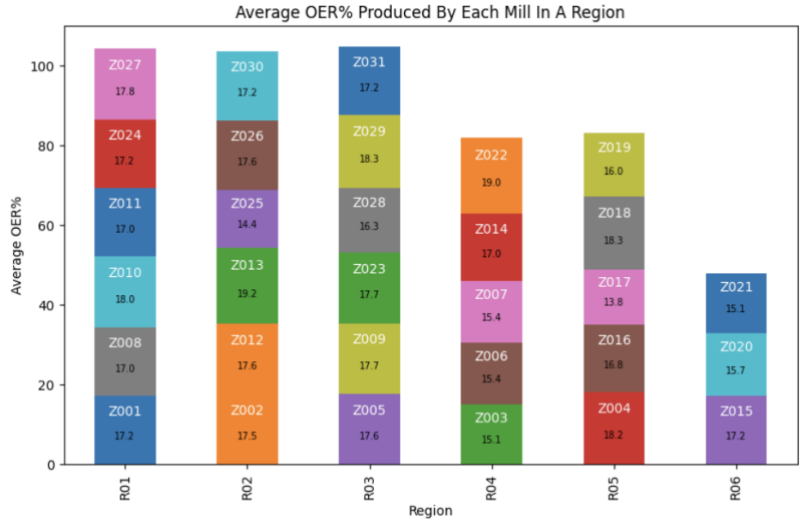
$$Seed\ A\ \% + Seed\ B\ \% + Other\ Seeds\ \% = 100\%$$

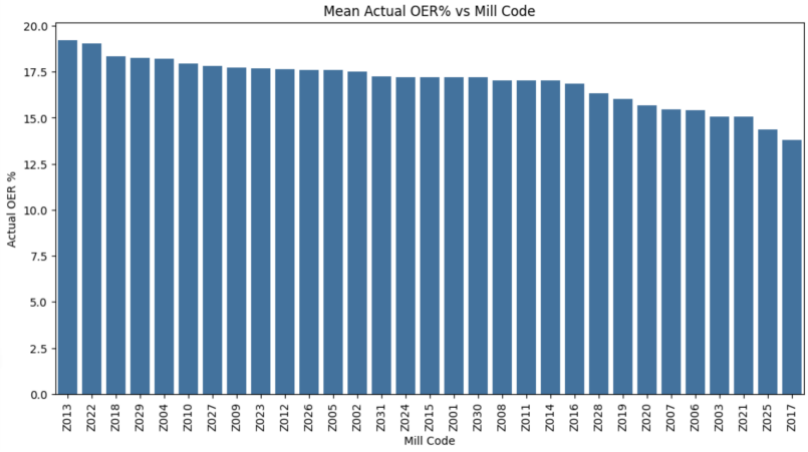
Relationship 2:

$$Coastal\ \% + Inland\ \% = 100\%$$

These insights suggest that **retaining only one feature from highly correlated pairs is necessary** to reduce redundancy and ensure the model can make predictions efficiently. By removing or consolidating redundant features, we can **minimize the risk of multicollinearity**, which could otherwise impact the stability and interpretability of the model.

iii. What is the OER% production in each mill and region?





From the charts above, we observe that, on average, Regions R01, R02, and R03 (top-performing regions) produce similar OER% compared to R04, R05, and R06 (bottom-performing regions). Additionally, we analyzed the average OER% of each mill from 2020 to 2024. To address our hypotheses, we focus on the **top two (ZO13, Z022)** and **bottom two (Z025, Z017)** performing mills. In this discussion, performance is defined by the average OER% achieved by a mill or region. Therefore, a top-performing mill or region is one that demonstrates a higher average OER% compared to a bottom-performing mill or region.

Hypothesis #1: Do top-performing regions have a higher concentration of top-performing mills?

Mill	Top Performing Mill?	Expected	Actual
ZO13	True	True	True (RO2)
ZO22	True	True	False (RO4)
ZO25	False	False	True (RO2)
ZO17	False	False	False (RO5)

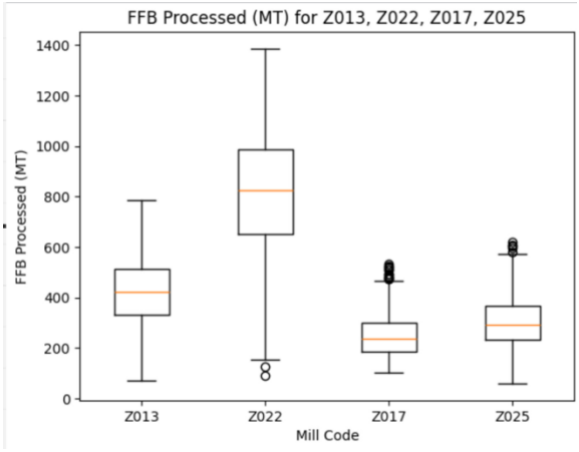
Based on the observations above, we cannot conclude that top-performing regions have a higher concentration of top-performing mills, as the expected and actual values do not align.

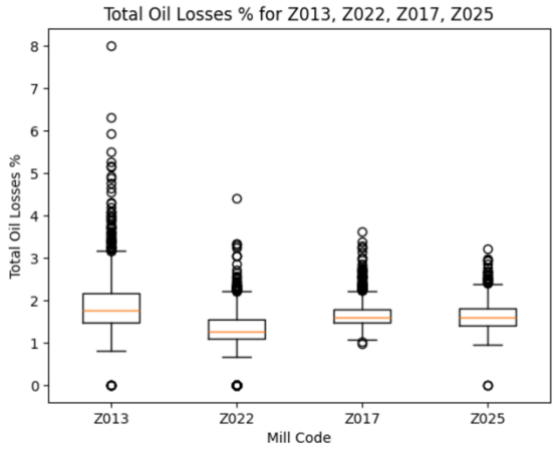
Region	No. of Mills	Top Performing Region?
RO1	6	True
RO2	6	True
RO3	6	True
RO4	5	False
RO5	5	False
RO6	4	False

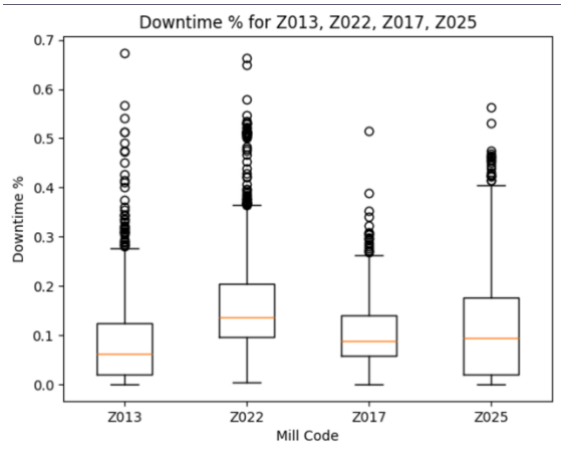
However, the insights above reveal that top-performing regions have a higher number of mills (~6 mills) compared to bottom-performing regions (~4.6 mills). This contradicts our initial hypothesis that top-performing mills correlate directly with top-performing regions. Instead, we can conclude that **the number of mills in a region significantly contributes to the region's higher overall OER%.**

Hypothesis #2: Do top-performing mills, on average, exhibit higher values for the features most strongly correlated with OER%?

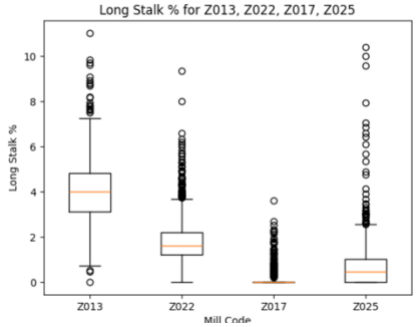
Initially, we discovered that Total Oil Losses (%), FFB Processed (MT), Downtime (%) have a positive correlation with the OER%.

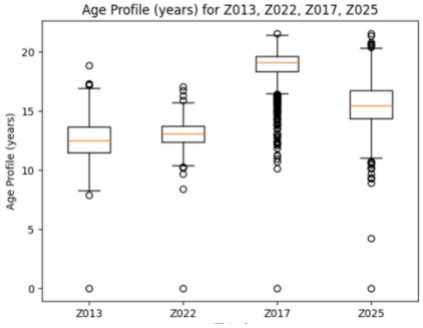






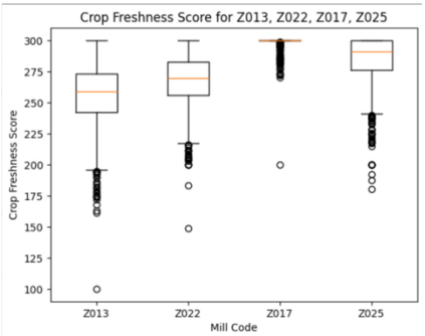
However, based on the box plots produced, we can only observe that median of **FFB Processed tend to be higher in top-performing mills**. In contrast, the medians for Downtime % and Total Oil Losses %, do not exhibit a clear or distinct pattern.

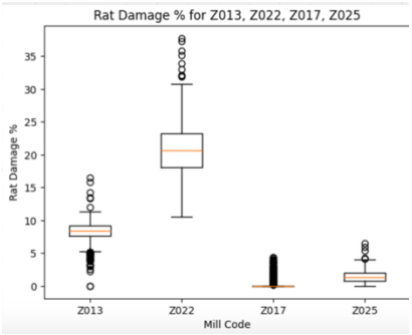




For other features, we observed distinct patterns with Long Stalk % and Age Profile (years). **Top-performing mills tend to have a higher median for Long Stalk % and a lower median for Age Profile (years) compared to bottom-performing mills.**

Although a higher Long Stalk % is typically associated with lower oil extraction, this may be attributed to the greater volume of FFB Processed in top-performing mills, which increases the likelihood of having a higher Long Stalk % in those plants. The lower median Age Profile in top-performing mills aligns with expectations, as older age profiles generally result in a lower Oil Extraction Rate (OER).





Furthermore, we observe that the **Crop Freshness Score** tends to be lower in top-performing mills compared to bottom-performing mills. This may be linked to the increased **Rat Damage %** observed in top-performing mills, which could, in turn, be attributed to the higher volume of FFB Processed in these mills. This suggests a potential trade-off between processing capacity and maintaining crop freshness. However, it is important to note that **correlation does not imply causation**, and further investigation is needed to establish any direct causal relationships.

Moving forward, we will build a Machine Learning model for the ZO22 mill to assess whether the feature importance in the ZO22 model aligns with that of the ZO13 model. Ultimately, the model developed for the ZO13 mill will guide us in identifying the parameters that need to be adjusted to improve the Oil Extraction Rate (OER).

SECTION 2: PREDICTIVE ANALYTICS

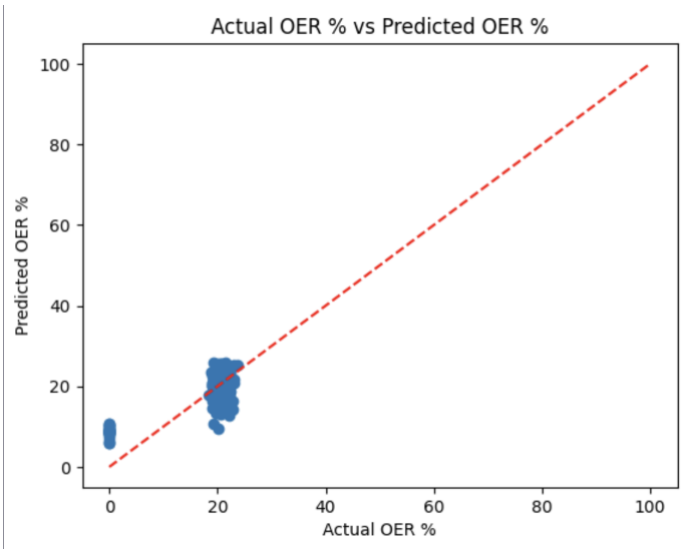
Our initial approach to building the model at the ZO22 begins by including all available features and using the **Ridge Regression** algorithm for training. Ridge Regression, also known as L2 regularization, is particularly effective when dealing with datasets that contain multicollinearity and reduces the influence of less relevant features, allowing the model to focus on the more impactful ones. We will retrain the model by removing the perfectly collinear features once we could get some understanding of the behaviour in the mill.

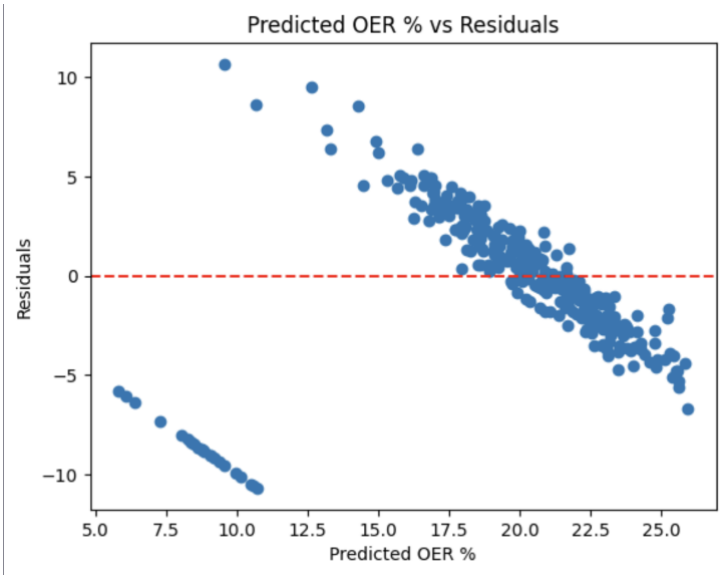
$$RSS_{L2} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^{P} B_j^2$$

However, we encountered several challenges while training the model. The model's performance metrics during training are summarized in the table below:

Accuracy Metric (Ridge)	Value
RMSE	3.71
R-squared	56.3%

Although the R-squared value is 56%, which indicates a moderate level of explanation for the variance in the target variable, the actual vs. predicted plot and the residuals plot reveal an unusual pattern.





The residuals (differences between predicted and actual values) should ideally be randomly scattered around the horizontal line at y = 0. In this plot, there appears to be a clear, **linear pattern** where residuals consistently increase or decrease across the predicted values. This indicates that the model is **failing to capture some systematic relationship** in the data, likely due to:

- Non-linear relationships between features and the target variable not being accounted for by the model.
- Model assumptions (e.g., linearity) not being met.

This suggests that the model may not be fully capturing the underlying relationships in the data or that certain assumptions of the model may not hold true. Further investigation was done to **address multicollinearity, presence of outliers or non-linearity**,

i. Addressing multicollinearity

As described in the previous section, some of the features in the model exhibit a clear linear relationship:

Relationship 1:

$$Seed\ A\ \% + Seed\ B\ \% + Other\ Seeds\ \% = 100\%$$

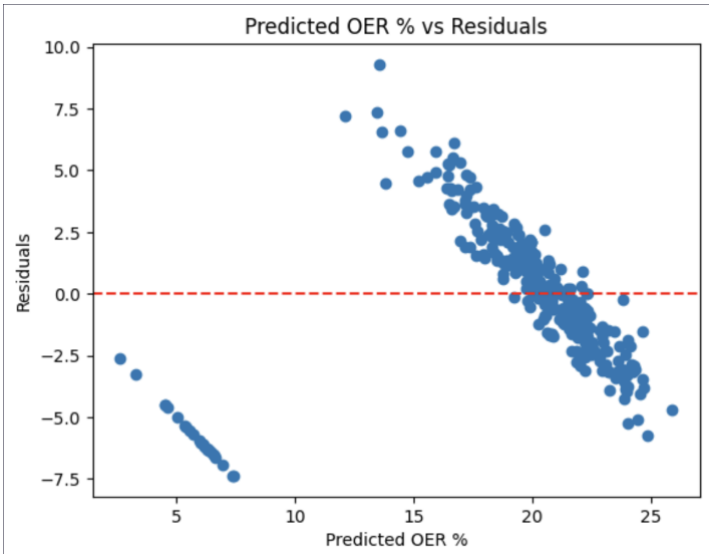
Relationship 2:

$$Coastal\ \% + Inland\ \% = 100\%$$

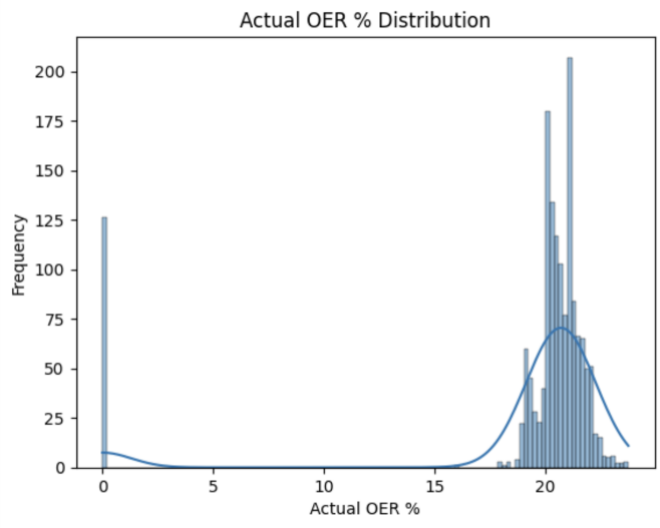
These linear dependencies suggest that **dimensionality reduction** can be applied by removing one feature from each relationship group. The features: Other Seeds% and Inland % have been removed and the model metrics are tabulated below:

Accuracy Metric (Ridge)	Value
RMSE	2.93
R-squared	72.7%

Although we observe an improvement in the accuracy metrics, the underlying issue persists: the residuals are not randomly scattered but instead exhibit a linear pattern. This indicates that the model is still failing to fully capture the true relationships within the data, despite the observed increase in performance. **This suggests that the improvements in metrics may not necessarily reflect a more accurate or reliable model.**

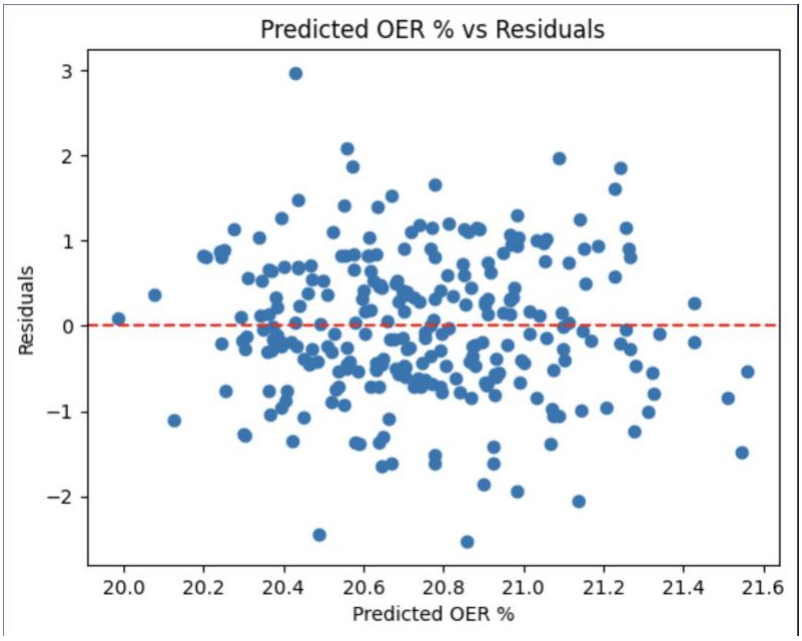


ii. Addressing outliers



From the histogram of the target variable shown above, the OER% distribution exhibits right-skewed data, with the majority of values concentrated around 20%. To effectively model a linear relationship, the outliers at 0% must be addressed. By applying the **Interquartile Range (IQR)** method to identify and remove outliers from the target variable, we obtained the following model metrics.

Accuracy Metric (Ridge)	Value
RMSE	0.83
R-squared	0.08%

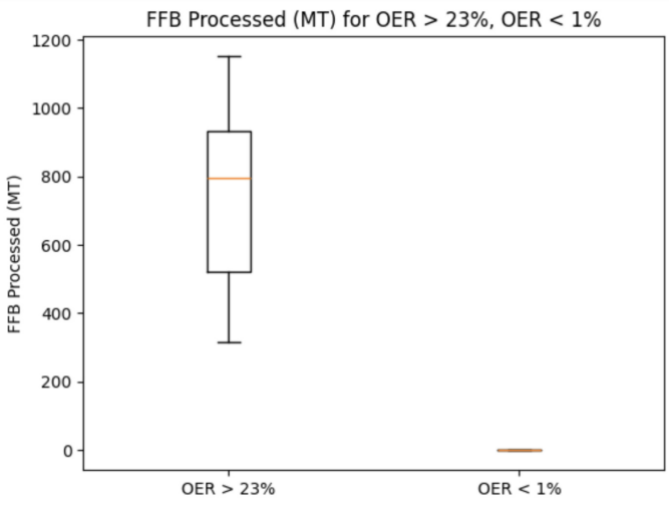


Although the residuals plot now appears satisfactory, with the residuals randomly scattered around $y = 0$ and showing no linear pattern, there was a significant performance drop in the R-squared value, decreasing from 56.2% to 0.08%. This suggests that **while the removal of outliers improved the residual distribution, it may have removed critical information that the model relied on**, resulting in a drastic reduction in predictive performance.

iii. Addressing non-linearity

The first step is to **determine whether the zeros in the dataset occur naturally and represent genuine values** or if they are artifacts of data collection or processing errors. This involves investigating the context of the data, reviewing the data collection methods, and verifying whether zeros are expected under specific conditions (e.g., periods of no production, equipment downtime, or data entry errors).

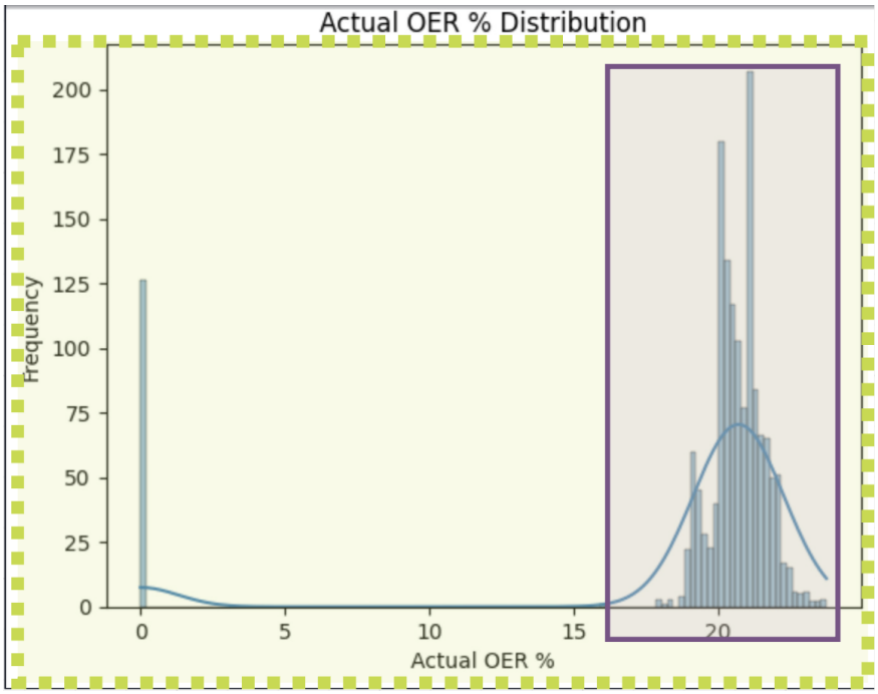
We analyzed the data at extremely high OER% values and extremely low OER% values. What we discovered was a distinct pattern in the FFB Processed in this regime.



The FFB Processed variable behaves like an “on-off” switch, where at high OER% values, FFB Processed has a non-zero value, and conversely, at low OER% values, FFB Processed is zero. According to the client,

higher FFB Processed corresponds to a higher utilization rate, which in turn leads to a higher OER%. This suggests that **the presence of zeros in the data is genuine** and reflects real operational conditions (periods of no Oil Extraction process), rather than being the result of significant data processing errors.

What we are encountering now is a **zero-inflated regression problem** and a simple linear regression will not work well in this situation



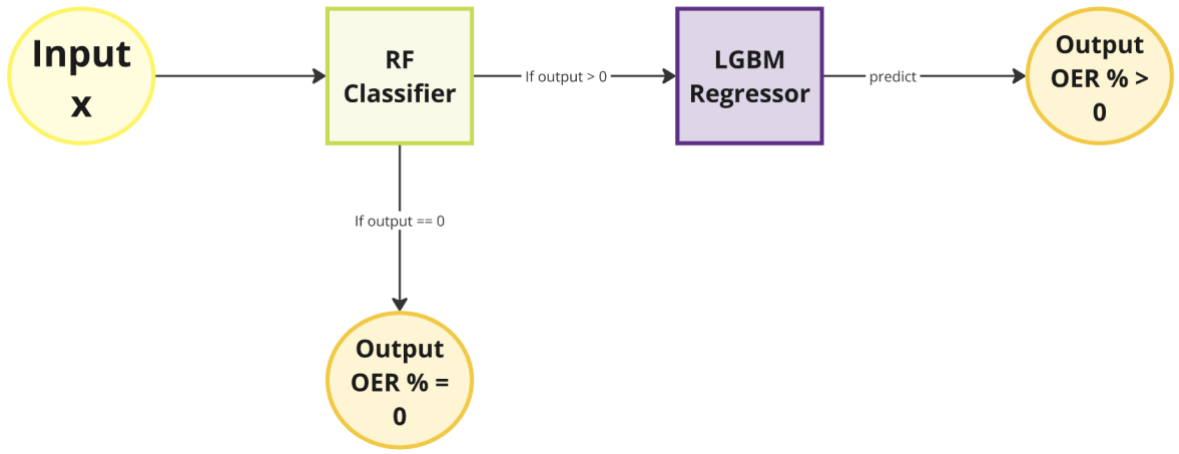
To address this, will develop a **meta-model** i.e. a general model that consists of many models. The model consists of 2 steps:

1. Given the input, x will the output of the first model be zero or non-zero. If the output is zero, the OER% output is zero (**classification problem: green dashed green line** in the chart)

2. If the output of the first model is non-zero, predict the output of OER% based on the training data (**regression problem: solid purple** line in the chart).

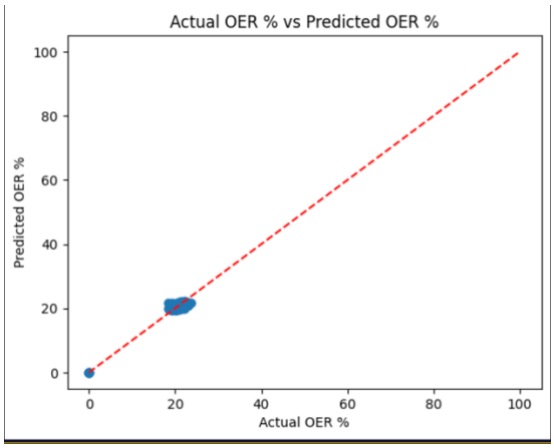
In the first step, a **Random Forest Classifier (sk-learn)** will be used to classify observations into two categories: zeros and non-zeros. Random Forest is well-suited for this task due to its robustness to imbalanced datasets and its ability to handle non-linear relationships, which are often present in zero-inflated data. This allows us to differentiate between the two groups.

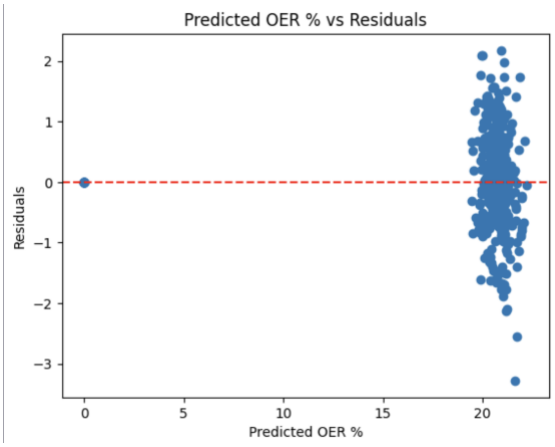
In the second step, for the non-zero observations, a **Gradient Boosting Regressor (LGBM)** will be employed to predict the continuous target values. Gradient Boosting is chosen for its ability to capture complex patterns and interactions in the data while minimizing prediction errors through iterative improvement.



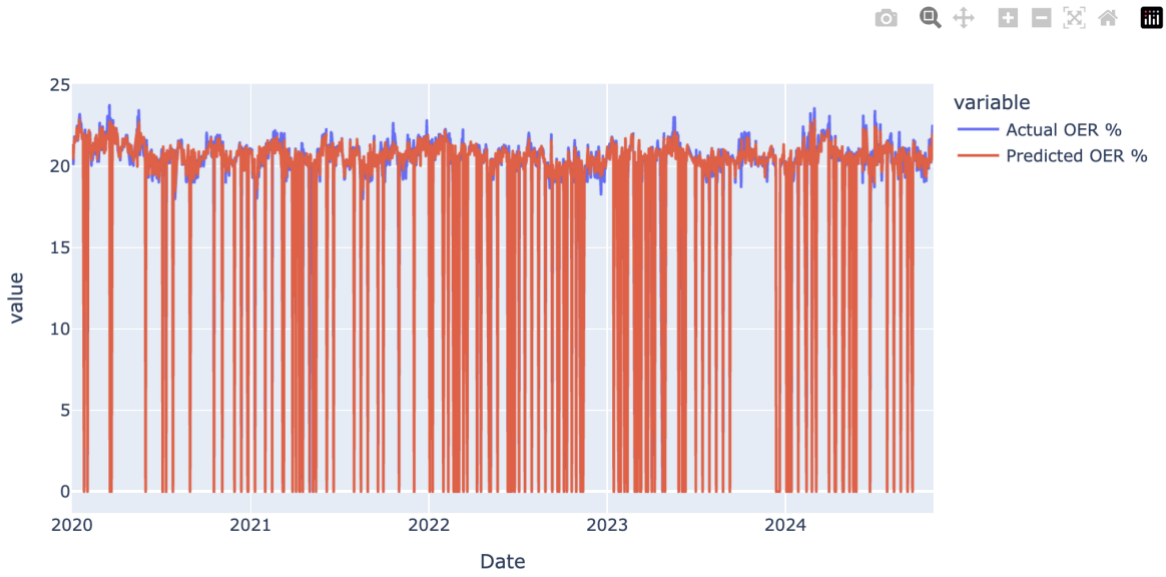
Using the the **scikit-lego library** to train the machine learning model, we obtained the following results of the trained machine learning model using a 70:30 training-testing split and with date range from 2020-2024:

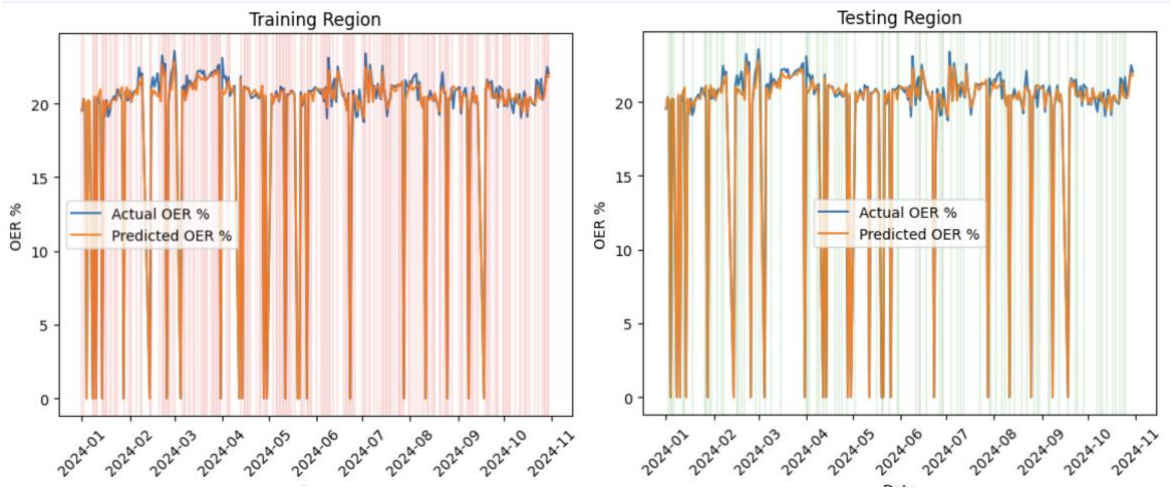
Accuracy Metric (meta-model)	Value
RMSE	0.77
R-squared	97.97%





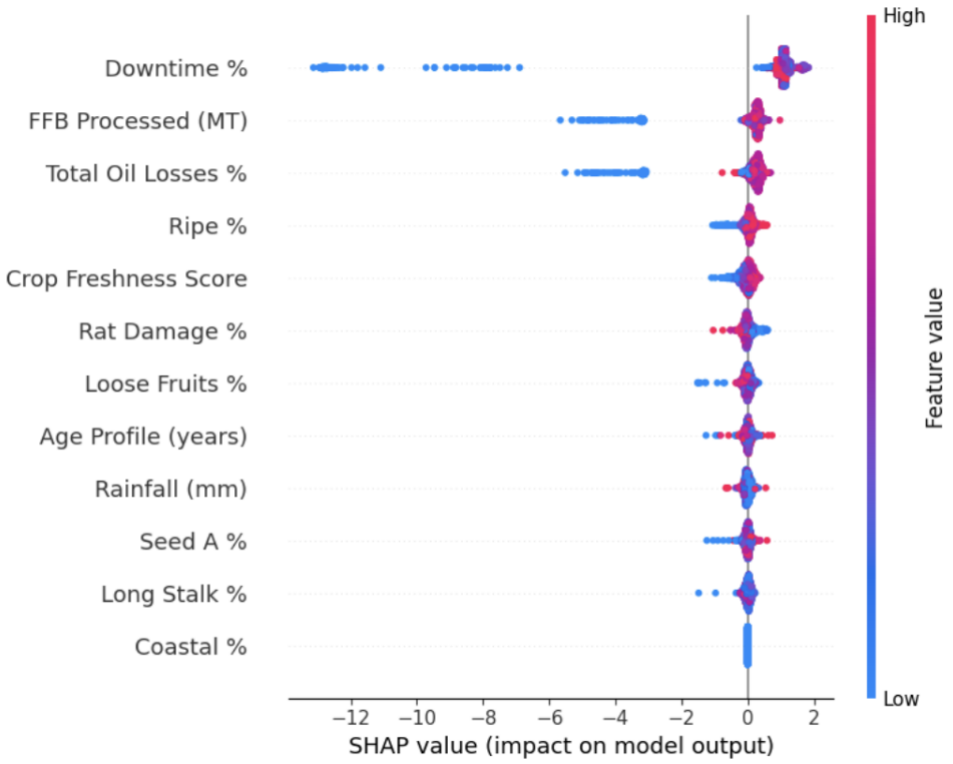
We successfully reduced the RMSE from 0.83 in the previous Ridge regression model to 0.77, while significantly improving the R-squared value from 0.08% to 97.97%. Additionally, the residuals show marked improvement, as they are now randomly scattered around the horizontal line at $y = 0$ without any apparent linear pattern, unlike the structured pattern observed with the Ridge model. These improvements indicate that the **meta-model approach better captures the relationships in the zero-inflated data and produces more reliable predictions.**





iv. Interpreting the Model Features

Tree-based models (Random Forest Classifier and Gradient Boosting Regressor) do not produce a straightforward equation like $y = mx + c$, making their predictions less interpretable in traditional terms. To gain insights into the model's behavior, we use **a SHAP plot, which visually explains how each feature impacts the predictions.**

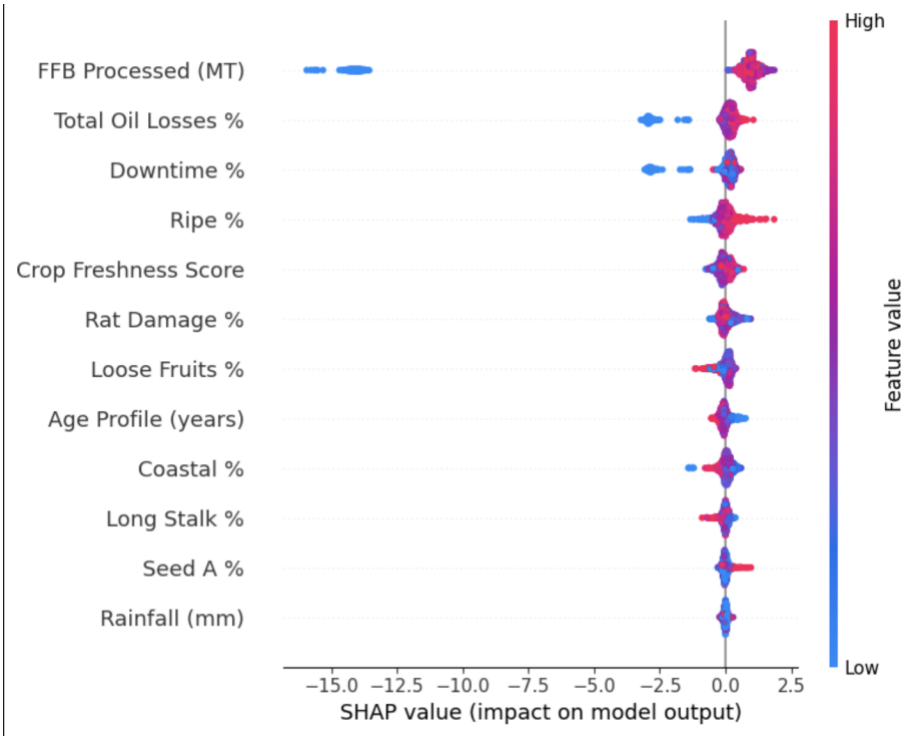


On the SHAP plot, the x-axis represents the SHAP values, indicating the influence of each feature. **Positive SHAP values drive the predictions higher (towards higher OER%), while negative SHAP values lower the predictions (towards lower OER%).** For instance, Downtime % exhibits a strong linear relationship, with higher downtime significantly reducing OER%. In contrast, Crop Freshness Score shows that higher freshness scores (indicated by red) contribute positively, boosting the OER%. This approach allows us to interpret the model's predictions and identify key drivers of performance.

From this insight, we have confirmed that FFB Processed plays a crucial role in determining whether the model will yield a prediction. This explains why FFB Processed appears at the top of the SHAP plot, as it has the most significant impact on the model's output compared to other features.

Section 3: Prescriptive Analytics

Using the same modeling approach applied in ZO22, we trained a similar model for the plant in ZO13. We observed that the model in ZO13 behaves similarly to the one in ZO22, with **FFB Processed**, **Total Oil Losses**, and **Downtime** consistently appearing as the most impactful features at the top of the SHAP plot. This consistency reinforces the importance of these features across different plants in predicting OER%.



We will now attempt to use the model in ZO13 and guide us on making the most optimized decision to maximize our OER%. An optimization problem has the general form of:

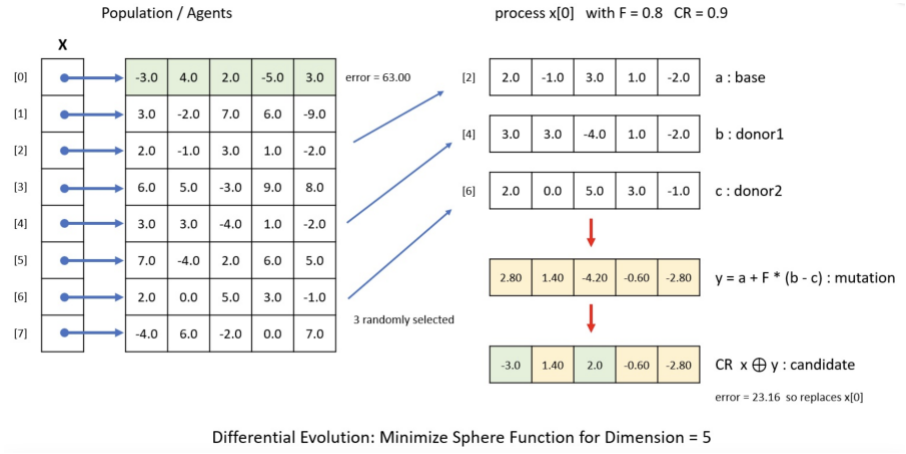
$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_j(x) = 0, \quad j = 1, \ldots, p \end{aligned}$$

Where:

- f(x) is the objective function to be minimized/maximized over the variable vector, x *(in this case, our **machine learning model**)*
- g(x) is the inequality constraints
- h(x) is the equality constraints

By integrating the model's outputs with optimization techniques, we aim to determine the combination of feature values (e.g., FFB Processed, Downtime %, and Total Oil Losses) that will yield the highest possible OER%, while adhering to the operational constraints of the plant. This approach bridges predictive modeling with actionable decision-making to enhance plant performance.

The predictive model is integrated with the **Differential Evolution Algorithm**, a metaheuristic optimization technique, to determine the variable combinations that maximize the output, specifically the OER%. This algorithm iteratively explores the solution space by simulating the process of natural evolution, such as mutation, crossover, and selection, to identify the **optimal set of variables**



We conducted an experiment using the data from **1st October 2024** with the optimization algorithm, incorporating specific thresholds for each variable to ensure operational feasibility. These thresholds, as detailed in the table below, were designed to define the acceptable ranges for each variable

Crop Freshness Score	Acceptable threshold : 280 and above
Ripe %	Acceptable threshold : 90% and above
Long Stalk %	Acceptable threshold : 5% and below
Rat Damage %	Acceptable threshold: 5% and below
Loose Fruits %	Acceptable threshold: 8% and above
Rainfall (mm)	Optimal rainfall level is 150mm/month (~5mm/day)
Age Profile (years)	N/A
Total Oil Loss (OL)	Acceptable threshold: 1.40 and below
Downtime %	Acceptable threshold: 6% and below
FFB Processed MT	N/A
Seed Type	N/A
Topography	N/A

As a result, we obtained **an improvement of OER% by 2.19%** by increasing/decreasing the following values:

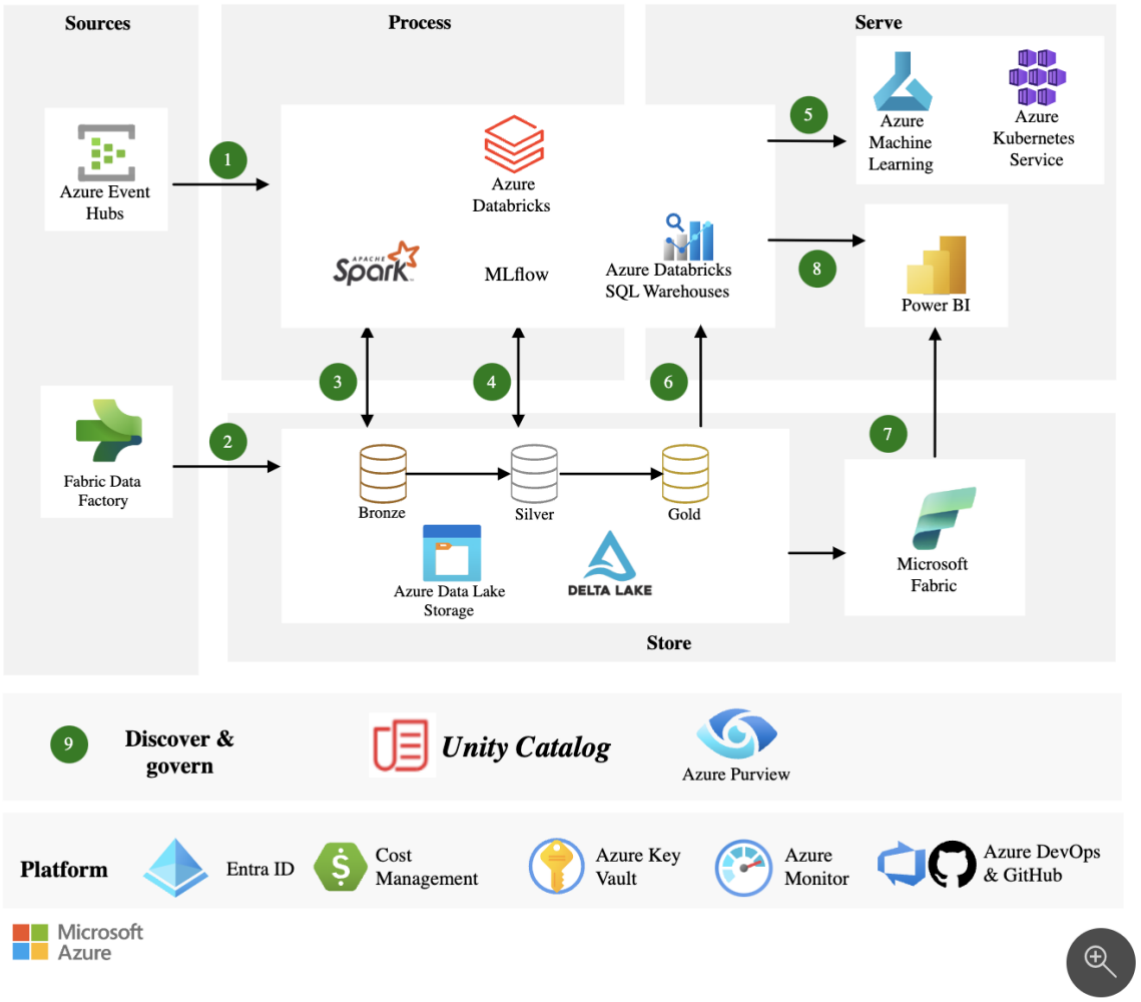
Date	Column1	Crop Freshness Score	Ripe %	Long Stalk %	Rat Damage %	Loose Fruits %	Rainfall (mm)	Age Profile (years	Total Oil Losses %	Downtime %	FFB Processed (MT)	Seed A %	Coastal %	Actual OER %
01/10/2024	Original	279.131268	84.190405	3.4003	9.355905	2.379274	3.502374	13.91828	1.920196	0.061842	415.135	0.442039	63.688545	20.390776
	Optimized	280.27	92.02	1.29	0.49	13.85	0.23	13.92	1.92	0.06	353.86	0.44	63	22.59
	Delta	1.138732	7.829595	-2.1103	-8.865905	11.470726	-3.272374	0.00172	-0.000196	-0.001842	-61.275	-0.002039	-0.688545	2.199224

Variable	Delta
Crop Freshness Score	+ 1.14
Ripe %	+ 7.83
Long Stalk %	-2.11
Rat Damage %	-8.87
Loose Fruits %	+11.47
Rainfall (mm)	-3.27
Age	0
Total Oil Losses %	0
Downtime %	0
FFB Processed	-61.28
Seed A %	0
Coastal %	0

CONCLUSION

In conclusion, we have successfully developed both **predictive and prescriptive model**s to optimize the Oil Extraction Rate (OER%). The predictive model provides insights into key factors influencing OER%, while the prescriptive model utilizes optimization techniques to recommend actionable decisions for maximizing efficiency.

The next critical step involves setting up a robust **deployment and monitoring** architecture to operationalize these models. This ensures that the models are seamlessly integrated into the production environment, delivering real-time predictions and optimization recommendations. Additionally, monitoring will allow us to track model performance, detect potential issues such as data drift, and facilitate regular updates to maintain accuracy and relevance. An example of such a deployment and monitoring architecture is shown below, illustrating how the models will be utilized effectively in practice.



References

https://builtin.com/data-science/metaheuristic-optimization-python

https://medium.com/towards-data-science/zero-inflated-regression-c7dfc656d8af

https://www.mpoc.org.my/challenges-faced-by-malaysian-palm-oil-the-way-forward/

https://poeb.mpob.gov.my/case-study-on-oil-extraction-rate-and-its-issues/

https://www.ibm.com/think/topics/ridge-regression

https://visualstudiomagazine.com/Articles/2021/09/07/differential-evolution-optimization.aspx?Page=1

https://www.amazon.com/Regression-Analysis-Intuitive-Interpreting-Linear/dp/1735431184

https://www.amazon.com/StatQuest-Illustrated-Guide-Machine-Learning/dp/B0BLM4TLPY

https://learn.microsoft.com/en-us/azure/architecture/solution-ideas/articles/azure-databricks-modern-analytics-architecture

https://www.sustainablepalmoilchoice.eu/how-is-palm-oil-produced/