Aalto University
School of Business

APPLICATION OF MACHINE LEARNING TO PREDICT OCCURRENCE OF ACCIDENTS AT FINNISH CONSTRUCTION SITES

Master's Thesis
Ina Häkkä
Aalto University School of Business
Information and Service Management
Spring 2022

Aalto University
School of Business

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of master's thesis

Author Ina Häkkä

Title of thesis Application of Machine Learning to Predict Occurrence of Accidents at Finnish Construction Sites

Degree Master of Science in Economics and Business Administration

Degree programme Information and Service Management

Thesis advisor(s) Pekka Malo

Year of approval 2022 **Number of pages** 54 **Language** English

Abstract

Construction industry is one of the most hazardous industries due to high frequency of accidents and several safety hazards on the construction site. Frequent safety hazards are caused by the unique, complex and dynamic nature of construction projects, including constantly changing physical environment on construction sites, physically demanding working conditions and continuously evolving construction technology that may relocate or reform the possibility of error rather than downright eliminating it. Efforts to decrease accident frequency and mitigate safety risks on construction sites require enhanced focus on safety culture, practices and equipment as well as proactive safety management while assessing the effects of individual, managerial and environmental factors on safety performance.

The use of Artificial Intelligence (AI) to develop diagnostic and predictive models is deemed the next revolutionary advancement for occupational safety in the construction industry. This study is conducted as a case study on one of Finland's most prominent construction companies, aiming at versatilely utilizing its data to develop a predictive Machine Learning (ML) model capable of predicting the occurrence of accidents. The data includes environmental, managerial and project-specific factors from over 600 construction projects in Finland. Several studies aiming at construction accident prediction with Machine Learning exist, however, majority of the previous studies have lacked access to data of non-accident cases to complement records of accidents. Thus, these studies have focused on, for example, predicting accident severity rather than occurrence of accidents, limiting their contribution to proactive safety-risk management and accident prevention. One of the greatest strengths and contributions of this thesis is its ability to combine data of accident and non-accident cases, thus providing Machine Learning models examples that enable predictive classification of incidents on construction sites.

The findings show that while uninterpretable models may have significant predictive skill for this research problem as the KNN model provided the most accurate predictions, interpretable models are desperately needed to increase awareness of factors that are most influential in predicting the occurrence of accidents. While obtaining mostly satisfactory results, none of the models was able to provide excellent results with selected predictor variables, which indicates that more extensive experimentation is required regarding feature selection. Furthermore, formulating accident occurrence prediction as a multiclass classification problem may improve accuracy and applicability of the model.

Keywords Machine Learning, Construction Industry, Occupational Safety and Health (OSH), Safety Risk Prediction, Construction Safety

Tekijä Ina Häkkä

Työn nimi Application of Machine Learning to Predict Occurrence of Accidents at Finnish Construction Sites

Tutkinto Kauppatieteiden maisteri

Koulutusohjelma Tieto- ja palvelujohtaminen

Työn ohjaaja(t) Pekka Malo

Hyväksymisvuosi 2022 **Sivumäärä** 54 **Kieli** Englanti

Tiivistelmä

Rakennusala on yksi vaarallisimmista toimialoista korkean tapaturmataajuutensa ja monien työmaalla esiintyvien turvallisuusriskien vuoksi. Lukuisat turvallisuusriskit ovat seurausta rakennushankkeiden uniikista, kompleksisesta ja dynaamisesta luonteesta, koostuen alituisesti muuttuvasta työympäristöstä, fyysisesti vaativista työolosuhteista sekä jatkuvasti kehittyvästä rakennusteknologiasta, joka saattaa luoda uusia olosuhteita virheen mahdollisuudelle sen eliminoinnin sijaan. Ponnistelut työmaiden tapaturmataajuuden laskemiseksi ja turvallisuusriskien lieventämiseksi vaativat tehostettua työturvallisuuskulttuurin, -käytäntöjen ja -välineiden tarkastelua sekä ennakoivaa työturvallisuuden johtamista, arvioiden yksilöllisten, johdollisten ja ympäristötekijöiden vaikutusta työturvallisuuteen.

Tekoälyn hyödyntäminen diagnostisten ja ennustavien mallien kehittämiseksi nähdään olevan seuraava mullistava kehitysaskel rakennusalan työturvallisuudelle. Tämä tutkimus on suoritettu case-tutkimuksena eräästä Suomen johtavista rakennusalan yrityksistä, tähdäten yrityksen tietolähteiden laajamittaiseen hyödyntämiseen tapaturmien ilmenemistä ennustavan koneoppimismallin kehittämiseksi. Tutkimuksessa käytetty data sisältää ympäristöllisiä, johdollisia ja projektikohtaisia tekijöitä yli 600 rakennushankkeesta Suomessa. On olemassa lukuisia koneoppimista hyödyntäviä, työmaan tapaturmien ennustamiseen tähtääviä tutkimuksia, mutta valtaosalla aikaisemmista tutkimuksista ei ole ollut käytettävissään tapaturmaraportteja täydentävää dataa tapahtumista, jolloin tapaturmaa ei ole sattunut. Näin ollen, aikaisemmat tutkimukset ovat laajalti keskittyneet esimerkiksi tapaturmien vakavuuden ennustamiseen tapaturmien ilmenemisen ennustamisen sijaan, rajoittaen niiden panosta ennakoivaan turvallisuusriskien hallintaan ja tapaturmien ennaltaehkäisyyn. Yksi tämän tutkimuksen merkittävimmistä vahvuuksista on kyvykkyys yhdistää dataa tapaturmista hetkiin, jolloin tapaturmaa ei ole tapahtunut, tarjoten koneoppimismalleille ennakoivaa luokittelemista varten tarvittavia esimerkkejä eri tapahtumista työmailla.

Löydökset osoittavat, että vaikka tulkitsemattomat mallit kuten KNN saattavat olla päteviä kyseisen tutkimusongelman ratkaisemiseksi, tulkittavia malleja tarvitaan erityisesti tapaturmien ennustamiseen vaikuttavien tekijöiden ymmärtämiseksi. Mallien tulokset olivat enimmäkseen tyydyttäviä, mutta yhdenkään mallin tulokset eivät yltäneet erinomaiselle tasolle käytössä olleiden selittävien muuttujien avulla. Tämä osoittaa, että laajempia kokeiluita tarvitaan selittävien muuttujien valinnassa. Lisäksi useamman kohdeluokan sisällyttäminen saattaisi lisätä mallin ennustetarkkuutta ja sovellettavuutta.

Avainsanat Koneoppiminen, Rakennusala, Työturvallisuus, Turvallisuusriskien ennustaminen, Työturvallisuus rakennusalalla

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Background and motivation

Construction industry is of great significance when it comes to securing the premises for a functional society. It contributes substantially to economic growth through building, infrastructure and industrial construction as well as real-estate development, creating many jobs along the way. However, construction industry is unfortunately also one of the most dangerous industries due to many fatal accidents at workplace caused by safety hazards on the construction site (Hämäläinen, Takala and Saarela, 2006; Mesaros, Spisakova and Mackova, 2019; Pinto, Nunes and Ribeiro, 2011). Worksite accidents and poor occupational health cause intolerable humanitarian and economic burden, massively taking a toll on sustainable economic growth and a company's reputation (International Labour Organization, 2014).

The physical environment on construction sites is constantly changing due to the complexity and dynamic nature of construction projects, often creating coordination challenges, which explains the premises for frequent worksite accidents in construction. Moreover, physically demanding working conditions, often including work at great heights or working outdoors in all weather conditions, contribute to high occupational safety risks in the industry, therefore requiring more effective safety management (Rozenfeld et al, 2010; Törner and Pousette, 2009). According to the International Labour Organization (ILO), the construction industry has a disproportionately high rate of recorded accidents worldwide. Each year over 60 000 fatal accidents take place on construction sites around the world, and although the construction sector employs between 6% and 10% of the workforce in industrialized countries, it is responsible for between 25% and 40% of work-related deaths (International Labour Organization, 2005). In fact, construction workers are three to four times more likely than other workers to die from accidents at work according to data from several industrialized countries (International Labour Organization, 2009).

In Finland, the frequency of accidents on construction sites has continuously decreased over the last few decades according to statistics (Confederation of Finnish Construction Industries, 2022; Finnish Centre for Occupational Safety, n.d.; Lantto and Räsänen, 2019). Furthermore, severe and fatal accidents on worksites in the construction industry have likewise taken a downward trend over the last few decades, although there is annually a lot of variation in the number of fatal accidents on construction sites per year (Confederation of

Finnish Construction Industries, 2022; Lantto and Räsänen, 2019). While construction industry is the most dangerous compared to other industries suffering from high frequency of accidents at work, occupational safety has also progressed most significantly in the construction sector (Lantto and Räsänen, 2019). Some of the most prominent causes behind this change are a positive shift in safety culture and attitudes towards occupational safety in the construction industry, improved development and use of protective equipment and the efforts of construction companies to improve occupational safety (Lantto and Räsänen, 2019).

However, there is still a lot that could be done to improve the safety and accident prevention on construction sites. Industry professionals believe that majority of the worksite accidents happening to this day, if not all, could be prevented (Immonen, 2015; Leskinen, 2018; Loukkola, 2013), which has sparked initiatives for zero worksite accidents in construction sector and the ideology that all worksite accidents are unacceptable, unnecessary and preventable (Confederation of Finnish Construction Industries, 2015; Lantto and Räsänen, 2019). Moreover, according to the International Labour Organization, 'those who sustain their efforts to prevent occupational accidents and diseases will find their efforts rewarded' (ILO, 2014, p. 1). This is highlighted by the fact that poor occupational safety and health (OSH) practices and substandard safety culture cause significant economic losses for enterprises and societies, such as lost productivity and decreased work capacity, which contribute to the estimated loss of four percent of the world's gross domestic product (GDP), lost due to various direct and indirect costs like compensation, medical expenses, property damage, lost earnings and replacement training (ILO, 2014).

The urgency and need for accident prevention have been widely recognized in the industry. At one of Finland's most prominent construction companies many measures have been taken and new technology has been implemented to improve work safety on construction sites. Nevertheless, there is still a lot of room for improvement especially in real-time risk-assessment and proactive risk-mitigation to ensure construction workers' safety. According to a report published by Finnish Institute of Occupational Health (TTL), the use of AI to develop diagnostic and predictive models may be the next revolutionary advancement for occupational safety in the construction industry (Lantto and Räsänen, 2019).

1.2 Research problem

This research is a case study focusing on utilizing data from a Finnish construction company's worksites and projects to build a predictive machine learning model for worksite accident detection. The model should take into consideration many factors, such as number and nature of safety hazards reported at a worksite (e.g. equipment lying around, slippery scaffolding etc.), project-specific, managerial and environmental factors.

RQ1:
What are the most influential factors when predicting the occurrence of an accident at a construction site?

RQ2:
Which Machine Learning approach is the most suitable for predicting the occurrence of accidents in the context of this research?

1.3 Research objectives

Based on current information collected on several attributes, the predictive model should be able to classify new datapoints as accidents or non-accident cases or it should, at the very least, give new insight on which factors contribute to emerged incidents at worksites.

The case company should gain new, actionable insights to help further develop and improve safety at worksites, attending to areas of special interest detected in the research. Furthermore, the study should give recommendations on improving data collection processes and quality of data as well as the model itself, making the solution more sustainable and actionable in the long run.

2 Literature review

This section will cover some of the most relevant research and literature as theoretical background to this study.

2.1 Occupational safety and health (OSH) in construction industry

As this research mainly focuses on improving occupational safety, it should be established what safety and health encompass and, furthermore, how occupational safety is defined. One way to distinguish safety and health is that safety is associated with injury-causing situations or hazards to humans caused by sudden severe conditions, while health deals with risk of disease or adverse reactions to exposure to toxic or otherwise dangerous hazards (Brueggmann, Roetting and Luczak, 2001; Goetsch, 2019). The Joint ILO/WHO Committee on Occupational Health adopted a comprehensive definition at its First Session (1950), later revised at its Twelfth Session (1995), which states that occupational health should "aim at the promotion and maintenance of the highest degree of physical, mental and social well-being of workers in all occupations" (International Labour Office, 2003). However, the most common title for this field nowadays is Occupational Safety and Health (OSH, also commonly referred to as OHS). Terms occupational health and occupational safety are often used as synonyms for the field and, therefore, OSH or OHS is likely the most comprehensive and descriptive title for it.

The definition and scope OSH have kept evolving over the years: as a field, it is multidisciplinary as it is concerned with the topics of safety, health, and welfare of people at occupation. With gradual changes and continuous development, the areas of concern for OSH have expanded beyond the workplace, through economic and national concerns, ultimately reaching the global arena and examining environmental concerns (International Labour Office, 2003). Other modern definitions of OSH agree with its far-reaching scope and practice. Friend and Kohn (2018) describe OSH as a field aiming at loss minimization through efforts towards preservation and protection of human and other physical assets in the workplace, addressing moral and economic issues while providing companies with important means of survival in the increasingly competitive global economy.

Hale and Hovden (1998) have divided the emergence and development of OSH into three distinct 'ages' in history. Hale and Hovden (1998) describe how OSH related concerns began to surface in response to fundamental changes in agricultural, industrial and manufacturing processes in consequence of the industrial revolution, requiring development of engineering and technological solutions to new hazards at workplace. According to Hale and Hovden (1998), limitations of purely technological approach became apparent, and thus the second age of safety included a greater focus on human factors. However, towards the end of the twentieth century, researched introduced and demonstrated the influence of management and organizational factors in OSH outcomes, which caused another shift in the focus of OSH. Therefore, Hale and Hovden (1998) outline the implementation of OSH management systems to be the focal point of the third age of safety.

The discoveries of Hale and Hovden (1998) are also linked to development and technological advancements in the construction industry, requiring a new set of standards and measures for safety. Dekker (2002) states that advancements in technology do not eliminate the possibility for error, but rather just relocate or reform the potential error. In his work he found that implementing new technology often increases the operational pressure, introducing new requirements and expectations that organizations impose on employees since the motivation behind investing in new technology often lies in desire to increase efficiency: to achieve more with less time or fewer resources (Dekker, 2002). This further supports Dekker's finding that people are constrained by factors more or less outside of their control, including time constraints as well as pressure to produce and perform highly, which lead to actions and choices that contribute to possibility of accidents caused by human error. Moreover, Dekker (2002) stresses that human error is a symptom of trouble deeper inside a system rather than the actual cause of accidents, and that it is vital to understand how the surrounding circumstances affect and steer people's assessment and actions at the time. Likewise, Hansen (2001) emphasizes the cruciality of identifying and eliminating the actual underlying root causes in order to reduce human errors since a variety of circumstances determines whether the people classify as actively learning, highly motivated employees or as resistant, non-participative and unhappy workforce.

The nuanced discussion on the impact of the human element is extremely relevant in the construction industry as several studies have recognized human-related factors to be the most significant cause behind worksite accidents (Haslam et al., 2005; Heinrich, 1959; Jaafar et al., 2017; Petersen, 1984). Before Heinrich's studies and theory presented in Industrial Accident Prevention (1959), practitioners of industrial safety were heavily focused on

physical aspects behind accidents, such as machinery and equipment, however, Heinrich's revolutionary concepts at the time permanently changed the focus of discussion in occupational safety (Petersen, 1984).

Jaafar et al. (2017) argued that due to the uniqueness, complexity and hazardousness of construction projects, the importance of sustaining an effective OSH management program for accident-prevention is emphasized. Furthermore, Jaafar et al. (2017) introduced a framework of causes of occupational accidents and illnesses in the construction industry as a central tool for identifying the source of the industry's safety hazards, helping with comprehending the safety risks that need to be addressed and managed. According to the framework, causes of occupational accidents and illnesses in the construction industry can be divided under four main categories; human elements, worksite elements, management elements and external elements, and these elements are further divided into multiple levels of more detailed subcategories describing the causes of occupational accidents. Jaafar et al. (2017) concluded that human and worksite elements are considered the most immediate causes of occupational accidents. Human element related causes consist of factors related to human physique, experience, attitude and behavior, whereas worksite element related causes of occupational accidents include variants of worksite conditions, poor site management, demanding tasks as well as equipment and materials (Jaafar et al., 2017).

Likewise, other studies highlight the importance of preventative approach to safety risks in construction industry. Overall, the construction industry has demonstrated major improvement in safety since establishing the Occupational Safety and Health Act in 1970 (Hinze et al., 2013). Nevertheless, due to the passive or reactive nature of the compliance practices, they fail to provide early warnings detected from the unique characteristics of a worksite. Thus, the emphasis of developing occupational safety in the construction industry should be shifted towards proactive safety strategies that are able to detect the risk of potential hazards in advance due to competence in identifying precursors of incidents (Esmaeili, Hallowell and Rajagopalan, 2015). Furthermore, Tixier et al. (2016) discovered that occurrence of construction injuries is not random but rather can be explained by underlying patterns and trends, which was further supported by discoveries by Poh, Ubeynarayana and Goh (2018).

A significant tool for preventative, proactive OSH management and comprehensive safety performance measurement in the construction industry has been incorporating leading and lagging indicators. Robson et al. (2017) defined lagging indicators (also known as trailing indicators) as measuring significant safety outcomes, such as injury frequency,

whereas leading indicators focus on measuring workplace activities, conditions, and events that may obtain significance in predicting, determining or contributing to safety outcomes, e.g. frequency of worksite inspections. Furthermore, Hinze, Thurman and Wehle (2013) distinguished two categories in leading indicators: passive and active. According to Hinze, Thurman and Wehle (2013), passive indicators give an implication of expected safety performance in the long-term and cannot usually be affected in a short period of time, while active indicators may be subject to change in a short period of time and thus enable prompt corrective actions. Toellner (2001) established that both leading and trailing indicators are required to obtain a holistic view of worksite's safety performance while allocating resources to preventative actions and, moreover, safety profession should aim at both short- and long-term improvement of safety performance.

Finally, when assessing the state of prevailing approaches with regard to the future of improving safety in the construction industry, Bhattacharjee et al. (2011) argued it should be considered how mitigation of occupational safety and health hazards and risks could commence with architects and designers prior to launching the actual construction phase. Swuste et al., (2020) stated that a dramatic decrease in accidents in the construction sector would be plausible by clarification of safety-related responsibilities and engaging construction designers to consider the actual reality of construction sites or involvement of construction workers in the design process. Zhou, Goh and Li (2015) discovered in their studies six research gaps in safety management studies in the construction industry including lack of hazardous behavior monitoring, deficiency in incorporating safety climate to accident prediction, unfamiliarity of quantitative relationship between project scale and construction safety, lack of construction safety research at the task level, excessive focus on construction phase and traditional building projects, and lack of innovative technology applications in construction safety practice. To a degree, safety performance is connected to the organization's ability to influence events and control what happens in the future and will therefore be leveraged for strategic advantage (Young, 1996).

2.2 Application of machine learning in predicting construction safety

Over the course of more than two decades that construction research has considered Machine Learning, many of the existing safety risk analysis techniques are limited due to inadequacy in considering the dynamic nature and complexity of construction sites and utilizing

empirical data (Tixier et al., 2016). Poh, Ubeynarayana and Goh, (2018) argued that machine learning (ML) approaches are not utilized widely enough in the construction industry, especially in the development of validated leading indicators for classifying construction sites' safety risk levels which the industry is currently lacking. According to Zhou, Goh and Li (2015), research regarding construction safety has had an overriding emphasis on analysis of lagging safety indicators. However, machine learning approaches introduced in this section concentrate on studies aimed to contribute to more proactive management of construction safety since they are more appropriate considering the scope and aim of this research.

Study conducted by Poh, Ubeynarayana and Goh (2018) aims at improving construction companies' project safety risk forecasting capabilities by using a ML approach to predict occurrence and severity of accidents, providing a validated leading indicator to measure construction safety. Results from the study imply that occurrence and severity of accidents do not happen randomly. Moreover, Poh, Ubeynarayana and Goh (2018) discovered that safety-related indicators need to be supplemented with project-related factors for effective prediction of construction accident occurrence and severity. Results from application of feature selection techniques in the study indicated the importance of effective project management to construction safety since six out of 13 selected input variables were project related. Additionally, although some project characteristics remain unchanged across time, ML models applied in the study take them into account when predicting the target variable. This may explain why some previous studies, including research by Tixier et al. (2016), have demonstrated absence of predictive skill for injury severity as an output variable.

Tixier et al. (2016) used Random Forest (RF) and Stochastic Gradient Tree Boosting (SGTB) machine learning algorithms to predict construction safety outcomes with essential construction attributes. The data utilized in the study was extracted from a wide collection of injury reports, and the models provided probabilistic forecasts to each degree of safety outcomes. Furthermore, the purpose of the models is to give an indication of which preventative or corrective measures to take, thus reinforcing proactive safety management. Although the models predicted injury type, energy type, and body part with high skill, Tixier et al. (2016) failed to obtain consistent results for forecasting injury severity, indicating that it is mainly random. Other studies have deemed this to be false (Baker, Hallowell, and Tixier, 2020; Poh, Ubeynarayana and Goh, 2018; Zhu et al., 2020).

In later research by Baker, Hallowell, and Tixier (2020), prediction of injury severity as an outcome has undergone notable advancement, comparing the performance of Random Forest, XGBoost and linear SVM as well as combinations of these through model stacking. Nevertheless, major limitation regarding the study's contribution to proactive safety management concerns the fact that prediction of the outcome of an incident made with this approach is conditional to the occurrence of an accident. Baker, Hallowell, and Tixier (2020) declare that incident occurrence prediction (i.e., success or failure) is a fascinating research topic that certainly has its place in research of construction safety prediction. However, striving for making such predictions would require access to non-injury reports, i.e., success cases to be used in training machine learning models. Baker, Hallowell, and Tixier (2020) argue that reports of near-misses or safety hazard observations are not enough nor necessarily desired for representing non-accident incidents (i.e. successes) since they still correspond to safety issues. Hence, notes of events in the construction environment where safety was at a commendable state are required according to Baker, Hallowell, and Tixier (2020).

Choi, Gu, Chin and Lee (2019) applied machine learning to identify the potential risk of fatal accidents at construction sites while comparing predictive performance of various machine learning models. The study utilizes publicly available national data from the Republic of Korea, however, due to Korean data protection laws the data lacks non-accident instances, which limited the scope of the research to predicting the likelihood of fatality in the event of an accident. Similar limitations due to absence of records on non-accident events on construction site exist in several other studies as well (Ayhan and Tokdemir, 2018; Baker, Hallowell, and Tixier, 2020; Zhu et al., 2020). After trialing numerous different machine learning prediction models, Choi, Gu, Chin and Lee (2019) found the highest prediction rate with Random Forest classification technique. The analysis revealed most significant independent variables to comprise of month, weekday, project size indicated by number of employees, employee age and employee service length, which supports Poh, Ubeynarayana and Goh's (2018) argument on importance of project-related attributes complementing safety-related indicators. However, Choi, Gu, Chin and Lee (2019) emphasize importance of seasonal factors to be dependent on the region or country since the risk of fatal accidents varies with the climate as can be deduced from results from previous studies based on data from other locations.

Zhu et al. (2020) conducted a study trialing eight different machine learning algorithms to analyze diverse combinations of various attributes and their effect on the performance of

predicting the severity of construction accidents. The study revealed that some factors remaining consistently of high importance across several models when evaluating the impact of different factors on the severity of accident were ‘Type of accident’ and ‘Safety culture’ which describes the state of safety awareness and safety knowledge or lack thereof according to data extracted from investigation reports for construction accidents. Furthermore, the results support findings from other studies (Baker, Hallowell, and Tixier, 2020; Poh, Ubeynarayana and Goh, 2018) regarding predictability of the severity of the accidents, demonstrating the predictive skill of machine learning algorithms in providing accurate forecasts of accident severity.

Ayhan and Tokdemir (2018) trialed Artificial Neural Network (ANN) and multiple regression models to predict outcome of construction incidents between five different categories from “At Risk Behavior” to “Fatalities”, describing the severity of the incident. Furthermore, the research included a decision step using the incident category as an input and thus proposing appropriate action as an outcome to help address the situation according to the severity of the incident. The study utilized construction site records with 17285 incidents, including accidents along with other types of incidents causing safety hazards at construction site. While results from the multiple regression model remained unsatisfactory, they obtained moderate success with the predictive accuracy of ANN.

As many other studies in the field applying machine learning for similar purposes, the study conducted by Ayhan and Tokdemir (2018) is limited by the available data. Since there are no records on situations at construction site where everything went well, which Baker, Hallowell, and Tixier (2020) deemed as a significant drawback in many of the recent studies in the field, the study can only provide insight on hazardousness or severity of an incident. Nevertheless, extending the model to recommend actions to be carried out based on predicted severity of the incident makes the study more applicable for taking concrete actions towards improving safety management at construction sites. However, the authors also acknowledged that the outcome of incidents may never be predicted accurately due to existence of factors other than the ones used in this study, including cultural, social, emotional, and environmental factors which are extremely difficult record in a format suitable for application of machine learning while remaining truly representative of the phenomena (Ayhan and Tokdemir, 2018). Moreover, according to findings by Tixier et al. (2016), studies on construction safety should include empirical data instead of strictly being qualitatively approached.

Abbasianjahromi and Aghakarimi (2021) developed a machine learning model to predict the safety performance of construction projects. The authors paid special attention to analyzing factors constituting the concept of safety performance, aiming at extracting the most influential criteria on safety performance to help determine the input variables to be used in training the model. The study accounted a total of over a hundred effective criteria discovered from construction safety performance, construction safety climate and construction safety culture related literature from 2013 to 2019 and reduced the number of criteria down to nine with the highest priority according to frequency of the factor's occurrence in literature and the factor's importance score provided by a panel of subject area experts. Nine most important safety performance indicating criteria discovered and utilized in Abbasianjahromi and Aghakarimi's study (2021) are presented in Table 1. The decision tree results demonstrated that specialized safety employees, training, rule adherence and management commitment were the principal criteria in predicting construction safety performance. Furthermore, Abbasianjahromi and Aghakarimi (2021) criticize many existing prediction models for safety performance of construction projects for their black-box approach, failing to provide interpretable results, i.e., information on criteria affecting the safety performance, which makes the results unactionable.

Table 1: Nine safety performance criteria with the highest priority identified in research by Abbasianjahromi and Aghakarimi (2021)

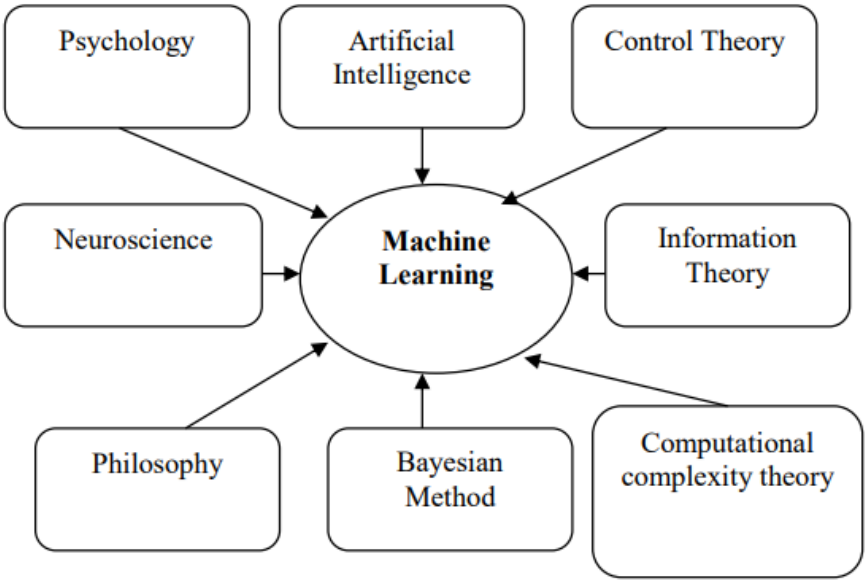
Criterion	Description
Management commitment	Management commitment to incorporating safety as a priority in projects
Safety rules	History of safety rule adherence among workers and attempts to properly implement safety rules in previous projects
Safety equipment	History of utilizing safety equipment in previous projects
Training	History of worker training to improve safety familiarity
Safety investigation	A regulated procedure in place to investigate project safety
Safety rules	A set of safety instructions and rules
Safety employees	Safety experts supervise safety and hazards in activities
Hazard awareness	Project workers' awareness of possible hazards
Safety budget	Budget assigned to safety management measured as a portion of the total project budget

Kang and Ryu (2019) proposed a random forest approach to classify and predict occupational accidents on construction sites, obtaining influential attributes to be used in the model via feature significance. The data of fatal and non-fatal accident reports included detailed information on total of 55 input variables, including age, work contents and unsafe behaviors. Results on feature importance from the study are considerably aligned with what Abbasianjahromi and Aghakarimi (2021) discovered: unsafe behaviors of employees (e.g., failure to comply with safety rules) and protective equipment (types of safety equipment worn by the employee) were some of the key factors in classifying construction accidents, and main reasons for majority of the accidents were lack of or deficiency in wearing safety equipment, lack of installed protection facilities, and missing footing on stairs or structures. Moreover, Kang and Ryu (2019) discovered that several accidents resulted from risky or incorrect handling or mis-operation of machinery, tools or other equipment as well as disorderly working conditions or other defects in work environment or gear, such as faulty structures or improperly installed protective devices. The aforementioned safety risks can be, to some extent, linked to lack of training or hazard awareness, some of the most important safety performance indicators according to Abbasianjahromi and Aghakarimi (2021). However, Kang and Ryu (2019) found season- and weather-related factors, such as month, temperature, humidity and wind speed, to be mostly insignificant, which contradicts the results obtained by Choi, Gu, Chin and Lee (2019) in their research.

2.3 Machine Learning methods

Ever since the establishment of the era of computers, researchers have strived for embedding human-like learning capabilities and processes in computers, making this pursuit the most demanding long-term goal of artificial intelligence (Michalski, Carbonell and Mitchell, 1983). Artificial intelligence (AI) is a wide-ranging field that merges several topics including but not limited to probability, logic and continuous mathematics as well as perception, reasoning and learning (Russell and Norvig, 2010). Machine learning (ML) is a category of artificial intelligence and, furthermore, a computational method that utilizes past information typically in the form of electronic data in order to improve performance or make accurate predictions (Mohri, Rostamizadeh and Talwalkar, 2018). Although described as a type of AI, machine learning can also be perceived as a multidisciplinary field of study combining theories from several research domains (Alzubi, Nayyar and Kumar, 2018), as displayed in Figure 1.

Figure 1: The Multi-disciplinary ML (Alzubi, Nayyar and Kumar, 2018)



2.3.1 History of Machine Learning

One of the first pioneers in machine learning was Arthur Samuel who in 1952 wrote a computer learning program for the game of checkers that quickly learned to become a more advanced player than its creator by studying moves constituting winning strategies and incorporating them into the program (Samuel, 1959; Russell and Norvig, 2010). Moreover, emergence and early history of machine learning is often associated with psychologist Frank Rosenblatt for he created one of the earliest applications of machine learning, a machine known as the ”perceptron” (1958) with the purpose of recognizing the letters of the alphabet, also perceived as a prototype of modern artificial neural networks (ANN) (Fradkov, 2020). However, due to primitive nature of computers in the mid-1900s, studies and development of machine learning during its early age were limited to theoretical level or experimental hardware systems developed for performing a specific task (Michalski, Carbonell and Mitchell, 1983; Russell and Norvig, 2010).

Shavlik and Dietterich (1990) divided the emergence and early development of machine learning into three periods in time consisting of exploration in the 1950s and 1960s, followed by development of practical algorithms in the 1970s and explosion of research directions in the 1980s. In the 1950s and 1960s, work on machine learning was heavily focused on single-task-oriented programs, such as simple symbol-processing systems, along with theoretical analysis of machine learning that was primarily influenced by psychological, neurophysiological, and biological research (Shavlik and Dietterich, 1990). In the 1970s,

machine learning evoked unforeseen interest with several new methods being proposed, and publication of P.H. Winston's dissertation on blocks-world learning (1970) can be perceived as one substantial instigator for the spiked interest in machine learning (Langley, 1996; Shavlik and Dietterich, 1990). Machine learning proceeded branching out in the 1980s with spread availability of various robust software packages in the community which enabled conducting comparative studies and building on previous systems (Langley, 1996).

Some of the most significant advancements in machine learning have been made over the past two decades, which is indicated by the rapidly spread commercial use of machine learning techniques (Jordan and Mitchell, 2015). Furthermore, Russell and Norvig (2010) discussed in their work how the emphasis of the studies regarding AI has recently shifted from examining the algorithm towards focusing more on the data due to increasing availability of very large datasets. From managerial perspective, this can be seen as vast expansion of data-driven decision-making within organizations due to drastically evolving data storage and processing technologies that have created new opportunities to collect and leverage data, making it available to managers (Brynjolfsson and McElheran, 2016).

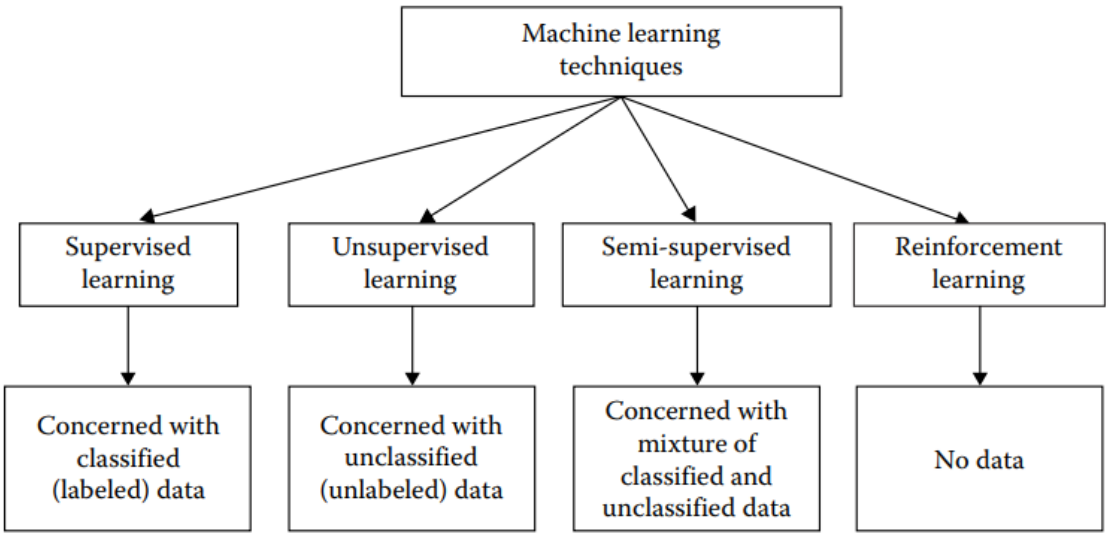
2.3.2 Differences between models

According to Alzubi, Nayyar and Kumar (2018), machine learning approaches can be categorized based on types of problems they are capable of solving, and the categories of describing these types of problems are classification, anomaly detection, regression, clustering and reinforcement problem. Classification problem represents a situation where classes for potential outcomes are know in advance, anomaly detection problem aims at detecting changes or anomalies in a certain pattern, regression problem attempts to explain continuous, numeric output with a set of predictor variables, clustering is utilized for forming clusters based on the similarity in the structure of data and reinforcement algorithms rely on past experiences of learning to make a decision (Alzubi, Nayyar and Kumar, 2018).

Another common way of differentiating machine learning algorithms is according to their learning scenarios, usually divided to supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning, which are defined by the training data available to the learner as depicted in Figure 2 (Mohammed, Khan and Bashier, 2016). In addition to these conventional machine learning scenarios, Mohri, Rostamizadeh and Talwalkar (2018) introduced in their work learning scenarios of transductive inference, on-

line learning and active learning while recognizing the possibility of encountering many other intermediate and potentially more complex learning scenarios in practice.

Figure 2: Different machine learning techniques and their required data (Mohammed, Khan and Bashier, 2016)



3 Data and research methodology

This section will introduce the data provided by the case company, assessing its sufficiency, reliability and feasibility for the application of machine learning in the context of this research. The source of the data and the processes or circumstances in which the data is produced will also be covered in this section. Beneficial proposals for improving the quality and usability of the data will be introduced and discussed iteratively as well as in the conclusion chapter.

After introducing the data sources and the data itself, the data will be observed through exploratory data analysis and visualizations to get a better view on its distribution across different values and development over time. Since data is being imported from several different tables, gaining an understanding of what kind of data each table contains must be achieved before combining the tables and preprocessing the merged data. Exploratory data analysis will also uncover deficiencies and errors in the data that must be eliminated before the application of selected Machine Learning methods.

Once each table has been studied on its own, the data can be combined and processed in such way that is appropriate for this study. In addition to various options for attribute selection, there are several ways to combine and structure the data to form a dataset that enables predictions made by Machine Learning, which will be assessed in this section.

3.1 Data introduction

Majority of the data used in this study is collected through a field tool used by construction personnel on the construction site. The field tool is a mobile application that is easily and timely accessed through a mobile devise, and it is in daily use for reporting work accidents and safety observations at each of construction sites across all the case company's projects. However, the information that is reported depends on the incident type (accident or safety observation), and thus the data is stored into several tables with different attributes. Moreover, the granularity and frequency of data varies between different projects depending on their size, duration and complexity.

The case company has enforced safety measures, processes, policies and KPIs that encourage and oblige employees and subcontractors to active usage of the field tool on a daily basis. The usage of the field tool is perceived as a central instrument for proactive and

preventative safety measures on construction sites, not only in the case company but as a standard practice in the industry in all of Finland. Its current and future potential for safety related data collection on construction sites is at a commendable level due to accessibility, level of detail and continuous support and feature development by the supplier.

Nevertheless, it is also important to note and consider how variation in management, safety culture and other factors on worksite affects the usage of the field tool and, furthermore, quality and frequency of data. If and when this type of variance in the data appears that cannot solely be explained by predetermined characteristics of the project such as size (described by number of personnel or budget) and duration, it will add up to relevant findings about the case company's internal safety culture. While further discussion about the case company's safety culture remains outside the scope of this research, the data may indicate essential points of reflection for the case company.

3.1.1 Accidents table

Accidents table contains data of each occupational accident causing some type of physical harm to an employee or subcontractor. Accidents that occur on the case company's Finland-based construction sites are mostly reported through the field tool mobile application. This table contains information about what will later be defined as the target variable, the occurrence of accident per project, to be predicted by other variables. Each row represents an occurrence of a new accident, containing information about the time and circumstances during which the accident took place, as well as descriptive data of the consequences and severity of the accident.

As can be detected from Table 2, many of the variables in the table have little use in predicting occurrence of accidents. However, many of the columns contain useful information for further research, such as refining the models to predict consequences like severity of the accident and, thus, more accurate cost of the accident. Variables that could have some importance in predicting the occurrence of accidents cannot be utilized at this time since they are not reported actively enough. For example, employee experience has not been recorded at all in the accidents table, and information on physical activity performed by the employee during the moment that led to the accident is only reported when an accident occurs, so there is no recorded physical activity for the class of "no accident".

The column MTI, medical treatment injuries, defines the accidents that are of interest for this study, being in the circle of influence for the case company. However, some rows

implied that the specifications for the column labelling are slightly outdated since some rows had incorrect labels and thus were falsely left outside of further inspection. Another major defect of this table is that an accident or a row is not assigned to a project through PROJECT_WID but Cost Object, that can be project-specific code or a code that refers to a higher organizational level. On some rows, the cost object selected by the reporter referred to business unit or division, not a specific project, which made it impossible to trace which project the accident belongs to. This error in process should be fixed in the system so that a reported accident must always be connected to a project.

Some potentially significant data enrichment for the accident table could be achieved by supplementing the information regarding the employee affected by the accident. Columns 'Experience' and 'Employee start date' do not contain enough non-null values to provide sufficient information to Machine Learning models. Moreover, there is no employee ID available that could connect the employee from the accident table to other tables containing information on the employee's worksite behavior, such as safety observation reporting frequency.

Table 2: Description of the contents of Accidents table

Variable name	Description	Variable type
ACCIDENT_WID	Unique ID for the specific accident in question	Unique ID
DATE_WID	Reporting date of the accident	Date
EMPLOYEE_TYPE_WID	Indicates whether the injured employee is internal or a subcontractor	Binary
BUSINESS_UNIT_WID	Indicates which business unit the project reporting the accident belongs to	ID
PART_OF_BODY_WID	Describes which part of body the injury has affected (head, torso, back etc.). The dimension includes a more specific location of the injury as a subcategory.	Categorical (ID connects to dimension)
INJURY_TYPE_WID	Information about the injury type, divided to several different categories (Bone fractures, Dislocations, sprains and strains, Concussion and internal injuries…)	Categorical (ID connects to dimension)

MODE_OF_INJURY_WID	Describes the circumstances of the accident (Fall of person, Slipping, Breakage of material, machine, object…)	Categorical (ID connects to dimension)
PHYSICAL_ACTIVITY_WID	Information on physical activity performed by the employee during the moment that led to the accident (Movement, Handling of objects, Operating of machine…)	Categorical (ID connects to dimension)
Cost Object	ID that allocates the accident to the right project	ID
Accident date and time	Date and time of when the accident occurred	Datetime
Accident Type	Indicates whether the accident has occurred on worksite or during commute	Categorical
Lost-Time Injury	Indicates whether the accident has caused loss of workdays for the employee	Binary
Injuries	Indicates how many injuries the accident resulted in	Continuous (int)
Lost Days	Number of lost workdays, i.e. duration of sick leave	Continuous (float)
Circumstance text	Specifies whether accident has happened e.g. at work, while performing work-related tasks, on break or outside of worksite	Categorical
Employee start date	Indicates since when the person has been employed by the case company	Date
Experience	Should describe the length or depth of the employee's work experience but consists of only NaN values	NaN
Sick Leave Estimation	Preliminary estimate of sick leave based on severity of the accident	Categorical
MTI	Column based on accident type and circumstance text that defines whether the accident is a Medical Treatment Injury (MTI) whose occurrence can/could be directly influenced by the employer (on worksite or while performing work-related tasks)	Binary

HasDuplicate	Whether the accident has duplicate rows based on ACCIDENT_WID	Binary
Granularity	Indicates if the reported accident time is accurate to the minute, day or month	Categorical
Is LTIF Accident	Indicates whether the accident has caused loss of workdays for the employee	Binary
dv_source	States the source system to which the accident was reported	Categorical

3.1.2 Project table

Project table contains information on multiple characteristics and background information on each of the case company's projects, whether in progress or completed. The table consists of 33 columns, and the information provided in the table enables to distinguish projects located in Finland from those outside of the country to align the data with the scope of this research. Several projects lack country code but can be traced to Finland with business unit code that is in use for Finnish projects.

The most relevant information gained from project table in addition to project country location includes project segment, project start date and scheduled construction completion date, which can be used to calculate estimated project duration at least until the end of construction phase. When combined with other tables, the construction phase duration can be used to calculate project completion rate (%) at the time of reported incident, which gives some insight of current time pressure at worksite that can be expected to increase towards the end of the construction phase.

As for project segment, each segment has their own characteristics that widely affect the planning and execution of the project. For example, each infrastructure project is unique due to vast variety of types of infrastructure (bridge, road, railroad, water infrastructure etc.), and thus cannot be planned and scheduled by the same rules as typical housing projects, which adds to the complexity of the construction phase. On the other hand, many housing projects include working in great heights which adds to safety risk factors at worksite. Therefore, due to divergence between projects of different segments it can be expected that project segment as an attribute is of great importance when predicting occurrence of accidents at worksite.

Information in project table could be further enhanced by including number of employees and subcontractors at worksite, project budget, physical size of construction site or actual building in square meters or height etc.

3.1.3 Observations table

Site notes from the construction site are collected through the field tool to a table of observations, where each row is a new note regarding an observation on the worksite. The site notes can be related to either safety observations conducted on a daily basis while performing regular work tasks or traversing the worksite, or they can be observations related to standardized routine safety and quality measurements performed by internal and external parties every few weeks. Note type ID distinguishes safety observations from site notes related to routine measurement checks, which have their own table containing the actual measurement results.

There are three types of observations, all of which are important to this study. 'Task' type observations are focused on factors causing safety risks at the worksite, such as equipment lying around, missing safeguard, risk of falling etc. 'Positive' type observations record notes on commendable circumstances at worksite, such as neatness and orderliness, clearly indicated passages and correct usage and maintenance of safety equipment. 'Nearmiss' observations are reported in the presence of serious safety hazards that are extremely likely to cause or nearly have caused an accident when encountered by a human. Examples of these can be falling of a heavy element or incautious operating of equipment or machinery, causing a serious safety hazard to other employees.

The occupational safety experts at the case company have previously discovered that activeness in reporting observations, whether positive or hazardous, can positively affect a project's safety performance by reducing the number of accidents. Additionally, it has been detected by the experts that a high ratio of positive safety observations to observations reported of safety hazards can actually indicate a higher risk of accidents in some cases. This may be either due to failure to observe many of the existing safety hazards or desire to report in such manner that gives an overly positive impression of the worksite conditions and safety management at the worksite. Therefore, it should not be assumed that the type of observation and number of certain types of cumulated observations provide unambiguous information about a worksite's safety.

Other relevant values from the observations table are when the note was reported and when it was actually observed since delay between these two time points can be an indication of defective safety reporting processes or shortcomings in other processes that affect reporting promptness of safety observations. The table also displays when the observation was modified through the field tool, which for majority of the observations tells when the status was changed to 'Accepted', meaning that required measures have been completed and inspected to correct the reported deficiencies in safety. The time between when the note was created and when its status was changed implies how quickly the worksite reacts to rectify the circumstance causing a safety hazard.

As Tixier et al. (2016) suggested in their research, further studies could aim at including of "non-accident" cases as well by, for example, recording random attribute-based observations of worksite conditions rather than making predictions solely conditional on the occurrence of an accident. This could enhance what could be learned from attribute and outcome data. Hence, data on daily observations at worksites is of great significance for this research.

3.1.4 Management walks and talks table

Management walks and talks table contains information about managerial visits performed to worksites. Management walks and talks are part of the case company's safety KPIs for 2022 and set target is 12 walks per director per year. As Toellner (2001) suggested, management safety walks can be a significant indication of management commitment to safety culture and performance while reinforcing safety hazard detection, motivating workers and leading by example. Nevertheless, there is no reference or target value for individual projects regarding how many management walks and talks should be performed each year or on the course of the construction phase. However, consistent management visits to worksites could improve safety culture by increasing accountability and enhancing communication between construction personnel and higher management.

Data regarding management walks and talks is recorded through the field tool, and the notes include general observations, agreed action points, feedback from the construction site as well as note class indicating whether note is documentation, a positive note or a task to be followed by agreed action points. These attributes may contain some useful information to machine learning models, however, extracting the needed information in a format suitable for prediction performed by machine learning models proposed for this research would

require implementation of natural language processing (NLP). Therefore, most useful information from management walks and talks table is the cumulated number of walks and talks per project per date.

3.2 Exploratory Data Analysis

3.2.1 Visualizing the data

When looking at number of accidents per year (displayed in Figure 3), it is evident that data from before 2019 is flawed. It was confirmed with the case company's occupational safety experts that number of accidents prior to 2019 should be close in figures to later years, mostly following the trend of national construction data. Most plausible explanation for the dramatic shift in number of accident between 2018 and 2019 is alterations in accident recording processes and/or systems. Nevertheless, this finding suggests data from 2018 and before should be omitted from the sample data since it does not accurately represent accidents from the time period. Notable drop in number of accidents in 2020 and 2021 results from the societal impact of corona virus, restricting number of ongoing projects and thus hindering growth since the construction sector is sensitive to economic cycles (ILO, 2021).

Figure 3: Number of accidents per year at construction sites

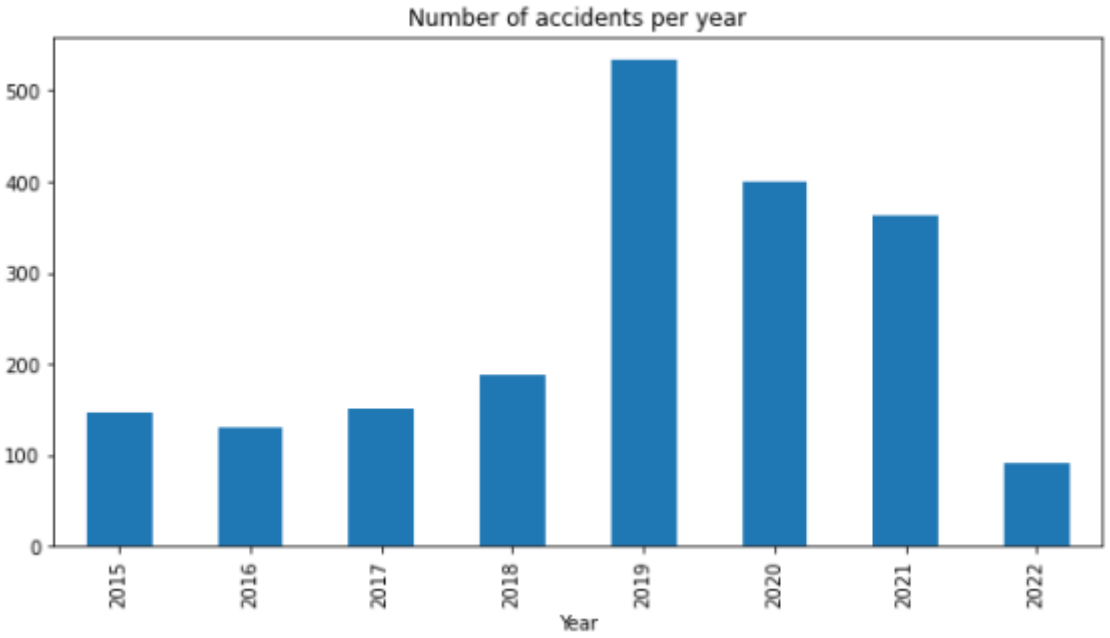
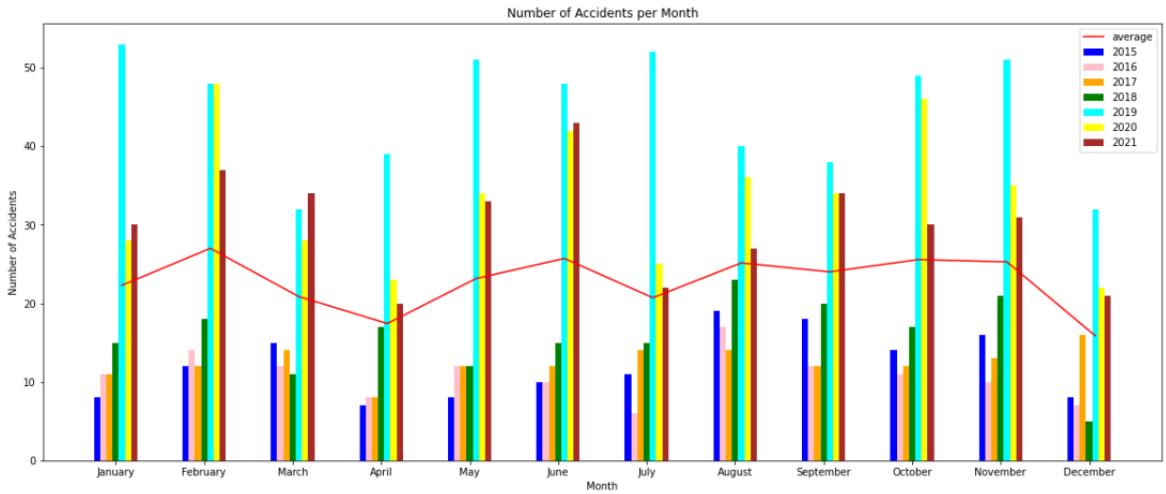


Figure 4 displays cyclical occurrence of accidents throughout the years. While years 2015-2018 lack data and, hence, remain underrepresented, their figures still demonstrate similar pattern for occurrence of accidents in different months as the rest of the data. The average

line illustrates this moderately cyclical pattern well although it is depressed by deficient numbers. There is some fluctuation around the pattern, and some years' monthly figures appear to follow the trendline better than others. Year 2022 is excluded from this graph since it contains data from only first four months of the year.

Figure 4: Number of accidents per month in 2015-2021

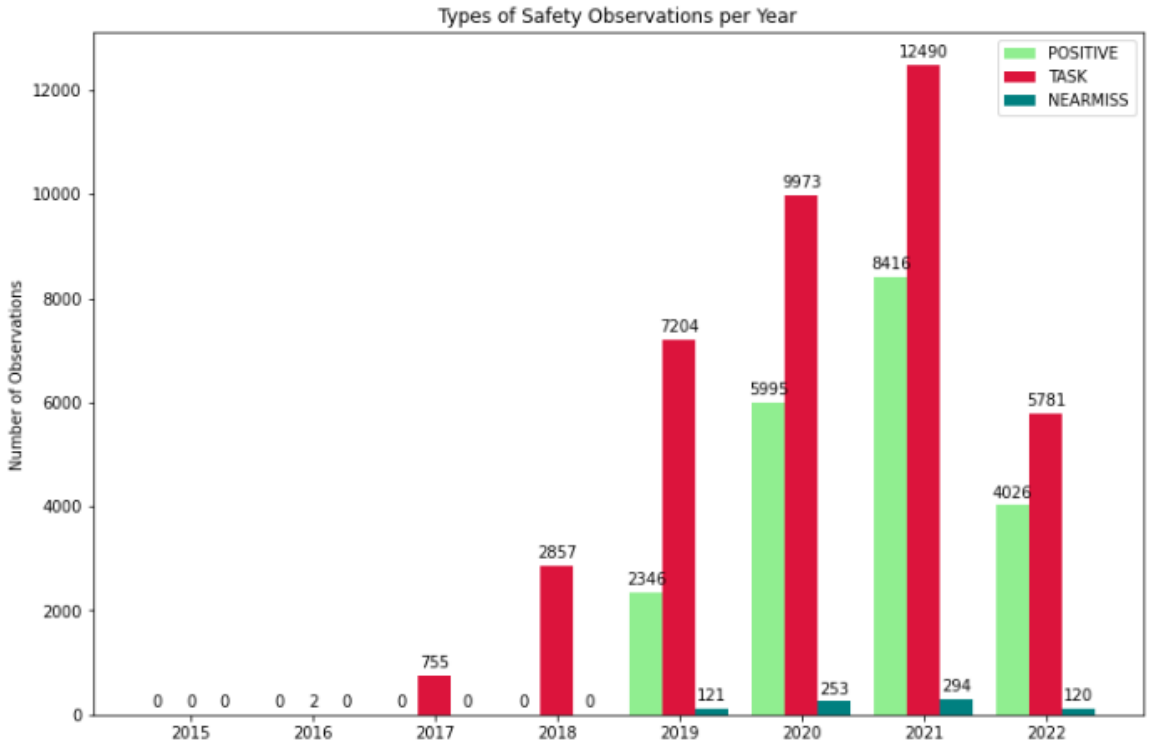


When considering construction accident data, it is relevant and essential to understand what types of actions or work activities and tasks are leading up to these accidents, although they may not be the sole direct cause of the accident. Figure 5 reveals how total number of accidents each year (2015-2021) is distributed between physical activity types preceding the accident. Beneath the vertical bar chart is displayed sum of lost days resulted from accidents grouped by their physical activity. Although physical activity is not reported prior to occurrence of an accident, it is an event preceding the accident and as such, is useful knowledge in assessing effective risk mitigation actions, especially those considering increasing the safety of daily work tasks.

Figure 5: Annual distribution of accidents between physical activity types preceding them (%) and sum of lost days per physical activity.

Safety observations represent non-accident cases, and especially positive observations are vital for this purpose since they are examples of success cases where safety conditions on worksite are close to impeccable. As can be detected from Figure 6, there are no reported positive observations or near-misses before 2019, which further supports limiting the data sample to years posterior to 2018. While it should be noted that 2022 only includes figures up to April, it seems that annual reporting activity on all safety observations has increased over the years.

Figure 6: Distribution of reported safety observations between different observation types



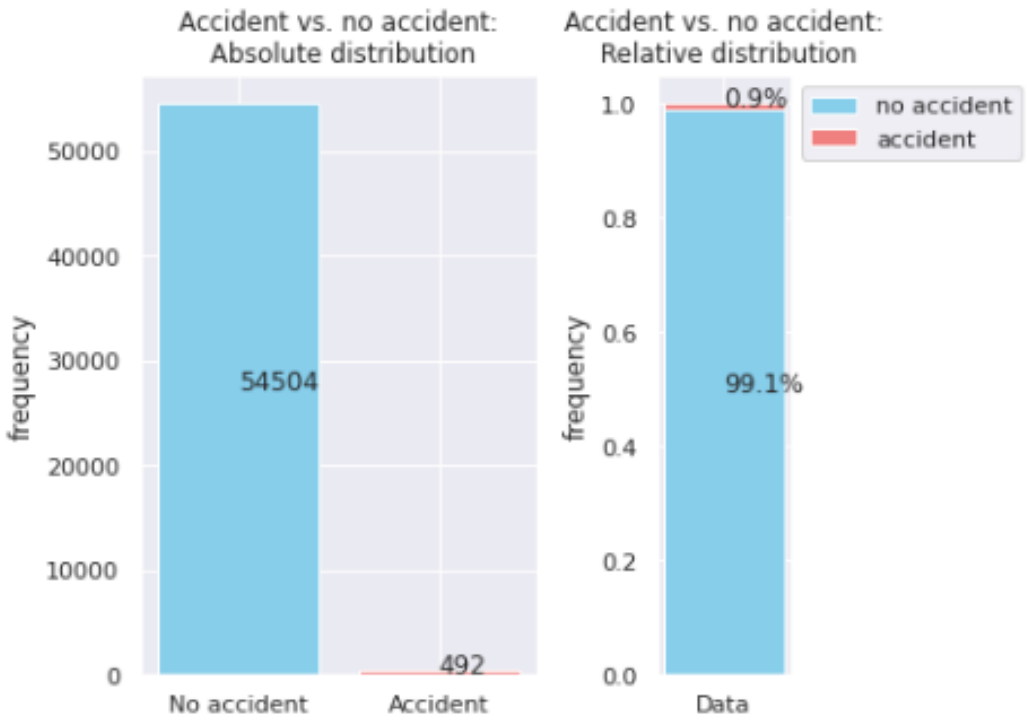
3.2.2 Processing the data

As has been established, the data used for this research consists of multiple tables that need to be combined. Accidents table will provide the binary target variable (accident vs. no accident) when combined with observations reported through field tool will, which represent non-accident cases. Additionally, counts of reported observation types per each project will be used to represent safety reporting activity but also work safety conditions on construction sites. Furthermore, reaction time and report delay to observations are considered as measures of how well safety processes are in control, and management walks and talks counts per site indicate management commitment to safety initiatives.

As Poh, Ubeynarayana and Goh (2018) suggested, project-specific factors should also be considered when making construction safety related predictions. Thus, planned project duration and project completion rate at the time of each incident represent project size and potential time pressure that is perceived to increase towards the estimated construction completion date. Project segment is also taken into consideration since complexity of projects varies between segments. Finally, seasonal factor represents weather conditions as well as changes in construction activity between seasons, such as increased amount of ongoing construction in the summer.

When processing the data, projects outside of Finland are omitted as well as data prior to 2019 due to lack of reported incidents from the time period. Categorical data (segment, season) is transformed to a suitable format for machine learning algorithms with one-hot encoding, and continuous variables are normalized to prevent confusion with feature importance. As displayed in Figure 7, class imbalance between accident and non-accident cases is extremely high with only approximately 1% in accident class out of all incidents. This is not an unusual situation when considering real-life ratio of how often or seldom an accident occurs at a given moment, but this natural imbalance in data can severely affect the performance of a machine learning model if not handled correctly.

Figure 7: Distribution of data between classes



All of the 16 final features used as input variables for the Machine Learning models are displayed in Table 3.

Table 3: The input variables

Variable Name	**Description**	**Values**
NEARMISS_count	Cumulated count of near-miss incidents	Discrete, int
POSITIVE_count	Cumulated count of positive safety observations	Discrete, int
TASK_count	Cumulated count of safety hazard related observations	Discrete, int
Completion rate	Project completion rate at the time of reported incident	Continuous, float between 0 and 1
Project duration	Project duration in days	Discrete, int
Mgmt_wt_counts	Cumulated count of management walk & talks on the construction site	Discrete, int
Obs_reactTime_meanH	Average time in hours it takes the project to react to reported safety hazard	Continuous, float
Obs_reportDelay_meanH	Average time delay in hours it has taken the project to report a safety hazard, compared to when it was detected	Continuous, float
Previous accidents	Cumulated count of previous accidents at the worksite	Discrete, int
Segment_Code MDS_BP	1 for projects in Business Premises segment	Binary, int
Segment_Code MDS_HO	1 for projects in Housing segment	Binary, int
Segment_Code MDS_IN	1 for projects in Infrastructure segment	Binary, int
Season_autumn	1 for incidents reported during autumn	Binary, int
Season_spring	1 for incidents reported during spring	Binary, int
Season_summer	1 for incidents reported during summer	Binary, int
Season_winter	1 for incidents reported during winter	Binary, int

3.3 Machine Learning methods

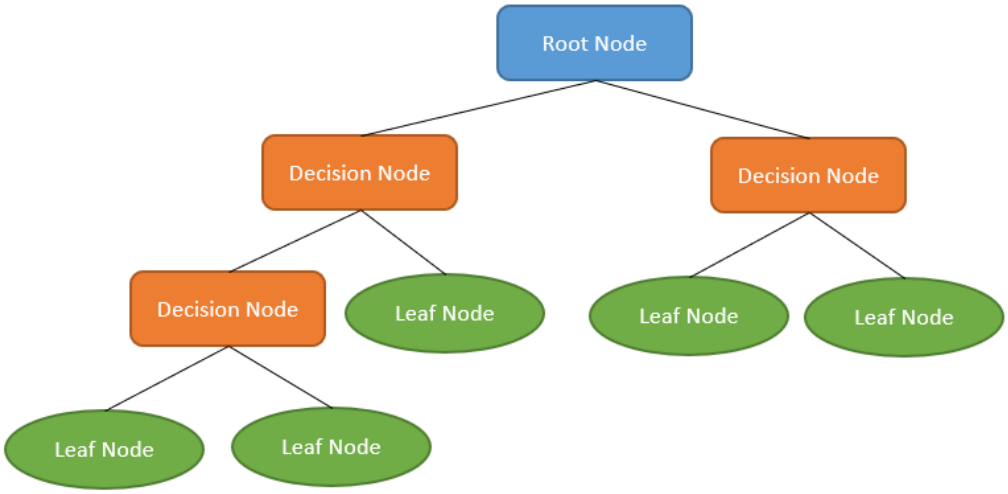
This section will introduce proposed models for predicting occurrence of accidents at construction sites. Since the research objective defines the prediction problem as a classification problem, all models proposed in this section have been selected due to their applicability for classification. Majority of proposed models consists of interpretable models that are able to provide insight into factors affecting accidents occurrence prediction, however, a few uninterpretable models were included due to their previous success in similar prediction problems.

3.3.1 Decision tree (DT)

Decision tree (DT) is a supervised learning algorithm that is widely utilized for classification problems, applicable to both categorical and continuous dependent variables. Some of its major advantages are interpretability and comprehensibility, showcasing learned knowledge, as well as capability of processing samples with missing attributes, ability to handle unrelated features and produce feasible results for large datasets in a relatively short time. Nevertheless, DT's significant shortcomings are its proneness to overfitting and bias for overrepresented classes which can, however, be avoided by model tuning.

Decision tree algorithm divides the data into two or more groups based on the most significant or independent variables. Its structure, as illustrated in Figure 8, contains a root node, decision nodes that split the data according to a certain parameter and leaves representing the final outcomes, i.e. classes in case of a classification problem. The two main attribute selection measures that decision tree utilizes as a principle for dividing the data

Figure 8: Structure of a Decision Tree



selecting the best attributes for the root node and sub-nodes are Gini Index and Information Gain.

Gini Index measures the impurity or purity of a node and it is defined below:

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Where *C* represents the total number of classes and p_i is the probability of picking the data point with the class i. Moreover, Gini Index or Gini Impurity is calculated by subtracting the sum of the squared probabilities of each class from one. Node impurity refers to node containing observations from more than one class, meaning that a node with Gini Index of 0.5 has a 50–50 split of classes in a group, equivalent for the worst purity.

Information Gain, on the other hand, is an attribute selection measure based on Entropy, which specifies randomness in data. Information Gain measures changes in entropy after the division of a dataset based on an attribute, and decision tree algorithm aims at maximizing the value of information gain. Entropy and Information Gain are calculated with following formulas:

$$Entorpy = \sum_{i=1}^{C} -p_i \log_2 p_i$$

$$Information\ Gain = Entropy(T) - Entropy(T, X)$$

Where *C* represents the total number of classes and p_i is the probability of picking the data point with the class i.

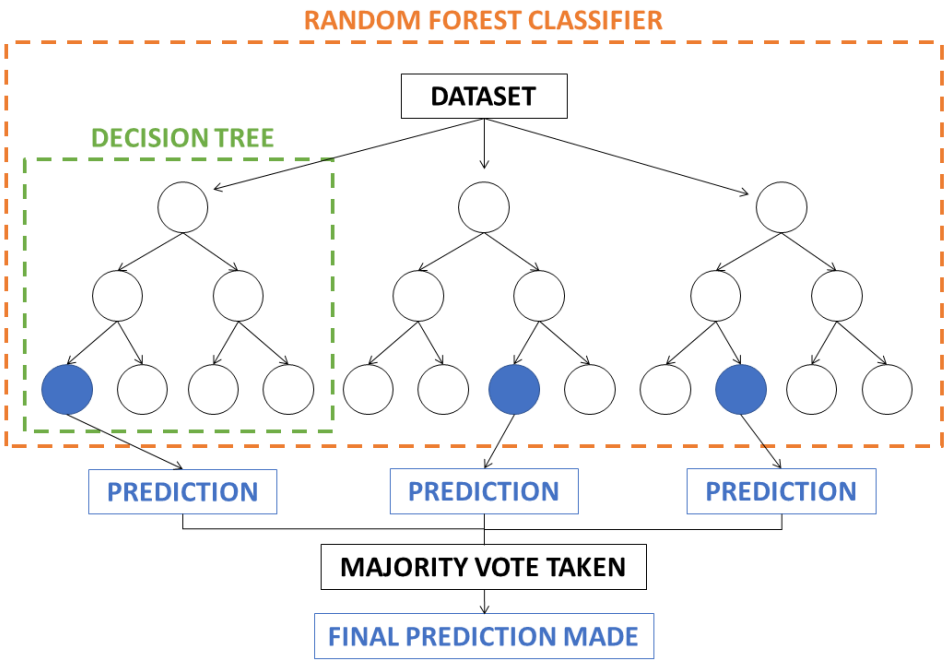
$Entropy(T)$ describes the entropy or randomness of the entire dataset or prior to the split, while $Entropy(T, X)$ is the remaining entropy of the dataset after splitting the data along the values of explanatory feature X.

3.3.2 Random forest (RF)

Random forest (RF) is a supervised machine learning algorithm based on ensemble learning that consists of a series of decision trees (forests) built on different samples from the data, and it can be utilized for both classification and regression problems. Classification made by each decision tree is referred to as vote, and the random forest model chooses the category obtaining the majority of votes to classify a new object according to its attributes. Random forest is capable of processing high-dimensional data while maintaining fast training speed. It diminishes some of Decision Tree's limitations by reducing overfitting and increasing precision, however, it may suffer from overfitting in some noisy classification or regression problems.

Random forest utilizes a principle called bagging, also known as Bootstrap Aggregation, which chooses a random sample from the original dataset and completes it with row sampling. Thus, each tree in the Random Forest is trained independently, generating different results that are later combined to produce a final output based Majority Voting or Averaging for classification and regression problems respectively. This technique contributes to Random Forest's most important characteristics: since not all attributes of a dataset are considered for each individual tree but each individual tree is different, diversity of the model is increased. Furthermore, feature space is reduced and stability of the model is improved through techniques implemented with Random forest model.

Figure 9: Random Forest classifier



3.3.3 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is an efficient implementation of the stochastic gradient boosting ensemble modelling technique suitable for both classification and regression problems. As established in the previous section, ensemble learning presents a systematic solution to compound the predictive power of several learners where the result is the aggregated output from multiple models.

Like bagging, boosting is an alternative approach to ensemble modeling where each model attempts to correct errors present in the previous model. In Gradient Boosting, every predictor is trained utilizing the residual errors of predecessor as labels, and XGBoost implements the technique of Gradient Boosting to decision trees. In XGBoost, individual trees are created sequentially, and eventually the weak learners produce a strong learner by training each sequence in the model with updated version of the residuals.

The boosting ensemble technique can be defined as follows:

$$F_m(x) \leftarrow F_{m-1}(x) + h_m(x)$$

Where m is describes the number of iterations, and F_m is produced by combining residuals from previous sequence F_{m-1} with the new model h_m.

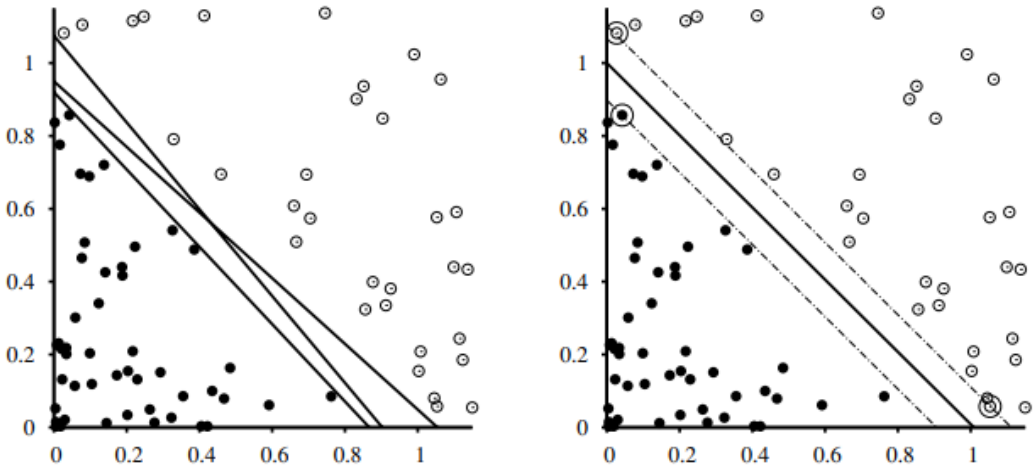
3.3.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) has been a popular supervised learning algorithm among construction machine learning research due to its predictive ability. It is used primarily for classification problems but can also be utilized for regression problems. SVM aims at forming a decision boundary called a hyperplane to segregate n-dimensional space into classes, ultimately categorizing each data point. The hyperplane which is inferred to perform classification tries to maximize the margin between the datapoints closest to the decision boundary, also referred to as Support Vectors for they support the hyperplane in between.

SVM can be linear or non-linear, and as the dimensionality in the data increases, it becomes more difficult to use a linear hyperplane to classify the data. Non-linear SVM uses a technique known as the kernel trick to perform complex data transformations, transforming

low dimensional input space to a higher dimensional space, thus converting a previously not separable problem to a separable problem.

Figure 10: SVM classification with two classes represented by black and white circles. Left graph contains three candidate linear separators. Right graph presents the maximum margin separator (heavy line), positioned at the midpoint of the margin (area between dashed lines). The support vectors (circled points on the dashed line) are the examples closest to the separator (Russell and Norvig, 2010).



Although SVM is deemed effective in high dimensional spaces, it suffers from decrease in performance with large data sets due to time required for training. Furthermore, one of SVM's most significant drawbacks is its lack of interpretability. SVM does not offer any insight on feature importance or other explanation for the model's output.

3.3.5 K-nearest neighbors (KNN)

The k-nearest neighbors (KNN) is supervised machine learning algorithm applicable to both classification and regression problems. KNN calculates the distance between data points and classifies the data points based on their proximity, assuming that similarity of data points is based on their calculated proximity to one another. KNN stores all cases and classifies new cases based on the class of majority of the surrounding k neighbors. KNN is generally based on the Euclidian distance that is defined as

$$d(x_i, x_l) = \sqrt{(x_{i1} - x_{l1})^2 + (x_{i2} - x_{l2})^2 + \cdots + (x_{ip} - x_{lp})^2}$$

Where x_i is an input variable with p features whose distance to data point x_l is being measured.

While easy to implement, the computation cost required by KNN is high for large datasets due to storing all the available data and calculating the distance between the data points for all the training samples.

3.3.6 Logistics regression (LR)

Logistic regression (LR) branches from linear regression, aiming at performing classification on a categorical target variable instead of regression. Logistic regression utilizes logistic function to model a binary output variable, and it does not require a linear relationship between predictor variables and output variables. Logistic function is defined below:

$$Logistic\ function = \frac{1}{1 + e^{-x}}$$

Where x is the input variable. Thus, logistic regression can be described as a transformation of a linear regression using the sigmoid function, producing an s-shaped curve. Formula of logistic regression is defined as follows:

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Where $\beta_0 + \beta_1 x$ can be perceived as corresponding to linear model function $y = ax + b$.

Mathematically, logistic regression relies on a probability of an event occurring divided by the probability of an event not occurring, concept referred to as the "odds" of the event. Logistic regression assigns weights to dimensions of input data, however, unlike in linear regression, the weights have an exponential relationship with the odds of the event rather than linear.

Logistic regression is interpretable, easy to implement, does not require significant amounts of training and makes no assumptions about distributions of classes in feature space. However, logistic regression assumes linearity between the dependent variable and the log odds, which impairs its applicability to real-life problems. Moreover, logistic regression requires average or no multicollinearity between independent variables. Logistic regression

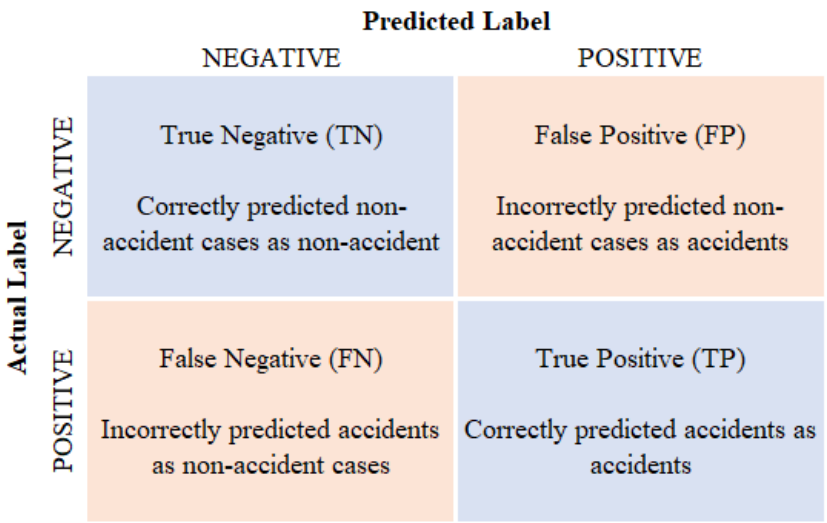
is also easy to underfit, meaning that built model is too simple to achieve a decent performance on the dataset.

3.4 Model evaluation

Classification problems require their own performance evaluation criteria for assessing how successful the model is in predicting the outcome. Furthermore, these evaluation metrics are instrumental in comparing the suitability of proposed models to the described research problem. This study will utilize common evaluation metrics suitable for binary classification models including confusion matrix, precision, recall, F1 score, ROC curve and ROC AUC.

Confusion matrix denotes how many times the classification model predicted correctly or incorrectly the class of the data point. As displayed in Figure 11, confusion matrix is a table that presents predictions for values that fall under true positive (TP), false positive (FP) true negative (TN) and false negative (FN). Class imbalance should always be considered when observing and interpreting performance metrics derived from confusion matrix. Confusion matrix reveals class imbalance better than accuracy score which only tells the proportion of correct predictions out of total number of predictions without considering class sizes. Moreover, if one of the classes is extremely small (accidents) compared to the other (non-accident cases), falsely labeled cases belonging to the smaller class will have very little impact on the accuracy score. Thus, accuracy is not an appropriate evaluation metric due to its intolerance for class imbalance.

Figure 11: Confusion matrix



Several other performance evaluation metrics can be derived from confusion matrix, such as precision, recall and F1 score. Precision focuses on how many of all cases predicted positive (TP+FP) are actually positive (i.e. accidents) whereas recall calculates how many of actual positive cases (TP+FN) are detected by the model. Formulas for calculating precision and recall are presented below:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

As can be deduced from the formulas, precision and recall range from 0 to 1, and an increase in precision would cause a decrease in recall and vice versa. Therefore, a choice needs to be made between prioritizing one of these metrics, which is often referred to as precision-recall trade-off. Detecting all accidents is of highest priority so that it can further be assessed how accidents could be more effectively prevented altogether. Thus, reducing False Negative is more important than minimizing False Positive, making recall more valuable performance evaluation criterion than precision for this study.

Another metric based on confusion matrix is F1 score, often referred to as the harmonic mean of the model's precision and recall, which is particularly useful when neither precision nor recall is perceived as more important than the other. Moreover, F1 score is applicable for imbalanced datasets where one classification appears much more often than another. As also indicated by precision and recall, the closer the F1 score is to 1, the better the model is. Formula for F1 score is defined as follows:

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall}$$

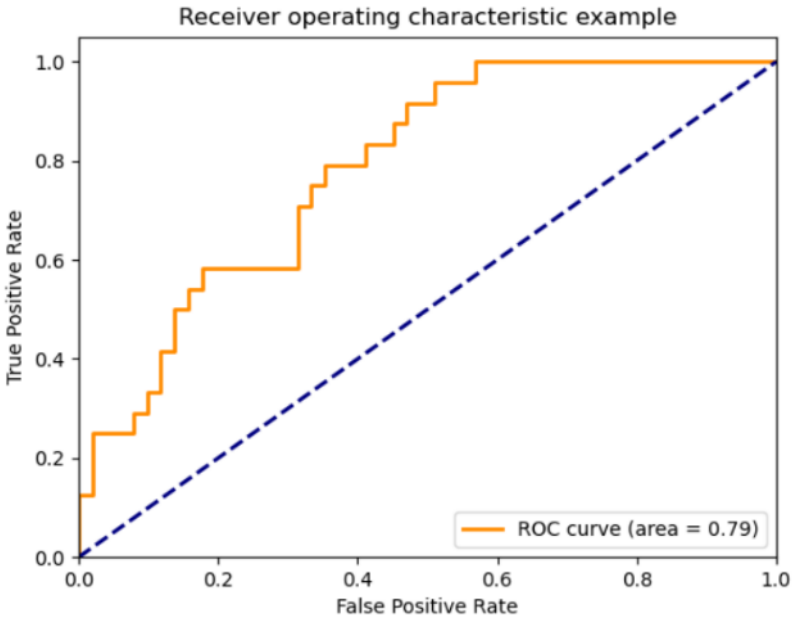
An ROC curve, also known as receiver operating characteristic curve, visualizes the performance of a classification model at all classification thresholds in a line graph forming a curve. More specifically, for each possible decision threshold, the ROC curve plots the False Positive Rate (FPR), fraction of negative instances that are incorrectly predicted as positive, versus the True Positive Rate (TPR), fraction of positive instances correctly classified as positive. Lowering the classification threshold causes more items to have

positive predicted class, thus increasing both False Positives and True Positives. From example in Figure 12 it is evident that initially the TPR increases at a higher rate than the FPR, however, after at a certain threshold, the increase in FPR outweighs the increase in TPR. True Positive Rate (TPR) and False Positive Rate (FPR) are defined below:

$$True\ Positive\ Rate\ (TPR) = \frac{TP}{TP + FN}$$

$$False\ Positive\ Rate\ (FPR) = \frac{FP}{FP + TN}$$

Figure 12: Receiver Operating Characteristic (ROC) Curve (Scikit-learn, 2022)



While an ROC curve displays the performance of a classification model across all decision thresholds, Area Under the Curve (AUC) measures the two-dimensional area below the ROC curve. With values ranging between 0 and 1, AUC of 1 represents a model whose predictions are 100% correct and thus, a model's AUC needs to be greater than 0.5 for it to perform better than the random guess. However, it should be noted that AUC is classification-threshold-invariant, meaning that it is not particularly helpful in minimizing one type of classification error which is a significant drawback since there are major disparities in the cost of false negatives vs. false positives.

In conclusion, costs of treating an accident and project potentially suffering from lost resources are higher than cost of taking preventative actions in case of predicted accident. Hence, since minimizing False Negative is prioritized higher than reducing False Positive, recall is one of the most significant performance criteria, however, it should not be the only performance metric when evaluating model performance.

4 Application of Machine Learning

This section will present the experiment setup prior to trialing proposed machine learning approaches and discuss the results of the trials utilizing several evaluation metrics established in previous chapter.

4.1 Experiment setup

Prior to application of machine learning methods, the data is split into train and test sets which is an important procedure for later when evaluating the performance of a prediction-based machine learning model. If the entire dataset is utilized in training the model and instead of dividing it into two subsets, the model gets highly biased to the dataset which often results in rewarding overly complex models, leading to an overfitted model that is synced too closely to the training data. In consequence, the model is unable to generalize on unseen data, thus performing poorly when implemented after training and testing phase. Therefore, it is recommended to use separate training and testing sets in model building instead of the entire dataset unless the initial dataset is already extremely small.

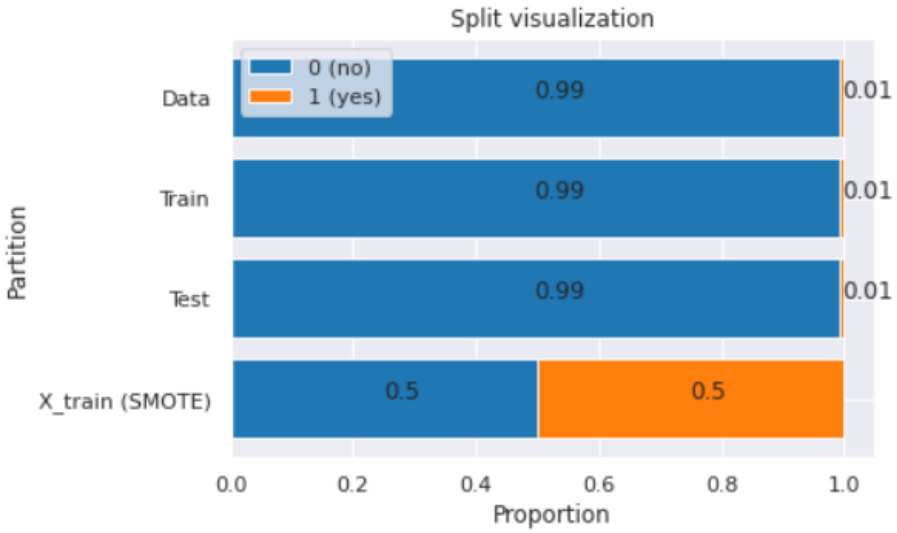
In train-test split, the training set is used for fitting the model, using the input and output or target variables from the data as a set of examples. The model in training takes the input variables to produce a prediction for output class, comparing its outputs with the actual target variables in the training data corresponding to the inputs and then adjusts its parameters according to the result, aiming at fitting the data closely enough. Once training the model is complete, the test data set is used to evaluate the model's fit, providing examples it has not seen before and then comparing the outputs predicted by the model with the actual target variables from the test set. It is often perceived that splitting the dataset with 70-80 % allocation to training and 20-30 % allocation to testing set generates the best results (Gholamy, Kreinovich and Kosheleva, 2018), hence this study uses 75/25 split between train and test data, respectively.

After splitting the data into train and test sets, SMOTE technique is applied for balancing the training set. As data processing revealed, the data is distributed between the two classes (accident vs. no accident) extremely unequally, providing very few examples of accidents compared to non-accident cases. This violates the assumption of an equal number of examples for each class, which most machine learning classification algorithms lean on.

However, this severe imbalance can be handled with Synthetic Minority Over-sampling Technique (SMOTE) which generates synthetic samples for the minority categories, i.e., accidents in the case of this research. All models will be trained using SMOTE balanced data to compare performance of machine learning approaches on balanced datasets.

As portrayed in the split visualization of Figure 13, SMOTE technique has created equal distribution of target variable examples between accident (label 1) and non-accident (label 0) cases in the training set, while the unbalanced train set has inherited the class imbalance of the original dataset.

Figure 13: Split visualization displaying distribution between target classes



4.2 Results

Predictive skill of each machine learning approach applied in this research is presented with several previously introduced performance evaluation metrics in the following table:

Table 4: Results from trialed machine learning models

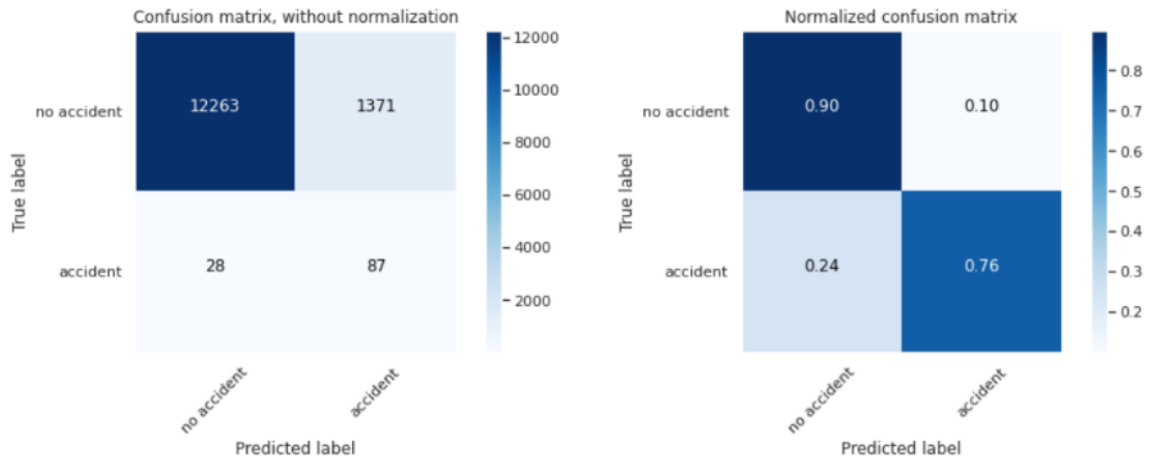
Model	**Precision**	**Recall**	**F1 Score**	**AUC**
Decision tree with SMOTE	0.027	0.557	0.051	0.693
Random forest with SMOTE	0.353	0.617	0.449	**0.804**
XGBoost with SMOTE	0.654	0.444	0.529	0.721
SVM with SMOTE	0.028	0.600	0.053	0.717
KNN with SMOTE	0.061	**0.757**	0.113	**0.829**

Logistic regression with SMOTE	0.019	**0.643**	0.036	0.678

As was established in the previous chapter, this study is primarily focused on developing a model that is capable of detecting as many accidents as possible rather than minimizing false positives. Thus, prioritizing recall over precision as a performance evaluation metric favors models that excel at predicting correctly majority of accident cases. Recall scores for all models apart from XGBoost presented in Table 4 are satisfactory with over 50% of accident cases correctly labeled as accidents, and the scores for best performing models according to recall, KNN and Logistic Regression, are highlighted in the recall column. Table 4 also exhibits how precision and, hence, F1 score for different models vary greatly, and regardless of deliberate precision-recall trade-off, it is important to understand why the precision metric is so low while recall scores are mostly acceptable.

Figure 14 displays confusion matrix for the KNN model which obtains the best predictive skill for this research problem with recall of 0.757, as summarized in the results table (Table 4). This means it is able to classify 75.7% of all accidents to the correct class, and as indicated by AUC, the model has 82.9% chance of making a correct prediction overall. However, KNN has a precision score of 0.019, so only 1.9% of incidents the model labeled as accidents actually were accidents. When looking at the confusion matrix, the false positive rate for KNN is only 10%, so only 10% of non-accident cases end up with incorrect class label, but due to high class imbalance, 10% of non-accident cases incorrectly labeled as accidents (false positives) significantly overpower the number of correctly labeled accidents (true positives), resulting in extremely low precision and F1 scores. Therefore,

Figure 14: Confusion matrix for KNN



although precision and F1 score are not the most optimal model performance evaluation metrics in the context of this study, they provide important and thought-provoking insight on the impact of class imbalance, indicating frequent efforts for treating false positives, also referred to as Type I error.

Although KNN is the best at detecting accidents and making correct predictions overall when comparing its recall and AUC scores to other trialed models, a significant drawback of the model is its lack of interpretability. Thus, it is only able to satisfy one of the two research questions of this study without increasing the amount of knowledge about most influential factors in predicting the occurrence of an accident at a construction site. Logistic Regression possesses the second-best ability to detect accidents with a recall score of 0.643, however, regarding all incidents, it has only 67.8% chance of making a correct prediction with the worst AUC score of all the models. Random Forest, on the other hand, obtains the second-best AUC score (80.4% chance of making a correct prediction) and the third-best recall score (able to classify 61.7% of all accidents to the correct class), so it may potentially overthrow KNN due to its interpretability and, thus, ability to serve the purpose of this study, satisfying both research questions.

Out of all of the models trialed in this study, only Decision Tree, Random Forest, XGBoost and Logistic Regression are interpretable. These four models provide further explanation for their predictions by including the relative importance of each feature or, in the case of Logistic Regression, coefficients that indicate how much a change (increase or decrease) of one unit in each of the input variables affects the odds of an accident occurring as explained in methods chapter. Feature importance refers to a calculation technique that ranks features based on the effect that they have on the model's prediction. Most important features according to these metrics are presented in Table 5; *Number of previous accidents* was the most influential factor in predicting occurrence of accidents according to all models, suggesting that a project is more prone to accidents the more accidents it has had happen in the past.

Additionally, *Reaction time to observations* appeared among the top three most important features in three out of four models and *Project completion rate* in two out of four models. These findings indicate that, firstly, the more quickly a project is able to react to and resolve a reported incident of a safety hazard on construction site, the less likely it is to suffer from an accident and, secondly, a project is more prone to accidents in the early stage of the construction phase. This could be due to lack of experience in navigating the circumstances

at the new construction site or it could be the consequence of potentially complicated or hazardous tasks and activities related to early stages of the construction phase.

Table 5: Most important features ranked by model's feature importance or coefficients

Model	**Most important features**
Decision Tree	***Number of previous accidents***, *Project completion rate, Reaction time to observations*
Random Forest	***Number of previous accidents***, *Project completion rate, Reaction time to observations*
XGBoost	***Number of previous accidents***, *Housing as segment, Summer as season*
Logistic Regression	***Number of previous accidents***, *Number of positive observations, Reaction time to observations*

To conclude, with the predictor variables used in this study, KNN obtains the highest predictive skill in predicting the occurrence of accidents given its recall and AUC scores. However, similarly to SVM, it offers no insight into feature importance regarding this classification problem. To satisfy RQ1 the model needs to be interpretable and considering all the interpretable models trialed in this study, Random Forest was the most successful at making correct predictions with AUC greater than 0.8. Additionally, it had the third-highest recall score overall, indicating decency in prognosticating accidents despite the deficiency in examples for accidents. According to feature importance, Random Forest also included all of the top three features most commonly appearing across all of the models, potentially conforming with earlier discussion regarding why these three features are influential in predicting occurrence of accidents.

5 Conclusion

The goal of this research was to examine factors that are potentially significant in predicting occurrence of accidents at Finnish construction sites, apply several machine learning techniques to utilize factors perceived meaningful in training a model capable of predicting occurrence of accidents and, ultimately, discover which technique is most suitable for this research problem while assessing which attributes are most impactful for the model. The data utilized in this research consisted of accident and safety observation data from over 600 construction projects across Finland, complimented with environmental, managerial and project-specific characteristics. As discovered in the literature review, access to data with ability to represent non-accident cases is a significant strength of this study, and lack of such data has prevented much of the previous research from attempting at predicting occurrence of accidents.

The Machine Learning methods trialed in this study included Decision Tree, Random Forest, XGBoost, Support Vector Machine (SVM), K-nearest Neighbors (KNN) and Logistic Regression. Although SVM and KNN are not interpretable and thus are incompetent in providing explanation for their predictions or insight into the impact of each input variable, they were included in this study due to their success in previous research for similar prediction problems and predictive accuracy. When considering the performance of proposed models, all machine learning approaches presented acceptable results when judging by the AUC score which remained between 0.6 and 0.9 for all models, meaning that they have over 60% chance of making a correct prediction which makes them better than the random guess. However, recall was also considered one of the most important model evaluation metrics for this problem due to the potentially high cost of treating accidents compared to the cost of enforcing preventative measures and risk-mitigation actions. In addition to KNN, Logistic Regression and Random Forest were the best at classifying accidents to the correct class.

Out of all the predictor variables selected for this research, *Number of previous accidents*, *Reaction time to observations* and *Project completion rate* appeared most frequently among the top three most influential features in predicting the occurrence of an accident across all of the interpretable models. Random Forest ranked all three in as most impactful features by feature importance, and it was the was the most successful at making correct predictions out of the interpretable models with AUC greater than 0.8, ranking it

second-best model overall when judging by this metric. Additionally, it was the second-best of interpretable models and third-best overall at correctly classifying accidents according to recall score of 0.617. Therefore, Random Forest is a considerable candidate for future development of this approach with the ability to produce satisfactory results while enabling further experimentation and examination of feature selection and influence different factors have on accident prediction. On the other hand, once knowledge and understanding of effect of potential predictor variables for proactive accident detection has been enhanced, uninterpretable models such as KNN may provide increased prediction accuracy for this cause.

5.1 Managerial implications

This study uncovered multiple means for improving the case company's proactive accident prevention, providing actionable insights to help further develop and improve safety at worksites. Recommendations for improving data collection and utilization at the case company include enriching data with background information on individual suffering from an accident (e.g. years of experience, role, start date on the construction site etc.) and more project-specific characteristics (such as project budget or number of employees and subcontractors at worksite); as Zhou, Goh and Li (2015) implicated, unfamiliarity of quantitative relationship between project scale and construction safety is a significant research gap in the entire industry. Data that is more accurately able to explain different safety outcomes could also be leveraged by, for example, employing and monitoring safety performance criteria with the highest priority identified by Abbasianjahromi and Aghakarimi (2021) in their research. Providing data on these criteria could improve the quality and versatility of the company's leading safety indicators, which are a vital part of obtaining a holistic view of worksite's safety performance while implementing preventative safety procedures (Toellner, 2001). Additionally, all accidents should be connectable to the correct project with the project code as key instead of cost object code that is currently used inconsistently by different users, making it unreliable.

As established earlier, the premise for this type of prediction requires data on non-accident cases to complement examples of accidents. This study utilized different types of reported safety incidents, including positive safety observations, and combined them with reported accidents. To obtain even more reliable results, the case company should pay attention to especially the continuity, quality and frequency of reporting positive safety

observations as those represent the true success cases predictive Machine Learning model needs for training (Baker, Hallowell, and Tixier, 2020). Although positive safety observations have been recorded since 2019, there is known to be significant variance in reporting activity between different construction sites. One reason for this according to case company personnel relies in lack of understanding of the reason and benefits of reporting positive safety observations. Therefore, if the case company plans on implementing a Machine Learning solution that is heavily dependent on records of positive safety observations, it needs to provide extensive training to personnel working on the construction sites to spread knowledge on how the process directly benefits the construction site by increasing its safety. As stated by Asanka and Ranasinghe, the commitment of all individuals involved in the construction project to safety objectives and practices is required to enhance the safety performance on construction sites (2015).

Construction accidents cause substantial direct and indirect costs to individual workers, employers and society in the form of lost income and reduction of professional capacity, loss of production and increase in production costs in addition to medical, legal and administrative costs and more (Asanka and Ranasinghe, 2015). Everett and Frank (1996) estimated the cost of accidents constitute between 7.9% and 15.0% of total construction costs. Previous research also shows how total benefits, such as savings on insurance premium, medical expenses, accident investigation and lost time far outweigh the costs of accident prevention by a ratio of approximately 3:1 (Ikpe, Hammon and Oloke, 2012), so the financial advantages of investing in accident prevention are evident.

5.2 Limitations

This research is limited by the selection of machine learning approaches proposed in this study, excluding some potentially suitable methods worth exploring. For example, suitability of neural networks for this research problem was not examined although Ayhan and Tokdemir (2018) obtained moderately successful results in trialing Artificial Neural Network (ANN) to predict outcome of construction incidents between five different categories. Additionally, the study did not employ any methods available for optimized feature selection, such as wrapper feature selection methods, or hyperparameter tuning. However, the current structure of the dataset limits opportunities for hyperparameter tuning due to deficiency in examples of accident cases. Hyperparameter tuning requires splitting the data to validation set in addition to train and test sets, and with already few accident

examples in the test set, continuing to split the data may have been unbeneficial in this experiment setup.

Furthermore, the data utilized in the study demonstrated major class imbalance and, moreover, it only modeled two types of outcomes (accident vs. no accident) instead of establishing a more diverse class distribution with milder incidents or near-misses as their own categories. Positive safety observations are the only records that truly represent success cases, whereas task and near-miss observations represent incidents of safety hazards at the construction site, so grouping these under the same class may disrupt the model.

Additionally, it should be noted that while the high class imbalance in the data was handled with SMOTE technique for the training set, it is possible that the test set contained too few examples of accident cases for effective evaluation of the model performance, leading to overly pessimistic evaluation. Furthermore, using cross-validation would have allowed utilizing the entire dataset for training, including a greater variety of examples of accident cases while potentially obtaining more stable results (Xu and Goodacre, 2018).

Finally, the best performing model according to the most significant performance evaluation metrics, KNN, is not an interpretable model and thus, it does not offer information on feature importance. The interpretable models trialed, however, indicated some of the input features to be fairly insignificant to the model although removing them did not improve the models' performance. The models may be underfitting the data, being unable to capture the relationship between the predictor variables and the target variables, which supports the recommendation of including more domain-specific features to the models.

5.3 Suggestions for further research

Regarding similar cases of predicting occurrence of accidents with examples of non-accident cases in the data, future research could experiment with multiclass prediction. This would, in turn, better conform with unique nature of construction projects and wide-ranging possibilities of different incidents that may unexpectedly occur on construction site due to the dynamic environment. Baker, Hallowell, and Tixier (2020) deemed incident occurrence prediction as an important part of research on construction safety prediction, and the lack of such research has created a void in utilizing Machine Learning in modern construction.

Additionally, predictive Machine Learning models developed in future research could include more information especially about human elements, which are substantial causes for occupational accidents (Haslam et al., 2005; Heinrich, 1959; Jaafar et al., 2017;

Petersen, 1984). Collecting more data on human factors at worksite could take inspiration from framework created by Jaafar et al. (2017). Moreover, future research should also consider and further examine what currently missing but potentially relevant new data could be obtained by implementing new collection methods or processes at construction sites or how current data collection methods or processes could be improved to enrich and improve the quality of existing data.

Finally, in addition to reformulating the research problem and expanding the data at hand, future research should trial models excluded from this research to see if they can yield more accurate results. Experiments with feature selection and hyperparameter optimization as well as improving the compatibility of data compiled from different sources may also increase the predictive skill of proposed models.

References

Abas, N., Yusuf, N., Suhaini, N., Kariya, N., Mohammad, H. and Hasmori, M., 2020. Factors Affecting Safety Performance of Construction Projects: A Literature Review. IOP Conference Series: Materials Science and Engineering, 713(1), p.012036. doi: 10.1088/1757-899X/713/1/012036

Abbasianjahromi, H. and Aghakarimi, M., 2021. Safety performance prediction and modification strategies for construction projects via machine learning techniques. Engineering, Construction and Architectural Management, vol. ahead-of-print no. ahead-of-print. doi: 10.1108/ecam-04-2021-0303

Alzubi, J., Nayyar, A. and Kumar, A., 2018. Machine Learning from Theory to Algorithms: An Overview. Journal of Physics: Conference Series, 1142, p.012012. doi: 10.1088/1742-6596/1142/1/012012

Asanka, W. A., and Ranasinghe, M., 2015. Study on the impact of accidents on construction projects. 6th International Conference on Structural Engineering and Construction Management, vol. 4, pp. 58-67. Available at: https://www.researchgate.net/publication/288022884_Study_on_the_impact_of_accidents_on_construction_projects [Accessed 19.6.2022]

Ayhan, B. and Tokdemir, O., 2018. Predicting the outcome of construction incidents. Safety Science, 113, pp.91-104. doi: 10.1016/j.ssci.2018.11.001

Baker, H., Hallowell, M. and Tixier, A., 2020. AI-based prediction of independent construction safety outcomes from universal attributes. Automation in Construction, 118, p.103146. doi: 10.1016/j.autcon.2020.103146

Bhattacharjee, S., Ghosh, S. and Young-Corbett, D., 2011. Safety Improvement Approaches in Construction Industry: A Review and Future Directions. Conference: 47th ASC Annual International Conference Proceedings

Brueggmann, M., Roetting, M. and Luczak, H., 2001. International Comparison of Occupational Safety and Health Research—A Review Based on Published Articles. International Journal of Occupational Safety and Ergonomics, 7(4), pp.387-401. doi: 10.1080/10803548.2001.11076509

Brynjolfsson, E. and McElheran, K., 2016. The Rapid Adoption of Data-Driven Decision-Making. American Economic Review, 106(5), pp.133-139. doi: 10.1257/aer.p20161016

Carter, G. and Smith, S., 2006. Safety Hazard Identification on Construction Projects. Journal of Construction Engineering and Management, 132(2), pp.197-205. doi: 10.1061/(ASCE)0733-9364(2006)132:2(197)

Choi, J., Gu, B., Chin, S. and Lee, J., 2019. Machine learning predictive model based on national data for fatal accidents of construction workers. Automation in Construction, 110, p.102974. doi: 10.1016/j.autcon.2019.102974

Confederation of Finnish Construction Industries (2022). Tapaturmatilanne marraskuu 2021. Online. Available at: https://www.rakennusteollisuus.fi/globalassets/tyoturvallisuus/tapaturmatilanneseurannat/marraskuu-2021-kuukausittain-seuranta.pdf [Accessed 27.1.2022]

Confederation of Finnish Construction Industries (2015). Rakennusteollisuuden kannanotto: Tavoitteena nolla tapaturmaa rakennusteollisuudessa 2020. Online. Available at: https://www.rakennusteollisuus.fi/globalassets/tyoturvallisuus/2015/turvallisuuskannaotto_syksy_2015.pdf [Accessed 28.1.2022]

Dekker, S., 2002. The Field Guide to Understanding 'Human Error'. Ashgate Publishing, USA.

Esmaeili, B., Hallowell, M. and Rajagopalan, B., 2015. Attribute-Based Safety Risk Assessment. II: Predicting Safety Outcomes Using Generalized Linear Models. Journal of Construction Engineering and Management, 141(8), p.04015022. doi: 10.1061/(ASCE)CO.1943-7862.0000981.

Everett, J. and Frank, P., 1996. Costs of Accidents and Injuries to the Construction Industry. Journal of Construction Engineering and Management, 122(2), pp.158-164. doi: 10.1061/(asce)0733-9364(1996)122:2(158)

Finnish Centre for Occupational Safety (n.d.). Työtapaturmat rakennusalalla. Available at: https://ttk.fi/tyoturvallisuus_ja_tyosuojelu/toimialakohtaista_tietoa/rakennusala/tyotapaturmat_rakennusalalla [Accessed 27.1.2022]

Fradkov, A., 2020. Early History of Machine Learning. IFAC-PapersOnLine, 53(2), pp.1385-1390. doi: 10.1016/j.ifacol.2020.12.1888

Friend, M. and Kohn, J., 2018. Fundamentals of occupational safety and health. 7th ed. Lanham: Bernan Press. ISBN: 9781598889826

Gholamy, A., Kreinovich, V. and Kosheleva, O., 2018. Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. Departmental Technical Reports (CS). 1209. Online. Available at: https://scholarworks.utep.edu/cs_techrep/1209 [Accessed 19.6.2022].

Goetsch, D., 2019. Occupational safety and health for technologists, engineers, and managers. 9th ed. Boston: Pearson. ISBN: 9780134695815

Hale, A.R. and Hovden, J., 1998. Management and culture: the third age of safety. A review of approaches to organizational aspects of safety, health and environment, in Feyer, A.M. and Williamson, A. (eds) Occupational Injury: Risk Prevention and Intervention, Taylor & Francis, London, pp. 129-65.

Hansen, R., 2001. Overall equipment effectiveness: a powerful production/maintenance tool for increased profits. 1st ed. New York, Industrial Press Inc.

Haslam, R., Hide, S., Gibb, A., Gyi, D., Pavitt, T., Atkinson, S. and Duff, A., 2005. Contributing factors in construction accidents. Applied Ergonomics, 36(4), pp.401-415. doi: 10.1016/j.apergo.2004.12.002

Hinze, J., Hallowell, M. and Baud, K., 2013. Construction-Safety Best Practices and Relationships to Safety Performance. Journal of Construction Engineering and Management, 139(10), p.04013006. doi: 10.1061/(ASCE)CO.1943-7862.0000751

Hinze, J., Thurman, S. and Wehle, A., 2013. Leading indicators of construction safety performance. Safety Science, 51(1), pp.23-28.

Hämäläinen, P., Takala, J. and Saarela, K., 2006. Global estimates of occupational accidents. Safety Science, 44(2), pp.137-156. doi: 10.1016/j.ssci.2005.08.017

Ikpe, E., Hammon, F. and Oloke, D., 2012. Cost-Benefit Analysis for Accident Prevention in Construction Projects. Journal of Construction Engineering and Management, 138(8), pp.991-998. doi: 10.1061/(ASCE)CO.1943-7862.0000496

Immonen, M., 2015. Työturvallisuudella säästötalkoisiin. [Blog] Rakennuslehti. Online. Available at: https://www.rakennuslehti.fi/blogit/tyoturvallisuudella-saastotalkoisiin/ [Accessed 28.1.2022]

International Labour Office, 2003. ILO standards-related activities in the area of occupational safety and health: An in-depth study for discussion with a view to the elaboration of a plan of action for such activities. International Labour Conference. [online] Geneva: International Labour Organization. Available at: https://www.ilo.org/wcmsp5/groups/public/---ed_protect/---protrav/---safework/documents/publication/wcms_177455.pdf [Accessed 3.2.2022].

International Labour Organization (2014). Safety and Health at Work: A Vision for Sustainable Prevention. Online. Available at: https://www.ilo.org/global/topics/safety-and-health-at-work/resources-library/publications/WCMS_301214/lang--en/index.htm [Accessed 27.1.2022]

International Labour Organization (2009). Facts on Safety and Health at Work. Online. Available at: https://www.ilo.org/global/topics/safety-and-health-at-work/resources-library/publications/WCMS_105146/lang--en/index.htm [Accessed 27.1.2022]

International Labour Organization (2021). Impact of COVID-19 on the construction sector. ILO Sectoral Brief. Available at: https://www.ilo.org/sector/Resources/publications/WCMS_767303/lang--en/index.htm [Accessed 9.6.2022]

International Labour Organization (2005). Facts on Safety at Work. Online. Available at: https://www.ilo.org/global/resources/WCMS_067574/lang--en/index.htm [Accessed 27.1.2022]

Jaafar, M., Arifin, K., Aiyub, K., Razman, M., Ishak, M. and Samsurijan, M., 2017. Occupational safety and health management in the construction industry: a review. International Journal of Occupational Safety and Ergonomics, 24(4), pp.493-506. doi: 10.1080/10803548.2017.1366129

Jordan, M. and Mitchell, T., 2015. Machine learning: Trends, perspectives, and prospects. Science, 349(6245), pp.255-260. doi: 10.1126/science.aaa8415

Kang, K. and Ryu, H., 2019. Predicting types of occupational accidents at construction sites in Korea using random forest model. Safety Science, 120, pp.226-236. doi: 10.1016/j.ssci.2019.06.034

Langley, P., 1996. Elements of machine learning. San Francisco, Calif.: Morgan Kaufmann, pp. 2-4.

Lantto, E. and Räsänen, T., 2019. *Rakennusalan työturvallisuuden kehitys.* NOLLA TAPATURMAA RAKENNUSTEOLLISUUDESSA 2020 -HANKE. Finnish Institute of Occupational Health (TTL). Online. Available at: https://www.rakennusteollisuus.fi/globalassets/tyoturvallisuus/2020_sekalainen/raportti_final.pdf?utm_source=eJokka+2/2020+%7c+Aiheina+mm.+nolla+tapaturmaa,+turvallisuusviikko,+koronaohjeita,+hiljainen+hetki&utm_medium=email&utm_campaign=eJokka [Accessed 27.1.2022]

Leskinen, M., 2018. Kajaanin sairaalatyömaalla on sattunut vain lieviä työtapaturmia – Kainuussa vakavat työtapaturmat ovat pysyneet kurissa. Yle Uutiset. Online. Available at: https://yle.fi/uutiset/3-10307390 [Accessed 28.1.2022]

Loukkola, P., 2013. Turvapuistossa opitaan turvalliset työtavat. Yle Uutiset. Online. Available at: https://yle.fi/uutiset/3-6908161 [Accessed 28.1.2022]

Mesaros, P., Spisakova, M. and Mackova, D., 2019. Analysis of Safety Risks on the Construction Site. IOP Conference Series: Earth and Environmental Science, 222, p.012012. doi: 10.1088/1755-1315/222/1/012012

Michalski, R., Carbonell, J. and Mitchell, T., 1983. Machine learning: An Artificial Intelligence Approach. Palo Alto, Calif.: Tioga Pub. Co., p.3. ISBN: 9781493303489

Mohammadi, A., Tavakolan, M. and Khosravi, Y., 2018. Factors influencing safety performance on construction projects: A review. Safety Science, 109, pp.382-397. doi: 10.1016/j.ssci.2018.06.017

Mohammed, M., Khan, M. and Bashier, E., 2016. Machine Learning: Algorithms and Applications. 1st ed. Boca Raton: CRC Press, pp.6-11.

Mohri, M., Talwalkar, A. and Rostamizadeh, A., 2018. Foundations of machine learning. 2nd ed. Cambridge (Mass.) [etc.]: The MIT press.

Petersen, D., 1984. Human-error reduction and safety management. Deer Park, N.Y., Aloray.

Pinto, A., Nunes, I. and Ribeiro, R., 2011. Occupational risk assessment in construction industry – Overview and reflection. Safety Science, 49(5), pp.616-624. doi: 10.1016/j.ssci.2011.01.003

Poh, C., Ubeynarayana, C. and Goh, Y., 2018. Safety leading indicators for construction sites: A machine learning approach. Automation in Construction, 93, pp.375-386. doi: 10.1016/j.autcon.2018.03.022

Robson, L., Ibrahim, S., Hogg-Johnson, S., Steenstra, I., Van Eerd, D. and Amick, B., 2017. Developing leading indicators from OHS management audit data: Determining the measurement properties of audit data from the field. Journal of Safety Research, 61, pp.93-103. doi: 10.1016/j.jsr.2017.02.008

Rozenfeld, O., Sacks, R., Rosenfeld, Y. and Baum, H., 2010. Construction Job Safety Analysis. Safety Science, 48(4), pp.491-498.

Russell, S. and Norvig, P., 2010. Artificial Intelligence: A Modern Approach. 3rd ed. Upper Saddle River, N.J.: Prentice Hall, pp.1-2. ISBN: 9780136042594

Samuel, A., 1959. Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development, 3(3), pp.210-229. doi: 10.1147/rd.33.0210

Shavlik, J. and Dietterich, T., 1990. Readings in machine learning. San Mateo, CA: Morgan Kaufmann, pp. 3-4.

Shin, Y., 2019. Application of Stochastic Gradient Boosting Approach to Early Prediction of Safety Accidents at Construction Site. Advances in Civil Engineering, 2019, pp.1-9. doi: 10.1155/2019/1574297

Snashall, D., 2005. Occupational health in the construction industry. Scandinavian Journal of Work, Environment & Health, [online] 31(2), pp.5-10. Available at: https://www.sjweh.fi/show_abstract.php?abstract_id=955&add_to_basket=1 [Accessed 3 February 2022]

Swuste, P., Groeneweg, J., van Gulijk, C., Zwaard, W., Lemkowitz, S. and Oostendorp, Y., 2020. The future of safety science. Safety Science, 125, p.104593. 10.1016/j.ssci.2019.104593

Tixier, A., Hallowell, M., Rajagopalan, B. and Bowman, D., 2016. Application of machine learning to construction injury prediction. Automation in Construction, 69, pp.102-114. doi: 10.1016/j.autcon.2016.05.016

Toellner, J., 2001. Improving safety & health performance: Identifying & measuring leading indicators. Professional Safety, 46(9), pp.42-47. ISSN: 00990027

Tsang, Y., Fung, I., Tam, V., Sing, C. and Lu, C., 2017. Development of an accident modelling in the Hong Kong construction industry. International Journal of Construction Management, 17(2), pp.124-131. doi:10.1080/15623599.2016.1222664

Törner, M. and Pousette, A., 2009. Safety in construction – a comprehensive description of the characteristics of high safety standards in construction work, from the combined perspective of supervisors and experienced workers. Journal of Safety Research, 40(6), pp.399-409. doi: 10.1061/(asce)0742-597x(1996)12:4(33)

Xu, Y. and Goodacre, R., 2018. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. Journal of Analysis and Testing, 2(3), pp.249-262. doi: 10.1007/s41664-018-0068-2

Young, S., 1996. Construction Safety: A Vision for the Future. Journal of Management in Engineering, 12(4), pp.33-36. doi:

Zhou, Z., Goh, Y. and Li, Q., 2015. Overview and analysis of safety management studies in the construction industry. Safety Science, 72, pp.337-350. doi: 10.1016/j.ssci.2014.10.006

Zhu, R., Hu, X., Hou, J. and Li, X., 2020. Application of machine learning techniques for predicting the consequences of construction accidents in China. Process Safety and Environmental Protection, 145, pp.293-302. doi: 10.1016/j.psep.2020.08.006