**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.

**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.**

# Chapter 2: Research Design

## Primary Data

## Problem Identification and Clarification

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

## Research Objectives

**Objective 1** -

**Objective 2** -

**Objective 3** -

## 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

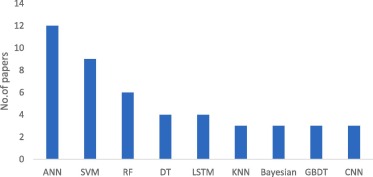
Reinforcement learning – using feedback mechanisms to reward positive action and punish the negative action.

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.



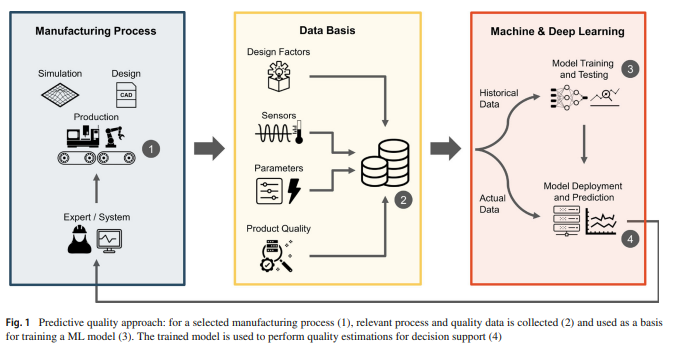
***Fig 1 Distribution of machine learning algorithms***

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 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 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.

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 Methodology and 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 Methodology and 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, based on the fact that 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.

6. Sampling Strategy

**5.1 Identify the appropriate population(s) of interest.**

To identify the appropriate population of interest for my research, I must examine my objectives and the context of my research. It is important to define and justify the populations of interest to guarantee that the research is focused and relevant which will lead to meaningful analysis and interpretation of the result. It is important to be aware that this population of interest / data input is coming from one manufacturing company using one manufacturing material specific production step.

For this research, the initial population of interest was process data of materials produced in a specific production area over a set time. This population as data was immense in size. This would take a lot of time and resources to examine. using my objectives as a tool to take a sample of this data which would be specific to my research. towards looking at materials described as common mucilage beverage products. There are 13 of these type of products in total produced between two mixing tanks. As part of their process, they have a unique step called deaeration. All steps in the production process are recorded for every batch that is made and is stored in a database. So, by using SQL language, we created a query to call out firstly all data on all products batches produced over a 2-year period and then added that we only wanted product batches that had the deaeration step. From this a new population of interest was formed where we have process batch information for all 13 materials of interest over 2 years for 2 mixing tanks.

The original population of interest was based on 2 years of process batch data for a particular production process area. This is a huge amount of data, so after a consulting with associates that work in the area, we decided that it was much more of benefit to look at batches that require a deaeration step in the process. So, batch data was extracted based this process step. This extract became the population of interest.

The Data population includes 13 materials /products that were produced between two mixing tanks several times over 2 years and require a deaeration step after mixing.

* To gain a better understanding on why there are different deaeration times for each batch through the following:
  1. A deeper understanding on the the common mucilage beverage product production and the various factor affecting the deaeration time
  2. Review of the data that is available and what it can tell us about each step in the process

Based on the research objectives, is sampling required here? One of the objectives is to understand the common mucilage beverage product production, so initially data was collated for batches produced over 2 years. This was large amount of data which was stored on the company’s databases. All process steps are recorded and stored for each batch produced, however due to time and financial constraints, this data is not reviewed. I needed to used sampling technique to reduce the amount of data I need to collect but considering only data from a subgroup rather than all possible cases of elements. To determine the subgroup of data, I narrowed down my research to only looking at batch production data that has a deaeration step. This data extraction was completed using an SQL query searching for the deaeration step on the shopfloor database. The resultant sampling show that there are 13 materials that require this step and were produced several times over the time of 2 years.

2 years was chosen to make sure we had produced these materials a few times.

This is type of sampling is call probability or representative sampling – where the chance of each case being selected from the population is known and is usually equal for all cases. As we know the materials that require deaeration and we also had a set time, the only thing we didn’t know if these materials were produced more then once. To get an accurate picture of the process, we need to look at more than one batch production for each material. To eliminate the possibility of bias.

Machine learning algorithms are only as unbiased as the data they are trained on, and there is a risk that biases and discriminatory practices may be intentionally reinforced using such algorithms. To avoid this, is to be aware of potential bias to ensure fairness and objectivity. To choose a population of interest randomly and to critically assess the impact of bias on the analysis results. The question of access to this data is important. This would impact on the ability to select a representable population of interest. An informal access was granted by the manufacturing company to allow data to be downloaded from their internal database systems on the basis that this data would be anonymised in the research.

6.1. A key part of any research project is deciding if a sampling strategy is necessary, and if it is, what is the appropriate strategy upon which to complete the analysis on. Looking at the two data inputs, the author determined that a different strategy will be needed for primary and secondary data collection. For primary data, namely in-depth unstructured interviews as outlined above, a non-probability sampling method will need to be applied. For secondary data, a probability-based sampling method is most appropriate. Both approaches are outlined in the following sections.

6.2. Primary Data

As outlined in Section 5, the chosen primary research methodology selected is that of in-depth unstructured interviews. This approach allows for participants to be selected based on their expertise with a given area, a form of sampling known as purposive or judgement sampling (Saunders et al, 2012). Purposive sampling does not have any statistically relevance as the chosen sample bears no resemblance to the overall population of the study. Saunders et al (2012) outlines that Heterogeneous sampling is that where ‘chosen participants with sufficiently diverse characteristics provide the maximum variation possible in the data collected’, (pp. 287). Saunders et al (2012) also advise that a minimum sample size of 5 in-depth unstructured interviews be carried out.

6.3. Secondary Data

When reviewing the secondary data requirements, it was clear that a different strategy was needed. Specifically relating to this project, the total population is approximately 3,000 employees worldwide. Narrowing the scope of the study to just employees in Ireland reduces the scope to approximately 550 employees. Focusing on just employees employed in the operations function again the population narrows to approximately 300 employees. As outlined above training records are collated from when a new employee starts work. Training records have been collated since the LMS was introduced in 2020. In addition, the timeframe by which to complete the analysis is limited to several weeks. Working with full populations of more than 50 is not recommended (Saunders, Lewis and Thornhill, 2012). Based on these reasons, it will be necessary to use a sample strategy to complete the research.

6.4. The author reviewed characteristics of the two sampling techniques of probability and non-probability closely. Each case has an equal chance of being selected for inclusion in the sample. Therefore, the author has identified that using a probability-based sampling strategy would best fit the research task, objectives, and population within scope of the project.

6.5. To clearly articulate the reasons behind this choice of sampling strategy, the author has utilised work completed by Saunders et al (2012) as a guide. In their book ‘Research Methods for Business Students’ they outlined a number of steps to help provide clarification when choosing a sampling strategy, namely identification of sampling frame, sample size, sampling technique and a check to ensure that the sample is representative of the overall population (Saunders, Lewis and Thornhill, 2012). Using this framework, the author will discuss each element comprehensively in the following sections.

**6.6. Sampling Frame**

6.6.1. Saunders et al (2012) outlines that a sampling frame is ‘a complete list of all the cases in the population from which your sample will be drawn’, (pp 293)

6.6.3. A further consideration would be that the data is stored within the LMS is for all employees working for the company both in Ireland and in other sites around the world. By filtering to the Irish site, it’s possible to say that employees contained there have an equal chance of being selected to take part in the study.

6.6.4. In addition to filtering the data to only Irish based employees, it will be necessary to consider the different working patterns of employees. Some employees work full time (100%), whist other employees work less than this on a range of flexible work patterns such as weekend or part time.

**6.7. Sample Size**

6.7.1. As sampling size will have an impact on any results, the research found considered what the appropriate sample size should be. The population of employees is approximately 300. Saunders et al (2012) outline that a sample size of 30 is a ‘rule of thumb’ in order to carry out statistical analysis on the group. This will allow a degree of confidence in the analysis and is based on best practice.

Using a sample size of 30 for this population would give a 10% sample size. The author is suggesting a sample of 60, which would give a sample size of 20%.

**6.8. Sampling Techniques**

6.8.1. The author reviewed the five sampling techniques common to probability sampling and identified that systematic sampling would be the most suitable. The technique is discussed in more detail below.

**6.9. Check for Representation**

**6.10. Applying the Sampling Strategy**

# **Chapter 5: Implementation**

To perform machine learning to predict downtime based on phase over runtimes, tank sizes, and materials the data needs to be structured into a Pandas Data Frame

The DataFrame df contains columns for 'TankName', 'Material', 'TankSize', 'PhaseOverRuntime', and 'Downtime'. Each row represents a specific instance of data with information about the tank, material, tank size, phase over runtime, and corresponding downtime.

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.

**Tank\_1: Encoded tank categories.**

**BATCHID: Batch ID for each entry.**

**Material: Material ID for each entry.**

**Phase\_duration: Duration of the production phase.**

**Phase\_overrun: Phase overrun value.**

**Phase\_start\_delay: Delay in starting the production phase.**

**Flowrate\_KGMIN: Flow rate of ingredients.**

**Target\_Phase\_duration: Target duration of the production phase.**

**Target\_Flowrate: Target flow rate of ingredients.**

# Chapter 6: Results

# Chapter 7: Discussion

# Chapter 8: Conclusion

# Appendix A: Workflow

# Appendix B: Interview Transcripts

# Appendix C: Data Permissions

# Appendix D: Consent Forms

# Reference List