**Demographic data and it’s use in a succession planning process - the role of data analytics.**

Sinéad M Duffy

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: Marina Soledad Iantorno

# 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:** | Demographic data and it’s use in a succession planning process - the role of data analytics. |
| **Supervisor Name:** | Marina Soledad Iantorno |
| **Student Full Name:** | Sinéad Duffy |
| **Student Number:** | SBA22229 |
| **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

Research Objectives

Validity Type

Ethical Considerations

Chapter 3: Literature Review

Chapter 4: Methodology

Chapter 5: Implementation

Chapter 6: Results

Chapter 7: Discussion

Chapter 8: Conclusion

# Chapter 1: Introduction

The author has been working in the area of Human Resource Management for nearly 20 years, specifically within a multinational company for nearly half that time. They have observed how much data is collected throughout the department such as

1. employee data in databases such as SAP SuccessFactors, Workday, ADP Workforce etc.
2. time and attendance data in time management systems
3. compensation and benefit data in benefit platforms
4. employee engagement data using employee experience systems, including performance management data
5. employee expenses in financial management systems
6. talent management software that allows management of the recruitment process, onboarding of employees as well as ongoing management of talent
7. learning management systems that structure learning experiences and ensuring compliance with training requirements or continuous professional development etc.

In the authors experience, all these systems operate independently of each other. For example, SAP and Workday may incorporate time and attendance tracking, talent management and some payroll processing or each may be a stand-alone system. A level of integration with APIs facilitating a connection to share basic data such as employee name and work number as well as work email can also be created. Beyond this, there appears to be very little integration into the wider financial governance of expenses, benefit management platforms or indeed platforms that track and detail the employee experiences.

The author wanted to utilise training data gathered from both local and corporate systems to see if there is any value in using this to support the succession planning process. As it currently stands, succession planning is largely a manual process, where Human Resource Business Partners speak with employees to identify areas that they would like to develop, what they feel their key skills are, identify if the employee is interested in moving within the company etc. How this information is gathered this information is individualized and based on the reporters’ own experiences. At different times during the year, this information is collated, and meetings are held functional area leads and other senior managers to hold succession planning conversations. Currently training data (such as courses completed) is not included as a metric in the process. The author would like to explore if there is a role for training data within the succession planning process. If it is possible to identify such a role, what it would look like.

Tambe et al in an article called ‘Artificial Intelligence in Human Resources Management: Challenges and a Path Forward’ (Tambe, Cappelli and Yakubovich, 2019) discuss the challenge that is faced when using HR data for machine learning. The article outlined that datasets from human resources can be small with not commonly repeating events (such as dismissals) or are influenced by external factors such as employment law (Equality Acts) or company policies (gender positive profiles). With small datasets taking account of influences outlined previously, there is an opportunity to look at relationships in the data to through the lenses of relationships rather than prediction from correlations of observed variables as in other areas of machine learning algorithms.

Initially, the author planned to use data from the learning management system (LMS) within the multinational company where they work. However, due to data privacy and ethical issues as well as the need for a non-disclosure agreement and limited access to proprietary information led the author to reevaluate this data source. An alternative data source was identified as an alternative (the Open University Learning Analytics Dataset) which will be discussed in more detail in following chapters.

The author was able to utilise the expertise of colleagues within the Human Resources community to discuss their experiences in the field of human resources, specifically in succession planning.

# Chapter 2: Research Design

## Primary Research - Data Collection

As part of the research proposal a quantitative approach to collecting primary research data was deemed the most appropriate method of gaining first person information from sources well versed in the succession planning process within a multinational company. Such data collection would be completed using in person interviews. Unstructured interviews were completed online using the Microsoft Teams platform which enabled transcription to be completed automatically. This method of 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 and adjust the flow of the interview depending on the interviewee’s reaction to the question (Kumar, 2011; Saunders, Lewis and Thornhill, 2012; Wilson, 2013). In essence, this method allows the author and interviewee to have a conversation that moves organically through topics giving an opportunity to probe for further understanding where necessary.

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 key in answering the research objectives posed as part of this research, namely;

• 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.

To overcome any potential bias that occurs on the interviewee’s behalf, the author sought input from different individuals who have taken part in the succession planning process over several years, either in the multinational company, or as part of their previous roles with other organisations. In terms of gaining a holistic view of the succession planning process, the interviewees have been identified from different part of the organisation. 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. Transcripts from the interviews can be seen in appendix XXX.

## Secondary Data Collection

As outlined previously, initially, the author hoped to use data from the company’s own LMS system. 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). The data file specifically selected for analysis was the ‘student\_info.csv’ file.

The dataset OULAD closely mimics data typically found in a commercial operation and extracted from an LMS within the company such as:

* **Student\_Id** = representation for employee number
* **Studied\_Credit** = as a simulation / substitute for level of interaction with the system i.e. the more credits gained, the more courses completed.
* **Tenure** = the addition of a column on tenure as a random variable to simulate the number of years the given employee / student has been with the company.
* **Gender**
* **Age**
* **Previous education**

The use of OULAD reduces the risks associated with data privacy and data ethics. It also allows for the data to be filtered to one module over one semester - thus allowing the data to represent one manufacturing site with a similar number of employees.

## Problem Identification and Clarification **FINISHED HERE ….**

**Thesis Title** - Employee demographic data and it’s use in a succession planning process - the role of data analytics.

## Research Objectives

**Objective 1** - Does employee tenure have an impact on a succession planning?

**Objective 2** - Can and employee’s previous level of education have an impact on succession planning?

**Objective 3** - Are there other demographic features which have an impact such as age or gender?

## Validity Type

Reference - CHAPTER 11 of Ranjit Kumar’s book

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.

## Ethical Considerations

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

In respect of the scheduled interviews, participants have been asked to take part, and have been given the option to withdraw their consent or have their data excluded at any stage of the process up to the final submission date. All interview participants are over 18 years of age and have not disclosed any medical condition or any other prohibition that will limit their ability to take part in the interview process. No incentives have been given to any participant in order to gain their support in the research process. As an added measure, all interviews have been transcribed for completeness and included in the appendix of this document. In the event that any participants as any questions at the end of the interview process, or in the time up to the submission date, the author has outlined a communication process that will allow for speedy resolution to these queries as quickly and sensitively as possible.

In respect of secondary data, due data protection and sensitivity issues the author decided to use dataset obtained from the Open University (OULAD). This dataset was selected as it closely mimicked an extract of the LMS system used within the selected Company. The OULAD dataset contains more the 34000 data points which have already be anonymised, thus limiting any potential data breach. The General Data Protection Regulations (GDPR) outlines companies’ requirements to protect the private data of individuals. It also enshrines the concept of privacy by design -where anyone working with, or handling data needs to have sufficient security measures in place to secure the data from any potential risks. The decision to use a widely available dataset instead of actual employee data a key reason that OULAD was selected for analysis. As such, the author has attempted to minimise any potential risks to the company’s data, whilst also maintaining compliance with the companies own internal GDR Procedures.

# Chapter 3: Literature Review

## Introduction

People analytics as defined by Ferrar et al (2021) is ‘*The analysis of employee and workforce data to reveal insights and provide recommendations to improve business outcomes*’. Numerous authors outline the importance of using data analytics to empower business decisions within the Human Resources Function (Ferrar et al. 2021, Mattox et all 2020, Rasmussen and Ulrich 2015). Rasmussen and Ulrich (2015) however point out the need to ask the ‘*right question*’ when reviewing data generated by HR and propose that this question should be incorporated into the end-to-end analytics process to identify and confirm the impact of people on decisions.

Learning analytics on the other hand, focuses on the effectiveness of a learner’s experience and is routed in basic training evaluation models such as the Four Levels of Evaluation model developed by Don Kirkpatrick (Mattox et al 2020). Specifically in this research paper, the author will focus on training provided solely within a corporate structure. Using a Learning Management Systems (LMS) has provided an effective way of gathering, analysing and reporting on learning related data (Sin and Muthu, 2015, Arka et al 2022, Mattox et al 2020). LMS’s such as Moodle have long been used in academic circles and have provided rich data sources in understanding how students learn and interact with systems (Sin and Muthu, 2015, Arka et al 2022).

This paper is an attempt to identify if a link or relationship can be found between training undertaken by employees and area’s such as succession planning within a manufacturing environment. Initial investigations into academic literature on the use of HR and learning data uncovered different themes which will be outlined in the following sections.

**Themes of the Literature Review.**

The Literature Review of academic and related papers helped to uncover several themes with the opportunity for further analysis of data held within the HR Department focusing specifically on data relating to learners. How the analysis should be conducted was, as expected, discussed at length with different approaches being taken. Four main themes that were identified by the author and have been outlined in more detail in the following sections.

## Opportunities for use of Human Resource Data

HR data provides a lot of opportunity for analysis within companies (Rasmussen and Ulrich, 2015; Mattox II, Parskey and Hall, 2020; Ferrar and Green, 2021). Mattox et al (2020) in their book ‘*Learning Analytics*’ outline the pressure from business leaders to provide better and more insightful information in a timely manner. The demand for information is coming not just from Senior Managers, but also from stakeholder who want to know more about the people function and how effective it is (Topno, 2012; Mattox II, Parskey and Hall, 2020; Ferrar and Green, 2021). David Ulrich outlines how people analytics can add value to companies by allowing teams to make informed decision led with data in support of the business (Ferrar and Green, 2021).

To balance out this desire, HR data is uniquely different from other types of data used for analysis (Tambe, Cappelli and Yakubovich, 2019; Bankins, 2021). By its nature, data gathered by HR is formed of generally small datasets where events that companies want to model or predict are infrequent and nonstandard (for example dismissal of employees) or the data is subject to interpretation such as performance management where employees with different roles and responsibilities cannot easily be compared (Chadwick and Dabu, 2009; Tambe, Cappelli and Yakubovich, 2019; Bankins, 2021). Another issue with HR data relates to external requirements on the company which are not evident in other functions. For example, the recruitment process is influenced by internal factors such as the company’s own recruitment goals, as well as external ones such as the statutory landscape (Tambe, Cappelli and Yakubovich, 2019). This fact forces companies to limit the use of historical data such as recruitment data as it’s use could make incorrect predictions based on outdated information, or based on practices that are no longer the same (Rasmussen and Ulrich, 2015; Tambe, Cappelli and Yakubovich, 2019). As Bhardwaj et al (2019) stated ‘*Human resource analytics is an area of study that uses the mix of art and science on human capital in order to get measurable return on investment*’, (Bhardwaj and Patnaik, 2019).

That being said, HR needs to prove its importance to the business, especially in terms of how impactful it’s action are on the overall financial health of the company (Dong, 2022; Losey, Meisinger and Ulrich, 2005), (pp 121). In monetary terms analysis has shown that small changes to processes can make cost savings for the business such as implementing training reminders to cut down on the amount of time to complete induction, or to uncover a link between engagement data and business performance (Ferrar and Green, 2021) (pp. 4). There is an opportunity for data analytics within HR, whilst also recognising the need for help from HR experts to interpret the results of any analysis (Edwards and Edwards, 2019) - (pp. 5). In truth HR need to refocus their role to become a ‘strategic partner’ of the business helping it to achieve its strategic goals (Bhardwaj and Patnaik, 2019; Dahlbom *et al.*, 2020; Losey, Meisinger and Ulrich, 2005) (pp 150). Academics are aligned on the need for HR to upskill and become ‘ambassadors’ for data analytics as a means of driving data driven decision making (Martin, 2019).

One could argue that the future for HR data is to become integrated into the wider information stream of the company as a method to identifying how individual’s performance affects the wider company performance (Rasmussen and Ulrich, 2015; Tambe, Cappelli and Yakubovich, 2019). Rasmussen et al (2015) outline that impactful HR analytics are about linking to strategic business operations rather than trying to identify patterns in big data (Rasmussen and Ulrich, 2015). Some academic’s espouse the opinion that to be used successfully, HR data must be taken away from the HR department for analysis (Rasmussen and Ulrich, 2015; Ferrar and Green, 2021). Experience in one case-study outlined by Ferrar et al (pp 20 - 26), recofirms that HR data is different to other types of data and to successfully analyse it HR must be included in system development (Ferrar and Green, 2021).

Use of machine learning or artificial intelligence to supplement the HR decision making process is another growing theme within the data analysis raises ethical issues and questions which should be considered as part of this research. Employee’s perception on the use of artificial intelligence

Bankins et al (2021) has proposed a framework to help with the ethical implementation of artificial intelligence within an organisation (Bankins, 2021).

Focusing on data gathered as part of the learning process and how such analysis might be completed is discussed in the following sections.

## Learning Analytics

A range of methods is used to create a dataset, and this is particularly true for online or distance learning (Sin and Muthu, 2015, Arka et al 2022). Systems such as Moodle allow analysts to follow a student’s learning path through a module or full course (Sin and Muthu, 2015, Arka et al 2022). Shen and Chi (2016) analysed how different levels of learners reacted to different methods of learning using such online interactions. In practice companies use systems such as LMS’ to collate learning data from employee interactions. An LMS (a Learning Management System) is a system that allows companies to manage training within the company, which then allows companies to run reports, track training requirements, assign learnings etc (*The LMS Guidebook : Learning Management Systems Demystified*, 2018)(Chapter 1). The advantages of using such a system is advanced features such as dashboards and reports created displaying high-level overviews of the data contained within the LMS as well as the ability to interlink with existing systems within the HR department (*The LMS Guidebook : Learning Management Systems Demystified*, 2018). The amount of data incorporated into an LMS means that large datasets can potentially be extracted, and it may be necessary to use data mining techniques to focus such big data sources (Sin and Muthu, 2015, Arka et al 2022, Mattox et al 2021).

The advent of LMS systems has led to a culture of self-directed learning by employees within companies (*The LMS Guidebook : Learning Management Systems Demystified*, 2018). Self-directed learning is where the employee is in charge of their own learning journey, a method of learning that is gaining traction in recent times (Araka *et al.*, 2022; Mustafa Yağcı, 2022). The drive to this new method of learning is coming from both companies as they roll out new technologies and employees themselves as they become more data savvy (Mattox II, Parskey and Hall, 2020; Araka *et al.*, 2022). The drive towards digitisation has only increased since the onset of Covid-19 and the need for companies and employees to adapt to increasing digital offerings (Almeida, Duarte Santos and Augusto Monteiro, 2020). Kokoc et al (2021) present the theory that by giving learners (employees) access to a dashboard to support their individual learning journey they will have more motivation to develop based on consistent feedback on their progress (Kokoç and Altun, 2021).

The development of this new area of learning has given more scope to allow machine learning to analyse the resulting data to help predict different outcomes - especially within educational settings (Araka *et al.*, 2022; Mustafa Yağcı, 2022). Analysis completed by academics chart learner performance against system access, and compare the results to final exam results (Araka *et al.*, 2022; Mustafa Yağcı, 2022). In companies, a different but similar approach is needed to gauge employee progress. For clarity, learning analytics has many definitions, but the one used in this paper is that learning analytics is the method of collecting, analysing, interpreting and reporting data to inform and understand learning methods and environments with a view to making improvements (Mattox II, Parskey and Hall, 2020; Kokoç and Altun, 2021; Araka *et al.*, 2022; Mustafa Yağcı, 2022). Educational data mining has emerged as a new field in which to access learning data stored in data warehouses or data lakes and seeks to work to open learning data to new analysis methods (*Learning Analytics – A Growing Field and Community Engagement*, 2015; Araka *et al.*, 2022; Mustafa Yağcı, 2022).

Deloitte in their 2017 Global Human Capital Trends outline that HR leaders, and specifically Learning & Development (L&D) leaders should reassess how they think about employees learning journey and ‘inspire’ employees to develop deeper skills with a view to enabling employees to change positions within their respective companies (‘2017 Deloitte Global Human Capital Trends’, 2017) (pp 36). The Deloitte report goes on to outline a case study about AT&T where they focus on career development for their employees and encourage them to change roles every four years as part of employees ongoing development (pp 36). Numerous sources outline reasons that employee should ideally be seeking new experiences every three to five years such as keeping in touch with outside trends, that employees become comfortable with change as some of the key items (Ryan, 2016; Christian, 2022).

As outlined in the research objectives above, the succession planning process is critical to the business’ ability to develop its employees. Huselid et al (2005) agree and outline that it is better to identify roles that are critical for the business and then spend time investing in the development of employees going into those roles to ensure that the right people are in place to drive the business forward (Huselid, Beatty and Becker, 2005).

## Causal Discovery

Causal Discovery is an area of analysis that has been growing steadily in the last number of years and numerous authors have made a study of using causal algorithms to help identify and infer relationships within data (Eberhardt, 2017; Malinsky and Danks, 2018). Eberhardt (2017) outlines in his article ‘Introduction to the foundations of causal discovery’ that these so called ‘causal relations’ are thought-provoking because of how they can be used to illustrate how a system or process will react if an intervention is put in place (Eberhardt, 2017). Eberhardt (2017) goes on to outline that when searching for a definition of causal discovery it is made up of three distinct elements - statistical inference (inference from data to the distribution ), causal discovery (inference of finding about the possible causal structure, given statistical quantities) and finally causal inference (is the deciding on the causal effects given the causal structure and associated quantities), (Spirtes and Zhang, 2016; Eberhardt, 2017).

Causal relationships and structures can be displayed with the use of graphs such as the Directed Acyclic Graphs (DAG) (Vowels, Cihan Camgoz and Bowden, 2023). An example of a DAG graph used by Vowels et al (2023) is displayed in Figure 1 below. The figure on the left-hand side indicates that B has an impact on both A and C, and that A also has an impact on A. Applying the CMC theory to the figure on the right-hand side, the external values of Ua, Ub and Uc are found to all have a causal relationship.

A picture containing sketch, diagram, drawing, circle

Description automatically generated

Figure 1- Transitioning from a typical DAG representation (left) to a structural equation model (right). Grey vertices are unobserved/latent random variables. Source - Vowels et al, 2023

As can be observed, causal discovery is routed in statistics and relies heavily upon the Causal Markov Condition with refinement such as adapting to include noise (Eberhardt, 2017), the concept of faithfulness (Malinsky and Danks, 2018) and fairness in the data (Loftus *et al.*, 2018) to name but a few. Malinsky et al (2017) define the Causal Markov Condition (CMC) as being ‘every variable X in V (the set of variables in the causal graph) is independent of its non-effects conditional on its direct causes. Malinsky et al (2017) goes on to explain Faithfulness as being ‘the only independencies among the variables in V are those entailed by the CMC’ (Malinsky and Danks, 2018).

A point raised in articles reviewed for this case study is that ‘causation is not correlation’, which is very true (Xiao *et al.*, 2022). However researchers have put forward the idea that there are links between different variables (dependent and independent), and the study of causation allows for the identification of such relationships as well as the opportunity to estimate the size and / or magnitude of the relationship (Vowels, Cihan Camgoz and Bowden, 2023).

Tambe et al (2019) put forward the argument that the HR Department generally contains small datasets which may not be suitable to use to clearly identify relationships with the dataset, but applying causal discovery to the data allows the data analyst to infer relationships (Tambe, Cappelli and Yakubovich, 2019). This is further complicated by any decision to use historical data for analysis, the danger with is that historical HR data my unwittingly contain a bias towards non-traditional employees within a workplace such as a bias towards men against women where historical data is largely collected on male employees as women were underrepresented at the time period (Loftus *et al.*, 2018; Tambe, Cappelli and Yakubovich, 2019; Vowels, Cihan Camgoz and Bowden, 2023). Therefore to limit potential bias when analysing HR data, whilst working with small datasets it is necessary to use other sources such as theory and prior research as a guide to identifying causal relations (Tambe, Cappelli and Yakubovich, 2019).

Tambe et al (2019) outline the benefits of using causal discovering and reasoning as a method for working with HR datasets, such allowing analysts to focus on the characteristics and behaviours of the variables in the dataset, decreasing the cost of data management as well as allowing users to articulate and display the relations between variables also (Tambe, Cappelli and Yakubovich, 2019). A natural disadvantage to the use of causal discover and reasoning is that the results are open to interpretation depending on how the data is read, they also are not strong on being able to predict outcomes to queries (Tambe, Cappelli and Yakubovich, 2019).

It is important to note that in some fields it is not ethical to seek causal relationships between variables, especially in scenarios where there may be ethical considerations (Eberhardt, 2017; Malinsky and Danks, 2018; Vowels, Cihan Camgoz and Bowden, 2023).

To facilitate causal discovery, several causal discovery algorithms have been developed to make the search for such casual relationships easier. This is the next theme that the author will review, including what kind of analysis is typically used in conjunction with causal discovery is also outlined.

## Casual Search Algorithms

Causal search algorithms, as defined by Malinsky and Danks (2017) are used to investigate hypothesis-based relationships between variables for example A and B based on the context of event C. They go on to explain that causal search algorithms are exactly the same as other better known analysis methods such as regression, the main difference being that assumption that results can be expressed as a causal graph such as a DAG (Malinsky and Danks, 2018). The benefit of using causal search algorithms is they help explain questions like ‘what makes a person intelligent’ by displaying all possible connections between subjects such as math, logic and writing test scores, leaving the analyst to select the best model based on the data and their own experience (Malinsky and Danks, 2018).

As expected, when applying causal search algorithms, there are a myriad of implementations possible, and some of these are detailed in the following sections. However, it would be helpful to understand the data preparation process that is recommended by Malinsky et al to help specifically with causal search algorithms. The first step is to assume that variables are ‘semantically independent’, i.e that they are capable of being manipulated independently. Therefore, it is important to remove any unnecessary or redundant variables before beginning analysis. The second step / assumption is that variable is continuous or categorical in nature, therefore any mixed datasets should be ‘cleaned’ and with the recommendation that values be placed within a scale to minimise any potential bias that could occur in the results. The third step is ensuring that any proxy or estimated values are as accurate as possible and of a single unobserved causal factor. The fourth step is to consider the timeframe that the data collected represents - is it month by month / day by day etc as the analyst must be able to outline if the measurement are for the same individual or for different ones over time. Finally, the fifth step is for the analyst to consider their own knowledge of potential causal relationships (Malinsky and Danks, 2018).

Malinsky and Danks, (2018) also outline the different types of search algorithms that may be applied to data including:

* Constraint based algorithms which display connections between the causal graphs and independencies that are found in the data.
* Score base algorithms which compares the models based on some measure of model fit such as the Bayesian Information Criterion.
* Causal search algorithms with semi-parametric assumptions which allows the model to use the assumptions to display the relationships in more detail.
* Clustering causal analogues which cluster variables as opposed to individuals. This helps to show or imply patterns of ‘observed correlation.

Example so applying such causal search algorithms are outlined in the beginning with Mӓkelӓ et al who applied a constraint based algorithms to their paper on Earth system sciences (Mäkelä *et al.*, 2022). The authors outline that knowledge experts can help beyond initial inputs and quality check the results by helping to interpret the results based on their own prior experience.

Qui et al (2019) used fuzzy cognitive maps to display causal relationships in emergency cases as an effective way of modelling knowledge representation (Qiu, Gu and Wang, 2019).

Dong (2022) proposes an integrated approach to analysis of HR data including use of neural networks, artificial intelligence etc in a proposal for an integrated system called the Human Resource Intelligence System (IHRMS) (Dong, 2022)

Assaad and Devijer (2022) applied causal discovery to time series data sharing a number of different approaches to this type of analysis. They outlined a number of issues applying theory to practice and warn that causal discovery and time series is still an active area for research (Assaad, Devijver and Gaussier, 2022).

Peters et al (2014) suggest using structural equation models rather than graphs as being helpful in using noise in data to highlight causal relationships between variables (Peters *et al.*, no date)

Finally, Kalainatha et al (2020) propose a method for using causal discovery within Python called the Casual Discovery toolbox (CDT) (Kalainathan, Goudet and Dutta, no date). CDT is an open-source library that allows users to first graph dependencies and then apply causal discover algorithms.

In respect of training data, different techniques which were used by authors to complete the analysis are outlined in the articles reviewed. Some of these techniques are outlined below:

* clustering / classification techniques (Araka *et al.*, 2022; Mustafa Yağcı, 2022)
* data mining based on time data (Araka *et al.*, 2022)
* Random forests / Support Vectors / Naïve Bayes techniques (Mustafa Yağcı, 2022)

As is expected, several alternative means of analysis is possible, and the author will need to experiment to identify the method and / or algorithm that will give the best results for the individual dataset.

## Conclusion

Reviewing work by other authors is a key step in completing any analysis. In this instance, the review was able to highlight the need for a more analytical mindset when considering HR data or learning data and outlined the benefits of same in terms of closer strategic connections to the business. Opportunities exist for such data to be used to allow better decision making by business leaders, but the HR department must enable this by being ‘ambassadors’ for their data. Systems such as an LMS should be used to leverage learning analysis within companies, as well as directing more effective and strategic learning outcomes-based on sound data analysis.

Analysis of LMS data can be completed using several causal machine learning algorithms to produce causal models that best suit the data at hand.

# Chapter 4: Methodology

Chapter 3 provides a strong foundation up on which the methodology for this research paper is based. The author reviewed the processes a few of several different research papers had previously taken when reviewing the OULAD dataset. This provided a launch point upon which this research could build upon. Specifically, the author was able to identify models already used to explore the dataset, as well as the focus of the article. Generally, research focused on developing models predicting student performance or engagement with the systems within a learning environment. This research paper however is concerned with the work environment, where learning opportunities may be more limited or directed by the company rather than the interest of the learner. As such, the author felt an opportunity exited to use prior learning and apply it in a new area.

## Dataset

As outlined in previous chapters, the dataset used for analysis was sourced from the Open University Learning Analytics Dataset. This was selected as it provides similar data structure and content to that found in LMS within commercial companies. Using an existing dataset allowed the author to clearly see analysis conducted by other researchers around learning analytics. This research paper differs from those others in that the main aim of this research is not to identify students at risk of dropping out or of poor performance, but rather to identify if opportunities exist to utilise learning data to aid succession planning. It also minimises exposure to GDPR considerations, as well as releasing the company from sharing any proprietary data.

## Python Libraries - Calculations, Graphs and Analysis

Python was the programming language that was used to complete the analysis for this research. Python was selected as it was relatively easy to learn and program for the researcher. The open source code libraries that are widely available within python and can be tailored depending on the needs of the programmer (M.Sc, 2023b, 2023a).

Figure 2- Python Libraries used for analysis

Numerous articles outline the values of different python libraries that are available for use, with six core libraries being identified as being of use in this analysis. The selected libraries are outlined below (M.Sc, 2023a, 2023b; *scikit-learn: machine learning in Python — scikit-learn 1.3.0 documentation*, no date; *SciPy*, no date; *An introduction to seaborn — seaborn 0.12.2 documentation*, no date).

* Pandas - used for ‘data wrangling’ or manipulation of data within a dataframe.
* NumPy - used for calculations quickly and easily.
* Matplotlib - helps display graphs and visualisations of the data using pandas and numpy
* seaborn - built to work alongside Matplotlib, the Seaborn library allows for statistical data to be graphed and displayed.
* 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 - enables a programmer to implement statistics and other mathematical computations with python.
* Keras

## Exploratory data analysis

As with all data analysis projects, the exploratory data analysis (EDA) stage is crucial to ensure that the resulting analysis will be valid and not biased one way or the other. The steps undertaken as part of EDA in Notebook 00 - EDA on - Student\_Info.csv - as outlined by Peng are shown below (Peng *et al.*, 2021).

Figure 3- Exploratory Data Analysis steps (Peng et al., 2021)

## Encoding Data Types

The data used for analysis consists of both ordinal (age band / tenure band etc) and nominal (gender / credit studied etc) data types. Thus, it was necessary to apply a suitable methodology to handle such data. Label Encoding was provided a simple solution to resolving data in the ‘gender’ column by converting the values from text to number (1 for Male, 0 for female) (*Categorical encoding using Label-Encoding and One-Hot-Encoder | by Dinesh Yadav | Towards Data Science*, no date) . One -Hot Encoding was selected as the basis to convert ordinal data into numerical values as it is relatively easy to implement, and allows for each categorical data value to give given it’s own column within the dataframe (Gefferth, 2023; *Categorical encoding using Label-Encoding and One-Hot-Encoder | by Dinesh Yadav | Towards Data Science*, no date; *sklearn.preprocessing.OneHotEncoder*, no date). Both encoding methods (Label Encoding and One-Hot Encoding) are part of the Scikit-Learn Python Library discussed previously.

## Machine Learning Models

All the models used as part of this research project were sourced from literature, as outlined in the figure below. Five different models were identified from the articles, and each are discussed in detail below.



Figure 4 - Articles and Models reference table

## Algorithm 1 - Logistic Regression

Logistic Regression is widely recommended as a starting point for data analysis models, especially when working with classification data (Kohnke, Foung and Chen, 2022; Python, no date). The first pass of the algorithm was completed without any parameter tuning. Hyperparameter tuning, specifically GridSearchCV was applied to the algorithm to improve results. Cross Fold Validation was applied in Kohnke et al (2022), the author applied GridSearchCV to the data as it automatically searches for all possible parameter combinations withing the data (*3.2. Tuning the hyper-parameters of an estimator*, no date).

## Algorithm 2 - Decision Trees

As logistic regression was not conclusive in terms of its results, the author decided to implement a similar model capable of working with categorical data. Decision Trees was selected as the second algorithm to test as it had been used by other researchers (Djoundourian, 2017; Hussain *et al.*, 2018). Different methods of hyperparameter tuning were applied in the form of cross validation and GridSearchCV from scikit-learn. Cross Validation was selected as it was used in previous papers (Kohnke, Foung and Chen, 2022) **INSERT ZHANG**), whilst scikit learn was selected as the library as both the model and hyperparameter tuning was selected from both. For Cross Validation, a range of features were tried to identify what number of features is optimal for the best results within the model. In addition, the author sought to look at tenure both as a grouped column, and an ungrouped column. This was completed to identify if the there was any difference in the results.

**Upload Zhang .. and check SVM / SVC**

## Algorithm 3 - Support Vector Machines

Support Vector Machines (SVM) is also able to work with classification data, and is able to identify outliers in the data (*1.4. Support Vector Machines — scikit-learn 1.3.0 documentation*, no date). Zhang et al (2023) also used SVM to carry out their analysis on OULAD dataset. Scikit-learn documentation highly recommend that data used for SVM analysis be scaled using the StandardScaler(), and the author intends to follow this advice (*1.4. Support Vector Machines — scikit-learn 1.3.0 documentation*, no date). In terms of other variables, a range of kernels are to be applied to identify which is the most suitable. The kernel criteria utilise different mathematical bases to transform the data across different vectors (into 3d to better view the data). Hyperparameter tuning was applied in the form of GridSearchCV as recommended by Scikit-Learn, and the applied further tuning by applying param\_grid as a further addition to the parameters to allow for a search to be completed over a wider sequence of parameter settings including ‘C’ and ‘gamma’. The cross-validation parameter in this instance was set to 5.

## Algorithm 4 - Random Forest

Based on the results achieved using Decision Trees, the author decided to use the Random Forest algorithm from scikit learn. Random Forest is an ‘ensemble’ algorithm which builds a ‘forest’ of decision trees to complete the desired analysis (*sklearn.ensemble.RandomForestClassifier*, no date; *1.11. Ensemble methods*, no date). Ensemble in this instance refers to the classifiers method of combing different estimators (Decision trees in this instance) to arrive at a single result (*1.11. Ensemble methods*, no date). It is possible to select one of three possible criterion to assess the quality of the splits in the data. Two of the methods were used to evaluate which was the most fitting - entropy and gini as the default. Only Entropy was selected as for the remaining criterion as both it and log\_loss use the same method of evaluation. As with previous models, no feature selection or hyperparameter turning was applied in the first run of the model. GridSearchCV and parm\_grid was selected as the most appropriate tuning methods based on the results obtained when working with algorithm two. The model was also run with a different n\_estimators or ‘trees’ in the forest of the model to determine what is the optimal number for the algorithm.

## Algorithm 5 - Multi-Layer Perceptron (MLP)

The final algorithm selected was that of a Multi-Layer Perceptron (MLP) neural network. This algorithm was selected as it can work with classification data. As with the previous algorithm, the author did not find an academic study which used this method for analysis. However, the author felt it was important to see if applying a neural network algorithm to the dataset would provide any insights to the study. While no hyperparameter tuning was applied to the data, the algorithm was run several times with different number of neurons in the hidden layers.

# Chapter 5: Implementation and Results

This chapter outlines how the various algorithms, python libraries were used to complete the analysis of the data. As with the previous chapter, implementation of each algorithm will be outlined on its own merits.

## Dataset

The OULAD data set was imported into Jupyter using the Panda’s library, where it was explored using numerous graphs and tables. The author included an additional column to the dataset as part of the EDA process to mimic employee tenure at the company. The author did not use any of the recommended tools for generation of synthetic data as there is no pattern to employee tenure. As such, the synthetic data created in this column was generated using random function randint, with a seed being set to keep the data consistent once created. The newly created tenure data was then grouped into bands to align with other categorized data in the dataset.

To follow the sampling strategy outlined as part of the research proposal, the author used the data from one semester and one course to complete the analysis. Semester 2013J was identified as the cohort that would more closely align with the number of employees at the manufacturing environment initially selected for evaluation. This resulted in a smaller dataset of approximately 382 rows of data with a new smaller dataset being created for ease of reference.

## Independent Variables

The independent variable selected for analysis was the data in ‘studied\_credits’. This column was selected as it more closely mimic’s employee interactions with training materials. Those with minimum interactions (such as only completing mandatory training) would have less credits earned than those with more credits would be seen as availing of the courses on offer more frequently. It is important to note here that employees are in control of their own learning journey, and trainings may be recommended by their manager as part of the employee’s own development.

Tenure … fill in here

Gender .. fill in here

## Algorithm 1 - Logistic Regression

As outlined previously, Logistic Regression was applied to the dataset, initially without hyperparameter tuning, then with GridSearchCV applied. The results of the analysis are outlined below.



Table 1 - Algorithm 1 Logistic Regression Results

## Algorithm 2 - Decision Tree

Based on previous research, Decision Tree was the next algorithm implemented. Initially the author used tenure as the independent variable and ran the model twice - including tenure grouped by band in version 1, and using tenure ungrouped in version 2. Both versions of the algorithm were run multiple times, altering the number of features that were selected each time. The result of this analysis is shown below. As can be seen, for grouped tenure four features need to be selected to attain the best accuracy for the model, while only one feature needs to be selected in the model using ungrouped tenure.



Table 2 - Algorithm 2 - Decision Tree Results (Tenure)

As ‘studied\_credits’ has been identified as the independent variable, the Decision Tree model was rerun using this as the target variable, as well as the ‘gender’. As can be seen below when hyperparameter turning was applied in the form of GridSearchCV, the result was more than that achieved using other variables.



Table 3- Algorithm 2 - Decision Tree Results (studied\_credits)

## Algorithm 3 - Support Vector Machines (SVM)

Continuing with algorithms used in previous literature, an SVM mode was the next one created. As with previous models, the initial model was run with no hyperparameter tuning applied. The Algorithm was changed with each of the four possible kernel’s being employed.



Table 4 - Algorithm 3 SVM

Applying GridSearchCV and param\_grid did not improve the accuracy of the results across any of the updated parameters used as part of the analysis. However, model output outlines that the following parameters are the best when completing the analysis.

Best Hyperparameters: {'C': 0.01, 'gamma': 0.001, 'kernel': 'linear'}

A close-up of a graph

Description automatically generated

Figure 5- SVM analysis with hyperparameter tuning applied.

## Algorithm 4 - Random Forest

The next algorithm to be applied was the Random Forest Classifier. As with previous implementations, the model was run first without hyperparameter tuning, focusing on the number of ‘trees’ within the random forest. The optimal number of trees for the first model is 150 trees, going up to 200 trees when hyperparameter tuning in the form of GridSearchCV is applied. The entropy criterion. The result for entropy shows to be the most accurate with 200 trees selected.

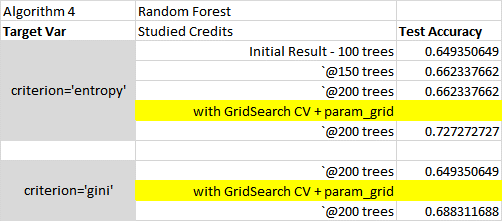


Table 5 - Algorithm 4 - Random Forest results.

The graph below ranks the important feature’s with GridSearchCv applied as hyperparameter tuning. Applying parm\_grid to the Random Forest with GridSearchCV identifies that:

Best Model Accuracy: 0.7012987012987013

Best Hyperparameters: {'C': 0.1, 'gamma': 0.001, 'kernel': 'linear'}

A graph with a bar graph

Description automatically generated

Figure 6 - Random Forest with hyperparameter tuning applied.

## Algorithm5 - Multi-Layer Perceptron (MLP)

The final algorithm selected for analysis was the Multi-Layer Perceptron or MLP neural network. The model was run with differing neurons in the hidden layers. The more neurons added to the layers, the more varied the results - see table 7 below.



Table 6 - Algorithm 5 - MPL implementation

|  |  |  |
| --- | --- | --- |
| **Test 1 - Loss** | **Test 2 - Loss** | **Test 3 - Loss** |
|  |  |  |
| **Test 1 - Accuracy** | **Test 2 - Accuracy** | **Test 3 - Accuracy** |
|  |  |  |

Table 7 - MLP implementation results of loss and accuracy

The addition of hyperparameter turning on test two was complicated by the decision to use the keras library. It was necessary to create a function to allow for turning to take place between the keras and scikit-learn libraries. This involved creating the model within a function, then wrapping it in the KerasClassifier from keras.wrappers.scikit\_learn, which is a standalone module that allow keras and sci-kit learn to work in tandem (Brownlee, 2016, 2022; ‘Hyperparameter tuning using GridSearchCV and KerasClassifier’, 2020). As with other algorithms, a list of variables was created using param\_grid with the intention of identifying the most appropriate variables that could be applied to the dataset. The output of the model is outlined in Table 8 in terms of accuracy and loss. It is clear to see that the application of hyperparameter tuning initially impacted on the accuracy of the model by the third epoch the model returned to a zero-accuracy result. In terms of the loss function, again the results are slightly better than previous models as the loss function gradually tuns to loss value rather than immediately doing so as its previous iterations.

|  |  |
| --- | --- |
| **Plot of Model Accuracy** | **Plot of Model Loss** |
|  |  |

Table 8 - MLP Results with hyperparameter tuning.

Overall, the best result for each algorithm selected is displayed in Table 9 below.

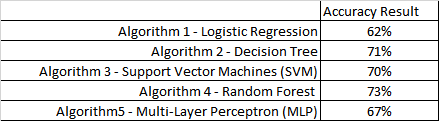


Table 9 - Best Accuracy Scores by Algorithm

# Chapter 6: Discussion

The methodology that was outlined in Chapter 4 was applied to each of the five algorithms selected for analysis with the implementation and results being discussed in the previous Chapter - Chapter 5.

# Chapter 7: Conclusion

# Appendix A: Workflow

# Appendix B: Interview Transcripts

# Appendix C: Data Permissions

# Appendix D: Consent Forms

# Reference List

## Algorithm 1 - Logistic Regression

## Algorithm 2 - Decision Tree

## Algorithm 3 - Support Vector Machines (SVM)

## Algorithm 4 - Random Forest

## Algorithm5 - Multi-Layer Perceptron (MLP)