**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

# 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.**

# Chapter 4: Methodology

As a rule, research methodologies form two distinct groups namely primary and secondary data (Saunders, Lewis and Thornhill, 2012). Primary data is collected specifically in order to answer the research questions posed at the start of this paper, secondary data is collected for some other use and is then applied to this data. As can be expected, within both areas there are different types of collection methods that can be applied. The purpose of this section is to outline the collection strategy or strategies that work best in the author’s opinion to answer the research questions posed in Section 1 above.

**Primary Research Data Collection**

The author identified that a quantitative approach to collecting primary research data would be the most appropriate method of gaining first person information from people well versed in the succession planning process. Such data collection would be completed using in person interviews. Interviews were to be completed online using the Microsoft Teams platform which enabled transcription to be completed automatically. Using this method of data collection allowed for in depth discussion on the area of succession planning. As well allowing a level of observation to be used by the Researcher to gauge reactions to questions posed, as well as allow the flow of the interview to be altered depending on the interviewee’s reaction to the question (Saunders et al, 2012). To overcome any potential bias that occurs on the interviewee’s behalf, the author has sought input from different individuals who have taken part in the succession planning process. The interviewees will all provide a different point of view. One interviewee manages the succession planning process for the manufacturing division of the company, another manages the process for office-based employees located around the wider European Economic Area. Yet another interviewee has experience in both areas of the business outlined previously and is well placed to share similarities and differences in the two processes. Carrying out a diverse range of interviews allows the author to construct a holistic view of the succession planning process currently used.

The author has opted to using unstructured / in-depth interviews which are informal in nature with no prepared questions rather there is a general area for discussion (Interview Techniques for UX Practitioners: A User-Centered Design Method, 2013), Saunders et al 2012). This allows the author and interviewee to have a conversation that moves organically through topics giving an opportunity to probe for further understanding where necessary.

Issues considered by the author when selecting this method of primary research including:

* Research Objectives - the objectives stated previously lean on utilising HR Data, specifically training data to enhance decision-making. Unstructured interviews with experts within the HR Community will enable allow for open discussion on the themes of this research as outlined previously in the literature review.
* Expertise of the Selected Participants - expertise of the HR Community is important to this research. As such, it is important to canvas a range of opinions to gain a holistic understanding of how training data can be utilised for data driven decision making in the succession planning process.
* Timeframe - it is important to recognise that the timeframe with which to complete the research project is limited to a maximum of 10 weeks. As such, a limited number of interviews could be carried out during the timeframe.
* Ethical Considerations - in the world of the General Data Protection Regulations (GDPR) gaining consent of data subjects is very important. This is no different when considering ethical considerations for a research project. As standard, all interviewees were informed of the confidentiality of the process, as well as highlighting their option to withdraw their consent to have their interview notes included as part of the finished research paper. Should a participant decide to withdraw their consent, the author will put measures in place to delete all contributions from the project.
* Bias - the author herself is an experienced HR practitioner and as such already has a view regarding the use of HR data to drive decision making. Therefore, the author is aware of the need to be impartial in asking questions of participants as well as noting their responses.

The availability of experts was an important consideration when deciding on a primary research method. Working with HR data can be a nuanced process and the author understand the importance of garnering a range of views so a holistic view can be considered of the succession planning process. Using the framework outlined above, the following points were considered:

• Research Objectives - would the subject matter of the research objectives lend themselves to facilitating participant interviews of 30 minutes or longer? Were themes outlined in the Literature Review of relevance for the HR Community as they stood?

• Expertise of the Selected Participants - again, a key point in considering who would be a suitable participant for the interview process. In the end, the selected participants were employees with the company who play a key role in integrating data from different systems to support the succession planning process. Other employees had integral knowledge of the company’s LMS - where if successful will be the source of future evolutions of any successful algorithm.

• Timeframe -having consideration to the timeframe of the data, the author chose to largely nominate experts within the company in which she works. This will encourage a greater response rate as the ‘collegiality’ of the request.

• Ethical Considerations - as outlined in the previous section, all participants will be asked for their consent prior to the interviews taking place and will be reminded of their ability to withdraw their consent at any time up to publication of this research paper.

• Bias - using inputs from one company will naturally lead to a degree of bias in the process. However, the author has selected those experts who are involved in either improving or implementing analytical systems, or who work in diverse geographies within the company. The author is confident that this will give the interview responses a rounded view, keeping bias (unconscious or otherwise) to a minimum.

The next step in the methodology is in selecting the experts to take part in the interview process. To ensure that the right mix of experts are chosen, Saunders et al (2012) outline that identifying the characteristics of the experts prior to selection will create a more rounded group of experts. To that end, the author has identified two characteristics that would be helpful in answering the research objectives posed at the start of the document. The three characteristics are:

• Have some involvement in implementing / improving processes with HR and the wider company.

• Be ‘outward looking’ in that they are knowledgeable of company strategy as well as best practices within the market.

A panel of experts has been selected and based on the criteria and framework outlined previously. Transcripts from the interviews can be seen in appendix XXX.

**Secondary Research Data Collection**

Initially, the author hoped to use data from the company’s own LMS system. However, the author was unable to gain sufficient data to work with. As an alternative, the Open University Learning Analytics Dataset (OULAD) was selected as the dataset upon which analysis could be completed (Kuzilek, Hlosta and Zdrahal, 2017). In order to closely mimic data typically found in a commercial operation, the employee added a column to the ‘student\_info.csv’ file called ‘Tenure’ to represent the number of years the employee / student is working within the company. The values of the ‘Tenure’ column were randomly generated using ‘***randint***’ with values between 0 and 20. These values were also subsequently grouped together into ‘bands’ to identify if tenure had an impact when analysing categorical data.

**Research Methodology and Validity**

Considering the research methodology outlined above, it is possible to say that the most relevant components of validity relevant to this research are accuracy, currency, and bias. It is however also possible to say all components of validity apply to the proposed research, some component’s more than others. The concepts of accuracy and currency are explored below.

Accuracy in this instance relates to comprehensive the data statistically is. In terms of primary data, accuracy does not apply as the data is not statistically based. The data captured from interviews will need to be transcribed and included in the appendices of this report. Furthermore, the main points and sentiments expressed will be used to verify if the proposed model will be useful or not.

Currency in this instance is a potential barrier to the methodology of this research. The author has chosen to use simulated data extracted from an educational institute learning management system. The data was originally released in 2017 and contains data from 2013. It is true to say that the data is not current, however, it closely mimics the data is contained within the company’s own LMS. That being the case, the author has chosen to accept the risk to the validity of the results of this research paper.

Bias has already been identified as a possible threat to validity when conducting in-depth interviews for primary research. The author will attempt to limit bias by ensuring that there is a clear purpose of the interview which is communicated in advance. By working with known participants there is already a degree of trust established between the parties to facilitate a frank discussion. Finally, the author will create several prompts based on key research themes that will help guide the interview process and stay within the research area.

Although three components have been listed, it is not unreasonable to assert that other components may also become more apparent as this research progresses.

Research Methodology and Ethics

As with all research, there are ethical considerations that will need to be planned for, some of which have been outlined above.

5.5.2. In respect of primary data collection, participants will be asked to participate, and will have the option to withdraw their consent to participate or have their data included at any stage of the process up to the submission data. Interviews will be conducted.in a professional manner with responses being confidentially and anonymously recorded for the purpose of the research. The author will have a master file noting the responses, but this will not be share as standard. This Master file will be stored in a secure location and will be password protected as added security. Finally, if participants have questions at the end of the interview process, or before, the author will undertake to resolve these queries as quickly and sensitively as possible.

5.5.3. In respect of secondary data collection taken as a download from the company’s LMS, the author will put in place the relevant requests necessary within the company to obtain access to the data. Again, any data received will be anonymised immediately so no data can be related back to any individual. Furthermore, the data will be stored in a safe location with the relevant passwords in place.

5.5.4. The author is also conscious of the General Data Protection Regulation’s (GDPR) and will put the necessary steps in place to ensure compliance at all stages of the research.

6. Sampling Strategy

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)

6.6.2. As outlined in previous sections, the population for this research is contained within an LMS system, which contains data for all employees working for the company. Employee records are created as part of the new hire process, and a learning profile is created. As employees complete training tasks and are signed off, their learning record is updated. For employees not in a manufacturing role, assigned learnings are marked ‘complete’ once the learning is completed. Other training sites within the company are linked to the LMS so employee records are automatically updated with no manual interaction.

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.

6.8.2. Saunders et al (2012) outline that the systematic sampling technique involves dividing the population based on sample based on a regular interval. The interval is calculated using a sampling fraction. The inputs to the sampling fraction are the size of the same and the total population. The benefit of using a sampling fraction is that is it relatively easy to implement and explain. However, it’s important to ensure that the population is not pre-sorted in any way prior to the sampling method being applied. This will ensure that the applied sampling method is truly random.

6.8.3. For this piece of research, the sampling faction calculation is worked out below.

Sampling Fraction = Actual sample size 60 = 1

Total population 300 5

6.8.4. Based on the workings above, every 5th record will be taken to form part of the sample for this research.

6.9. Check for Representation

6.9.1. The final check of the sample is to assess the possibility of bias. This can be done in multiple ways, with one method being to draw another sample from the same data set, using a different sampling fraction to compare the results.

6.10. Applying the Sampling Strategy

6.10.1. The sampling strategy outlined above is normally applied to primary data collection. For the purposes of this research, a sampling strategy will be applied to both primary and secondary data collection.

6.10.2. Secondary data collection for the purpose of this research will be taken from the LMS system in the form of a download. The data downloaded will conform to the sampling strategy outlined above.

6.11. Now that a sampling strategy and methodology have been developed, it is necessary to consider the ethical implications associated with the research, and how the author proposes to work with them.

# Chapter 5: Implementation

# 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