**Employee learning data and demographic information as an aid in the 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:** | Employee learning data and demographic information as an aid in the 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:** | Friday, 22 September 2023 |

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 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 (Application Programming Interface) facilitating a connection to share basic data such as employee name and employee number as well as work email, manager, work area etc. 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 such as engagement.

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 is individualized and based on the reporters’ own experiences. At different times during the year, this information is collated, and meetings are held at functional area lead level who, in conjunction with and other senior managers to hold succession planning conversations. Succession Planning is a long-term planning strategy within HR, where every existing role within the company has a successor identified. It is an important planning process and helps the company develop internal talent as well as minimising business interruptions in the case where there is an unplanned departure (*Importance of Succession Planning (With Benefits and Tips) | Indeed.com Canada*, no date). 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 balance policies). With small datasets taking account of influences outlined previously, there is an opportunity to look at relationships in the data 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 OULAD - the Open University Learning Analytics Dataset which will be discussed in more detail in subsequent chapters. To aid their understanding of the succession planning process, the author was able to utilise the expertise of colleagues within the Human Resources community to discuss their experiences both at the current employing company and beyond.

# Chapter 2: Research Design

## Problem Identification and Clarification

The lack of any visible link between learning data and succession planning led the author to consider if some form of data analysis could be utilised to enhance the process. Experts in the area outline attempts to digitise the process by using standard templates to upload data to analytics tools used such as Power BI and Tableau (**Interviewee 1**). The author could not identify any attempts to create this link within the company, or the wider HR community, or even within literature. Indeed, the author could not find any literature where analysis was attempted on learning data within a work environment.

In the author’s own experience, employee interaction with learning systems can be mixed. Some employees complete mandatory learnings when assigned, others are assigned learnings as a development aid to their job. There is another category of learners that need to be considered though, and that is those who complete learnings based on their own interests, or to increase their knowledge of how the company works. It’s these outliers that interest the author as managers and human resources may not be familiar with either employees’ areas of interest, the learnings or are out the scope of their normal job. It is also possible that additional studies undertaken are done so outside the company, and additional learning is completed to support this.

There are other influences that can be analysed to determine their impact on an employee’s choice such age, gender, education level, access to education etc. The selected dataset from OULAD, has many of these features that can be used for such analysis such as age, education, result, gender, and the number of credits that are being studied for. There is a column for region which identifies where the student is from at the time that they registered for the course. Also included is the column imd\_band, which identifies the Index of Multiple Depravation that plots the areas in the UK based which are relatively deprived based on socio-economic factors (Alhakbani and Alnassar, 2022; ‘Multiple deprivation index’, 2023). The author had discounted the region and imd\_band columns, as in practice, employees generally live within commuting distance to the workplace, and as such this information is less relevant. The addition of a column on tenure will help mimic the relative experience the employee has with the company and will be included for analysis.

Taking all of this into account, the focus of this study will be on determining if learning data and demographic information for employees can be used as an aid in the succession planning process using data analytics to complete the analysis.

## Research Objectives

Now that the focus of the research project has been identified, it’s key to clarify what the research objectives are going to be support this analysis. The research project is focusing on the impact of learning data on succession planning using demographic features. Traditionally, demographic features are defined as age, gender, ethnicity, disability and family status (Tsui and Gutek, 1999; Clair *et al.*, 2019). It has to be acknowledged that current thinking within organisations show a more inclusive view of the definition of demographic traits such as gender fluid employees, whilst academia recognises that demographic traits are more fluid than previously defined (Clair *et al.*, 2019). The researcher acknowledges these definitions. However, the dataset selected for use as part of this analysis does not reflect such updated thinking, and standard demographic information will be used for analysis. This research paper will focus on identifying if there is a machine learning algorithm that is effective in accurately predicting with accuracy, the success of an employee in the succession planning process using traditional demographic information.

Specifically, the initial research objectives identified are:

**Objective 1** - Using is an employee’s gender a reference point for succession planning?

**Objective 2** - using studied\_credits as a substitute for number of courses completed, what machine learning algorithm will provide an accurate measure for succession planning?

**Objective 3** - Does employee tenure have an impact on a succession planning?

## 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 specifically within a multinational company. Such data collection was 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 author to gauge reactions to questions posed and adjust the flow of the interview depending on the interviewee’s reply 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 or 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 itself, or as part of their previous roles with similar 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** = the gender of the employee / student
* **Age** = grouped by age bands
* **Previous education** = highest level of education achieved by the student / employee.

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 within a defined period. It also allows the author to closely mimic the number of employees within a manufacturing environment.

## Validity Type

Considering the research objectives outlined in the previous chapter, it is possible to say that the most relevant components of validity, relevant to this research are accuracy, currency, and bias (Kumar, 2011; Saunders, Lewis and Thornhill, 2012). It is however also possible to say all components of validity apply to the proposed research, some component’s more than others. The author has attempted to minimise overall validity by rooting concepts and models utilised in this research within a academic literature as well as practitioners experiences (Kumar, 2011). The concepts of accuracy, currency, and bias are explored below.

Accuracy in this instance relates to how comprehensive the data statistically is (Kumar, 2011; Saunders, Lewis and Thornhill, 2012). 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 (Kumar, 2011; Saunders, Lewis and Thornhill, 2012). The author has chosen to use data extracted from an educational institute learning management system. The data was 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. With the addition of simulated data in the form of the tenure column the author has chosen to accept the risk to the validity of the results of this research paper. As part of the future work of this research, the author advocates the need to compare results of this research with real world data.

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 to gain their support in the research process. As an added measure, all interviews have been transcribed for completeness and included in **appendix XX** of this document. If any participants have 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 to data protection and sensitivity issues the author decided to use a dataset obtained from the Open University (OULAD) (Kuzilek, Hlosta and Zdrahal, 2017). 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 been anonymised, thus limiting any potential data breach. The General Data Protection Regulations (GDPR) outlines companies’ requirements to protect the private data of individuals (*General Data Protection Regulation (GDPR) – Official Legal Text*, no date). 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 is a key reason that OULAD was selected for analysis. As such, the author has attempted to minimise potential risks to the company’s data, whilst also maintaining compliance with the companies own internal GDPR and ethics procedures.

# Chapter 3: Literature Review

## Human Resources and data analytics

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).

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, Diez et al (Rasmussen and Ulrich, 2015; Diez, Bussin and Lee, 2019; Mattox II, Parskey and Hall, 2020; Ferrar and Green, 2021). 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. Diez et al (2019) in their book Fundamentals of HR Analytics outline how questions for HR are evolving and why there is now a need for HR speak the same language as senior management and other functions within the company. The specific fields within HR are called out in Figure 1 below.

A white rectangular box with text

Description automatically generated with medium confidence

Figure 1- Charting the change in management requests for HR. Source: Diez et al (2019) page 5

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 businesses 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).

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).

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), confirms 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).

## Learning Analytics

Building on this, learning analytics on the other hand, focuses on the effectiveness of a learner’s experience (Mattox II, Parskey and Hall, 2020; ‘Handbook of Learning Analytics - Second edition’, no date) 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, an area with limited research carried out in (‘Handbook of Learning Analytics - Second edition’, no date). Implementing a Learning Management Systems (LMS) has provided an effective way of gathering, analysing, and reporting on learning related data (Katrina Sin and Loganathan Muthu, 2015; Mattox II, Parskey and Hall, 2020; Karakolis *et al.*, 2022; ‘Handbook of Learning Analytics - Second edition’, no date). 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, and are becoming the ‘backbone’ of analysis within companies (Sin and Muthu, 2015, Arka et al 2022, (Katrina Sin and Loganathan Muthu, 2015; Mattox II, Parskey and Hall, 2020; Karakolis *et al.*, 2022; ‘Handbook of Learning Analytics - Second edition’, no date).

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 multinational environment. The literature review of academic and related papers has helped to uncover several opportunities for further analysis of data held within the HR Department focusing specifically on data relating to learners. The author could uncover very little research into learning data held outside of academia with Allison Littlejohn provided the closest match in her chapter on Professional Learning Analytics as part of the Handbook of Learning Analytics (Handbook of Learning Analytics - Second edition’, no date). Most articles found outlined analysis based on data from educational institutions **REFERENCES HERE**. How analysis of education related data should be conducted was, as expected, discussed at length with different approaches being taken as outlined in Figure 5.

Focusing on data gathered as part of the learning process and how such analysis might be completed is discussed in the following sections. 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 of study (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 an employee’s 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). 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 previously, the succession planning process is critical to the business’ ability to develop its employees (*Importance of Succession Planning (With Benefits and Tips) | Indeed.com Canada*, no date). 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).

## Data Analysis of HR Data

Outlined in the previous section is the need for HR to become more integrated and aligned with Senior leaders within business (Ferrar et al. 2021, Mattox et all 2020, Rasmussen and Ulrich 2015, Diez et al (Rasmussen and Ulrich, 2015; Diez, Bussin and Lee, 2019; Mattox II, Parskey and Hall, 2020; Ferrar and Green, 2021). How can this be accomplished though?

Tambe et al (2019) put forward the argument that data within the HR Department generally contains small datasets which may not be suitable to use to clearly identify relationships or meaningful insights within the dataset (Tambe, Cappelli and Yakubovich, 2019). This fact is further complicated by any decision to use historical data for analysis, with the danger being that historical HR data may 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 identify models and potential relationships in the data (Tambe, Cappelli and Yakubovich, 2019).

In addition to bias, It is important to note that in some fields it is not ethical to seek relationships between variables, especially in scenarios where there may be moral considerations to be taken into account (Eberhardt, 2017; Malinsky and Danks, 2018; Vowels, Cihan Camgoz and Bowden, 2023). Bankins et al (2021) has proposed a framework to help with the ethical implementation of artificial intelligence within an organisation with scope to use for implementation of any type of machine learning or predictive analysis (Bankins, 2021).

Working forward, Muslim et al (2023) completed a review of literature on ‘Open Learning Analytics’ outlining key frameworks within the field of learning analytics (Muslim, Chatti and Guesmi, 2020). That paper showcased that in terms of implementation, when applying machine learning algorithms for learning analysis, there are a myriad of implementations possible, and indeed in use (Muslim, Chatti and Guesmi, 2020). Techniques used by authors across the research area are outlined in below:

* clustering / classification techniques (Araka *et al.*, 2022; Mustafa Yağcı, 2022)
* data mining based on time data (Araka *et al.*, 2022; Assaad, Devijver and Gaussier, 2022)
* Random forests / Support Vectors / Naïve Bayes techniques (Hussain *et al.*, 2018; Mustafa Yağcı, 2022; Zhang *et al.*, 2023)
* Neural networks such as MLP, ANN and CNN (Kulala and Rani, 2017; Chunqiao Mi, 2019; Poudyal, Mohammadi-Aragh and Ball, 2022; Wang, Guo and Shen, 2022; Al-azazi and Ghurab, 2023)

An example of where HR data has been implemented within an integrated approach to analysis including the use of neural networks, artificial intelligence, has been proposed by Dong (2022) for an integrated system called the Human Resource Intelligence System (IHRMS) (Dong, 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 chosen.

## 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.

# Chapter 4: Methodology

Chapter 3 provides the strong foundation on which the methodology for this research paper is based. The author reviewed the processes that several different research papers had previously taken when using the OULAD dataset, amongst others. This provided a launch point upon which this research could be built upon. Specifically, the author was able to identify models already used to explore the dataset, as well as the research focus of the article. Generally, research was dedicated to developing models predicting student performance or engagement with the systems within a learning environment. This research paper however is concerned with learning data accumulated within the work environment, where learning opportunities may be more limited or directed by the company rather than only at the interest of the learner. In addition, previous research focused on interaction with the system by students as a method of assessing engagement or performance. Neither of these areas are a focus for this research paper. However, the author felt an opportunity existed to use prior learning and apply it within this different area of study.

## Dataset

As outlined in previous chapters, the dataset used for analysis was sourced from the Open University Learning (Kuzilek, Hlosta and Zdrahal, 2017). 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 proprietary or otherwise sensitive data.

In terms of the dataset, it was necessary to apply filters to reflect the employee numbers within the target company. As outlined in the sampling strategy adopted for this research, the author is proposing to filter the dataset so that it reflects the employee makeup at the selected operating site. To implement this, the data will be filtered by module\_code, and again by semester. This will allow the author to select the best representative sample similar in number to the employment location.

As outlined previously, 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 within a defined period. It also allows the author to closely mimic the number of employees within a manufacturing environment.

## 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 .

* Pandas - used for ‘data wrangling’ or manipulation of data within a dataframe (M.Sc, 2023a).
* NumPy - used for calculations quickly and easily (M.Sc, 2023a).
* Matplotlib - helps display graphs and visualisations of the data using pandas and numpy (M.Sc, 2023a).
* seaborn - built to work alongside Matplotlib, the Seaborn library allows for statistical data to be graphed and displayed (*An introduction to seaborn — seaborn 0.12.2 documentation*, no date).
* Scikit-Learn - allows a programmer to quickly implement a range of machine learning algorithms in conjunction with other libraries such as pandas and matplotlib (*scikit-learn: machine learning in Python — scikit-learn 1.3.0 documentation*, no date).
* Scipy - enables a programmer to implement statistics and other mathematical computations with python (*SciPy*, no date).
* Keras - facilitates deep learning within Python by working in conjunction with TensorFlow to provide a simple but flexible platform for analysis (Team, no date).

## 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 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 its 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. The models selected can work with classification data, contained within the OULAD dataset.



Figure 4 - Articles and Models reference table

## Algorithm 1 - Logistic Regression

The Logistic Regression method 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 then 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; Zhang *et al.*, 2023), 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.

## 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 (for example into 3D to better view the data). Hyperparameter tuning was applied following the first run in the form of GridSearchCV as recommended by Scikit-Learn, and further tuning was applied using a parameter grid (param\_grid) as a further addition to the parameter search to allow for a wider sequence of parameter settings to be included such as ‘C’ and ‘gamma’. The cross-validation parameter in this instance will be 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 criteria 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 a parameter grid will be implemented to help identify the most appropriate tuning methods based on the results obtained when working with algorithm two. The model will 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 is that of a Multi-Layer Perceptron (MLP) neural network. This algorithm was selected as it can work with classification data and is also relatively simple to implement as a baseline neural network. As with the previous algorithms, the author did not find an academic study which used this method for analysis in similar circumstances. 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 to identify if an impact can be seen.

# Chapter 5: Implementation and Results

This chapter outlines how the various algorithms and 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 first variable selected for analysis was that of gender to study if an employee’s gender has any impact on learning and may be useful in helping with the area of succession planning. Note that this data set only records gender as being male or female. The author recognises that more gender types are now common and not reflected in the selected dataset. When implementing the algorithm using real or company data, it will be necessary to reflect employee gender types as they appear within the company’s employee database.

The second 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, or they may source courses on their own initiative.

The final independent variable selected is that of tenure, which reflects the length of time the employee has been with the company. This is synthetic data and added using a random number generator as defined above.

## 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. As the first model run, the test and train results are more than 50% accurate and will act as a base level for test following afterwards.



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 identified in previous literature, an SVM mode was the next one created. As with previous models, the initial run through was completed with no hyperparameter tuning applied. The algorithm was changed with each of the four possible kernels’ being employed. This was completed manually to confirm the best kernel and baseline results.



Table 4 - Algorithm 3 SVM

As with the other algorithms, GridSearchCV and yet another parameter grid was applied to the model, but an improvement in the accuracy of the results across any of the updated parameters used as part of the analysis was not seen. However, model output below show how each kernel performed once parameters were identified as part of the analysis.

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 result for entropy is most accurate with 200 trees selected.

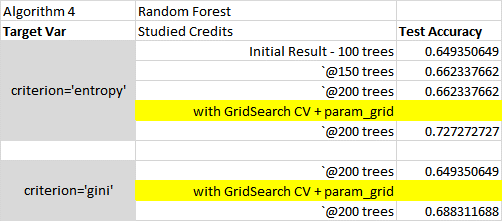


Table 5 - Algorithm 4 - Random Forest results.

The graph below ranks the important feature’s with GridSearchCv applied as hyperparameter tuning. Applying a parameter grid to aid in the search for the Random Forest with GridSearchCV identifies that the best parameters to be applied for the analysis. Figure 7 (below) identified the most important features in the model. Education level appears to be the least important feature withing the model.

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 6 below.



Table 6 - Algorithm 5 - MPL implementation

|  |  |  |
| --- | --- | --- |
| **Test 1 - Loss** | **Test 2 - Loss** | **Test 3 - Loss** |
| A graph with a blue line  Description automatically generated | A graph of a graph  Description automatically generated | A graph of a graph  Description automatically generated |
| **Test 1 - Accuracy** | **Test 2 - Accuracy** | **Test 3 - Accuracy** |
| A graph with a line  Description automatically generated | A graph with a line  Description automatically generated | A graph with a line  Description automatically generated |

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 a parameter 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** |
| A graph with a line  Description automatically generated | A graph with a line going up  Description automatically generated |

Table 8 - MLP Results with hyperparameter tuning.

## Analysis of Selected Variables

Once all the algorithms had been tested, the versions with the best results were run against the independent variables selected as part of the research paper. The results for each algorithm by the relevant variable are displayed in Table 9 below and will be discussed in detail in the following chapter.

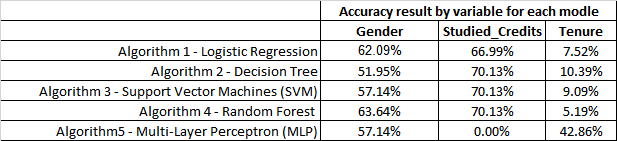


Table 9 - Accuracy Scores by Algorithm and by Target Variable

Reviewing the results of the analysis of all the algorithms it is possible to make some observations. Looking each variable on its own, for example, Gender, Algorithm 4 Random Forest provides the highest level of accuracy, whilst Algorithm 2 - Decision Trees is least accurate at 51.95%. For studied\_credits, 3 algorithms give the same level of accuracy - namely Decision Trees, Support Vector Machines and Random Forests. Algorithm5 on the other hand, the MLP network fails to give any degree of accuracy in respect of its results. In contrast, the variable for tenure has its strongest performance in MLP, whist results in all other variables are well below, with just one result (Decision Trees) achieving an accuracy result of just over 10%.

The initial results in Table 9 show that three Algorithms, 2 and 4 received the same accuracy result for studied\_credits. The author sought advice on how to test for statistical alignment or difference. The dataset was not suitable for an ANOVA analysis to be carried out - mainly in terms of the randomness of the sample (*One-way ANOVA - An introduction to when you should run this test and the test hypothesis | Laerd Statistics*, no date). A further subset of the OULAD dataset was created using the same sampling strategy as before - taking one module over one semester as the sample group. The idea being that for the test group of data to confirm if there is any clear over of under fitting of the data, or some other anomaly easily shown. The test group of data was larger than the initial module. The results of the initial run and the test results are outlined below.

A yellow and black data

Description automatically generated

Table 10- Initial results v's Test results

As can be seen, the output of studied\_credits are again consistently the same in results for algorithms 2 and 4. Algorithm 3 also has the same results. The gender variable performs much better in the test dataset than the initial run. This may be down to a larger dataset being used for analysis. The outcomes of the results will be discussed in more detail in the following section.

# 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 Chapter 5.

Initial thoughts viewing the results of the overall analysis in Table 10, the most successful algorithm was Algorithm 4 - Random Forest as it achieved the highest results for two variables - notably gender and studied credits. For the tenure variable, Algorithm 5 was the most accurate, although results were not above the 50% accuracy level. However, it’s necessary to review the analysis based on the research questions posed in chapter 2.

## Objective 1 - Gender Variable

As a brief refresh, this objective asked if an employee’s gender would provide a reference point for data analysis carried out as part of the succession planning process.

Using the initial dataset, and then the test dataset, Random Forest proved to be the most accurate predictor of success using gender, with a 63.64% result in the initial test, and 79.33 % when the test data was applied. What is interesting about these results is that Algorithms 4 and 5 gave the same percentage accuracy during initial analysis. The difference in results between the initial data and the test data could be because of scale, with the initial dataset being more realistic in style for the environment proposed for the completed model. As part of future work in respect of this research, the author would propose to confirm within the Jupyter notebook if the data is overfitting or underfitting and adjust accordingly. Strangely, the SVM algorithm performed adequately in the smaller dataset, but with the least amount of accuracy with the larger test dataset, despite using a parameter grid to identify the most appropriate variables for analysis.

Figure 7 - Graph of Results by Gender.

**Limitations** that were identified in respect of this variable are outlined below.

* It is not possible to compare results gained with any previous work or research in this area, as the author was unable to locate a similar study.
* Using openly available data set - OULAD- as this is not fully reflective of the employee population within the work environment. The timeframe is also not current, as the contents of the dataset was released in 2017 but relate to the time frame 2013 to 2014.
* Focusing on demographic data such as gender / studied\_credits / tenure could unnecessarily import a bias into interpretation of results with, in this instance, a larger number of males being present over females. The dataset also does not consider the wider definition of gender that is accepted with HR circles in 2023.
* Limitations within the algorithms selected for analysis, namely.
  + Algorithm 1 - Linear Regression -the inclusion of too many features may affect the outcome of the model.
  + Algorithm 2 - Decision Tree - does not work well with outliers, which are key to the analysis of this dataset.
  + Algorithm 3 - Support Vector Machines (SVM) - works best for binary classification, so it may struggle with non-binary classification models.
  + Algorithm 4 - Random Forest - this model was slow to run, due to the size of the dataset. Once the Test dataset was applied, it took a noticeably longer time for the results to be displayed.
  + Algorithm5 - Multi-Layer Perceptron (MLP)

**Future work** identified within this research objective include:

* As outlined previously, opportunities exist to repeat the analysis taking into overfitting or underfitting of the data.
* The findings will also need to validate the findings against real data from the company as opposed to a freely available open dataset.
* Using Causal Discovery to investigate if causal algorithms, identified as part of this research paper will be a potential method of analysis for HR data as it apt in analysis of small datasets where expert knowledge is important in terms of interpretation.

Based on the points outlined above, the author can confirm that it is possible to accurately predict succession planning using gender as the target variable.

## Objective 2 - Studied\_Credits Variable

To recap, objective 2 of this research papers queries if it is possible to use studied\_credits as a substitute for the number of courses completed, and to identify what machine learning algorithm would provide the most accurate measure for succession planning.

Figure 8 below shows the accuracy results with studied\_credits as the independent variable for use in the selected algorithms. In terms of the initial results, algorithms 2 and 4, Decision Trees and Random Forest respectively, provide the most accurate results at 70.13%. The most inaccurate result was received by algorithm 5 - MLP which returned a 0.00% accuracy which can be interpreted as being either an issue with the data on which the model is run, or that the model is a poor choice for analysis. As other variables within the dataset achieve a much higher accuracy result, the author can conclude that the model is not a fit for this variable.

When checking for

Figure 8 - Graph of Studied\_Credits.

## Objective 3 - Tenure Variable

**Objective 3** - Does employee tenure have an impact on a succession planning?

* Weakest indicator

# Chapter 7: Conclusion

This research paper was undertaken to determine if it would be possible to use learning data to support the succession planning process. Three research questions were proposed to direct the research, could demographic and learning data be used as the basis for analysis.

future work / application

* need to validate the finding against real data from the company
* investigate if causal discovery could be applied to this type of data
* potentially used to support talent discover in an alternative way such as for short term assignments / experiences to increase employee experiences.

# Appendix A: Plan of Work

# Appendix B: Interview Transcripts

# Appendix C: Sample Consent Form

Actual signed consent forms available on request.

# 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)