**Enhancing Beverage Production Process Efficiency: A Machine Learning Approach**

Michelle M. Moran

A Thesis Submitted in Partial Fulfilment

of the requirements for the

Degree of

Master of Science in Data Analytics

Diagram

Description automatically generated with medium confidence

September 2023

Supervisor: Kislay Raj

# CCT College Dublin

Assessment Cover Page

To be provided separately as a word doc for students to include with every submission.

|  |  |
| --- | --- |
| **Module Title:** | Capstone Project |
| **Assessment Title:** | Enhancing Beverage Production Process Efficiency: A Machine Learning Approach |
| **Supervisor Name:** | Kislay Raj |
| **Student Full Name:** | Michelle Moran |
| **Student Number:** | SBA22220 |
| **Assessment Due Date:** | Friday, 22 September 2023 |
| **Date of Submission:** | TBC |

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

**Acknowledgement and Dedication**

# Table of Contents

# Table of Figures

# Table of Tables

# Abstract

# Chapter 1: Introduction

**Research Background and Context:**

The proposed approach holds significant potential, particularly for traditional industries that have not yet capitalized on the advancements of Industry 4.0. These industries are now beginning to explore the possibilities of using data analytics and machine learning to optimize their production processes.

Modern Beverage production industries have state of the art equipment to help ensure the consistent delivery of high-quality beverage products for their customers. Over the years automation, computerization and data analytics has transformed manufacturing, enabling precision in ingredient addition, quality control and efficiency. As the beverage industry advances, the incorporation of smart manufacturing technologies, sustainable practices, and consumer preferences will shape its future direction, ensuring its ongoing significance in the global market.

Smart manufacturing

Briefly introduce the field of study, particularly the domain of production tank operations or related industrial processes.

**Explain the relevance and importance of efficient production tank operations in various industries.**

An efficient production operation results in better quality product, lower waste levels, lower operating cost and a better decision making.

**Highlight the challenges and complexities** associated with managing and optimizing production processes.



**Problem Statement:**

Clearly state the specific problem or gap in the current practices that your research aims to address.

Problem Statement:

In the beverage manufacturing industry, specifically the production of non-alcoholic beverages, there exists a challenge in achieving optimal efficiency and quality during the production process. Traditional methods often lack the ability to harness the potential of Industry 4.0 technologies, such as data analytics and machine learning. As a result, many production processes still rely on manual adjustments and outdated practices. This situation leads to inconsistencies in mixing and deaeration times, which directly impact product quality and production efficiency.

The lack of an integrated and data-driven approach hinders the industry's ability to adapt to changing consumer preferences, maintain competitive advantages, and reduce operational costs. The industry faces a need to explore new strategies that leverage smart manufacturing techniques to optimize the production tank operations for beverage base preparation. Therefore, the research seeks to address this gap by developing and testing machine learning algorithms that can predict optimal mixing and deaeration times. The goal is to provide a comprehensive solution for traditional industries to enhance their production processes and remain competitive in an ever-evolving market landscape.

**Emphasize the significance of this problem** in terms of its impact on production efficiency, cost-effectiveness, or quality control.

**Research Objectives:**

Clearly state the main objectives of your research. For example, these could include optimizing phase durations, reducing phase overruns, and improving resource utilization in production tanks.

**Discuss the research questions that will guide your investigation.**

**Motivation for the Study:**

**Explain why you chose this topic for your master’s.**

**thesis**. Did personal experience, observations, or industry trends motivate your interest in this area?

**Highlight any potential real-world applications** of the research outcomes.

**Scope and Limitations:**

**Define the scope of your research**. Clarify the specific aspects of production tank operations that you will focus on.

**Identify any limitations or constraints** that may affect the scope or generalizability of your findings.

**Methodology Overview:**

**Provide a brief overview** of the research methodology you plan to use, including the use of machine learning algorithms.

**Mention any data sources**, tools, or frameworks you'll employ in your analysis.

**Expected Contribution:**

**Discuss the potential contributions** your research will make to the field. This could involve improving existing processes, enhancing decision-making, or providing insights into tank-material pairing optimization.

**Highlight the novelty of your approach**, especially the application of machine learning techniques.

**Significance and Benefits:**

**Highlight the potential benefits of optimizing production tank operations using machine learning**.

**Discuss how your research could lead to increased efficiency**, reduced costs, improved resource utilization, or enhanced quality control in industrial processes.

Some of the key benefits of analytics and visualization in the IIoT ecosystem include:

· Improved decision-making: By providing real-time insights into the performance of the physical asset, analytics and visualization tools enable stakeholders to make data-driven decisions.

· Optimization of asset performance: By analyzing data from the digital twin model and using prescriptive analytics, optimization techniques can be used to improve the performance of the physical asset.

· Predictive maintenance: By analyzing data from the digital twin model using predictive analytics, machine learning algorithms can be used to forecast when maintenance is required, reducing downtime and increasing productivity.

· Cost reduction: By optimizing the performance of the physical asset, analytics and visualization tools can help reduce costs associated with maintenance, repairs, and downtime.

**Thesis Structure:**

Provide a roadmap for the structure of your thesis, briefly outlining the content of each chapter or section.

**Summary:**

**Summarize the key points covered in the introduction.**

**Reiterate the importance of the research topic and its potential impact on industrial practices.**

**Data Acquisition**

The Production batch data in the organisation is acquired from software called FactoryTalk® Batch. It provides a flexible batch control system. It allows you to specify procedures or recipes and enforce their execution in production.

The FactoryTalk Batch Server operates FactoryTalk® Batch software and coordinates the following functions:

* Creating a Batch: Transforms the configured recipe into an executable.
* working recipe.
* Executing a Recipe: Communicates with the process-connected.
* devices to execute phases.
* Arbitrating Equipment: Allocates resources based on recipe and
* operator requirements.
* Collecting Data: Gathers and stores production information for
* reporting and archiving.
* Performing Client Communications: Transfers data between the
* process-connected devices (PCDs), operator displays, Human Machine
* Interfaces (HMIs), databases, and various other software packages.

This FactoryTalk® Batch software uploads batch data from each of the production activity known as phases per batch to a Microsoft SQL database which is managed by SQL Server Management Studios. One of the components of this software is the FactoryTalk Event Archiver which translate the FactoryTalk batch tab-delimited ASCII electronic batch record files to a user-specified file type. This organisation uses Microsoft SQL database which is managed by SQL Server Management Studios to maintain files for each batch created.

Microsoft SQL Server Management Studio (SSMS) is a software application developed by Microsoft that is used for configuring, managing, and administering all components within Microsoft SQL Server. A central feature of SSMS is the Object Explorer, which allows the user to browse, select, and act upon any of the objects within the server.

# Chapter 2: Research Design

## Primary Data

## Problem Identification and Clarification

**Thesis Title** - Enhancing Beverage Production Process Efficiency: A Machine Learning Approach

## Research Objectives

**Research Objective 1:** Exploration and Quantification of Production Downtimes

Explore the beverage batch manufacturing process to identify and quantify instances of mixing and deaeration downtimes across different production tanks. This investigation seeks to provide a foundational understanding of the existing inefficiencies or gaps in the current production regime.

**Research Objective 2:** Efficiency-driven Machine Learning Analysis of Batch Data

Employ machine learning techniques to efficiently analyze the batch production data. The aim is to underscore the importance of such analysis and highlight potential areas of optimization, particularly focusing on minimizing production steps and shortening the overall process. This objective further seeks to provide evidence-based insights into why investment in time and resources for such analysis can be beneficial for the overall production strategy.

**Research Objective 3:** Predictive Modelling of Production Downtimes for potential Enhanced Scheduling

Develop and validate machine learning models designed to predict downtimes associated with mixing and deaeration across different production tanks. Leveraging these predictions, the objective is to propose optimized scheduling processes for batches, ultimately aiming to minimize process downtimes and improve overall production efficiency.

## Validity Type

## Ethical Considerations

# Chapter 3: Literature Review

**Introduction**

**Themes of the Literature Review.**

**4.1 Introduction**

In this section we will review existing literature on the application of machine learning methods for optimizing production processes by looking at production data (Big Data). This will give a solid foundation for this study and hopefully identify areas for further research and application in other production areas.

The world of manufacturing has the potential to utilize machine learning to extract patterns from existing datasets, which can serve as a basis for predicting future system behaviour, (Alpaydin et al 2010, Nilsson et al , 2005). Alpaydin also claims “Storage data only becomes useful when it is assessed and translated into knowledge that we may use, for example to develop predictions”. This is true of this research. There are huge amounts of data collected and stored in the manufacturing process. By assessing and examining it through machine learning models, it may aid in the future decision making of the process owners.

We will use themes structurally for this literature review to present the existing research in a structured manner to give a clear understanding of the state of knowledge and supporting the development of the research objectives. Through the literature review I hope to highlight the importance of data and the different machine learning models that have been applied in the manufacturing process situations.

## 

**4.2 Big -Data - Machine Learning in Production Process Optimization:**

According to Kovalev et al, 2019, data is at the head of the process of digital transformation part of the industry 4.0 revolution. Digital transformation is the approach used by industry for the optimization of production. Data must be collected, stored, aggregates and transferred to various levels. Data has become described as Big DATA due to the huge exponential volume, e.g., terabytes, that is generated by various systems in production.

Big Data is worthless on its own, the manufacturing industry requires efficient processes to be able to derive valuable information from it. The following are processes of examining big-data, /complex datasets to uncover hidden patterns and correlations. These processes can include data mining, machine learning, natural language processing. High Quality data and large data sizes can increase the accuracy of machine learning models. (Kang et al 2020),

The biggest advantage to industry is this increased understanding gained from the processed data leads to benefits in production operations such as costs reduction and time efficiencies. In a 2016 report by McKinley, it states “Big data’s potential just keeps growing. Taking full advantage means companies must incorporate analytics such as machine learning into their strategic vision and use it to make better, faster decisions. Manufacturing companies worldwide are investing huge amounts of money in developing these big-data driven techniques.

According to McKinsey, Machine Learning accounts for between $3.5 t0 $5.8 trillion in annual value. Machine learning can be applied successfully in operations, productions, and post-production activities in manufacturing. (H., XIA et al, 2022.)

**4.3 Traditional Process Optimisation Methods**

Prior to machine learning, traditional methods used for improving production efficiency included manual inspection, statistical tools, expert systems, and mathematical modelling, (Wang et al, 2018) However as earlier mentioned, with the increased volumes of data, Big- Data being accumulated, these traditional methods are greatly impacted. An example of statistical tools traditionally used are statistical process methods such as control charts used to detect defects. These control measures are applicability and simplicity but again not able to keep up with the increasing complexity of production and volume of data being gathered as a result. (Ismail et al 2022).

Arif et al ,2023, states most existing quality monitoring models only look at one manufacturing state and the data gathered is not processed until after the product is made or manufacturing process is over. This has a negative effect on resources and time and this production performance.

Three areas that make traditional methods obsolete are:

* Information and communication technologies – mode of production has changed, large-scale tasks, operating performances and environments are more complex, (Wang et al, 2018a)
* Increased demand for real time dynamic self-adaptive and precise production management (Arashpour et al, 2018, Lamon et al, 2010.)
* The completion of various kinds of information systems deployed in manufacturing enterprises. E.g., CAPP, computer-aided process planning, (Papananias et al, 2019.)

Machine learning cannot completely replace traditional methods. One can learn from the other while machine learning can aid the identification and modification of the parameters of the traditional methods to improve processes (H., XIA et al, 2022.)

Also, another consideration is the time and financial cost of developing and programming models of machine learning, this may be out of reach of small-scale manufacturers, however for large scale manufacturers and their various production lines, it is the varying complexities and lack of efficiencies these models can bring. It can be challenging to implement AI – machine learning in an entire organisation with existing processes and systems if the company lacks robust technology infrastructure and collected data. Its important that there is clear strategy stemming from top management to achieve goals otherwise AI will fail (Kang et al 2020, Heio et al, 2021).

But there is an increasing importance to improve / optimize the effectiveness and efficiency of decision making in a production process, by extracting/mining the production data both online and offline using more efficient techniques. (H., XIA et al, 2022.

**4.4 Machine Learning**

**“*Machine Learning algorithms are the lenses through which we view and interpret the world of data* “– Bernard Marr**

This quotation by Bernard Marr (a business consultant who specialises in helping organizations leverage data and analytics to drive business success) highlights the transformative role of machine learning algorithms in our understanding and interpretation of data. It suggests that machine learning algorithm serve as a lens that enables us to gain insights and make sense of the vast amount of data available to us.

Machine Learning is a part of the world of Artificial Intelligence that enables production to analyse and interpret the vast amount of data generated. Simply It models the complex relationship between input and output data, (Wang et al, 2018b). Machine Learning model systems learn from data, identify trends or patterns from data, make decisions based on structure feedback and then perform tasks on their own, with continued improved performance and problem-solving skills without human intervention (Helo et al, 2019)

Using Machine learning algorithmsfor the improvement in manufacturing process and quality optimization can be divided into 4 stages, (Koksal et al., 2011):

* Description of product
* Process quality by identifying and ranking the most significant variables and factors related to the quality
* Prediction of the quality
* Classification of quality
* Optimization of the manufacturing parameters

Rai et al 2021 states that Machine learning techniques can augment the manufacturers effort to select the optimal set of parameters related to a given manufacturing process, hence enabling process improvement and optimisation. The type of machine learning to use is also governed by the type of data.

**4.5 Machine Learning Methodology**

The common methodology in using Machine learning algorithmsfor the improvement in manufacturing process and quality optimization are as follows, (Ismail et al, 2021, Kant et al.,2015, Aksa et all.,2021, Ahmad et al.,2018, Kulkarni et al.,2020)

1. Data Collection – collect data on the performance of the chosen parameter under different operation conditions, e.g., flow rates, temperatures etc
2. Data Prepossessing – removing outliers and missing values ensure the data is clean and accurate
3. Feature selection – selecting the relevant features that can help predict the performance of deaeration
4. Model development – develop the predictive model and evaluate it using metrics such as accuracy, precision, and recall
5. Model Optimization – to improve its performance using hyperparameters or ensemble learning techniques
6. Model Deployment – deploying the model in a real-world setting and validate its performance

Lieber et al, 2013, states that even though the interest manufacturing optimization surveys by KDnuggets annually show that production related projects in manufacturing are still underrepresented in comparison to more well-established fields such as banking or fraud detection.

Machine learning can be divided up into 3 unsupervised, supervised and reinforcement types and 5 common techniques:

Supervised – a function is derived between an input and output from a set of labelled training data. There is more human interaction, and it requires more data processing for feature selection

Unsupervised – The relationship between the input variables is not known. There is no labelled data. The output is more a pattern of input variables, or a cluster built on input data

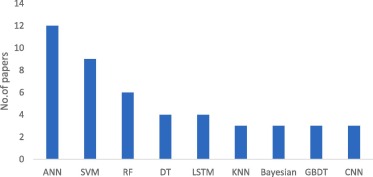
Reinforcement learning – using feedback mechanisms to reward positive action and punish the negative action.

There are 5 common machine learning techniques –

* Regression – where the input value is numerical continuous variable. These algorithms are used to optimize the coefficients of each independent variable to achieve a minimum error in the prediction.
* Classification – mapping input features to one of the discrete output variables.
* Clustering – dividing data points into relevant groups
* Data reduction- reducing the number of features. Use mainly with regression and classification
* Anomaly Detection – used in unsupervised methods. Like clustering, grouping unknowns together

According to Kang el al, 2020 a literature review on machine learning applications in the manufacturing industry, the majority (80%) of papers reviewed were concerned with supervised learning, due to the abundance of production data available. Regression is the main task applied for quality optimization problems while classification and anomaly detection are mainly applied for product failure detection.

Artificial Neural Networks is the most frequently applied machine learning algorithm applied in production lines. See fig 3 for details on all machine learning algorithms that were review in terms of popularity. This popularity of ANN is due to the complexity of the production lines and again the available of large amounts of data.



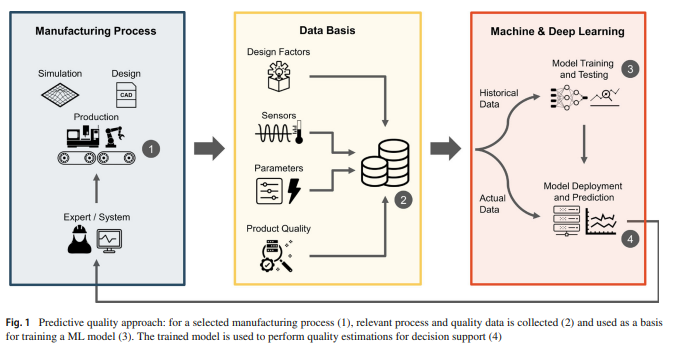
***Fig 1 Distribution of machine learning algorithms***

ML has successfully been applied in industry, these are summarized below, (Diez-Olivan et al ,2019, Aksa et al, 2021, Menezes et al, 2019):

* **Descriptive** – accounting and analysis of historical data. e.g Fault detection and diagnosis
* **Predictive** – Considers near past data to predict coming future trends, biases tendencies and behaviours through causation and correlation.
* **Prescriptive** – finds or prescribes the best mode route manner or moves to operate based on given data (output) and models (inputs). help make decisions on what to do and how to minimize failure impact
* **Preventative** – determine the potential of failures happening
* **Detective analytics** – makes diagnostics of collected data to eliminate and rectify inappropriate values used in predictive analytics
* **Cognitive analytics** – automated predictions, prescriptions and detections for smarter decisions over time

**4.6 Process Optimization by Machine learning**

Through this literature review, there was no direct research on the process optimization by machine learning on the deaeration process in beverage manufacturing. However, there was plenty of research reviews on process optimization in the manufacturing industry using various machine learning (Monostori et al, 1996, Md et al , 2022, Paturi et al, 2021). Fig 1 below gives an illustrated diagram of the predictive approach for a selected manufacturing process.(Helo et al, 2022)



***Fig 2 Predictive quality approach for a selected manufacturing process***

Tercan et al collated the occurrences of machine learning models that were used as baselines for predictive quality in research publications between the 2020 and 2021, Fig 1.0. Many of the research publications compared several models where he described the model that performed the best as the prime model and the others as the baseline models. The prime models included MLP – multilayer perceptron and neural networks. There is a wide range of models represented in fig 2. Most models are regression type. The following are predictive process studies using regression models as machine learning.



***Fig 3 Examples of baseline models used in Predictive quality processes (***Tercan et al)

**4.7 Regression models as machine learning in process optimization**

Most manufacturing processes involve multistage steps to produce accurate products. Papananias et al, 2019 developed a Bayesian linear regression model to estimate part quality of and associated uncertainties given in process monitoring data. The predicted results compared well with the experimental measurements and further a neural network model was developed which also showed similar results. It highlighted challenges to the model of evolution of more complex products, big data, and manufacturing intelligence. It also gave the suggestion to look at self-organising maps (SOM, Lieber et al, 2013) ,or principal component analysis for the consideration of process variables such as high sampling rates or large measurement uncertainties to reduce them.

A research paper concerned with engine oil aeration process step looked at a gaussian regression model used to correlate the identified features to measure oil aeration. The results were successful in the prediction of oil aeration to an uncertainty of +/-0.02 from the measured oil aeration values. The model was trained using previous oil pressure data. The results also highlighted that importance of looking at sampling measure as the cases used showed overfitting. This was calculated from using the metrics of RMSE, root mean square deviation. (Kulkarni et al 2021)

Under Predictive tool performances, logistic regression as a machine learning model has been used for the reliability estimation for cutting tools using the variable vibration signals. It is using correlation analysis approach to estimate the reliability and failure time of the cutting tools. It concluded that by further looking at the mechanical knowledge and probability density functions of other variables could further enhance the model’s predictions on machine tool performances. (Chen et al, 2010)

IBM Research has developed a process and system regression optimization service for optimising set points for process controls. The Process and System Regression Optimization service consists of three main components -- (1) Regression component and (2) Single Process Optimization component, and (3) System-wide Optimization component. The three components are used to model the problem at hand with the regression component being used to train one or more regression models based on historical data from the process or asset, and one of the optimization components used in combination with the trained model(s) to optimize the control points.

Datasets are trained on the API and then it uses various models of regression to make predictions requested.

It uses Mean Squared error and R2 as metrics to compare different regression algorithms.

Phan et al, 2021, used this IBM AutoAI toolkit for automated machine learning to search for the right learning algorithm and optimize its hyperparameters. They were able to restrict their models to decision tree, multivariate adaptive regression splines. Using cross validation to pick the best performance model.

Hassani et al, looked at the efficiency of the equipment rather than the product process in manufacturing, the methodology included a case study where the data was based on results from 2 machines. The applied methodology included data preparation, exploratory analysis of the data and creating training sets and applying various models. Including support vector regression, random forest. Model accuracy was calculated with and with cross validation. It is the same methodology for optimization of equipment as is process

**4.8 Deep Learning via Neural Networks**

Deep learning provides advance analytics tools for processing and analysing big manufacturing data. It is seen as a breakthrough solution to the challenges of multimodal data, the high dimensionality of feature space and multicollinearity among data measurements. It has applications in speech recognition, image recondition, it allows automatically processing of data towards highly nonlinear and complex feature abstraction via a cascade of multiple layers, (Wang et al, 2018, Rivas et al, Manami et al, 2023, Trask, A.W., 2019, Chollet, et al., 2021). The following review looks at its application in process optimizations in manufacturing.

Pfrommer et al, 2018, used ANN as a surrogate model to optimize the draping process of textiles. He used the surrogacy process to eliminate the costly process of trial experimentation that is often the approach for process optimization. The study achieved improved results using surrogacy with deep ANN but highlighted the importance of the training data and the sampling.

A combination of a neural networks with a genetic algorithm to predict critical parameters in a particle board and axial flow designs manufacturing process. The Genetic algorithm was then applied to the trained NN to determine the optimum values using the parameters successfully. (Cook et al, 2000, Liu et al 2023)

Fang et al, 2019, learned from Chen and Wang et al, 2013 studies to forecast the cycle time in a wafer fabrication factory using a fuzzy back propagation network. Chens et al studies showed that the NN models lacked generalization and fitting capabilities to deal todays big manufacturing data. It also highlighted the importance of considering expertise for feature extraction to reduce the input dimensionality. Fang et al, 2019 considered all this in their study of shop floor real time job remaining time prediction. They used a Deep stacked sparse Autoencoder to deal with every changing manufacturing floor. In its conclusion it suggested looking at LSTM, long short time models to analyse time series problems and find bottlenecks in the production line. It also suggests reinforcement learning to help with self-learning in the production control and optimization.

A Comparison between ARIMA, LSTM and GRU models was done on time series forecasting for bitcoin technology. ARIMA, Autoregressive integrated Moving average, turned out to be the better model, followed by the GRU model gated recurrent network whereas LSTM allows the tracking the dependencies of new observations with past ones. This study gave details on the methodology on creating and comparing types a recurrent network, with certain parameters and evaluation metrics. It highlighted the importance of data normalization, using Scikit-learn making it easier for the network model to learn. Another highlight from these studies was the importance of the architecture of the neural network itself and optimisation of these methods (Yamack et al, 2019, kuric et al, 2022, Sun et al, 2019).

The deaeration time prediction task is essentially a time series prediction task. Time series can be defined the time of occurrence of the series of quantities of some investigated parameter, (Kuric et al, 2022). The time series is used to monitor time changes and monitor trends in the development of the examined parameter. The most used models in research and practice for time series prediction are linear autoregressive models (AR and ARX), LSTM neural networks, moving average mode (MA) and autoregressive moving average (ARMA), (Kuric et al.,2022).

Yamack et al, 2019 conducted a comparison between three different machine learning models in making a time series prediction. The three were ARIMA (Auto regressive integrate moving average), LSTM (Long Short-term Memory) and GRU (Gate recurrent Unit). The topic was predicting the price of Bitcoin. This research compared all three models by accuracy (measured by MAPE and RMSE) and time. It concluded that the outcomes could have been of effected by the parameters that were chosen and the total amount of data used.

Fang el al.,2019 examined the time predictions to complete a job in a discrete manufacturing system using a deep learning-based approach. A deep Stacked sparse autoencoder (S-SAE) model was designed to enable a machine algorithm to learn highly varying status of manufacturing for prediction. In this research the performance was compared to different models, such as linear regression, back propagation neural networks, multi- layer neural networks and deep brief neural networks. To do this, the above models were trained and tested using the same dataset via a fivefold cross-validation. He concludes that future work, includes using LSTM neural networks to analyse time series in production.

**4.9 Data sources, collection, and Preparation for machine learning in production process optimization**

These are the most important activities of building a machine learning model. It is important to collect the relevant data and create a proper dataset. Data sources can be external and internal. In this research, support for data and information communications come from various industrial information systems which is integrated. Support systems are Manufacturing execution systems (MES), supervisory control and data acquisition (SCADA) system and the programmable logic controller system (PLC) which directly controls the reaction parameters of machines and warehouse management system (WMS). The SCADA system is where production data is acquired, and it is stored on the relational database system called shopfloor system. (Min et al, 2019)

This data can be unstructured, and need to be prepared, by screening for duplicate data, missing data, irrelevant data records. Extraction of indicators and features by labelling the data that will be needed in the learning/training process. Data needs to be mapped depending on the knowledge of what it is going to be used for. (Lee et al, 2019, Min et al , 2019).

The data for this research was originally stored on the manufacturing shopfloor database system. For each batch produced, each production process time step was recorded. By selecting the production time, we were able to request via SQL query all material batch details that had a deaerating step in their process. Therefore, reducing the amount of data was sent and stored as a CSV file. This data will need to be prepared as stated above by Lee et al before it can be passed to a machine learning model.

**4.10 Model training**

Once the dataset is established, it needs to be split into three subsets: Training, validation, and test sets. There is no optimal proportional for splitting, according to research, the rule of thumb was 70% for training and 15% for both validation and testing. It depends on the size of the dataset, (Barkov, 2019., Rajasekaran et al, 2022) The training set is used to build the model, validation set will be used to choose the machine learning algorithm and find the best values for hyperparameters, and test set will be used to assess the model,

To train the machine learning model, the collected data must be divided up into training and test sets, which are randomly divided by 70:30 ratio of the entire data set. (Cavalcante et al, 2019, Goli et al, 2019).

In only one research literature, it mentions the two types of infrastructure, data collection, which we discussed earlier, and it readily applied in industry, however the second type is model training. It consists of cloud computing, big data analysis and machine learning. Cloud computing is used to support the collection and selection and analysing of data from ambient environments using centralised methods. To successfully implement machine learning, it requires a training set. It also requires the ability to continuously learning from training data to improve. For this, a presence of good computing memory so that the knowledge discovered by the trained datasets will be well -stored. In terms of IT infrastructure, security, privacy, and resource constraints need to be considered. (Sing et al 2020)

**4.11 The Deaeration Process in Beverage Industry**

It plays a crucial role in ensuring the quality and shelf life of the final product which was documented by Feilner et al. It involves the removal of dissolved gases such as oxygen from the beverage. The deaeration time must be carefully calibrated to achieve optimal results with the goal of preserving the desired sensory characteristic and stability of the beverage (Paquin et al, 2009).

There are many process parameters that can affect the length of deaeration time a beverage batch needs. Examples of process parameters include the type of final product to be produced, the equipment used, the initial methods of addition and nature of raw material added, temperature and pressure factors. Optimization of these parameters can lead to a lower deaeration time thus an increased efficient process without compromising the final product quality and stability.

Industries acquire huge amounts of data throughout the process production stages through sensors and other data collection forms. To optimize these parameters, this data needs to be interpreted and analysed. One method of doing this is Machine Learning, subfield of artificial Intelligence.

Machine Learning has successfully been applied to many areas in manufacturing, the focus area for this research is the prediction and improvement and manufacturing process optimization, (Ismail et al, 2021).

* 1. **Conclusions**

When researching the use of machine learning in manufacturing, the largest theme running is Industry 4.0 and smart factories. This was introduced in 2016, otherwise known as the fourth industry revolution, (Rai et al, 2021). Machine learning has a major role in this revolution where it enables the generation of actionable intelligence by processing the collected data to increase manufacturing efficiency with changing the required resources or process. In other words, make manufacturing smarter, (Aksa et al, 2021).

From the above literature review, the most common thread is the use of more than one machine learning model to assess data and make predictions. A important aspect of all, is the data itself, its collection and preparation in order to understand the relationships and trends from it and that’s even before applying any machine learning model to it. The type of data, whether is labelled, structured , or unstructured will affect the machine learning model of choice. The methodology applied to the various research uses the same steps and will be helpful in this research as will the metrics suggested being used to evaluate the predictions and trends.

# **Chapter 4: Methodology**

* **Research Methodology**

Research methodologies typically fall into two main categories: primary and secondary data collection (Saunders, Lewis, and Thornhill, 2016). For the research questions poised at the start of this paper, a combination of primary and secondary data collection will be utilised.

Within both categories, various methods of data collection can be utilized. This section aims to outline the chosen collection strategy or strategies that, in the author's view, best serve to answer the research questions presented in Section 1 above.

**Primary Research Data Collection**

The author determined that employing a qualitative and quantitative approach for collecting primary research data would be the most suitable technique to gather first-hand insights, approaches as observing and interviewing individuals knowledgeable in beverage process production and data analytics.

Factors considered by the author when choosing this primary research method encompassed:

* Research Objectives – the appropriateness of the research method involved.
* Author status – role/access in the organisation if applicable
* Expertise of the Selected Participants – availability and validity
* Timeframe – how long this research method could take.
* Ethical Considerations – gaining consent of the participants to partake and the use the data collected.
* Bias - ensuring a diverse range of inputs by involving individuals with distinct roles in the beverage production process.

Two approaches are as follows:

**Data Collection through Observation**: **Qualitative Approach**

The author works full-time in a beverage production company and had the opportunity to be a participant as observer in the production area. This qualitative method was deemed appropriate as it helped develop a better understanding of the production process which is where the root of the research questions originated.

To gain access to the production area, an informal request to observe the production process was made to the appropriate production manager and associate. It was clearly stated as to the reasons why this request was made. The author explained in detail the research that was been conducted and the potential benefit to this observation would make to the research. Such benefit for example would be a clearer picture for the author of what is going on in the process.

The Author also checked prior to the visit that it would be ok to write notes on any observations or informal conversations had. It was also highlighted to all participants that they would remain anonymous. These notes will be in the appendices.

Potential errors in this type of research approach considered were:

* Observer Error could occur here due to the authors lack of understanding of the process area. There is importance on the interpretation of the data collected.
* Observer Effect where the author could affect the behaviour of those being observed thus posing a threat to the reliability and validity of the data collected. This also known as the Hawthorne effect.

The observational data collected, and its potential errors outlined above would be further clarified in further primary method of in person interviews.

The observational approach proved successful as the author gained invaluable insight into the physical workings of the process and made connections with experts in the process area which went on to prove useful in the next primary research approach – in person interviews. One such insight was the terminology used in the process area and how it related to the secondary data collected.

**Data Collection through in person interviews:**

In-person interview can take many forms but the one used for this research is an unstructured informal interview, and for practicality and ease of transcription, these interviews were conducted online via the Microsoft Teams platform. By employing an informal interview using open- ended questions as this data collection method, it facilitated a more comfortable scenario allowing for in depth discussions surrounding the beverage production process and the potential use of data analytics. Chosen Participants were able to freely talk about the process and the author was able to observe the participant reactions and adapt the interview flow, accordingly, as suggested by Saunders et al. (2012). This is a more flexible approach.

To address potential respondent bias, the author ensured a broad range of perspectives by engaging individuals from various roles within the beverage production process. Each interviewee contributed a distinct viewpoint: one oversaw the production process, another actively worked in the production area, and the third played a role in data analytics within the organization. Through interviews with this diverse group, the author aimed to achieve a comprehensive understanding of the existing beverage production process, encompassing both common practices and individual variations.".

Transcripts from the interviews can be seen in appendix XXX.

**Secondary Research Data Collection**

The author determined that employing a quantitative approach for collecting secondary research data would be the most suitable technique.

Factors considered by the author when choosing secondary data research method encompassed:

* Research Objectives – the purpose and compatibility of the secondary data
* Author status – role/access in the organisation if applicable
* Expertise of the Selected Participants – availability and validity
* Timeframe – how long this research method could take.
* Ethical Considerations – gaining consent of the participants to partake and the use the data collected.
* Bias - ensuring a diverse range of inputs by involving individuals with distinct roles in the beverage production process.

The secondary data source used was internal to the organisation, from various industrial information systems which is integrated. The support systems are FactoryTalk batch software and SQL Server Management Studios.

This data can be unstructured, and need to be prepared, by screening for duplicate data, missing data, irrelevant data records. Extraction of indicators and features by labelling the data that will be needed in the learning/training process.

The data for this research was originally stored on the manufacturing shopfloor database system. For each batch produced, each production process time step was recorded. By selecting the production time, the author was able to request via SQL query all material batch details that had a deaerating step in their process. Therefore, reducing the amount of data was sent and stored as a CSV file.

**Research Validity**

To determine the extent of which the primary and secondary data collection above accurately assesses the research problem objectives is the basis of validity. We need to validate if the data collected for this research answers the research questions poised and are these answers trustworthy and meaningful.

The most relevant components of validity:

**Internal Validity**: Addressing potential biases and threats to the reliability of the collected data. The use of diverse participants from various roles within the beverage production process helps mitigate bias, and employing both observation and interviews allows for cross-validation of findings, enhancing the robustness of the research outcomes.

**External Validity**: The extent to which the findings can be generalized to other contexts. By involving individuals from different roles within the production process and considering multiple data sources such as SCADA and MES, the research aims to capture a broad representation of the beverage production domain, increasing the potential for generalizability.

**Ecological Validity**: Ensuring that the research settings and conditions closely resemble real-world situations. The use of actual production environments and interactions with experts in the beverage production area through observation and interviews enhance the ecological validity of the study.

**Content Validity**: Ensuring that the data collected accurately represent the content domain of the research. The qualitative approach of observing the production process and conducting unstructured interviews with knowledgeable individuals helps capture a holistic and nuanced understanding of the beverage production process and its data analytics.

**Face Validity**: The extent to which the research methods appear to measure what they intend to measure. The use of multiple data collection methods, combined with clear communication with participants about the research objectives, helps establish the face validity of the research approach.

By addressing these components of validity, the research aims to enhance the credibility and reliability of its findings, contributing to a more robust and meaningful exploration of the beverage production process and data analytics.

**Research Ethics**

Ethics in research is important because it ensures that the participants are treated fairly and respectfully. It will also help protect their privacy and well-being. It also ensures that the data gathered is honest and trustworthy.

In the context of the primary data collection, the expert participants will be invited to join the research with the freedom to withdraw consent or exclude their data contributions at any point before the submission date. During the interviews a professional approach will be maintained with confidentiality and anonymity of the participants and the data information given preserved for research purposes. A master file containing the responses will be kept by the author and will not be shared. This master file will be securely stored, and password protected as an enhanced security. Additionally, if the participant has any queries during or after the interview, they will be answered promptly and sensitively.

The secondary data source used was internal to the organisation and it is stored on the relational database system called shopfloor system. The author followed proper protocols when gaining access to the data by engaging in the relevant channels within the organisation. Such channels include the organisations legal counsel and ethics point software where you can determine if a non-disclosure is relevant. However, since the data received will be swiftly anonymized to prevent linkage to the individual participants and organisation, there was no legal or ethical worries. The safeguarding of the data will be ensured through secure storage and the implementation of passwords.

Furthermore, the study acknowledges the importance of adhering to the General Data Protection Regulation (GDPR) and is committed to implementing necessary measures to uphold compliance throughout all stages of the research process. This demonstrates a commitment to ethical research practices and the protection of participants' privacy and rights.

1. **Sampling Strategy**

An important part of any research is the collected data and how to determine how much if it is needed to answer the questions posed by the research. The data collected through the primary and secondary data collection as described above constitutes the population. Depending on the research’s objectives, it may not be necessary to analyse the entire population. A sampling strategy may be needed to be employed to reduce the population but still maintain its representativeness.

The type of sampling strategy will be different for the data collected through primary and through secondary. Primary data was collected through in-depth unstructured interviews which is qualitative research and thus nonprobability sampling will be applied. For Secondary data collection, the type of sampling examined would be probability sampling where every item of the population has an equal chance of being included in the sample.

**Primary Data**

In-depth non structured interviews were used as the primary data collection methodology and therefore only a non-probability sampling strategy could be applied. The participants were purposely selected because of their expert knowledge and distinct perspectives in the research area. The type of non-probability sampling strategy used here is known as purposive or judgemental sampling. This strategy is convenient and appropriate for exploratory research design, but unlike random sampling it doesn’t try to represent the whole of the population but focuses on the depth of the information provided.

**Non-Probability Sampling Strategy: Purposive sampling**

**Validity and Generalizability:**

The use of purposive sampling might not provide a representative sample of the entire population, it does offer high validity for addressing specific research questions. The insights gained from participants with specialized knowledge contribute to a deeper understanding of the topic. However, the findings may not be easily generalizable to broader populations.

**Advantages and Disadvantages:**

Purposive sampling is advantageous for its ability to provide rich and targeted insights from individuals with expertise. It is particularly useful when the goal is to explore complex or specialized topics. However, its limitations include potential selection bias and reduced generalizability to broader populations.

**Conclusion:**

Purposive sampling is a valuable non-probability sampling strategy that allows researchers to gather in-depth and contextually relevant insights. By intentionally selecting participants with specific expertise, researchers can obtain valuable information that addresses the research objectives. While the findings may not be easily generalizable, the depth and quality of insights obtained make purposive sampling an important approach in research design.

**Secondary Data**

The secondary data collection source used was internal to the organisation and it is stored on the relational database system called shopfloor system. Here all data relating to production of all batches in the organisation are stored. This type of data would be described more recently as Big -Data due to the shear volume and velocity of it. This is a massive population size and not feasible to work with so therefore a sampling strategy is needed to be applied.

The type of strategy initially examined was probability sampling where every item in the population which was the database has an equal chance of being included in the sample. But this was too much data so to narrow down the population, based on the research objectives, the type of materials specifically those containing mucilage and a time of 2 years was chosen. This changed the sampling strategy to purposive or judgemental sampling, an example of non-probability sampling. A focus on a specific time range could yield possible trends, patterns, or changes during that period by examining variables involved in the production of the batches.

**Non-Probability Sampling Strategy: Purposive sampling**

**Sampling Frame:**

The sampling frame is ‘a complete list of all the data in the population from which the sample will be drawn’. So, in the case of the secondary data collection, the complete list is the organisation database of all the production batches produced and the sample is the filtered production data containing only mucilage containing materials, it will be necessary to consider the different production tanks used.

**Sample Size:**

Based on the research objectives, only mucilage containing materials produced in the last 2 years were chosen to be filtered from the main organisations databases. These materials, 46 in total, can only be produced in a range of 16 production tanks varying in capacity. The scheduling of these materials depends on customer demands. A total number of 347 batches were produced in this time over various tanks. The number of batches per tank produced were in the range from as low as 6 to 54, this allowed all batches to be included in the sample.

**Sample Technique:**

Under Purposive sampling, 6 techniques were reviewed including extreme case is only applicable to more unusual research. Homogenous sampling was deemed appropriate where focus is given to a particular subgroup in this case is the production tanks used in which all the materials produced are similar in that there are mucilage containing. The characteristics of tanks are similar which allows them to be explored in a greater depth.

**Check for Representation:**

The Author carefully chooses production batch material data with specific ingredient such as mucilage component. This approach enhances the depth and breadth of the insights gained.

**Validity and Generalizability:**

Like primary data collection, the use of purposive sampling might not provide a representative sample of the entire population, it does offer high validity for addressing specific research questions. The insights gained from the mucilage containing material production data chosen will contribute to a deeper understanding of the research topic. However, the findings may not be easily generalizable to broader populations.

**Advantages and Disadvantages:**

Purposive sampling is advantageous for its ability to provide rich and targeted insights from individuals with expertise. It is particularly useful when the goal is to explore complex or specialized topics. However, its limitations include potential selection bias and reduced generalizability to broader populations.

**Applying the Sampling Strategy**

Purposive sampling is a valuable non-probability sampling strategy that allows researchers to gather in-depth and contextually relevant insights. By intentionally selecting participants with specific expertise, researchers can obtain valuable information that addresses the research objectives and selecting mucilage containing materials over 2 years, the author can obtain invaluable insight in the production of these materials in various production tanks under various conditions.

While the findings may not be easily generalizable, the depth and quality of insights obtained make purposive sampling an important approach in research design.

The sampling strategies outlined above will be applied to primary data collection and secondary data collection.

**Experimental Methodology**

This chapter offers a detailed explanation of the experimental methodology adopted for answering the research objectives of Exploration and Quantification of Production Downtimes, applying machine learning models and predictive modelling of production downtimes.

Below is a breakdown of the main goals and how they were approached.

Research Objective 1: Understanding Production Downtimes

First, we wanted to understand the current process better. We investigated how often and why there were pauses (downtimes) in the various phase stages in the production manufacturing tanks. This gave us a clear picture of where there were production phase overruns for each tank.

Research Objective 2: Using Machine Learning to Analyse Data

Next, we turned to machine learning to analyse the production batch data. By studying the data, we wanted to point out where the process could be made more efficient. This step showed the value of using advanced tools and methods to analyse production data.

Research Objective 3: Predicting and Planning for Downtimes

Lastly, we created machine learning models to predict when these downtimes might happen. With these predictions, we aimed to plan production schedules better, reducing the number of pauses and making the whole process faster.

The comprehensive approach that follows integrates data collection, pre-processing, model selection, training, evaluation, and optimization.

**3.2 Data Collection**

**3.2.1 Source of Data**

The author acquired historical production process batch data through the Microsoft SQL Server Management Studio (SSMS). A SQL request was created with the following requirements:

* Time period – 2 years
* Production batch Material – Mucilage beverage containing batches that require a deaeration phase.
* All production phase details including time duration details.

This information was received in the excel format and was converted to a CSV file for better management and transferring of data because it preserves the original data values. More importantly CSV format is easily read and written by python programs.

**3.2.2 Features and Target Variables**

The features integrated into the model include:

* **Quantity** (Kg)
* **Phase\_duration** (min)
* **Phase\_start\_delay** (L/min)
* **Flowrate\_KGMIN (KGMIN)**

The primary target variable is **Phase\_overrun (min)**

**3.3 Data Preprocessing**

**3.3.1 Software , Libraries, Web Applications**

For this research, the programming language employed was Python. Its selection stemmed from its user-friendly nature and the ease with which the researcher could learn and utilize it. The vast availability of open-source libraries in Python further enhances its adaptability, allowing customization based on the programmer's requirements, (McKinney,2012). One of the major advantage of using python and its libraries, is that they can be used together, in tandem for data analysis , scientific computing and machine learning. The following are examples of the open-source python libraries that were used:

* Pandas – provides data structures like DataFrames and Series that make it easy to handle structured data. It allows for quick data cleaning, visualisation, and statistical analysis. In this research its main use was in the handling of missing data, filtering rows and aggregation of data.
* NumPy -employed for its basic mathematical and array operations.
* Matplotlib – A plotting library – creates and display graphs and visualisations of the data working alongside other libraries such as pandas.
* TensorFlow – Deep learning frameworks – used for building and training neural network models.
* Seaborn – works along with Matplotlib to create statistical graphics.
* Scikit-Learn - allows a programmer to quickly implement a range of machine learning algorithms in conjunction with other libraries such as pandas and matplotlib.
* Scipy – a advanced scientific computing library built on NumPy for scientific computational tasks. It adds additional functionality.

All the libraries above were used in Jupiter notebook which is another open-source web application that allows the creation and sharing of documentation that contain live code, equations, visualisations, and narrative text.

* + 1. **Data Cleaning**:
       1. **Handling missing values**:

Missing values are entries in the dataset, such as 0 or not a number (NaN). Using the panda’s library in python, the amount and type of missing values should be determined, (S.Xu et al , 2015). What columns in the data contain missing values and will it have an impact on the model used are questions that will be answered.

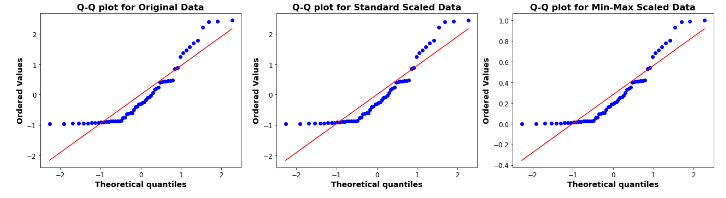
* + - 1. **Removing duplicates**

To protect data integrity, the dataset needs to be checked for duplicate rows of data. Duplicated data can distort the actual data analysis and give inaccurate results, skew distributions. Using panda’s library, duplication can be determined, and before removing by drop function, the data results need to be check in case they are valid repetitions.

* + - 1. **Data Normalization**

This pre-processing technique that involves the transformation of all numeric variables in the dataset to a standard scale. This is necessary when the data features in the dataset have different units or varying scales which can give skew interpretations from machine learning models.

An investigation using the Quantile-Quantile (Q-Q) Plot was looked at to determine the effect of the different types of normalisations such as standard scaling and minmax scaling have on the data. Using the data from the production tank 22MT – reference ProductionTank22\_df2, the following graphs shows results of the distribution of the phase overrun data. Using two examples of standard scalar and min-max scaler , there is a S – Shaped curve which indicates that the data may be following a logistic or s curve distribution rather than a normal distribution.



From the data above , there was no difference between the two scalar methods and below table gives details of the machine models and their sensitivity to scaling.



The StandardScaler function from Scikit-learn was used to execute Z-score normalization, ensuring each feature contributes equally to model efficiency. This a solid general-purpose choice and had broad applicability,

* 1. **Exploratory Data Analysis (EDA)**

Exploratory Data Analysis is an approach of analysing data sets to summarize their main characteristics. It is a critical process of performing initial investigations on data. It is used to discover trends or patterns, to spot anomalies and check statistical assumptions with the help of summary statistics and graphical representations. This is aided by the python libraries: Pandas/ NumPy/ Matplotlib/ Scikit-Learn/ Seaborn.

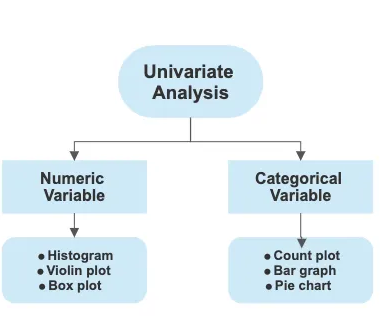
As Phase overrun is the target variable, the EDA was used to gain understanding of its distribution and characteristics in relation to the other feature variables for each production tank in the DataFrame.

Further EDA was examined for the following using phase overrun as the target variable:

* Ingredient Addition – gum material addition
* Agitation Phase times
* Deaeration phase times

EDA was examined under the following headings:

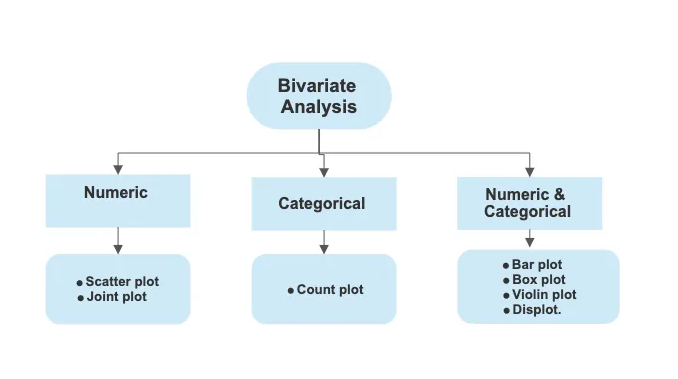
**3.4.1 Univariate Analysis:**



**Numerical Features**: Histograms, box plots, and descriptive statistics determined to understand the central tendency, spread, and shape of the distributions of the target and feature variables.

**Categorical Features**: Use frequency tables or bar plots to understand the distribution of each category.

**Bivariate and Multivariate Analysis:**



**Correlation Analysis**: Compute correlation matrices for numerical features to determine if there were any linear relationships between the variables.

**Scatter Plots**: For visual inspection of relationships and trends between pairs of numerical features.

**6. Identification of Outliers**:

**Visual Methods**: Use box plots to visually inspect for outliers.

**Statistical Methods**: Techniques such as the Z-score or IQR method can help in a more formal identification.

**9. Reporting and Visualization**:

Leverage visualization tools like Matplotlib, Seaborn, or Plotly in Python for plotting.

Documentation: Summarize the findings, insights, and any patterns observed during the EDA in a clear and concise manner, preferably complemented with visual aids.

* 1. **Machine Learning Application**

An evaluation of multiple machine learning models was performed, see table for details of models examined and their advantages and disadvantages.

For each production tank group data, the following steps were completed to apply each model to the data.

**3.5.1 – Data loading and Processing**

The data was loaded into a pandas dataframe. The features and target variable were defined.

**3.5.1 Dataset Split**

Using the train\_test\_split function in Scikit-learn, the dataset was partitioned into 80% for training and 20% reserved for testing.

* + 1. **Models Initialization**

– see table below for details of the models investigated.

* + 1. **Cross -Validation**

for each model in the table, a 5- fold cross-validation was performed, (CV=5). The mean and standard deviation of the mean square error was calculated.

A list of the machine models is creating so that a loop can iterate over each model to perform the cross validation. The model is trained and evaluated 5 times (CV  
=5), each time with a different split of data in training and validation sets.

The scoring parameter is neg\_mean\_squared\_error which means the mean squared error will be calculated for each of the 5 crosses.

* + 1. **Evaluation Metrics** –

The predictions from the models are compared to the true values to calculate the Mean Squared Error (MSE) and the R-squared (R2) score for both training and testing datasets. The MSE provides a measure of the prediction error, whereas the R2 score provides a measure of the proportion of the variance in the dependent variable that is predictable from the independent variables.



* + 1. **Hyperparameter Tuning** –

For models like Ridge Regression, Lasso Regression, Random Forest Regressor, Gradient Boosting Regressor, Decision Tree Regressor, and Bagging Regressor, hyperparameter tuning is performed using GridSearchCV. This ensures that the best parameters are chosen for the model, enhancing its performance. The best parameters are outlined below and in table :

* Ridge and Lasso Regression: The regularization strength, denoted by alpha, is tuned.
* Random Forest and Gradient Boosting Regressor: Parameters like the number of trees (n\_estimators), and the maximum depth of the trees (max\_depth) are optimized.
* SVR: Parameters like the regularization strength (C), epsilon, and the kernel type are optimized.
* KNN: The number of neighbors (n\_neighbors) is tuned for optimal performance.

|  |  |
| --- | --- |
| **Breakdown per Machine Learning Model for Hyperparameter Tuning** | |
| **Model** | **Key Parameters/Steps** |
| **Linear Regression** | Fit, predict, evaluate, and store the results. |
| **Ridge Regression** | Perform hyperparameter tuning using GridSearchCV with specified alphas. Fit, predict, evaluate, and store the results.  Set alpha=1.0 |
| **Lasso Regression** |
| **Random Forest Regressor** | Perform hyperparameter tuning using GridSearchCV with specified n\_estimators and max\_depth. Fit, predict, evaluate, and store the results. Note the feature importance. |
| **Gradient Boosting Regressor** | Perform hyperparameter tuning using GridSearchCV with specified n\_estimators, learning\_rate, and max\_depth. Fit, predict, evaluate, and store the results. Note the feature importance. |
| **Decision Tree Regressor** | Perform hyperparameter tuning using GridSearchCV with specified max\_depth. Fit, predict, evaluate, and store the results.  Set random\_state=42 |
| **Bagging Regressor** | Using Decision Trees by default. Perform hyperparameter tuning using GridSearchCV with specified estimators, max\_samples, and max\_features. Fit, predict, evaluate, and store the results.  Set estimators=100 and random state=42 |
| **AdaBoost Regressor** | Fit model with specified estimators. Perform hyperparameter tuning using GridSearchCV with specified estimators and learning rate. Fit, predict, evaluate, and store the results.  Set estimators=100 and random state=42 |
| **SVR** | Perform hyperparameter tuning using GridSearchCV with specified C, epsilon, and kernel type (e.g., 'linear', 'rbf'). Fit, predict, evaluate, and store the results. |
| **KNN** | Perform hyperparameter tuning using GridSearchCV with specified neighbours. Fit, predict, evaluate, and store the results. |
| **Neural Networks** | Define the architecture (number of layers, neurons per layer, activation functions). Compile the model specifying loss, optimizer, and metrics. Train the model using fit. Predict, evaluate, and store the results. Optionally, adjust hyperparameters such as learning rate, batch size, or epochs. |

* + 1. **Individual model training –**

For predictive modelling, each model goes through a structured process of training, prediction, and evaluation. The process ensures that models are validated on unseen data to provide a reliable measure of their expected performance on new data.

**Training Process:**

Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regressor, Gradient Boosting Regressor, Decision Tree Regressor, Bagging Regressor, AdaBoost Regressor, SVR, KNN, and Neural Networks:

**Training:** The model is trained using the fit() method with the scaled training dataset.

**Prediction**: Once the model is trained, predictions are made on both the training dataset and the testing dataset. This helps in understanding how well the model has learned from the training data and how it generalizes on unseen data.

**Evaluation:** This was facilitated using the metrics module in Scikit-learn the predictions from the models are compared to the true values to calculate the Mean Squared Error (MSE) and the R-squared (R2) score for both training and testing datasets. The MSE provides a measure of the prediction error, whereas the R2 score provides a measure of the proportion of the variance in the dependent variable that is predictable from the independent variables.

**Neural Networks require additional steps:**

**Architecture Definition**: The architecture of the neural network is defined, including the number of layers, neurons in each layer, and activation functions.

**Compilation**: The neural network model is compiled specifying the loss function, optimizer, and evaluation metrics.

**Training**: The model is then trained using the fit() method. Hyperparameters such as the learning rate, batch size, and number of epochs can be adjusted based on performance requirements.

**Summary:**

Through this structured training process, each model is systematically trained, tested, and evaluated. The results provide insights into the performance of each model and help in selecting the best model for predictions on new, unseen data.

**3.8 Limitations**

The study's potential limitations encompass:

A strong reliance on the accuracy and consistency of data extracted from the XYZ production plant through the PlantLog V2.3 system.

The inherent assumption that the highlighted features have a linear impact on the product quality.

**3.9 Conclusion**

This chapter presented a comprehensive methodology for leveraging machine learning to enhance production parameters. The following chapter will discuss the results obtained from implementing this methodology.

# **Chapter 5: Results and Discussion**

**5.1 Introduction**

This chapter explains the impact of the primary data and secondary data on the research objectives, through the information received from the three expert participants and the production batch data downloaded from the organizations database system.

The author examines various machine models to determine if the production downtime as phase overrun variable can be predicted for the various production tanks used to produce mucilage containing beverage materials. It leverages historical data and relationships between features and a target variable to determine if predictions about phase downtime in the future is possible. The downtime was examined for three phases of the production: Agitation, Deaeration and the addition of Gum ingredient.

**5.2 The Primary Data**

From the in-depth interviews, responses from the participants with expert knowledge and distinct perspectives in the research area was reviewed and summarised here. The actual interview transcripts are in appendices. .

**5.2.1 Participant no. 1: Data Scientist**

Participant number 1, a Data Scientist, played a crucial role in enhancing the existing system's objectives. Initially, the system solely focused on executing batches, with data collection lacking a structured approach for analytics. To rectify this, the participant was involved in refining tracking mechanisms by developing a stored procedure that summarized batch data comprehensively, covering aspects like start times, consumption, problems, weights, agitation times, deaeration times, and homogeneity. They further established targets for different phases through historical data analysis, providing clear objectives for each batch and aiding in issue identification when batches fell short. Resource management was addressed by quantifying departmental needs for shared resources like mobile tanks. Operational Efficiency (OE) calculations, focused on equipment effectiveness rather than usage, were employed to pinpoint areas for improvement. The participant also highlighted the company's emphasis on data analytics, mentioning the creation of a Digital Performance Management (DPM) system for data visualization and future consideration of machine learning where it genuinely adds value. Moreover, they stressed the significance of preventative maintenance, citing examples of early issue detection through thermal balance tracking and the importance of team awareness and training in efficient problem resolution.

The key insights from the Data Scientist's contributions revolved around enhancing batch process tracking, target setting, issue resolution, and emphasizing the role of data analytics and preventative maintenance in optimizing organizational processes.

**5.2.1 Participant no. 2: Production Manager**

Participant 2 highlighted the company's data-driven approach, where they utilize operational data to calculate Overall Equipment Effectiveness (OE) figures for each batch, which are then reviewed during daily 9:15 meetings involving cross-functional teams. They stressed the importance of acting based on these figures, with diligent follow-up to ensure completion. OE calculations are performed at the phase level, considering prerequisite times, enabling a granular assessment of phase efficiency. The interviewee discussed various batch phases, such as agitation, deaeration, and ingredient additions, outlining their timing and criteria for proceeding to the next phase. Quality checks before batch finalization were emphasized to meet quality standards. Initiatives to reduce downtime were mentioned, including pre-weighing bulk powders, and optimizing temperature and pump speed. Downtime challenges like liquid filling delays and breakdowns were addressed. Recording downtime reasons and their impact on OE data were explained. Setting targets for parameters like flow rate involved collaboration between different teams. Tank assignments, numbering, and capacity details were provided, with some batches requiring multiple tanks. Different systems, ranging from 20-tonne to medium-to-small systems, were discussed based on production volume needs. An efficiency improvement example involving juice barrel handling was presented. The interviewee reiterated the importance of continuous collaboration, data analysis, and process enhancement to minimize downtime and boost productivity.

**5.2.1 Participant no. 3: Production Operative**

From the participant's point of view, the conversation primarily revolved around production downtimes and efficiency in a manufacturing process, specifically related to gum mixing. The participant discussed the challenges of interruptions during production, which could result in extended downtimes, impacting operational efficiency (OE). They emphasized the importance of accurately recording extra downtime for batches where gum mixing takes longer due to poor agitation, as this data is crucial for analysis. Additionally, the participant acknowledged that target times for production phases should be more realistic to account for batch-to-batch variations and highlighted potential solutions like improving agitation and pre-mixing highly concentrated gum batches. The conversation also touched on the phases of a typical batch, including ingredient addition, agitation, and transitioning to the high-pressure phase.

The key insight from the participant's point of view is the focus on addressing production downtimes and operational efficiency challenges in gum mixing, emphasizing the need for accurate data recording, realistic target times, and potential solutions to improve the manufacturing process.

**5.2 The Dataset**

The original production csv file contained data on beverage batches that contain a mucilage ingredient produced in various production tanks with differing capacities over time of 2 years. In compliance with the organisations data usage agreement, specific data points have been omitted to maintain the confidentiality. Table 1 below provides an overview of the dataset’s column data and explanations. The dataset is called. ProductionDateupdated1



**Table 1 Dataset Column Headings and Detail**

Over 46 different type of beverage mucilage containing materials with a total of 367 beverage batches were produced across 16 production tanks of varying capacities. Which tank they were produced in; was dependent on the type of material and Quantity required. For the purposes of this data analysis , The batches will be examined , as not all materials would be produced in each tanks.

**5.2.1 Beverage Process Production Tanks (MT- Mobile Tanks)**

Table 2 shows the division of materials and their batches produced per production tank over 2 years. Some tanks are utilised more than others, this is determined based on the production schedule. Some of the production tanks are used as destination or holding tanks and may not be used for beverage production.

The production tanks were grouped together according to their capacity and the number of batches produced. The number of batches produced per tank over the 2-year time selected was low and would be an issue for the performing any accurate machine model on prediction. It would limit the training of the model and therefore affect the accuracy of the prediction.

For example, for production tank 22MT02, the number of batches produced was 13 which is a small number to perform any data analysis on, whereas for all the tanks in the group 22MT, the total number of batches were 47.



**Table 2 Production Tank /Materials/Batch Dataset Details**

**5.2.2 Production Phases (Instruction Steps)**

Table 3 below details what each Production Phasesmeans in the production of mucilage containing batches. The list of instruction steps is common to all batches detailed for each of the production tanks. So, for some materials, an extra HP (homogenisation may be required). This list of instructions or recipe is automatically activated at the start of the batch production.

For each of the instruction steps, the phase start time and finish time is logged. This data is converted into the total phase duration time, the start of the phase time, the phase overrun time which is all calculated based on the Target times logged into the system in the background. These target times for the phase duration and overrun times are historically determined and applied to each material. Each ingredient addition (Step1\_cons), the quantity and flowrate are also logged.

**Table 3 Instruction Steps**



|  |  |  |
| --- | --- | --- |
| **Instruction\_Step/Phase Step** | **Details** |  |
| S3\_BATCH\_IN\_PROGRESS | Batch process is in progress inside the tank. |  |
| STEP1\_CONS | 1st step in the consumption process where ingredients or materials are added to the tank: Treated Water |  |
| PLEASE VERIFY BULK ADDITION | A prompt to check and verify the bulk addition of materials or ingredients. |  |
| STEP1\_CONS |  |
| STEP1\_CONS |  |
| STEP1\_CONS |  |
| STEP1\_CONS |  |
| STEP1\_CONS |  |
| STEP1\_AGITATION | The agitation or mixing process that takes place after the first set of ingredients is added to ensure uniform distribution. |  |
| STEP2\_AGITATION |  |
| STEP3\_AGITATION |  |
| HP | High-Pressure phase or operation: homogenize the batch. |  |
| SELECT\_DESTINATION\_TANK | Indicates a phase where the finished or semi-finished product is transferred to a different tank, |  |
| S4\_BATCH\_COMPLETE\_QA\_PENDING | Signifies that the batch processing is complete and is now pending Quality Assurance (QA) checks. |  |
| TAKE A SAMPLE AND SUBMIT FOR QA. | An instruction for the operator to take a sample from the batch and submit it for quality assurance testing. |  |
| SAMPLE TO LAB. RESULTS OK? (NO TO HOMOGENISE) | A prompt suggesting that the sample has been sent for testing. |  |
| STEP8\_AGITATION | The agitation or mixing process |  |
| S7\_RELEASED\_TO\_FILLING | Indicates the batch has passed all processing stages and is now ready or has been released for filling |  |

**5.2.3 Ingredients and their Quantities**

The ingredients and their quantities are important to the phase overrun times, any delay in their addition reflects in the phase duration times. The table below lists the main ingredient components in the batches in the dataset. The quantities are approximated average, they depend on the recipe for each of the batch produced. The main ingredients that have the most quantities are treated water which is delivered via the automated bulk delivery system, and the mucilage /gum ingredient. The dry ingredients are all added manually via bags through the manifold on the top of the tank. So, for the gum ingredient, this is very labour intensive. Not every batch has a colour ingredient addition, but for those that do, there is a significant quantity to be added.



**Table 4. Ingredients Categories /Quantities**

**5.4 Data Preparation and Preliminary Analysis**

To prepare the data in a way that maximises the performance of the subsequent machine learning models using data pre-processing and feature selections.

The Original DataFrames which given actual details of the beverage containing mucilage batch production was reviewed, and some data was removed as it was deemed unnecessary for the purpose of answering the objectives of this study. But most importantly removed to preserve the anonymity of the data. Column heads were updated for easy of manipulation. Once this was performed, the dataset was saved and used for the research, reference Productiontankupdated 1. The total number of columns are 22 and rows at 9487. Initially exploratory data analysis was performed to determine the relationships and any trends between the variables.

The beverage batch data from each group of production tank were examined under the following instruction step/phases and these investigations were saved as follows also:

* Production Tanks: All Phases
* Production Tanks: Agitation Phase/
* Production Tanks: Deaeration Phase/
* Production Tanks: Gum Addition Phase

Each section above was prepared, and preliminary analysis was performed using pythons’ pandas library. The data processing step includes data cleansing and transformation, to ensure the quality and accuracy of the data.

* Data Cleansing was completed and involved removing or correcting errors, inconsistencies, and duplications in the dataset to ensure its accuracy and completeness. ensuring the quality of the data used for analysis.
* Data transformation was completed and involves converting the data into a usable format and structure for analysis. This includes aggregating data, splitting data, or merging data from different sources to create a unified dataset relative to the production tank and phase was being examined.

**5.4.1.1 Handling Missing Values:**

There were missing values in some of the rows of data. This was to be expected as the rows were for each of the instruction steps/phases for each batch produced. So, for example for rows relating to agitation, there wouldn’t be any values for ingredient addition, or quantities. The entries were in the form of Nan (not a number) or just 0. For each of the notebook investigations, the missing values were checked, and values changed to a 0. as the missing value signified that there were no occurrences/events, and zero is an appropriate replacement.

* + - 1. **Duplicate Values**
* There were no duplicate values in the dataset.

**5.4.2. General Visualisations from the main dataset: productiontankupdated1**

**5.4.2.1 Univariate Analysis:**

* **Numerical Features**:

1. The seaborn library(‘sns’) was used to visualise distribution of batches via a count plot overall production tank, (Fig 1). The height of each bar corresponds to the number of records for each tank. The 25MT03 and 04 tanks produced the most batches, they have a capacity of 10 tonne.

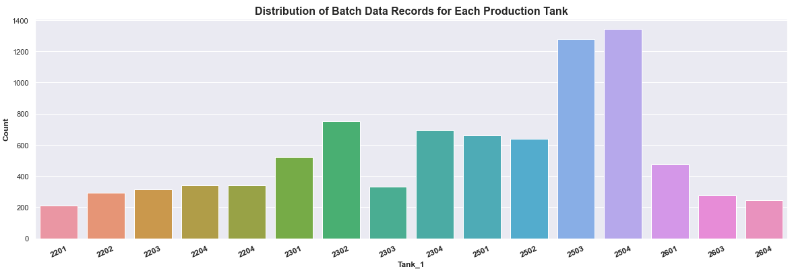
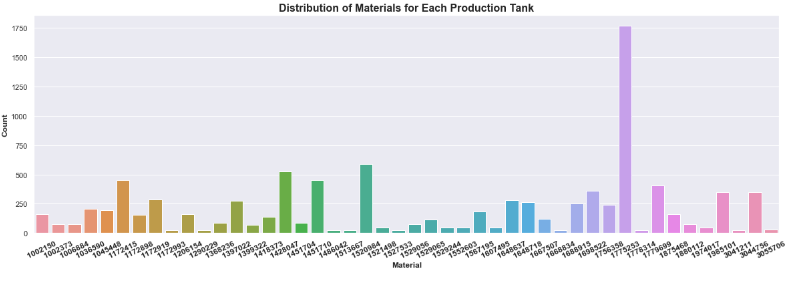


Fig 1 Distribution of Batches for each Production Tank

1. Fig 2 shows the distribution of material data records for each tank and allows you to see which materials were produced the most. This was beverage batch material 1756358 – produced 61 times.



**Figure 2: Distribution of materials overall production tanks**

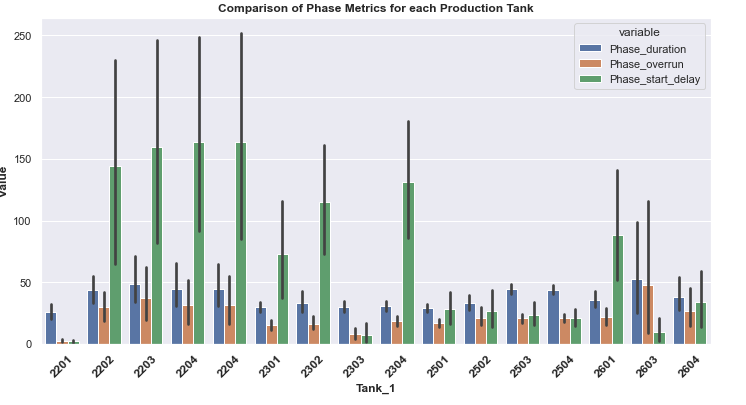
**5.4.2.2 Multivariate Analysis:**

The Bar Chart, in Fig 3, gives visual representation of how phase duration, phase overrun, and phase start delay varies across the different production tanks. Each bar represents a tank, and the height is the phase metric for that tank. All the 22MT and 23MT tanks which have the capacity of 20 tonne, show the highest values for the metric phase start delay, i, e it takes longer for the phase to start. for a production of a batch. In comparison to the lower capacity tanks of 25 and 25 MT tanks.

Phase duration metric is higher for 22MT than it is 23 MT tanks, showing that the batch production is slower in the 22MT tanks than 23MT tanks. There is phase overrun times (downtimes) for every production tank but higher in the 20 tonne tanks in specific 22 MT.

Notable outliers are Tanks 22MT01 and 23MT03 for low phase overrun, however these tanks were destination tanks and only a small number of batches were produced. In line with the information from participant 3, the production operative, he mentioned that tanks 22MT01 and 23 MT05 were the idea tanks to use for production , this is show in the graph where the 3 variables are the lowest.

This chart gives an overview of the tanks, however the number of batches per tank need to be considered, giving from table 2 earlier, more batches were produced in the 25MT tanks, yet the phase duration, overrun and phase start delay metrices were all lower than for those tanks with bigger capacity and less batches produced. This is the reason why the production tanks will be examined in their capacity groups and not individually.



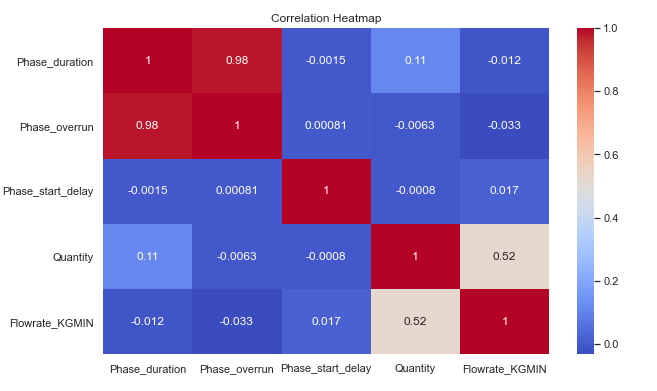
**Figure 3: Comparison of Phase Metrics for each Production Tank**

**5.4.2.3 Correlation Analysis**:

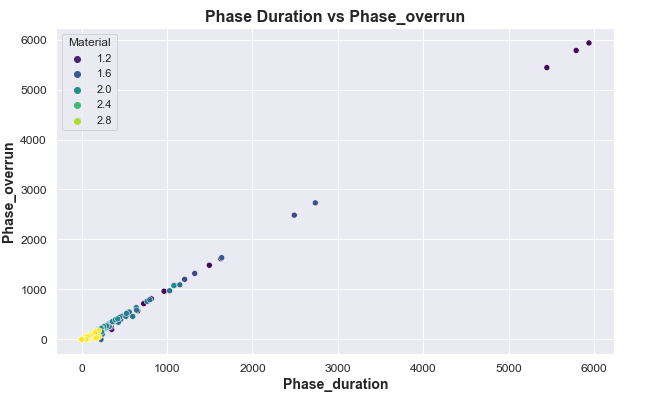
The main objective of this research is to determine the downtime for the production tanks. The variable phase overrun is representative of the downtime that can be experienced by beverage mucilage batch production. Phase start delay is another variable that be measured as downtime.

The correlation heatmap was created to determine the relationships between the variables. The values closest to I have the strongest positive correlation which is seen in the relationship between the phase overrun and phase duration variable, value of 0.98. Another close relationship highlighted is batch quantity and the flowrate, 0.52.

To further examine the relationship between the phase duration and phase overrun, a scatterplot was created, and linear relation was observed, figure 4



**Table 5. Correlation Heatmap**



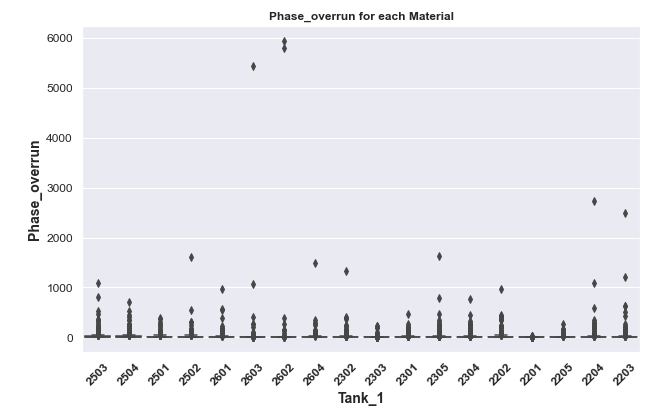
**Figure 4: Phase Duration vs Phase Overrun for all materials produced in all Tanks.**

In conclusion there is a strong relationship between the phase overrun and phase duration. Given the high correlation between phase duration and phase overrun, a linear regression model might be a good starting point. This allows prediction of the phase overrun based on the phase duration.

In comparison there is only a moderate relationship between batch Quantity and flowrate, value of 0.52, it indicates that there are other factors affecting the tank flowrate.

**5.4.2.4 Handling Outliers**

To determine the presence of outliers, boxplots were utilised. Figure 7 displays the distribution of data overall the production tanks for the phase overrun variable. There is the presence of outliers for each tank, with points present outside the range. Under each investigation of the selected phases and production tanks, outliers were assessed and removed prior to machine modelling. Table below shows a summary of the no. of outliers removed from each group production tanks for each batch produced.



**Figure 7: Boxplot of Phase Overrun for each production tank.**



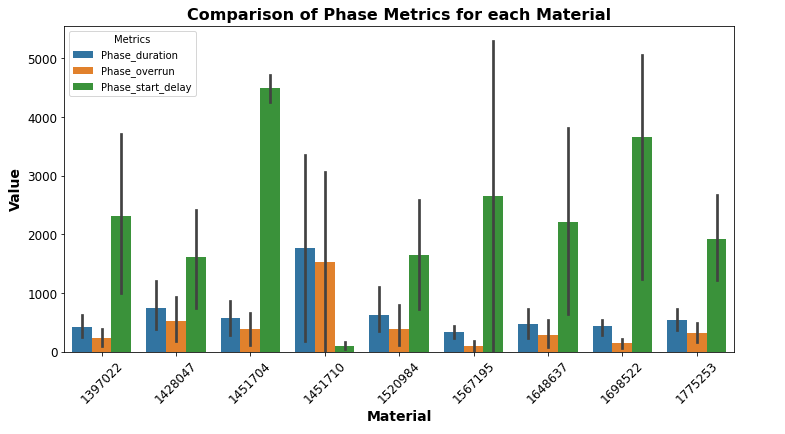
1. **Exploratory Data Analysis for 22 MT. production Tanks**
2. **Model Evaluation for Predicting Phase Overrun for 22 MT. production Tanks for the All phases.**

|  |  |
| --- | --- |
| **Evaluation Details** | |
| **Instruction Step** | **All phase** |
| **Production Batches** | 73 |
| **Production Batches after Outlier removal** | 59 |
| **Target Variable** | Phase Overrun |
| **Instruction steps/Phases per batch** | 27 |

NB: For each batch produced, there are two production tanks used and thus included in this analysis. One for the actual production and the other for storage as in the destination which has itself a phase step and a phase overrun metric.

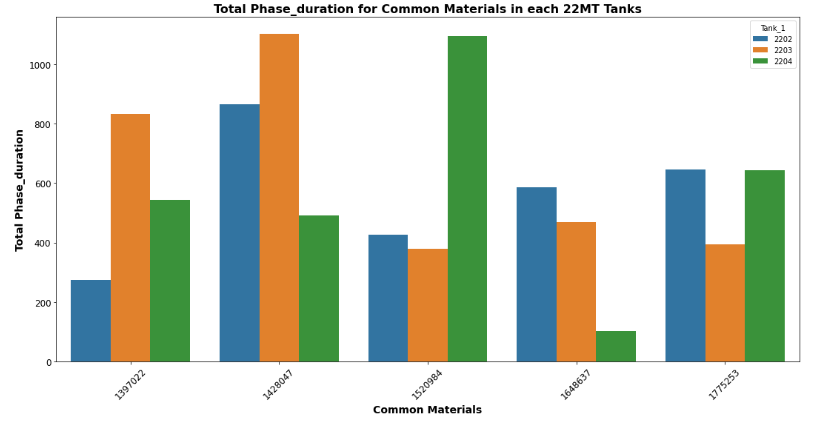
**Univariate Analysis:Numerical Features**:

The comparison of Phase Metrics for each Material for the 22MT TanksThis histogram gives a performance overview of each of the 22MT tanks, looking at common materials that were produced. Each material produced in the 22MT tanks experienced phase start delay, phase overrun, and the production time was different between all tanks.

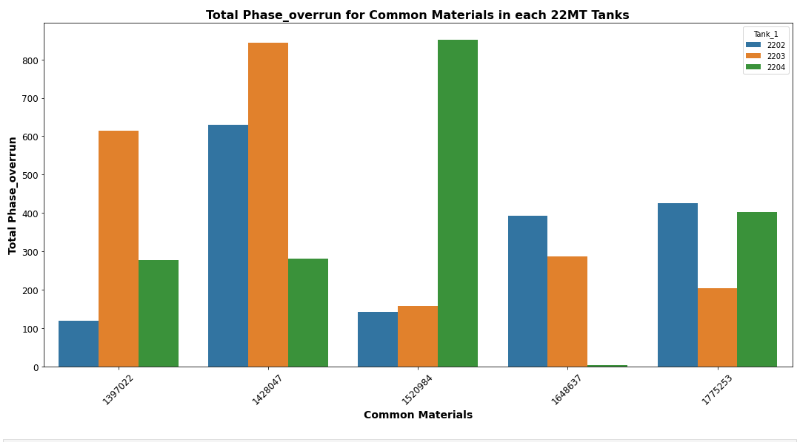


**F****igure 8: Bar chart of 3 phase matrix phases for each material in the 22MT production tanks.**

The phase duration and phase overrun for common materials in each 22 MT tanks are shown in figure 9 and 10 below, its shows that there is difference in production time for materials between all 3 production tanks. For example, for material 1297022, the histogram shows that it would be better to produce in tank 2202 as the phase duration is the lowest and the phaser overrun time is lowest for this tank. Important to note, that for each of the materials there is downtime logged, phase overrun,

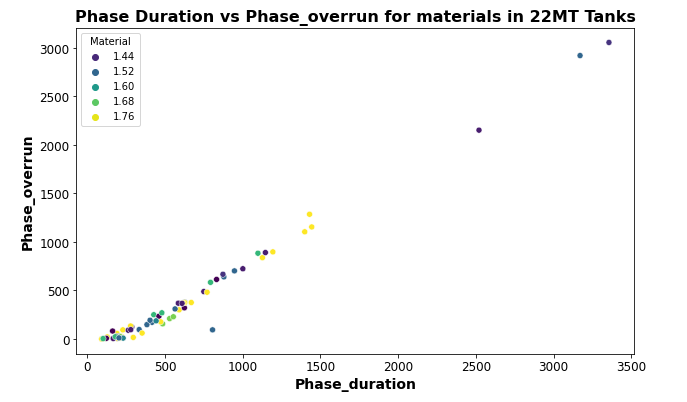


**Figure 9: Bar chart of phase duration phases for each material in the 22MT production tanks.**



**Figure 10: Bar chart of phase overrun phase for each material in the 22MT production tanks.**

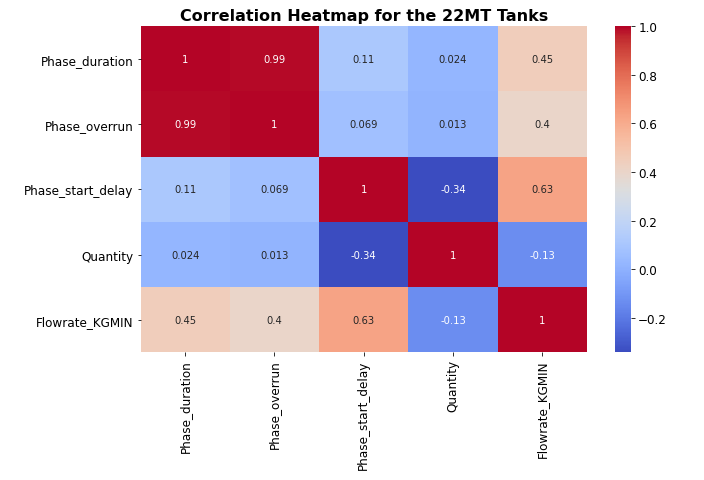
Figure 11 highlights the linear relationship between the target variable phase overrun and phase duration time for all 22 MT production tanks.



**Figure 11: Scatter chart of phase overrun phase vs phase duration 22MT production tanks.**

**Bivariate and Multivariate Analysis:**

**Correlation Analysis**:

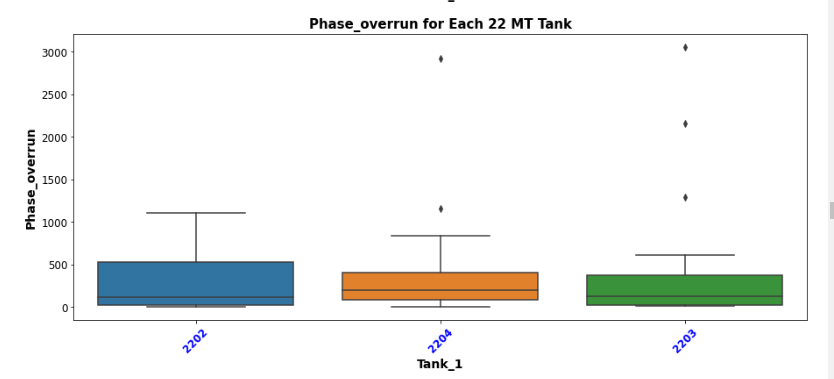


**Figure 12: Correlation analysis heatmap for 22MT production tanks.**

There was high correlation between the Phase overrun and phase duration variables, 0.99.

**Handling outliers**

Figure 13 a boxplot of the phase overrun batch data for each of the 22MT production tanks. There is the presence of points above the highest observations, the horizontal line call whiskers. Also majority of data are distributed above the median line for all three 22 MT tanks



**Figure 13: Boxplot for each 22MT production tanks. For the target variable Phase overrun**

The Interquartile range (IQR) method was used to remove outliers from the production tank data.

There were 14 rows of production data removed from the dataframe as potential outliers.

**Data standardization**



Based on the descriptive statistics for each of the production tank batch data, there was wide variation in the data for phase duration, phase\_start\_delay and the phase overruns., which can affect certain algorithms. Based on this, the data was standardised using the Z-score normalization using StandardScaler from Sklearn.preprocessing library. Another advantage is that its consistent and highly interpretable.

**Applying Machine Learning Models**

Various Machine learning models were applied to the dataframe, reference: Production 22\_df1. Full details of these are in the appendix.

Results for each of the investigations below show the top and poor performing models. Visual results of the top performing models are given in the predicted vs actual graphs and the residual plots.

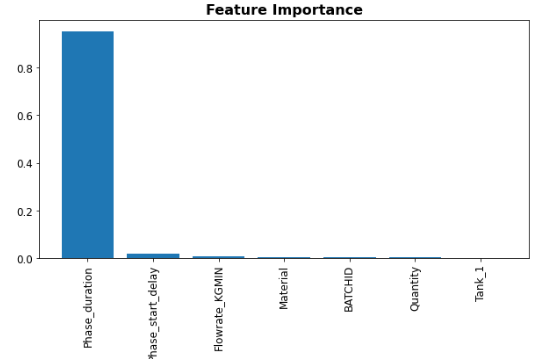
1. **Model Evaluation for Predicting Phase Overrun for 22 MT. production Tanks for the All phases.**

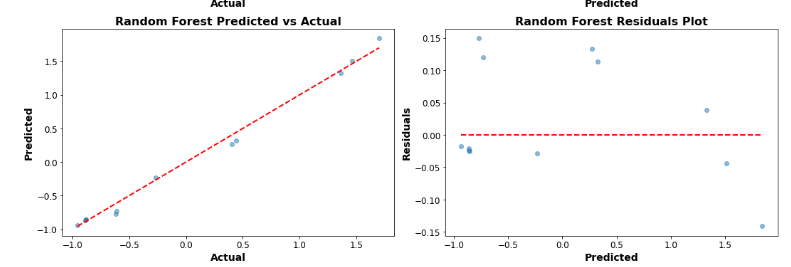
|  |  |
| --- | --- |
| **Evaluation Details** | |
| **Instruction Step** | **All phase** |
| **Production Batches** | 73 |
| **Production Batches after Outlier removal** | 59 |
| **Target Variable** | Phase Overrun |
| **Instruction steps/Phases per batch** | 27 |

NB: For each batch produced, there are two production tanks used and thus included in this analysis. One for the actual production and the other for storage as in the destination which has itself a phase step and a phase overrun metric.



Random Forest Regressor achieves almost perfect scores on both the training and testing datasets before applying tuning or cross validation. This regressor model provides stability and is less likely to overfit. Another of its advantage is interpretability of feature importance, the graph below shows which feature phase duration as the driver of the predictions. The scatterplot and residual plot show good model performance, with no obvious outlier points.







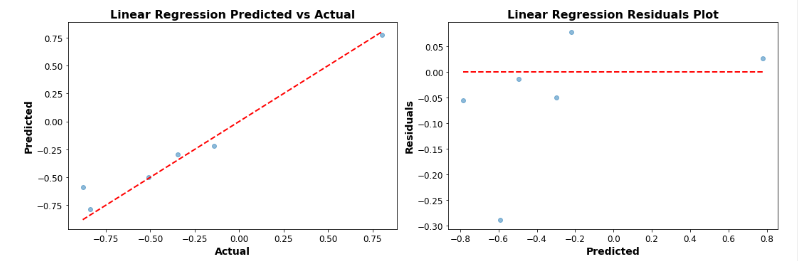
The poorest performing model was the LSTM Neural Network with a poor accuracy score R2 of -0.14 for R^2. This is highly likely due to the target variable not been a time series or text format. LSTM require a lot of data for the training, and this could affect the hyperparameter choices as the top performing models are much simpler models, the complexity of using LSTM is redundant.

1. **Model Evaluation for Predicting Phase Overrun for 22 MT. production Tanks for the deaeration phases.**

|  |  |
| --- | --- |
| **Evaluation Details** | |
| **Instruction Step** | **deaeration phase** |
| **Production Batches** | 38 |
| **Production Batches after Outlier removal** | 30 |
| **Target Variable** | Phase Overrun |



Given the results, the linear regression model, a perfect accuracy score R2 of 1.00 on training data and a score of 0.94 on test data. This indicates that this model can explain 100% of the variance in the training dataset and 94% in the test dataset. This indicates that these models can explain 100% of the variance in the training dataset and 94% in the test dataset, reflecting an excellent fit. Furthermore, the Mean Squared Error (MSE) for these models is 0.00 for both training and test datasets, showcasing the model's precision in estimating phase overrun in production tanks. The consistent performance in both tuned and untuned metrics shows the strength of linear regression model in capturing the linear relationships in the given dataset for production tank group 22MT, making it a prime choice for predicting phase overrun for the deaeration production phase. The accuracy is shown in the following graphs





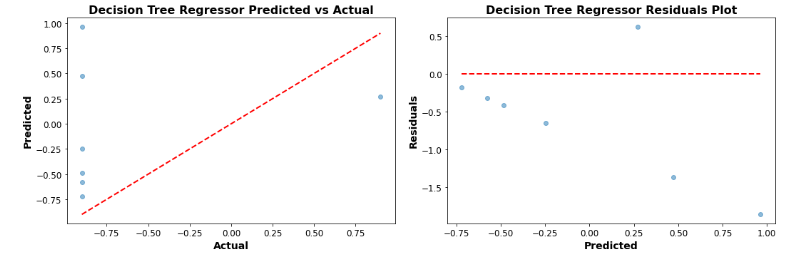
The Long Short-Term Memory (LSTM) Neural Network, a type of Recurrent Neural Network (RNN), presented mixed results when predicting phase overrun in production tanks. The training values R2 were notably negative, standing at -0.91 and -0.79 for the untuned and tuned models respectively. This indicates that the LSTM model failed to capture the underlying patterns in the training data. On the test side, the values further declined to -1.24 and -1.10 for the untuned and tuned models, respectively, suggesting the model's predictions were worse than a basic horizontal line mean prediction. Additionally, the high Mean Squared Error (MSE) values, 1292.83 for training and 424.13 for testing in the untuned model, underscore its lack of precision. While tuning did bring about some improvements, they were marginal. In this context, the LSTM model appears unsuitable for the task, possibly due to its inherent complexity and tendency to model sequential or time-dependent data, which might not align with the nature of the provided dataset.

**Model Evaluation for Predicting Phase Overrun for 22 MT. production Tanks for the Agitation phase phases.**

|  |  |
| --- | --- |
| **Evaluation Details** | |
| **Instruction Step** | **Agitation phase** |
| **Production Batches** | 47 |
| **Production Batches after Outlier removal** | 34 |
| **Target Variable** | Phase Overrun |



The Decision Tree Regressor top performing model when assessing the Agitation Phases for Production Tanks 22MT. Both the tuned and untuned models presented good results with a Train and Test score R2 of 1.00, indicating that the model perfectly explained the variance in the target variable. The Mean Squared Error (MSE) for both training and testing was impressively low at 0.00 and 0.01 respectively, highlighting the model's precision. These results suggest that the Decision Tree Regressor, without any need for parameter tuning, can capture the inherent patterns and relationships in the dataset with high accuracy.





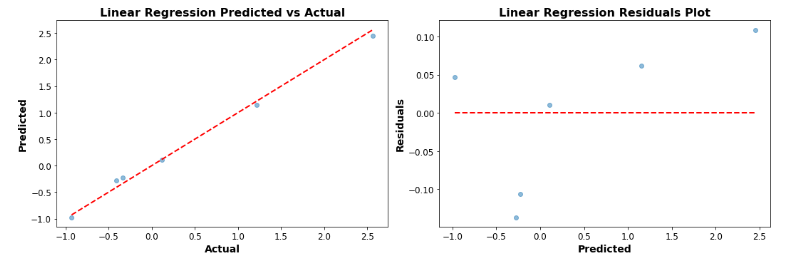
On the other hand, the Dense Neural Network (FCN) displayed the weakest performance among all with a R 2 -of0.72 which declined significantly for the test data, registering a poor R 2 of -2.14. Tuning the model parameters led to some improvements in the training. R 2, but not the test scores. The negative values, R 2, especially for the test data, point towards the model's inability to predict the phase overrun in Production Tanks 22MT reliably. Moreover, the high MSE values reinforce the model's lack of precision in its predictions. Despite the inherent capabilities of neural networks, in this context, the Dense Neural Network failed to generalize well to unseen data.

**Model Evaluation for Predicting Phase Overrun for 22 MT. production Tanks for the Gum addition phase**

|  |  |
| --- | --- |
| **Evaluation Details** | |
| **Instruction Step** | **Gum addition phase** |
| **Production Batches** | 43 |
| **Production Batches after Outlier removal** | 29 |
| **Target Variable** | Phase Overrun |



For the Gum Addition process in Production Tanks 22 MT, the Linear Regression model demonstrated superior performance. Both the tuned and untuned variants exhibited excellent predictive capabilities, as evident from the Train and Test R 2 scores which consistently hovered around 0.98 and 0.99 respectively. These scores suggest that the model was able to explain almost 99% of the variance in the test data. The MSE values further bolster the model's credibility with figures as low as 0.02 for training and 0.01 for testing, indicating accurate predictions with minimal error. In this context, Linear Regression, without any need for parameter tuning, seems to adeptly capture the underlying relationships in the dataset, offering a reliable model for predicting outcomes in the Gum Addition process. This is visualised in the graphs below.

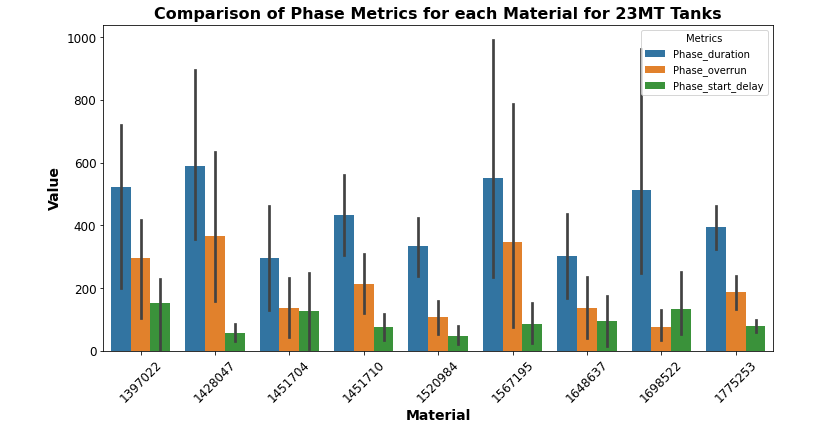




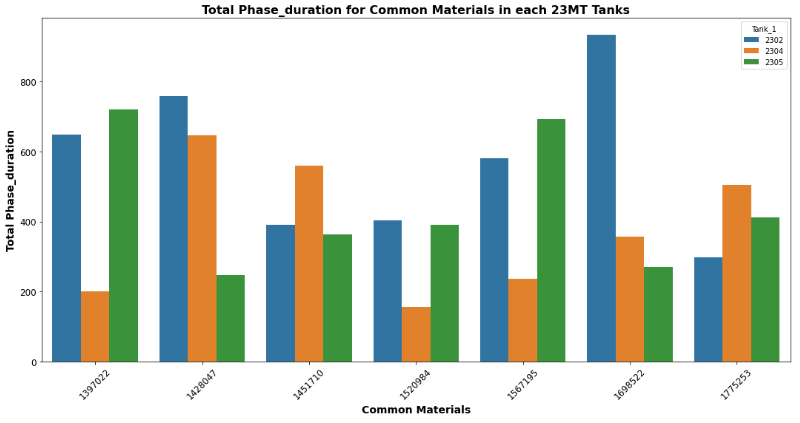
On the opposite end of the spectrum, the Simple Neural Network model proved to be the least efficient in predicting outcomes for the Gum Addition process. While the untuned model presented a discouraging Train R 2of -0.71, the situation worsened after tuning, dropping the score to -0.47. This negative indicates that the model's predictions are worse than simply predicting the mean of the target variable. The test data didn't fare much better with R 2scores of -1.10 and -0.82 for the untuned and tuned models respectively. Furthermore, the high MSE values, especially in the test data, emphasize the model's imprecision. Despite the inherent power of neural networks, in this specific scenario, the Simple Neural Network seems ill-equipped to generalize or make accurate predictions.

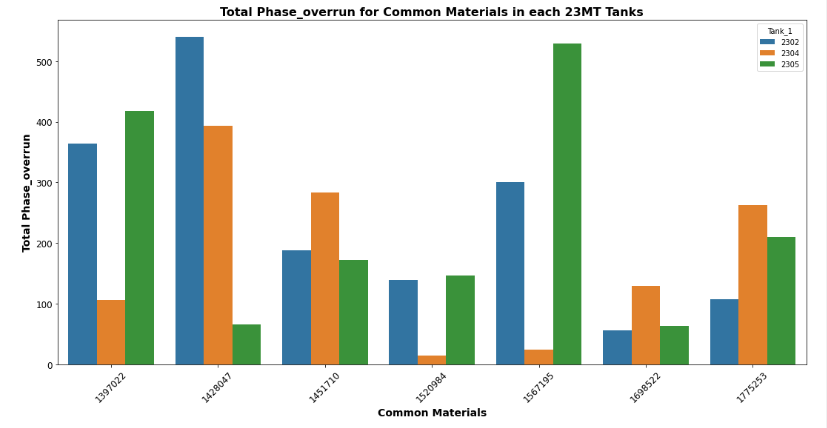
**Exploratory Data Analysis Production Tank 23MT**

**Univariate Analysis: Numerical Features**:



For all the materials produced in the 23 MT production tanks , the bar chart above shows that there was little delay in the starting of the production phases for all tanks , however again there was phase overrun downtimes noted

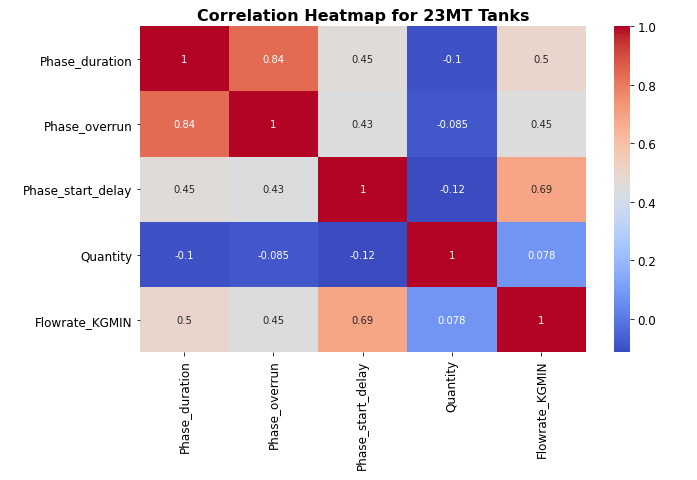




Bar Chart giving visual representation of how phase duration and phase overrun varies for the common materials produced across the 23MT production tanks. Each bar represents a tank, and the height is the phase metric for that tank. There is a lot of variability, and each material would have to reviewed separately, but for material 1567195, it seems that tank 23MT 04 would be the better tank to use, as the phase overrun times were lower in this tank. In general, for tanks that have the same capacity and materials with the same quantity and ingredient addition, there is a lot of variation in the production time and the down time.

**Bivariate and Multivariate Analysis:**

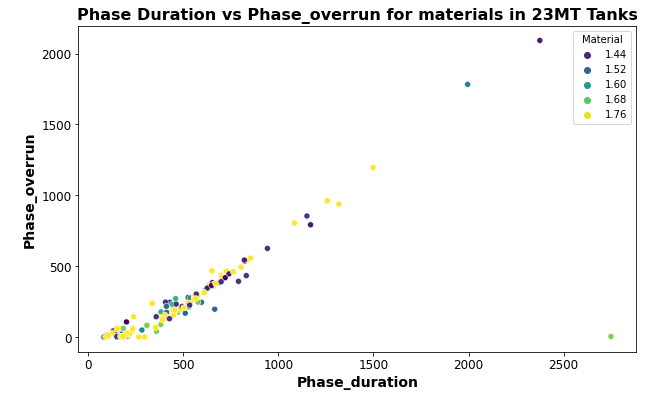
**Correlation Analysis**: A correlation matrices for numerical features to determine if there were any linear relationships between the variables.



The correlation heatmap was created to determine the relationships between the variables. The values closest to I have the strongest positive correlation which is seen in the relationship between the phase overrun and phase duration variable, value of 0.84. Another close relationship highlighted is phase start delay and the flowrate, 0.69.

To further examine the relationship between the phase duration and phase overrun, a scatterplot was created, and linear relation was observed, figure 4

**Pair Plots & Scatter Plots**: For visual inspection of relationships and trends between pairs of numerical features. Looking at the relationship between the phase duration vs phase overrun times , which is linear as per the results of the correlation chart



**Handling outliers#**

**Boxplots**

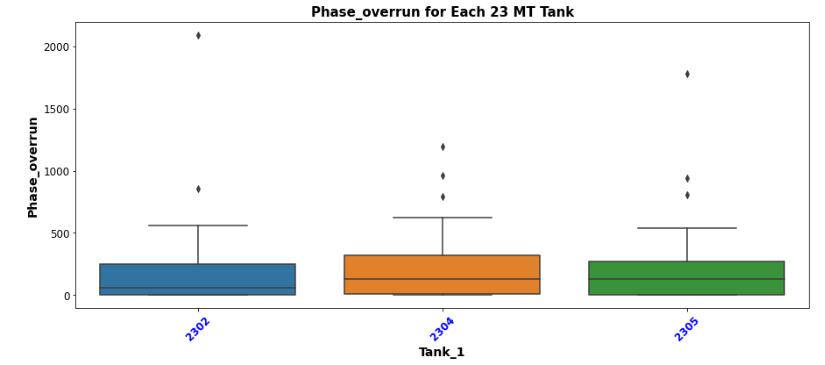


Figure 13 a boxplot of the phase overrun batch data for each of the 23MT production tanks. There is the presence of points above the highest observations, the horizontal line call whiskers. Also, majority of data are distributed above the median line for all three 23 MT tanks. The Interquartile range (IQR) method was used to remove outliers from the production tank data. There were 7 rows of production data removed from the dataframe as potential outliers.

**Data standardization**



Based on the descriptive statistics for each of the production tank batch data, there was wide variation in the data for phase duration, phase\_start\_delay and the flowrates.

These wide variations could impact the performance of certain machine learning models, a single feature disproportionately in scale could influence models that use distance to compute. Based on this, the data was standardised using the Z-score normalization using StandardScaler from Sklearn.preprocessing library. Another advantage is that its consistent and highly interpretable.

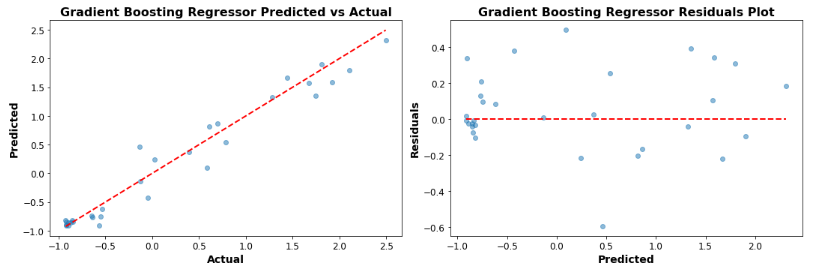
**Model Evaluation for Predicting Phase Overrun for 23 MT. production 23 Tanks for the Agitation phase.**

|  |  |
| --- | --- |
| **Evaluation Details** | |
| **Instruction Step** | **All phase** |
| **Production Batches** | 162 |
| **Production Batches after Outlier removal** | 39155 |
| **Target Variable** | Phase Overrun |
| **Instruction steps/Phases per batch** | 27 |

NB: For each batch produced, there are two production tanks used and thus included in this analysis. One for the actual production and the other for storage as in the destination which has itself a phase step and a phase overrun metric.



The Gradient Boosting Regressor is the star performer for the Production Tanks 23 MT dataset. Even without tuning, it exhibits a pristine R 2score of 1.00 on the training set, indicating a flawless fit. Upon tuning, this model maintains its high fidelity on the training data and demonstrates robust predictive power on the test set, achieving an R 2score of 0.95. The optimal parameters that accentuate its performance include a learning rate of 0.2, a max depth of 3, and 300 estimators. This model's ability to systematically build trees by adjusting to the errors of the previous ones gives it an edge in capturing intricate data patterns.





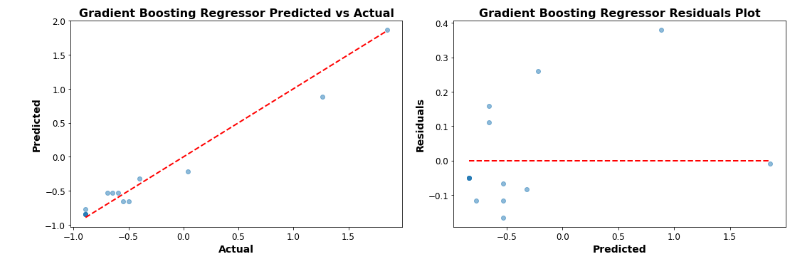
The K-Nearest Neighbors (KNN) model is considered the poorest performer because, despite its perfect training score of 1.0 after tuning, it failed to replicate this high performance on the test data, only achieving a score of 0.70. This large discrepancy between the training and test scores indicates potential overfitting. Overfitting occurs when a model performs exceptionally well on the training data but poorly on new, unseen data. The model becomes too tailored to the specific details and noise of the training set, making it less generalizable to new data. In practical applications, a model's performance on test or unseen data is more crucial than on the training data, hence why the KNN's subpar test score positions it as a weak performer in this context.

**Model Evaluation for Predicting Phase Overrun for 23 MT. production Tanks for the Deaeration phase.**

|  |  |
| --- | --- |
| **Evaluation Details** | |
| **Instruction Step** | **Deaeration phase** |
| **Production Batches** | 83 |
| **Production Batches after Outlier removal** | 67 |
| **Target Variable** | Phase Overrun |
|  |  |



The Gradient Boosting Regressor stands out as the best performer. It demonstrates near-flawless results with a Train R 2 score of 1.00 and a Test R 2 score of 0.99. After tuning with a learning rate of 0.1, a maximum depth of 3, and 200 estimators, it consistently maintains its top-tier performance, showing its robustness and adaptability.



The Gradient Boosting Regressor stands out as the best performer. It demonstrates near-flawless results with a Train R 2 score of 1.00 and a Test R 2 score of 0.99. After tuning with a learning rate of 0.1, a maximum depth of 3, and 200 estimators, it consistently maintains its top-tier performance, showing its robustness and adaptability.



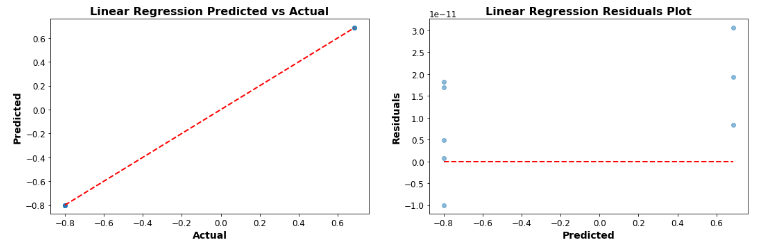
The K-Nearest Neighbors (KNN) model falls short in its test performance. While it achieves a perfect Train R 2 score of 1.00 after tuning, it plummets to a Test R 2score of just 0.4. This indicates that the model might be overfitting the training data. Despite an exhaustive tuning process involving 160 fits to pinpoint the best parameters, like using the 'auto' algorithm, selecting 3 neighbours, and distance-based weights, the model doesn't translate its training success to the test set. This stark difference emphasizes the importance of a model's ability to generalize beyond its training data.

**Model Evaluation for Predicting Phase Overrun for 23 MT. production 23 Tanks for the Agitation phase.**

|  |  |
| --- | --- |
| **Evaluation Details** | |
| **Instruction Step** | **Agitation phase** |
| **Production Batches** | 82 |
| **Production Batches after Outlier removal** | 39 |
| **Target Variable** | Phase Overrun |
|  |  |



The **Linear Regression model** stands out with perfect results. It achieves a Train R 2score and a Test R 2score both of 1.00, indicating that it can explain 100% of the variance in the target variable for both training and test datasets. This is mirrored in the extremely low Mean Squared Error (MSE) values, which approach zero. No tuning was necessary for this model, suggesting that a simple linear relationship was likely sufficient to capture the patterns in the data.





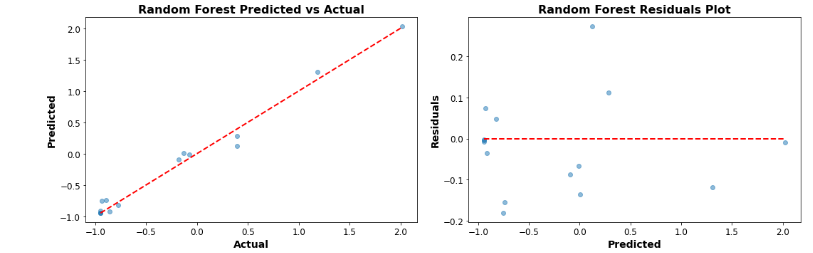
The **Lasso Regression** model has shown the least promising performance before tuning. Its Train R 2 score of 0.17 and Test R 2score of 0.02 suggests that it struggled to capture the variance in the target variable initially. However, after tuning with an alpha value of 0.01, the performance improved drastically to an R 2score of 1.00 for both training and test sets. This indicates a significant improvement, but the stark difference in performance before and after tuning might raise concerns about the model's robustness and its ability to generalize across different datasets.

**Model Evaluation for Predicting Phase Overrun for 23 MT. production Tanks Gum Addition phase**

|  |  |
| --- | --- |
| **Evaluation Details** | |
| **Instruction Step** | **Gum Addition phase** |
| **Production Batches** | 82 |
| **Production Batches after Outlier removal** | 73 |
| **Target Variable** | Phase Overrun |

:





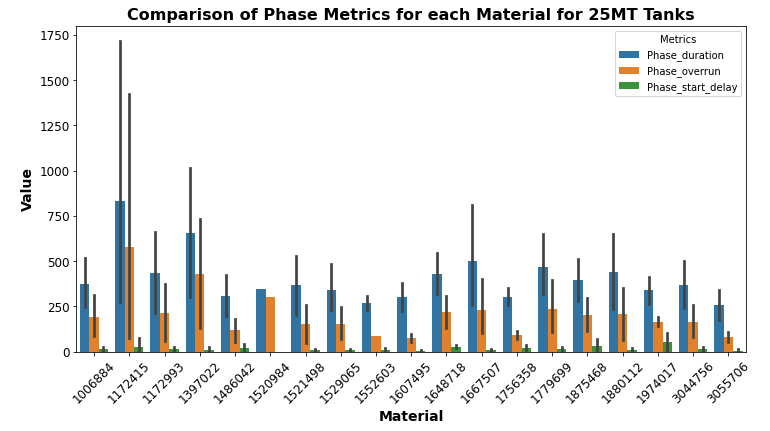
The **Random Forest Regressor** stand out with perfect results. They achieve a Train R2 score and a Test R2 score both of 1.00, which suggests that these models can capture almost all the variance in the target variable for both training and test datasets. The very low MSE values, approaching zero, further confirm this excellence in prediction. The best parameters indicate that for the Random Forest, the optimal number of trees (estimators) is 200 and there's no restriction on depth. These models seem well-suited for predicting phase overrun during the GUM addition phase.



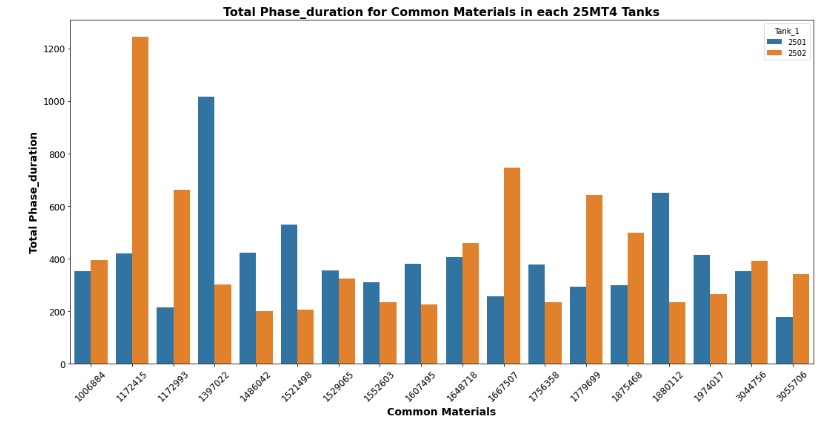
The **Lasso Regression** model initially performed poorly with a Train R2 score of 0.10 and a negative Test R2 score of -0.13, indicating that the model was worse than a horizontal line. However, after tuning with an alpha value of 0.01, the performance dramatically improved to an R2 score of 0.99 for training and 0.97 for testing. Despite this improvement, the initial poor performance might cause concerns about the reliability and robustness of this model for this specific dataset.

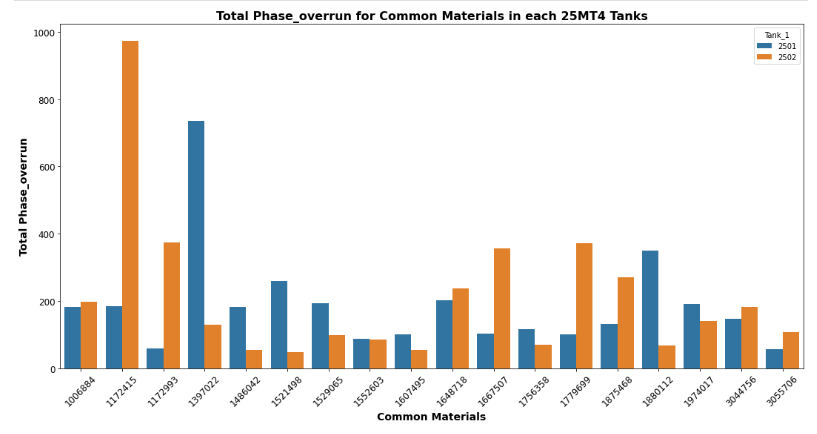
**Exploratory Data Analysis Production Tank 25MT 4:**

**3.4.1 Univariate Analysis:Numerical Features**:



1. The distribution of Material data records for each tank and allows you to see which tanks produced materials in the dataset. There is phase overruns for every material produced in these tanks ,

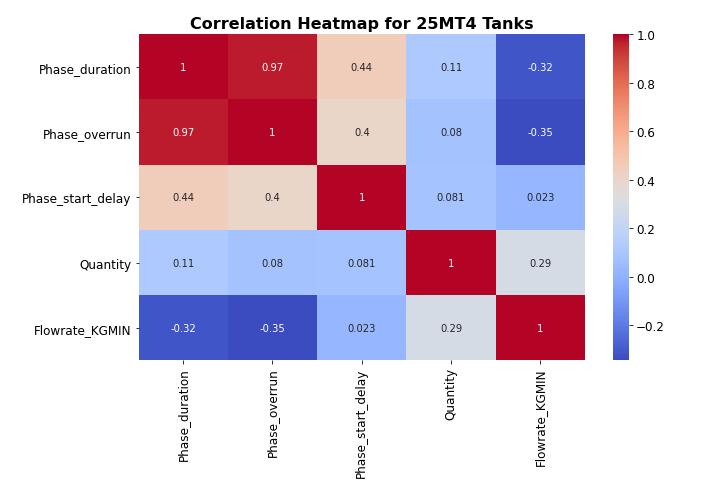




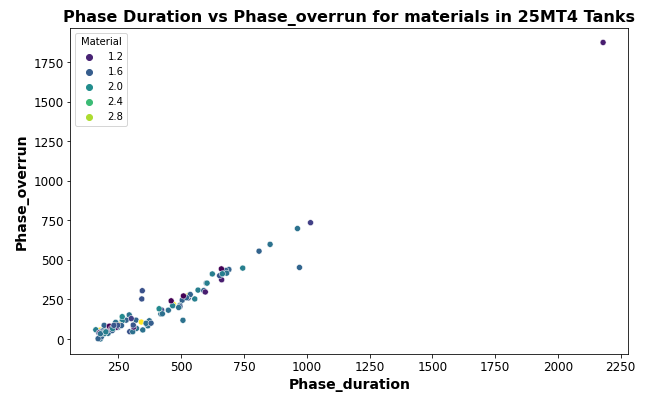
Bar Chart giving visual representation of how phase duration and phase overrun varies across the different production tanks. Each bar represents a tank, and the height is the phase metric for that tank. Apart from 2 high phase overrun values in both tanks , the rest of the materials produced have low phase overrun values

**Bivariate and Multivariate Analysis:**

**Correlation Analysis**: Compute correlation matrices for numerical features to determine if there were any linear relationships between the variables.A high correlation between phase overrun and phase duration variables.

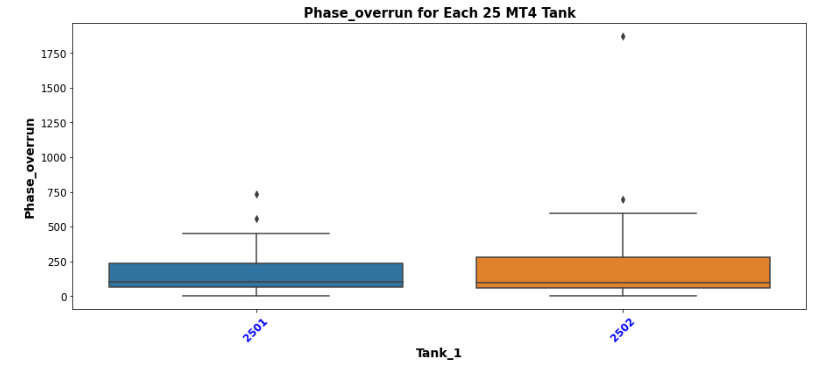


**Pair Plots & Scatter Plots**: For visual inspection of relationships and trends between pairs of numerical features of phase overrun and phase duration



**Handling outliers#**

**Boxplots**



Based on the boxplots in the EDA, outliers were determined to be presence in the data for both tanks of the 25 MT4

**Data standardization**



Based on the descriptive statistics for each of the production tank batch data , there was wide variation in the data for phase\_duration , phase\_start\_delay and the flowrates.

These wide variations could impact the performance of certain machine learning models, a single feature disproportionately in scale could influence models that use distance to compute. Based on this, the data was standardised using the Z-score normalization using StandardScaler from Sklearn.preprocessing library. Another advantage is that its consistent and highly interpretable.

**Model Evaluation for Predicting Phase Overrun for 25 MT. 4 production Tanks All phases**

|  |  |
| --- | --- |
| **Evaluation Details** | |
| **Instruction Step** | **All phases** |
| **Production Batches** | 98 |
| **Production Batches after Outlier removal** | 81 |
| **Target Variable** | Phase Overrun |
| **Instruction steps/Phases per batch** | 27 |

:

NB: For each batch produced, there are two production tanks used and thus included in this analysis. One for the actual production and the other for storage as in the destination which has itself a phase step and a phase overrun metric.



The Random Forest Regressor stood out as the star performer for predicting phase overrun across all production phases in Production Tanks 25MT. It showcased superior accuracy with an R2score of 1.00 on the training data and an impressive 0.98 on the test data. These metrics reflect the model's excellent capacity to capture the underlying patterns in the data, positioning it as an ideal choice for predicting phase overruns in this context.





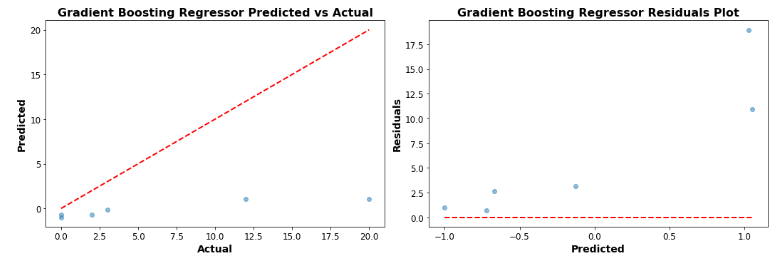
Conversely, the LSTM Neural Network struggled significantly in this predictive task. With an R2score of -0.57 on the training set and an even more concerning -1.26 on the test set, the model displayed a conspicuous inability to make accurate predictions for the given data. This underwhelming performance suggests that the LSTM, in its current configuration, might not be suitable for predicting phase overruns for Production Tanks 25MT.

**Model Evaluation for Predicting Phase Overrun for 25 MT. 4 production Tanks Deaeration phase.**

|  |  |
| --- | --- |
| **Evaluation Details** | |
| **Instruction Step** | **Deaeration phase** |
| **Production Batches** | 49 |
| **Production Batches after Outlier removal** | 30 |
| **Target Variable** | Phase Overrun |



In the Deaeration Phase of Production Tanks 25MT4, the Gradient Boosting Regressor is the top performer. Displaying minimal error with an almost perfect Train R2of 0.99 and a commendable Test R2of 0.63, it proves its capacity to understand and predict the underlying patterns effectively. Moreover, the model’s performance remains consistent after tuning, with the R2metric still closely hugging the 1.00 mark for training and a stable 0.63 for the test set. The best results were achieved with a learning rate of 0.2, a max depth of 3, and 100 estimators.





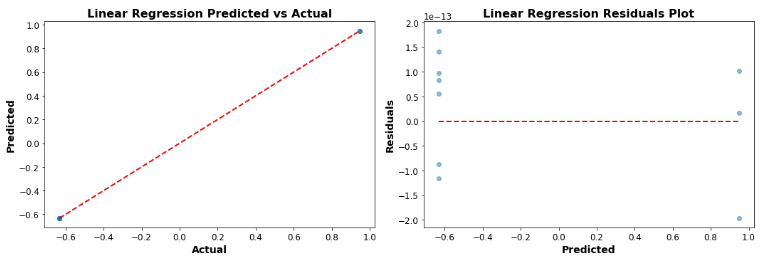
On the other hand, the LSTM Neural Network faced significant challenges during this phase. Originally presenting a negative Train R2 of -0.67 and a Test R2of -0.75, it demonstrates a lack of fit and poor prediction capabilities. While tuning did improve its performance, bringing the Train R2 up to 0.85 and Test R^2 to 0.77, the discrepancy between its original and tuned performances suggests potential overfitting issues or that the model might not be optimally designed for this task.

**Model Evaluation for Predicting Phase Overrun for 25 MT. 4 production Tanks Gum Addition Phase**

|  |  |
| --- | --- |
| **Evaluation Details** | |
| **Instruction Step** | **Agitation Phases** |
| **Production Batches** | 51 |
| **Production Batches after Outlier removal** | 50 |
| **Target Variable** | Phase Overrun |



In the Agitation Phase for Production Tanks 25MT 4, the Linear Regression model demonstrated top performance. Both its Train and Test Mean Squared Error (MSE) are practically at zero, showcasing the model's excellent capability in fitting the training data and generalizing to unseen data. A perfect R2score of 1.00 for both training and test sets, even after tuning, highlighting its predictive power.





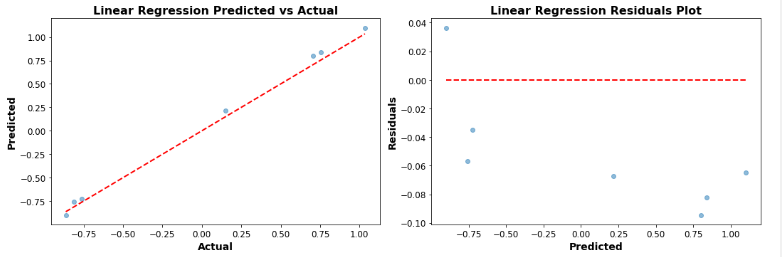
Conversely, the Instance-based K-Nearest Neighbors model underwhelmed in its performance for the Agitation phase. Its Train MSE was 0.43, and it only achieved a low R2score of 0.04 for training and an even more diminished 0.07 for testing. Tuning only marginally improved its R2score, reaching 0.13 for training and remaining at 0.04 for testing. The model's performance suggests that it struggled to find relevant patterns in the dataset for this phase, even after considering 9 neighbors and opting for uniform weighting.

**Model Evaluation for Predicting Phase Overrun for 25 MT. 4 production Tanks Gum Addition Phase**

|  |  |
| --- | --- |
| **Evaluation Details** | |
|  |  |
| **Instruction Step** | **Gum Addition** |
| **Production Batches** | 50 |
| **Production Batches after Outlier removal** | 35 |
| **Target Variable** | Phase Overrun |
|  |  |



In the Gum Addition Phase for Production Tanks 25MT 4, the Linear Regression model emerged as the top performer. The model demonstrated almost perfect fitting with a Train Mean Squared Error (MSE) of 0.00, which held consistently even after tuning. Its testing performance was equally outstanding, with a Test MSE of 0.00 both pre- and post-tuning. The R2score, a measure of the model's predictive power, achieved a flawless 1.00 for both training and testing datasets, highlighting its exceptional efficacy in this phase.



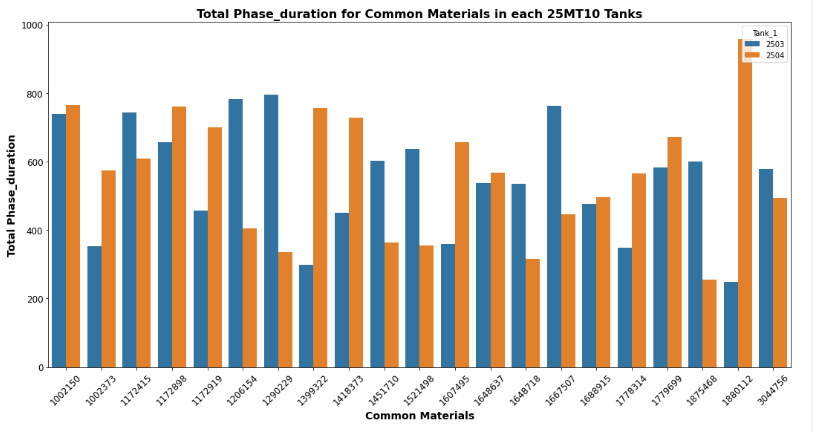


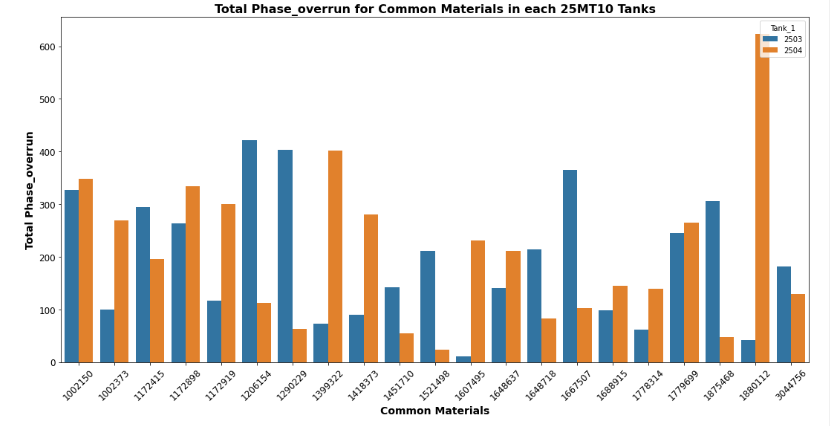
Neural Network (RNN) model using LSTM displayed poor performance. Before tuning, the Train MSE soared to an alarming 3116.96, and the Test MSE reached 2442.13. The R2score, registering at -0.81 for training and an even lower -1.64 for testing, indicated that the model's predictions were substantially worse than simplistic, mean-based predictions. Even after tuning, while there was a marginal improvement in the scores, the results remained unsatisfactory with the Test R2 still lingering at -1.19, underscoring the model's struggle to capture the inherent patterns of the Gum Addition Phase effectively.

**Exploratory Data Analysis : Production Tank 25MT 10**

**Univariate Analysis: Numerical Features**:

The distribution of Material data records for each tank and allows you to see which tanks produced materials in the dataset. The graph below shows the little difference in the total phase duration times between the 2 production tanks of 25 MT10

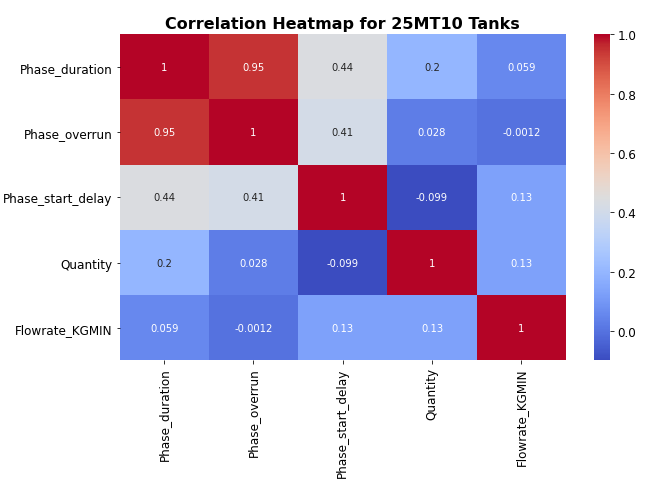




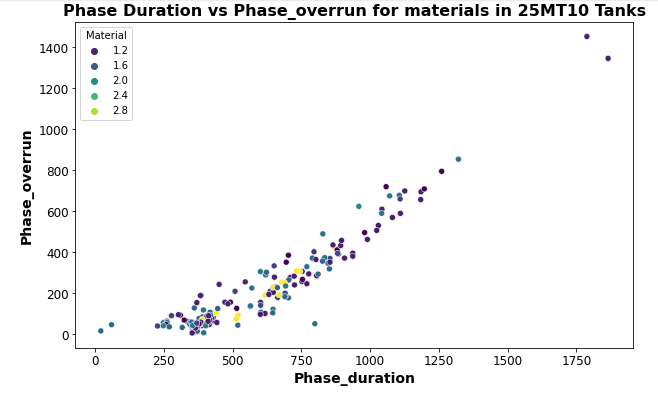
Bar Chart giving visual representation of how phase duration, phase overrun and phase start delay varies across the different production tanks. Each bar represents a tank, and the height is the phase metric for that tank. All materials exhibit Phase overruns in these tanks.

**Bivariate and Multivariate Analysis:**

**Correlation Analysis**: Compute correlation matrices for numerical features to determine if there were any linear relationships between the variables. Good correlation between the phase overrun and phase duration variables.

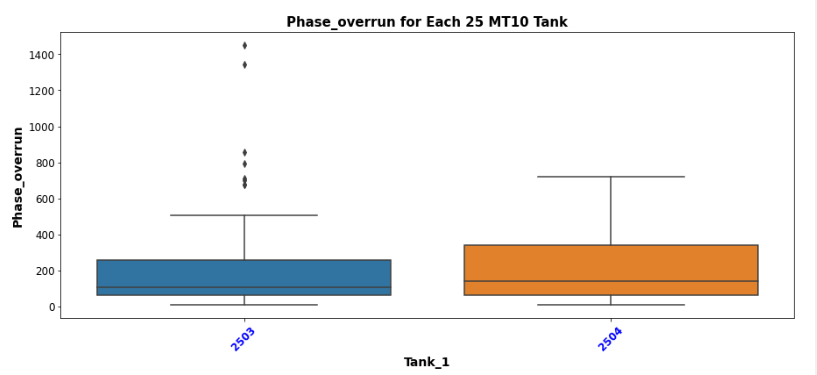


**Pair Plots & Scatter Plots**: For visual inspection of relationships and trends between pairs of numerical features. Good linear relationship observed between the phase duration and phase overrun variables



**Handling outliers**

**Boxplots**



Based on the above boxplots outliers were determined to be presence in the data

**Data standardization**



Based on the descriptive statistics for each of the production tank batch data , there was wide variation in the data for phase\_duration , phase\_start\_delay and the flowrates.

These wide variations could impact the performance of certain machine learning models, a single feature disproportionately in scale could influence models that use distance to compute. Based on this, the data was standardised using the Z-score normalization using StandardScaler from Sklearn.preprocessing library. Another advantage is that its consistent and highly interpretable.

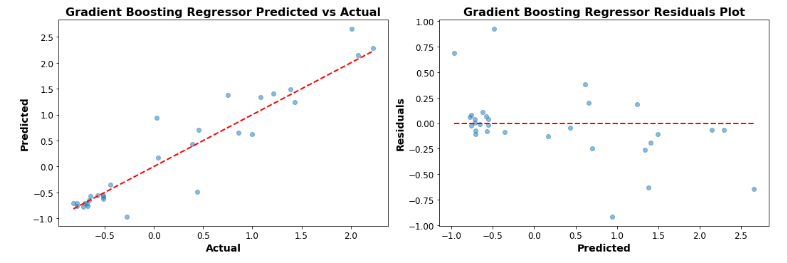
**Model Evaluation for Predicting Phase Overrun for 25 MT. 10 Production Tanks– All Phases**

|  |  |
| --- | --- |
| **Evaluation Details** | |
|  |  |
| **Instruction Step** | **All Phases** |
| **Production Batches** | 194 |
| **Production Batches after Outlier removal** | 150 |
| **Target Variable** | Phase Overrun |
| **Instruction steps/Phases per batch** | 27 |

NB: For each batch produced, there are two production tanks used and thus included in this analysis. One for the actual production and the other for storage as in the destination which has itself a phase step and a phase overrun metric.



The Gradient Boosting Regressor showcased best performance in the modeling for All Phases for Production Tanks 25MT 10. Before tuning, it had a negligible Train Mean Squared Error (MSE) of 0.001 and a Test MSE of 0.088. The model's ability to explain the variance in the data was almost flawless with a training R2score of 0.999 and a testing R2score of 0.905. After tuning, the model managed to achieve a perfect training R2score of 1.000, although with a slight drop in the test R2to 0.887. This was accomplished using a learning rate of 0.2, a maximum depth of 4, and 300 estimators.





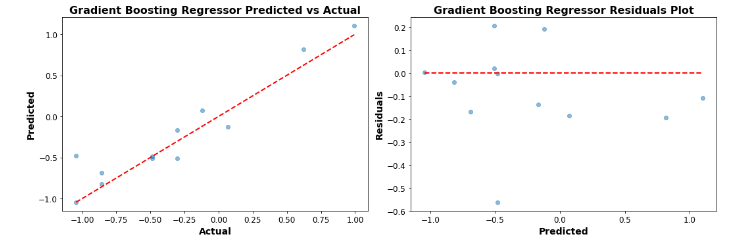
In stark contrast, the Neural Network (RNN) model using LSTM Neural Network displayed significant underperformance. Prior to tuning, it registered a massive Train MSE of 51437.30 and an even larger Test MSE of 63756.90. The R2scores were notably poor, with -0.98 for training data, indicating that the model's predictions were drastically worse than basic mean predictions. The test R2score was at a dismal -1.69. Despite tuning efforts, the results remained subpar with the Test R2improving only slightly to 0.19. The best parameters for the tuned model included 30 neurons for LSTM, 100 epochs, and a batch size of 64.

**Model Evaluation for Predicting Phase Overrun for 25 MT. 10 Production Tanks– Deaeration**

|  |  |
| --- | --- |
| **Evaluation Details** | |
|  |  |
| **Instruction Step** | **Deaeration** |
| **Production Batches** | 92 |
| **Production Batches after Outlier removal** | 58 |
| **Target Variable** | Phase Overrun |
|  |  |



The Gradient Boosting Regressor demonstrates good performance for the deaeration results of the Production Tanks 25MT10. The model, before tuning, delivered a Train MSE of 0.01 and a Test MSE of 0.04. The R2 scores were highly commendable with a score of 0.99 for training data, indicating the model's proficient ability to explain 99% of the variance. Its testing R2 score stood at 0.88. Remarkably, after tuning, the training R2 reached a perfect score of 1.00. The model achieved this exceptional performance with a learning rate of 0.01, a max depth of 4, and 300 estimators. This is also seen in the plots below with a good predictions and residuals.





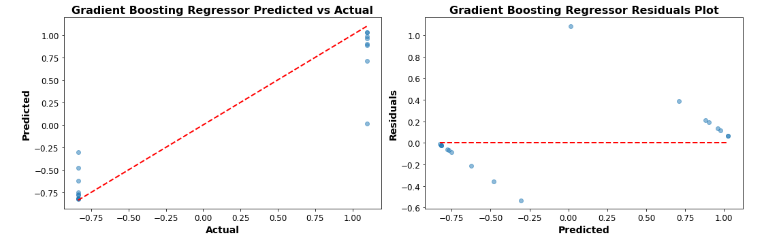
The LSTM Neural Network, a form of RNN, notably struggled in modelling the deaeration results. Pre-tuning, it registered a Train MSE of 54.79 and a much worse Test MSE of 19.09. The R2 values were concerning: -0.67 for training and -0.75 for testing, indicating the model's predictions were significantly worse than a naive mean-based approach. Fortunately, tuning improved the model to some extent, bringing the training R2up to 0.85 and the testing R2to 0.77. This improvement was accomplished with 50 LSTM neurons, 100 epochs, and a batch size of 16.

**Model Evaluation for Predicting Phase Overrun for 25 MT. 10 Production Tanks– Agitation**

|  |  |
| --- | --- |
| **Evaluation Details** | |
|  |  |
| **Instruction Step** | **Agitation** |
| **Production Batches** | 97 |
| **Production Batches after Outlier removal** | 90 |
| **Target Variable** | Phase Overrun |
|  |  |



The Gradient Boosting Regressor stands out as the top-performing model for the agitation phases of the Production Tanks 25MT10. Prior to tuning, it exhibited a Train MSE of 0.01 and a Test MSE of 0.09. Its R2scores are strikingly good, with a score of 0.99 for the training data, which indicates the model's superior capability to account for 99% of the variance. For testing data, it scored 0.90. After optimization, both the training and testing R2scores reached a perfect 1.00. The model achieved this stellar performance using a learning rate of 0.01, a max depth of 4, and 300 estimators.





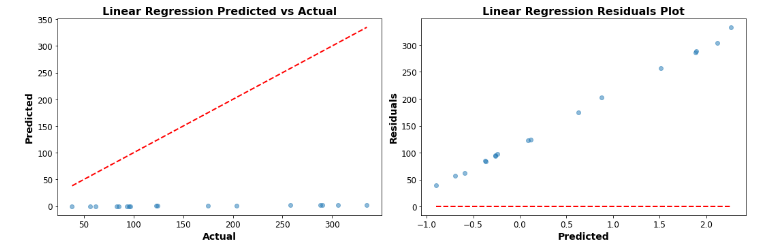
The Lasso Regression model, before tuning, struggled in this task. It reported a Train MSE of 0.90 and an even worse Test MSE of 1.10. The R2scores were not promising: 0.11 for the training set and -0.19 for the testing set. Negative R2 values, especially for the test set, suggest the model's predictions were considerably worse than a simplistic mean-based strategy. While tuning did improve the performance slightly, raising the training R2to 0.32 and the testing R2 to just -0.01, the results were still far from satisfactory. This improvement was possible due to an alpha value of 0.01.

**Model Evaluation for Predicting Phase Overrun for 25 MT. 10 Production Tanks– Gum Addition**

|  |  |
| --- | --- |
| **Evaluation Details** | |
|  |  |
| **Instruction Step** | **Gum Addition** |
| **Production Batches** | 96 |
| **Production Batches after Outlier removal** | 81 |
| **Target Variable** | Phase Overrun |
|  |  |



The top-performing model appears to be the **Linear Regression** under the "Linear" model type. This model exhibits a perfect Train and Test performance with both MSE (Mean Squared Error) and R2 (R-squared or Coefficient of Determination) values. Specifically, it has an MSE of 0.00 for both training and testing phases, and its R2 score stands at 1.00, indicating that the model perfectly predicts the outcomes. Even after tuning, the model retains its superior performance, demonstrating its robustness.

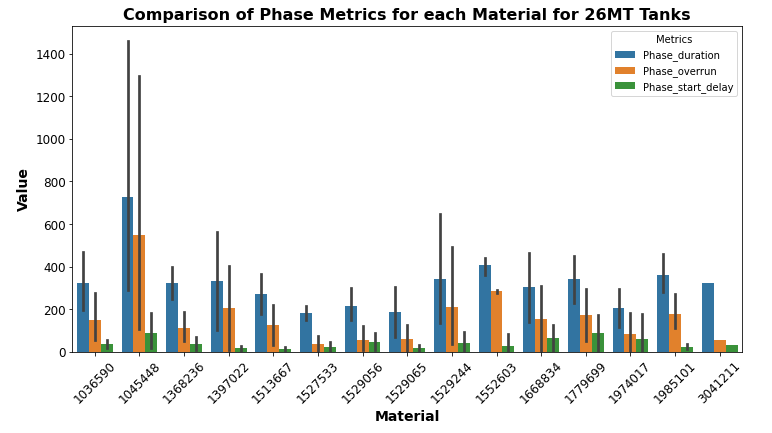




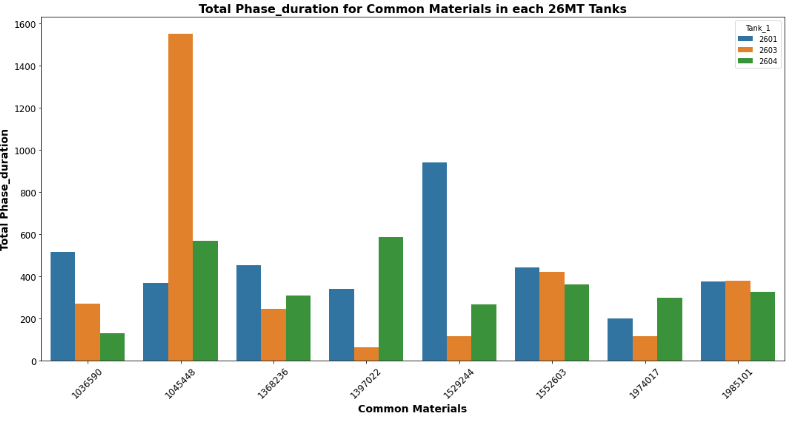
**LSTM Neural Network** under the "Neural Network (RNN)" category showed the poorest performance. Before tuning, it had a Train MSE of 18366.10 and a Test MSE of 34292.40. The R2 values for training and testing were -1.45 and -2.66 respectively. Even after tuning, though there was a reduction in the MSE values (Train MSE tuned to 11195.51 and Test MSE tuned to 20525.29), the R2 scores did not show significant improvement, with -0.50 for training and -1.19 for testing. The negative R2 values indicate that the model is performing worse than a simple horizontal line (mean-based model), signalling that the model might not be the right choice for this particular dataset or task.

**Exploratory Data Analysis** **for** **Production Tank group 26 MT** –

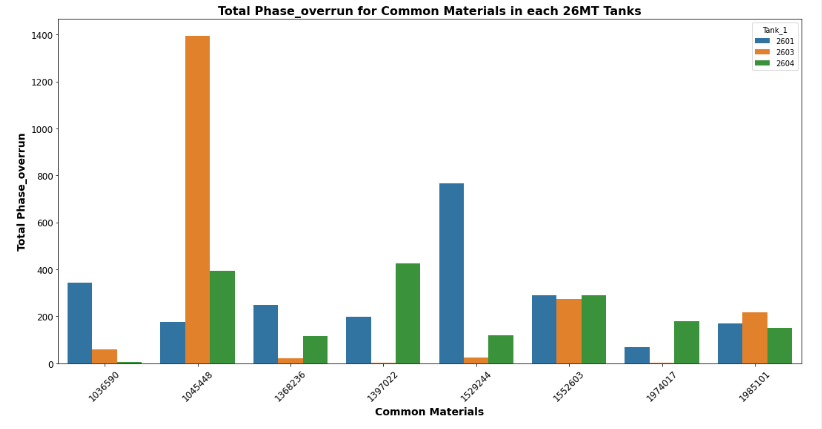
**Univariate Analysis:Numerical Features**:



There are production phase overruns for each material produced in all 26MT tanks. The phase start delay is the lowest metric for all tanks.

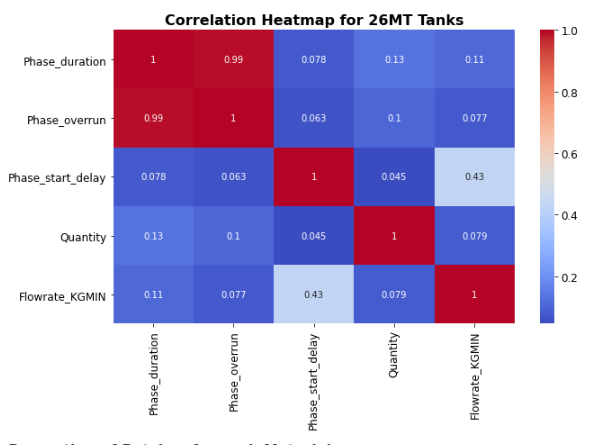


The phase overrun values for common materials between each 26 MT production tanks is seen in the following graph. Bar the unusual overrun value for tank 2603, it’s the tank 2601 that has higher phase overrun times for the common materials.

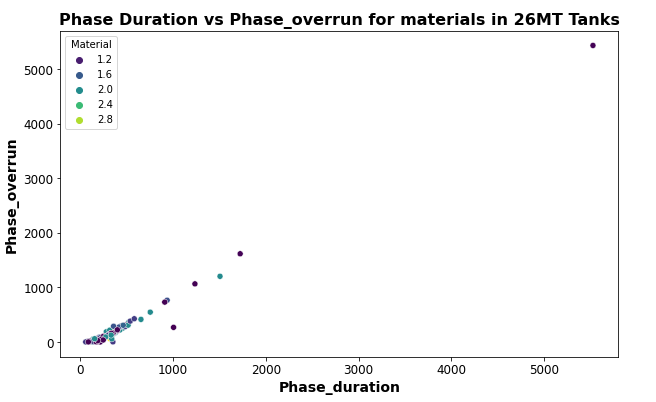


**Bivariate and Multivariate Analysis: Correlation Analysis**:

Compute correlation matrices for numerical features to determine if there were any linear relationship between the variables. The highest correlation is between the phase duration and the phase overrun variables.



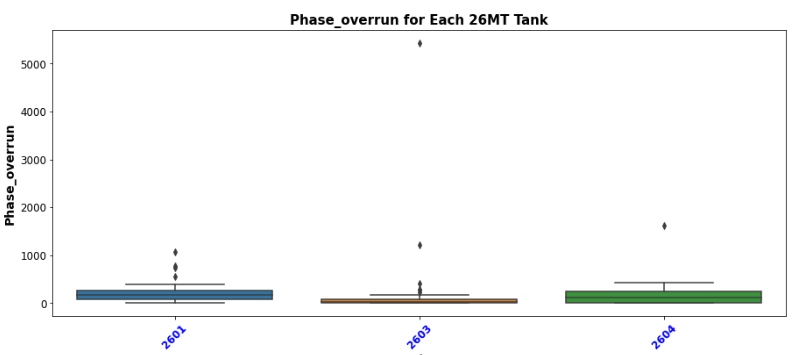
**Pair Plots & Scatter Plots**: For visual inspection of relationships and trends between pairs of numerical features. There is a linear relationship between the phaser overrun and phase duration variables in the dataset.



**Handling outliers#**

**Boxplots –**

There is the present of outliers in each of the 26 MT production tanks by the appearance of points about the observation line.



The outlier measurement and removal method used was based on the Interquartile Range (IQR). method is robust to extreme values and is preferred over methods like standard deviation-based outlier detection especially if there isn’t a gaussian data distribution.

**Data standardization**



Based on the descriptive statistics for each of the production tank batch data, there was wide variation in the data for phase\_duration , phase\_start\_delay and the flowrates.

These wide variations could impact the performance of certain machine learning models, a single feature disproportionately in scale could influence models that use distance to compute. Based on this, the data was standardised using the Z-score normalization using StandardScaler from Sklearn.preprocessing library.

**Machine Learning Results**

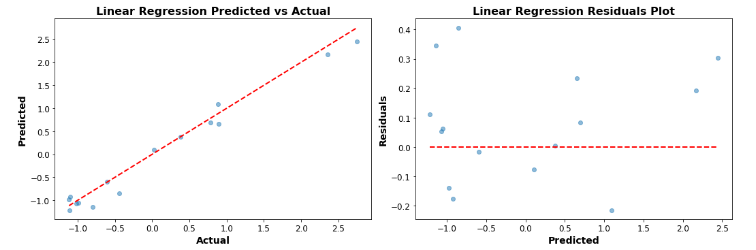
**Model Evaluation for Predicting Phase Overrun for 26 MT Tank for All Phases**

|  |  |
| --- | --- |
| **Evaluation Details** | |
|  |  |
| **Instruction Step** | All Phases |
| **Production Batches** | 46 |
| **Production Batches after Outlier removal** | 27 |
| **Target Variable** | Phase Overrun |
| **Instruction steps/Phases per batch** | 27 |

NB: For each batch produced, there are two production tanks used and thus included in this analysis. One for the actual production and the other for storage as in the destination which has itself a phase step and a phase overrun metric.



In the analysis of Production Tanks 26MT All Phases, the Linear Regression model is the better model. Before tuning, the model yielded a training *R*2of 0.96 and a test *R*2of 0.97, suggesting it could explain approximately 96% and 97% of the variance in the training and test datasets respectively. Post-tuning, the model further improved its test *R*2to an impressive 0.99, indicating an excellent fit to the unseen data and emphasizing its potential in predictive capacities for this dataset. The Mean Squared Error (MSE) metrics had values close to zero for both training and test sets after tuning. Overall, the Linear Regression model stands out as a potent tool for predictions in the context of Production Tanks 26MT All Phases, achieving high predictive accuracy and reliability this is shown in the below plots for the predicted vs actual values and the residual plots





The poorest performer is the LSTM Neural Network from the "Neural Network (RNN)" category. Its initial Train MSE is a 24976.80, and its Test MSE is even higher at 34481.50. Furthermore, the *R*2 values are deeply concerning: -1.35 for training and -0.91 for testing. These negative *R*2 scores imply that the model's predictions are far worse than a simple horizontal line, or a basic mean model. Even after tuning, while there's an improvement in the metrics, they're still quite poor. The best configuration for the LSTM model includes 70 LSTM neurons, 100 epochs, and a batch size of 16.

**Model Evaluation for Predicting Phase Overrun for 26 MT Tank for Deaeration Phase**

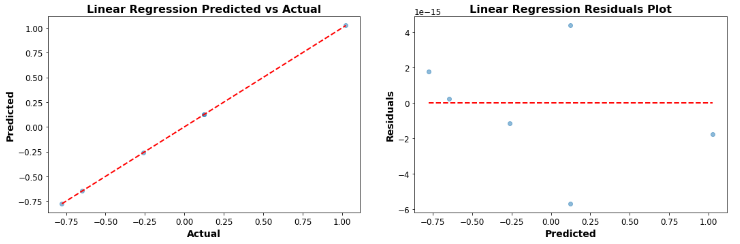
|  |  |
| --- | --- |
| **Evaluation Details** | |
|  |  |
| **Instruction Step** | Deaeration |
| **Production Batches** | 46 |
| **Production Batches after Outlier removal** | 27 |
| **Target Variable** | Phase Overrun |
|  |  |



Linear Regression exhibits good accuracy for this dataset. It achieved a perfect *R*2value of 1 for both the training and test datasets. Additionally, the Mean Squared Error (MSE) for this model is extremely low, nearly zero for both the training and test sets.

The Linear Regression model's performance great for this dataset. To achieve a perfect *R*2 , may indicate a potential for overfitting. Overfitting means the model may not generalize well to unseen data. This perfect score also raises concerns about possible data leakage, where the model might have unintentionally accessed the target variable during training. Other regression methods were looked at , that have regularizations such as lasso and ridge , they also showed good accuracy and precision.

The plots below also highlight how good the predictions made by the linear model area and also low variances in the residuals





The LSTM (Long Short-Term Memory) Neural Network showed a poor performance with the given dataset. The training *R*2value stands at -0.82 and, the test *R*2at -2.38. Such values suggest that the model does a poorer job at predictions than a basic horizontal line. The MSE values further validate this, being relatively high at 129.78 and 73.03 for the training and test datasets, LSTM is not the model for this prediction study

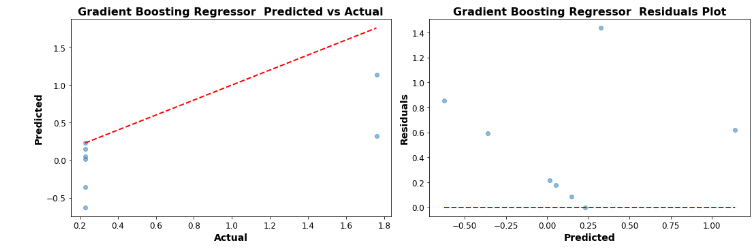
|  |  |
| --- | --- |
| **Evaluation Details** | |
|  |  |
| **Instruction Step** | Agitation |
| **Production Batches** | 46 |
| **Production Batches after Outlier removal** | 40 |
| **Target Variable** | Phase Overrun |
|  |  |

**Model Evaluation for Predicting Phase Overrun for 26 MT Tank for Agitation Phase**



The Gradient Boosting Regressor top performing in terms of model accuracy. With a R2 value of 0.97 for the training set, it indicates that the model captures 97% of the variance in the target variable. The test R2 is lower at 0.28, it's still the highest positive test R2 value among all the models. Its Mean Squared Error (MSE) for both the training and test datasets is low, further indicating its high predictive capability. It provides a good balance between fitting the training data and generalizing to the test data. There is potential for overfitting as there is a difference between the training and test R2 values. Further improvements with hyperparameter tuning, more batch data, or regularization techniques could help this.

The graph below shows that predictions using this model gives errors and the residuals plot shows that there is non linearities in the data





The LSTM Neural Network has a training R2 of -0.42 and a much worse test R2 of -5.51. The negative R2 values, especially one as low as -5.51, imply that the model is doing an exceptionally poor job at predicting the target variable. The high MSE values, 0.62 for the training dataset and 1.22 for the test dataset, further affirm the model's inadequacy.

The LSTM model's poor performance suggests it's not the right choice for this dataset. LSTMs, as a subtype of recurrent neural networks, are primarily designed for sequence prediction problems. The LSTM model might need significant architecture and hyperparameter adjustments.

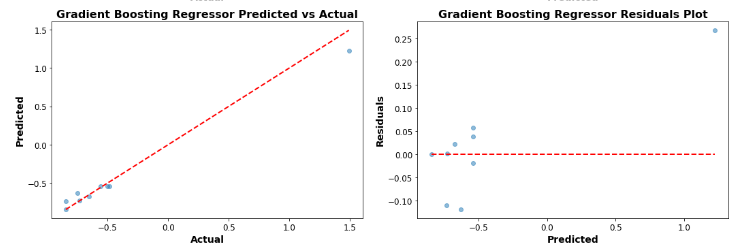
**Model Evaluation for Predicting Phase Overrun for 26 MT Tank for Gum Addition Phase**

|  |  |
| --- | --- |
| **Evaluation Details** | |
|  |  |
| **Instruction Step** | Gum Addition |
| **Production Batches** | 44 |
| **Production Batches after Outlier removal** | 43 |
| **Target Variable** | Phase Overrun |
|  |  |



The Gradient Boosting Regressor, after tuning, displayed good performance. The Train MSE of 1.07513 *E*−13 - 1.07513*E*−13, it essentially got close to a perfect fit for the training data. Moreover, the R2 value of 1.00 for both training and 0.98 for the test dataset indicates a good fit of the model to the data. The test MSE is also considerably low at 0.011. The hyperparameters provided, including a learning rate of 0.2, max depth of 3, and 300 estimators, seem to be optimally tuned for the dataset.

This can be also enhanced by the results from the plots below where there is good alignment of points along the line for the predicted vs actual and good distribution of points around the line in the residual plot





The LSTM Neural Network is ill-suited for this dataset. The high Train MSE of 5461.17 (even after tuning) and the Test MSE of 1773.12 highlight the inadequacy of the model. Although the LSTM model improved its test R2 from -0.34 to 0.01 after tuning, it's still barely better.

These results could be since LSTM models are primarily designed for time-series data where the sequence and temporal dependencies are important.

**Summary of all Machine Learning Models applied for all Production Tank Group under the different phases.**



**Feature Selection:**

The target or dependent variable is the Phase overrun duration and the independent variables are:

* Flowrate\_KGMIN,
* Target\_Phase\_duration,
* Target\_Flowrate,
* Phase\_duration,
* Phase\_start\_delay.

A correlation heatmap was used to visualise the correlation between the numeric variables and the highest correlation was between the phase duration and the phase overrun durations meaning when one increases so will the other.

**4.3 Machine Learning Model Selection**

Based on the nature of the production data and the target variable of downtime via phase overrun duration times is continuous, the model selection was a regression problem. The following is a list of selected Machine learning models applied to the data :

**Regression Algorithms:**

**Linear Regression**: As a basic approach, if you want to predict downtime or overrun times based on some features.

**Ridge/Lasso Regression**: These are extensions of linear regression that incorporate regularization, which might be helpful if there's multicollinearity or overfitting concerns.

**Tree-based Algorithms:**

**Decision Trees and Random Forests**: Can capture non-linear relationships and are useful for both regression (predicting time values) and classification (categorizing into different levels of downtime).

**Gradient Boosted Trees (like XGBoost, LightGBM):** Can improve predictive accuracy by iteratively correcting errors from previous trees.

**Support Vector Machines (SVM**): With the regression version (Support Vector Regression), it can be applied to predict downtime or overrun times.

**Ensemble Methods:**

**Bagging:** Uses multiple instances of a model to achieve better predictive performance.

**Boosting:** Aims to convert weak learners into strong learners by focusing on the misclassified predictions.

Data Split: 80% of the data was used for training, and the remaining 20% was reserved for testing.

Hyperparameter Tuning: GridSearchCV was used to identify optimal parameters for the Random Forest Regression model. For the RNN, different numbers of layers and neurons were tried to achieve the best result.

Validation: A 10-fold cross-validation was applied during the training phase to prevent overfitting and ensure model generalization.

Results Parameters :

Columns:

* **Model**: The name of the machine learning model.
* **Train MSE**: Mean Squared Error on the training dataset. This metric tells us how well the model fits the training data. Lower is better.
* **Test MSE**: Mean Squared Error on the testing/validation dataset. This tells us how well the model generalizes to new, unseen data. Lower is better.
* **Train R^2**: *R*2 (R-squared) score on the training data. It indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. Closer to 1 is better.
* **Test R^2**: *R*2 (R-squared) score on the testing data. Closer to 1 is better.

4.8 Conclusion

This chapter offered a comprehensive view of the steps undertaken to implement machine learning methodologies aimed at enhancing beverage production efficiency. Subsequent chapters will discuss the results, insights, and recommendations derived from these models.General Method outline:

1. Feature Engineering: Decide which features (columns) are relevant for predicting 'Downtime' and create any additional derived features that could be useful.
2. Data Preprocessing: If your features include categorical variables like 'TankName' and 'Material', you'll need to convert them into numerical representations using techniques like one-hot encoding.
3. Train-Test Split: Split your data into a training set and a testing set to evaluate your machine learning model's performance.
4. Select a Model: Choose a suitable machine learning algorithm for regression tasks. Linear regression is a common choice for predicting continuous values like 'Downtime', but you can explore other algorithms as well.
5. Train the Model: Fit the chosen algorithm to your training data.
6. Evaluate the Model: Use your testing data to evaluate how well your model predicts 'Downtime'. Common evaluation metrics for regression include Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.
7. Fine-Tuning: You can fine-tune your model's hyperparameters to improve its performance.
8. Prediction: Once your model is trained and evaluated, you can use it to make predictions on new data.

In this section, we present the implementation details of the predictive models developed to achieve the objectives outlined in Section 3. These models aim to predict the phase overrun in a production process based on various independent variables, including phase duration, phase start delay, flowrate, and quantity. We utilized three distinct machine learning techniques: Gradient Boosting Regressor, Support Vector Machine (SVM), and Artificial Neural Network (ANN). The implementation was carried out using the Python programming language and relevant libraries such as scikit-learn and Keras.

4.1 Data Preprocessing

Before building the predictive models, the dataset was preprocessed to ensure that it is suitable for training and evaluation. This involved handling missing values, scaling features, and splitting the data into training and testing sets. The dataset consists of various features, including Phase\_duration, Phase\_start\_delay, Flowrate, and Quantity. The target variable is Phase\_overrun, which represents the time by which a production phase exceeds the expected duration.

4.2 Gradient Boosting Regressor

The first model implemented was the Gradient Boosting Regressor, a powerful ensemble method known for its ability to capture complex relationships in the data. We trained the model using the scikit-learn library and experimented with hyperparameters such as the number of estimators, learning rate, and maximum depth. The model was evaluated using metrics such as Mean Squared Error (MSE) and R-squared on the test data.

4.3 Support Vector Machine (SVM)

The second model employed was the Support Vector Machine (SVM), a non-linear regression technique suitable for capturing intricate relationships between variables. We utilized the SVM implementation from scikit-learn and opted for a radial basis function (RBF) kernel due to its flexibility in handling non-linear data. The SVM model was trained using the scaled features and evaluated using metrics like MSE and R-squared.

4.4 Artificial Neural Network (ANN)

The third model implemented was an Artificial Neural Network (ANN), a deep learning approach capable of discovering complex patterns in the data. We used the Keras library to build the ANN architecture, which included input, hidden, and output layers. The input features were standardized, and the model was trained using the Adam optimizer with mean squared error loss. The ANN's performance was evaluated using the same metrics as the previous models.

4.5 Model Evaluation

After training each model, we evaluated their predictive performance using appropriate metrics. The evaluation involved making predictions on the test dataset and computing the Mean Squared Error, Mean Absolute Error, and R-squared values. These metrics provided insights into the accuracy of the models in predicting phase overrun.

4.6 Comparison and Interpretation

The results from the three models were compared to assess their effectiveness in predicting phase overrun. The objective was to determine which model best captures the relationships between the independent variables and the target variable. Additionally, the interpretation of model outputs was performed to gain insights into the contribution of each feature in predicting phase overrun.

5. Results and Discussion

The results obtained from the implementation of the three predictive models are discussed in the subsequent section. The findings are analyzed in the context of the research objectives to draw meaningful conclusions.

**Outliers :**

The decision of whether to include outliers in your dataset when predicting phase overrun using machine learning depends on several factors and should be made based on the specific characteristics of your data and the problem you're trying to solve. Here are some considerations to help you decide:

1. Impact on Model Performance:

Outliers can significantly influence the training of machine learning models, particularly linear models. They can pull the model's fit towards them, resulting in poor generalization to new data.

For non-linear models or models robust to outliers (e.g., Random Forests, Gradient Boosting), the impact of outliers might be less pronounced.

2. Data Quality:

Consider whether the outliers are genuine data points or errors in data collection. Genuine outliers might contain valuable information about rare scenarios or anomalies in your data.

If outliers are errors, it might be beneficial to remove or correct them to avoid misleading the model.

3. Problem Context:

If predicting phase overrun is critical for both common and rare cases, outliers could be valuable to capture those rare instances.

On the other hand, if you are more interested in predicting typical cases and outliers are considered unusual events, you might want to treat them separately.

4. Robustness vs. Precision:

Including outliers might increase the robustness of the model to handle various situations, but it might come at the cost of predictive accuracy for the majority of cases.

5. Scaling and Transformation:

If outliers are included, consider scaling or transforming your features appropriately to reduce their influence.

6. Testing and Validation:

Evaluate the model's performance both with and without outliers using appropriate validation techniques, such as cross-validation, to understand how they affect predictions.

7. Outlier Detection:

Before deciding, you might want to perform outlier detection techniques to identify and analyze potential outliers. This can help you make an informed choice about including or excluding them.

In summary, whether to include outliers in your machine learning model depends on your problem context, the type of outliers, and the behavior of the chosen algorithm. It's recommended to experiment with both scenarios and compare their performance using appropriate evaluation metrics to make an informed decision.

# Chapter 8: Conclusion

# Appendix A: Table of Results

**Tank 22 MT**

1. All phases



1. Agitation Phase



1. Gum Addition Phase



1. Deaeration Phase



Tank 23 MT

1. All phases



1. Agitation Phase



1. Gum Addition Phase



Tank 25 MT 4

1. All Phases



2- Agitation Phase



1. Deaeration Phase



Gum Addition Phase



25MT10

All Phases



Deaeration



Agitation Phase



Gum Addition Phase



26MT

All Phases



# Appendix B: Interview Transcripts

# Appendix C: Data Permissions

# Appendix D: Consent Forms

# Reference List