Lund University
Department of Political Science

WPMM43
Tutor: Nils Droste

# Digitalization and Technostress

*An Exploratory Data Analysis*

*of the Norwegian Workforce*

Andre Hansen

# Abstract

Digitalization has grasped almost every aspect of work in Norway in similarity to the developed world. Although many advantages are followed with the implementation of digital tools to bolster the work effort, there is a widespread downside, technostress. This phenomenon engulfs large parts of the workforce and poses a challenge for any foreseeable future, and there is a general lack of research and knowledge on the topic. However, from the little empirical work available about technostress, it is known that there are psychosocial consequences such as burnout taking a heavy toll on the workforce. Without sufficient understanding of the extent and perks of technostress, mitigation, and management of it become unlikely. In order to enhance the ability to ameliorate technostress and the consequences of it, it is necessary to enrich the understanding of the phenomenon. Henceforth, this thesis abductively explores the sole data available on the topic in Norway while generating tentative hypotheses. These include combinations of the factors that are significantly associated with technostress. The main tentative findings suggest that in the Norwegian workforce, a combination of time-pressure, work-overload, and availability is caused by the use of digital tools as the most common recipe for technostress.

*Keywords*: Digitalization, Technostress, Exploratory Data Analysis, Machine Learning, Tree-Based Methods
Words: 19182

# Table of contents

# 1 Introduction

Technostress, a consequence of digitalization, has been mentioned in literature since Brod (1984) coined the term, describing it as “a modern disease of adaptation caused by inability to cope with new computer technologies in a healthy manner”. Approximately a third of Norwegian workers are affected (Torvatn, Kløve & Landmark, 2017), potentially leading to the same consequences as in Germany where technostress accounted for up to 10% of sick leaves (Sicking, 2011). Furthermore, the development of digitalization continues its upward curve as both the public sector is digitalizing its services and work procedures, so is the private sector (Regjeringen, 2018). The intentions of which are aimed at utility and effectiveness, which has proven fruitful (Østvold & Rehbinder, 2017), yet the backside has also been taking hold according to the prominent authors on technostress.

Although the first mention of the term came in 1984, the topic is still in development as in only entered the mainstream research literature in 2007, from which it only starts building theories (Tarafdar, Cooper, & Stich, 2019, p. 7). Henceforth, technostress has yet to enter the political and managerial agenda. Earlier research has discovered an exhaustive list of causal mechanisms behind technostress, and its ameliorators finding some consensus (Ayyagari, Grover, & Russell, 2011; Ragu-Nathan, Tarafdar, Ragu-Nathan, & Tu, 2008; Tarafdar, Tu, & Ragu-Nathan, 2011; Tarafdar, Pullins, & Ragu-Nathan, 2014). This research is backed by epidemiological cognitive appraisal research, establishing the medical relevance of the phenomenon (Arnetz, 1997; Fox, Dwyer, & Ganster, 1993; Lazarus, 1991; Shultz, Wang, & Olson, 2010). At this point, there is an established, yet limited international research with no widely accepted theories expressing the exact mechanisms behind technostress in the work-life.

In the case of Norway, from both the perspectives of a highly digitalized country sporting high count of digital tools and connectivity in the workplace while holding a social-democratic welfare model (Esping-Andersen, 1990) that incorporates an “explicit legislation relating to the psychological work environment” (Bambra, Lunau, Van der Wel, Eikemo, & Dragano, 2014, p. 136). These can be partially jeopardized with unacceptably high work-demands from digitalization and technostress timely, leading to burnouts. Which makes it a delicate balance with the regulations restricting poor working conditions, such as working outside of paid hours, and being dangerously overworked during these hours, while digital tools facilitate these backdrops (Mensah & Adjei, 2020, p. 3). One survey conducted by SINTEF about the negative consequences of digitalization in the workforce (Torvatn, et al., 2017). Given this knowledge and likewise, from foreign reports and research, there is a motivation to pinpoint the variable combinations that are associated with technostress. This is to provide a better understanding of the patterns leading to this ever more relevant phenomenon. Provided the lack of time-series data, established theories, and limited knowledge, this thesis utilizes an abductive exploratory data analysis (EDA) approach utilizing tree-based methods. This is also with respect to the data that in the original form sports more than 80 variables, for which machine-learning algorithms can effectively aid the discovery of patterns in the data. From which the tentative hypotheses are generated to facilitate subsequent confirmatory research.

The thesis starts with summarizing the background in detail, then moving towards the status quo knowledge literature on the phenomenon of technostress breaking down the terminology that reflects the categories in the interview-based data. The following section orients about the methodological consideration, such as critical realism epistemology and ontology, followed by EDA application, and the tree-based methods. Afterward, when the decision tree algorithms and random forest functions are introduced, the paper continues to present the data, breaking it down to its components for the subsequent analysis. This analysis starts with descriptive statistics and concludes with the aforementioned machine-learning methods. The principle target is the variable transformed into a dichotomous perception of digital tools, namely technostress, and not. The thesis is then concluding with the discussion of the findings.

The EDA utilizing machine-learning methods on the data supplied by SINTEF (Torvatn et al. 2017) yielded multiple hypotheses. Among the most notable, are the (1) combination of digital tools causing perceptions of time-pressure, workload, and availability as the combination that has the most probable association with perceptions of technostress; (2) stress due to workload and availability regardless of it being caused by digital tools or not, enhances the risk of perceiving technostress; (3) When combined with time-pressure, and workload, perceiving availability due to digital tools increases the chances of perceiving technostress, especially if availability is perceived positively. These hypotheses have no confirmation in such covariation among the response variables as described above in the existing literature and serves as a platform for future confirmatory studies.

## 1.1 Aim of the Thesis and the Research Questions

The thesis has the aim of exploring the data available on the phenomenon of technostress in the Norwegian workforce to create new hypotheses and build on the existing theories of technostress and digitalization. By conducting this work, a more specific and unexpected view of the technostress phenomenon fueled by digitalization can lead to future findings and facilitate accurate further research. Which is the goal of the exploratory data analysis. Provided the very limited knowledge on the topic in general, not to speak of Norway that has only a single survey on the matter, this approach is a perfect match. By using the exploratory data analysis approach informed by earlier quantitative and international literature, theoretical, and qualitative research this thesis has the aim to provide an unmatched oversight.

Technostress literature covers lists of all the predictors associated with technostress and its amelioration, timely with a correlation coefficient to approach covariations among multiple predictors and the response. This thesis utilizes tree-based methods for this purpose and aims, in line with EDA practice, to generate hypotheses of combinations of factors for future confirmatory research. Tree-based methods are suitable to observe covariations among predictors that have an association with the response, technostress in this case. Hence, given that this is not a deductive approach, but abductive, having the literature only as reference of context, the research questions are as follows.

RQ1: What hypotheses can be generated about unexpected patterns in the data regarding technostress?

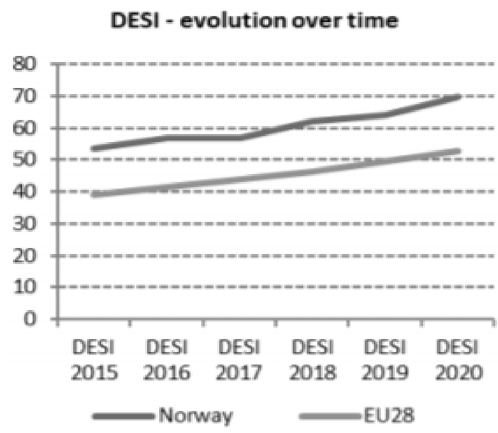
RQ2: What hypotheses can be generated regarding the covarying predictors association with technostress?

This setup facilitates the most openminded EDA fashion in order to remain data-oriented as it is not based on predetermined theory. Technostress will remain a problem, to be addressed from multiple perspectives, and will have growing relevance both intra and extra-scientifically. As a better understanding serves a purpose for anyone in a work environment with digital tools for any foreseeable future, and the intra-scientific purpose is the enriching of an emergent topic in a distinctive style. This is the only research utilizing EDA and machine-learning methods in the sphere of technostress, hence, providing unique insights given this configuration. Earlier research lacks observing the most likely combinations that can help better undestand, detect or avoid technostress The gap comes with identifying the factors that in combination produce the most homogenous groups technostressed and not. Provided the utility of the tree-based methods, this thesis fills the gap in the literature than has so far not produced such tentative findings.

## 1.2 Background

Provided the expanding use of technology and ICT tools throughout the society, the term technostress has been established as a thorn in the gut of the otherwise optimistic outlooks for the benefits of the advances in technological tools. Although this does not overshadow these benefits, neither is the thesis aimed at condemning the digitalization efforts and advantages. Moreover, according to the International Labor Office, “digitalization is one of the main drivers of technological change in the foreseeable future”, mentioning the ever more relevance of the topic due to the expansive nature of it (Walwei, 2016). Norway is among the top digitalized societies in the world, according to the International Digital Economy and Society Index (IDESI 2017). The score that reflects the factors such as internet usage, connectivity, integration of digital services, and digital public services are putting Norway in an above-average score of EU28, and 4th most digitalized in the world (Figure 1). This reflects the governments spending of considerable funds into digitalization, likewise the private sector investments, leading to a highly digitalized society. Norwegian policy aspirations with this spending are to be best in the world when it comes to digital solutions, and the government is increasing the funds of public and private ICT projects (Regjeringen, 2018). Hence, we can expect the dynamic trend to continue in line with the predictions of IDESI. The projects related to coordination, capacity, and flexibility that so far have been prognosed to have the potential of savings of 65 billion NOK between 2017 and 2025 in the public sector alone (Østvold & Rehbinder, 2017). In other words, digitalization is and will remain important, and there is considerable debate regarding digitalization policies and a wide consent on the necessity to

*Figure 1*



implement and facilitate more effective digital solutions and exploit that potential. An example of the case of Norway on the matter is the digitalization policy for 2019-2025 where the target is formulated as an all-grasping, coordinated ecosystem integrating the public services with the private sector (Astrup & Helgesen, 2019). Henceforth, the prospects of both the statistical data of digitalization, and the policy efforts have and remain pointing only in an expanding direction for frontrunning Norway in the international trend.

However, due to the requirement of personal cognitive, social, and intellectual skills, IT requires from the user, one toxic downside effect is perceived stress caused by digital tools. In the case of Norway, Torvatn et al. (2017) discovered that about 1/3 of the workforce is negatively affected by digitalization, including perceived technostress, about nine hundred thousand workers in total affected (p. 48). Additionally, international literature has pointed out that technostress has a relationship with burnouts that constitute up to one-tenth of all sickness-leaves in the workforce in Germany (Sicking, 2011), and perceived 77% had burnout at some point and 51% more than once in US (Deloitte, 2018). The majority of the burnouts are likely to be caused or exacerbated by technostress (Ayyagari, Grover, & Russell, 2011; Khedhaouria & Cucchi, 2019; Berg-Beckhoff, Nielsen, & Ladekjær Larsen, 2017).

Additionally, technostress hinders productivity, performance, commitment to the job, and wellbeing among affected (Tarafdar, Tu, & Ragu-Nathan, 2011). Not surprisingly, following the findings, the syndrome technostress was “recognized as an occupational disease” in 2007 (Chiappetta, 2017, p. 360). Regardless of these connections, it is important to state that digital tools delicately affect technostress, as it has both “positive and negative aspects in relation to burnout and stress” (Berg-Beckhoff, Nielsen, & Ladekjær Larsen, 2017). Hence, there is a monetary cost to digitalization, it is hard to pinpoint, yet burnouts do have a monetary price attached to degrading wellbeing of the individuals affected. An example is digital interruptions that are caused by high connectivity, Gupta et al. estimated that a 1000 employee company loses over 2 million USD due to workload and interruptions by additional emails a monthly (2006, pp. 948-949). Likewise, Maier et al. (2015) pointed out that provided most of the burnouts caused by digital tools, technostress costs the German economy billions of euros (p. 368).

Nevertheless, the best-known consequences of technostress are the psychosocial costs such as burnouts, dissatisfaction, reduced productivity and performance, and absenteeism (Patel, Ryoo, & Kettinger, 2012; Ragu-Nathan, Tarafdar, Ragu-Nathan, & Tu, 2008; Tarafdar, Pullins, & Ragu-Nathan, 2014). Both the monetary and psychosocial costs are the consequences of technostress that provide the underlying motivation to conduct the thesis aimed at the topic. Therefore, when conducting research about technostress, it is essential to retain a balanced view of the findings, provided the underlying empirical knowledge. This paper takes both sides of the spectrum into account when considering the underlying theoretical framework and empirical findings. In this way, the previous knowledge serves as a fertilizer for the hypotheses this thesis serves as the outcome.

The literature applicable to the thesis focuses on digitalization, the underlying occupational stress paradigms (epidemiological and cognitive appraisal), and subsequent technostress research. The chapter describes both the research behind technostress and the terminology used through the existing literature and technostress research. First, the discussion goes about the fundamental characteristics of digitalization that has spawned the phenomenon. Then occupational stress research paradigms and its relevance, a detailed walkthrough the technostress literature, and finally technostress terminology that is implemented in the thesis.

# 2 Literature Review

## 2.1 Digitalization

Digitalization is a term that covers a wide range of ICT developments, “Digitalization, meaning the growing use of information and communication technology”, that leads to effectivization of working tasks through the use of more advanced ICT equipment (Cijan, Jenič, Lamovšek, & Stemberger, 2019, p. 4). However, the term is defined differently in literature, depending on the use, including interchangeably with the term digital transformation (Mergel, Edelmann, & Haug, 2019). It can also be understood as the original process of changing from analogue to digital and reorganized into a format compatible with digital processing. Consequently, this facilitates the “communication between people, machines and workpieces” (Fleischmann, Oppl, Schmidt, & Stary, 2020, p. 10). Nonetheless, the proliferation of digital technology and its implementation throughout the society affecting every organization and almost every individual, where the usage of digital tools becomes inevitable for the majority. The way this takes place is under the usage of digital tools. Hence, in this thesis, digitalization is observed through the introduction and use of digital tools, such as software, automatic registration scanners, computers, smartphones, etc. (devices, hardware, and software) in accordance with Torvatn et al. (2017). The digital tools have the advantages that save time, resources, and create the opportunity for more interwoven, interconnected, solutions to handle and pass large amounts of information. The relevance of digitalization is in the ways it affects the perceived experience of technostress. Hence, the relevant parts of digitalization (tools in use) are the ones that have a relationship with technostress, a type of occupational stress spawned from the digitalization of the workplace.

## 2.2 Occupational Stress Paradigms

The existing literature on technostress builds upon existing research related to occupational stress at work. Notably, these are rooted in both the epidemiological and cognitive appraisal theoretical paradigms as the symptoms have similarities from wide stress-related literature and more specific occupational stress research (Fox, Dwyer, & Ganster, 1993). The epidemiological paradigm relates to the medical, objective observations of symptoms, this research established the basics of understanding of how stress is produced or the variables that matter, and symptoms of the stressed including the subsequent perceptions (Shultz, Wang, & Olson, 2010).

Based upon these findings, the deriving cognitive appraisal paradigm focuses on the conditional individual perceptions of occupational stress. The paradigm builds around that coping with stressful work conditions is neither static, as in an individual may find quite similar conditions stressful at different times, such as being worn out over time, a transactional, contextual, or process outlook (Lazarus, 1991, p. 3). Regardless of the acceptance of the epidemiological paradigms' generalizable assumptions, such as from the psychological effects of industrial technology and driving conclusions from medical records or psychology, Lazarus argues that there must be a balanced emphasis. This balance is context-specific while considering prior knowledge from medicine and contextual studies while stressing that "to overemphasize personal agendas is autism, and to overemphasize the environmental realities is to abandon one's personal identity" (p. 6). Likewise, the theoretical framework for technostress in this thesis is informed by the epidemiological paradigm, such as objective biological or visual observation of the symptoms (Muter, Furedy, Vincent, & Pelcowitz, 1993; Riedl, Kindermann, & Javor, 2012). These inform the cognitive appraisal paradigm, where the individual circumstances and perceptions are targeted using survey data to find correlations. In accordance with Ayyagari et al. (2011) comprehensive study, stress is "neither emerging from the individual nor the environment" but rather a "phenomenological process reflected in the relationship between the two" (p. 833). Thus, when conducting research on an occupational stress topic, it is necessary to have both paradigms in mind provided the synergic relationship. Technostress literature is conducted in a similar fashion, having experimental research, on the one hand, seeing the symptoms, and opinionated on the other.

## 2.3 Technostress

called technostress, which by the pioneer of the term was defined as "inability to cope with the new computer technologies in a healthy manner" (Brod, 1984) or "anxiety, avoidance, fear, stress, negative attitudes and cognitions" (Self & Aquilina, 2013). Another close relative of technostress in the literature is technophobia, which often also is a byproduct, precondition, or consequence of technostress, or simply used as a synonym (Sami & Pangannaiah, 2006). Hogan (2005) describes it as being "negative global attitudes about computers, their operation or their societal impact; and/or specific negative cognitions or self-critical internal dialogues during actual computer interaction or when contemplating future computer interaction" (p. 60). This is rooted in the consequences of using ICT provided the necessity of a degree of mental durability and cognitive skills involved when using the tools (Ragu-Nathan, Tarafdar, Ragu-Nathan, & Tu, 2008). For those unable to cope with the pressure stemming from the tools develop the aforementioned views or symptoms such as feeling the transgression into their private life (Moore, 2000), inability to cope or learn using the tools (Ayyagari, Grover, & Russell, 2011), and perceive an *information overload* (Tarafdar, Tu, & Ragu-Nathan, 2011). The causes of which leads to technostress through exhaustion (Ayyagari, Grover, & Russell, 2011), lower performance or satisfaction (Tarafdar, Tu, & Ragu-Nathan, 2011), absenteeism from work or intentions to quit as a whole (Laumer, Maier, Weitzel, & Eckhardt, 2012). These consequences of technostress are called strain that can be both psychological and behavioral. With psychological strain referring to "dissatisfaction with the job, depression, and negative self-evaluation" and behavioral strain to "reduced productivity, increased turnover and absenteeism, and poor task performance" (Tarafdar et al, 2011, p. 307).

The studies on the topic have been developed further throughout the years, documenting psychophysiological, mental, psychosocial, and or bodily reactions related to technostress (Arnetz, 1997, pp. 100-102). However, as in accordance with the information attained by SINTEF, their nuances to the use of digital tools that work both to the advantage and disadvantage in the form of user-friendliness (Torvatn et al. 2017). As in accordance to the epidemiological paradigm, Muter et al (1993) psychophysically observed the stress symptoms, establishing the clear relationship to digital tools as a source of stress. Likewise, Riedl et al. (2012) conducted an neurobiological technostress experiment, with cortisol-levels as dependent variable to single out acute non-conscious technostress. By which, the authors confirmed the health-related consequences of technostress and sources. Furthermore, the newer research has capitalized on the early benchmarks of literature and expanded on the academically well-established terminology to distinguish between unhealthy and acceptable types of technostress (Gaudioso, Turel, & Galimberti, 2017; Ragu-Nathan, Tarafdar, Ragu-Nathan, & Tu, 2008).

However, technostress in this research relates to the unhealthy types of stress coming from the digital tools, forced upon the workers as an integral part of their job leading to "unintended consequences of these ICTs that could be counterproductive" (Ayyagari, Grover, & Russell, 2011, p. 832). In other words, the ICT's both hardware and software intertwined in the workplace have the negative downside effect of inducing stress upon the participants. Technostress is produced under different circumstances; the technological tools have shown that because of their properties, they can produce stress in either different patterns than other stressors or exacerbate the existing ones (p. 833). This adds a new dimension of stress upon the workforce, as reflected in the disproportionately high count of burnouts or overall dissatisfaction due to technostress, as mentioned in both Norwegian and international literature. Nevertheless, this does not mean that the very same digital tools can serve as inhibitors, as this depends on factors such as experience, training, and personal background of the individual. Moreover, provided the nature of the data, the findings may also differ due to the cultural dimension, as Tu et al. (2005) state that there is a different pattern in Chinese workers' technostress in comparison to the United States. Meaning that the patterns and conditions for technostress vary according to the individual, organization, or culture. Meaning that context stretches outside of only the tasks, work-conditions, and other technical matters, as perceptions of technostress also differ culture-wise. Therefore, we can expect that to technostress is mitigated or averted depending on the aforementioned factors during the adaptation or use of the necessary ICT tools in the workplace. These factors can be distinguished as techno-stressors and techno-ameliorators that will be discussed in detail in the sub-chapters

## 2.4 Stressors and Ameliorators

The abovementioned paragraph discusses that digital tools can both enhance and decrease (or have little effect on) technostress. This idea is taken from wide literature focusing on the various factors that have either a positive or negative relationship with technostress. The categories are drawn up based upon the underlying models made by Ragu-Nathan et al. (2008) and Ayyagari et al. (2011), Shu, Tu, and Wang (2008), and Tarafdar et al. (2011;2014). The articles share many common definitions yet fulfill one another in several gaps that are necessary in exploring the data of this thesis. Ragu-Nathan et al. (2008) divides the factors into technostress creators and inhibitors, while Ayyagari et al. (2011) into *technology characteristics* and *stressors*. Regardless of their sub-divisions that were aimed at the models used in the respective research, this thesis utilizes them across the variables in the data due to the open-minded EDA approach. However, it is necessary to create a breakdown of the factors to comprehend the empirical relevance from the terms their relationship with the variables, and relevance to studying technostress.

These stressor categories are in accordance to Ayyagari et.al. (2011), framework, work-home conflict, invasion of privacy, work overload, role ambiguity, and job insecurity and enhanced by technostress. Furthermore, the earlier findings also points to *usability of features* that reduces technostress and *pace of change* that works bi-directionally (pp. 839-842). This framework is more related to general stress and strain and individual perceptions unlike epidemiological paradigm measures. While Ragu-Nathan et al. (2008) found similar characteristics directly relevant for technostress, as techno: overload, invasion, complexity, insecurity, and uncertainty. The distinction is the way Ayyagari et al. distinguishes between *technology characteristics* or features, however for the purpose of this thesis, they will only be distinguished as stressors, or ameliorators. Ayyagari et al. (2011) stressors include: *complexity*, *presenteeism*, *pace of change*, *work-home conflict*, *invasion of privacy*, *work overload*, *and role ambiguity*, and *job insecurity.*

*Table 1*

| Techno – Stressors | | Techno – Ameliorators | |
|---|---|---|---|
| Ayyagari et al. (2011) | Ragu-Nathan et al. (2008) | Other Authors | Ragu-Nathan et al. (2008) |
| Complexity | Techno-Complexity | Usefulness; End-User Satisfaction Tarafdar et al. (2011) | Literacy Faclitation |
| Presenteeism/ Invasion of Privacy/ Work-Home Conflict | Techno-Invasion Connectivity | Reliability (Ayyagari et al. 2011) | Support Provision |
| Pace of Change | Techno-Uncertainty | Involvement (Tarafdar 2011) | Involvement Facilitation |
| Work-Home Conflict | Invasion/Overload | | Job Satistfaction |
| Work Overload | Techno-Overload | | Continuance Commitment |
| Role Ambiguity | Techno-Insecurity | | Organizational Commitment |
| Job Insecurity | | | |

These factors translate into stressors positively or negatively and eventually lead to strain, which again is related to the negative outcome of technostress (2011, p. 839). While Ragu-Nathan divides his definitions into two, the aforementioned stressors, and the following ameliorators: *literacy facilitation*, *technical support provision*, *involvement facilitation*, *job satisfaction*, *organizational commitment*, *continuance commitment* (2008, pp. 426-427). These concepts are foundational for the thesis as the two papers are among the most cited literature on the topic, and the abovementioned features are of considerable relevance for the utilized dataset. In this section, the technostress creating features will be summed up with the similarities and differences between the authors. Some definitions are unique to the respective author; however, they are largely based on earlier occupational stress literature, hence many similarities exist, such as the Ayyagari et al. (2011), Ragu-Nathan (2008), and SINTEF study by Torvatn et al. (2017). Although one could guess the description, they describe “the adoption and use of technologies", the technostress literature often speaks of the lack of reliability, too much *complexity* as having a positive causal relationship with technostress (p. 836).

### 2.4.1 Stressors

#### 2.4.1.1 Complexity and Techno-Complexity

Complexity is related to the difficulty in coping with the technologies, or the task solved by the technology being complex in it-self. The reasoning behind two different labels, are the research-designs implied by Ragu-Nathan (2008) utilizing techno-complexity looking only at users with digital tools, and Ayyagari (2011) comparing both, hence, complexity as a term. The discussion is about the occupational or task related stress symptoms regardless. Wang, Shu, and Tu (2008) states that the stress of complexity is related to a feeling of incompetence (p. 3004), in addition to what feeling of an extra work effort (Tarafdar et al., 2011, 2014; Ragu-Nathan, 2008), and knowledge barrier effort (Ayyagari et al. 2011). The complexity also relates to the general dissatisfaction of tools being unnecessary complex to perform a task that otherwise could have been solved with more user-friendly tools. Ragu-Nathan et al. (2008), has an identical characteristic called techno-complexity, likewise describing the stressful mismatch between the user skill and the technology in question.

#### 2.4.1.2 Presenteeism, Invasion of Privacy, Work-Home Conflict, Techno-Invasion

Ayyagari et al. breaks up the concept into a technology characteristic and intrusive feature called presenteeism where the technology creates a loophole and cycle of communication and interruptions with the end-user being unable to disconnect (2011, p. 840). The three terms are closely related, as they speak of the feature of being on GSM, wi-fi, or otherwise connected to other people, automatic notifications, alarms, etc., through digital tools beyond work. This connectivity, fragmentation of the working process and off work time is instead a cause of *invasion of privacy, work-home conflict,* or what Ragu-Nathan et al. (2008) calls *techno-invasion*. Tarafdar et al. (2011) states that it appears as a "blurring between work-related and personal contexts" (p. 310). Additionally, according to Ayyagari et al. (2011), summarizes the consequences as the end-user with high connectivity receives interruptions of tasks for example through sporadic e-mails and feels spent (p. 841). Provided a situation where the user is unable to finish the preceding task and consequently experiences strain, or "information fatigue" (Ragu-Nathan et al. 2008, p. 421). Again, this is possible without digital tools, however, the infrastructure of internet and other communication devises and possibilities of programming greatly facilitate and/or enhance these hazards.

#### 2.4.1.3 Work-Overload, Techno-Overload

The most important of the two terms is techno-overload, as it is the specific type of work-overload related to digital tools and their technical features. Such can be the necessity of multitasking, simultaneously conducting and processing several information sources and objectives, either increasing the workload, making it more complicated, or/and more intense (Tarafdar, Tu, & Ragu-Nathan, 2011, p. 312). The term work-overload relates to expectations from peers or superiors of effectiveness, productivity, and performance (Ayyagari et al. 2011, p. 841), while techno-overload is related to digital tools. In similarity to the other techno-stressors, these features have more precise definition than the general stressor of work overload.

#### 2.4.1.4 Pace of Change, Techno-Uncertainty, Job-Uncertainty, Role-Ambiguity, Job-Insecurity

In Ayyagari et al (2011) magazine of technology characteristics is the *dynamic characteristic,* pace of change, that describes the swiftness of introduction of new technology which Ragu-Nathan et al. (2008) calls techno-uncertainty among his technostress creators. Both these terms refer to new developments, changes, and upgrades in the digital technology at the workplace and refer to enhancement of technostress. Hence, this section discusses the terms relating to technostress with the common feature of discontent or unease about change. In Ayyagari et al. (2011) pace of change is linked to the *stressors*: job insecurity, role ambiguity and work load as these changes cause more work, feeling of ambiguity or threaten the job as whole when the end-user feels underperforming (pp. 839-841). Note that, Ragu-Nathan et al. (2008) considers techno-uncertainty as a stress factor, while Ayyagari et al. (2011) as a technology characteristic that causes the ensuing stressors that eventually lead to strain. Job-

Insecurity, is related to being replaced, either by technologies capable of automating the task one performs, or not being able to catch up with technology in order to perform in the workplace, while uncertainty is the unpredictability about new updates (Tarafdar et al., 2011, p. 310). The relationship between job-uncertainty and insecurity, and pace-of change, is that the pace can result in both, or one of the features, as technological developments can both outdate a type of work, or the worker may not cope with the development. The last is ambiguity, which is closer kin to uncertainty, as it makes the role of the worker less relevant, such as developing the task of the worker from a task performer to error controller of an automatic process (Ragu-Nathan et al., 2008).

### 2.4.2 Ameliorators

#### 2.4.2.1 Techno-Usefulness, Techno-Literacy Facilitation, End-User Satisfaction

Techno-usefulness and literacy facilitation, and end-user satisfaction are a part of the same construct, where belief of the ICT usage is being advantageous, or training and support that allows the use to master the otherwise complicated and confusing tools. Ayyagari et al (2011) describes them as usability characteristics (p. 839). Ayyagari et al. mapped out usefulness and reliability among the features that ameliorate the stressor called work overload (p. 839). However, Ragu-Nathan et al. (2008) integrated technical support provision, literacy facilitation and involvement as *technostress inhibitors.* These are described as “organizational mechanisms and adjustments through which negative outcomes from ICT use can be alleviated” (p. 422). Literacy facilitation is a close follow-up training, where user experience has shown that a high training standard reduced stress and anxiety (p. 427). If the literacy is facilitated, usefulness perceived, then end-user satisfaction is subsequently achieved, the result is less perceived technostress, and “productivity and innovation” (Tarafdar et al, 2011, p. 306). Hence, the satisfaction is the outcome of usefulness, and literacy facilitation, while the performance is the bi-effect of end-user satisfaction along with lessening the likelihood of technostress.

#### 2.4.2.2 Implementation Involvement, Involvement Facilitation, Support Provision

Implementation involvement, and facilitation is about the employees having a say in what types of tools are to be introduced into the workplace, and implementation facilitation is referring to "keeping users informed about the rationale for introducing new ICTs" (Ragu-Nathan et al., 2008, p. 427). These are similar, and therefore can be used synonymously, as the main point is allowing the users to comprehend the intentions of implementation and therefore also understand what features to get with them. It is only chronologically ameliorating step ahead of the natural support provision that allows the users to exchange experiences, and experts to guide proper use instead of needlessly bottle-necking work-processes and spurring negative perceptions (Tarafdar et al., 2011, p. 304). Hence, it's a design of amelioration through allowing participatory activities during the implementation process, informing and testing the rationale behind implementation, with additional support provided during later use. This is of course situationally relevant, however, the literature describes these as necessary and typical ameliorating activities.

#### 2.4.2.3 Job-Satisfaction, Continuance Commitment, Organizational Commitment

These last terms are only mentioned in Ragu-Nathan et al. (2008) as attitudes or work-ethics of the individuals, and list them as generally important factors to prevent technostress. However, these also may be lowered in case of higher technostress, as many of the factors are related to ethical task management and demands for the employees. Tarafdar et al. (2011) states that increase in stress and strain stemming from "increased perceived work demands, and reduced job control" are relevant factors to lack of continuance and organizational commitment, and job-satisfaction (pp. 307-308). Hence, the terms continuance and organizational commitment may both be a predeceasing attitude, like loyalty, curbing technostress, or their decrease is a result poor utilization of the other ameliorators.

# 3 Methodology

## 3.1 Epistemology and Ontology

The thesis does take a quantitative method in analyzing large data. However, when it comes to epistemological and ontological considerations, giving the whole thesis its meaning, its important to keep in mind the aim of the project its crucial to take a balanced stance. Hence, the thesis takes a critical realist view to as it is necessary when conducting an exploratory analysis where the goal is to create hypotheses with a wide utility for future research. Mingers (2002) stated that critical realism "can be very useful in the exploratory stage in detecting particular patterns within the data […] [t]he results, though, will merely be the starting point for more substantive investigations" (p. 301). Referring back to the aim of the thesis, the point is to explore the data for the hypotheses generation. As epistemologically the theoretical understandings have its roots in the human cognitive process of the earlier studies, and it is an abductive process guided by the conceptual framework that exists on the topic. Hence, the underlying epistemology does not reject the possibility of interpretation of the context-sensitive reality through empirical observations. This is a critical realist assumption of the epistemic fallacy, or the decoupling of ontology and epistemology. For this thesis, the ontological consideration does not rest with a socially constructed reality, there is one reality regardless of our social underpinnings, but our understanding depends on our filters, and presents them tentatively. Hence, this thesis takes a critical realist stance on methodology where the data analysis serves as a vessel for future confirmatory studies and not to formulate a conclusive study.

The reasoning behind this is the aim of the thesis, the same for why the method takes an investigating function, where it is important to be informed by theory and empirical evidence yet remain skeptical and flexible (or critical). Although keeping the philosophical core of realist ontology, critical realism shares the perception that reality exists regardless of our interpretation or knowledge (Archer, 1998). However, this reality is only attainable through our limited ability of sensing it. Hence, the only reality we can describe and learn about is what we know, from our senses, theory, and personal opinions, and it always has a potential for at least some bias, or "choices made on subjective grounds such as experience [and] usefulness" (Mingers, 2002, p. 301). Likewise, the critical realist ontology has a focus on reflexivity, a necessity to guide and orient empirical investigations, while being transparent about assumptions that sprang from the earlier literature. In the work about ontology and critical realism, Bhaskar (1998) stated that its essential to sum the “assumptions fully and explicitly at the beginning of some piece of work so as to put the reader (and possibly also the writer) on their guard” (p. 62). This is due to the likelihood the influence from the assumptions when conducting the research, and necessity for awareness of that possibility. As it has the potential to limit the findings that otherwise open-minded or less biased researcher would discover. Although the research is conducted abductively and theories and concepts are orientational (see literature review), it is a necessity to remain openminded and reflexive throughout the research to effectively create suggestive hypotheses.

Furthermore, critical realism, epistemologically, acknowledges a limit to our understanding of reality: “even though there is one reality it does not follow that we, as researchers, have immediate access to it or that we are able to observe and realize its every aspect” (Zachariadis, Scott, & Barrett, 2013, p. 857). This idea follows the logic of that ontology is guiding the epistemology of critical realism, or what is otherwise called epistemological relativism (Lawson, 2011, p. 162). Or, that the way the real world is, determines the way and to the extent to which we can know, interpret, and learn about it. Hence, the epistemological considerations are secondary to ontological in critical realism, which is crucial for the researcher when exploring the data as that is where the least biased hypotheses may be found about the phenomenon.

However, this thesis follows the positivist idea of applying laws of natural science through a statistical analysis towards the social phenomenon of technostress. As mentioned earlier, this is due to our inability to know every aspect of the phenomenon and therefore only create tentative findings, patterns, and hypotheses. Because it only serves as tentative suggestion, a hypothesis in the thesis created for a future study, is not a confirmation of a presupposed ready-to-use theory. Likewise, Eastwood, Jalaludin and Kemp (2014) states that “if patterns exist within a set of observations then there must be some underlying structures, mechanisms, or constraints that may prove to be a useful starting point for critical realist investigation” (p. 5). Furthermore, critical realism also rests on the idea of epistemic fallacy, separating the epistemic and ontological analysis. This is likewise reflected in the method which does not aim to establish a single reality. Regular natural science and therefore also orthodox objectivists would emphasize the method as being synergic with the epistemology and therefore have ontology as secondary. The method in this research is flexible and does not follow a formal set of rules, it has any possibility open. These possibilities go in hand with the critical realist ontological dimensions of multiple realities: the empirical, actual, and the real

(Sayer, 2000, p.13). Additionally, reality is stratified consisting of "hierarchically ordered levels where a lower level creates the conditions for a higher level" and "[e]ach stratum is separate and distinct and may interact with the layer above or below to produce new mechanisms, objects and events" (Eastwood, Jalaludin, & Kemp, 2014, para. 9). The term is collectively called emergence, an interplay of facts existing on a different level than observable through the filters of the process. Hence, the ontological considerations acknowledge the unlikelihood of knowing everything, or creating perfect predictions based on a set of data about a real-world phenomenon. With this consideration, exploratory purpose of the thesis is thus, an investigation of the data with the state of mind of not being over-optimistic, but inventively create tentative hypotheses.

Moreover, the issue at hand is related to a social phenomenon that has been proven to behave differently or at least being understood differently in various places. Or as Bhaskar stated that, a phenomenon can be unique, recurring, or context dependent, and likewise, the experience of them can be different based on involved individual perception (Archer, Bhaskar, Collier, Lawson, & Norrie, 1998, p. 14)

When it comes to the assumptions of critical realism, reality has three dimensions that are tangible, fagmentable. Social facts have an objective reality. However, we know that the technostress phenomenon is a social phenomenon that behaves, or is interpreted differently throughout organizations, cultures, branches. Including with conflicting evidence, such as the gender dimension that has shown different results in different studies, potentially based upon a different paradigm application in confirmatory studies. There is a necessity for flexibility and understanding of a balance between the two when researching such a dynamic social phenomenon. To understand it, requires open mindedness and consciousness about the limitations of a researcher viewing the work from a critical realists lens. In the context of EDA especially, as is summarized in Table 2, data has primacy ahead of other aspects mentioned in the sub-chapter. Notice that many of the points are explicit, they are to be treated as general, as an EDA process is not to be restricted by a pre-decided framework. Yet, it is a summary of the underlying considerations taken throughout the thesis

*Table 2*

| Mode of Inquiry |
|---|
| *Assumptions*<br>- **Reality (three realities)**<br>- **Knower and known are independent**<br>The knower and the known are as in positivist assumptions independent. Yet, the knower can only interpret according to own ability<br>- **Primacy of data**<br>Flexibly explore data to create suggestive findings, patterns<br>Interpretation of the data<br>- **Variables can be identified, and relationships measured**<br>To the extent possible, due to the quantitative nature of the thesis<br>- **The inquiry is objective, value-free**<br>Note that critical realism rejects that inquiry can be completely value-free, hence, some bias will exist due to the researcher's human agency |
| *Purposes*<br>- Contextualization, understanding and visualizing the patterns of the data<br>- Suggestive causality<br>- Facilitation of further confirmatory analysis |
| *Approach*<br>- Begins with former research and theory for orientational purposes. Ends with hypotheses<br>- Abductive<br>- Search for patterns, correlations, covariate associations with the response variable<br>- Emergence and portrayal<br>- Treating the data<br>- Descriptive write-up |
| *Researchers Role*<br>- Detachment and impartiality<br>- Objective (as possible) portrayal<br>- Etic (Outsiders point of view) |

*This is based on Yilmaz (2013)*

## 3.2 Method

Although the aforementioned literature review clearly provides expectations about the data, there are always new patterns to look for, the core task of exploratory data analysis (EDA). This is conducted by presenting the data through suitable visualizations of different combinations of variables on the data (Sailem, Sero, & Bakal, 2015). Or as the name of the method states, explore the data to make more precise application of the data at hand for tentative hypotheses and further research. Although the theories and earlier empirical work orients the researcher, when using EDA it is important to conduct more of an investigation and remain open to new findings and critical to the earlier ones. This, however, does not mean it is an absolute free-for-all where everything is presented from the lengthy process, Behrens et al. (2012) states that "we abduct only those that are more plausible for subsequent confirmatory experimentation" (p. 39). The chosen procedure, therefore, includes features such as sampling, and cross-validation, to ensure that the hypotheses generated are likely to remain fruitful for later CDA (confirmatory data analysis) use. Hence, the process follows the steps of graphical presentations of the overall data, then moves to machine learning methods, including sampling methods to generate the hypotheses.

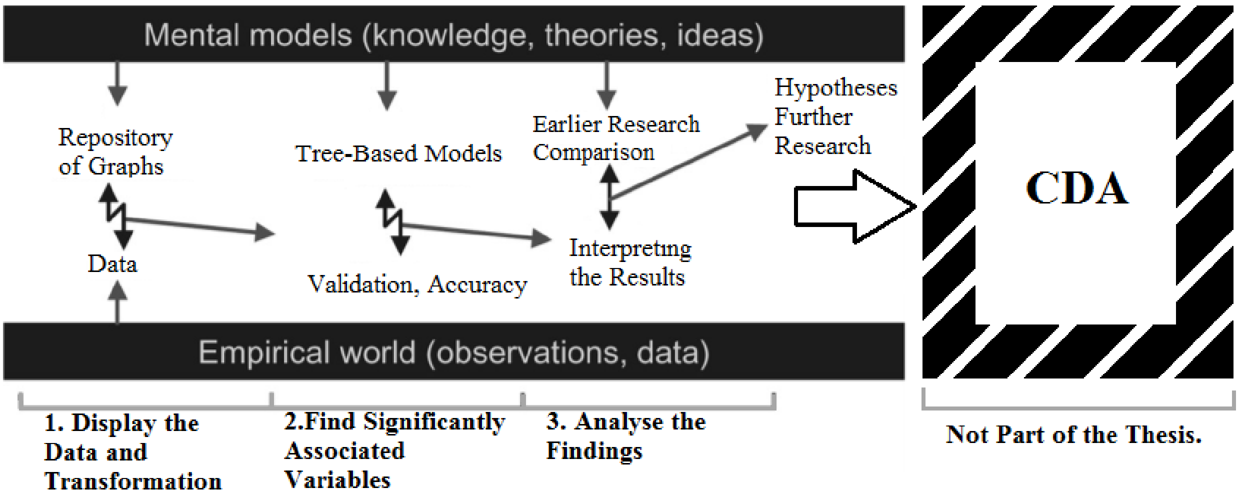
The procedure of EDA goes in hand with visual data exploratory forms to inform both the audience and guide the researcher's ability to observe patterns through visual representation, what Hong and O'Neil (1992) calls a presented mental model. The definition of this model is "a representation formed by a person, which is based on previous experience and knowledge, as well as on current observation and learning" (p. 150). However, they are not conceptual models built of existing components to test existing hypotheses, when utilized in an EDA setting, the practice is more goal-oriented and inventive with hints of earlier theory. Hence, theories do inform and inspire, yet are to be treated with a portion of skepticism to achieve the maximum utility from the method. As Behrens (1997) states, "the researcher entertains numerous hypotheses, looks for patterns, and suggests hypotheses based on the data, with or without theoretical grounding" (p. 133). EDA method follows optional, or customized procedures, often creating graphical visualization of the data to explore, analyze, and discover its secrets. Therefore, EDA goes in hand with what Shneiderman (2003) calls "Visual Information-Seeking Mantra" which consists of an "Overview […] Zoom […], Filter […] Details-on-Demand […], Relate […], History […], Extract" procedure (p. 365). This procedure inspires the steps taken in this thesis, where the descriptive statistics present an overview and a slight zoom that slightly filters the information and the details. Then tree-based methods provide a detailed visualization with details about the relationships, and finally through sampling, extracting the core message. At the same time, relation and history are generally spread out based on how detailed the discovery is.

The target of EDA is "discovery, its goal is to maximize the use of the data; it is not limited to single sets of hypotheses, prior research, or the personal limits of the researcher" (Jebb, Parringon, & Woo, 2017, p. 271). It is a process of utilizing the value of the data at hand for future, multiple confirmatory kinds of research in a synergic fashion, where the EDA serves as a broad fundament upon which CDA may structure according to its academic purpose. To achieve this goal, provided the utilization of machine-learning methods, where EDA also incorporates some validation, as the method "aims to yield predication rather than theoretical explanations of the relationships between variables" of CDA (Ho, 2010, p.18).

Gromelund and Wickham (2016) add that "EDA is not a formal process with a strict set of rules. More than anything, EDA is a state of mind", with the intention of presenting the worthy findings in the data at hand (par. 2). The concept of the method is that it urges the researcher to examine the data as it is without remaining sanguinely orthodox about the spawned findings of the predetermined theoretical assumptions. As an abductive method, not a classical quantitative deductive null-hypothesis rejection test. In a setting of critical realism and the acknowledgment of the existence of individual subjectivity formed by the concepts that inform the researcher, the thesis takes the openminded and utile EDA model.

Inspired by the EDA model by De Mast and Kemper (2009) the thesis, takes the follows the logic of displaying, identifying, and interpreting. This option allows for optimal flexibility and the absence of strict prohibition of deviating from a predetermined path. It instead focuses on telling a story about the detected patterns found in the data, how they can be understood, and, therefore, how they serve as a fundament for successive CDA. As a part of the process, it "encourages the development of mental models" (Behrens, Dicerbo, Yel, & Levy, 2012, p. 35).

*Figure 2*




Based on De Mast & Kemper, (2009, p. 369)

The display is the presentation of the data, descriptive statistics, and summaries, optimally, they are easy to interpret, avoiding obfuscations of needless complexity. Although simple visual models are preferable, as opposed to long texts, some information cannot be shared in such form, while keeping it comprehendible for the human mind. Schneidermann

(2003) states that a “page of information is easy to explore, but when the information becomes the size of a book […] or even larger, it may be difficult to locate known items” (p. 365). Hence, the non-graphical displays will be limited to up to a page of size if possible.

However, in this thesis, a former report utilizing the data is known, with no other data yet available, whilst implementing machine-learning in a data-mining context using resampling. This process demands a model, wherein the combination with EDA has the capacity for “suggesting and validating a model at the same time” (Ho, 2010, p. 18). The method helps to understand the data, the relationships between the variables, and, therefore, also observe the problematic aspects of the data in combination with future methods. Therefore, doing this preparatory work that facilitates more accurate and robust CDA as a follow-up, while presenting a detailed map of the findings of the data with tentative hypotheses. The three steps Ho (2010) describes in a data mining context are “detecting clusters, screening variables, and unearthing hidden relationships” (p. 14).  Despite EDA appearing as the opposite of classical deductive, hypothesis testing methods, it does not compete with confirmatory research, as “the modes are complementary rather than antagonistic” (Behrens, 1997, p. 132). In similarity to a criminal case, the EDA is the detective work searching and considering all the evidence, while CDA is the court trial judging for and against and delivering the verdict. It is a necessity during a research process, a preliminary step that facilitates further research.

### 3.2.1 Tree-Based Methods

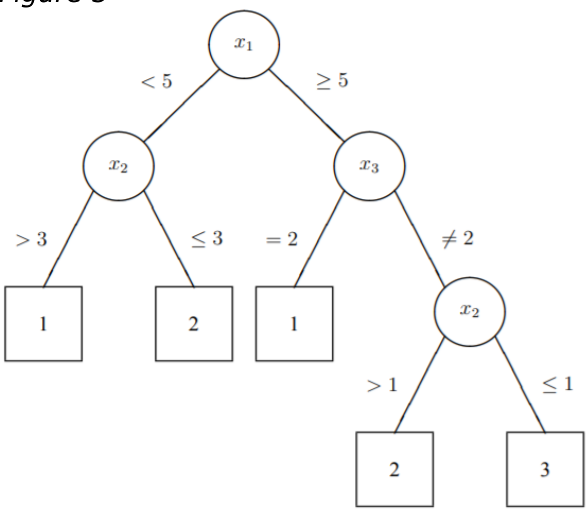
Provided the quite large dataset at hand, it is important to visualize and sort the information at hand for the audience of the EDA. Regardless of EDA being described as a state of mind in the literature, which it is, it is necessary to tailor it in the best way possible according to the aim of the research. In this thesis, the aim is to find patterns of digitalization that potentially lead to technostress in the Norwegian workforce. By utilizing a combination of descriptive statistics, accompanied by Tree-Based methods, or conditional inference trees and random forests in Rstudio.

Tree-Based Methods, especially decision trees, are visually interpretable, like flow-charts or inverted tree-shaped algorithms that allow to decide or show a result of each of its nodes. Shortly, it is a method where we ask a binary question about the data using the predictor variables and test whether there is a significant association among the covariates and the outcome variable (Hothorn, Hornik, & Zeileis, 2006). The reasoning behind the use of Tree-Based Models are the numerous advantages, such as excellent interpretability, ability to handle qualitative variables, and appropriately handle big volumes of data while tracking the covariation among the predictors on the response. According to Ho (2010) decision-trees are specifically "robust against outliers, because the data set is partitioned into many nodes during the exploratory process, and as a result, the effect of outliers is confined into their own nodes" (p. 15). Furthermore, another advantage is confirming the multicollinearity in the data, that at worst may risk of being overseen in a multivariate regression analysis that is sensitive to outliers, producing biased estimates (p. 11). Likewise, as a follow-up to the regular tree's, the thesis also employs random forest, which thanks to its ability to rank the variable importance, avoids the instability of logistic regression "which are known to be affected by order effects" (Strobl, Malley, & Tutz, 2009, p. 324). Thereby, by utilizing the tree-based models, we enhance the observations of the digitalization and technostress related variable combinations that have significant associations.

#### 3.2.1.1 Decision Trees

The anatomy of a decision tree has a starting point or root node at the top, from where the divisions start forming into new branches representing an outcome of the intended test for the data, such as yes or no, or a range, ending in the terminal-nodes (Sharma & Kumar, 2016, p. 2094). Alternatively, the partitioning of the nodes can be based on numerical data such as above a certain value or below, or a value different than a selected value (Figure 2). As seen in Figure 2, the algorithm is conducted through a recursive binary splitting into two new branches, concluding in a leaf-node when the best possible split is found based on the most relevant variable and no improvement is possible. This improvement is set by a minimum significance level (α), obtained through cross-validation to find the lowest possible classification error rate, with the error being the lowest when the total variance is lowest among the classes (James, Witten, Tibshirami, & Hastie, 2013, p. 316). This is also called "impurity reduction", where "each split in the tree-building process results in daughter nodes that are more pure than the parents node" (Strobl, Malley, & Tutz,

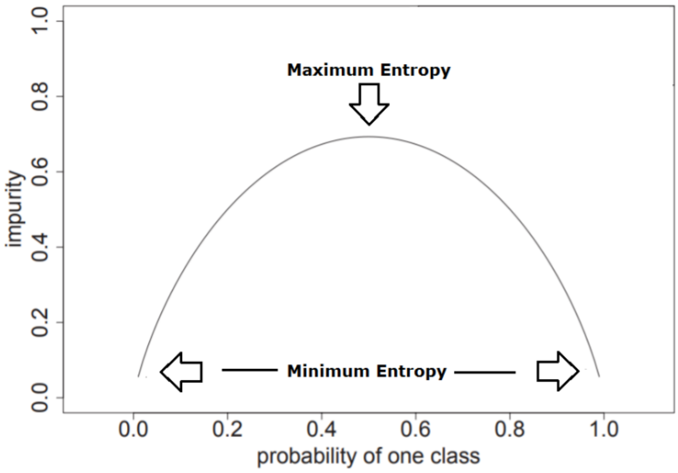
*Figure 3*




(Breiman et al. 1972, p. 256)

2009, p. 326). Hence, resulting in easily interpretable homogenous leaves with logical splitting and ordering of the variables.

*Figure 4*




The decision tree allows for good visualization of the data and the correlations with the response variable. In this case, we will determine which variable(s) are the most important for the respondents in relationship to the outcome variable. In this case, as mentioned, either to experience technostress or not. The observations form the descriptive data produced some initial suggestions. However, the decision tree method is applied here to gain a lot more information utilizing as opposed to the descriptive plots. The information we gain is related to the entropy of the partitioning of the attributes into the most homogenous subsets based upon their properties, to the point where a further division of the tree does not gain additional information (Wong, 2017). That information is gained by an attribute that has the highest likelihood of a relationship with the outcome variable. An example is 100% or 0% percent to correlate with the outcome variable, therefore being ‘pure’ in both cases, as opposed to 50%, which makes it irrelevant, as the intent is to understand whether it is positively or negatively relevant for the outcome. The ‘pure’ attribute is when the entropy is equal to zero, which happens at either a 100%, or 0% impurity, that being all the respondents in the attribute unanimously produce the either/or answer. As seen in figure 4 the minimum entropy is on the extremes left and right, while the maximum is on the top, where the probability of classification is a 50% chance. In this thesis, the ‘rpart’, and ‘party’ - ‘ctree’ version of Rstudio is utilized to generate decision trees. With the ‘ctree’ or what is called conditional inference tree, the version we also get the p-values as you can see in the root, and leaf nodes, represents the association with the response variable, having the highest entropy (Strobl et al., 2009, p. 327). Additionally, we get the count in each terminal node along with the proportions or percentage-wise distributions from the overall sample.

#### 3.2.1.2 Random Forest

In addition to the trees, the mentioned ensemble method of random forest combining hundreds of trees with a random sampling of variables for each split is utilized to observe the importance of the variables that a single tree cannot. Although the process is less suitable for visualization methods dictated by EDA, it provides insightful and more accurate observations by using a limited set of random variables tested for each split during training and offers variable of importance. The necessity of variable of importance in the context of EDA, provides additional insight for the global overview of all the covariations among the predictor variables and the response, which the tree overlooks (Strobl et al., 2009). Random forests are actually ctree's, just many of them with random samples tried at every cut point or split. Which is the main difference between classification trees and random forest are the advantages of smoothing out decision boundaries (Strobl et al., 2009, pp. 331-333). This method randomly samples a subset from the data frame, automatically picks a set of predictors for each split in each tree, resulting in very diverse trees, delivering scores of results. In the randomForest package of Rstudio this is reflected by the 'mtry' function, where for example, 16 different variables instead of all are tested for each split. A process followed by this production of numerous trees, the end result is utilized to classify the most common predictions in a majority vote, where the majority represents the final classifications (Wang, Zhang, & Yu, 2020, p. 143). To find the optimal number of trees and randomly selected variables to contest each split is called tuning, where the predictive accuracy is measured, providing the best tuning parameter of trees and 'mtry'.

### 3.2.2 Applied Method

Because of the magnitude of variables, the tree-based models allow to visualize a large number of combinations from which it is feasible to create a multitude of hypotheses regarding their relationships, the principal goal of EDA. Additionally, the approach is abductive, as EDA requires, the tree-based models allow the researcher to deploy a **"**data-driven approach" (p. 325). The observation in the data has priority, not earlier theories. In contrast, a lone CDA, without prior EDA, utilizing a linear model would fail to see these covariations, and other similar models have a poorer visual representation and requires prespecified hypotheses and a coactive model specification (James, Witten, Tibshirami, &

Hastie, 2013, pp. 320-324). Thus, facilitating the design of a future linear model based CDA, or a non-linear model strengthened with the presence of formerly conducted EDA on the topic. This is essential provided the data at hand with over 90 variables, some of which can only be treated as factors to produce meaningful results. Likewise, the nature of the research topic, being a under-researched and an overall fresh phenomenon of the digital age (Groes, 2017, p. 1482). It is therefore a thankful process to prepare for wider research of the phenomenon, while employing methods that accommodate the data in an effective fashion.

Provided the already existing report by SINTEF about the relationship between digitalization and stress, technostress, advantages, and potential solutions to the issue; the method include improving the accuracy and generalizability of the covariate associations with the outcome. A crucial difference between tree-based models and linear models, are the possibilities of observing the multiple associations without the need of having the for specifying the association rules. As Strobl et al (2009) points out, that includes “nonlinear and even nonmonotone association rules, which do not need to be specified in advance” (p. 325). Regardless of which, the results must be treated as tentative, as the data is based on one set of interviews and not repeated several times, prohibiting the opportunity of truly confirmatory experimental results. Torvatn et al. (2017) mentioned that the topics the report addressed require more research and eventually more data to truly understand the nature of digitalization and its challenges in Norway. Therefore, to offer the best opportunity to facilitate this without any more available data, EDA in combination with the best tools for handling large amounts of variables. Hence, tree-based models are opted for in this thesis, including the utilization of resampling methods such as 10-fold cross-validation with bootstrapping.

Although, some researchers deny the necessity in using any statistical inference methods to control for the accuracy, generalizability when conducting EDA, and warning researchers from slipping into CDA when doing so. The method in this thesis does indeed utilize these methods, as the data available is too thin to be considered confirmatory to establish empirical and robust relationships. Henceforth, the accuracy and generalizability of the important covariates is to be exposed in the EDA analysis and is not to be treated as a final correlation. Likewise, when utilizing tree-based models, that is too slight changes in the factor variables, it is a necessity to cross-validate in order not to provide misguided or inaccurate suggestions for the subsequent CDA. Especially provided the discernible cross-validation method, and bagging available to through Rstudio in the party, rpart, and randomForest packages. Hence, the technique is somewhat similar to Acquuah-Mensah, Leach, and Gida (2006), Steger, Brenning, Bell, Petschko, and Glade (2016), and Ho (2010). These authors combined the use of machine-learning techniques, including cross-validations with EDA to present tentative findings. Ho, brands it the “new EDA” stating the following:

> “Traditional EDA techniques might pass the initial findings (suggested factors or hypotheses) to CDA for further inquiry. However, with the use of resampling, new EDA can go beyond the initial sample to validate the finding.” (Ho, 2010, p.12)

The logic of Ho’s EDA approach is based upon the purposes of preparing the ground for a solid CDA (Ho, 1994). Hence, there are clear advantages in the context of data-mining methods to ensure some general accuracy and validation, instead of exclusively present unprecise models that will fall-through in the initial CDA or misguide future induction or abduction. This has two reasons, some of the models, such as the decisions trees, are cost-sensitive with the rare classes, and may have a positive-class bias that may wrongly include variables in the tree (Zhang, et al., 2017, p. 34; Tian & Zhang, 2019). And second, since “the researcher still has to construct a theoretical model in the context of CDA”, it is only advantageous to ensure accurate estimates beforehand (Ho, 2010, p. 18). Hence, in the thesis, the patterns found in the transformed data will have validation following the initial trees, as that can help ensure accurate CDA in the aftermath.

However, there are concerns when using tree-based models. One of the main issues about using tree-based models is the risk of producing inaccurate predictions, overfitting, high sensitivity, and bias. First, in accordance to Strobl et al. (2009), for the regular simple trees, the most dangerous issue is the instability when only the training set or the entire data available. This method is extremely sensitive and produces high variability even to slight changes, and therefore lacks generalizability. Additionally, “variables with many categories and numeric variables or, even more unintuitively, variables with many missing values are artificially preferred” (p. 342). This issue is handled by imputations, in accordance with the explanation in the Data Preparation sub-chapter replacing the missing values. Overfitting is addressed by the random forest as it produces a vast amount of trees with only a set of potentially best predictors and 10-fold cross-validation. Likewise, by sampling, the “Out-of-bag data (OOB) can obtain an unbiased estimate of the true error” and curb the bias for random forests (Wang, Zhang, & Yu, 2020, p. 143). In addition, the bias problem for the trees is solved by using 10-fold cross-validation with bootstrapping to assess the sampling variation and learn the accuracy of the model. Hence, to cope with the downsides of the tree-based models, cross-validations, bootstrapping, and tuning for random forests will be implemented in this thesis. Henceforth, avoiding overfitting, high-variability, bias, and lower the classification error while creating the best possible hypotheses from the data.

# 4 Data

The datasets used in this thesis were supplied by Hans Torvatn et al, (2017) from SINTEF ahead of being made public, the survey was conducted through telephone interviews by the agency called Respons Analyse AS with 50 questions and additional background variables with 2393 respondents. A total of 96 variables are made in the dataset in total based upon the questionnaire (Appendix 3). The sample is aimed at a representation of the national distribution of the entire workforce based on the background variables such as gender, age, educational level, years of work-experience, salary-level, and branch table 3.

*Table 3*

| | |
|---|---|
| Background Variables | • Gender, Education, Age, Income<br>• Role: Leader, Union Representative, Health and Safety Representative<br>• Branch (List of 18) |
| Work-Tasks with Technology | • Type: Core, Rapport, Coordination, other |
| Techno - Ameliorators | • Functioning Technology<br>• Implementation Involvement<br>• Collaboration (Support-provision)<br>• Productivity (User-Satisfaction/Reliability)<br>• Training (Literacy Facilitation)<br>• Leader support (Support-Provision)<br>• Experienced Flexibility (Job-Satisfaction, related to positive availability) |
| Techno - Stressors | • Use of technology<br>• Technology implementation<br>• Demanded Availability<br>• Registration of Work |

Then, tasks or positions are questioned, followed by techno-ameliorators and stressors, accompanied by direct questions about general stress and technostress. The background questions are asked due to the intentions of imitating the Norwegian workforce while catching their perceptions about digital tools and stress. The questions are formulated to answer the background categories, followed by a Likert-scale from 1 to 5, or 1 to 6 (see Appendix 3). The limitation, as noted by the authors, is that in some types of work-branches had considerable difficulty in obtaining enough respondents, as the survey aimed at collecting a sample that replicates the shares of the branches proportionally. Examples of this limitation are sales and automotive-workshops, health and social work that are slightly under-represented in comparison to the overall workforce share (Torvatn et al., 2017, p. 9). Likewise, technical support jobs and logistics are below 80% and 60%, respectively, of the intended survey population. Additionally, the lower paid and educated employees are underrepresented, and therefore the more educated and higher paid being overrepresented in comparison to the distribution in the total workforce (p. 10). Torvatn et al. (2017), stated that one of the main problems is that some categories of jobs are generally of a character that makes them unavailable for a phone interview during the working hours.

Furthermore, the survey was apart from that oriented towards mapping out the technologies in use per employee and newly introduced digital tools, thereby facilitating the observation of the extent of digitalization and the connection between digitalization and technostress. Furthermore, the survey addressed the type of work and whether digital technology is a part of their performance of their core or regular tasks and what type(s). Therefore, enabling the observation of the most digitalized workers who use digital tools a lot, and/or a variety of tools.

Many of the questions are in line with the mainstream technostress literature and aim at the earlier observed and researched the technostress phenomenon. The examples are especially such as the workload, time-pressure, availability, work-demand related questions that are attempted linked to digital tools. Provided the large sample size, the dataset serves a good fundament for data-exploration and hypothesis generation. Therefore it is possible to categorize the variables in the data in accordance with the Literature Review. Table 4 below summarizes the topics of the questions and the variable name. Example, S_12 series are represented by variables, S_12a, S_12b, and S_12c. They ask about the types of tasks performed by the respondent distributed by percentages. The background variables are just questions about identity, such as gender, which is translated below, UTD representing education, etc. More details about the question can be found in Appendix 3.

*Table 4*

| | |
|---|---|
| Background Variables | •KJONN (Gender)<br>•UTD (Educational level)<br>•Aldr_kat (Age category)<br>•Alder (age)<br>•Region, series (Region)<br>•Kommnr (Municipality Number)<br>•S_06-S_08 (Position: Leader, Elected Union Representative, Elected Health and Safety Representative, Other) |
| Work-Tasks with Technology | •S_12 Series (Type of tasks: Core, Repport, Coordination, other) |
| Ameliorators | •S_14 Series (Reliability, User-Friendliness, Utility)<br>•S_19 (Implementation-Involvement)<br>•S_20-S_22 (Implementation-Involement)<br>•S_23 (Support Provision)<br>•S_24 Series (Implementation, User-Friendliness)<br>•S_25 series (Implementation, Literacy Facilitation)<br>•S_26 (Training, Literacy-Facilitation)<br>•S_27 series (Litaracy Facilitation,Support-Provision)<br>•S_28 (Literacy-Facilitation)<br>•S_29 (Support-Provision) |
| Use of Technology | •S_09 series (What type of technology do you use eg. PC, Handy, Tablet, etc. 19 alternatives)<br>•S_10 series (Which do you use most, second, and third most)<br>•S_11 (New Technology introduced?)<br>•S_13 series (Time used with ICT on tasks) |
| Stressors | •S_15 series (digital technology causing stress or strain: Work-Load, Time-Pressure, Technostress, Availability/Presenteeism)<br>•S_16 (Availability, Presenteeism)<br>•S_17 series (Availability conditional)<br>•S_18 (Automatic registration, Ambguity)<br>•S_30 (General stressors: work-load, time-pressure, mistakes, salary-size, digital tools, availability. Similar to S_15 except not exclusively related to digital tools) |

The overview in the table 4 below breaks down the variables in accordance to the terminology offered in the literature review, making it easier to identify the type of cause, condition, or effect that may have an association between covariates and the response. The questions in the appendix 3 supplies additional information that is helpful for deeper interpretation. The data is made in both numerical and nominal forms (see Appendix 1). However, provided the numerical dataset that are best suitable for modification of the data, as they represent the answers the respondents gave are represented by a number. Especially since the other answers are in Norwegian and provide no advantage other than additional edits. In the next section the explanation of how the data was edited to best fit the Tree-based models utilized in this thesis.

## 4.1 Data Preparation

Provided the function of decision tree's and logic of numerical variables a lot of the data could not retain its original form and had, therefore, to be permuted using the mutate and cut functions. As an example, gender or the variable 'KJONN' is represented by "1" and "2", suggesting that one of the genders is worth more than the other. Or branches in the variable 'BRANSJE' are numbers from 1-19. Likewise, if only recategorized as factor variables tree-based models might be misleading, as there could be a strong association between two predictors by chance regardless of the meaning of the answers being contradictory, such as, by accident answer 1, 2, and 6, on the scale have an association with technostress, while 3 and 5 does not. Hence, in the initial experimental trees, these and many other potential variables were transformed into binary with a 'None' option for the imputed 'NA's'. Examples are variable S_28, asking to what extent the respondent perceives the training as adequate in the workplace, which has a Likert-scale from 1-6 where, 1 is 'Always', 5 'Never', 6 'I don't know/Not sure'. To make sure we capture the answers that are closer to each other in meaning, producing either/or and an 'not sure/don't know/NA alternative variables that appeared important are muted into binary with a 'None' alternative. Such as S_28 is removed and replaced with S_28Binary with the following alternatives: 1-3 merged to "Digital Training", 4-5 "No Training", 6 and NA "None". The "None" alternative refers to none of the alternatives for the entire dataset and the NA, as it represents the respondents answers the best. The exceptions are made if any of the alternatives represent something equally contrasting to one of the sides of the scale. For example if we are trailing who experience technostress in S_15D, where only alternatives 4 and 5 represent some technostress, then all the other alternatives are labelled 'No Stress', including NA and the 'not sure/I don't know' alternative. Importantly, to keep the data as consistent as possible, NA's were included into the neutral ‘I don’t know/Not sure/Skip alternative in favor of the functions of the different packages in Rstudio instead of omitting the package. Therefore, some packages such as ‘ctree’ that run well with missing values still run with merged NA’s into the neutral alternative to ensure consistency across the various models.

### 4.1.1 Variable Alterations

After the experimentation with the variables using the ‘ctree’ in Rstudio, several variables were found to be problematic in their numerical form as the tree could over-react due to the sensitivity to many classes. For that reason, the data was converted into a factor variable data frame, with the following examples where the respective variables merged to form binary alternatives instead of all the numbers as factors. The entire process is represented in the code-snippet in appendix 2. However, to gain an overview, some examples are listed in table 5 below.

*Table 5*

| S_09 Series "A" to "Q" | |
|---|---|
| Original: 1-6 scale ⇨ | New: Digital, Not Digital |
| **S_10 series** | |
| Orignal: 19 classes ⇨ | New: PC, Handy, eCard, Other |
| **S_30E** | |
| Original: 1-7 scale ⇨ | New: Available ok, Available Stress, None |

The duplicate versions were at first applied to control for the different behavior of the tree, and if they appeared, they were removed from the data frame that was applied to the final tree in use for this thesis. However, if we follow the logic of the mutations and cuts above, the following considerations were taken for practical reasons. S_09 series represented a good avenue for mapping out digital users, as the series ask about the extent of use of specific digital tools. Hence, they were merged into one ‘S_09Binary’ variable in addition to the series, forming the new variable with options ‘Digital’ and ‘Not Digital’. As those who use the tools at least monthly, are defined as ‘Digital’ users. The rest of the series is kept in the factored form, with each number on the scale representing one factor. Note, both versions of classification trees react with “measuring the association between responses and covariates is the basis for an unbiased selection among covariates measured at different scales” (Hothorn, Hornik, & Zeileis, 2006, p. 2). However, the random forest model is biased towards covariates with many factors. Meaning, the more factors, more biased towards their significant associations, hence, to create an overall suitable data frame, it is necessary to avoid such variables by permuting them by limiting the number of classes (Zhang, et al., 2017, p. 34). This is provided by such variables entering the tree as the most significant in their original form. Some such variables were not permuted due to their irrelevance, regardless of having many factors. Hence, not having singled out any of the individual S_09 as significantly associated with the covariates and the response, as there is no hazard for the tree regardless of mutating the variable into dichotomous or any other number of classes.

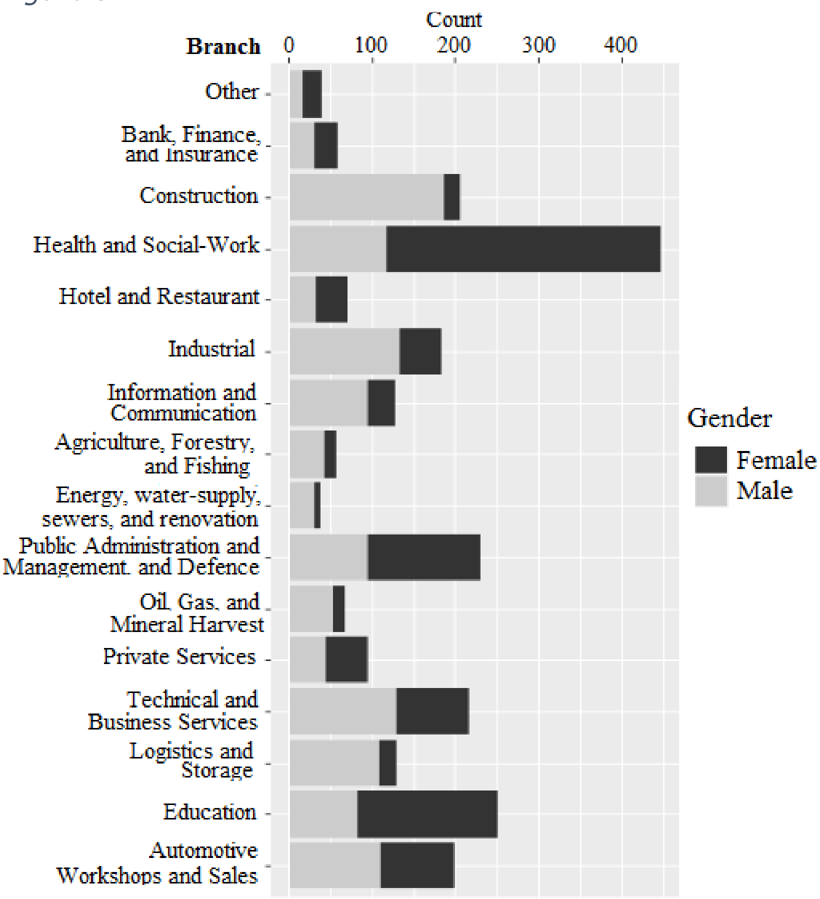
Likewise, the mutations of other variables are unnecessary if they will not produce a different tree regardless, as the tree will defenestrate due to the lack of significant association. The examples are the variables that did not gain any significant association in their original form. Therefore, only those that did, and especially those that provided inconsistent splits, were mutated or transformed. That is, if from a scale where 5 is “agree” and 1 is “disagree” in the interview, it makes little sense if answer 4 and 2 have a strong association, with not being stressed and 2, 3, 5, does as that would only yield mixed results. However, in a setting where 5, 4 and 1, 2, the has an association, one can argue that regardless if the person strongly agrees or strongly disagrees (to some question), it has an association with the outcome. An example of this is made for perceptions about availability provided by digital tools in variable S_15E, where the extremes on both sides had a significantly associated relationship with technostress. Hence, the variable was permuted into S_15EBinary in the data frame, with ‘Availability Positive’ and ‘Availability Negative’ instead of retaining the original Likert-Scale from 1 to 6, see appendix 3. That, however, is the sole example, for the rest of the variables, the results only tended to be mixed and uninterpretable in case of pre-mutation associations, or had no associations regardless, hence kept in the original form.

Then, the S_12 series mutation represented the type of tasks on a numerical percentage range, hence the high count of ‘Levels’ in table 5. These were redefined into above 50% and below as displayed in the snippet above. A similar choice of simplification for variables with a high count of ‘Levels’ was done for the S_10 series, where the respondents where asked what of the alternative tools they use the most. Provided the 16-19 alternative answers among the respondents, only the most common were included, namely ‘PC’, ‘Handy’, and ‘eCard’, the remaining have a low percentage count (see Figure 6) and therefore labeled ‘Other’. The ‘kommstr’ variable represents a rough partition into rural and urban workplaces with inhabitants ranging from below 50.000, 50.000-100.000, or 100.000 and above. These were divided into ‘small’, ‘medium’, and ‘large’. This is due to that the population counts are taken from the municipal register, where some may be rural, yet hold quite large populations regardless, making it fairer to only label them into sizes. The remaining variables alterations are the Likert-scale questions, where the considerations are taken in accordance to the interpretation where the answers represent either/or simplifications. They were rebranded to serve as the best overall for all the Tree-Based Models, and not specifically for each model, to retain consistency throughout the analysis.

## 4.2 Descriptive Statistics

This section is a presentation of the overall demographics of the sample, starting with the distribution among the branches, income-levels, and digital tools used by the target population. This is in order to provide an overview of the data at hand in an informative fashion, offering some exploratory interpretations. It includes the more basic demographics, digital tool usage, and relationships to technostress attainable through descriptive statistics.

*Figure 5*




As mentioned earlier, the data provided by SINTEF is a rough representation of the overall workforce in Norway, and it does so to the extent the survey company ‘Respons Analyse AS’ performed during the telephone interviews. In figure 5 on the right we can see the respective branches as they are defined by the statistical central bureau, and the overall count in the sample, divided by gender. Some branches have an overall low representation making tentative generalizations about the branches less accurate and risk over-fitting in case of attempting individual branch analysis. Nevertheless, the presentation here signifies the unequal distribution of the sample among the various branches. Again, not being hazardous for the overall purpose of exploring the digitalization and technostress throughout the data, yet a roughly similar to the distribution in the workforce. Another property worth mentioning, is that the gender distribution is quite fair in the sample compared to the national distribution of the branches. In the sample there is a female majority in education, and health and social work, and male domination in industrial, information and communication, logistics and storage, and construction. That distribution is proportionally reflective.

However, there is a somewhat underrepresentation of the lower income-group in the distribution. The ones earning less than 200.000 and between 200.000-400.000 NOK (Norwegian Crowns) are proportionally quite few in comparison to the overall population. Nevertheless, the average salary being around 550.000 NOK, is represented to a good degree, likewise the age groups are distributed in accordance to the national as we see few below 30 years old in the higher earning groups and few older in the low income in Figure 6. Furthermore, the digital tools, in accordance to what was expected ahead of the survey are generally used to a large extent in the Norwegian workforce. In the sample, approximately 2/3 of the respondents use some kind of tools at least on a weekly basis as opposed to those who use them less than weekly (Figure 7). Provided the fair distribution among the respondents in accordance to the overall work-force it gives us a good estimate if whether technostress can potentially be more prominent for those who use the tools a lot.

*Figure 6*

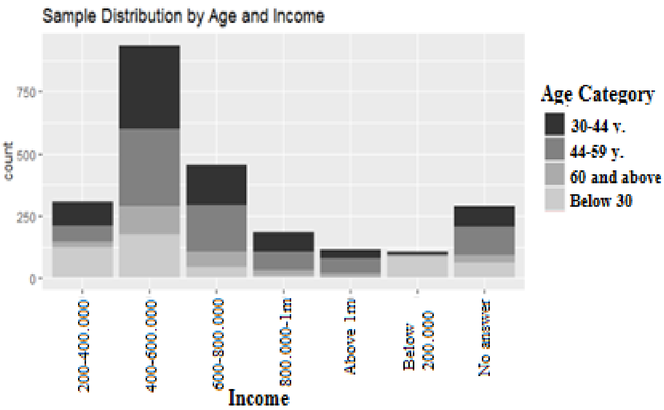
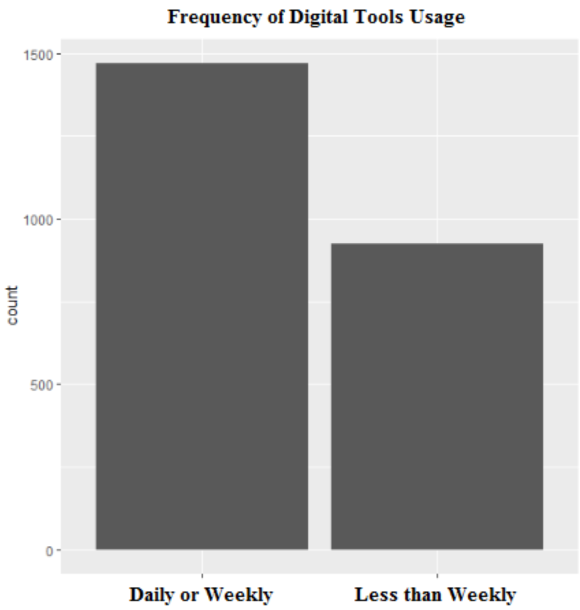
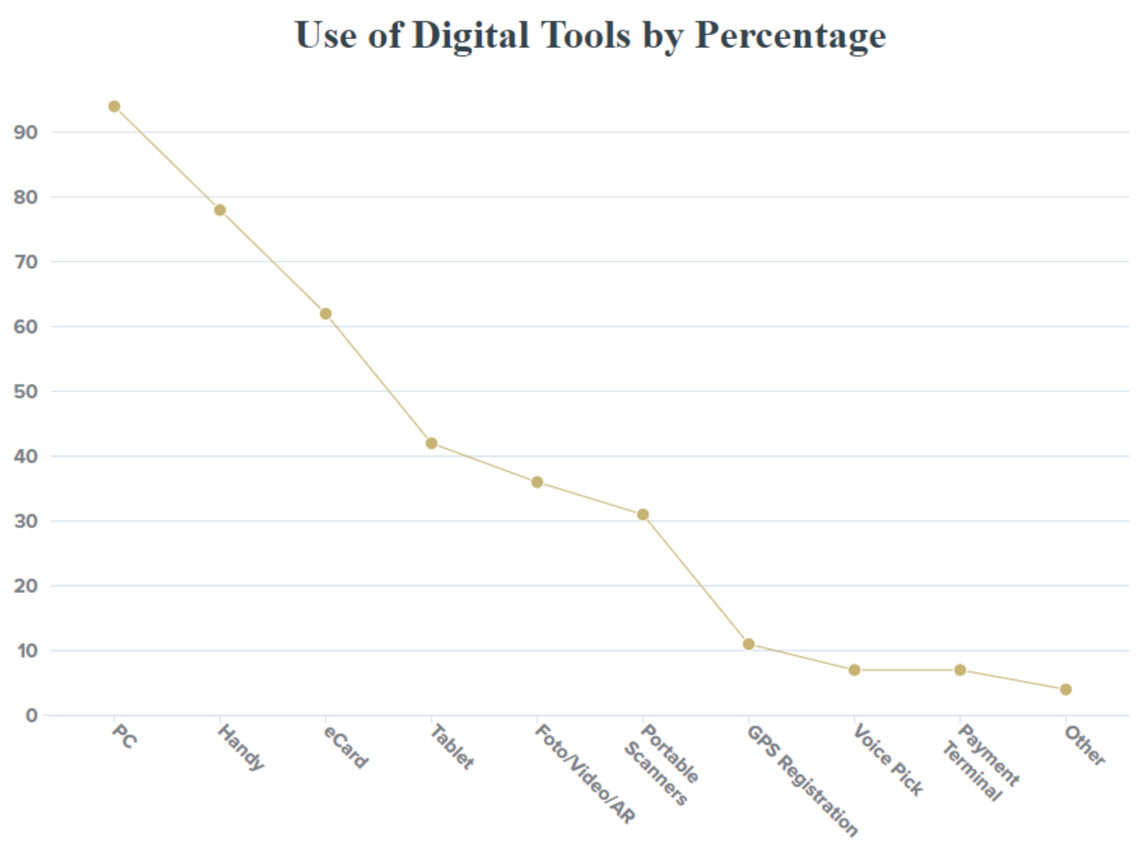



*Figure 7*



Furthermore, the distribution among the respondents seem to almost unanimously use some sort of digital tools in relation to their job with PC being almost unanimously (Figure 7 and 8), with other tools being used to various degrees (Figure 8). However, there is a considerable mix of tools in the workforce, it is dominated by smartphones, PC's, tablets, and electronic entry-cards, being the tools close to half of the entire workforce. This reflects the overall digitalization of the workforce and is a metric at the IDESI ranking also, confirming the extent of digitalization in the Norwegian workforce and additionally supplementing it with the composition of the tools in use. Provided the higher percentages or the top three tools, it is also therefore so that a large proportion of the workforce uses multiple tools in relation to their work, especially a combination of PC, smartphone, and electronic access card/ eCard. Hence, to conclude the descriptive overview, we can observe that the distributions among the working population is what Torvatn et al. (2017) states as quite close to representative of the real workforce. And the expectations about the extent digital tools in the workplace is also large, the vast majority uses digital tools, and as Figure 8 suggests, a large proportion uses a combination, as PC has above 90% users, then cell-phone/Handy, and eCard all above 60% as marked by the sloping chart.

*Figure 8*



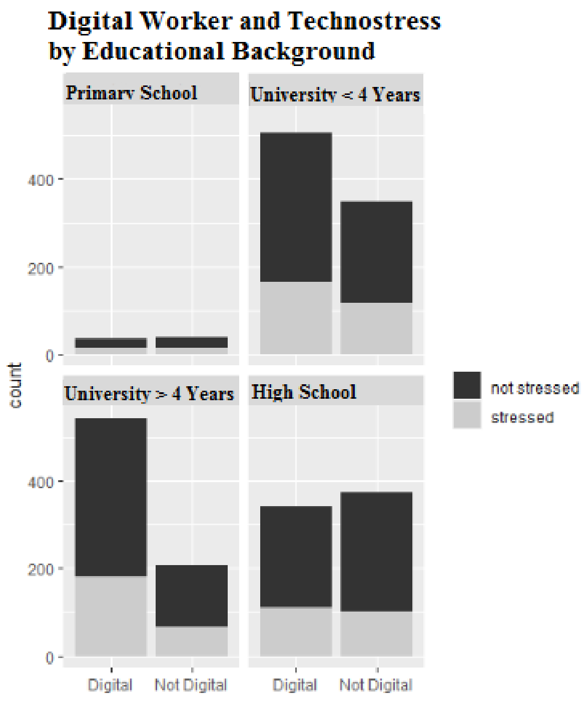
# 5 Exploratory Data Analysis

## 5.1 Descriptive Statistical Data Analysis

Provided that the principal target of the study is the extent of technostress, as earlier mentioned the phenomenon is only possible with regards to digital tools. Hence, the natural follow-up is the extent of the perceived technostress and severity among the respondents in the workforce. First off, is the metric that was found to be the strongest among the background variables, namely educational level combined with the frequency of digital tools usage and technostress. The overview among the respondents divided by educational level and both being digital (using digital tools in the workplace monthly, weekly, or daily) and being stressed by tools we can see some variations (Figure 9).

First off, the general population have few respondents that only completed primary school and therefore has around 40 respondents. However, we can see that among the respondents that have up to 4 years of higher education or only high school both the digital workers and the non-digital are quite equally stressed (or not). Yet, the above 4 years of education digital workers seem to outnumber the not digital at a larger ratio. Note also that the more education, the more likely the digital workers are to perceive technostress, as the ratios grow from high school (almost 1 to 1 ratio), through 'University < 4 years' (1.2 to 1 ratio) education to the 'University > 4 years' (2.8 to 1 ratio). Which suggests that educational level for the digital worker has higher than average likelihood of perceiving technostress. Nevertheless, it is fair to add that also that group is by far more likely to be in the 'Digital' worker category than being 'Not Digital', as the above 4 years of educated digital workers outnumber the 'not digital' at a ratio of 2.5 to 1. Although no correlation is to be claimed, we know based on the descriptive statistics that the above 4 years of educated digital worker represents an above average likelihood of perceiving technostress. However,

*Figure 9*



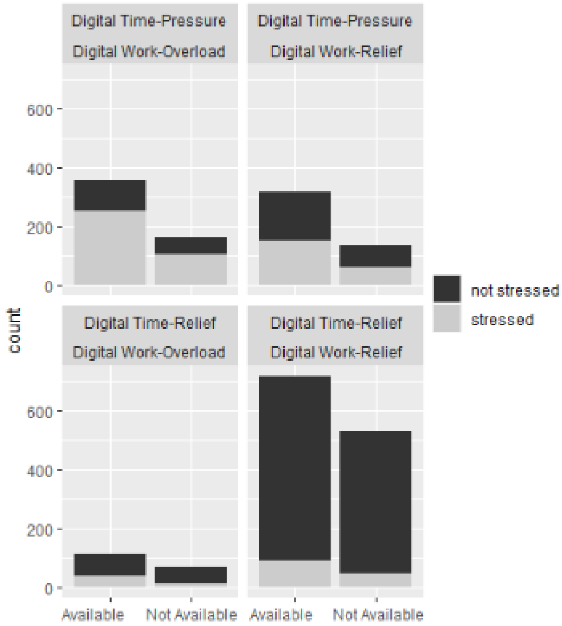
all University educated workers are more likely to be ‘Digital’. Another suggestion is that there are a lot of techno-stressed workers among the ‘Not Digital’ workers generally, and equally as many that perceive technostress as the ‘Digital’ for the ‘High School’ educated. Which points at the high likelihood that the non-extensive users of digital tools experience technostress in the few encounters they have with digital tools at work, or the work-related tools are perceived equally stressful as the voluntary private tools. Regardless of the reasoning behind it, both these are potential hypotheses worth considering for closer data exploration. In the following chapter utilizing decision trees this thesis will work through the categories of among these education and other background-related (e.g., age, branch, gender) variables to see what other factors seem to increase the likelihood for technostress.

Furthermore, apart from background variables where educational level appeared to be the most influential towards technostress, there are also other technostress-related variables that increase the likelihood of experiencing technostress. In this thesis, its among the primary targets to find the various combinations that collectively increase the likelihood of experiencing technostress. One such combination was found among the respondents that experience digital workload, digital time-pressure, and availability. Digital workload is when the digital tools have inflicted workload to the extent that the respondent considers them an additional burden, be it quantity and/or quality. Likewise, digital time-pressure is when the respondent considers the digital tools in the workplace to be increasing the time pressure. The opposites are called ‘Digital Time-Relief’ and ‘Digital Work-Relief’. Meaning that the digital tools spare the time or either simplify or reduce the perceived amount of work the respondent perceives doing. Combined with the ‘Available’ workers, those who according to the questionnaire, either are demanded, expected, or by their peers, leaders, customers, or voluntarily become ‘Available’ outside of working hours to perform or do a good job. Note, the questionnaire (see Appendix 3) did not specify this availability to be exclusively related to digital tools, hence the ‘Available’ are those who answered ‘Agree’ or ‘Partly Agree’ on the question: ‘Do you conduct work-related tasks outside of working hours such as answering email, texts, take phone-calls, etc.?’. The answers to the questions were originally on a scale from 1 (disagree) to 5 (agree), transformed to dichotomous either/or answers. Where e.g., the Digital Time-Pressure experience are represented by answers 4 (somewhat agree) and 5 (agree), and the rest categorized as ‘Digital Time-Relief’. Meaning that also those who were neutral are in the ‘Digital Time-Relief’ category, the same transformation was done for ‘Availability’, and ‘Digital Work-Overload’.

We can see that the 'Digital Work-Overload', 'Digital Time-Pressure and 'Available' category respondents are experiencing technostress, in pale-grey 'stressed' section. No other observed category has such an overrepresentation of respondents perceiving technostress. They see the highest proportion of technostress, at a ratio close to 4 out of 5. Seemingly there is a correlation of these factors also on the opposite side of the scale. As the respondents perceiving a 'Digital Time-Relief', and 'Digital Work-Relief' are rarely perceiving technostress, regardless of their availability outside of working hours as we can see in the bottom right corner of Figure 8. Furthermore, also observe the overall count differences. We can see that in accordance to our expectations, the majority experiences advantages in regard to digital tools, even though a few of them perceives them stressful, either always or more or less.

*Figure 10*

**Digital Time-Pressure, Work-Load, Availability, and Technostress**




Nonetheless, there is also another technostress-prone group of people among those who experience 'Digital Time-Pressure' and 'Digital Work-Relief' simultaneously. Leading to the first (1) hypothesis: Those experiencing 'Digital Time-Pressure' and 'Digital Work-Overload' represent the largest group of people perceiving technostress. And the (2) second being: Those experiencing 'Digital Time-Relief' and 'Digital Work-Relief' perceive the least stress. Suggesting a combination of task-tool-user combinations where some tasks or tools are a good match for relieving a lot of work, while using a lot of time in the process. Proportionally, we observe about half of them experiencing technostress regardless if they are 'Available' outside the working hours or not. The same cannot be stated regarding those who see experience 'Digital Time-Relief', but also 'Digital Work-Load', as they are proportionally less stressed, especially if they are the 'Not Available' outside of working hours. Note that 'Digital Work-Load' refers also to personal performance related pressures, not only a higher number of tasks in total, it is a combination of both quality and/or quantity.

## 5.2 Tree-Based Analysis

This chapter represents the main analysis yielding the tentative hypotheses an EDA research entails while utilizing conditional inference trees in Rstudio. The two types of classification trees are the ‘party’ conditional inference tree ‘ctree’, and ‘rpart’ classification tree. Although using similar logics of entropy and information gain, they use slightly different machine learning algorithms which have a slightly different performance and results from each other. Hothorn, Hornik and Zeileis (2006) stated that “the partitions induced by rpart trees are structurally different from the partition induced by conditional inference trees” and “may lead to different conclusions about the influence of certain covariates on the response” (p. 669). Therefore, it is wise in an EDA procedure utilizing decision trees, that the results are cross-checked by presenting both algorithms, comparing them to create best possible hypotheses from the data. This chapter will provide the details about the two different trees, a procedure of utilizing the entire dataset followed by partitioning, k-fold cross-validation, and pruning if necessary, to ensure accuracy, and avoid overfitting and bias. Then, continuing for a check with random forest and rounding up with the hypotheses generated from the entire analysis.

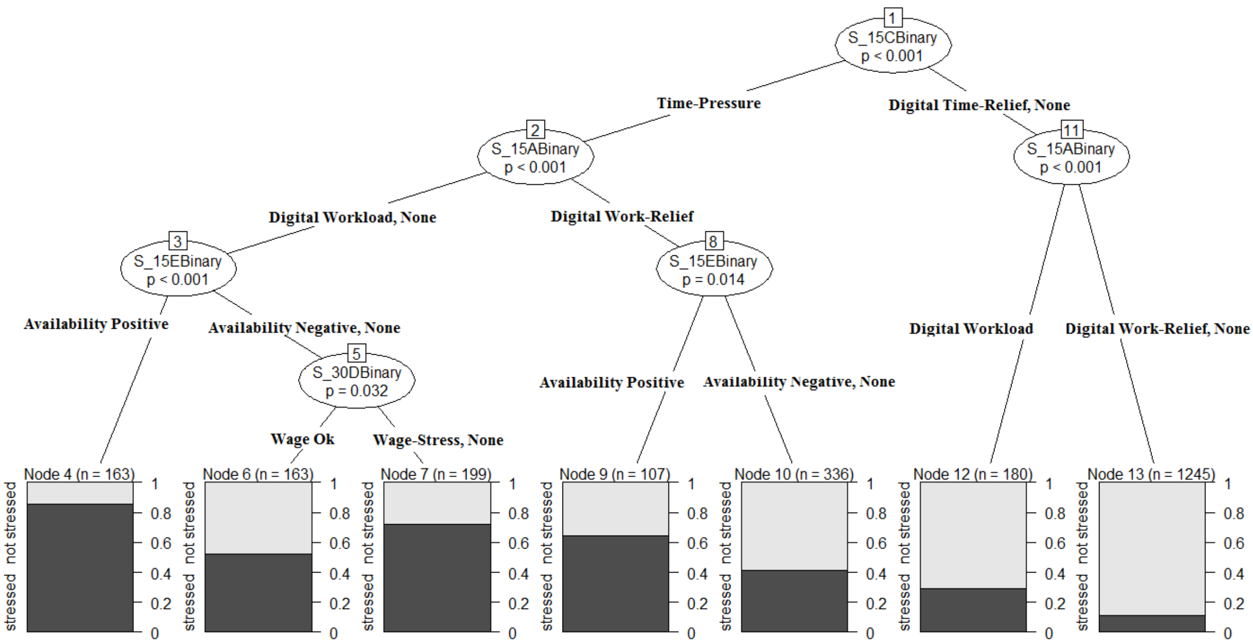
### 5.2.1 Ctree

The conditional inference tree, ‘ctree’ uses an algorithm, that was design to overcome the employment of “greedy search strategies, directly comparing all possible split points in all available covariates” (Schlosser, Hothorn, & Zeileis, 2019, p. 1). Which is described in the previous paragraph, starting with the statistical inference of the variable utilizing p-value, then, it divides the data based upon the most “homogenous sub-groups based on a set of covariates” (p. 3). This means that regardless of how many variables are available in the data frame, it will pick the most homogenous attribute based on the entropy of the variable that follows (Quinlan, 1985, p. 100). The response variable is dichotomous in this case, being the respondent experiencing technostress and thereby being relevant for either the ‘stressed’ or ‘not stressed’ respondents. If any of the predictors are irrelevant the process terminates, as it is based on the “significant association between any of the covariates and the response” (Hothorn, Hornik, & Zeileis, 2006, p. 652). For example, if the next variable provides an irrelevant answer in the data frame, say, the geographic location of the respondent’s workplace, it is not selected. Provided that it is so, that geography has less significance than any better alternative, or is generally insignificantly associated. However, if the variable in the node, and the following variable significantly increases likelihood for technostress, it will be included based on all the possible options in the data frame. Say, experiencing of ‘Availability’ outside of working hours and ‘Digital Overload’ are related to technostress according to the data, then ‘Availability’ is added with ‘Digital Overload’ as a follow-up of and split the tree accordingly. Likewise, if a variable ameliorates technostress, such as ‘No Digital Overload’, and reduces the likelihood of the respondent experiencing technostress, the tree will split in the opposite direction which is ‘not stressed’ until the most homogenous result possible is achieved. This is related to the entropy of the attributes, less the entropy, the more relevance the variable has for the outcome variable, and hence more information gain.

The tree in Figure 11 is built with the data that coincides with earlier literature, and likewise inspired the descriptive presentation in the previous sub-chapter. However, more information and findings are available thanks to the structure of the tree-plot. First off, the root-node on the top, number 1, is ranked as the most important in the division followed by the most relevant variables in the following nodes (Figure 11). Note that the S_15 series dominate the tree, these questions were formulated as sub-categories of the main question regarding ‘the consequences of digital technology’. The interviewees were asked to explain ‘to what degree they agree or disagree with each claim’. The S_15CBinary is a variable built out of the claim ‘digital technology has resulted in a higher time-pressure for me’ on a Likert-scale from 1 ‘completely disagree’ to 5 ‘completely agree’ and 6 being ‘not sure/don’t know’. However, the data in the tree was reconfigured to a binary answer and a third ’none’ option. Hence, the results in the tree are reflected as ‘Time-Pressure’ for the scores 4, and 5, and ‘Digital Time-Relief’ for 1,2, and 3. The split is therefore performed upon either of the two answers with ‘None’ joining one or the other branch. The ‘None’ is either alternative 6, or NA imputations, which was necessary for the later cross-validations. The other variables starting with S_30 series have a similar dichotomous configuration based upon their content as listed in the table attached to the plot (Figure 11). Except running from 1 to 7, with 7 being ‘None’ or the ‘I don’t know’ alternative, which in similarity to the other variables that include the NA imputation. The S_30 series are introduced to the respondents as ‘How much stress/frustrations do you experience in relation to the following conditions in your work-place?’. Followed up by the alternatives, on a Likert-scale from 1-7, where 1 is ‘No Stress’ and 6 ‘A lot of Stress’ and the in between are blank, with 7 being ‘I don’t know/Not sure’. In the case of the tree below, the S_30DBinary, or the binary configuration of the ‘d’ part of the question is included. In that configuration, 1-3 are coded as ‘Wage Ok’, or that wage does not correspond to higher amounts of stress, while 4-6 are coded ‘Wage Stress’ and 7 as ‘None’.

This initial tree is based on the entire data with 2393 respondents, which is an important step of the EDA, where the data is twisted and turned in multiple ways with transparent presentations of the data in terms with Haig's (2005) suggestions of EDA procedures. Furthermore, the criterions for each leaf-node is set at minimum 200 respondents with a minimum p-value of .05 which includes only the more significant values. The terminal nodes in the 'ctree' version are set with proportions, with dark-grey representing the 'stressed' and light-grey 'not stressed', based upon the regular S_15 series scale distribution for the binary reconfiguration. The question sounds 'Do you consider digital tools a source of stress in your workplace'? Which was designed as the technostress variable in the original report by SINTEF and therefore also the main outcome variable in this thesis.

*Figure 11*




| Questions | Answers: original scale, and dichotomous reconfigurations | | | | | | |
|---|---|---|---|---|---|---|---|
| S_30 series (Original) | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| S_30DBinary (In the tree) | "Wage Ok" | | | "Wage Stress" | | | "None" |
| S_15 Series (Original) | 1 | 2 | 3 | 4 | 5 | 6 | |
| S_15C (In the tree) | "Digital Time-Relief" | | | "Time-Pressure" | | "None" | |
| S_15A (In the tree) | "Digital Work-Relief" | | | "Digital Workload" | | "None" | |
| S_15E (In the tree) | "Availability Negative" | | | "Availability Positive" | | "None" | |

What we can see from the tree that the most important variable in relation to technostress is on the top in the root-node, 15CBinary, which describes the stress of time-pressure experienced due to digital-tools. The branch stretching leftwards from the root-node follows into the bin of those most likely to perceive technostress, with a high digital workload and perceiving the availability provided by digital tools as positive (node 2-4). Interestingly, those who fancy the availability from digital tools are also those who are the most stressed, while, in nodes 5, 6, 7, where those perceive it negatively are slightly better off, especially if wage size is not a source of stress. The difference is staggering as, the respondents are having below 50% chance of perceiving technostress as reflected in node 6, as opposed to node 4 which is above 80% likely of perceiving technostress. Based on the findings done in the report by SINTEF utilizing the same data, the time-pressure was not among the main stressors, however, it is listed as the most important when utilizing the 'party' based 'ctree'. It appears very important, as in comparison S_15A appears on the right side of the tree with 'Digital Work-Relief', where even with experience of 'Digital Workload' the likelihood of correlation with technostress is slightly around average of 30% in node 12. As for availability, which was considered the strongest variable in the SINTEF study, it still is among the most important out of 93, yet the second least important for variables at .05 of p-value limit in the tree of Figure 9. Hence, producing a high information gain, and entropy, and being significantly associated with the response.

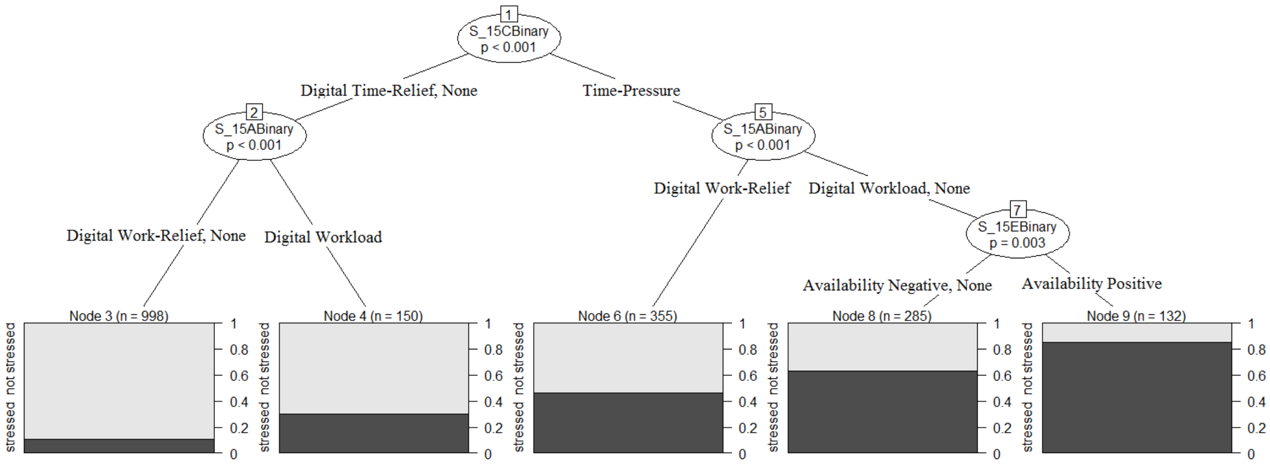
On the opposite side of the scale, we can also trail the most likely combination of the least likely of experiencing technostress, those who experience 'Digital Time-Relief' and 'Digital Work-Relief', with below 10% chance of perceived technostress in node 13. Note, that more than half of the respondents are in that category, 1245 out of 2393. Hence, there are a lot of those who experience technostress among them, just for other reasons than those most likely for the entire population as all other terminal nodes are ranging from 107 to 336 respondents. This will be addressed in more detail. However, apart from the initial findings from the tree, it is important to point out the accuracy, in Table 6, being at 78,14% with 195 observations of the 'stressed' and 328 'not stressed' miscategorized.

*Table 6*

```
Confusion Matrix and Statistics

              Reference
Prediction     not stressed stressed
  not stressed         1433      328
  stressed              195      437

               Accuracy : 0.7814
                 95% CI : (0.7643, 0.7979)
    No Information Rate : 0.6803
    P-Value [Acc > NIR] : < 2.2e-16
```

Unfortunately, this issue plagues the Tree-Based models despite the numerous advantages. This issue is addressed as a standard procedure when plotting trees by bagging, random forests, and boosting, as will be described in detail in the analysis. To comply with the procedure, first the data will be split into a random training and random test set with an 80% and 20% distribution. With 1920 observations in the training set and 473 in the test set.

*Figure 12*



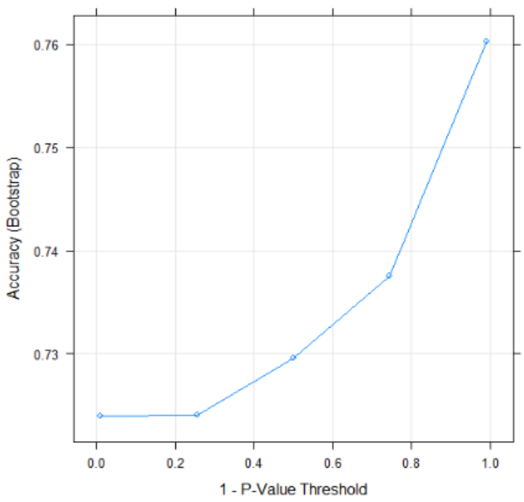

The tree reacted slightly differently with only the training data at hand, the S_30DBinary variable disappeared, while S_30FBinary reflecting to the extent the respondent perceives stress due to availability regardless of that being connected to digital tools in Node 6 in Figure 12. However, the same contrasts are visible, as both 'Digital Time-Relief', and 'Digital Work-Relief result in the lowest likelihood of perceived technostress. Likewise, for the opposite side of the scale, where digital 'Time-Pressure', 'Digital Workload' with digital 'Availability' both positive and negative produce the highest proportions of technostress in node 9, 10, and 11. Again, with the digital 'Availability Positive' producing the largest close to 90% in node 10 of likelihood of technostress provided the other covariate associations from node 1, and 5.

*Table 7*

```
Confusion Matrix and Statistics

              Reference
Prediction     not stressed stressed
  not stressed         1159      264
  stressed              153      344

               Accuracy : 0.7828
                 95% CI : (0.7637, 0.8011)
    No Information Rate : 0.6833
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4723
```

As for the confusion matrix and statistics in Table 7, we can observe similar distributions and accuracy, only with lower overall counts. Furthermore, this tree was pruned based on the bootstrapped resampling of 25 repetitions with increasing P-value's in the tuning parameter, using a higher P-Value threshold based on the bootstrapping performed to achieve the higher accuracy the marginal improvement confirms the extent of the generalizability of the non-resampled data.

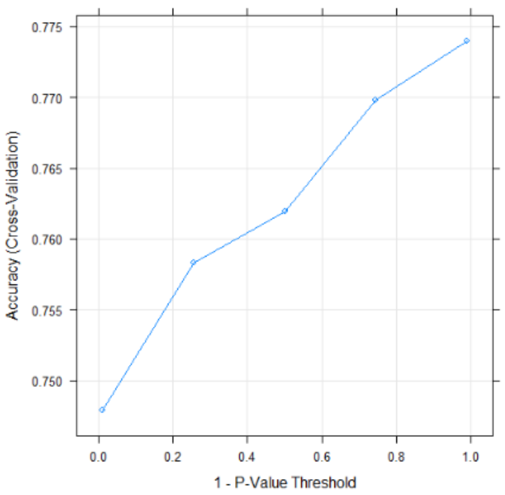
*Table 8*




```
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 1920
Resampling results across
tuning parameters:

  mincriterion  Accuracy   Kappa
  0.010         0.7239827  0.3518112
  0.255         0.7240172  0.3506653
  0.500         0.7295695  0.3624140
  0.745         0.7375511  0.3810514
  0.990         0.7602301  0.4327464

Accuracy was used to select the
optimal model using the largest value.
The final value used for the model was
mincriterion = 0.99.
```

This re-interpretation relies on modifying the variables to a larger extent to see if there are other avenues of discovery are possible when it comes to the associations that perform with better accuracy. Hence, obtaining accurate and validated tentative results, and exposing them in the EDA of the sample. As observable in Table 8 above, below “tuning parameters”, the sample was bootstrapped 25 times, and achieved the best accuracy of 76% on a p-value of 0.99 which is “mincriterion” in the table. Similar approach was conducted utilizing 10-fold cross-validation achieving slightly better results, achieving 77,4% accuracy at 0.99 of “mincriterion”, setting the p-value at .01.

*Table 9*



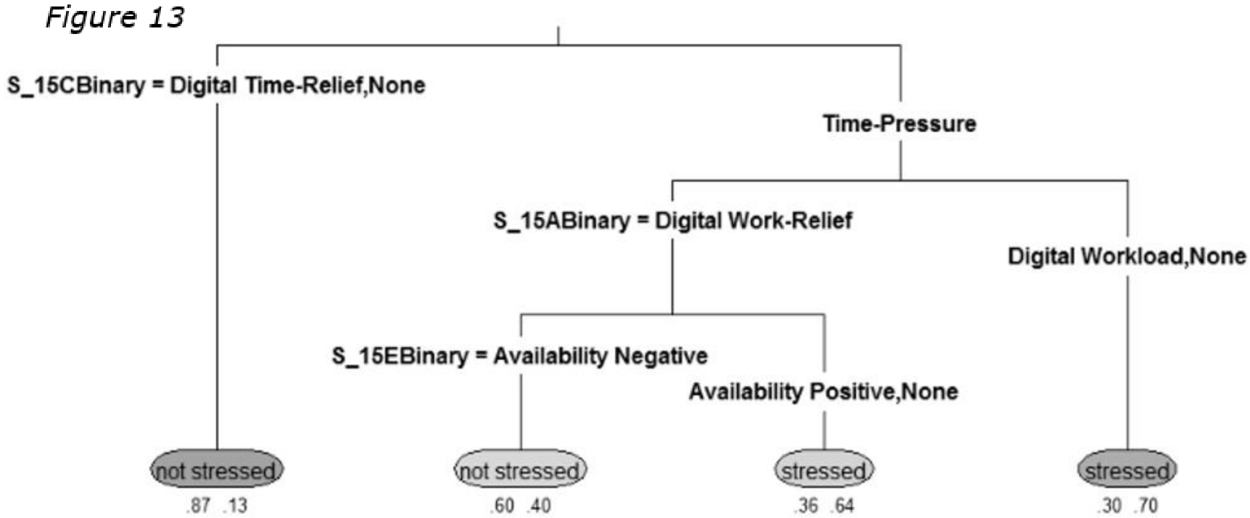

```
1920 samples
87 predictor
2 classes: 'not stressed', 'stressed'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1728
Resampling results across
tuning parameters:

mincriterion  Accuracy   Kappa
0.010         0.7479125  0.3813293
0.255         0.7583265  0.4194665
0.500         0.7619670  0.4323700
0.745         0.7698066  0.4495530
0.990         0.7739625  0.4412579

Accuracy was used to select the optimal
model using the largest value.
The final value used for the model was
 mincriterion = 0.99.
```

The tree looks identical to the tree in Figure 11, as none of the leaf-nodes have a p-value higher than .003. With the comparison of the two methods, we can observe that the optimal accuracy levels achieved on bootstrapping of 76% were already met for the 10-fold cross validation at much lower p-value threshold (ca 0.4). Suggesting that the number of variables in the data provides both better overall accuracies, and potentially more of the significant associations among the variables when using the 10-fold cross-validation method. Regardless of which, the first model introduced shows the entire data frame, with more associations among the predictors and response variable with low entropy. This as a process of EDA is worthy to mention for future CDA, however, it is not as generalizable as the second tree with 10-fold cross-validation, bootstrapping resampling to ensure generalizability to the extent of the accuracy level. The initial accuracy of the trees, especially for the training set in Table 6, may produce optimistic results which must be controlled with OOB error of random forest (Strobl et al., 2010, p. 335). Likewise, Cohens Kappa reliability statistic, range up to approximately 0.45, which indicates “moderate strength” prediction for the generalizable classification agreement according to Landis and Koch’s benchmarks (1977, p. 165). Hence, the results yield some predictive power improvement potential for future CDA applications.

### 5.2.2 Rpart

Unlike ctree, rpart does not utilize the significance test in the effort of bypassing the issues of overfitting. The ctree is based on selection by permutation-based significance test, which is observable as the p-values in the nodes, which are absent for the rpart version. Normally, rpart would also have the tendency of selecting variables with many missing values, this problem is however skipped in this analysis as the missing values are all imputed with the ‘I don’t know’ or ‘neither’ alternative based on the question. The results are therefore slightly different for the rpart. Which poses no hazard in the context of both trees in EDA.

*Figure 13*



As mentioned earlier, the rpart tree in the figure 13, produced slightly different results, with the most homogenous of the 87% 'not stressed' group on the left, only referring to 'Digital Time-Relief' and 'None' options. Additionally, we have three groups, experiencing 'Time-Pressure' due to digital tools distributed among the most 'stressed' at 70% for those experiencing 'Digital Workload' and 'None' in addition to 'Time Pressure'. Followed by those experiencing 'Digital Work-Relief' and seeing 'Availability [as] Positive' and 'None' at 64% stressed. And finally, as observed with the ctree utilizing identical training set, the respondents considering 'Availability Negative' are less 'stressed'.

*Table 10*

```
Conditional Inference Tree

1920 samples
  87 predictor
   2 classes: 'not stressed', 'stressed'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1728, 1728, 1727
Resampling results across tuning parameters:

  mincriterion  Accuracy   Kappa
  0.010         0.7510282  0.3924915
  0.255         0.7526016  0.4105497
  0.500         0.7572973  0.4252305
  0.745         0.7703319  0.4494505
  0.990         0.7677249  0.4321182

Accuracy was used to select the optimal model
using the largest value.
The final value used for the model was
mincriterion = 0.745.
```

## 5.3 Random Forest Analysis

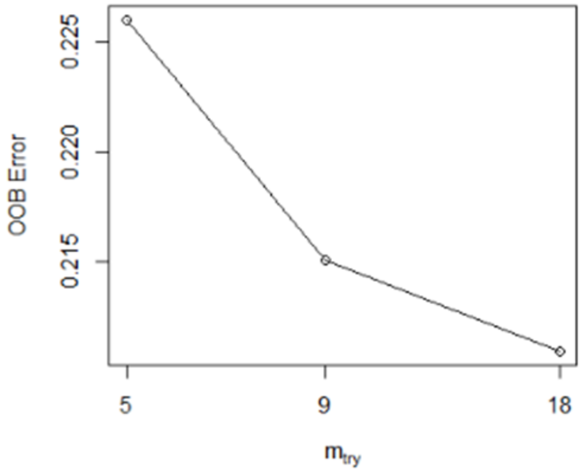
Lastly, the random forest model which I will start with a 1000 trees testing a variety of randomly selected variables, and provides an out-of-bag error estimate which is similar to the bootstrapping or bagging of the prediction error of the training set. As mentioned in section 3.2.1.2, this method us utilized to reduce variance and overfitting by randomly sampling the data and averaging the importance of each variable through the majority voting. In similarity to decision trees, it is a powerful method when the amount of covariates is large, like in this thesis. Provided, the robustness of the earlier classification trees against bias, yet weakness against variance, random forest is a technique, to curb the variance at the expense of some bias. Therefore, in an EDA approach it is crucial to utilize both methods for comparison, although 10-fold cross validation of the decision tree may produce similar, but not identical results. Moreover, the decision tree's have the ability to visualize, as opposed to random forest, as there is "no such thing as an average tree with a simple structure, that could be visualized for interpretation" (Strobl et al., 2009, p. 335).

*Table 11*

```
Call:
 randomForest(formula = S_15DBinary ~ ., data =
tFtrain, ntree = 1000)
               Type of random forest:
classification
                     Number of trees: 1000
No. of variables tried at each split: 9

        OOB estimate of  error rate: 22.19%
Confusion matrix:
             not stressed stressed class.error
not stressed         1188      124   0.0945122
stressed              302      306   0.4967105
```

*Table 12*

```
mtry = 9  OOB error = 21.51%
Searching left ...
mtry = 5      OOB error = 22.6%
-0.05084746 0.02
Searching right ...
mtry = 18     OOB error = 21.09%
0.01937046 0.02
```



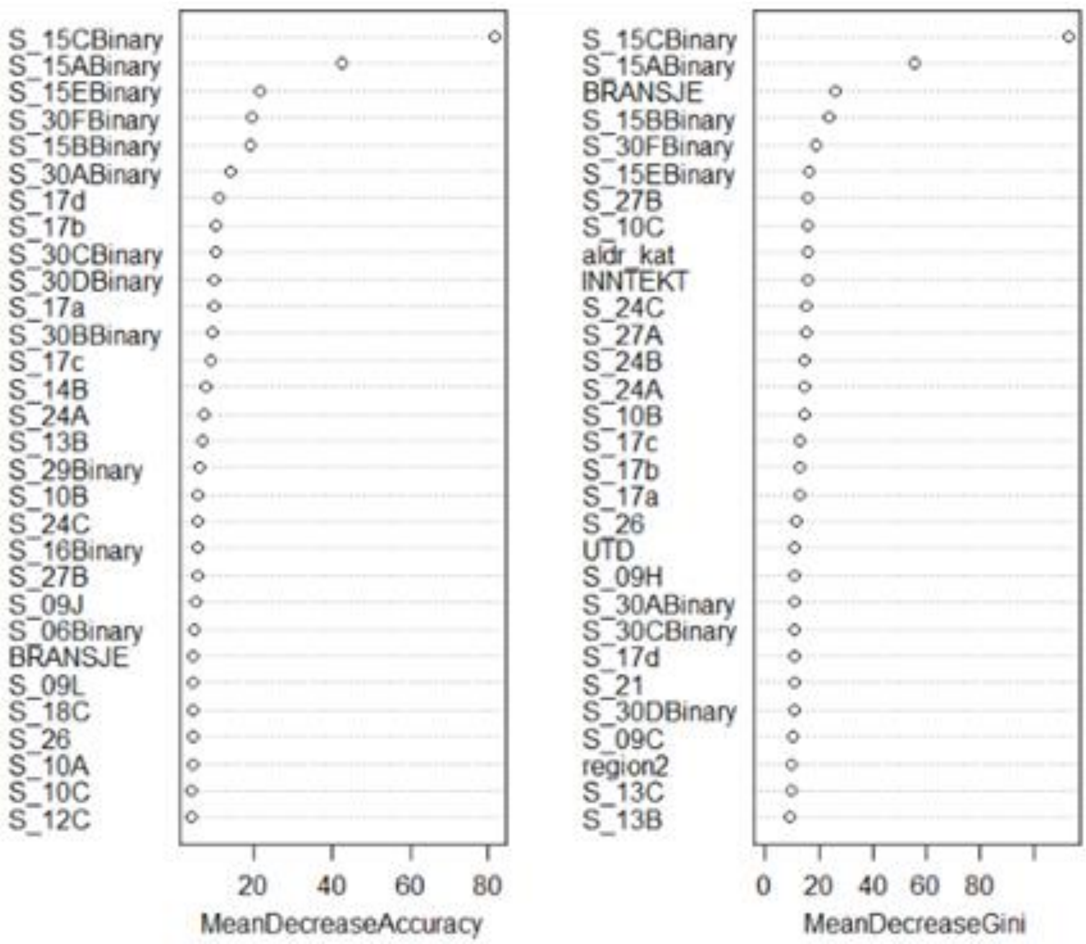

In the table 11 on the right we can see that the random forest algorithm utilized the parameters of 1000 trees with 9 mtry, or variables attempted at each cut point. Upon the initial trials of application of random forest on the training data, the results are close to excellent for the positive class, and staggeringly poor for the negative 'stressed' class with close to 50% classification error. Hence, also the majority of the out-of-bag error estimate lies with the negative class. In the confusion matrix in the lowest part of the table, we can see that 302 of 'stressed' were predicted, while they are 'not stressed', the real number should have been '306'. Hence, the accuracy especially for that class is poor. Henceforth, it is important to tune the forest before proceeding to the variable of importance plot, the take-away from random forest in the purposes of creating hypotheses. The tuning parameters are generated in the table 12 above. With the 'mtry = 18' scoring the lowest possible out-of-bag-error estimate. Henceforth, during the extraction of the variable of importance, these parameters were utilized as seen in

*Table 13*

```
Call:
 randomForest(x = rf.train, y = rf.label, ntree = 1000, mtry = 18,
 importance = TRUE, proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 18

        OOB estimate of  error rate: 21.46%
Confusion matrix:
             not stressed stressed class.error
not stressed         1165      147   0.1120427
stressed              265      343   0.4358553
```

the Table 13 above. The accuracy increased slightly overall, and about 6% for the negative 'stressed' class. Provided the data, this is the best accuracy achievable on the outcome. The table below represents the variable of importance plot produced with the random forest with the same covariates and outcome variable as in the tree models. Hence, we can observe the similar situation as in the tree-plots. With S_15CBinary, S_15ABinary, S_15E Binary crowned as the most important on the mean decrease accuracy scale. Generating the hypothesis (3) of their covariate association pattern with technostress. Which refers to their predictive ability in regard to the outcome variable S_15DBinary referring to technostress. Only the top 30 most important are presented, as the remaining have less clear relationship with the outcome at average for all the 1000 trees in the ensemble or forest.

*Figure 14*




Note that for the mean decrease Gini, BRANSJE or branch plays the third crucial role. This is however due to the number of classes in the variable towards which mean decrease Gini is biased. As opposed to all the other variables, BRANSJE was not easily mutated, neither would it have any drastic effects, as it was consistently not making it to the trees. As it is the way someone uses digital tools in the type of work that is important. Additionally, there is no mention about what particular task one has in the branch, such as working in the oil and gas, and mining industry, without specification of what the person does. It could range from service-desk jobs to drilling, or managerial tasks. A second notice for the mean decrease Gini, which measures entropy, more decrease, less entropy, less entropy means better prediction for each listed variable overall throughout the ensemble. The variables mentioned often as important mitigators of technostress, made it to top 10 most important, S_27B refers to training, support provision, literacy facilitation, depending on the interpretation. However, this is a thankful find, as it is the only active mitigator to have made it among the most contributive of all variables to choose from, see Table 4, chapter 4. A hypothesis that would require interpretation of how and why these two perceptions among the respondents matched. However, one apparent view is that they perceive literacy facilitation as useful, and provided the prediction accuracy for 'not stressed' in the random forest application, more often for those who share that perception.

Furthermore, an interesting finding is the relevance of S_30FBinary, S_15BBinary, and S_30ABinary. S_30FBinary, refers to stress due to availability, note, unlike S_15EBinary it does not reflect digital tools explicitly. It is general stress due to availability, regardless if that refers to digital tools. One can question what other possible means one has, yet that is the formulation of the question. Similarly, S_30ABinary does not mention digital tools while referring to workload. Which provides an interesting insight into the possibility of general stress due to workload, and availability and the similar variables referring to digital tools, as both seem to have a relationship with technostress regardless. A hypothesis to draw from this, is that technostress has a synergic effect with other forms of stress (5). Or, that those who already experience stress, experience it likewise with digital tools if the question relates to availability and workload. Lastly, the S_15BBinary referring to digital work-demands or demands for increased concentration provided the digital tools in the workplace is also added among the top 5 most important variables in the mean decrease accuracy. Which until now has not been observed by any of the trees. Yet, the congruence between the individual decision trees, and the ensemble method points out the combination of digital tools causing time-pressure, and workload. However, the majority of the variables have little associations with technostress, hence, the last hypothesis (6) reads: The majority of the variables have little relationship with perceived technostress.

However, it is worthy keeping in mind two more conditions when interpreting the random forest variables of importance. These are the variables that have the lowest entropy, and optimal information gain. But the confusion matrix also suggested mispredictions for the ‘stressed’ to be manifold of the ‘not stressed’. We know from the mutations that it produces a binary option. And with classification error being lower for the ‘not stressed’, it is the opposite side that is likely best suited in most cases. Henceforth, as a mitigator of technostress, the absence of availability related stress and workload, the likelihood for being ‘not stressed’ is very high. Similarly, the ‘Digital Time-Relief’ and ‘Digital Work-Relief’ of S_15ABinary and S_15CBinary, the two most importantly associated variables with the response had very few misclassified predictions for ‘not stressed’. Henceforth, the top 5 variables serve as a double-edged sword of hypothesis generation for the utility of potential follow-up confirmatory studies. Especially, provided that these findings are exclusive to this EDA, and the only hypothesizing about the synergic effects of the most significantly associated covariates with technostress, in a ranked order.

The whole analysis yields the following hypotheses answering RQ1 and RQ2:

- Those experiencing ‘Digital Time-Pressure’ and ‘Digital Work-Overload’ represent the largest group of people perceiving technostress. (1)
- Those experiencing ‘Digital Time-Relief’ and ‘Digital Work-Relief’ perceive the least stress. (2)
- Digital Workload, Time-Pressure together with digital Availability are the three most common perceptions for those experiencing technostress. (3)
- Those experiencing positive sides of availability due to digital tools are more prone to technostress than those who do not (4)
- Stress and technostress have similar patterns when it comes to workload, and time-pressure in association to perceived technostress (5)
- There is little relationship with technostress for large proportions of variables (6)

# 6 Discussion

The technostress issue, due to being a fresh phenomenon does not have clear premanufactured variables to test for its existence, although the relevant literature in the review offered some optimism about the relationships. Nevertheless, as an EDA approach, it is yet important to remain open to new ideas, as they are to be spawned from the data, not predetermined by the findings elsewhere. Hence, the combination of digital tools causing time-pressure and additional workload leading to technostress is the most plausible hypotheses generated from the data. As that was the prophecy in agreement for all the tree-based methods utilized, and suggested without further details in the descriptive statistical analysis. However, with the exception of 'rpart', the other models also suggested availability due to digital tools another last attachment to the earlier combination, especially if the respondent viewed this availability as advantageous.

However, more notable results were found in the tuned random forest model where other stress-factors, not referring to digital tools were found significantly associated with technostress. Providing the hypotheses suggesting either there is a synergy between the non-digital stress and technostress, especially when it comes to workload and availability. To illustrate the importance of retaining an open-minded outlook when exploring for associations such as in this thesis. Comparatively, SINTEF study for which the data was made, completely different findings were emphasized as ameliorators of technostress such as training and functioning technology (Torvatn et al., 2017, p. 44). Note, that SINTEF used a different approach, where several of the variables were ensembled into one category, functioning technology is a construct of reliability related variables, user-friendliness, and time-relieving. While, in this thesis, the time-relieving variable was treated by it-self. Henceforth, SINTEF did suggest something similar, however, a different combination was discovered. Which is expected from utilizing different methods, and even more so, not predetermining any combination of variables based upon earlier literature, instead testing the predictability of each in the various tree-based methods. This yielded an interesting, yet tentative finding, that till now was hidden in the data.

Furthermore, SINTEF also had a construct of training and support variables as an mitigator of technostress. However, the tree-based models only found any of them to be of far lighter relevance, the training, literacy-facilitation variables S_27b and S_28 made it however to the variable importance top 30, however ranking as far less important in predicting the outcome than the S_15 and S_30 series variables. The findings in the SINTEF study are credible, and logical, however, it is essential to see it in a broader perspective. Such as to what extent can training or literacy facilitation, implementation-facilitation/implementation involvement sufficiently curb technostress? A scenario of how it could work can be imagined, but the extent to which it would play out and empirically proven, requires time-series experimentation, and likely a qualitative-quantitative approach. As we would have to compare the quality, types of these activities facilitating the managerial coping strategies. The hypotheses generated in this thesis suggest that the relevance of those mitigators that were mentioned by Torvatn et al. (2017), Ragu-Nathan et al. (2008), Ayyagari et al. (2011), and Tarafdar et al. (2011), are consensual. Noting, that many of the researchers build on the previously available material, predetermining and guiding their efforts. Often, masterly, however, based off a research topic that is still in development. Which requires more EDA style approaches that can spawn wider confirmatory work.

Hence the extent of their effectiveness, say of the mitigators suggested in earlier literature, is hard to judge, especially since neither appeared decisive in the tree-based models. These findings are neither disproven, yet their extent is unknown, the effectiveness is uncertain. For instance, Tarafdar, et al., (2019) mentions this limitation to existing technostress literature when it comes to interdiciplinarity of the topic (p. 13). This, limitation leads to a lot unconfirmed assumptions in research designed for confirmatory purpose, provided the few pioneering terms the topic has so far. Likewise, this thesis found drastic differences in significance of covariation and predictive power on the response, suggesting that some have only slight, moderate effects. Potentially, that most of explanations are still beneath the layer the methods cannot observe, a different kind of knowledge than the current varibales, that are made out of the only literature available at this point. Nevertheless, that is a finding in it self in a EDA setting, that the relevance appears for most part of the variables, vague, and have little relationship with technostress as whole.

Future research can capitalize on the hypotheses deducted, either for conducting similar EDA with different datasets but similar variables, or run a confirmatory study. Accordingly, the findings can be interpreted from various perspectives in terms of technology development, education, psychological, and social sciences, in similarity to Kozyreva, Lewandowsky, and Hertwig, (2020, pp. 8-10). That is an approach oriented at influencing the structures surrounding the workers psychological wellbeing, and long term policy options. This literature can also be enhanced when considering the existing legislations, and potential implications they may have. This thesis adressed data from Norway, a social democratic welfare state sporting its own legislation protecting the psychological aspects of working conditions  (Bambra, et al., 2014), where relevance of technostress hazards in regards to policy implications are clearer. However, as both this research and earlier works have shown, technostress is quite wide-spread and broadly relevant. With the particular factors being increasingly exposed, further research in regards to both policy, and management adressing mitigation of this hazard is better facilitated. Likewise, for other research, in various fields, ranging from psychology to information systems, as mentioned by Tarafdar et al. (2019) is required to identify stronger associated variables to bolster technostress literature. By utilizing machine-learning, and tree-based models this EDA approach achieved paving the way for future research by generating the necessary hypotheses.

# 7 Conclusion

No research or experiment of this configuration on this topic has yet been explored, neither or confirmed in a CDA style in as to the best knowledge of this author. It tightened the otherwise wide research-gap in technostress related research. The intention was to generate hypotheses from the unexpected patterns in the data regarding technostress and regarding covarying predictors associating with technostress. The unexpected patterns are the chronologically ordered covarying predictors that together result in the most homogenous groups of techno-stressed, and not stressed respondents. Of course, as EDA dictates, the method only allows for tentative findings, hence, with the predictive accuracy being overall spread, such as Kappa values around .50, it would require additional modeling in a CDA setting. However, such is EDA, without predetermined theories, neither seeking to confirm, it helps the researcher gaining ideas.

These ideas resulted in the covariate association of digital time-pressure, workload, and availability, and the response representing the perceived technostress. These results are especially visible in the tree-plots in Figures 11, 12, and 13. Furthermore, when conducting the random forest algorithm, some other interesting hypotheses were generated. First, the best-ranked variables reflected the most important variables of the antecedent trees, the ones in the root node at the top. However, the forest observed what the trees could not, the other stress coinciding with the digital ones that were more predictive, suggesting a type of synergy between the stress types. However, as the goal was not to establish causality for which neither variable of importance is made, neither the goal of EDA, it is only a “rough assessment of the importance of potentially theoretically interesting variables" (Jones & Linder, 2015, p. 14).

Furthermore, another interesting finding is the small trees grown from the initial models, even the first tree in Figure 9. Is relatively small compared to the number of variables in a dataset made specifically to measure technostress and its mitigators. Likewise, the random forest variable of importance showed that many of the variables, even in their prediction, are only loosely associated with the outcome. This is likely due to the fragmentation of the earlier research finding only bits and pieces of relevant definitions without having a deeper understanding of technostress. Nevertheless, the goal was to generate hypotheses that do not yet exist in the literature, and the research has produced and listed the ones visible from the tree-based models. Henceforth, the hypotheses generated in the thesis can fruitfully add to the literature in a quite broad perspective, and especially if the most important and relevant associations are key components of future confirmatory research.

# 8 References


Acquaah-Mensah, G. K., Leach, S. M., & Guda, C. (2006). Predicting the Subcellular Localization of Human Proteins Using Machine Learning and Exploratory Data Analysis. *Genomics, Proteomics & Bioinformatics*, pp. 120-133.

Archer, M., Bhaskar, R., Collier, A., Lawson, T., & Norrie, A. (1998). *Critical Realism: Essential Readings.* New York: Routledge.

Arnetz, B. B. (1997). Technological stress: phychophysiological aspects of working with modern information technology. *Scandinavian Journal of Work, Environment & Health, Vol. 23 (3)*, pp. 97-103.

Astrup, N., & Helgesen, G. M. (2019). *Én digital offentlig sektor.* Oslo: Regjeringen.

Ayyagari, R., Grover, V., & Russell, P. (2011, December). Technostress: Technological Antecendents and Implications. *MIS Quarterly, 35(4)*, pp. 831-858.

Bambra, C., Lunau, T., Van der Wel, K. A., Eikemo, T. A., & Dragano, N. (2014). WORK, HEALTH, AND WELFARE: THE ASSOCIATION BETWEEN WORKING CONDITIONS, WELFARE STATES, AND SELF-REPORTED GENERAL HEALTH IN EUROPE. *International Journal of Health Services, 44(1)*, pp. 113-136. doi: http://dx.doi.org/10.2190/HS.44.1.g.

Baskhar, R. (1998). Chapter 2: Societies. In R. Baskhar, *The Possibility of Naturalism (3rd ed.)* (pp. 27-78). London: Routledge.

Behrens, J. T. (1997). Principles and Procedures of Exploratory Data Analysis. *Psychological Method, 2(2)*, pp. 131-160.

Behrens, J. T., Dicerbo, K. E., Yel, N., & Levy, R. (2012). Exploratory Data Analysis. In I. B. Weiner, J. A. Schinka, & W. F. Velicer, *Handbook of Psychology, Volume 2, Research Menthods in Psychology* (pp. 34-70). Tampa: John Wiley & Sons, Inc.

Berg-Beckhoff, G., Nielsen, G., & Ladekjær Larsen, E. (2017). Use of information communication technology and stress, brunout, and mental health in older, middle-aged, and younger workers - results from a systematic review. *International Journal of Occupational and Environmental Health, 23(2)*, pp. 160-171. https://doi.org/10.1080/10773525.2018.1436015.

Brod, C. (1984). *Technostress: The Human Cost of the Computer Revolution.* Reading: Addison Weslety.

Chiappetta, M. (2017). The Technostress: definition, symptoms and risk prevention. *Senses & Sciences, 4(1)*, pp. 358-361. doi: 10.14616/sands-2017-1-358361.

Deloitte. (2018). *Analysis: Workplace Burnout Survey*. Retrieved from Deloitte Development LLC: https://www2.deloitte.com/us/en/pages/about-deloitte/articles/burnout-survey.html

Eastwood, J., Jalaludin, B. B., & Kemp, L. A. (2014). Realist explanatory theory building method for social epidemiology: a protocol for mixed method multilevel study of neighbourhood context and postnatal depression. *Springerplus, 3(12)*, pp. 1-12.

Esping-Andersen, G. (1990). *The Three Worlds of Welfare Capitalism.* Princeton, NJ: Princeton University Press.

Fleischmann, A., Oppl, S., Schmidt, W., & Stary, C. (2020). Digitalization. In A. Fleischmann, S. Oppl, W. Schmidt, & C. Stary, *Corporate Governance and Business Process Management* (pp. 1-16). Cham: Springer Nature Switzerland.

Fox, M. J., Dwyer, D. J., & Ganster, D. C. (1993). Effects of Stressfull Job Demands and Control on Physiological and Attutudal Outcomes in a Hospital Setting. *Academy of Management Journal, 36(2).*, pp. 289-318.

Gaudioso, F., Turel, O., & Galimberti, C. (2017). The mediating roles of strain facets and coping strategies in translating techno-stressors into adverse job outcomes. *Computers in Human Behavior 69*, pp. 189-196.

Groes, S. (2017). Information overload in literature. *Textual Practice, 31(7)*, pp. 1481-1508.

Gupta, A., Sharda, R., Ducheneaut, N., Zhao, J. L., & Weber, R. (2006). E-Mail Management: A Techno-Managerial Research Perspective. *Communications of the Association for Information Systems, 17*, pp. 941-961.

Haig, B. D. (2005). An abductive theory of scientific method. *Psychological Methods, 10(4)*, pp. 371-388. doi: 10.1037/1082-989X.10.4.371.

Haig, D. B. (2018). Exploratory Data Analysis. In D. B. Haig, *The Philosophy of Quantitative Methods: Understanding Statistics* (pp. 13-40). Oxford Scholarship Online.

Ho, Y. C. (1994). Abduction? Deduction? Induction? *Paper presented at the Annual Meeting of the American Educational Research Association* (pp. 1-28). New Orleans: Educational Resources Information Center.

Ho, Y. C. (2010). Exploratory data analysis in the context of data mining and resampling. *International Journal of Psychological Research, 3(1)*, pp. 9-22.

Hogan, M. (2005). Technophobia Amongst Older Adults in Ireland. *Irish Journal of Management, Special Issue*, pp. 57-77.

Hong, E., & O'Neil, H. J. (1992). Instructional Strategies to Help Learners Build Relevant Mental Models in Inferential Statistics. *Journal of Educational Psychology, 84(2)*, pp. 150-159.

Hothorn, T., Hornik, K., & Zeileis, A. (2006, September). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics, 15(3)*, pp. 651-674.

James, G., Witten, D., Tibshirami, R., & Hastie, T. (2013). Tree-Based Models. In G. James, D. Witten, R. Tibshirami, & T. Hastie, *An Introduction to Statistical learning with Applications in R* (pp. 307-340). New York: Springer Inc.

Jebb, A. T., Parringon, S., & Woo, S. E. (2017). Exploratory data analysis as a doundation of inductive research. *Human Resource Management Review, 27*, pp. 265-276. http://dx.doi.org/10.1016/j.hrmr.2016.08.003.

Jones, Z., & Linder, F. (2015, April 16). Exploratory Data Analysis using Random Forests. *73rd Annual MPSA conference*, pp. 1-31.

Khedhaouria, A., & Cucchi, A. (2019, May 6). Technostress creators, personality traits, and job burnout: A fuzzy-set. *Journal of Business Reseach*, pp. 349-361.

Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens Versus the Internet: Confronting Digital Challenges with Cognitive Tools. *Invited submission to the Psychological Science in the Public Interest*, pp. 1-126. Preprint https://doi.org/10.31234/osf.io/ky4x8.

Landis, R. J., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics Society, 33(1)*, pp. 159-174. DOI: 10.2307/2529310.

Laumer, S., Maier, C., Weitzel, T., & Eckhardt. (2012, January). The Implementation of Large-Scale Information Systems in Small and Medium-Sized Enterprises--A Case Study of Work-and Health-Related Consequences. *45th Hawaii International Conference on System Sciences*, pp. 3159-3168.

Lawson, T. (2011, January 20). Theorizing Ontology. *Feminist Economics*, pp. 161-169.
Lazarus, R. S. (1991). Psychological Stress in the Workplace. *Journal of Social Behaviour and Personality, 6(7)*, pp. 1-13.
Maier, C., Laumer, S., & Eckhardt, A. (2015, May). Information technology as daily stressor: pinning down the causes of burnout. *Zeitschrift für Betriebswirtschaft, 85(4)*, pp. 349-387.
Mensah, A., & Adjei, N. K. (2020). Work-life balance and self-reported health among working adults in Europe: a gender and welfare state regime comparative analysis. *BMC Public Health*, pp. 1-18. https://doi.org/10.1186/s12889-020-09139-w.
Mergel, I., Edelmann, N., & Haug, N. (2019). Defining digital transformation: Results from expert interviews☆. *Government Information Quarterly, 36*, pp. 1-16.
Mingers, J. (2002). Real-izing Information Systems: Critical Realism as an Underpinning Philosophy for Information Systems. *ICIS Proceedings, 27*, pp. 295-303.
Moore, J. E. (2000, March). One Road to Turnover: An Examination of Work Exhaustion in Technology Professionals. *MIS Quarterly, 24(1)*, pp. 141-168.
Muter, P., Furedy, J. J., Vincent, A., & Pelcowitz, T. (1993). User-Hostile Systems and Patterns of Psychophysiological Activity. *Computers in Human Behavior, 9*, pp. 105-111.
Østvold, R., & Rehbinder, E. (2017). *Accenture.* Retrieved from Vårt Nye Digitale Norge: https://www.accenture.com/t20170920t205302z__w__/no-en/_acnmedia/pdf-59/accenture-digitale-norge-2.pdf
Patel, J., Ryoo, S., & Kettinger, W. (2012). Theorizing the Dual Role of Information Technology in Technostress Research. *AMCIS 2012 Proceedings, 12*, pp. 1-9.
Quinlan, J. R. (1985). Induction of Decision Trees . *Machine Learning, 1*, pp. 80-106.
Ragu-Nathan, T., Tarafdar, M., Ragu-Nathan, B. S., & Tu, Q. (2008, December). The Consequences of Technostress for End Users in Organizations: Conceptual Development and Empirical Validation. *Information Systems Research, 19 (4)*, pp. 417-433. doi 10.1287.
Regjeringen. (2018, October 8). *Statsbudsjettet 2019: Tidenes største satsing på digitalisering*. Retrieved from Regjeringen: https://www.regjeringen.no/no/aktuelt/tidenes-storste-satsing-pa-digitalisering/id2614074/
Riedl, R., Kindermann, H., & Javor, A. (2012). Technostress from a Neurobiological Perspective: System Breakdown Increases the Stress Hormone Cortisol in Computer Users. *Business & Information Systems Engineering*, pp. 61-69. DOI 10.1007/s12599-012-0207-7.
Saganuwan, M. U., Khairuzzaman, W. I., & Ungku, U. A. (2015, February 12). Conceptual Framework: AIS Technostress and Its Effect on Professionals' Job Outcomes. *Asian Social Science, 11(5)*, pp. 97-107.
Sailem, H. Z., Sero, J., & Bakal, C. (2015). Visualizing cellular imaging data using PhenoPlot. *NATURE COMMUNICATIONS*, pp. 1-6. DOI: 10.1038/ncomms6825.
Sami, K. L., & Pangannaiah, N. B. (2006). "Technostress" A literature survey on the effect of information technology on library users. *Library Review, 55(7)*, pp. 429-439. DOI 10.1108/00242530610682146.
Sayer, A. (2000). Key Features of Critical Realism in Practice: A Brief Outline. In A. Sayer, *Realism and social science* (pp. 10-28). London: SAGE Publications Ltd.
Schlosser, L., Hothorn, T., & Zeileis, A. (2019, June 24). *The Power of Unbiased Recursive Partitioning: A Unifying View of CTree, MOB, and GUIDE.* Retrieved from Cornell University: https://arxiv.org/abs/1906.10179

Sharma, H., & Kumar, S. (2016, April). A Survey on Decision Tree Algorithms of Classification in Data Mining. *International Journal of Science and Research, 5(4)*, pp. 2094-2097.

Shneiderman, B. (2003). The Eyes Have it: A Task by Data Type Taxonomy for Information Visualizations. In B. B. Benderson, *Interactive Technologies: THe Craft of Information Visualization* (pp. 364-371. https://doi.org/10.1016/B978-155860915-0/50046-9). San Fransisco: Morgan Kaufmann Publishers.

Shultz, K. S., Wang, M., & Olson, D. A. (2010). Role Overload and Underload in Relation to Occupational Stress and Health. *Stress and Health, 26*, pp. 99-111. DOI: 10.1002/smi.126.

Sicking, M. (2011, April 26). *Heise online*. Retrieved from Burnout weiter auf dem Vormarsch: https://www.heise.de/resale/artikel/Burnout-weiter-auf-dem-Vormarsch-1230727.html

Steger, S., Brenning, A., Bell, R., Petschko, H., & Glade, T. (2016). Exploring discrepancies between quantitative validation results and the geomorphic plausibility of statistical landslide susceptibility maps. *Geomorphology, 262*, pp. 8-23.

Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological Methods, 14(4)*, pp. 323-348.

Tarafdar, M., Cooper, C. L., & Stich, J.-F. (2019). The technostress trifecta - techno eustress, techno distress and design: Theoretical directions and an agenda for research. *Information Systems Journal, 29*, pp. 6-42.

Tarafdar, M., Pullins, B. E., & Ragu-Nathan, T. (2014). Examining impacts of technostress on the professional salesperson's behavioural performance. *Journal of Personal Selling & Sales Management, 34(1)*, pp. 51-69. http://dx.doi.org/10.1080/08853134.2013.870184.

Tarafdar, M., Tu, Q., & Ragu-Nathan, T. (2011). Impact of Technostress on End-User Satisfaction and Performance. *Journal of Management Information Systems, 27(3)*, pp. 303-334.

Tian, Y., & Zhang, W. (2019). THORS: An Efficient Approach for Making Classifiers Cost-Sensitive. *IEEE Acces, 7*, pp. 97704-97718.

Torvatn, H., Kløve, B., & Landmark, D. A. (2017). *Ansattes syn på digitalisering.* Trondheim: SINTEF Teknologi og samfunn.

Tu, Q., Wang, K., & Shu, Q. (2005). Computer-related technostress in China. *Communications of the ACM, 48(4)*, pp. 77-81.

Walwei, U. (2016, September). *Digitalization and structural labour market problems: The case of Germany.* Retrieved from International Labour Office: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---inst/documents/publication/wcms_522355.pdf

Wang, K., Shu, Q., & Tu, Q. (2008). Technostress under different organizational environments. *Including the Special Issue: Electronic Games and Personalized eLearning Processes, Computers in Human Behavior, 24(6)*, ss. 3002-3013. DOI: 10.1016/j.chb.2008.05.007.

Wang, Y., Zhang, Y., & Yu, X. (2020). A Comparative Assessment of Credit Risk Model Based on Machine Learning: A case study of bank loan data. *Procedia Computer Science, 174*, pp. 141-149.

Wong, K. (2017). Chapter 24: Decision Trees. In D. Alexander, & S. Carlos, *A Language not a Letter: Learning Statistics in R.* Chicago, Illinois , USA: https://ademos.people.uic.edu/Chapter24.html.

Zachariadis, M., Scott, S., & Barrett, M. (2013, September). Methodological Implications of Critical Realism for Mixed-Methods Research . *MIS Quarterly, 37(3)*, pp. 855-879.
Zhang, X., Li, Y., Kotagiri, R., Wu, L., Tari, Z., & Cheriet, M. (2017). KRNN: k Rare-class Nearest Neighbour classification. *Pattern Recognition, 62,* pp. 33-44.

# Appendix 1

Data format, variables in the numerical dataset, original labels, types.
(Digitalt_stress_varen2019_numerisk, see attachment)

| Variable | Levels | Type | Variable | Levels | Type | Variable | Levels | Type |
|---|---|---|---|---|---|---|---|---|
| kommnr | 328 | Numeric | S_09P | 6 | Numeric | S_18E | 3 | Numeric |
| fylke | 20 | Numeric | S_09Q | 6 | Numeric | S_19 | 6 | Numeric |
| region | 3 | Numeric | S_09R | 6 | Numeric | S_20_1 | 2 | Numeric |
| region2 | 4 | Numeric | s_09r_open | 262 | Character | S_20_2 | 2 | Numeric |
| region3 | 4 | Numeric | S_10A | 16 | Nominal | S_20_3 | 2 | Numeric |
| kommstr | 3 | Numeric | S_10B | 19 | Nominal | S_20_4 | 2 | Numeric |
| FODT | 57 | Numeric | S_10C | 18 | Nominal | S_20_5 | 2 | Numeric |
| alder | 57 | Numeric | S_11 | 3 | Nominal | s_20_open | 256 | Character |
| aldr_kat | 4 | Numeric | S_12A | 41 | Numeric | S_21 | 5 | Numeric |
| KJONN | 2 | Numeric | S_12B | 37 | Numeric | S_22 | 3 | Numeric |
| UTD | 4 | Numeric | S_12C | 39 | Numeric | S_23 | 7 | Numeric |
| INNTEKT | 7 | Numeric | S_12D | 56 | Numeric | S_24A | 6 | Numeric |
| BRANSJE | 16 | Numeric | S_13A | 6 | Numeric | S_24B | 6 | Numeric |
| Bransje_open | 30 | Character | S_13B | 6 | Numeric | S_24C | 6 | Numeric |
| S_06 | 4 | Numeric | S_13C | 6 | Numeric | S_25_1 | 2 | Numeric |
| S_07 | 2 | Numeric | S_14A | 6 | Numeric | S_25_2 | 2 | Numeric |
| S_08 | 2 | Numeric | S_14B | 6 | Numeric | S_25_3 | 2 | Numeric |
| S_09A | 6 | Numeric | S_14C | 6 | Numeric | S_25_4 | 2 | Numeric |
| S_09B | 6 | Numeric | S_15A | 6 | Numeric | S_25_5 | 2 | Numeric |
| S_09C | 6 | Numeric | S_15B | 6 | Numeric | S_26 | 5 | Numeric |
| S_09D | 6 | Numeric | S_15C | 6 | Numeric | S_27A | 6 | Numeric |
| S_09E | 6 | Numeric | S_15D | 6 | Numeric | S_27B | 6 | Numeric |
| S_09F | 6 | Numeric | S_15E | 6 | Numeric | S_27 | 6 | Numeric |
| S_09G | 6 | Numeric | S_16 | 5 | Numeric | S_28 | 6 | Numeric |
| S_09H | 6 | Numeric | S_17a | 5 | Numeric | S_29 | 6 | Numeric |
| S_09I | 6 | Numeric | S_17b | 5 | Numeric | S_30A | 7 | Numeric |
| S_09J | 6 | Numeric | S_17c | 5 | Numeric | S_30B | 7 | Numeric |
| S_09K | 6 | Numeric | S_17d | 5 | Numeric | S_30C | 7 | Numeric |
| S_09L | 6 | Numeric | S_18A | 3 | Numeric | S_30D | 7 | Numeric |
| S_09M | 6 | Numeric | S_18B | 3 | Numeric | S_30E | 7 | Numeric |
| S_09N | 6 | Numeric | S_18C | 3 | Numeric | S_30F | 7 | Numeric |
| S_09O | 6 | Numeric | S_18D | 3 | Numeric | vekt | 16 | Numeric |

# Appendix 2

Rstudio Code Snippet for Variable Mutations, and ‘cuts’

```
df<-read_xlsx(paste0(here(),"/","Digitalt_stress_varen2017_numerisk.xlsx"))
df_tree<-df

df_tree$S_15DBinary <- cut(df_tree$S_15D,c(0,3,5,6),labels=c("Not Stressed","stressed", "None"))
df_recat <- df_recat %>% mutate(S_06Binary = ifelse(S_06==1 | S_06==2 | S_06==3, "Leadership", "No Leader")

df_tree$S_09Binary <- cut(df_tree$S_09A & df_tree$S_09B & df_tree$S_09C & df_tree$S_09D & df_tree$S_09E
& df_tree$S_09F & df_tree$S_09G & df_tree$S_09H & df_tree$S_09I & df_tree$S_09J & df_tree$S_09K &
df_tree$S_09L & df_tree$S_09M & df_tree$S_09N & df_tree$S_09O & df_tree$S_09P &
df_tree$S_09Q,c(0,3,5,6),labels=c("Digital","Not Digital", "None"))

df_tree <- df_tree %>% mutate(S_12A = ifelse(S_12A >= 50, "Core Often", "Rare"))
df_tree <- df_tree %>% mutate(S_12B = ifelse(S_12B >= 50, "Report Often", "Rare"))
df_tree <- df_tree %>% mutate(S_12C = ifelse(S_12C >= 50, "Coord Often", "Rare"))
df_tree <- df_tree %>% mutate(S_12D = ifelse(S_12D >= 50, "Other Often", "Rare"))

df_treeFactor <- df_treeTest
df_treeFactor <- data.frame(lapply(df_treeFactor,as.factor))

df_treeFactor$S_10A <- cut(df_treeTest$S_10A,c(0,1,2,3,19),labels=c("PC","Handy","eCard","Other"))
df_treeFactor$S_10B <- cut(df_treeTest$S_10B,c(0,1,2,3,19),labels=c("PC","Handy","eCard","Other"))
df_treeFactor$S_10C <- cut(df_treeTest$S_10C,c(0,1,2,3,19),labels=c("PC","Handy","eCard","Other"))
df_treeFactor$kommstr <- cut(df_treeFactor$kommstr,c(0,1,2,3),labels = c("small","Medium","Large"))

df_treeTest$S_14A <- cut(df_treeTest$S_14A,c(0,3,5,6),labels=c("Unreliable Tech","Reliable Tech", "None"))
df_treeTest$S_14B <- cut(df_treeTest$S_14B,c(0,3,5,6),labels=c("Not User-Friendly","User-Friendly", "None"))
df_treeTest$S_14C <- cut(df_treeTest$S_14C,c(0,3,5,6),labels=c("Slow Tech","Fast Tech", "None"))
df_tree$S_15ABinary <- cut(df_tree$S_15A,c(0,3,5,6),labels=c("Digital Work-Relief","Digital Workload",
"None"))
df_tree$S_15BBinary <- cut(df_tree$S_15B,c(0,3,5,6),labels=c("Digital Relief","Digital Demand", "None"))
df_tree$S_15CBinary <- cut(df_tree$S_15C,c(0,3,5,6),labels=c("Digital Time-Relief","Time-Pressure", "None"))
df_tree$S_15EBinary <- cut(df_tree$S_15E,c(0,3,5,6),labels=c("Availability Positive","Availability Negative",
"None"))
df_tree$S_16Binary <- cut(df_tree$S_16,c(0,3,4,5),labels=c("Available","Unavailable", "None"))
df_tree$S_19Binary <- cut(df_tree$S_19,c(0,4,5,6),labels=c("Involved","Rare-Never", "None"))
df_tree$S_23Binary <- cut(df_tree$S_23,c(0,3,5,6,7),labels=c("Supported","Rare-Never","No need","None"))
df_tree$S_29Binary <- cut(df_tree$S_29,c(0,3,5,6),labels=c("Digital Support","Not Supported","None"))
df_tree$S_30ABinary <- cut(df_tree$S_30A,c(0,2,6,7),labels=c("Workload ok","Stressed","None"))
df_tree$S_30BBinary <- cut(df_tree$S_30B,c(0,2,6,7),labels=c("Time ok","Time-Stress","None"))
df_tree$S_30CBinary <- cut(df_tree$S_30C,c(0,2,6,7),labels=c("Mistakes ok","Mistake-Stress","None"))
df_tree$S_30DBinary <- cut(df_tree$S_30D,c(0,2,6,7),labels=c("Wage ok","Wage-Stress","None"))
df_tree$S_30EBinary <- cut(df_tree$S_30E,c(0,2,6,7),labels=c("Techno ok","Technostress","None"))
df_tree$S_30FBinary <- cut(df_tree$S_30F,c(0,2,6,7),labels=c("Available ok","Available-Stress","None"))
df_tree$S_29Binary <- cut(df_tree$S_29,c(0,3,5,6),labels=c("Digital Support","Not Supported","None"))
```

# Appendix 3

Interview Questions by Torvatn et al., 2017, own translation.

The following questions are a translation of the questions seen as related to this thesis, and based upon a replication and translation of the SINTEF study (Torvatn, Kløve, & Landmark, 2017).

Questionnaire

1. Are you currently employed in a paid job? In case you are, what branch? 1. Agriculture, forestry, fishing 2. Oil and gass 3. Industry 4. Energy or water supply 5. Construction 6. Commodity sales, car repair 7. Transport and storage 8. Hotel and restaurant 9. Information and communication (including media) 10. Bank, finance and insurance 11. Technical og business related services 12. Public administration and management (including defence) 13. Education 14. Health and Social Sector 15. Personal consultancy 16. Other 17. I don't work

201. When were you born?

2. What is your highest completed education?

1. Basic obligatory school

2. Gymnasium

3. University up to 4 years

4. University more than 4 years

3. What is you annual income (gross)?

1. Less than 200.000

2. 200-400.000

3. 400-600.000

4. 600-800.000

5. 800-1 million

6. More than 1 million

7. Will not reply

4. Gender?

1. Male

2. Female

6. Do you have leadership responsibilities?

1. Personel responsibilities

2. Professional responsibilities

3. Both

4. No

This questionnaire is aimed at your usage of digital technologies in relationship to your work and includes all kinds of ICT related tools and systems that create, communicate and share information digitally. That includes PC, tablets, cellphones, control-systems, scanners, cameras, informationsystems and tracing devices, and any other systems ou can come up with.

7. How often do you use the following systems in your work? On a scale from 1 to 5 (1: Daily, 2: Weekly, 3: Mothly, 4: Less than once a month, 5: Never, 6: Don't know.)

a) PC b) Smartphone c) Tablet d) Portable registration equipment/Scanner e) Voice pick f) Smart glasses/ Helm g) Smart watch h) Foto/Video i) 3d – Printer j) Mobile payment terminal k) Static cashier machine/Terminal l) GPS register for transport, electronic chaffeurs journal, courer book m) Robots in manufacturing

Andre Hansen Final Research Proposal STVN 14 30.01.2020 Welfare Policies and Management Lund University

15

n) Robots in transport o) Tracing technology like RFID, barcodes and QR for production/ and or storage and logistics p) Digital control systems for machinery q) Electronic access card r) Do you use other technologies? 1: Yes – Specify 2: No

8. Which one of the previous technologies do you use the most?

9. In relationship to what tasks do you use them?

- Core tasks

- Reporting and documentation

- Information and coordination tasks

- Other

10. Think of the technology in your workplace, and consider whether you agree with the following claims. (1-5 scale, 1. Disagree … 5 Completely agree, 6. Not sure)

a) The digital technology I use in my workplace is reliable

b) … User friendly?

c) The digital technology in my workplace works fast enough for my purpose

11. Agree/Disagree with the following claims regarding the consequences of digital technology on a scale from 1-5.

a) Digital technology has resulted in more workpressure for me

b) Digital technology has increased the demands for concentration in the workplace

c) Digital technology has increased the time-pressure in my workplace

d) Digital technology increases my stress-level

e) The demand for availability and flexibility that digital technology provides is positive in my view

12. Do you experience a requirement of availability also outside of the paid working hours?

1. Yes, often or always

2. Sometimes

3. Rarely

4. No

Andre Hansen Final Research Proposal STVN 14 30.01.2020 Welfare Policies and Management Lund University

16

5. Not sure

13. Does the following types of registration happen at your workplace? (Scale 1 – 3; 1. Yes, 2, No, 3, don't know)

a) Automatic registration during task completion through a data system/control system?

b) Automatic registration of work completion through sensors?

c) Automatic registration of work completion through navigation and tracing devices?

d) Automatic registration of work time?

e) Self-report of the work time?

14. When your workplace introduced new digital technology. Which ones of these training-forms did they provide or if not, which did you use yourself?

1. Educational course?

2. I read my self on the web or otherwise

3. My colleagues tought me

4. Digital training (e-learning, games or something similar)

5. Not sure / Don’t know

15. Did the training allow you to perform your tasks with technology in a better way? (Scale 1-5, 1: Completely disagree, … 5. Completely agree, 6. Don’t know)

a) Workload

b) Demands for concentration during work

c) Time-pressure during my work

d) Increases my stress levels

e) Availability the digital tools provide are positive for me

16. Do you experience being available for work outside of paid working hours?

(Scale: 1-5, 1.Yes, Always/Often … 5. No, Never/Rare)

17. Under which conditions are you working outside of paid working hours?

(Scale: 1-5, 1.Yes, Always/Often … 5. No, Never/Rare)

a) Leadership demands it

b) Customers demand it

c) My colleagues demand it

d) My ambitions dictate it, it is voluntary

18. Does the following registration take place at your workplace?

(Scale 1-3, 1.Yes, 2. No, 3. Don’t know/Not sure)

a) Automatic registration of work

b) Automatic registration through sensors

c) Automatic registration of work through tracking systems (GPS)

d) Automatic registration of time use

e) Self-registration of time use

19. Are the employed (or any representative) in your workplace involved when new technology is planned to be implemented?

(Scale: 1 – 6. 1: Always …. 5. Never, 6. Don’t know/Not sure

20. Who is participating when new technology is being planned for implementation?

1. Elected Work Representative?

2. Elected Representative of Health and Safety?

3. Selected employees delegated to participate by the leadership?

4. Other (open) Relfected as S_20_r_open

21. At what time do the elected representatives participate usually?

(Scale 1-5. 1. Before it is accepted, 2. In process of the specific demands, 3. In process of evaluation, 4. In process of local adjustments of the technology, 4. I don’t know/Not sure)

22. Do they evaluate the consequences of the work-environment when choosing new technologies?

(Scale 1-3, 1. Yes, 2, No. 3. I don’t know/Not sure)

23. Are you supported by your peers when new digital technology is introduced or when you ask for support?

(Scale 1-7. 1. Always …. 6. Never. 7. I don’t know)

24. Remember the last time digital technology was introduced that was relevant for you in your work. To what degree did it contribute to…?

(Scale 1-6. 1. To a large degree. 5. Very small degree. 6. I don’t know)

a) Work quality increase?

b) Speed increase?

c) Enabled you doing something that was before its implementation impossible?

25. Remember the last time your workplace introduced a technology. What literacy facilitating forms did you use?

1. Courses together with the other employees

2. I learned my self, through internet or otherwise

3. My colleagues taught me

4. Digital e-based learning

5. I don't know/Not sure

26. Which forms of literacy facilitation did you find the most appropriate?

1. Courses together with the other employees

2. I learned my self, through internet or otherwise

3. My colleagues taught me

4. Digital e-based learning

5. I don't know/ Im not sure

27a. To what degree do you agree or disagree to the following claims?

(Scale: 1-6. 1.Absolutely agree … 5. Absolutely disagree. 6. I don't know/ Not sure)

a) Last time we introduced a new technology changed the way we worked in the organization to utilize the technology optimally

b) We need more effective ways to engage co-workers in technology implementation and improvement

c) Do you get support, training, literacy facilitation when you require it?

   1. Yes, 2. No. 3. Don't know/Not sure

28. Do you think that you generally get appropriate training in the use of technological programmes in your work place?

(Scale. 1-6. 1. Always … 5. Never. 6. I don't know/Not sure)

29. How often do you get support and help from the closest leader when introducing digital tools when you require literacy facilitation?

(Scale 1-5. 1. Always …. 5. Never. 6. I don’t know/ Not sure)

30. How much stress/Frustrations do you experience when it comes to the following tasks? (Scale 1-6, 1. No stress, … 6. A lot of stress, 7. Don’t know, not sure)

a) Workload

b) Time pressure

c) Making mistakes

d) Salary size

e) Digital equipment/Systems

f) Demand for Availability