UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Data Analytics in HR Management

Diploma Thesis

Ioannis Astli

Supervisor: Michael Vassilakopoulos

February 2023

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Data Analytics in HR Management

Diploma Thesis

Ioannis Astli

Supervisor: Michael Vassilakopoulos

February 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Αναλυτική Δεδομένων στην Διαχείριση Ανθρώπινων Πόρων

Διπλωματική Εργασία

Ιωάννης Άστλι

Επιβλέπων: Μιχαήλ Βασιλακόπουλος

Φεβρουάριος 2023

Approved by the Examination Committee:

Supervisor **Michael Vassilakopoulos**
Professor, Department of Electrical and Computer Engineering, University of Thessaly

Member **Panagiota Tsompanopoulou**
Associate Professor, Department of Electrical and Computer Engineering, University of Thessaly

Member **Eleni Tousidou**
Laboratory Teaching Staff member, Department of Electrical and Computer Engineering, University of Thessaly

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

«Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I also declare that the results of the work have not been used to obtain another degree. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism».

The declarant

Ioannis Astli

Diploma Thesis

Data Analytics in HR Management

Ioannis Astli

Abstract

In the era of digitization, **Human Resources** (HR) is in the glimpse of a structural change. Managers, by making use of sophisticated algorithms and **Artificial Intelligence** (AI), have the opportunity to give a stronger position in HR Management by making use of the ever growing Big Data. Translating people as mere data, consists of a counter productive move, thus in this thesis we build the argument that with careful and thorough examination of an algorithm, results that can optimize essential HR functions can be achieved. By further presenting the available tools, and choosing to implement a supervised regression model, we achieve to predict the **turnover** of a virtual company by the IBM data set. In the process we define the base variables that contribute in the effective prediction, via the use of a **Random Forests** implementation. Finally giving us the opportunity to potentially suggest a strategy that can be applied in order to increase retention.

Keywords: Human Resources, Human Capital, Analytics,People Analytics, Data Mining, Statistical Regression, Machine Learning, Artificial Intelligence, Random Forests, Turnover

Διπλωματική Εργασία

Αναλυτική Δεδομένων στην Διαχείριση Ανθρώπινων Πόρων

Ιωάννης Άστλι

Περίληψη

Καθώς διανύουμε την ψηφιακή εποχή, η διαχείρισή **Ανρθώπινου Δυναμικού** (ΑΔ) βρίσκεται σε φάση δομικής αλλαγής. Οι διευθυντές, με τη χρήση εκλεπτυσμένων αλγόριθμων και της **Τεχνητής Νοημοσύνης** (ΤΝ), έχουν την ευκαιρία να ισχυροποιήσουν τη θέση του τμήματος Ανθρώπινου Δυναμικού με την εκμετάλλευση του όλο και αναπτυσσόμενου τομέα των δεδομένων. Το να εκφράζουμε τους ανθρώπους ως απλά δεδομένα αποτελεί ένα αντιπαραγωγικό εγχείρημα, για αυτό σε αυτήν την εργασία επιχειρηματολογούμε ως προς τη προσεκτική και εκτενή εξέταση των αλγόριθμων, φαινόμενο που αποσκοπεί στην επίτευξη της βελτιστοποίησης θεμελιωδών διαδικασιών του ΑΔ. Συνεχίζουμε με την παρουσίαση των διαθέσιμων εργαλείων, και επιλέγοντας ένα επιβλεπόμενο μοντέλο παλινδρόμησης, καταφέρνουμε να προβλέψουμε τον δείκτη κινητικότητας και αντικατάστασης υπαλλήλων σε μια εικονική εταιρεία, χρησιμοποιώντας το σετ δεδομένων της IBM. Κατά τη διάρκεια της διαδικασίας ορίζουμε τις βασικές μεταβλητές που επιδρούν στην αποτελεσματική πρόβλεψη, με τη χρήση των **Random Forests**. Στο τελικό στάδιο, έχουμε την δυνατότητα, υποθετικά να προτείνουμε μια στρατηγική, όπου με τη χρήση της θα αύξανε τον δείκτη διατήρησης προσωπικού.

Λέξεις-κλειδιά: Ανθρώπινο Δυναμικό, Ανθρώπινο Κεφάλαιο, Αναλυτική, Αναλυτική Ανθρώπων, Εξόρυξη Δεδομένων, Στατιστική Παλινδρόμηση, Μηχανική Μάθηση, Τεχνητή Νοημοσύνη, Random Forests, Δείκτης Κινητικότητας Υπαλλήλων

Table of contents

Abbreviations

ML	Machine Learning
HRM	Human Resource Management
HRIS	Human Resources Information System
AI	Artificial Intelligence
RF	Random Forests

Chapter 1

Introduction

In the era of data, HR meets the challenge to become a part of a system that changes day by day. By integrating sophisticated analytical techniques and tools, HR has the opportunity to become a true mechanism of an organization, not only in operational level, but as a strategic mechanism as well. In order for that to be achieved, several complications should be surpassed.

First, is the technical capabilities of the HR personnel. Mainly HRs are professionals from a theoretical background, with an enhancement in psychology and socio-economic studies, thus not having the needed statistical background, analytics require. Of course this can be surpassed with according studies and training. Though it needs to be implemented in college level, so as the professionals of tomorrow to be equipped with knowledge that will enhance the capabilities of HR departments. To be fair, there is not needed for every HR to be accustomed in sophisticated mathematics and statistics, but a strong basic knowledge is a prerequisite for departments and managers to communicate in an organizational level. In addition, professionals now can be trained in new techniques in order to elevate their capabilities and tackle the problem of digitization. With single training in a corporate level, their background in Human Resources can give extensive meaning in the data they acquire, as well as means to manipulate and extract further observations from them.

Second, data of human capital are in general not in a good form. Mostly unstructured and not easily comprehensible, making the job of analytics difficult and in general slowing the whole process. Other than that, one of the strongest debates in human resources, is that employee data are easy to succumb in privacy issues. Most of the times analysts are entangled in a process where from the one hand they have to be careful not to dig too deep, and on the other hand to give thorough analytics of the employees from a given organization. Despite that, the problem of the data sufficiency is what complicates HR Analytics even in academic level, since the data sparsity makes the research level limited, because organizations are not allowed to disclose these sensitive material. All of this, add up to make solutions in HR Analytics more personalized and company-driven. Which from a point of view is good enough. But that it is, good enough. Without further exploration in academical level, there cannot be substantial evolution and harness the full power of HR Analytics.

Third, the trust that is given to the existing solutions from the managers. In most organizations, HR does not have a significant role in the strategic decisions of a company's course, even though the human capital is maybe the most important asset of an organization of a company. The old ways of management are quite strong even now, and executives are hard to trust analytical tools in order to progress in decisions based on algorithms. Which, from a point of view is quite understandable, as Artificial Intelligence in a plethora of circumstances has shown a quite biased perspective in exploring human data. In several occasions, AI algorithms tended to be even racist, thus compromising the whole process. Despite that, in order something to be in working form needs to pass several levels of research, and especially AI that gets better from the continuous training. Managers need to include HR officials in strategic decisions, as HR data can give insights that organizations can benefit further from them. And with the continuous development in Machine Learning and Analytics, better understanding and good usage of employee data can become a beneficial asset for both employees and organizations.

1.1 Subject of Thesis

In this thesis we begin by building up the whole background of how HR operates. By giving an extensive view in the processes and the way management is structuring the sector amidst an organization. The definition of what is the the role, and it stands is of critical importance in order to understand the role and the potential, as technology evolves.

By presenting the various tools, that a manager can have in her disposal, the role of employees and management acquires a new meaning in their cooperation, as an organization evolves. Besides, the role of human capital is important, if not the most, in an organization. Thus by optimizing the processes of employee management, an organization benefits in the long run.

Despite the human capital per se, we theorize upon what consists good leader, and try to picture the managerial decisions, as human centered, with the help of the available tools, analytics and machine learning provide. Furthermore, it is drawn a line, upon the usage of computer algorithms remain friendly for the data they manipulate, without crossing the line of human exploitation.

Either way, technological evolution is a phenomenon that is more difficult to grasp, day by day. As of its tremendous speed, managers are needed to stay on track and learn how to use effectively the ever growing data.

1.1.1 Contribution

Surpassing the theory, a thorough presentation upon the programming language R is made as it consists on of the best tools for statistical programming. The variety of packages and the effectiveness of its structure, makes the job for programmers and analysts better and optimized. Complementing this, the presentation of several machine learning algorithms is made, as to give the reader a complete view. Thus giving the reason why Random Forests was chose for the practical implementation of this thesis.

Finally a model for predicting turnover, in the IBM HR data set, is presented in the final section. The results that were produced, amplify the main argument of this thesis, that by having analytics optimize the HR processes, management has the room to implement strategic decisions, eventually boosting tenure, and taking place in the organization evolution. The contribution of this thesis is presented as followed :

1. Theorizing about what consists a good leader, with multiple abstract arguments, thus setting the stage for the appropriate usage of analytics

2. Setting a clear view about the contribution of computer algorithms, thus eliminating the disunity about their use in human processes

3. Establishing strong arguments about the benefits of using analytics in HR processes

4. Thorough presentation upon the programming language R

5. Comparison of different R dialects and arguing about each one's pros and cons

6. Presentation of basic machine learning regression algorithms.

7. Make the case for the choosing of Random Forests Algorithm
8. Building a statistical model for Turnover prediction
9. Theorizing about the contribution of the model, and the potential implementation of a strategical plan

1.2 Volume Structure

Main operations,structure and potential of Human Resources in the digital era, in addition main difficulties that are observed are found in Chapter 2. In Chapter 3, the definition of People Analytics takes place, as well as what it really means for people to be evaluated as data. Analytics are presented for the solutions they can provide in key fields of HR, thus being transformed in a predictive toolbox in Chapter 4. Whereas in Chapter 5 is presented, one of the strongest tools that can have an analyst in his disposal, which is the programming language R. To evolve the capacity of the HR optimization, statistical concepts and machine learning algorithms are given as a solution model in Chapter 6. Finally all of the previous chapters, add up to the creation of a sophisticated statistical model, that manipulates the IBM data set and predicts the turnover of a virtual company in Chapter 7.

Chapter 2

Human Resources Management

2.1 Operation in the era of data

In the base of this thesis lies the definition of human resources management as the central actor that gives power to the science of data and statistics. Human resources management nowadays is becoming slowly a science. As more and more organizations fill the void of intelligence in the field of people management with analytics, HRM gains a new perspective in organizational, as well as in academic level. By implementing this definition, we can further examine how in our age HRM is gaining more interest, as big data and machine learning become an inseparable tool of today's managers.

Of course there is a lot of skepticism around the usage of computer processing regarding human capital, and as we will examine the solutions analytics provide, we will see the pitfalls that are hiding as well. Nevertheless, data are everywhere, and they are vast. In order to maintain advantage, today's managers have to learn how to use them, and how to navigate in this unexplored world of human data. Because as we will see, the opportunities that lie ahead, are countless.

2.1.1 Human Resources Management: A general notion

We should begin by defining the role of HR in an organization and how this sector works. As Snell and Morris put it, in their book Managing Human Resources (18th Ed),"human resources management is the process of managing human talent in order to achieve the organizations objectives" [1] .

In a more practical manner, HRM is far more. From acquiring a new talent, to train it, engage in organization's culture, until the newcomer to establish a connection with the whole organization's community. HR is responsible for that person to be compensated fairly, acquire knowledge and return the investment to the given organization. Also, HR department is responsible for the member to be heard and give solutions, if a member has trouble inside the organization or needs to advance and grow. Despite that, HR has a role in strategic decision making. As the connection between employees and executives, is responsible to promote ideas in the objective an organization has, by extracting intelligence and managing the human capital for the benefit of the organization [2].

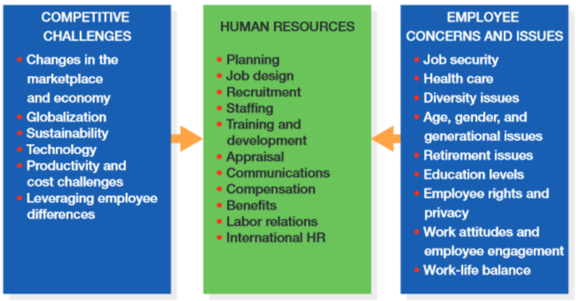


Figure 2.1: HRM Framework [1]

To understand better the role of HR, we need to understand how an organization works, even though this is not our case in this thesis to explain foundational definitions of how an organization stands in the business world, we need to see the main problems HR solves in our age. There is a need for clarification, because the management of human capital has changed in our age. Not so long ago, HR had a more administrative role [2], in contrast with the present, where it engages with more active processes inside a company.

2.1.2 HR Manager's Competencies

In order for HR to be productive and beneficial for the organization, it's managers need to be competent in the below sectors [1]:

Business Mastery: HR professionals must have a thorough understanding of the operations and strategies of their organizations. To help a firm shape and achieve its strategic direction and adjust it as necessary, this necessitates an understanding of an organization's customers and economic and financial capabilities.

HR Mastery:The behavioral science experts in the company are the HR professionals.They ought to become experts in hiring and training, as well as have strong interpersonal skills.

Personal Credibility:HR specialists must develop their personal credibility with both internal and external stakeholders of the company, just like other management professionals.

2.1.3 HR Manager's Responsibilities

On the other hand though, HR managers need to comply with a set of responsibilities as well, and as they are clearly defined by Snell and Morris, we understand the connection that bridges the qualifications and the tasks that need to accomplish.

Strategic Advice and Counsel: HR managers are an invaluable resource for making decisions due to their familiarity with internal employment data and productivity metrics as well as their awareness of external trends like economic and unemployment data as well as new legal and regulatory issues. Chief compliance or ethics officers help employees navigate murky waters when it comes to right and wrong in some businesses, usually larger ones. Also they make sure employees abide by the rules and laws that apply to their respective industries.

Service: Recruiting, choosing, testing, and planning/conducting training programs are just a few of the services HR managers provide. For the purpose of designing and implementing talent-management programs, HR managers must have technical expertise in these fields.

Policy Formulation and Implementation: To address recurring issues or foresee issues, HR managers typically suggest and draft new policies or revisions to existing ones.

Employ Advocacy: In order to ensure that the interests of employees and those of the company are compatible with one another, one of the enduring roles of HR managers is to act as an employee advocate. In this capacity, they listen to employees' worries and represent their needs to managers.

2.1.4 From a Manager to an HR Manager

As we defined above the main aspects of an HR Manager, we begin to establish how the role inside an organization stands. A good HR Manager needs to fuse the planning of their department, with the strategic planning of their organization. This process is achieved, with the strategy formulation and execution. At first, the available resources (talents) need to be taken under consideration, and with the appropriate formulation to match the potential of a task to be achieved with the availability of the human capital. Countless examples from the business world show us the importance of the organization's employees, because without them no strategy can be planned and executed [1].

2.1.5 A new tool: HR Information System

As technology is an inseparable mean of our age, HR in today's corporations has its own information system. More and more companies comply with this practice, either making their own software tools, or maintaining one by third party firms. Human resources information systems are used for everything, from automating payroll processing to administering benefits programs. The systems allow managers to access employee records for administrative purposes and employees to access and change their own benefits and other personal information on either an intranet or a secure website.Firms use human resources information systems to recruit, screen, and pretest applicants online before hiring them as well as to train, track, and promote employees once they have been hired.

One of the newer HRIS applications is the use of big data. Big data is a buzzword that describes the massive amounts data available online and offline today that can be "crunched" to make decisions. Marketing departments have very successfully used big data to detect people's buying patterns. By analyzing its customers' buying habits, Target was able to predict which of them were pregnant, sometimes before they had even told their families. The company then sent the customers ads and coupons for baby products [3].

Now companies are doing the same thing to analyze HR information, a process that's referred to as workforce (HR) analytics. Using HR data, such as employee demographic information, performance ratings, pay, employee surveys, academic history, years of service, and so on; a firm can definitely answer various questions about their employee's performance.

Traditionally, questions such as these have been answered based on anecdotal evidence or the "gut" feelings of HR professionals. But workforce analytics can provide more definitive answers and provide further insights, that stand according to facts. Although the biggest fuzz is made by the gathering of the HR data, many software providers, include data analysis tools in order to optimize the process. Such tools, are scorecards, dashboards, etc.

2.1.6 HRIS and Big Data

As mentioned above big data is the logical evolution of the HRIS. Companies gather tremendous amount of data from their employees, that are not in complete competence to understand and benefit from them [4].

Big data is gradually reshaping the field of HRM. Day by day, vast amount of data is gathered, with the need to maintain a pace in this change to be of huge importance. The HR

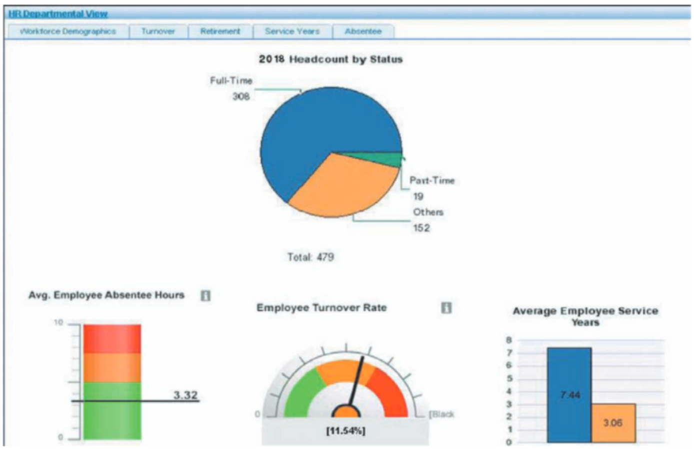


Figure 2.2: An example of an HR Dashboard [1]

department will initially experience changes in how it operates, and it won't be until after this change that all employees within the company will feel the effects of big data. As a result, the influence at the human resource level comes before the influence at the human level, even though the HR department itself already has an influence at the human level [5].

Although HRM's technological competency is currently underdeveloped, we could argue that the connection between HRM and big data is something that aspires great interest. As we examine the possibilities of the data that are being provided, by employees, organizations via HR can mutually benefit and optimize the main challenges that are blocking HR to be a strategical component of an organization. From analyzing employees' behaviour for example, and pushed their employees' to maintain a positive attitude towards customers, Starbucks achieved to establish one of the best corporate cultures and thus see an increase on their revenue [3].

There are a lot of examples of how corporations by extracting useful insights of their people's data, achieved an increase in productivity at first and on revenue in second phase. Another example, of how a corporation can benefit from their data, is that of Alcoa. Paul O'Neil, Alcoa's CEO, by examining employees' safety data, decided to put safety first and maintain a no-accident policy. By cooperating with managers in various positions in the corporate ladder, established a system where every employee could declare a problem or had the choice to immediately stop when he judged that it was not safe to continue. By giving power at first, to the data that was there, and later to the employees, Paul O'Neil achieved to make Alcoa one of the safest organizations to work, and nearly doubled its stock price [3].

HRM plays an integral role in organizations and the general strategy. There is a fit between the work of the HR department and the strategy implementation within the organization; consequently, strategic HRM contributes towards the competitive advantage of an organization [6]. Therefore, if the HR department is a source for strategic decisions and, by that, contributes to the competitive advantage, this HR department needs to have a high differentiation in its architecture [7]. Furthermore, Becker & Huselid mention that in order to contribute towards the strategic direction and the potential competitive advantage, the HRM focuses on its system rather than operational tasks. The strategic goal of HRM is to contribute to a sustainable competitive advantage. However, the focus will not purely lie at the organizational level but also at the individual level [8]. Strategic HRM is, therefore, a link between the strategic direction of the organization and the impact of such strategy at the individual level.

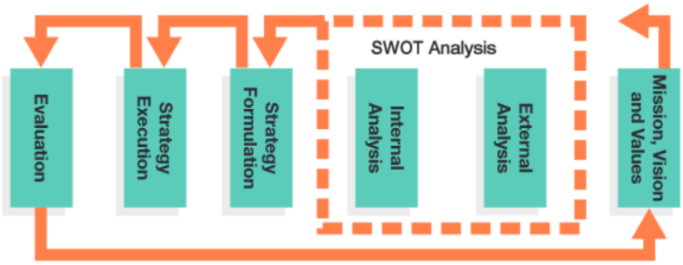


Figure 2.3: SWOT Analysis on HRM [1]

Due to the reason that big data influence the organization extensively, strategic HRM will deal with those changes in a strategic way to generate a competitive advantage out of big data. It is important to highlight that in this case, the competitive advantage is generated by combining people with big data. Big data aligned with the current digitization resemble a paper by Lepak and Snell talking about the virtual HR department. Big data enable the HR department to have access to all the relevant information as well as communicate with every employee everywhere.

2.2 Main difficulties and potential

2.2.1 Skepticism around HRM

In the digital era, as opportunities rise according to the technological advance, a blur of what is possible and ethical rises as well. In any revolution, as history suggests, in order for a change to be implemented, doubt and skepticism accompany the process. There is a need for trial and error, in order to accept something as a change in our society. Furthermore, this trials need to be successful and despite the doubts, benefits need to be clear. The reason, for the difficulty of HRM to integrate in a more extensive manner the big data, is that academic research is relative low and there is not sufficient literature to back up the theory with empirical data [9].

As Marler and Boudreu, suggest in an evidence based review on HR analytics, by examining three large databases to seek empirical research on the field (Academic Search Complete, Business Source Complete, Scopus) found that from the 60 articles that were available, only 14 made it through their evaluation process. Most of the studies suggested were non-empirical or stand alone case studies. In order for managers to trust the new way of things more than data-driven research needs to be made. In addition for the corporations to implement the new strategies according to big data, evidence of ROI (Return of Investment) is what makes a new implementation to work.

Let us take a step back here, and understand from where this skepticism originates. In his study Scholz, classifies the HRM in two categories, the *anti-guess workers* and *neo-luddites*. This classification derives from the main notion, that HR managers work with their “gut instict”, as opposed to data-driven analysts. It is clear, that the marriage of these to clans is in its foundation contradictory. People management is not something computers should have the power to touch, as the skepticists theorize. The unethical way to process people like numbers, hides many traps, as managers through their experience understand better of what an employee needs or if he/she is a good fit for the organization [10].

However, data-driven decisions are more solid in their perspective. Strategic decisions need facts, and facts in order to be thorough, processes on data must take place. The anti-guess workers, support the notion that data-driven decisions will take place of biased gut-feeling decisions. If by automating some processes gives more room to managers to center their attention in the implementation of strategies, then it is a measure to be followed. As Project Oxygen suggests, a people analytics experiment from Google, by analyzing performance reviews and team surveys, teams could make decisions for what makes a good leader and if he had a vision to employ with the team, as well as his strategic competence. This is one of the many isolated examples of how people analytics worked for the better.

On the other hand, the neo-luddites, support that HRM can work without taking such an intensive approach on the big data. The term luddite is derived from the anti-technological-progress movement in the beginnings of the industrial revolution [11]. Their main fear is that, as we go further in the usage of big data for people management, human resources will end up being just resources. In addition, as data give a holistic view of employees, fears of surveillance get a hold.

2.2.2 Somewhere in the middle

As the two sides have a strong core of beliefs and maybe exaggerate in their own way, a mutual cooperation would be of best fit. An augmentation of the data in the managerial gut feeling, could be the optimal case in HRM, as Scholz suggests.

The theory derives, from the word itself. Augmentation derives from the Latin word *augmentare* and means to gain, add, foster, or increase. Thus, by implementing the power of data in managerial decisions, we enhance the abilities of today's managers. This solution gives humanity to computer processing, and mistake avoidance, as well as solidity to the so called managerial gut feeling.

Augmentation also describes a certain direction of use. The human actor is augmented by technology to become better which does not exclude a data-driven approach rather than narrowing it down. A human is responsible for the decisions made and is only augmented by big data in order to make the best decision possible in a distinct case. There is room to be humane in certain cases, but also the potential of supporting decisions with big data. Big data and HRM can work together and their collaboration can be superior to either one working alone. The superiority of such collaboration has been proven in chess, in which the most successful combination is human and machine together [12] [13].

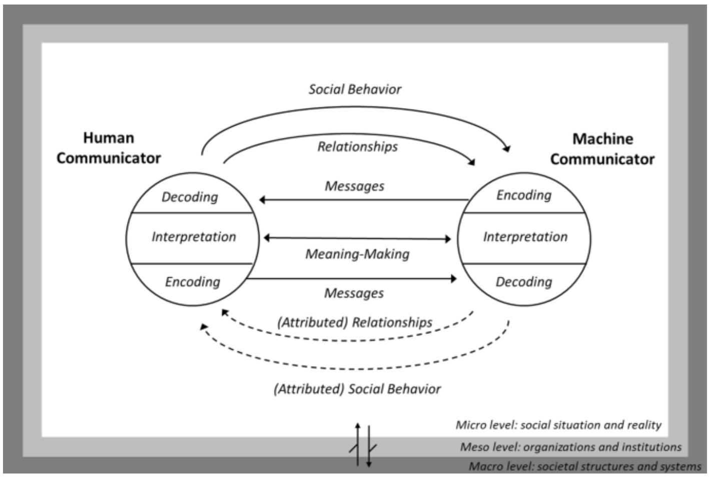

Social Behavior
Human Communicator
Relationships
Machine Communicator
Messages
Decoding
Interpretation
Encoding
Encoding
Interpretation
Decoding
Meaning-Making
Messages
(Attributed) Relationships
(Attributed) Social Behavior
Micro level: social situation and reality
Meso level: organizations and institutions
Macro level: societal structures and systems

Figure 2.4: Human-Machine Communication, source:Google

Another scope to this theory, is that data now is omnipresent in our world. By understanding and making good usage of them, we enhance our way of living as humans. The everyday tasks that mostly take time and do not require our critical thinking, can be automated and delegated. Thus, giving the area available for humans to invest their energy in critical thinking aspects and strategic decisions. If for example an algorithm in a given HRIS makes the compensation plan of an organization, the manager will then invest his/her energy in implementing the strategy and deciding who gets what, according to the insights of the data.

Many scholars have defined this new era of the human-machine. As Harari puts it, in his book Sapiens, a good scenario of the future is that of the *Homo Deus*. Where he defines humans and machines as a new hybrid. There people will incorporate the full power of data and computers, to become something God-like [14].

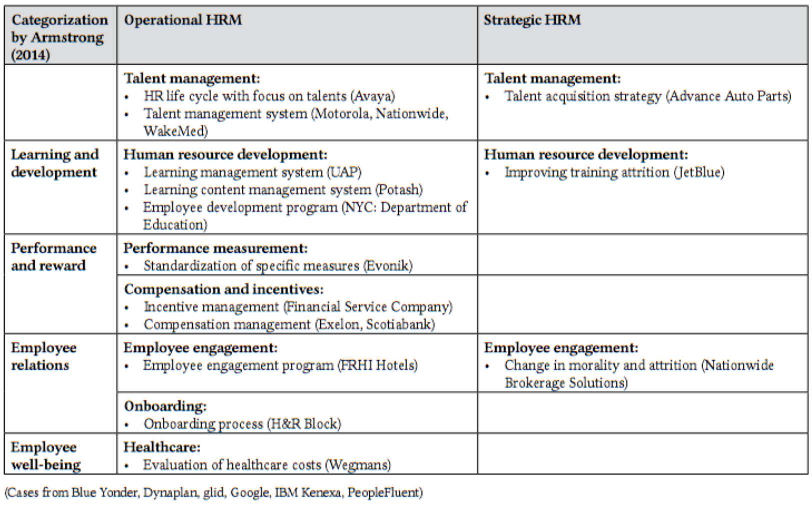
Another example of this augmentation, takes place in Wiener's book, The Human usage of Human Beings, where he explores a machine that can communicate with people. In this example communication takes another approach, as the means needed for two entities to establish rapport. In this approach senses become more abstract, as there is an effort to build a hearing device for a deaf person. Thus communication, does not need to be direct, but rather a mutual translation of the language or senses of the one part to other. By this derivation, we can stimulate the thought of data communicating with managers in means that optimize their cooperation [15].

Big data is here, and is here to stay as the digital era imposes. It is our challenge to find ways in order to make this marriage work. The potential of human processes enhanced with computer dynamic is huge. As there are drawbacks to be considered, data-driven decision can optimize many things in society and organizations [16].

2.2.3 Potential of data-driven HRM

For this thesis to dig deeper in the connection of data mining and human resources management, the dynamic of the potential that this combination has, should be established. Various projects by organizations show the huge benefits of the data-driven HRM has to offer. Though digitization changes our perspective, and thus HRM will change, if it has not done it [17].

In the below image Scholz is categorizing main HR processes, as how they could be transformed in the future. As we see in the table below, actors of the organizations change responsibilities and usage. From an operational standard, HR transforms its position, and acquires a more dynamic and strategic role within it.



Categorization by Armstrong (2014)	Operational HRM	Strategic HRM
	Talent management: • HR life cycle with focus on talents (Avaya) • Talent management system (Motorola, Nationwide, WakeMed)	**Talent management:** • Talent acquisition strategy (Advance Auto Parts)
Learning and development	**Human resource development:** • Learning management system (UAP) • Learning content management system (Potash) • Employee development program (NYC: Department of Education)	**Human resource development:** • Improving training attrition (JetBlue)
Performance and reward	**Performance measurement:** • Standardization of specific measures (Evonik)	
	Compensation and incentives: • Incentive management (Financial Service Company) • Compensation management (Exelon, Scotiabank)	
Employee relations	**Employee engagement:** • Employee engagement program (FRHI Hotels)	**Employee engagement:** • Change in morality and attrition (Nationwide Brokerage Solutions)
	Onboarding: • Onboarding process (H&R Block)	
Employee well-being	**Healthcare:** • Evaluation of healthcare costs (Wegmans)	

(Cases from Blue Yonder, Dynaplan, glid, Google, IBM Kenexa, PeopleFluent)

Figure 2.5: Categorization of new HR [10]

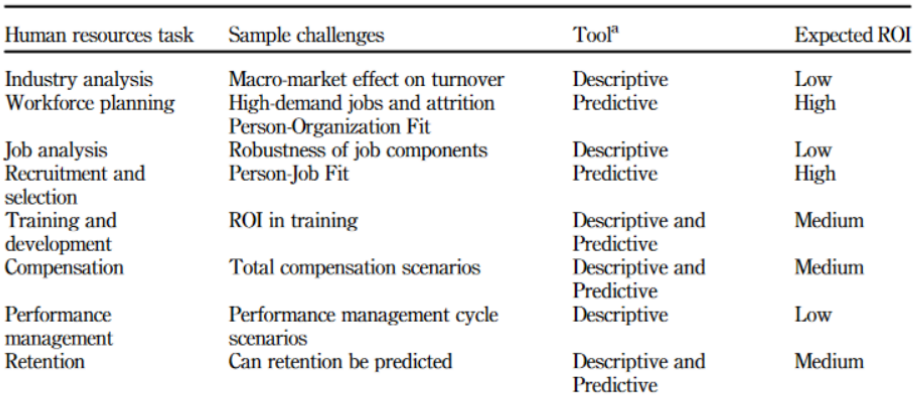
Scholz is taking under consideration Armstrong's categorization and gives a new texture in the HR management. The connection that is not clearly shown in this table, is analytics. Via these tools humans will accomplish to demonstrate the impact of the knowledge that is gained from the resources (employees), to further optimize their role in the organization, as well as the environment of the organization in order for employees to thrive (Compensation, Training, Development).

For this purpose various studies have shown the ROI of HR analytics, in various sectors of HR, thus mapping the practical opportunities that lie in the field. Furthermore, Chalutz, has estimated the ROI in research level for HR analytics. As we saw, previously the need of research in the field of HR analytics, a ROI approach articulates the importance of HRA as a tool for managerial decision making [18].

From their research in ESBCO database, they concluded that 86% of the research was made in the time frame 2011-2016, articulating it as a growth period. Which generally makes sense, because in the same period the computer power, as well as Artificial Intelligence saw a huge growth. This further enhances the argument that human capital is of huge importance for organizations. As technological means expand and availability is increased, the key factor that separates organizations is the human capital. So, those who invest in the optimization of human resources, will have an advantage in the business world.

2.2.4 ROI of HR analytics in key challenges

By theorizing over the importance of the HR analytics, Chalutz, combines his extensive research and outline some key challenges with proposed solutions and the ROI that accompnies them.



Human resources task	Sample challenges	Tool[a]	Expected ROI
Industry analysis	Macro-market effect on turnover	Descriptive	Low
Workforce planning	High-demand jobs and attrition Person-Organization Fit	Predictive	High
Job analysis	Robustness of job components	Descriptive	Low
Recruitment and selection	Person-Job Fit	Predictive	High
Training and development	ROI in training	Descriptive and Predictive	Medium
Compensation	Total compensation scenarios	Descriptive and Predictive	Medium
Performance management	Performance management cycle scenarios	Descriptive	Low
Retention	Can retention be predicted	Descriptive and Predictive	Medium

HR analytics as a potential solution

In the picture are shown, some of the features ot HR analytics, for further understanding, Chalutz, defines the available tools:

Descripitve Analytics: descriptive statistics, graphs and plots, benchmarking tools, KPIs-based methods (scorecards), business intelligence (BI) dashboards and advanced survey analytics.

Predictive Analytics: regression and parametric modeling (including logistic regression), time series analysis, classification methods (e.g. decision trees, SVM, discriminant analysis, neural networks,deep learning), clustering (K-nearest neighbors, K-means) anomaly detection, profiling, association rules, link-analysis, causality modeling (Bayesian networks), text analysis and NLP and Attrition Modeling.

In conclusion, HR analytics articulate a huge potential for organizations and employees as well. It is a bridge that gaps the need for the vast amount of data to be classified in a progressive and beneficial way, for both parties [19].

Even though a lot skepticism has been built around this topic, it essential to understand the reason behind it. Of course in any change, there will be criticism. But the dangers that are hiding in this topic are many. So by taking under consideration these dangers, data integration in the people management needs to be smooth and viable for both sides.

We will further explore, the dangers and the challenges that are hidden. Now that we have mapped the general idea of HR Analytics, we will explore the newly made term of People Analytics, and in which circumstances is useful and in which managers can trust their gut more.

Chapter 3

Human Capital as Data

3.1 The emergence of People Analytics

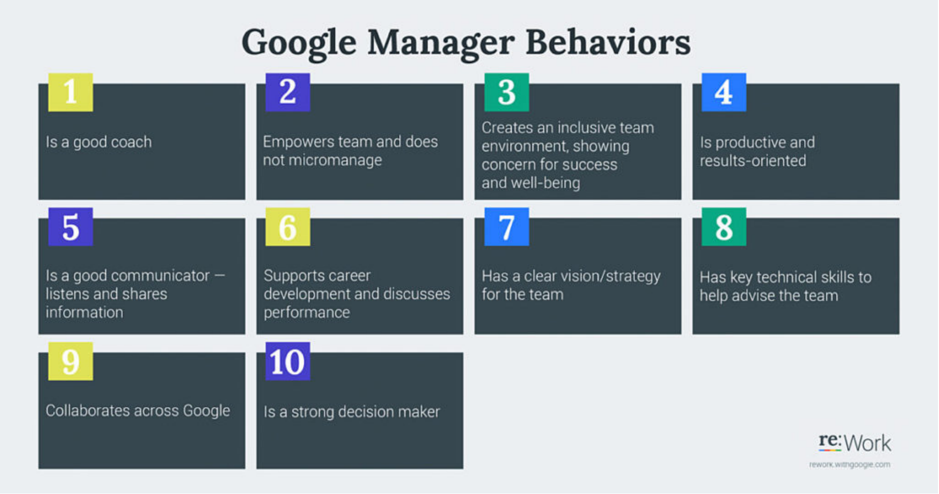
People are the driving force of an organization, be it a small enterprise or Google and Amazon. And as globalization takes place, people move around the globe for work, bringing their cultures along the way [20]. This huge movement is what brings different kind of cultures and people to work together, cooperate and communicate in a daily basis. As technology advances as well, the practice of remote work is a status quo in today's society. What is more interesting though, is the amount of data that is being produced from this never ending migration. From travel distances, to demographics and personal identifications to everyday browsing digital footprints. And in an organization data is money. Those who can make the most out of it, have a huge advantage over their competitors. Most importantly, it is not the capitalization of any data what makes an organization successful, but especially its members data [5].

3.1.1 Project Oxygen

We will begin this section with the example of Google, and its people analytics Project Oxygen. Google is widely known for its innovative culture, and when it comes to HR, has led the foundation for something unique as well. With the establishment of People & Innovation Lab (Pi-Lab), gathered a team of psychologists, decision scientists and researchers to approve policies based on data and academic research regarding people management. The result is Project Oxygen [21].

In this multi-year research project, leadership traits of managers were identified and ranked. It involved analyzing qualitative comments from employee surveys, employee expectations from managers, complaints and praises mentioned in performance reviews, and phrases in top manager awards. This research provided an in-depth look at the qualities of successful managers, offering valuable insights into the traits and characteristics of exemplary leaders [21].

This analysis provided managers with an understanding of how to manage their employees and provided guidance on their roles and responsibilities. In addition to professional guidance, employees were found to respect their superiors for coaching and mentoring.



Google Manager Behaviors

1 Is a good coach

2 Empowers team and does not micromanage

3 Creates an inclusive team environment, showing concern for success and well-being

4 Is productive and results-oriented

5 Is a good communicator — listens and shares information

6 Supports career development and discusses performance

7 Has a clear vision/strategy for the team

8 Has key technical skills to help advise the team

9 Collaborates across Google

10 Is a strong decision maker

re:Work
rework.withgoogle.com

Figure 3.1: Behaviors of a good manager, source: Google

Furthermore the algorithm was shown to improve retention along the ranks of Google. By examining various aspects of what made employees leave, Google was ready to tackle this problem, by planning compensation programs and further analyze the needs of its employees [21].

Hiring process was under optimization as well. With further analyzing applications of what made a good fit with the company, Google was in place to examine applications that otherwise would decline.

What we can get as an observation, is that the right usage of analytics to optimize day-to-day tasks can give an organization an environment of trust and cooperation for employees and executives. Thus giving space for managers to make more qualitative work and give their attention in strategic decision making [22].

3.1.2 HR takes a new approach

Until this point, the importance of analytics in human resources has opened a new perspective, of how organizations manage employees [23]. People analytics, the definition that derives from the fusion of HRM and analytics, defines its purpose to put HR in the scope of evolution and further extend its power as a department in the organization [24].

Despite the various recommendations, companies have made many steps forward in various sectors, but not sufficient when it comes to human resources. And that is one reason why people analytics can boost this leap of faith. From Google's success, there is a lesson to be learned for the other organizations, that analytics can work for us, and make a huge impact in productivity and overall work environment [25].

Until 2010s, HR mainly performed with a more operational and descriptive way. Software like HRIS, that was produced to extend the capabilities of HR, maintained a descriptive role in the toolbox of each HR manager, achieving only to manage what already was known in a more cohesive manner. No matter the evolution, that is not enough, for a contemporary organization which has to manage a tremendous amount of data, along with rapid technological evolution, a need for further implementation of analytics and sophisticated algorithmic automation is crucial in order to keep up [26].

To connect the dots, there is a need for further explanation in how people analytics works. The main notion, is that people analytics is a tool and not something to disengage and replace human decision. Although, the old ways of human management that are based on judgement and gut-feeling, sometimes are right, they hide the danger of bias. Managers are prone to mistaking situations, with their past experience and rush into conclusions that are ruled from their biases. Data driven decision making, solves this implication by giving insights that are not visible in the first glance. Furthermore, by analyzing data, employee's voices can be heard, because the immediate effect of change can be traced and put under consideration.

In a more practical manner, Petersen, articulates some major use cases for people analytics [27]:

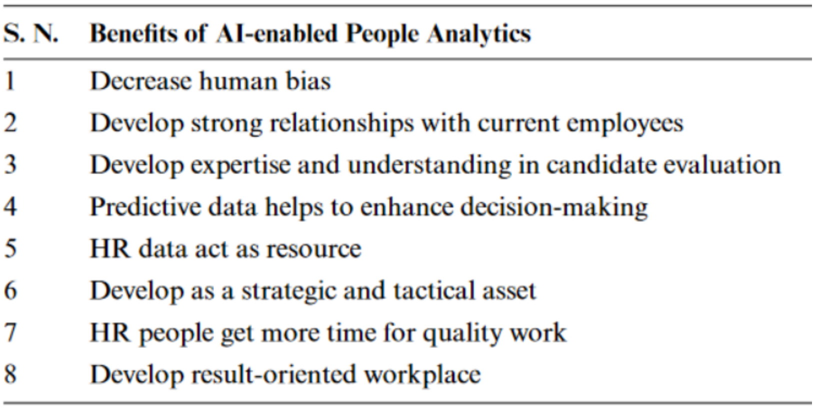
- The early detection of quitting intentions of employees
- The projection of future leadership ranks
- Stress level analysis for workplace health management
- Detailed KPI metrics of individual employees

- Network analyses of internal communication
- The creation of personas for modeling successful employees based on actual employee data

An economy driven by data allows for better placement of employees and fair opportunities according to their abilities [28]. The neutrality of data and algorithms, compared to humans, is another argument in favor of data-driven HR over traditional HR. Despite their experience, even the most experienced HR managers cannot completely avoid human biases. In contrast to human recruiters, selection and hiring algorithms that focus solely on key requirements lead to more diverse and functional teams [29] [28]. Analyzing the people in an organization is another valuable tool for supporting change management [30]. In addition, predictive analytics is also useful in the early detection and prevention of work-related conditions, including burnout [30].

3.1.3 AI to support the change

In the previous chapter, the importance of technology in HR was underlined, with some practical observations. In addition, the work of people analytics can be further enhanced with the contribution of Artificial Intelligence. Rai and Singh, made a thorough classification of the existing literature over the usage of AI in HR processes. In the table below we see the benefits of AI powered functions [31].



S. N.	Benefits of AI-enabled People Analytics
1	Decrease human bias
2	Develop strong relationships with current employees
3	Develop expertise and understanding in candidate evaluation
4	Predictive data helps to enhance decision-making
5	HR data act as resource
6	Develop as a strategic and tactical asset
7	HR people get more time for quality work
8	Develop result-oriented workplace

Figure 3.2: Benefits of AI [31]

By facilitating team effectiveness and real-time communication, AI is assisting in the transformation of the HR department. This procedure has several limitations because the organization is still fairly new to these technologies. In order to improve the decision-making process, individuals are teaching the machine how to extract the data quickly and accurately. According to Escolar-Jimenez et al., a smart system will evaluate different parameters to determine a candidate's suitability for a particular position [32].

The success of an organization depends on its people, and in the digital age, data empowers individuals. According to Hemalatha et al., AI significantly affects HR operations, particularly recruiting and selection. They discussed four AI capabilities—natural language processing, automation of machine vision, augmentation, and their impact on hiring and selection practices in a business [33]. According to the research, AI competencies have a major impact on the recruitment and selection process, leading to benefits like bias-free selection, decreased workload, time savings, cost effectiveness, accurate results, satisfied candidates, and increased efficiency.

The use of AI in hiring, compliance, employee coaching, mentoring, and performance evaluation is essential. AI-enabled systems manage a number of HR tasks, including interviewing, selecting candidates, communicating, learning, and developing employees (robots). They collaborate with the organization's HR as cobots (collaborative robots). Networks carry out a variety of HR tasks like hiring for career routes and learning and development. It is well acknowledged that HR plays a significant part in the long-term growth of an organization. According to Kim et al., technology has changed how the organization manages its employees and day-to-day operations [34].

People analytics powered by AI entails numerous risks in addition to its many benefits. Since technology is developing so quickly, employees are expected to complete their work quickly. The risks resulting from people analytics are presented by Giermindl et al. even if it is thought to be the answer for the contemporary business in the digital age. There are various advantages and hazards associated with AI-enabled people analytics, including privacy concerns, ethical concerns, and risks associated with integrating all the processes [35].

3.1.4 Drawbacks: Are People equal to data?

Many HR professionals are skeptical, conflicted, or uncomfortable because they don't believe that people can be reduced to statistics and are unsure of how to handle employees' data in a way that is morally right [9]. Additionally, it is challenging for HR managers to process (non-anonymized) personnel data while adhering to laws and data protection requirements and maintaining the public's trust [36].

In recent years, there is a slight emergence of studies that try to formulate the dangers and pitfalls of AI in HR management. As presented by Erb , a huge danger is the fear of algocrary. Where as he defines the role of AI in human resource processes, the danger of the machine bias is present due to the bias that preexists in the data, the algorithm is fed. And because of that reason, a need for caution is of huge importance [37]. As Morozov presented, that due to personal data being disclosed to organizations, a threat of surveillance over employees is a serious danger [38].

Analyzing extensively the existing literature over people analytics, Giermindl et al. classified the dangers that empowering AI in employee data accompanies. In a thorough review, they classified risks in the below categories.

People analytics can bring about an illusion of control and reductionism. Algorithms tend to give the false illusion of control to managers. Because of their objective nature, the results are misinterpreted and users believe that have complete control over them, something that is not true [35]. Also human data, do not represent wholly the person that derive from, thus making the process of decision making a false and underrated outcome [18].

People analytics can lead to estimated predictions and self-fulfiling prophecies. Because of their role as a predictive tool, managers can confuse the outcomes not as a possible event, but a final event. As a consequence, decisions based on an explicit outcome, discriminate over the reality and possibilities of an event occurring. Furthermore, when prediction of say, retention gives a certain outcome, and managers build strategies for further training on those who stay based on that prediction, as opposed to those who are meant to turn over, inevitably they manipulate the prediction to happen [39] .

People analytics can foster path dependencies. While taking into account past events to predict the future, there is a pitfall of continuing a mistake to future as well. Based on the profiles of successful recruits over the previous ten years, Amazon used a prescriptive analytics tool to identify the top candidates in order to automate its talent search. The program suggested that male candidates were more likely to be a good fit because the data revealed that the vast majority of successful hires in the past had been (white) men; hence, the algorithm improperly excluded women [40].

People analytics can impair transparency and accountability. After the implementation of sophisticated machine learning algorithms, only a small number of knowledge workers with extremely specialized expertise and technical training can reconstruct choices [41]. Overall, it appears challenging to determine who can be held responsible for substantial errors, significant failures, and misconduct of a system if the deployed learning algorithms and AI operate as a black box [42].

People analytics can reduce employees' autonomy. By substituting teams' interactive processes, collaboration, and cooperation practices with predefined goals and key performance indicators (KPIs) condensed in algorithms, people analytics might hinder employees' true discretion in decision-making and work habits [43]. Previous studies have dubbed this revival of the Taylorist paradigm as digital Taylorism and described how it further controls the employees in worrying ways while "involves creative and intellectual tasks being subject to the same process as chain labour" [44].

People analytics can marginalise human reasoning and erode managerial competence. Human capabilities and managerial skill will inevitably be devalued if organizations believe algorithms to be superior and more objective than purportedly emotional, subjective, and flawed human processes and they believe managers' only reliable source of knowledge should be measurable, and ideally quantitative facts [18]. In the end, using and relying on learning algorithms and AI as utilized in autonomous analytics is likely to exacerbate the erosion of managerial skills and may even reduce their job to that of the machine's stooge [45].

People analytics is a very potent tool that has become essential in contemporary HR, especially when it is based on AI. But the purpose of quantitative models is to supplement human judgment, not to replace it. You must continuously monitor how the application is operating in real time, what explicit and implicit criteria are being used to make decisions and train the tool, and whether outcomes are affecting different groups differently in unintended ways in order to get the most out of AI and other people analytics tools [24].

3.1.5 Modeling People and their Networks

In this section we cite the modeling of humans as actors. In his book The Model Thinker, Page explores the modeling of people as *rational* or *rule-based* actors. The purpose of this section is to give an intuition of how difficult is to model a human behaviour, in the sense of analytics that manipulates human data and making decisions upon those data.

As a first step a further classification of the set of rules is made, on how the actors behave

(fixed rules or adaptive). Then the notion of the modeling depends on what purpose does the model serve [46].

Page gives the following overview of various difficulties in modeling people: We are unique, socially influenced, prone to errors, purposeful, adaptable, and endowed with agency. Depending on the model's objective, the rational-actor model may serve as a gold standard, a straw man, or something in between. The rational actor will perform better as a tool for communication, action evaluation, and policy formulation than for predicting human behavior. People are difficult to model because they resist easy categorization, whereas models call for low-dimensional representations. People are diverse, socially influenced, prone to mistakes, purposeful, and able to learn new things. People also have agency—we have the power to take action.

Human models will inevitably be incorrect due to the difficulties of diversity, social impact, cognitive errors, purpose, and adaptation, which is why a many-model strategy is preferred. Austere behavioral models are appropriate in some contexts and enable attention to be directed toward other environmental factors. Better data will make richer behavioral models more suitable [46].

Another tool that Page uses in modeling of human behaviour is **Network Theory**. Although the purpose of this thesis does not align with the expanding of this theory, it is useful to mention the importance of this newly found science. By understanding network theory, human (nodes) find purpose in their rationality, as with the relationships(edges) that have build inside the community (cluster), they act and learn through interaction with the network. A lot of aspects of human behaviour can be traced in the comprehension of their respective networks and the interaction they build.

3.2 Digitization of People

Closing this chapter we are citing the methodology of how the digitization of people and their relations inside an organization happens. In order to achieve intelligence over employees, with the use of computer intellect, several steps must be followed. First, an organization must have in its disposal a clear database. Nowadays, every organization has a database of its employees, but in what condition is it. As business analytics evolves, and more companies acquire and put the knowledge into practice, we see more sophisticated approaches in business related issues with the use of analytics. That does not guarantee,it is happening simultaneously when it comes to people management. Either way, it is known at this phase, that one of the biggest problems HR has, is the condition of unstructured data. As it is stated by Duval, in one of the best platforms about Human Resources, to build a quality database for employees the following steps must be followed:

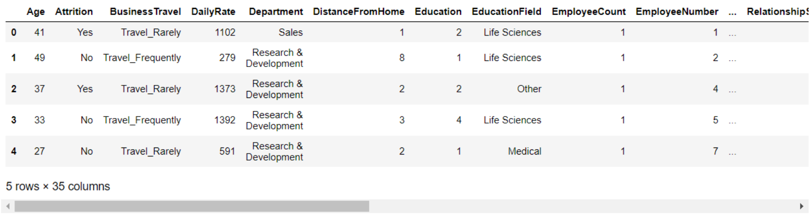
1. An employee database should be designed with a specific purpose in mind.
    - Improving reporting accuracy.
    - Automating manual HR processes.
    - Integrating payroll with processes for recruitment, benefits, and other areas.
    - Keeping track of HR metrics such as overtime expenses, time to hire, and workplace diversity.
    - Giving managers and employees easier access to data that supports their engagement and performance.
2. Determine your specific data requirements. (Employee Life Cycle)
    - Recruitment and onboarding
    - Performance management
    - Training and development
    - Retention and engagement
    - Off-boarding
3. Key stakeholders involvement
4. Possible outcome evaluation
    - Cloud-based, to offer your employees 24/7 access, wherever they are.
    - Dedicated support to get you up and running and provide day-to-day support after implementation.
    - Modules to support the entire employee lifecycle
    - Navigable reporting and dashboards to support better analysis and workforce decision-making.

5. Information and policies relevant to the policy should be gathered.

6. Provide employees with training and empowerment.

7. Utilize employee data in new and innovative ways.

Second, the data collection needs a thorough and extensive approach. For this to happen, we can rely either on cloud-based solutions, employee surveys or data scientists to collect and structure the data. For example a CSV or Excel file, that has as less as possible NA entries, and has an extensive collection of variables, that make the attributes of the employee. According to that collection greater insights can be given from the analysis of the data.



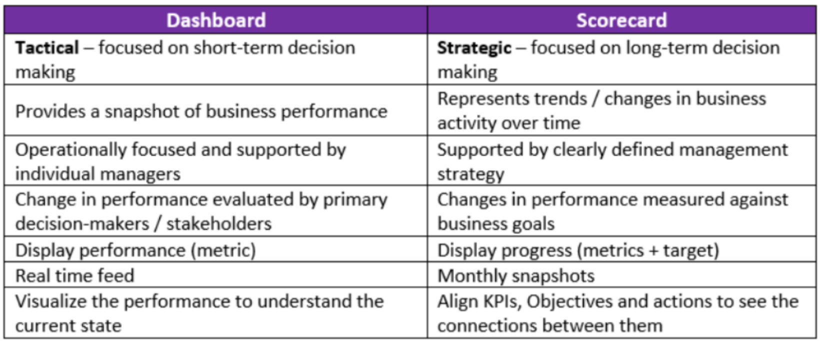
	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	...	Relationship
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	...	
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	...	
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	...	
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	...	
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	...	

5 rows × 35 columns

Figure 3.3: IBM's HR dataframe, source: Kaggle

As it is shown, on the image above, this dataframe contains 35 variables. This plethora, gives the opportunity to data analysts to have a quality over their analysis and extract information that are not clear to a manager that can only observe the data.

Third comes the interpretation over the data. It is of huge importance, for analysts to make a thorough categorization of the data, and the value they hold for the organization. Understanding the importance of these insights and acting accordingly is crucial since informed decisions are made based on data rather than preconceptions. The collected data could be used to build various forms of reports, graphs, and charts. Two of the most usual tools are the HR Dashboard and Scorecards.



Dashboard	Scorecard
Tactical – focused on short-term decision making	**Strategic** – focused on long-term decision making
Provides a snapshot of business performance	Represents trends / changes in business activity over time
Operationally focused and supported by individual managers	Supported by clearly defined management strategy
Change in performance evaluated by primary decision-makers / stakeholders	Changes in performance measured against business goals
Display performance (metric)	Display progress (metrics + target)
Real time feed	Monthly snapshots
Visualize the performance to understand the current state	Align KPIs, Objectives and actions to see the connections between them

Figure 3.4: Differences of Dashboards and Scorecards, source: LinkedIn

The final step is the application over our analysis. In this stage data scientists gather the insight and implement further processes. From predictive analytics and forecasting, to the implementation of machine learning algorithms and AI. In this point, according to the strategy and the goals of a given organization, models are created and thus along to the managerial decision-making process, strategies are implemented.

Chapter 4

Analytics vs HR

4.1 Key Fields of HR that Analytics help

In this section we are in a position to present the major aspects of HR, that analytics could best work, and alongside the flow of this thesis to begin and formulate the problem that we are proposing have a solution. So far, it is clear how HR management works and what are the dangers of implementing algorithms to process people data. Taking that under consideration, we know what to avoid and which solutions are capable of accepting an algorithm. In order to achieve a strategic advantage for the HR department in an organization, solutions that come out of the box, should be offered. And along with the capability of predictive analytics, this can be achieved in the long run. Below are examined some of the basic functions HR management has, and how can they be further optimized via analytics.

4.1.1 *Employee Recruitment*

The recruitment process is one of the most basic and valuable functions of the HR sector. An organization that can benefit from a well run recruitment process, can be found ahead in the business competition by securing valuable members that make a good fit. Some of the metrics that are used now are explained accordingly by Scott & Snell [1]:

- *Time to Fill*. The days between the moment a job opening is approved and the date an applicant accepts and/or starts work are referred to as the time-to-fill metrics. By enhancing this metric, with say dashboard and prescriptive functions, the process gets optimized.

- *Quality of Fill*. Companies have made an effort to create a quality-of-fill statistic that assesses how quickly new employees get "up to speed," how effectively they are functioning, and how long they stay with the company. Below is the formula that computes the quality-of-fill metric:

$$Quality - of - Fill = \frac{PR + HP + HR}{N}$$

  Where

  PR= Average job performance rating of new hires

  HP = Percentage of new hires reaching acceptable productivity within acceptable timeframe

  HR = Percentage of new hires retained after 1 year

  N= Number of performance indicators

- *Yield Ratio*. Yield ratios can be used to identify the recruitment channels that produce the most qualified job seekers. The proportion of applicants from a certain source who advance to the next round of the selection process is known as the yield ratio.

- *Cost of Recruitment*. It is relatively easy to calculate the average cost of hiring a new employee. The sum of the firm's overall recruitment expenses across all channels—including advertising, travel charges, referral bonuses, etc.—is divided by the quantity of hires. When accompanied with the Yield Ratio, managers have a thorough image of what directions can be used in the recruitment process. Thus it is relatively easy for an HR analytics tool, to calculate and automate these functions. An applicant tracking system (ATS) is another tool, that helps managers detect best candidates. In the example of Google for instance, the company uses algorithms and predictive analytics in order to make the process more diverse and suited for the needs of the organization.

4.1.2 *Employee Performance*

Creating a work environment where employees may provide their best effort in order to achieve a company's objectives is the process of performance management. An organization's goal informs the entire work system. Performance evaluations are the outcome of a process through which a management assesses an employee's performance in relation to the demands of his or her position and the goals established with that manager. The manager then uses the information to show the employee where and how to improve. Organizations can utilize the reviews as a tool to help employees grow [1].

A usual error in this process, that managers often do, is the distributional error. Because, some raters are either too good or to hard on their employees, there is a distribution error of what rating consists to be true. A usual solution for this problem, is averaging the distribution. With the help of analytics, companies like IBM, have integrated a 360 evaluation of its employees (From customer and peer reviews, to supervisor and manager reviews), tackling further the problem. Another useful analytical tool is the **Balanced Scorecard**, where the objectives of the company and the performance of a given employee are mapped, further resulting in a clear view of the rating [1].

4.1.3 *Workforce Planning*

It is highlighted that firms must adapt to global competitiveness, use technology effectively, increase efficiency, and keep costs under control, among other significant strategic and competitive problems. Managing a diverse workforce with a range of educational backgrounds, recognizing employee rights, and responding to employees' new work attitudes and desires for work-life balance are also some extremely significant employee problems that must be addressed. It is well known that the best organizations go beyond merely striking a balance between these occasionally conflicting demands; they develop work environments that integrate these issues to simultaneously get the most out of their employees and meet their needs while achieving the organization's short- and long-term goals [1].

Several occasions where firms have achieved better results in their workforce planning, an optimization in communication and inter-personal relations was observed. Every company that wants to achieve better employee conditions should give attention in making an egalitarian environment and encourage initiatives for every level of the corporate ladder. By implementing algorithms that utilize network theory and analytics for the employee satisfaction rates, managers can achieve better circumstances inside the organization, as well as encouraging climate. In the end, by these means the organization benefits, by having employees that accept the philosophy of the work environment and strive for further growing.

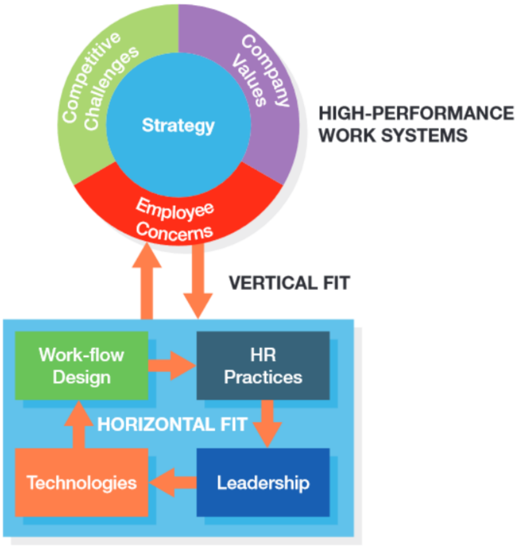


Figure 4.1: Strategic Alignment [1]

4.1.4 *Turnover*

One of the major challenges an organization has to tackle is the prediction of employee turnover. Simply put, turnover of employees refers to a person's departure from a company. It is frequently cited as one of the reasons why American worker productivity rates have not kept up with those of overseas rivals. Furthermore, it influences the labor supply in a significant way. A company's labor supply decreases as employees leave even if everything else remains the same. The organization will incur both direct and indirect expenditures as a result. The US Department of Labor has recommended the following formula:

$$Turnover = \frac{NumberOfSeparationsDuringPeriod}{TotalNumberOfEmployeesAtMidPeriod} * 100$$

The process of hiring a replacement employee takes time and money. Costs can typically be divided into three categories: costs associated with parting ways with an employee, costs associated with hiring a replacement, and costs associated with training a new employee. These expenses, which are conservatively estimated at two to three times the departing employee's monthly compensation, do not include indirect expenses like low productivity before leaving, decreased morale, and overtime for other employees as a result of the vacant position. As a result, a company may save a lot of money by lowering turnover [1].

By evaluating the importance of the turnover prediction, nowadays organizations are implementing sophisticated algorithms and machine learning for this purpose. It is the job of a skillful data scientist to find the best algorithm and evaluate the data, so as to make accurate predictions and identify trends in a sector where most organizations bleed. Below we see some of the best tools for turnover prediction:

Logistic Regression. The technique observed being utilized most frequently to forecast turnover is logistic regression. One method to begin gaining understanding into issues like “why do people leave,” “what can we do to affect turnover rates,” or “who is most likely to leave in the next year” is through the use of this technique. Any type of regression aims to forecast a result using one or more additional parameters. Using our workplace data, which are the independent factors, we wish to predict turnover, the dependent variable. Simply put, logistic regression is a type of regression that is utilized when trying to predict an outcome that will either be a 1 or a 0.

Survival Analysis.The sophisticated statistical method of survival analysis, sometimes referred to as event history analysis, is used to calculate the likelihood that an event will occur over time. The medical sciences have a history of using this method to forecast patient survival. Survival analysis can be used to determine the likelihood of attrition for a certain employee at a particular point in time, as opposed to the early results of logistic regression. This is a fantastic metric since it has the added capability of producing survival charts and attrition likelihood over time fast.

Tree Methods. When compared to logistic regression or survival analysis, which start at the individual level straight away, a decision tree model starts with all of the employees and divides them into smaller and smaller groups according to their propensity to attrition. As a result, each path develops a tree-like form with a central node and numerous leaves. The decision tree model is then expanded upon by the random forest method. At a high level, random forest creates its own decision trees out of random samples of data from your collection. To generate a prediction, it then takes the average of every tree it has created. The concept is that by combining a number of smaller forecasts, one larger prediction can be made.

4.2 Forecasting and Predictive Analytics

4.2.1 Forecasting: A leader’s skill

A key derivation that emerges from the portrayal of what consists a good manager is his competence in implementing the right tools for the right task. As we discuss the usage of analytics in HR management, forecasting is a unique attribute that enables managers to implement strategies preemptively for the needs of the organization. In this point an abstract point of view of what consists a good forecast, is imperative so as to get a better grasp of the concept.

In their book, Tetlock & Gardner, Superforecasting [47], solidify the notion of what consists a good forecaster, by implementing behavioral science and mathematics. A general theme is extracted by the book, that good forecasters, despite their knowledge in mathematics, the occasionally chose to make use of them. Thus, the insist on using their judgement and intuition instead. And that is a good argument, given that this thesis embraces the use of analytics in people management, however insists on the usage of them alongside human judgement. We will try to correspond the observations of this book, with our example of a manager that tries to predict the turnover of his organization. Below are listed the main aspects of a good forecaster, as Tetlock & Gardner map them:

1. *Triage* . A good forecaster, should ask the right questions that serve better the purpose that he tries to forecast. For example, a manager cannot be in a position to forecast, how many employees will turn over in the next 3 years. And in an essence it is unnecessary. Thus, it would suit better, to try and forecast for the next 6 months or a year, in order to be more effective.

2. *Break an initial problem to sub-problems*. In our case, where we want to predict the turnover, we would better break it to further sub-conditions. For instance, what makes an employee dissatisfied, or if the average time he spends on traffic exhausts him. (The Fermi Approach)

3. *Balance of inside and outside views*. It is about the uniqueness of every situation, or the absence of it. No matter the circumstances, a forecaster should take under consideration, of what happens generally, for example employees that do not get paid what their peers do eventually leave the organization, and of what happens especially, say the same employees of our example, balance the dissatisfaction of payment with less responsibilities.

4. *Balance of reaction over evidence*. Using predictive methods for managing requires a good intellect over the data and seer judgement of how to use them. If say, a manager who was informed of harder emigrant policies from the US government, hastens to

believe that employees under his disposal will leave, would most likely fall in the trap of over-reacting. It needs patience and further review of data and external factors, in order to best predict a given outcome.

5. *Each problem has clashing causal forces to look*. Taking under consideration the clashing arguments of every instance, is what can assure the situation to be tackled accordingly. An employee could churn because of the environment at work not being ideal for his standards, but a slight recession is going on simultaneously.

6. *The degrees of doubt should not surpass the dynamic of the problem*. Managerial decisions, as well as strategies cannot fall to impossible or certain circumstance. There are several nuances that make situations unique, but should not be too unique. Employees come and go in organization, but there are the little details that make them leave or not.

7. *Balance between under and over-confidence*. In statistics, confidence interval is used to measure if the estimates will fall again in the same range after a repeat of a test. Accordingly managers by examining previous predictions, can have the insights of previous mistakes or successes, making in addition novel strategies for future decisions.

8. *Revision of mistakes, avoidance of bias*. The danger of over information about a previous fallacy, hides the danger of scraping a strategy that was in general terms working.

9. *Embrace of cooperation*. A good manager is as good her subordinates are. By investing in team management, predictive decision making gains more value, and has a more solid nature.

10. *Balance of trial and error*. To gain expertise in something, deep and deliberative practice is needed. Along with continuous observations of the error. An algorithm or a system to be successful, needs constant revision and test. As well as a manager who practices upon the strategies of an organization.

11. *There is not a manual in forecasting*. As we train our algorithms, and nowadays predictions seem to be achieved with high level of certainty, no system is solid in an ever changing environment as work environments are.

Gaining an extensive view of what consists a good forecaster, we are ready to encompass the practical means in order to achieve a holistic view in this thesis. Predictability is a difficult task to be achieved, and a cooperation of the humans and machines is what will make this task to be successful. Without the knowledge background of humans, inconsiderate use is a prominent danger, as the power of computers in the digital era, sometimes excel our expectations.

4.2.2 Analytical Prediction

The goal of predictive analytics, which is based on the correlations between variables, is to provide answers to complex queries like: What will happen to Z if X and Y inputs change? The future can be forecast using the information already available, often with a high degree of certainty. Even though great insights can be gained from observing the graphs that are made through descriptive statistics there are several benefits from predictive analytics :

- Identifies the most reliable predictors and discards those that have no bearing
- Calculates how much a predictor affects an outcome measure's rise or decrease in order to quantify the influence of such predictors.
- A mathematical model to describe the current condition is provided
- Predicts future outcomes

A plan comes first when analyzing data. In this instance, the goal is to look at how efficiency, effectiveness, and outcome factors relate to one another. The strategy begins at the most fundamental level by incorporating all cases in the study. From there, it proceeds on to subgroup analyses, such as those of business units A, B, and C or campus hires (such as new grads) versus experienced employees (e.g., candidate with job experience). Three analytic techniques are more widely used:

1. **Correlation**. It investigates the connection between two variables. It responds to the question: What transpires to Y if the value of X rises? A complete positive association exists if both X and Y rise by 1. The statistic r, which goes from 1 to +1, describes a correlation. Zero means there is no relationship. A value of 1 denotes a proportional reduction in Y as X rises. A +1 denotes a proportional increase in Y as X increases.

2. **Multiple Linear Regression**. Although it is a slightly more sophisticated statistical method than correlation, its fundamental idea is the same: it looks at how variables

covariate. The employment of many simultaneous predictors is the main distinction. Similarly, a number of factors, such as time to hire, degree received, university attended, major, job experience, industry experience, tenure in role, and others, can be utilized to forecast new hire quality. Regression looks at all of the correlations between the variables and chooses the ones that are most strongly correlated with the output variable (e.g., productivity or profitability). Additionally, it eliminates predictor overlap, ensuring that each variable's predictive ability is distinct.

3. **Structural Equation Modeling**. It is a great approach to simultaneously evaluate several hypotheses and identify causal chains. Large data sets are needed for it, and it is based on confirmatory factor analysis. Regression is a much simpler analysis than this one, which calls for sophisticated tools like Lisrel or AMOS. SEM is a favored technique if the data set is amenable to analysis since it can produce the best-fitting model of the relationships between all the variables, offer trustworthy insights about the influence of numerous factors on one another, and produce an outcome measure.

Furthermore in their presentation of predictive analytics, Fitz-enz & Matox [48], they consider the impact in HR, that data analytics will make. In a future scenario that they present, their view of an organization that has eliminated the involuntary termination completely. Because of the power that predictive analytics will have, talent retention will be the major force of the HR department, thus picking and knowing in advance who to hire and for how long will the employee stay.

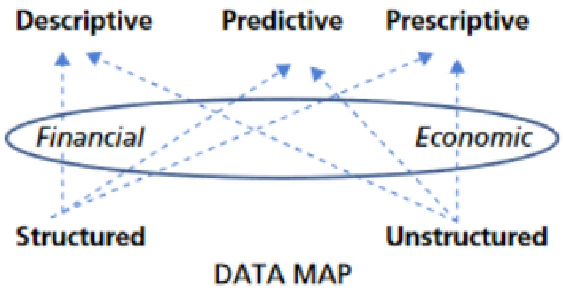


Figure 4.2: Paths of HR Analytics [48]

Chapter 5

Predictive Analytics Toolbox: R language

Until this point we have a established a view on HR management and how analystics can be implemented, in order for an organization alongside with computer intelligence, to make strategic decision upon the data that are under its disposal. Furthermore, we explained some key aspects of how a forecast can be structured and thus use and not be guided by the algorithms available.

One of the most powerful tools, a data analyst can have is the knowledge of R programming language. A programming language with a purpose in delivering statistical outcomes and manipulation of data. In this section, furthermore, we are presenting this tool, and what needs to be known when it comes to the usage of this language. We are exploring the packages that are available for use,and finally map the process of our implementation in a practical manner.

5.1 A quick introduction to R

R is a language and environment for visual design and statistical computing. It is a GNU project that is comparable to the S language and environment that John Chambers and colleagues created at Bell Laboratories (previously AT&T, now Lucent Technologies). R could be thought of as an alternative S implementation. Although there are some significant differences, much of the code created for S works flawlessly under R.

R offers a wide range of graphical and statistical tools, including time-series analysis, classification, clustering, and linear and nonlinear modeling. It is also very extendable. R offers an Open Source alternative for those interested in participating in statistical methods research, which frequently uses the S language as its preferred vehicle.

5.1.1 Packages

Extensions to the R statistical programming language are known as packages. Users of R can install R packages, which are standardized collections of code, data, and documentation, generally through a centralized software repository like CRAN (the Comprehensive R Archive Network). R has been widely used in data science due in large part to the abundance of packages that are available for it and how simple they are to install and use.

R packages are subject to a more stringent set of requirements than libraries in other programming languages. For R source code, data, documentation, and package metadata, the Writing R Extensions manual provides a standard directory structure that enables them to be installed and loaded using R's built-in package management tools. Additional requirements must be met before packages can be distributed on CRAN. While these demands ”impose enormous duties” on package developers, according to John Chambers, they enhance the usefulness and long-term stability of packages for end users.

The R Foundation supports the Comprehensive R Archive Network (CRAN), which serves as the main software repository for R. It includes documentation, donated R packages, and archives of both the most recent and earlier iterations of the R distribution. For Windows and macOS, it offers both source packages and pre-compiled binaries. There are more than 16,000 packages available as of November 2020. Kurt Hornik and Friedrich Leisch developed CRAN in 1997. The name was inspired by previous early packaging systems like TeX's CTAN (published in 1992) and Perl's CPAN (released 1995). It is still kept up to date as of 2021 by Hornik and a group of volunteers. The master site, which is mirrored on servers all over the world, is housed in the Vienna University of Economics and Business.

5.2 Dialects of R

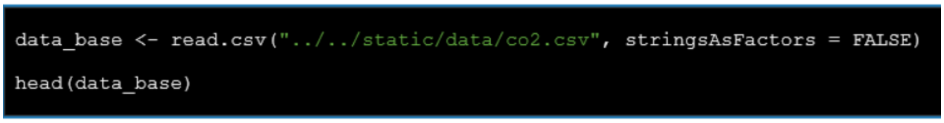
R gives the opportunity to programmers to use a variety of syntaxes. There is an unspoken rule, when it comes to the usage of each syntax. This rule goes according to the size of the data set, or the level of the programmer. We will see below the different dialects and we will choose for our implementation which one to use. Let us note, that we are going to show basic formulations of these dialects, without overextending. The notion here it to understand the different structures in basic operations.

5.2.1 Base R

Base R was the only way of writing R, until the emergence of the next two.Due of Base R's extreme stability, the code may be capable of enduring the test of time for those longer-term projects. Additionally, because Base R is more similar to a "pure" programming language, some of the fundamental abilities are more adaptable to other languages.

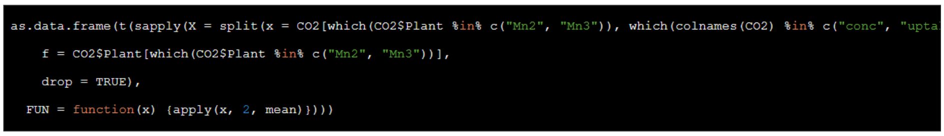
Below are shown some basic usages of base R:

The import of the data frame

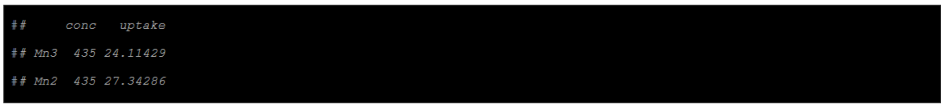


```
data_base <- read.csv("../../static/data/co2.csv", stringsAsFactors = FALSE)
head(data_base)
```

The piping method

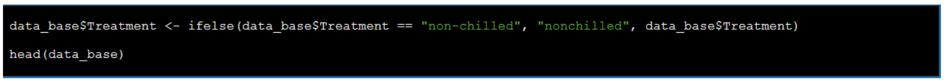


```
as.data.frame(t(sapply(X = split(x = CO2[which(CO2$Plant %in% c("Mn2", "Mn3")), which(colnames(CO2) %in% c("conc", "upta
    f = CO2$Plant[which(CO2$Plant %in% c("Mn2", "Mn3"))],
    drop = TRUE),
  FUN = function(x) {apply(x, 2, mean)})))
```



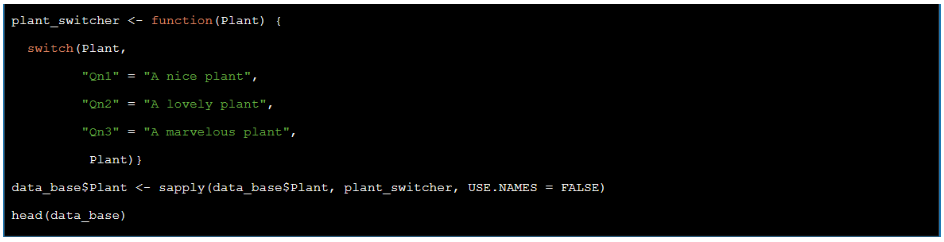
```
##      conc   uptake
## Mn3  435 24.11429
## Mn2  435 27.34286
```

If else conditional manipulation



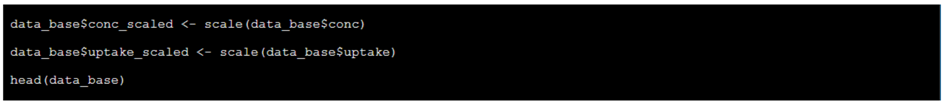
```
data_base$Treatment <- ifelse(data_base$Treatment == "non-chilled", "nonchilled", data_base$Treatment)
head(data_base)
```

Multiple conditional manipulation



```
plant_switcher <- function(Plant) {
  switch(Plant,
         "Qn1" = "A nice plant",
         "Qn2" = "A lovely plant",
         "Qn3" = "A marvelous plant",
          Plant)}
data_base$Plant <- sapply(data_base$Plant, plant_switcher, USE.NAMES = FALSE)
head(data_base)
```

Column add

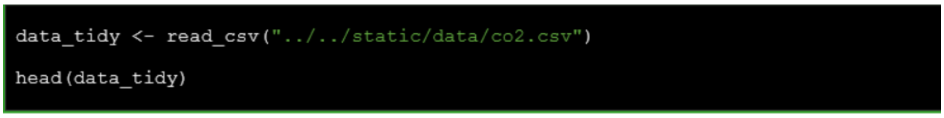


```
data_base$conc_scaled <- scale(data_base$conc)
data_base$uptake_scaled <- scale(data_base$uptake)
head(data_base)
```

5.2.2 Tidyverse

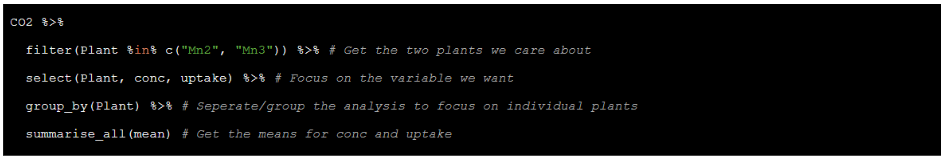
The tidyverse derives from the name of the tidyr package and the world universe. It is a collection of packages that define a dialect in the R language. The way it is structured, gives a more flexible and readable outcome for the programmers. Mainly the tidyverse dialect is used for smaller data sets. The easy interpretation of the code, helps in visualizations more than any other dialect. The special character (pipe) %>% is what differentiates this dialect, and what gives the readability advantage.

The import of the data frame

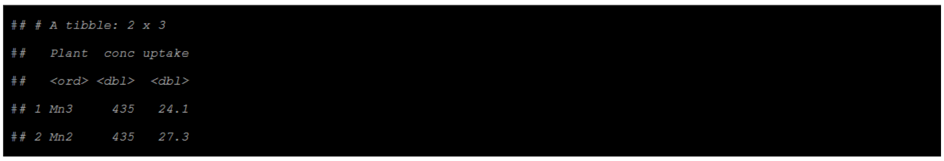


```
data_tidy <- read_csv("../../static/data/co2.csv")
head(data_tidy)
```

The piping method

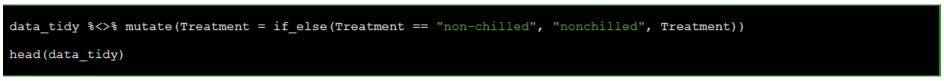


```
CO2 %>%
  filter(Plant %in% c("Mn2", "Mn3")) %>% # Get the two plants we care about
  select(Plant, conc, uptake) %>% # Focus on the variable we want
  group_by(Plant) %>% # Seperate/group the analysis to focus on individual plants
  summarise_all(mean) # Get the means for conc and uptake
```



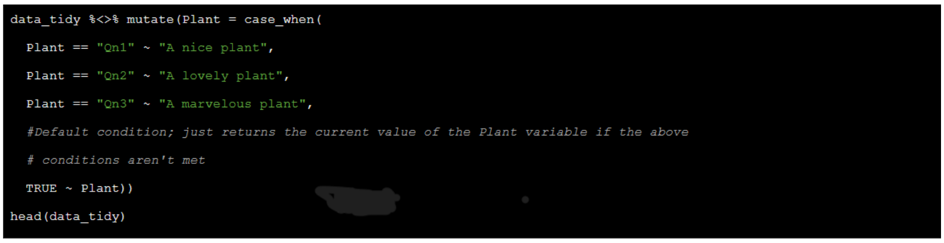
```
## # A tibble: 2 x 3
##   Plant  conc uptake
##   <ord> <dbl>  <dbl>
## 1 Mn3     435   24.1
## 2 Mn2     435   27.3
```

If else conditional manipulation



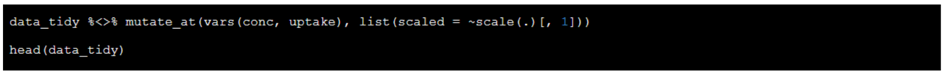
```
data_tidy %<>% mutate(Treatment = if_else(Treatment == "non-chilled", "nonchilled", Treatment))
head(data_tidy)
```

Multiple conditional manipulation



```
data_tidy %<>% mutate(Plant = case_when(
  Plant == "Qn1" ~ "A nice plant",
  Plant == "Qn2" ~ "A lovely plant",
  Plant == "Qn3" ~ "A marvelous plant",
  #Default condition; just returns the current value of the Plant variable if the above
  # conditions aren't met
  TRUE ~ Plant))
head(data_tidy)
```

Column add

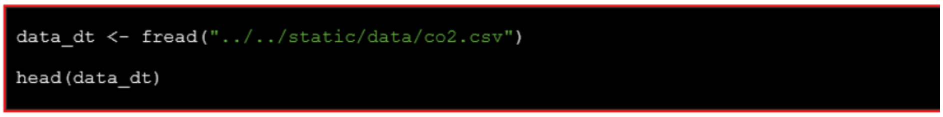


```
data_tidy %<>% mutate_at(vars(conc, uptake), list(scaled = ~scale(.)[, 1]))
head(data_tidy)
```

5.2.3 Data.table

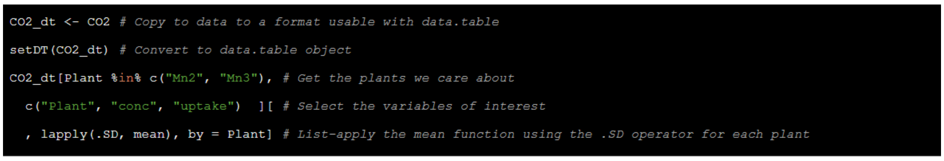
Data.table is the fastest way to write R. Besides, the way that the code is structured gives R room, to process faster the data and is capable of using large data sets (>100GB). The special ingredient of this dialect is bracket []. In addition, data.table uses tables as opposed with data frames of base or tibbles of tidyverse. In some machine learning algorithms, tables are the only that guarantee smooth operation.

The import of the data frame

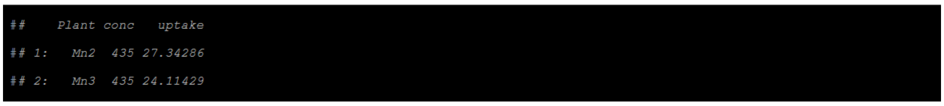


```
data_dt <- fread("../../static/data/co2.csv")
head(data_dt)
```

The piping method

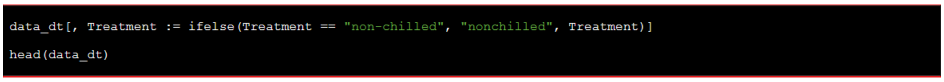


```
CO2_dt <- CO2 # Copy to data to a format usable with data.table
setDT(CO2_dt) # Convert to data.table object
CO2_dt[Plant %in% c("Mn2", "Mn3"), # Get the plants we care about
  c("Plant", "conc", "uptake")  ][ # Select the variables of interest
  , lapply(.SD, mean), by = Plant] # List-apply the mean function using the .SD operator for each plant
```



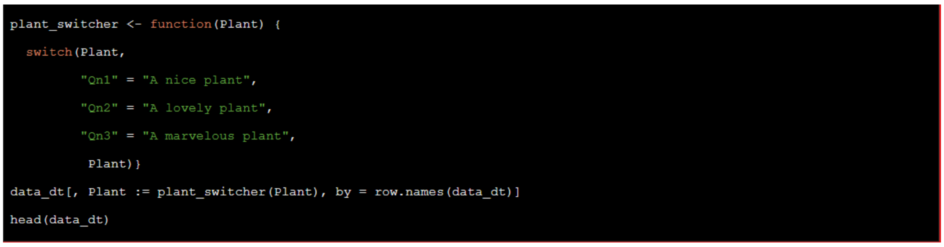
```
##    Plant conc   uptake
## 1:   Mn2  435 27.34286
## 2:   Mn3  435 24.11429
```

If else conditional manipulation



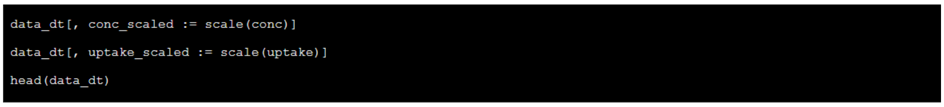
```
data_dt[, Treatment := ifelse(Treatment == "non-chilled", "nonchilled", Treatment)]
head(data_dt)
```

Multiple conditional manipulation



```
plant_switcher <- function(Plant) {
  switch(Plant,
         "Qn1" = "A nice plant",
         "Qn2" = "A lovely plant",
         "Qn3" = "A marvelous plant",
          Plant)}
data_dt[, Plant := plant_switcher(Plant), by = row.names(data_dt)]
head(data_dt)
```

Column add



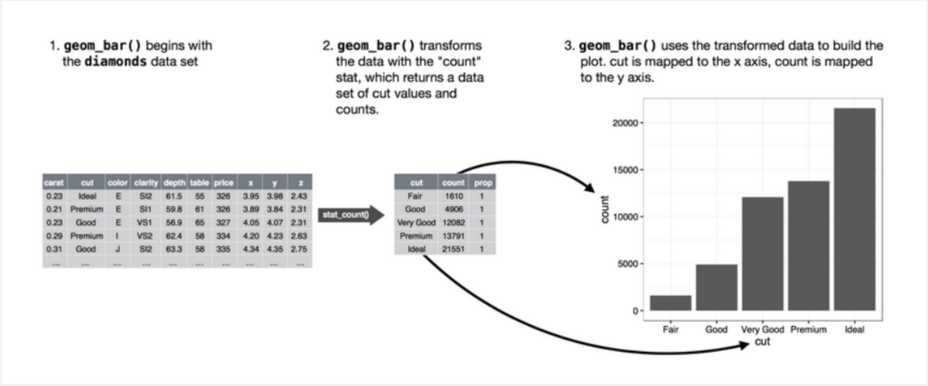
```
data_dt[, conc_scaled := scale(conc)]
data_dt[, uptake_scaled := scale(uptake)]
head(data_dt)
```

In this stage we need to clarify that each of these dialects has its pros and cons. Programmers tend to adjust the project and their goals accordingly. Either way, the flexibility of R gives the initiative to use multiple packages in consideration with the needs of the task, for example some programmers combine features of the tidyverse with the data.table like it is presented in this article (https://martinctc.github.io/blog/using-data.table-with-magrittr-pipes-best-of-both-worlds/). In this thesis, as we will see in the next chapters, data.table was used as a dialect for its fast processing and compact syntax, plus the compatibility issues that occur with tibbles (tidyverse data frames) and machine learning algorithms.

5.3 Data Visualization

To create a complete view over the data, an analyst utilizes various visualizations (charts, plots). R utilizes a package called **ggplot2**, which refers to grammar of graphics, a a logical framework for describing and creating graphs. With ggplot2, a plot is created with the function *ggplot()*. *ggplot()* creates a coordinate system that layers can be added to. The first argument of *ggplot()* is the dataset to use in the graph. So *ggplot(data = DATA)* creates an empty graph. The graph is completed by adding one or more layers to *ggplot()*. The function *geom_point()* adds a layer of points to the plot, which creates a scatterplot. **ggplot2** comes with many geom functions that each add a different type of layer to a plot. Below a template of code is shown for the *ggplot()*:

```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```



1. **geom_bar()** begins with the **diamonds** data set

carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...

2. **geom_bar()** transforms the data with the "count" stat, which returns a data set of cut values and counts.

cut	count	prop
Fair	1610	1
Good	4906	1
Very Good	12082	1
Premium	13791	1
Ideal	21551	1

3. **geom_bar()** uses the transformed data to build the plot. cut is mapped to the x axis, count is mapped to the y axis.

Figure 5.1: Statistical Transformation of the built-in diamond data set, source:https://r4ds.had.co.nz

Another functionality of the *ggplot()* is the statistical transformation. When adding layers with the *geom()* function, a hidden algorithm *stat()* is making the transformation, into a chart. The process is shown in the image.

Chapter 6

Theoretical framework for the model development

6.1 Statistical Background

The main goal for this thesis is to give a thorough view of the HR analytics. By expanding the available literature, as well as theorizing in several aspects, to give the prerequisites and tools available for anyone who is interested in this topic. Furthermore, this thesis offers the development of theory to praxis and gives the needed knowledge a basis for anyone who is interested in researching the field of HR analytics. Since the goal is clear, we will build our framework on fundamental concepts of statistics and machine learning, in order for the researcher to have a complete view on the topic.

Starting with statistics fundamentals, in this section the theoretical background will be presented, making the transition to the practical implementation comprehensible. It is noted, that we will not present the whole background of statistics, meaning the more basic concepts, thus given the thesis aims for researchers who have a strong mathematical background, while simultaneously giving the prerequisites of our work.

6.1.1 Scales of Measurement

Variables are categorized and/or quantified using measurement scales.Each scale of measurement satisfies one or more of the following properties of measurement:

- **Identity**.The measurement scale’s values each signify something distinct.

- **Magnitude**. The values on the measurement scale are related to one another in an organized manner. In other words, certain values are higher while others are lower.

- **Equal intervals**. Scale elements are equivalent to one another along the scale. As an illustration, this implies that the difference between 1 and 2 is equivalent to the difference between 19 and 20.

- **A minimum value of zero**. There is a true zero point on the scale, below which there are no values.

Besides the properties, its measurement falls in one of the following categories. The definition of scale, gives us a better categorization over our variables.

- **Nominal**. Only the identical property of measurement is satisfied by the nominal scale of measurement. Although values assigned to variables serve as descriptive categories, they lack any inherent magnitude-related numerical value. For example a variable that describes gender, is nominal.

- **Ordinal**. The identity and magnitude properties of the ordinal scale apply. Each value on the ordinal scale has a distinct meaning and is related to the other values in an ordered manner. For example job satisfaction (with values below, average or above), would fall in this category.

- **Interval**. Identity, size, and equal intervals are characteristics of the interval scale of measurement. An example variable of this category would be the job evolution of an employee. In regards with another employee, this variable can distinguish how much employee A has jumped in the corporate ladder against employee B.

- **Ratio**. All four of the qualities of measuring—identity, magnitude, equal intervals, and a minimum value of zero—are satisfied by the ratio scale of measurement. A simple example of this category, would be compensation.

6.1.2 Correlation Coefficient

The degree to which two variables are correlated is indicated by the correlation coefficient. The strength of the linear link between variables recorded on an interval or ratio scale is determined by the most often used correlation coefficient, known as the Pearson product-moment correlation coefficient.

The sign and the absolute value of a correlation coefficient describe the direction and the magnitude of the relationship between two variables.

- The value of a correlation coefficient ranges between -1 and 1.
- The greater the absolute value of the Pearson product-moment correlation coefficient, the stronger the linear relationship.
- The strongest linear relationship is indicated by a correlation coefficient of -1 or 1.
- The weakest linear relationship is indicated by a correlation coefficient equal to 0.
- A positive correlation means that if one variable gets bigger, the other variable tends to get bigger.
- A negative correlation means that if one variable gets bigger, the other variable tends to get smaller.

The formula to calculate the correlation coefficient is the below, which stands as the most common.

$$r = \sum(xy)/\sqrt{(\sum(x)^2) * (\sum(y)^2)}$$

where x,y are the values of the respective observations

6.1.3 Linear Regression

The independent variable in a cause-and-effect relationship is the cause, while the dependent variable is the effect. A technique for predicting the value of a dependent variable Y based on the value of an independent variable X is known as least squares linear regression.

Simple linear regression is appropriate when the following conditions are satisfied.

- The relationship between the dependent variable Y and the independent variable X is linear.It is needed that the residual plot displays a random pattern and that the XY scatterplot is linear to verify this.

- The probability distribution of Y has the same standard deviation for any value of X. When this criterion is met, a residual plot can be used to quickly verify that the residuals *(the difference between the observed value of the dependent variable (y) and the predicted value (ŷ) is called the residual e)* variability is essentially constant for all values of X.

- For any given value of X:

  - The residual plot displays a random pattern, which suggests that the Y values are independent.

  - The Y values are basically distributed regularly (i.e., bell-shaped). If the sample size is large, a little skewness is acceptable. The distribution's form will be displayed using a histogram or dot plot.

A bivariate data set's observations are best represented by a straight line, known as the least squares regression line (LSRL), according to linear regression. Let's assume that X is an independent variable and that Y is a dependent variable. Following is the population regression line:

$$Y = B_0 + B_1 X$$

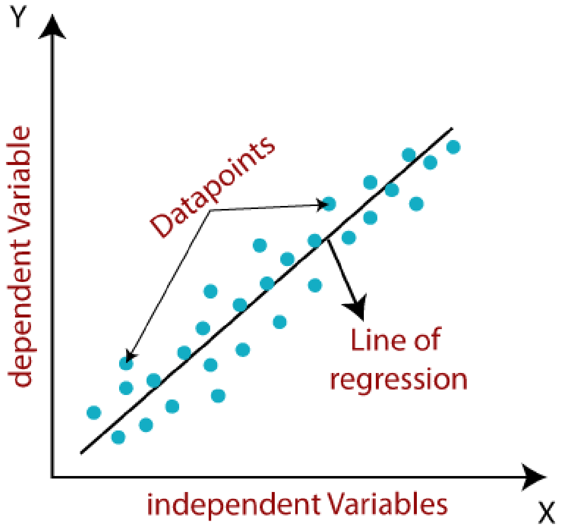


Figure 6.1: A graph of linear regression, source: Google

We will skip the further explanation on how a least squared regression line is produced, as computer algorithms come in handy in this situation, automating the process.

6.1.4 Residual Analysis

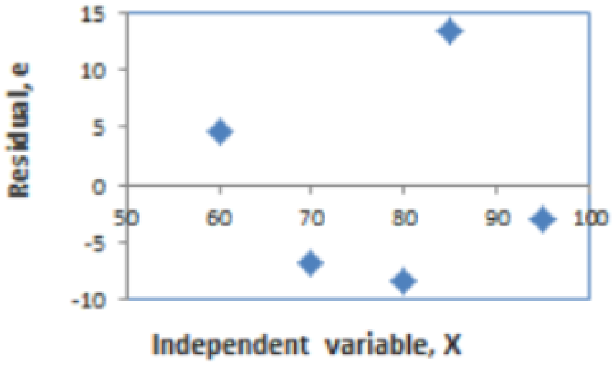


Figure 6.2: Residual plot, source:Google

Due to the fact that a linear regression model is not always suitable for the data, the model's suitability could be evaluated by defining residuals and analyzing residual plots. A residual plot is a graph where the independent variable is on the horizontal axis and the residuals are displayed on the vertical axis. A linear regression model is adequate for the data if the residual plot's points are randomly distributed across the horizontal axis; otherwise, a nonlinear model is preferable.

6.2 Machine Learning Algorithms

By presenting some key concepts of statistics, the view of this thesis is formed piece by piece. For this section, we encounter the concept of Machine Learning. As we discussed previously, in order to make optimizations in HR management, we implement concepts of statistics in algorithms and therefore the processes of HR have a more strategic role in an organization.

Utilizing sample data or prior knowledge, machine learning involves programming computers to optimize a performance criterion. A model has been defined up to a certain point, and learning is the process of running a computer program to optimize a model parameter using training data or prior knowledge. The model could be descriptive to learn from the data or predictive to make future predictions. Because drawing inferences from samples is its primary function, machine learning builds mathematical models using the science of statistics. First, we need effective algorithms during training to solve the problem that has been optimized as well as to store and handle the enormous amount of data that we typically have. Second, a model's representation and algorithmic solution for inference must be effective once it has been learnt. The effectiveness of the learning or inference algorithm, specifically its space and temporal complexity, may be just as crucial in some applications as the predictive accuracy [49].

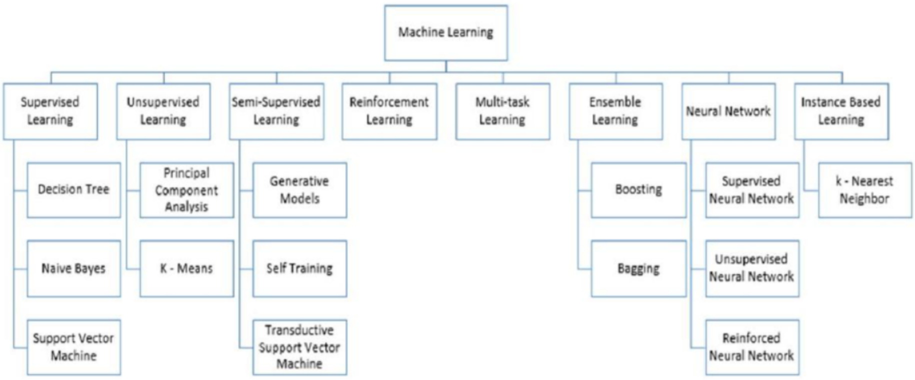


Figure 6.3: Categorization of ML [50])

According to Zhang, machine learning algorithms fall in the further classification according to their learning behaviour [49]:

- **Supervised Learning**. Where a function that links inputs to desired outputs is generated by the algorithm. An example of a typical supervised learning job is the classification problem, where the learner must learn (or approximation the behavior of) a function that divides a vector into a number of classes by examining many input-output samples of the function.

- **Unsuprevised Learning**. This simulates a group of inputs: there are no instances with labels.

- **Semi-supervised Learning**. It generates a suitable function or classifier by combining instances that are labeled and unlabeled.

- **Reinforced Learning**. Where an action strategy is learned by the algorithm based on observations of the outside world. Every action has an effect on the environment in some way, and the environment gives the learning algorithm feedback.

- **Transduction**. Similar to supervised learning, but does not explicitly build a function; rather, it attempts to predict new outputs based on training inputs, training outputs, and new inputs.

- **Learning to Learn**. Where, depending on prior knowledge, the algorithm develops its own inductive bias.

The work that was followed in this thesis, and the needs of the HR analytics, indicate the path of supervised learning as to the algorithms that should be used. Furthermore, supervised learning is best used for classification and regression, which are the main components of a predictive analysis. Below are presented a handful of supervised machine learning algorithms. Note, that because of the goal of this thesis is, by implementing predictive analytics as regards to predicting employee turnover(binomial outcome), we focus on the supervised machine learning algorithms.

6.2.1 Logistic Regression

To solve a classification issue, logistic regression is performed. Based on the values of the input variables, it provides the binomial result, which indicates the likelihood that an event will occur or not (in terms of 0 and 1) [51]. For example if an employee is going to leave or not the organization.

$$y = \frac{e^{b_0+b_1X}}{1+e^{b_0+b_1X}}$$

where x is the input value, y the predicted value, b0 the bias and b1 the coefficient of x

Key properties of logistic regression [52]:

- The dependant variable in a logistic regression follows the ”Bernoulli distribution.”

- Maximum Likelihood is the basis for estimation and prediction.

- The coefficient of determination (or R-squared), which is measured in linear regression, is not evaluated in logistic regression. Instead, a concordance is used to evaluate the model’s fitness.

6.2.2 Support Vector Machine

SVMs are capable of handling both classification and regression issues. The decision boundary for this method is the hyperplane, which must be determined. A decision plane is required to divide a collection of objects into their many classes. If the objects cannot be separated linearly, kernels—complex mathematical functions—must be used to separate the objects that belong to various classes. The goal of SVM is to correctly identify the objects using examples from the training data set [51].

SVMs can effectively do non-linear classification in addition to linear classification by implicitly mapping their inputs into high-dimensional feature spaces. This technique is known as the kernel trick. It essentially draws lines between the classes. The margins are created in such a way as to minimize classification error by increasing the distance between the margin and the classes [50].

The maximum margin problem, which is the distance between the decision border and the supporting hyperplane, must be solved in order to select a hyperplane. Somvashi et al., have made a formula for this purpose [53]:

$$margin = \arg\min x\epsilon D d(X) = \arg\min x\epsilon D \frac{|x - w + b|}{\sqrt{\sum_{i=1}^{d} w_t^2}}$$

6.2.3 Naive Bayes

This algorithm, which is based on conditional probability, is straightforward. This method uses a probability table as the model, which is updated using training data. When looking up the class probabilities to forecast a new observation, one must consult the "probability table," which is based on its feature values. The fundamental presumption is conditional independence, which is why it is referred to as "naive." The idea that all input features are unrelated to one another in a real-world setting can scarcely be true.

The following benefits of Naive Bayes (NB) include easy implementation, good performance, working with less training data, scaling linearly with the number of predictors and data points, handling continuous and discrete data, ability to handle binary and multi-class classification problems, and the ability to make probabilistic predictions. Both continuous and discrete data are handled. It is insensitive to unrelated factors [51].

In the event of having binomial outcomes, naive bayes has a strong relation to the logistic regression. Naive Bayes classifiers form a generative-discriminative pair with (multinomial) logistic regression classifiers.

Below is the formula of the Naive Bayes:

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)}$$

where k are possible outcomes of classes Ck

6.2.4 Decision Tree

By continually dividing data based on a certain parameter, Decision Tree is a Supervised Machine Learning approach to handle classification and regression issues. Decisions are made in the leaves, and the nodes divide the data. While the decision variable in a regression tree is continuous, the decision variable in a classification tree is categorical (the outcome takes the form of a Yes/No). The following benefits of the decision tree include: suitability for classification and regression problems, simplicity of interpretation, handling of quantitative and categorical values, ability to fill in missing values in attributes with the most likely value, and high performance due to the effectiveness of the tree traversal algorithm. Random Forest, which is based on an ensemble modeling approach, is the solution for Decision Tree's potential over-fitting issue [51].

Having a clear view of some of the basic algorithms in machine learning, we will approach the practical implementation of this thesis with the optimized version of the decision tree, the random forests. Due to its simple implementation, yet accurate results, random forests was chose. In the next section we gonna map, a more extensive view of this algorithm.

6.3 Why Random Forests

Creating prediction models can be done using the well-liked machine learning technique known as random forest. Random forests are a collection of classification and regression trees [54], which are straightforward models that use binary splits on predictor variables to derive outcome predictions. Breiman first developed random forests in 2001 [54]. In reality, decision trees are simple to use because they provide a straightforward way to predict outcomes by dividing "high" against "low" values of a predictor connected to the outcome. Though decision tree methodology has numerous advantages, it frequently produces subpar accuracy for complicated data sets (such as huge data sets and data sets with complex variable interactions).

Each tree's results are combined to provide an overall prediction for each observation. Because of this, random forest frequently offers greater accuracy than a single decision tree model while retaining some of the advantages of tree models (such as the capacity to understand correlations between variables and outcome) [55]. In the context of classification, random forests regularly provide among the highest prediction accuracy when compared to other models [25].

The ability to handle data sets with a large number of predictor variables is a key advantage of utilizing random forests for prediction modeling; nevertheless, in reality, it is frequently preferable to minimize the number of predictors needed to produce outcome predictions. For instance, when constructing a medical prediction model, one could prefer to use only a subset of the most crucial factors rather than all the variables present in the electronic medical record.Finding the most crucial predictors to include in a condensed, efficient model is frequently of interest in prediction modeling. This may be done by doing variable selection, which involves identifying optimal predictors based on statistical traits like relevance or accuracy. Using variable selection while creating prediction models may ease the difficulty of data collecting and increase the practicability of prediction. Given that many modern datas ets comprise hundreds or thousands of potential predictors, variable selection is frequently a crucial step in the construction of prediction models.

6.3.1 Mathematical Framework

To encapsulate the greater purpose of the random forests algorithm, we need to show the mathematical definition of it. In addition, we need to see the theoretical framework that was build for this algorithm, and understand the way it works. Breiman, who was the creator of the random forests algorithm, defines it as given:

A random forest is a classifier consisting of a collection of tree-structured classifiers {h(x, Θ_k), k = 1, . . .} where the {Θ_k } are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x [54].

We will focus the presentation of random forests in regression analysis. Even if it is a suitable algorithm for supervised classification as well [56].

Given this, it is assumed a training sample $D_n = ((X_1, Y_1), ..., (X_n, Y_n))$ of independent random variables distributed as the independent prototype (X,Y). So as to make the regression function $m(x)$= $\mathbb{E}[Y|X = x]$. To find the jth tree, of a randomized collection M of regression trees, given a $\Theta_1, .., \Theta_m$ the independent random variables. Where Θ is used to resample the training set prior to the growing of individual trees and to select the successive directions for splitting.

$$m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n) = \sum_{i \in \mathcal{D}_n^\star(\Theta_j)} \frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)} Y_i}{N_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)},$$

where, $D_n(\Theta_j)$ is the set of data points, that were selected before the construction of the tree, $A_n(x; \Theta_j, D_n)$ is the number of the points that fall to $A_n(x; \Theta_j, D_n)$. Combining all of this to the forest estimate :

$$m_{M,n}(\mathbf{x}; \Theta_1, \ldots, \Theta_M, \mathcal{D}_n) = \frac{1}{M} \sum_{j=1}^{M} m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n).$$

6.3.2 Algorithm

As was classified by Biau & Scornet, the algorithm has three important parameters[56]:

1. $a_n \in 1, ..., n$: the number sampled data points in each tree
2. $mtry \in 1, ..., n$: the number of possible directions for splitting at each node of each tree
3. $nodesize \in 1, ..., a_n$: the number of examples in each cell below which the cell is not split

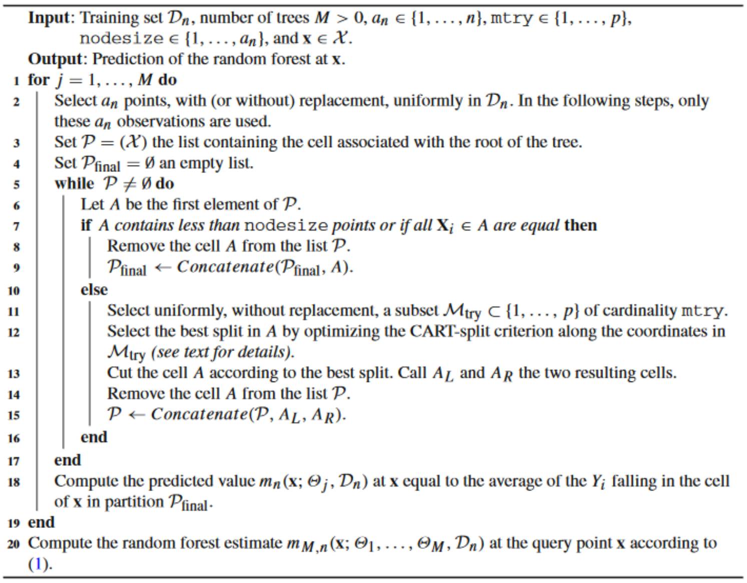
The algorithm grows M distinct (random) trees in the manner described below. a_n observations are chosen at random, with (or without), replacement, from the original data set prior to the building of each tree. The tree construction takes these—and only these—a_n observations (possibly repeated) into consideration. Then, a split is carried out at each cell of each tree by maximizing the CART-criterion over $mtry$ directions evenly selected at random from the p original ones. Lastly, construction of individual trees is stopped when each cell contains less than $nodesize$ points. . For any query point $x \in X$, each regression tree predicts the average of the Y_i (that were among the a_n points) for which the corresponding X_i falls into the cell of x [56].

Furthermore, the CART-criterion is defined as follows. Let A be a generic cell, and denote by $N_n(A)$ the number of the data points falling in A. A pair (j,z), which is a cat of A. The form of CART-criterion is : *where* $A_L = x \in A : x^{(j) < z}$, $A_R = x \in A : x^{(j) \geq z}$, and $\overline{Y_A}$ is the

$$L_{\text{reg},n}(j,z) = \frac{1}{N_n(A)} \sum_{i=1}^{n} (Y_i - \bar{Y}_A)^2 \mathbb{1}_{\mathbf{X}_i \in A} - \frac{1}{N_n(A)} \sum_{i=1}^{n} (Y_i - \bar{Y}_{A_L} \mathbb{1}_{\mathbf{X}_i^{(j)} < z} - \bar{Y}_{A_R} \mathbb{1}_{\mathbf{X}_i^{(j)} \geq z})^2 \mathbb{1}_{\mathbf{X}_i \in A},$$

average of the Y_i such that X_i belongs to A, with the convention that the average is equal to 0, when no point X_i belongs to A. For each cell A, the best cut (j_n^*, z_n^*) is selected by maximizing $L_reg, n(j,z)$ over M_try.

As shown below the complete algorithm of the Random Forests, first designed by Breiman [54] :



```
Input: Training set $\mathcal{D}_n$, number of trees $M > 0$, $a_n \in \{1, \dots, n\}$, mtry $\in \{1, \dots, p\}$,
       nodesize $\in \{1, \dots, a_n\}$, and $\mathbf{x} \in \mathcal{X}$.
Output: Prediction of the random forest at $\mathbf{x}$.
for $j = 1, \dots, M$ do
    Select $a_n$ points, with (or without) replacement, uniformly in $\mathcal{D}_n$. In the following steps, only
    these $a_n$ observations are used.
    Set $\mathcal{P} = (\mathcal{X})$ the list containing the cell associated with the root of the tree.
    Set $\mathcal{P}_{\text{final}} = \emptyset$ an empty list.
    while $\mathcal{P} \neq \emptyset$ do
        Let $A$ be the first element of $\mathcal{P}$.
        if $A$ contains less than nodesize points or if all $\mathbf{X}_i \in A$ are equal then
            Remove the cell $A$ from the list $\mathcal{P}$.
            $\mathcal{P}_{\text{final}} \leftarrow Concatenate(\mathcal{P}_{\text{final}}, A)$.
        else
            Select uniformly, without replacement, a subset $\mathcal{M}_{\text{try}} \subset \{1, \dots, p\}$ of cardinality mtry.
            Select the best split in $A$ by optimizing the CART-split criterion along the coordinates in
            $\mathcal{M}_{\text{try}}$ (see text for details).
            Cut the cell $A$ according to the best split. Call $A_L$ and $A_R$ the two resulting cells.
            Remove the cell $A$ from the list $\mathcal{P}$.
            $\mathcal{P} \leftarrow Concatenate(\mathcal{P}, A_L, A_R)$.
        end
    end
    Compute the predicted value $m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)$ at $\mathbf{x}$ equal to the average of the $Y_i$ falling in the cell
    of $\mathbf{x}$ in partition $\mathcal{P}_{\text{final}}$.
end
Compute the random forest estimate $m_{M,n}(\mathbf{x}; \Theta_1, \dots, \Theta_M, \mathcal{D}_n)$ at the query point $\mathbf{x}$ according to
(1).
```

Figure 6.4: Breiman’s random forests algorithm [54]

Concluding, this section we tried to give a view of the available algorithms that are used in supervised machine learning, with a more focused approach in regression. Random Forests was chosen, due to the straightforward implementation, as well as for the speed and effectiveness. As the sector of HR maintains massive and complex data sets, today necessitates the development of algorithms that ensure global competitiveness, achieving both computational efficiency and safety with high-dimension models and huge numbers of samples [57]. This is because statistical applications are becoming more and more sophisticated. Forests and its essential notions (such as “divide and conquer,” “resampling,” “aggregation,” and “random search of the feature space") are straightforward yet fundamental concepts that may benefit from cutting-edge algorithms.

In addition, we built up to this point a background both in how HR operates its major processes, as well as the mathematical background that is needed for the practical implementation. In order for a reader to understand the model and the implementation, all these steps we needed for the whole picture to be mapped.

Chapter 7

Predicting turnover with R

7.1 A general analysis over the data

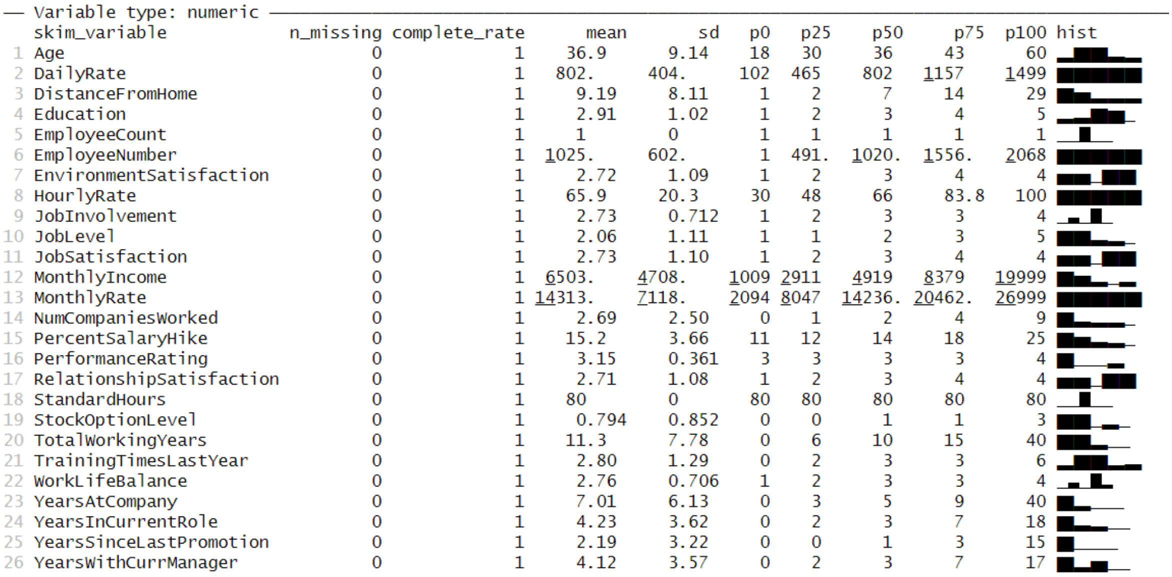
In this section we are going to discuss the implementation of an R model, utilizing the random forest algorithm, to try an predict the turnover rate of the IBM data set. The capabilities of the R language are numerous, given that is a statistical programming language, it gives as power over the manipulation of data and extracting findings over them.

We will begin our model with inserting the IBM data set, that was made for exercising in hr analysis. The given data set has 1470 observations with 35 attributes (variables). Although it is a made-up data set, with the plethora of the variables, it gives the opportunity to explore the many aspects of an organization, thus deploying a well defined model to be ready and used for real-world data.

As we explained in the previous section, we are using the data.table package, to make our work easier in means of speed, space and adaptability. As we use the caret package, the work is done better with data tables, than tibbles that uses the dplyr package.

```
library(data.table)
library(skimr)
library(ggplot2)
library(caret)
library(MLmetrics)
library(dplyr)
library(magrittr)
library(ggthemes)
library(cowplot)
current_date <- Sys.Date()

#Enter the IBM dataset
HRdfIBM <- fread("WA_Fn-UseC_-HR-Employee-Attrition.csv")

#With skim() we get intuition over our data
skim(HRdfIBM)
```

First we enter our data and get the current date in order to define a certain time frame. With the *skim()* function from the **skimr** package we get some useful insight over our data. As we see, *skim()* gives some useful values (mean,sd) and a histogram for every variable.



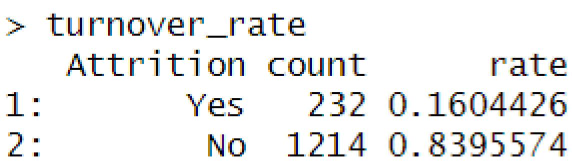
```
── Variable type: numeric ──────────────────────────────────────────────────────────────────────
   skim_variable            n_missing complete_rate    mean     sd   p0  p25    p50    p75  p100 hist
 1 Age                              0             1    36.9   9.14   18   30     36     43    60
 2 DailyRate                        0             1   802.  404.    102  465    802   1157  1499
 3 DistanceFromHome                 0             1     9.19  8.11    1    2      7     14    29
 4 Education                        0             1     2.91  1.02    1    2      3      4     5
 5 EmployeeCount                    0             1     1     0       1    1      1      1     1
 6 EmployeeNumber                   0             1  1025.  602.      1  491.  1020.  1556.  2068
 7 EnvironmentSatisfaction          0             1     2.72  1.09    1    2      3      4     4
 8 HourlyRate                       0             1    65.9  20.3    30   48     66     83.8  100
 9 JobInvolvement                   0             1     2.73  0.712   1    2      3      3     4
10 JobLevel                         0             1     2.06  1.11    1    1      2      3     5
11 JobSatisfaction                  0             1     2.73  1.10    1    2      3      4     4
12 MonthlyIncome                    0             1  6503.  4708.  1009 2911   4919   8379  19999
13 MonthlyRate                      0             1 14313.  7118.  2094 8047  14236. 20462. 26999
14 NumCompaniesWorked               0             1     2.69  2.50    0    1      2      4     9
15 PercentSalaryHike                0             1    15.2   3.66   11   12     14     18    25
16 PerformanceRating                0             1     3.15  0.361   3    3      3      3     4
17 RelationshipSatisfaction         0             1     2.71  1.08    1    2      3      4     4
18 StandardHours                    0             1    80     0      80   80     80     80    80
19 StockOptionLevel                 0             1     0.794 0.852   0    0      1      1     3
20 TotalWorkingYears                0             1    11.3   7.78    0    6     10     15    40
21 TrainingTimesLastYear            0             1     2.80  1.29    0    2      3      3     6
22 WorkLifeBalance                  0             1     2.76  0.706   1    2      3      3     4
23 YearsAtCompany                   0             1     7.01  6.13    0    3      5      9    40
24 YearsInCurrentRole               0             1     4.23  3.62    0    2      3      7    18
25 YearsSinceLastPromotion          0             1     2.19  3.22    0    0      1      3    15
26 YearsWithCurrManager             0             1     4.12  3.57    0    2      3      7    17
```

It widely accepted that employee satisfaction is related to compensation. We will define a categorical variable, that takes value according to the compensation that every employee gets in relation with the median compensation of its employee's peers. That is a straightforward observation, as in many organizations employees usually value their earnings according to those of their colleagues.

```
#We gonna classify some attributes by their respective values
HRdfIBM[ , `:=`(median_compensation =
                    median(MonthlyIncome)),by = .(JobLevel) ]
```

```
HRdfIBM[ , `:=`(CompensationRatio =
   (MonthlyIncome/median_compensation)), by =. (JobLevel)]
HRdfIBM[ , `:=`(CompensationLevel =
             factor(fcase(CompensationRatio
                      %between% list(0.75,1.25), ”Average”,
                      CompensationRatio
                      %between% list(0,0.74), ”Below”,
                      CompensationRatio
                      %between% list(1.26,2), ”Above”),
                  levels = c(”Below”,”Average”,”Above”))),
       by = .(JobLevel) ]
HRdfIBM <- na.omit(HRdfIBM)

#Here we count by attrition the turnover
turnover_rate <- HRdfIBM[, list(count = .N, rate =
   (.N/nrow(HRdfIBM))), by = Attrition]
turnover_rate
```

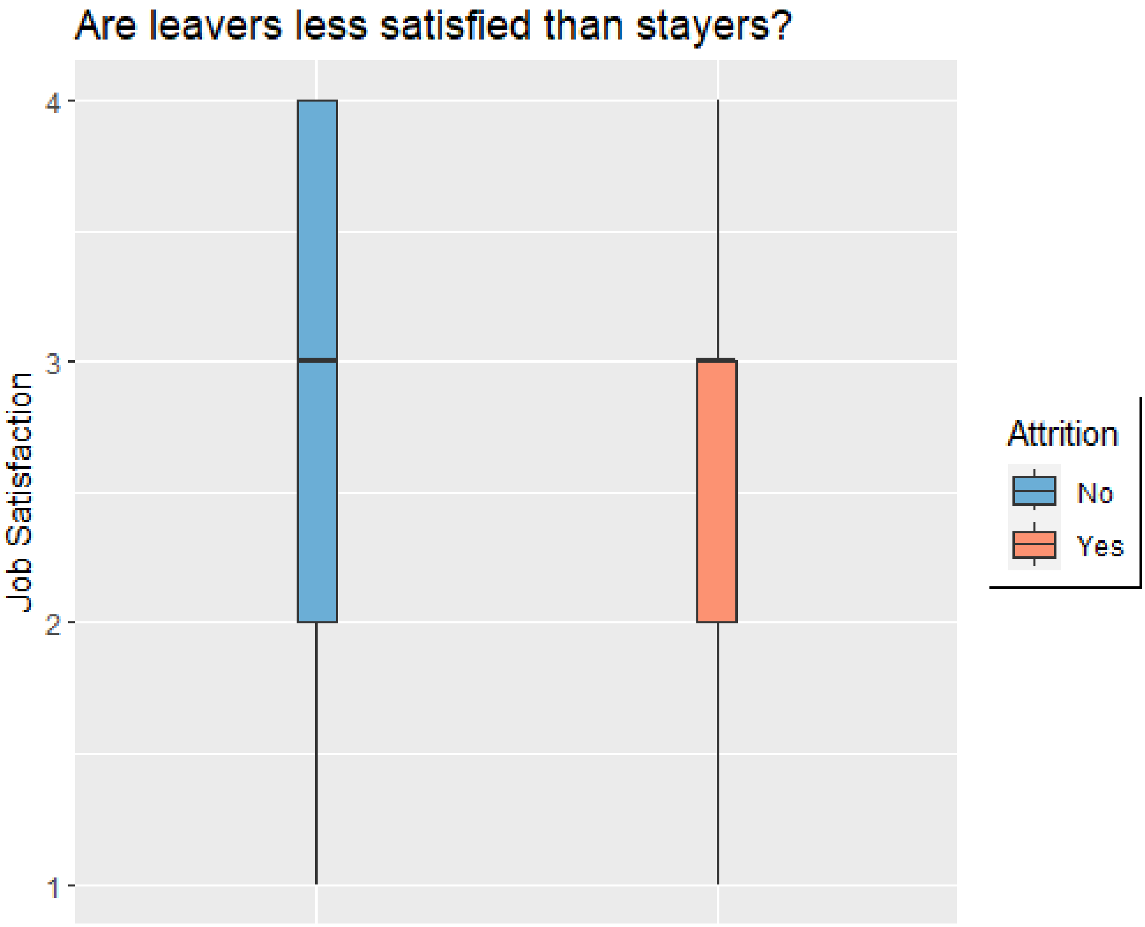
Within our given time frame we can see how many employees have left the organization. As we calculate the turnover based on the attrition variable that is given to us, we see that 232 employees have left. In order though, to get a more realistic outcome, we gonna split the data and sample them in order to get a view of a smaller organization. By getting 500 of our observations, we will test the ML model in a smaller group to get a notion of how a ML algorithm behaves in this case.



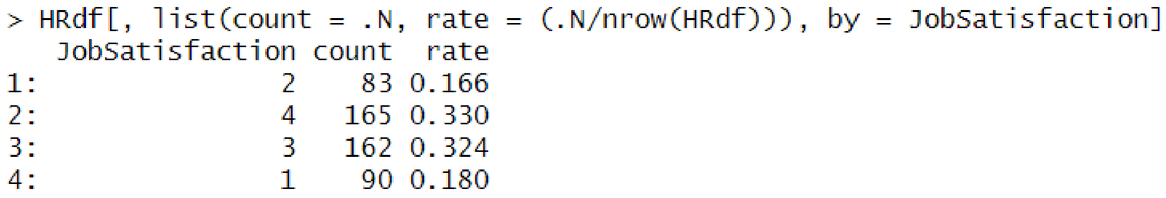
```
> turnover_rate
   Attrition count      rate
1:       Yes   232 0.1604426
2:        No  1214 0.8395574
```

```
set.seed(123)
HRdf <- HRdfIBM[sample(.N, 500)]

#We gonna see how many employees are satisfied with their jobs
theme_custom <- function(){
 theme(

   # background
   strip.background = element_rect(colour = ”black”, fill =
      ”lightgrey”),

   # x-axis already represented by legend
   axis.title.x = element_blank(),
   axis.ticks.x = element_blank(),
   axis.text.x = element_blank(),

   # legend box
   legend.box.background = element_rect())

}

# specify colors
library(RColorBrewer)
myCol <- rbind(brewer.pal(8, ”Blues”)[c(5,7,8)],
           brewer.pal(8, ”Reds”)[c(4,6,8)])




# Correlation: The higher employee’s job satisfaction,the lower
   the turnover rate.
# Boxplot
plot_jobsatisfaction <- ggplot(HRdf, aes(x = Attrition, y =
   JobSatisfaction,fill = Attrition))+
 geom_boxplot(width=0.1)+
 scale_fill_manual(values = myCol)+
 ylab(”Job Satisfaction”)+
 xlab(”Employee Attrition”)+
 theme_custom()+
 ggtitle("Are leavers less satisfied than stayers?")

ggsave(plot_jobsatisfaction,
   file=paste0(current_date,"_","Distribution_Jobsatisfaction.png"),
   width = 15, height = 10, units = "cm")
plot_jobsatisfaction
```

Now, let's examine how job satisfaction and compensation interact to determine employee churn. A typical hypothesis derived from the literature states that higher job satisfaction is associated with a lower likelihood of employee turnover — unhappy employees typically have more reason to leave because they expect to be happier somewhere else and do not feel as emotionally committed to their current organization, making it more desirable and easier to leave as soon as attractive alternatives are found [58].



```
HRdf[, list(count = .N, rate = (.N/nrow(HRdf))), by =
   JobSatisfaction]
```



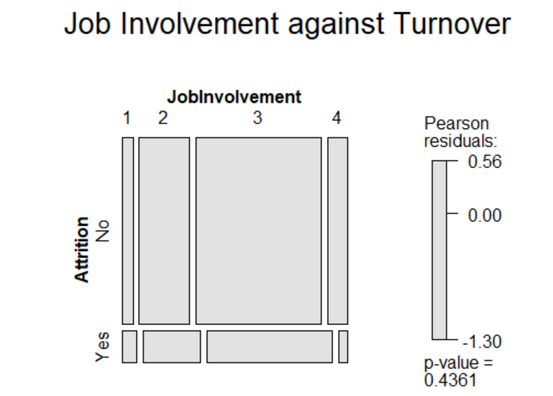
```
> HRdf[, list(count = .N, rate = (.N/nrow(HRdf))), by = JobSatisfaction]
   JobSatisfaction count  rate
1:               2    83 0.166
2:               4   165 0.330
3:               3   162 0.324
4:               1    90 0.180
```

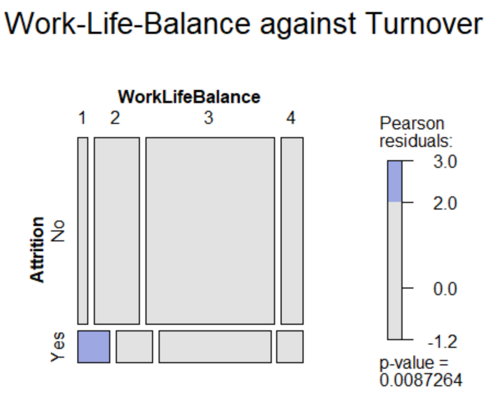
As we see from the Job Satisfaction count, we have more satisfied employees in the data set, as they amount in about 65%. That gives us the notion that job satisfaction is not the only thing that contributes to a person leaving the organization.

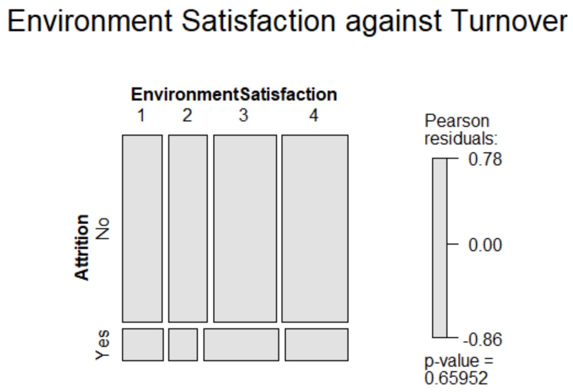
```
#We get some further insight from our analysis in attrition based
   on other variables
library(vcd)
library(vcdExtra)

mosaic(~ Attrition + EnvironmentSatisfaction, data = HRdf,
     main = "Environment Satisfaction against Turnover", shade =
        TRUE, legend = TRUE)
mosaic(~ Attrition + JobInvolvement, data = HRdf,
     main = "Job Involvement against Turnover", shade = TRUE,
        legend = TRUE)
mosaic(~ Attrition + WorkLifeBalance, data = HRdf,
     main = "Work-Life-Balance against Turnover", shade = TRUE,
        legend = TRUE)
mosaic(~ Attrition + RelationshipSatisfaction, data = HRdf,
     main = "Relationship Satisfaction against Turnover", shade =
        TRUE, legend = TRUE)
options(repr.plot.width=8, repr.plot.height=7)
```

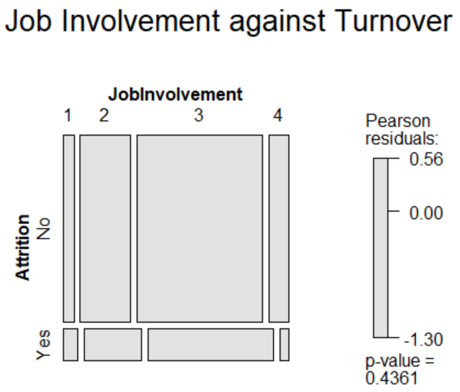
Furthermore, we can see some more analysis on various attributes that could signal an employee's choice in leaving the organization. With the importing of the **vcd** package, using the *mosaic()* function, we test whether the frequencies in our sample could have been generated by simple chance. By performing a Chi-squared test behind the scenes. The area of the rectangle represents the percentage of cases for any given combination of levels, and the color of the tiles indicates the degree relationship among the variables. The more the color deviates from grey, the more we must doubt statistical independence between the various factor combinations.

Only in the case of Work-Life Balance we see a minor deviation of statistical independence in the cases of those who are leaving the company. This can be explained further as the tiles that get near dark blue colors represent more random occurrence than expected. So despite the factors that contribute in an employee to leave, sometimes work-life balance can change everything solely.









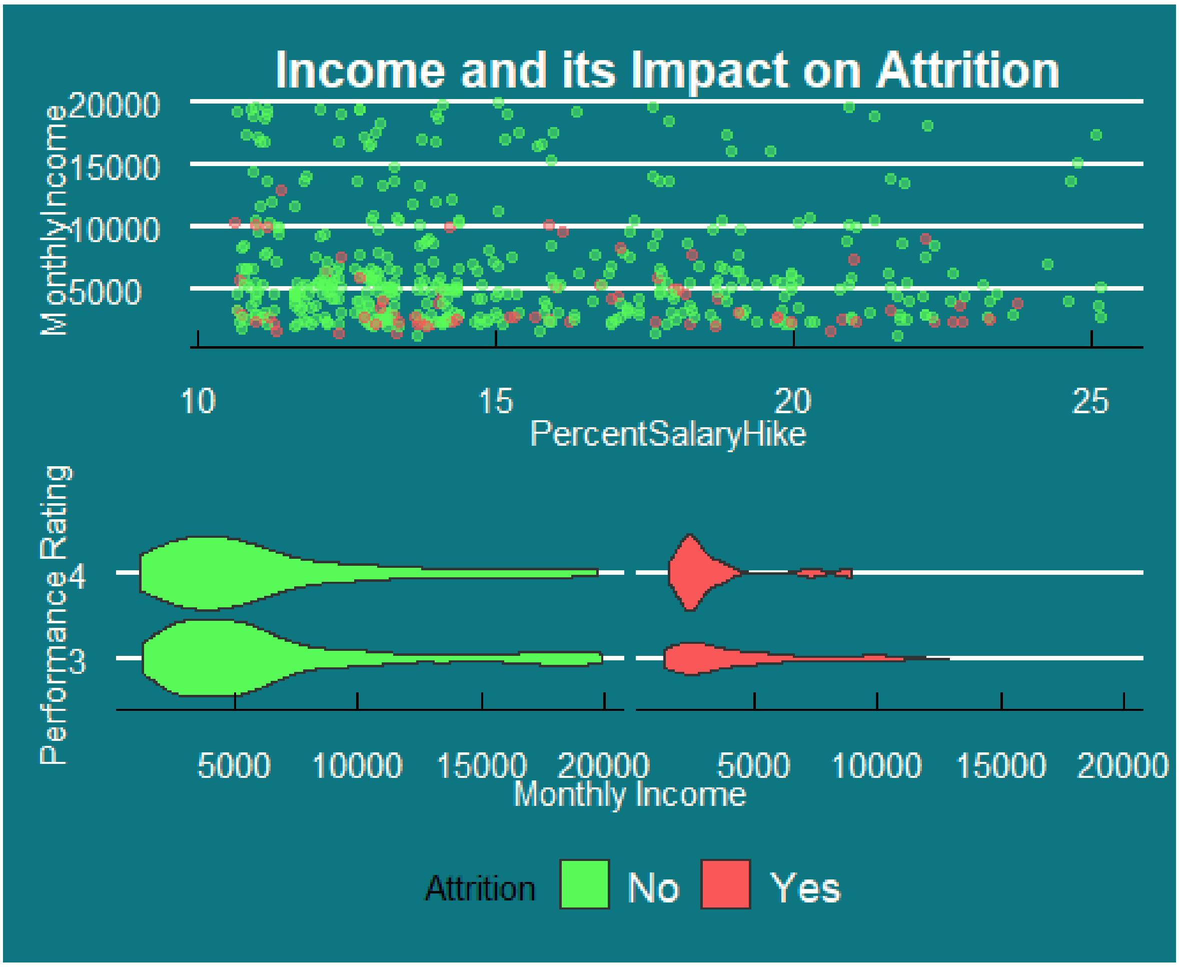
```
#This visualization took place with dplyr in order to be easier to
  read
per.sal <- HRdf %>% select(Attrition, PercentSalaryHike,
  MonthlyIncome) %>%
 ggplot(aes(x=PercentSalaryHike, y=MonthlyIncome)) +
   geom_jitter(aes(col=Attrition), alpha=0.5) +
 theme_economist() + theme(legend.position=”none”) +
   scale_color_manual(values=c(”#58FA58”, ”#FA5858”)) +
 labs(title=”Income and its Impact on Attrition”) +
   theme(plot.title=element_text(hjust=0.5, color=”white”),
   plot.background=element_rect(fill=”#0D7680”),
    axis.text.x=element_text(colour=”white”),
      axis.text.y=element_text(colour=”white”),
    axis.title=element_text(colour=”white”))

perf.inc <- HRdf %>% select(PerformanceRating, MonthlyIncome,
  Attrition) %>% group_by(factor(PerformanceRating), Attrition)
  %>%
 ggplot(aes(x=factor(PerformanceRating), y=MonthlyIncome,
   fill=Attrition)) + geom_violin() + coord_flip() +
   facet_wrap(~Attrition) +
 scale_fill_manual(values=c(”#58FA58”, ”#FA5858”)) +
   theme_economist() +
 theme(legend.position=”bottom”, strip.background =
   element_blank(), strip.text.x = element_blank(),
    plot.title=element_text(hjust=0.5, color=”white”),
      plot.background=element_rect(fill=”#0D7680”),
    axis.text.x=element_text(colour=”white”),
      axis.text.y=element_text(colour=”white”),
    axis.title=element_text(colour=”white”),
    legend.text=element_text(color=”white”)) +
 labs(x=”Performance Rating”,y=”Monthly Income”)
```

```
plot_grid(per.sal, perf.inc, nrow=2)
```

As examining the employee behaviour in a given organization, the strongest impact we can conclude is that of the income. People that are not happy with their pay are more likely to give up on a company that does not provide them with what makes them happy.

With the help of an other plot provided by The Economist team, we can see a beautiful visualization of how income impacts the performance of each organization's members.

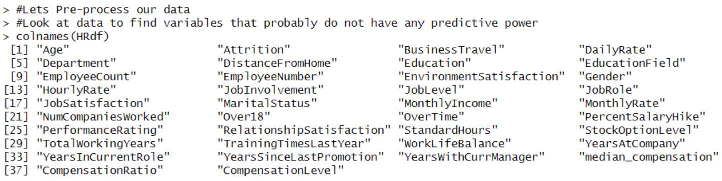
In the below graph, employees that did not leave the organization are represented with green, while those who left with red. A great insight is shown, as the members that get compensated lower are more likely to turn over. So, that gives us another observation. That employees who are paid more, they likely give attention to different aspects in order to stay in an organization (i.e. Work-Life, Management, Job Satisfaction). This is something that makes sense, as the prior need is payment, and if that aspect is satisfying employees can then turn their attention in more self-fulfilling factors in their given job.



7.2 The Pre-processing of the data

We now prepare our data set for the actual modeling phase. All variables that are highly unlikely to have any predictive value will be eliminated as a first step. Employee-ID, for instance, won't meaningfully explain any variation in employee turnover, so it should be removed for the time being along with some other variables. Variables that have a high degree of overlap with other features and could consequently cause multicollinearity problems are another example (e. G. Monthly income, hourly wage, etc.). By correctly converting all string variables to factors concurrently.

```
#Lets Pre-process our data
#Look at data to find variables that probably do not have any
   predictive power
colnames(HRdf)
```



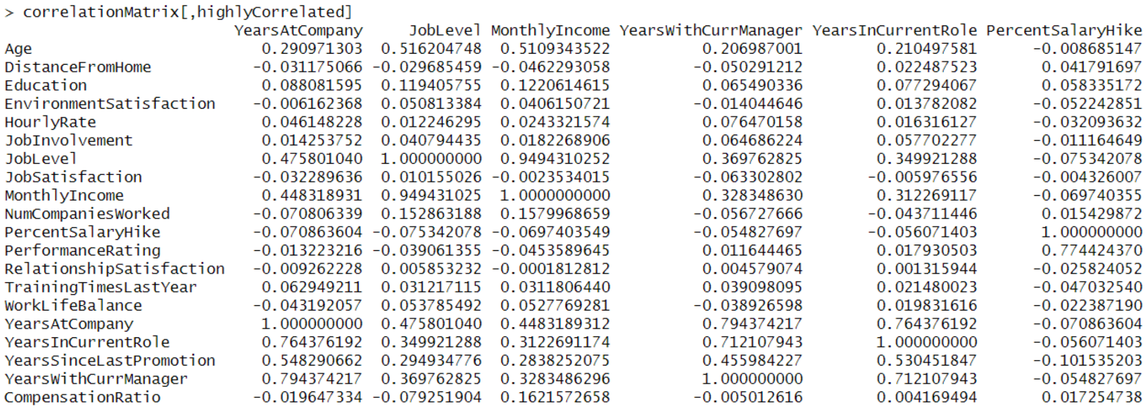
```
> #Lets Pre-process our data
> #Look at data to find variables that probably do not have any predictive power
> colnames(HRdf)
 [1] "Age"                    "Attrition"                "BusinessTravel"           "DailyRate"
 [5] "Department"             "DistanceFromHome"         "Education"                "EducationField"
 [9] "EmployeeCount"          "EmployeeNumber"           "EnvironmentSatisfaction"  "Gender"
[13] "HourlyRate"             "JobInvolvement"           "JobLevel"                 "JobRole"
[17] "JobSatisfaction"        "MaritalStatus"            "MonthlyIncome"            "MonthlyRate"
[21] "NumCompaniesWorked"     "Over18"                   "OverTime"                 "PercentSalaryHike"
[25] "PerformanceRating"      "RelationshipSatisfaction" "StandardHours"            "StockOptionLevel"
[29] "TotalWorkingYears"      "TrainingTimesLastYear"    "WorkLifeBalance"          "YearsAtCompany"
[33] "YearsInCurrentRole"     "YearsSinceLastPromotion"  "YearsWithCurrManager"     "median_compensation"
[37] "CompensationRatio"      "CompensationLevel"
```

```
#Clean up data
HRdf<- HRdf[,-c(”DailyRate”,”EducationField”,
   ”EmployeeCount”,”EmployeeNumber”,”MonthlyRate”,”StandardHours”,
”TotalWorkingYears”,”StockOptionLevel”,”Gender”,”Over18”,
   ”OverTime”, ”median_compensation”)]
HRdf_reduced <- as.data.frame(unclass(HRdf),stringsAsFactors=TRUE)

# find numeric values
nums <- unlist(lapply(HRdf_reduced, is.numeric))
# save numeric variables for later
HRdf_nums <- HRdf_reduced[,nums]
# show numeric variables
head(HRdf_nums)
# calculate correlation matrix
```

```
correlationMatrix <- cor(HRdf_nums)
# summarize the correlation matrix
correlationMatrix
# find attributes that are highly corrected (ideally >0.75)
highlyCorrelated <- findCorrelation(correlationMatrix, cutoff=0.5)
# print colnames of highly correlated attributes
colnames(HRdf_nums[,highlyCorrelated])
correlationMatrix[,highlyCorrelated]
```

We will automatically identify and eliminate any highly correlated variables to be fairly certain we have not missed any. To achieve this, we first determine whether there are any numerical variables, compute a correlation matrix, and look for correlations greater than 0.5.

So, there are indeed some variables that were flagged by our code — we should take a closer look at: “YearsAtCompany”, “JobLevel”, “MonthlyIncome”,“YearsInCurrentRole” and “PercentSalaryHike”.



```
> correlationMatrix[,highlyCorrelated]
                         YearsAtCompany     JobLevel MonthlyIncome YearsWithCurrManager YearsInCurrentRole PercentSalaryHike
Age                         0.290971303  0.516204748  0.5109343522          0.206987001        0.210497581      -0.008685147
DistanceFromHome           -0.031175066 -0.029685459 -0.0462293058         -0.050291212        0.022487523       0.041791697
Education                   0.088081595  0.119405755  0.1220614615          0.065490336        0.077294067       0.058335172
EnvironmentSatisfaction    -0.006162368  0.050813384  0.0406150721         -0.014044646        0.013782082      -0.052242851
HourlyRate                  0.046148228  0.012246295  0.0243321574          0.076470158        0.016316127      -0.032093632
JobInvolvement              0.014253752  0.040794435  0.0182268906          0.064686224        0.057702277      -0.011164649
JobLevel                    0.475801040  1.000000000  0.9494310252          0.369762825        0.349921288      -0.075342078
JobSatisfaction            -0.032289636  0.010155026 -0.0023534015         -0.063302802       -0.005976556      -0.004326007
MonthlyIncome               0.448318931  0.949431025  1.0000000000          0.328348630        0.312269117      -0.069740355
NumCompaniesWorked         -0.070806339  0.152863188  0.1579968659         -0.056727666       -0.043711446       0.015429872
PercentSalaryHike          -0.070863604 -0.075342078 -0.0697403549         -0.054827697       -0.056071403       1.000000000
PerformanceRating          -0.013223216 -0.039061355 -0.0453589645          0.011644465        0.017930503       0.774424370
RelationshipSatisfaction   -0.009262228  0.005853232 -0.0001812812          0.004579074        0.001315944      -0.025824052
TrainingTimesLastYear       0.062949211  0.031217115  0.0311806440          0.039098095        0.021480023      -0.047032540
WorkLifeBalance            -0.043192057  0.053785492  0.0527769281         -0.038926598        0.019831616      -0.022387190
YearsAtCompany              1.000000000  0.475801040  0.4483189312          0.794374217        0.764376192      -0.070863604
YearsInCurrentRole          0.764376192  0.349921288  0.3122691174          0.712107943        1.000000000      -0.056071403
YearsSinceLastPromotion     0.548290662  0.294934776  0.2838252075          0.455984227        0.530451847      -0.101535203
YearsWithCurrManager        0.794374217  0.369762825  0.3283486296          1.000000000        0.712107943      -0.054827697
CompensationRatio          -0.019647334 -0.079251904  0.1621572658         -0.005012616        0.004169494       0.017254738
```

```
#remove highly correlated variables to overcome multicollinearity
colnames(HRdf_nums)
highlyCorrelated <- c(1,7,11,16,17,19)
HRdf_nums <- HRdf_nums[,-highlyCorrelated]
```

We must convert these factor variables into dummy variables so that our machine learning algorithms can function. We will have a separate variable indicating whether or not the specific participant fits into this category for each factor level (e.g. a person who rarely travels would receive a 1 rather than a 0). First, all categorical variables will be found, with the exception of our target (attrition). Then, using caret’s *dummyVars()* function and our dataset, we'll build a new dataframe with our chosen set of numeric variables, dummy variables, and attrition (yes/no). Please take note that the **caret**-function *dummyVars()* converts the variables into a complete set of dummy variables, ensuring that no factor level is overlooked and that all are covered. This method does not apply to linear models, where the output is always compared to a reference level.

```
# select factor variables to convert, but leave Attrition out
vars_to_dummy <- HRdf_reduced[,sapply(HRdf_reduced, is.factor) &
   colnames(HRdf_reduced) != ”Attrition”]
head(vars_to_dummy)
# Create dummy variables with caret
dummies <- dummyVars( ~ ., data = vars_to_dummy)
# New dataframe to work with later
HRdf_sample <- data.frame(HRdf_dummy, HRdf_nums, Attrition =
   HRdf_reduced$Attrition)
View(HRdf_sample)
```

The variables that offer no predictive value will then be eliminated using the **caret**'s *nearZeroVar()* function. It applies to predictors with only one distinct value (i.E. E. An "absolute zero predictor"). In our statistical model, a constant would be created if all of our employees traveled frequently, leaving all other options (rare or none) blank. It also holds true for predictors with very low frequency distributions and only a few distinct values (e.g., if one out of every 100 workers got divorced). This could result in a model crash or an unstable fit for many models (aside from tree-based models).

```
# remove near zero variables (except for attr)
remove_cols <- nearZeroVar(HRdf_sample, names = TRUE)
remove_cols
# Get all column names
all_cols <- names(HRdf_sample)
# Remove from data
HRdf_final<- HRdf_sample[ , setdiff(all_cols, remove_cols)]
# make sure that factor levels of attrition are correctly ordered
levels(HRdf_final$Attrition)
HRdf_final$Attrition <- factor(HRdf_final$Attrition, levels =
   c(”Yes”, ”No”))
# double check
levels(HRdf_final$Attrition)
```

The cleaned-up version of our final data set is now prepared for modeling. Due to our small sample size, we risk over-fitting our model to the point where we are unable to use it in the future with new employee data. Over-fitting is a problem that frequently arises and may be the cause of our inability to replicate previously discovered effects. To get around this problem with machine learning models, we frequently divide our data into training and validation/test sets. The model is trained using a training set, and it is tested on data that it has never seen before using a validation/test set. If we applied the conventional 80/20-split to our use case, model performance would be largely dependent on chance because it would change each time the algorithm randomly selected 25 people to test. Assuming that only about 20% of employees left the company, this means that our model would most likely be tested on about 5 departing employees and 20 remaining, raising the question of whether the algorithm would behave consistently across cases. Additionally, the outcome would easily overestimate the model’s actual performance because there are so many examples of success to use as a benchmark.

```
#5fold cross-validation
myFolds <- createFolds(HRdf_final$Attrition, k = 5)
```

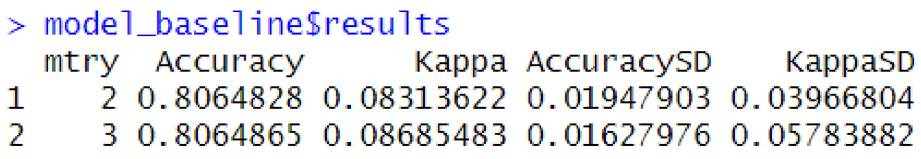
7.3 Model development: A Random Forest implementation

To build our machine learning models, we will set up a reusable train Control object with the same settings: repeated cross-validation ensures that we run our 5-fold-cross-validation process 5 times. Additionally, we ask **caret** to include class probabilities in our model output along with the results of our final predictions because we want to track the development of our modeling procedure.

The model we will be using is Random Forests. As basic decision trees are interpretable models that are constructed in a tree-like manner: the branches are feature combinations, and the leaves are the class labels of interest (e.g. (Yes or No). By combining the strength of numerous weak learners to make a collective prediction, random forests give us an advantage over simple decision trees. Because the final prediction is not dominated by a small number of significant predictors, it is more robust than simple decision trees.

```
#Control function for the model
myControl <- trainControl(
  method = ”cv”,
  number= 15,
  classProbs = TRUE, # IMPORTANT!
  verboseIter = TRUE,
  savePredictions = ”final”,
  index = myFolds
)

#Fit random forest model
model_baseline <- caret::train(
  Attrition ~ MonthlyIncome + JobSatisfaction +
      MonthlyIncome*JobSatisfaction,
  data = HRdf_final,
  method = ”rf”,
  tuneLength = 10,
  trControl = myControl
)



model_baseline$results
summary(model_baseline)
```

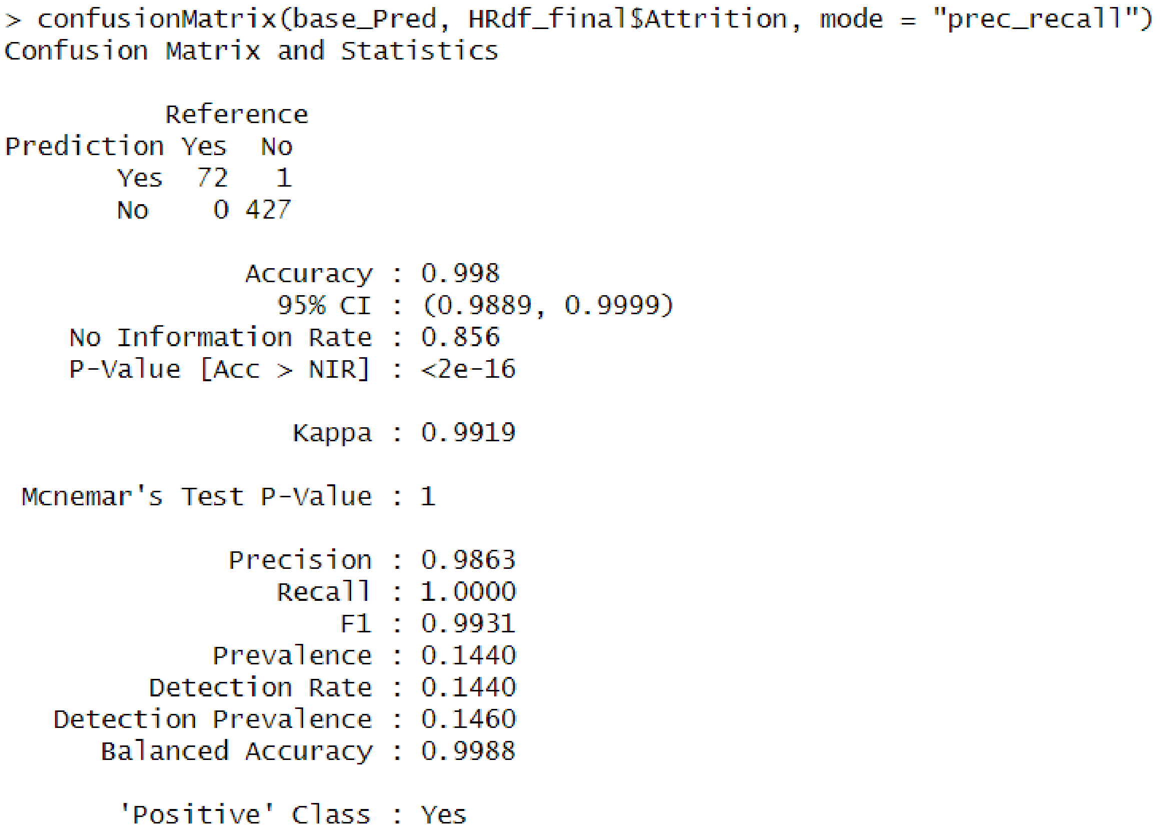
As we get some insight about the accuracy and Kappa value of our model, we observe that we have a good behaviour of our model when mtry is 2 and 3. This is an important value for the random forest algorithm, as it determines the random split according to the number of variables.



```
> model_baseline$results
  mtry  Accuracy      Kappa AccuracySD    KappaSD
1    2 0.8064828 0.08313622 0.01947903 0.03966804
2    3 0.8064865 0.08685483 0.01627976 0.05783882
```

```
# Create confusion matrix
base_Pred <- predict.train(model_baseline, HRdf_final)
confusionMatrix(base_Pred, HRdf_final$Attrition, mode =
   ”prec_recall”)
```

After implementing the model, the confusion matrix gives us the final observation. The results are outstanding. As we see all the predictions fall in true positive or true negative, except for one case that we have a false positive (an employee that is mistakenly believed to leave the organization). It is natural for these cases to occur, because we have to manage people data and the factors that affect an employee to leave are complicated. Although this difficulty our algorithm gives an accuracy of 0.98. The random forest is one of the best algorithms to predict, and when the outcome is fixed in two values, the job gets done easier.

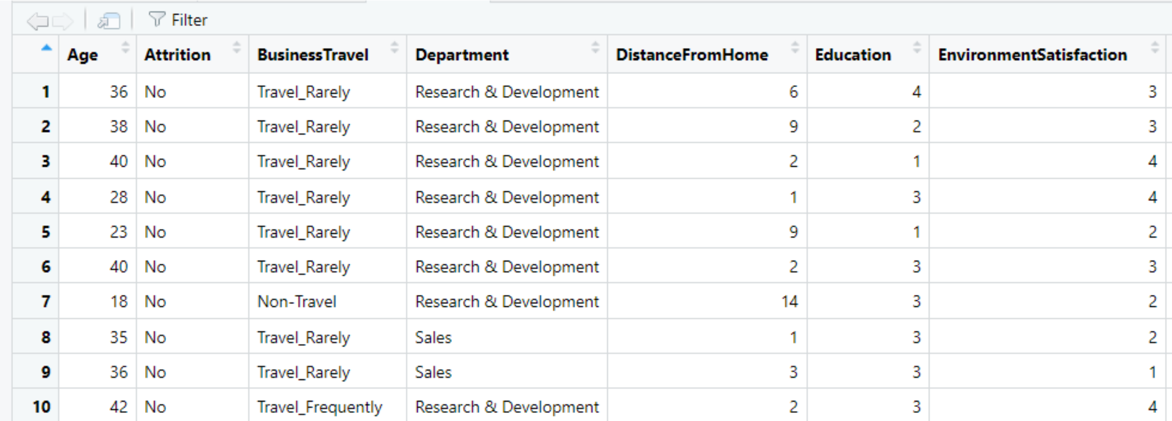


```
> confusionMatrix(base_Pred, HRdf_final$Attrition, mode = "prec_recall")
Confusion Matrix and Statistics

          Reference
Prediction Yes  No
       Yes  72   1
       No    0 427

               Accuracy : 0.998
                 95% CI : (0.9889, 0.9999)
    No Information Rate : 0.856
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9919

 Mcnemar's Test P-Value : 1

              Precision : 0.9863
                 Recall : 1.0000
                     F1 : 0.9931
             Prevalence : 0.1440
         Detection Rate : 0.1440
   Detection Prevalence : 0.1460
      Balanced Accuracy : 0.9988

       'Positive' Class : Yes
```

```
#Get indices of employees that still work for the company
still_active <- setDT(HRdf_final)[Attrition == ”No”, which = TRUE]
probs <- predict.train(model_baseline, HRdf_final, type = ”prob”)
```

```
# Save predicted probs for still active employees
risks <- probs[still_active,]
emp_active <- HRdf[still_active,]
# Save row index
risks$index <- 1:nrow(risks)
#find employees who may be at risk leaving the company
emp_risk <-setDT(risks)[order(-Yes)]
emp_risk
# show employee data who could leave soon
emp_indices <- emp_risk$index
top_10 <- head(emp_active[emp_indices,],10)
View(top_10)
```

7.4 From Prediction to Strategy



	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EnvironmentSatisfaction
1	36	No	Travel_Rarely	Research & Development	6	4	3
2	38	No	Travel_Rarely	Research & Development	9	2	3
3	40	No	Travel_Rarely	Research & Development	2	1	4
4	28	No	Travel_Rarely	Research & Development	1	3	4
5	23	No	Travel_Rarely	Research & Development	9	1	2
6	40	No	Travel_Rarely	Research & Development	2	3	3
7	18	No	Non-Travel	Research & Development	14	3	2
8	35	No	Travel_Rarely	Sales	1	3	2
9	36	No	Travel_Rarely	Sales	3	3	1
10	42	No	Travel_Frequently	Research & Development	2	3	4

As it is not shown in the picture the employees that were selected by the algorithm fall in the Average or Below categories of Compensation Level

The final step is to have an opportunity in gaining power over our data. The predictions would serve a company to implement a strategy in order to avoid any loss. We hope to use the model to improve retention in our small business. In order to achieve this, we will first obtain the active employee indexes and then use our model to predict the likelihood that these employees will leave. Then, in addition to the actual employee data, we will save these probabilities. The top 10 workers who pose the greatest risk of leaving the company will be identified last. In order to give the company a chance to intervene, these are probably the people that should be recruited first to find out what they need to be happier and how they would like to develop in the future. Hopefully, this will allow us to address any voluntary turnover. After all, it’s good to give employees the chance to get rid of critical feedback that could improve the working environment. In the end, we will give the managers a complete list of employees to talk to, ranked by their risk to leave.

Chapter 8

Conclusions and further discussion

Human Resources management has the opportunity to gain a strategic place when it comes to organization evolution. Despite the importance of the human capital, in our days there is still low involvement of the HR when it comes to the strategy of a given organization.

Structuring the importance of HR, researchers begin to give attention to further aspects, that can be optimized, by implementing behavioural science and computer algorithms. Making HR a new science, and emphasize a new era, where management alongside computer algorithms, will build the environment that makes employees feel safe and productive. Despite the fears of malicious use, algorithms can give an extension to the powers of HR, and with proper use, differentiate the role of employees.

As it is shown, in this thesis, we built up in the scenario where AI and analytics can maintain employee dignity and not peculate them for the sake of efficiency. A serious pitfall, which many critics of analytics fear, as numerous examples of AI wrong use occurred. In addition, the seriousness of the privacy of the employee data, is a one of the most critical challenges analytics have to surpass. Without respecting employee privacy, and using their sensitive data managers can make unfair decisions and compromising the whole function of HR.

Nevertheless, a strong background is being built, so as analytics to have the role they deserve. Being a tool for the management and giving insights, where human bias blocks, and not leading major processes and manipulating employees as mere data.

The results of the model in predicting turnover contribute in the main argument of this thesis, to make HR a strategic partner. By predicting the potential turnover of a company, HR is in a position to plan the next moves. Given that recruiting is one the costliest procedures in any organization, by being in a position to predict who may leave, management has the opportunity to implement a handful of strategies:

- Optimizing retention. Given that a model can predict employees who want to leave, an organization can come to terms to make that person stay by reaching a mutual agreement.

- Optimized recruiting. If say, a company has predicted that 5 employees are about to leave by the end of the year, can begin preemptively the process of filling up the vacant positions. In addition, if the turnover is based upon a certain reason, targeted recruitment would be useful strategy.

- Work conditions. A model is in a position to show the manager, which is the most critical reason, employees leave. If this reason is related to job circumstances, then a change must be done in order to keep employees inside.

- Management satisfaction. A common issue for employees to leave, is that they are not satisfied with management conditions. Usually executives maintain strong attitude against the employees, and lowering their morale. That is why organizations that make use of more humane approaches and applaud feedback, give employees the needed confidence to build upon the organization culture, and stick up with it.

Emphasizing the importance of the human capital, a smooth augmentation analytics in the HR management gives huge benefits. Further research is critical, in order to make the most out of analytics integration in strategic decision making about human development. From one side behavioural science and cognitive psychology need to be major counselors in the process of HR development. A strong background in those fields is needed when it comes to human managing. From the other side, further research is needed in how well machine learning algorithms can be trained, when manipulating human capital data. As of issues on privacy are raised, a thorough research is of critical importance in how effectively algorithms can be unbiased when performing HR processes.

Finally, this thesis can have a role for staging a basis for those who are interested in the field of HR analytics, and how data mining algorithms can optimize processes that we are not used to seeing them perform. Even though the availability of this data is scarce, even for research reasons, made up data sets can be used as a basis in testing various models and further contributing to this evolving sector of computer science.

Bibliography

[1] S. Snell and S. Morris. *Managing Human Resources.* Boston MA USA: Cengage Learning;, 18 edition, 2018.

[2] Mary Anne Devanna, Charles Fombrun, and Noel Tichy. Human resources management: A strategic perspective. *Organizational Dynamics*, 9(3):51–67, 1981.

[3] C. Duhig. *The Power of Habit.* Random House Books, 1 edition, 2013.

[4] Stone, Dianna, Deadrick, Lukaszewski, Kimberly, Johnson, and Richard. The influence of technology on the future of human resource management. *Human Resource Management Review*, 25, 06 2015.

[5] Weena Yancey Momin, , Kushendra Md.Abdul, and Mishra. Hr analytics as a strategic workforce planning. *Internationl Journal of Applied Approach*, 1(4):258–260, April 2015.

[6] Brian Becker and Mark Huselid. Strategic human resources management: Where do we go from here? *Journal of Management - J MANAGE*, 32, 12 2006.

[7] David P. Lepak and Scott A. Snell. The human resource architecture: Toward a theory of human capital allocation and development. *The Academy of Management Review*, 24(1):31–48, 1999.

[8] Barry Gerhart. Human resources and business performance: Findings, unanswered questions, and an alternative approach. *management revue. The International Review of Management Studies*, 16:174–185, 01 2005.

[9] D. Angrave, A. Charlwood, I. Kirkpatrick, M. Lawrence, and M. Stuart. Hr and analytics: why hr is set to fail the big data challenge. *Human Resource Management Journal*, 26(1):1–11, 2016.

[10] Tobias M. Scholz. *Big Data in Organizations and the Role of Human Resource Management: A Complex Systems Theory-Based Conceptualization*. Peter Lang AG, ned - new edition edition, 2017.

[11] Jon Baggaley. Global educational technology: A luddite view. In Zoraini Wati Abas, Insung Jung, and Joseph Luca, editors, *Proceedings of Global Learn 2010*, pages 1141–1149, Penang, Malaysia, May 2010. Association for the Advancement of Computing in Education (AACE).

[12] Simon Kelly. Towards a negative ontology of leadership. *Human Relations*, 67(8):905–922, 2014.

[13] Jackie Ford. Going beyond the hero in leadership development: The place of healthcare context, complexity and relationships; comment on “leadership and leadership development in healthcare settings – a simplistic solution to complex problems?”. *International journal of health policy and management*, 4:261–3, 04 2015.

[14] Yuval N. Harari. *Sapiens : a Brief History of Humankind*. New York: Harper, 1 edition, 2015.

[15] Norbert Weiner. *The Human Use of Human Beings*. Houghton Mifflin Company, 1 edition, 1949.

[16] P .Dahlbom, N. Siikanen, P. Sajasalo, and M. Jarvenpää. Big data and hr analytics in the digital era. *Baltic Journal of Management*, 15(1):120–138, 2020.

[17] A. Levenson Ph.D. Harnessing the power of hr analytics. *Strategic HR Review*, 4(3):28–31, July 2005.

[18] H. Chalutz Ben-Gal. An roi-based review of hr analytics: practical implementation tools. *Personnel Review*, 48(6):1429–1448, July 2019.

[19] S. Malisetty, R.V. Archana, and K.V. Kumari. Predictive analytics in hr management. *Indian Journal of Public Health Research & Development*, 8, July 2017.

[20] Sugandha Agarwal and Khalid Mohammed Saif Al Qouyatahi. Hrm challenges in the age of globalization. *International Research Journal of Business Studies*, 10:89–98, 2017.

[21] Shrivastava, Shweta, Nagdev, Kritika, Rajesh, and Anupama. Redefining hr using people analytics: the case of google. *Human Resource Management International Digest*, 26:3–6, 03 2018.

[22] Liu, Liyuan, Akkineni, Sanjoosh, Story, Paul, Davis, and Clay. Using hr analytics to support managerial decisions: A case study. 02 2020.

[23] Dr. Abdul Quddus Mohammed. Hr analytics: A modern tool in hr for predictive decision making. *Journal of Management*, 6(3):51–63, June 2019.

[24] David Anderson, Margrét V. Bjarnadóttir, and David Gaddis Ross. Using people analytics to build an equitable workplace. *Harvard Business Review*, 01 2022.

[25] Fernandez, Vicenc, Gallardo, and Eva. Tackling the hr digitalization challenge: key factors and barriers to hr analytics adoption. *Competitiveness Review: An International Business Journal*, ahead-of-print, 07 2020.

[26] Michael DiClaudio. People analytics and the rise of hr: how data, analytics and emerging technology can transform human resources (hr) into a profit center. *Strategic HR Review*, 18, 02 2019.

[27] Petersen, Emily January, Martin, and Breeanne Matheson. Misuse, play, and disuse: Technical and professional communication's role in understanding and supporting website owners' engagement with google analytics. In *2015 IEEE International Professional Communication Conference (IPCC)*, pages 1–5, 2015.

[28] Don Peck. They're watching you at work. *The Atlantic*, 2013.

[29] Henri de Romrée, Bruce Fecheyr-Lippens, and Bill Schaninger. People analytics reveals three things hr may be getting wrong. *McKinsey Quarterly*, 2016.

[30] Ben Waber. *People analytics: How social sensing technology will transform business and what it tells us about the future of work.* FT Press, 1 edition, 2013.

[31] A. Rai and L.B Singh. Artificial intelligence-based people analytics transforming human resource management practices. *Emerald Studies in Finance, Insurance, and Risk Management*, pages 229–244, 12 2023.

[32] Caryl Jimenez. A neural-fuzzy network approach to employee performance evaluation. *International Journal of Advanced Trends in Computer Science and Engineering*, 8:573–581, 06 2019.

[33] A. Hemalatha, P.Barani Kumari, Nishad Nawaz, and Vijayakumar Gajenderan. Impact of artificial intelligence on recruitment and selection of information technology companies. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 60–66, 2021.

[34] Sunghoon Kim, Ying Wang, and Corine Boon. Sixty years of research on technology and human resource management: Looking back and looking forward. *Human Resource Management*, 60(1):229–247, 2021.

[35] Lisa Marie Giermindl, Franz Strich, Oliver Christ, Ulrich Leicht-Deobald, and Abdullah Redzepi. The dark sides of people analytics: reviewing the perils for organisations and employees. *European Journal of Information Systems*, 31(3):410–435, 2022.

[36] van den Heuvel, Sjoerd, Bondarouk, and Tanya. The rise (and fall?) of hr analytics: A study into the future application, value, structure, and system support. *Journal of Organizational Effectiveness: People and Performance*, 4, 06 2017.

[37] Benjamin Erb. Human resource management in the age of big data. 09 2016.

[38] Evgeny Morozov. Information consumerism the price of hypocrisy. 2013.

[39] Bo Cowgill and Catherine Tucker. Economics, fairness and algorithmic bias. *SSRN Electronic Journal*, 01 2019.

[40] R.H. Hamilton and William Sodeman. The questions we ask: Opportunities and challenges for using big data analytics to strategically manage human capital resources. *Business Horizons*, 63, 10 2019.

[41] Paul Dourish. Algorithms and their others: Algorithmic culture in context. *Big Data & Society*, 3, 2016.

[42] Mike Ananny and Kate Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3):973–989, 2018.

[43] C. Kagan, M. Burton, P. Duckett, R. Lawthom, and A. Siddiquee. *Critical Community Psychology: Critical Action and Social Change*. Routledge, 2 edition, 2019.

[44] W. David Holford. The future of human creative knowledge work within the digital economy. *Futures*, 105:143–154, 2019.

[45] Samer Faraj, Stella Pachidi, and Karla Sayegh. Working and organizing in the age of the learning algorithm. *Information and Organization*, 28(1):62–70, 2018.

[46] Scott E. Page. *The Model Thinker: What You Need to Know to Make Data Work for You*. Basic Books, Inc., USA, 2018.

[47] Ph.E. Tetlock and D. Gardner. *Superforecasting: The Art and Science of Prediction*. Crown Publishers, 1 edition, 2015.

[48] Jac Fitz-enz and John R. Mattox. *Predictive Analytics for Human Resources*. The Wiley & SAS Business Series, 1 edition, 2014.

[49] Yagang Zhang. *New Advances in Machine Learning*. IntechOpen, Rijeka, Feb 2010.

[50] Batta Mahesh. Machine learning algorithms -a review. 01 2019.

[51] Susmita Ray. A quick review of machine learning algorithms. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMIT-Con)*, pages 35–39, 2019.

[52] H. Tsukimoto. *Logical Regression Analysis: From Mathematical Formulas to Linguistic Rules*, pages 21–61. 10 2005.

[53] Madan Somvanshi, Pranjali Chavan, Shital Tambade, and S. V. Shinde. A review of machine learning techniques using decision tree and support vector machine. In *2016 International Conference on Computing Communication Control and automation (IC-CUBEA)*, pages 1–7, 2016.

[54] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 10 2001.

[55] Jaime Lynn Speiser, Michael E. Miller, Janet Tooze, and Edward Ip. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, 134:93–101, 2019.

[56] Biau, Gérard, Scornet, and Erwan. A random forest guided tour. *TEST*, 25, 11 2015.

[57] Kliegr, Tomáš, Bahník, Štěpán, Fürnkranz, and Johannes. Advances in machine learning for the behavioral sciences. *American Behavioral Scientist*, 64:145–175, 02 2020.

[58] Ryan D. Zimmerman, Brian W. Swider, and Wendy R. Boswell. Synthesizing content models of employee turnover. *Human Resource Management*, 58(1):99–114, 2019.

APPENDICES

Appendix

In this section we will give some guidance for anyone that is interested in replicating the same experiment:

1. In the first place, the environment has to be established. Open the Rstudio, by downloading and installing it from `https://posit.co/download/rstudio-desktop/`
2. Following, there is needed to start project. The steps: File>New Project > New Directory. After that, a new directory for the project will be created.
3. Go to `https://github.com/JohnAstli/Turnover-Prediction-R`. To find the code of this thesis. We added the data set as well, in order for anyone interested to have it in one place. Download the needed files.
4. Afterwards, the code file and the data set have to be placed in the same directory. THe directory that was created by the Project in RStudio.
5. Before running the code, the needed packages must be installed. The user will find them in the top of the code. The packages that are used, are the same that must be installed as well, via the command ***install.package(Name Of Package)***.
6. By running the whole code, as a source(The option is given by the RStudio), the same results will be produced.
7. Finally the user is free to make changes and produce different results as well.