WORKER'S COMPENSATION CENTER

Predictive modeling of workplace accident outcomes utilizing XGBoost and Tree-Structured Parzen Estimator

Publication of Workers' Compensation Center
3/2022

Workers' Compensation Center | Publication 3/2022

Publisher:

Finnis Workers' Compensation Center (TVK)
www.tvk.

Helsinki 3/2022

Author:

Antton Koskinen
Master's Thesis

Degree:
Master's degree

Degree programme:
Information and Service Management

Thesis advisor(s):
Pekka Malo

Aalto University School of Business

ISSN: 2343-4295

ISBN: 978-952-7496-01-5

Foreword

Mika Tynkkynen
director, TVK

TAccording to Finnish Workers' Compensation Act, "The Finnish Workers' Compensation Center acts as the joint body for the implementation and development of the insurance". An important part of this role is to produce diverse and useful data and knowledge about the occupational accidents and diseases as well as the overall business environment of the workers' compensation scheme.

Risk and uncertainty are two fundamental concepts among the insurance business. According to the simple universal definition, the risk could be defined as the combination of the probability of occurrence of a defined threat or loss and the magnitude or severity of the consequences of the occurrence. Thus, comprehensive understanding of the factors affecting the probability and severity of the compensable losses is needed in order to manage successfully the risk based workers' compensation insurance risks.

This study deals with the latter component of the risk, that is, the severity. The aim of the study was to develop a binary classification model to predict the severity of a workplace accident based on the variables defined in the European Statistics on Accidents at Work (ESAW) methodology to describe the circumstances of occurrence of occupational accidents. One could list many valid and favorable arguments to find this aim worth pursuing. Briefly, gaining knowledge about the dynamics of the underlying causes affecting the severity of the accidents after the occurrence of the compensable claim events would help insurance companies act efficiently, swiftly and focus rehabilitation efforts on people with a higher risk of prolonged absence from work. This would benefit both insurance companies and employers due to rapid return to work and lower disability costs – not to mention the insured employee who gets proper medical treatment immediately after the accident and is

able to return to work and daily routines soon. Moreover, reliable knowledge of the factors affecting the severity of the accidents is highly valuable from the primary prevention point of view enabling efficient allocation and focusing of the occupational safety resources and procedures.

In this study, workplace accident outcomes were predicted using Tree-Structured Parzen Estimator optimized XGBoost algorithm. Outcomes were divided into serious and non-serious accidents based on the absence from work. Cases where the absence from work was more than 30 days, were considered serious.

The aims of the study were achieved. Despite its limitations, the TPE-optimized model could predict serious accidents with an accuracy of 73% and non-serious accidents with an accuracy of 77%. Less surprisingly, wounds and superficial injuries and bone fractures were found to be the most important features predicting the workplace accident outcome. In addition to the potential financial benefits and other practical implications, the model itself and overall findings of the study enhance knowledge and gain our understanding about the complex dynamics of the factors affecting the severity of the accidents. Moreover, this study represents a new approach to promote our efforts to find new methods in analyzing workers' compensation insurance data.

Abstract

Antton Koskinen: Predictive modeling of workplace accident outcomes utilizing XGBoost and Tree-Structured Parzen Estimator

Aalto University School of Business
Master's Thesis, pages 61+12
Spring 2022
Degree: Master's degree
Degree programme: Information and Service Management
Thesis advisor(s): Pekka Malo
Keywords: XGBoost, TPE, Workplace accidents, ESAW

Workplace accidents induce a cost of hundreds of millions of euros for insurance companies annually and indirectly even higher costs for employers and society, including human suffering. However, most of these costs are driven by the employee's recovery time, and by advancing employees returning to work, the cost of workplace accidents can be reduced.

From the insurance companies' perspective, employees' return to work can be advanced by helping the accident victim get appropriate care as soon as possible by improving the administrative process.

One way to advance the administrative process is by giving priority to cases involving a higher risk of prolonged absence from work. Therefore, this research aims to develop a prediction model to identify victims of workplace accidents that are likely to suffer a prolonged absence from work to help direct resources where most needed.

The seriousness of the accident was defined based on the absence from work, and accidents, where the absence was more than 30 days, were considered serious. Following this, a binary classification model was developed utilizing ESAW variables described in the accident notice submitted to the insurance company.

The model used in this research was XGBoost, and it was optimized using Tree-Structured Parzen Estimator (TPE). In addition, the model was trained using accident notifications delivered to insurance companies and collected by the Finnish Workers Compensation Center.

The model could predict serious accidents with an accuracy of 73% and non-serious accidents with an accuracy of 77%.

Abstrakti

Antton Koskinen: Predictive modeling of workplace accident outcomes utilizing XGBoost and Tree-Structured Parzen Estimator

Aalto University School of Business
Master's Thesis, pages 61+12
Spring 2022
Degree: Master's degree
Degree programme: Information and Service Management
Thesis advisor(s): Pekka Malo
Keywords: XGBoost, TPE, Workplace accidents, ESAW

Työpaikkatapaturmat aiheuttavat vuosittain satojen miljoonien eurojen kustannukset vakuutusyhtiöille, sekä epäsuorasti vielä suuremmat kustannukset työnantajille ja yhteiskunnalle, mukaan lukien inhimillisen kärsimyksen. Suurin osa näistä kustannuksista riippuu työntekijän työkyvyn palautumisesta ja nopeuttamalla työntekijän työhön paluuta työpaikkatapaturmien kustannuksia voidaan pienentää.

Vakuutusyhtiöiden näkökulmasta työntekijän työhön palaamista voidaan edistää tarjoamalla tapaturman uhrille soveltuvaa hoitoa mahdollisimman nopeasti nopeuttamalla hallinnollisia prosesseja.

Yksi tapa parantaa hallinnollisia prosesseja on asettaa tapaukset etusijalle, joissa riski pitkittyneeseen työpoissaoloon on korkea. Tämän pohjalta tutkimuksessa kehitetään ennustemalli tunnistamaan kyseiset korkean riskin työpaikkatapaturmat.

Työpaikkatapaturmien vakavuus määriteltiin työkyvyttömyyden pituuden mukaan, jolloin tapaturmia, jotka johtivat yli 30 päivän työkyvyttömyyteen määriteltiin vakaviksi. Tämän määrittelyn perusteella kehitettiin binäärinen luokittelumalli hyödyntämään vakuutusyhtiöille toimitettavien vahinkoilmoitusten sisältämiä ESAW-muuttujia.

Tutkimuksessa sovellettiin XGBoost-algoritmia, joka optimoitiin käyttäen Tree-Structured Parzen Estimator-algoritmia. Malli opetettiin käyttäen Tapaturmavakuutuskeskuksen aineistoa vakuutusyhtiöille toimitettavista vakuutusilmoituksista.

Malli pystyi ennustamaan vakavia työpaikkatapaturmia 73% tarkkuudella ja ei-vakavia työpaikkatapaturmia 77% tarkkuudella.

Table of content

List of tables

List of figures

1. Introduction

In 2021, insurance institutions paid compensation for around 91 800 workplace accidents (Finnish Workers' Compensation Center, 2021) concerning more than 3% of the employed workforce of Finland (Statistics Finland, 2021). Considering this large number of victims, it is no surprise that workplace accidents induce a cost of hundreds of millions of euros for insurance companies annually and indirectly even higher costs for employers and society, including human suffering.

Most of the costs associated with workplace accidents are driven by the employee's recovery time, and by advancing employees returning to work, the cost of workplace accidents can be reduced. Furthermore, the employee's return to work can be advanced by improving the care chain, including the administrative, diagnosis, and treatment processes, and especially by providing appropriate medical care as soon as possible. However, there is still much to improve in the care chain concerning workplace accidents. For example, in the case of knee injuries, only 55% of the patients get MRIs within two weeks of the accident, and only 5 % receive surgery within 1-to 2 weeks as recommended (Pietilä, 2018).

From the insurance companies' perspective, employees' return to work can be advanced by helping the accident victim get appropriate care as soon as possible by improving the administrative process. One way to advance the administrative process is by prioritizing cases involving a higher risk of prolonged absence from work by using machine learning to predict such accidents. In the event of a workplace accident, the employer must report the claim event to the insurance company. These accident notices can be used to predict the outcome of a workplace accident.

Finnish Accidents at Work and Occupational Diseases Act requires employers to notify the insurance company about the claim incident and set requirements for the accident notification variables. However, the accident notification form must also follow the requirements set by European Commission. The required variables describing the workplace accident by the European Commission Regulation No 349/2011 are referred to as ESAW-variables.

European Statistics on Accidents at Work (ESAW) project was launched to harmonize data on accidents at work for all accidents resulting in more than three days of absence from work. Its primary purpose was to provide up-to-date descriptions and references among members of the European Union. According to the framework, the data must cover the characteristics of the injured person, injury, enterprise, workplace, and accident. Furthermore, the characteristics of the accident must include the sequence of events characterizing the causes and circumstances of the accident (Eurostat, 2013, p. 5).

Finnish Workers' Compensation Center has collected accident notices submitted to insurance companies for over 20 years, being the official authority for statistics concerning occupational accidents and diseases in Finland. In this research, we are going to use this data to predict workplace accident outcomes that can help the insurance companies to advance their administrative process leading to a shorter recovery time for the accident victims.

1.1. Objective and research questions

This research aims to develop a binary classification model to predict the outcome of a workplace accident based on the ESAW variables described in the accident notice submitted to the insurance company. In this research, accident outcomes were classified as serious and non-serious where accidents that lead to more than 30 days absence from

work are considered serious. According to the presented research objective the research questions are as follows:

1) How well can ESAW variables be used to predict the outcome of a workplace accident?

2) What are the most crucial ESAW variables for predicting the outcome of a workplace accident?

3) Can the prediction model provide financial benefits to an insurance company?

2. Background

This part chapter will discuss the definition of workplace accidents, the care chain concerning workplace accidents, the effects of delays in the administrative -and treatment process, the cost of workplace accidents, and the previous research on accident outcome prediction.

2.1. Definition of a workplace accident

Finnish Accidents at Work and Occupational Diseases Act defines workplace accidents as incidents that result in an effect causally linked to the incident and are suffered by employees during activities generally associated with being at the workplace. This definition also includes accidents that occur during work-related travel, performing the duties of an employee representative, or performing tasks on behalf of the employer (Finnish Accidents at Work and Occupational Diseases Act 2015/459 § 21-24).

An accident means a sudden and unforeseen event arising from an external factor that causes the employee to be injured or develop an illness (Finnish Accidents at Work and Occupational Diseases Act 2015/459 § 17). Under certain circumstances, muscle or tendon pain can also be considered an accident, even when an apparent external factor is not demonstrated. In most cases, the effect is an apparent physical injury, but the impact may also be psychological, such as an acute stress reaction to a threatening event at work (Salo, 2015, p. 46).

2.2. The care chain

The care chain includes all possible administrative- and medical steps from handling the insurance claim to treatments supporting a workplace accident victim's recovery. This

part will cover the essential elements of the care chain, including the insurance claim process and steps after and before the compensation decision.

2.2.1. The process of handling insurance claims

When the employer pays or has agreed to pay wages amounting to more than 1300 euros in a calendar year Workers' Compensation Act requires employers to insure their employees against accidents at work and occupational diseases (Finnish Accidents at Work and Occupational Diseases Act 2015/459 § 3).

In the event of a workplace accident, the employee must report the claim event to the employer, who notifies the insurance company. Thus, the claim is formally initiated by the employer's notice, launching the claim process. The employer must file the claim without delay and no later than ten working days from when the employer was made aware of the accident (Finnish Accidents at Work and Occupational Diseases Act 2015/459 § 110-111).

According to the Workers' Compensation Act, the claim must specify the name and contact details of the employer, name of the injured employee, personal identity code, contact information, and details of the accident (Finnish Accidents at Work and Occupational Diseases Act 2015/459 § 111). Furthermore, the accident details are reported according to the standards set by the European Statistical Office (Eurostat) following European Statistics on Accidents at Work (ESAW) framework.

The insurance company must issue its compensation decision no later than 30 days after receiving adequate information to resolve the matter. If the time limit is exceeded, the insurance company will pay an increase for the delay (Finnish Accidents at Work and Occupational Diseases Act 2015/459 § 127). In this research, we chose the time limit for the compensation decision as a threshold for serious accidents.

To advance the recovery, it is essential to direct the workplace accident victim to appropriate medical care as soon as possible. From the insurance company's perspective, this can be done by helping the injured person to get diagnosed as quickly as possible and making the compensation decision as soon as possible based on that information.

2.2.2. Before compensation decision

After the accident, the employee can choose either private or public medical institutions for the initial treatment and diagnosis. In both cases, the employer gives an insurance certificate to the employee who can use it to receive free of charge treatment in the medical institution. Although the employee will have to pay for the treatment and the medication without the insurance certificate, the insurance company will compensate for the expenses against receipt (Salo, 2015, p. 83-84).

Generally, the insurance company must be notified before receiving medical treatment; otherwise, the treatment is not compensable (Finnish Accidents at Work and Occupational Diseases Act 2015/459 § 146, 110). However, in the case of general care or emergency care, the insurance company does not need to be notified beforehand. Emergency care means treatment cannot be delayed without worsening the injury. General care consists of medical appointments and minor treatments like ultrasonography or radiography that cost less than 300 euros in a private institution. (Finnish Accidents at Work and Occupational Diseases Act 2015/459 § 43-44).

General care or emergency care does not require payment commitment from the insurance company (Finnish Accidents at Work and Occupational Diseases Act 2015/459 § 44). In other cases, medical care received in private medical institutions will not be fully compensated and can only amount to the service fee charged in public medical institutions (Finnish Accidents at Work and Occupational Diseases Act 2015/459 § 45).

The medical institution formulates a plan for medical rehabilitation, treatment, or medical examinations that it sends to the employee's insurance company together with the medical report. The employer's insurance company can then make the compensation decision based on the medical and accident reports (Salo, 2011, p. 84).

Although the payment commitment is needed for further treatment in a private medical institution, this is not the case with public medical institutions since the municipality provides medical care for the patient. Public medical institutions cannot delay the treatment even if the insurance company has not made the compensation decision (Occupational Health Care Act 1326/2010).

2.2.3. After compensation decision

In favorable compensation decisions, insurance companies are fully responsible for the cost of the treatment process, including medical rehabilitation (Finnish Accidents at Work and Occupational Diseases Act 2015/459 § 37). Apart from the medical cost, insurance companies are also responsible for compensating for the loss of income incurred by the workplace accident (Finnish Accidents at Work and Occupational Diseases Act 2015/459 § 56). In most cases, compensation claims lead to a favorable compensation decision. For example, in 2018, only 10 percent of claims reported to insurance companies were rejected on legal or medical grounds (Finnish Workers' Compensation Center, 2021).

Insurance companies can direct the employee to a private medical institution to support recovery. In such a case, the insurance company will give the employee a payment commitment to a different medical institution in these cases. Payment commitment will ensure that the accident is compensable (Finnish Accidents at Work and Occupational Diseases Act 2015/459 § 45, 42). Apart from medical treatments and operations, the supporting processes include rehabilitation counseling, therapies to improve and

maintain functional capacity, medical rehabilitation aids, adaptation training, and rehabilitation episodes in institutional or outpatient care.

In the event the employee is unable to return to work, the employee is generally compensated in the form of daily allowance or pension (Finnish Accidents at Work and Occupational Diseases Act 2015/459 56, 59 §). In addition, the employee will also be compensated in the case of a general permanent functional limitation caused by the claim event (Finnish Accidents at Work and Occupational Diseases Act 2015/459 83 §) and for the cost of vocational rehabilitation if returning to work requires it (Finnish Accidents at Work and Occupational Diseases Act 2015/459 89 §).

2.3. The cost of workplace accidents

The direct cost of workplace accidents to the insurance companies can be divided into temporary and permanent costs. Temporary costs are realized within one year of the accident, and permanent costs during future accounting periods.

Table 1: Compensation paid by insurance companies concerning workplace accidents

Year	2017	2018	2019	2020
Workplace accidents	102 226	102 274	103 131	86 606
Medical care	86 204	87 812	86 171	82 699
Daily allowance	101 214	100 051	96 137	111 094
Rehabilitation allowance	559	791	622	802
Other	-503	2 283	6 117	2 569
All temporary costs	**187 475**	**190 938**	**189 046**	**197 164**
Compensation for functional limitation	8 844	9 582	7 838	10 538
Disability allowance	21 849	22 509	18 922	24 391
Confirmed pensions	113 954	117 754	114829	136 406
Rehabilitation allowance	6 095	6 283	6 567	8 783
Funeral assistance	621	429	475	513
The lump-sum compensation for disability	84	58	93	111
Rehabilitation	12 943	12 301	11 242	11 640
Other	1 261	1 782	1 968	3 141
All permanent costs	**165 653**	**170 698**	**161 933**	**195 523**
Other	88 667	92 633	142 076	107 612
Total cost	**353 128**	**361 636**	**493 056**	**392 687**

[Values are approximated from the occupational accident statistics by the proportion that workplace accidents represent them (Finnish Workers' Compensation Center, 2021). All units are in thousands except workplace accidents. Other cost that are neither permanent nor temporary consist of index increments]

2.3.1. Temporary cost

The main types of temporary costs consist of medical care, daily allowance, rehabilitation allowance, and another cost.

Medical care includes emergency care, examination, diagnosis and treatment of the injury or illness, medications, treatment supplies, and medical rehabilitation. Medical rehabilitation includes rehabilitation guidance, assessment of rehabilitation needs, therapies to improve and maintain functional capacity, medical rehabilitation aids, adaptation training, and rehabilitation in institutional or outpatient care (Finnish Accidents at Work and Occupational Diseases Act 2015/459 37 §).

In an occupational accident, the employee is entitled to a daily allowance if the employee's work capacity has lowered more than 10 percent. If the absence from work is less than four weeks, the daily allowance is the same as the regular income. When the absence from work is more than four weeks, the daily allowance is 1/360 of the yearly income, and it can be paid for one year (Finnish Accidents at Work and Occupational Diseases Act 2015/459 § 56). Daily allowances are the most significant expenditure and have the most considerable potential for cost savings (Table 1). Focusing rehabilitation efforts on people with a higher risk of prolonged absence from work makes it possible to advance returning to work and thus decrease the cost of daily allowances.

The injured person has the right to claim a rehabilitation allowance during vocational rehabilitation. The rehabilitation allowance corresponds to the total amount of the daily allowance for one year from the claim event date (Finnish Accidents at Work and Occupational Diseases Act 2015/459 69 §). In addition, the injured person is also compensated for costs of vocational rehabilitation required to help the injured person continue in their previous work or occupation or transfer to new work or an occupation (Finnish Accidents at Work and Occupational Diseases Act 2015/459 89 §).

Other cost includes travel and accommodation costs arising from compensated medical treatment (Finnish Accidents at Work and Occupational Diseases Act 2015/459 50 §), compensation for the unavoidable additional costs of housekeeping (Finnish Accidents at Work and Occupational Diseases Act 2015/459 53 §), compensation for medical aids used by the injured person and damaged in connection with the claim event. Other costs include torn clothes or broken rings during medical treatment (Finnish Accidents at Work and Occupational Diseases Act 2015/459 54 §).

2.3.2. Permanent cost

The main types of permanent costs are compensation for functional limitation, disability allowance, rehabilitation allowance, funeral assistance, rehabilitation, pensions, and other costs. Rehabilitation allowance, rehabilitation, and other costs are defined as before, but have been paid later than a year after the accident.

Compensation for functional limitation is paid to an injured person who suffers a general permanent functional limitation because of an injury or illness caused by the claim event. Compensation for functional limitation does not compensate for deterioration of the ability to work caused by the claim event, costs arising from the need for care or assistance, or other injuries compensated separately (Finnish Accidents at Work and Occupational Diseases Act 2015/459 83 §).

When the employee's work capacity has lowered more than 10 percent for more than a year, the employee is entitled to an occupational injury pension (Finnish Accidents at Work and Occupational Diseases Act 2015/459 § 59). Disability allowance concerns cases where the injury pension has not yet been confirmed, but work absence has lasted more than a year. Out of all forms of compensation, pensions are the highest cost (Table 1).

2.3.3. Indirect cost

Workplace accidents also induce indirect costs for employers. These costs can include reduction in production, decreased sales, and cost for replacing the employee in paid overtime and temporary staff. Employers may also be liable to pay compensation if the employer's negligence caused the workplace accident. Indirect costs have been estimated to be around 3-4 times bigger, corresponding to 1.5-2 billion euros annually (Rissanen and Kaseva, 2014, p. 6). Indirect costs are also linked to the length of the absence from work, and so by advancing the return to work, it is possible to decrease the indirect costs as well.

2.4. Consequences of administrative delays

Administrative delays in the compensation claim process may delay the employee's recovery, inducing costs for insurance companies, employees, and employers alike. There have been some studies where the consequences of administrative delays have been examined.

Sinnott assessed compensation claims from the California Workers' Compensation Institute (CWCI) to determine whether claims acceptance or administrative delays influenced outcomes for individuals with acute back injuries. Beyond the first two weeks, each interval of administrative delay was associated with increased odds of developing chronic disability. For example, between 2 and 4 weeks, the adjusted odds ratio was found to be 1.433 (1.327-1.547) (Sinnott, 2009, p. 694).

Stover identified predictive factors of long-term disability using administrative data from the Washington State Department of Labor and Industries. Workers with four or more days of work disability resulting from workplace injuries were followed for approximately six years. More than 20 days delay from the first medical visit to claim receipt was discovered to be a significant predictor of work disability with an adjusted odds ratio of

1.36 (1.29-1.45). In contrast, days to injury to first medical visit between 11 and 20 days, the odds ratio was 1.43 (1.32-1.54) and 1.81 (1.70-1.92) for more than 20 days (Stover et al., 2007, p. 35-36).

Even though these results are not completely applicable to the Finnish insurance systems where every employee is insured, and thus the barriers to seeking medical help are lower, they still highlight the importance of an efficient insurance claim process.

2.5. Consequences of delays in the treatment

There exist a multitude of medical research focusing on the effects of delayed treatment. However, since we cannot cover all possible injury types, we focus on knee and shoulder injuries that are the most injured body parts among serious accidents.

Hantes et al. discovered that early repair of traumatic rotator cuff tears (RCT) provides better results in terms of shoulder function in comparison with delayed repair. A delayed diagnosis of a traumatic RCT leads to difficulties in surgery and less good results (Hantes, 2011). Peterson's research likewise implicates that earlier RCT repairs are associated with better recovery results (Peterson 2011). Vastamäki (2002) points out that swift diagnosis of RCT is essential since the delays significantly affect the outcome of the surgery even though contemporary research suggest physiotherapy for RCT (Rotator cutoff: Current Care Guidelines, 2014).

Knee injuries are essential to diagnose and treat as soon as possible to avoid long absence from work or complications (Ristiniemi, 2018). Ristiniemi suggests surgery within 2 or 3 weeks for anterior cruciate ligament (ACL) trauma and lateral collateral ligament (LCL) trauma. Research conducted by Lin et al. also indicates that ACL-injured patients should undergo ACL reconstruction as early as possible (within one month) to lower the risk of knee osteoarthritis (Lin et al., 2017). Hohmann et al.'s research points in the same direction that early surgical intervention in multi-ligament injuries of the knee

produces a significantly superior clinical outcome compared to late reconstruction (Hohmann et al., 2017). Even in the less severe cases of acute dislocations of knee or ACL trauma, MRI should be done immediately to identify possible associated injuries (Ristiniemi, 2018).

In the light of prevailing research, shoulder and knee injuries should be diagnosed and treated soon as possible to support the patient's recovery to work. Contrary to this knowledge Pietilä, who analyzed Finnish occupational accidents and diseases statistics, discovered that surgeries that were recommended to do within two months from the time of the accidents were often delayed. In the case of RCT injuries, only 5 % of patients received surgery within recommended 1-2 weeks as recommended (Pietilä, 2018, p. 129), and for ACL injuries, only 15% within recommended 3-5 weeks (Pietilä, 2018, p. 126). It is not clear why the treatment is often delayed. One reason could be shortcomings in the diagnosis process. In the case of knee injuries, only 37% of the patients had MRI within one week of the accident and 55% within two weeks, and for shoulder injuries, only 38% of the patients had MRI within one week of the accident and 66% within two weeks (Pietilä, 2018, p. 125). Even though Pietilä's research concerns all occupational accidents, we can assume that the results also apply to workplace accidents.

According to the research presented in this chapter, it is safe to say that delays in treatment affect the recovery for all injuries to a varying degree, obscure the correct diagnosis, and affect the time of surgery or other medical procedures. Therefore, it seems that the care chain concerning the treatment of workplace accidents in Finland could be improved by providing patients with faster diagnosis and treatment.

2.6. Previous research on accident outcome prediction

This part will discuss previous research on accident outcome classification and pattern extraction. Predicting safety outcomes has been done before, but primarily for specific industries. Here we focus on studies where the data size was at least 1000 accidents.

Anurag et al. (2020) used Logistic regression (LR), Decision tree (DT), Random Forest (RF), and Artificial neural network (ANN) models to predict mining accident outcomes in the United States utilizing structural data like ESAW-variables. Accident outcomes were divided into nine different classes based on the length of absence from work. ANN achieved the best prediction accuracy with structured data with an accuracy of 78% and an F1-score of 0.67.

Sobhan et al. (2019) optimized SVM and ANN algorithms using particle swarm optimization (PSO) and genetic algorithms (GA) to predict accident outcomes in the Indian steel industry. Accident outcomes were classified as injury, near misses, and property damage. The data consisted of 15 categorical features and text variables. SVM outperformed ANN with both optimization methods, and particularly PSO-SVM performed the best with an accuracy of 90.67%.

Matias et al. (2008) analyzed workplace falls in Spain using decision tree (DT), support vector machine (SVM), extreme learning machine (ELM), and Bayesian network (BN) algorithms. In their research, BN was the best classifier in predicting accidents and extracting important factors behind accidents.

There have been some studies where data mining techniques have been applied to Finnish occupational accidents and diseases statistics data. For example, Nenonen (2011) examined association rules related to slipping, stumbling, and falling accidents at work. Furthermore, Rojas et al. (2018) used the same technique and data to study construction accidents in Finland.

3. Data

In this chapter, we will describe the data used in this research and the justification for selecting it.

3.1. Description of the data and variables

Finnish Workers Compensation Center's occupational accident data set consists of around 2.7 million workplace accident notices from 1999 to 2020. The data set has more than 20 variables, including, in addition to the ESAW variables, time of the accident, industry, occupation, gender, age group, corporate id, sector, regional state administrative agency, municipality, and the absence from work resulting from the accident. However, most of these variables were not considered to affect the accident outcome and were discarded.

The data set includes eight ESAW variables that describe the causes and circumstances of accidents at work: working process, specific physical activity, deviation, contact, and mode of injury, material agent associated with the mode of injury, workstation (until 2015), working environment (from 2016), type of injury, and part of the body injured.

Over the years, there have been several changes to the Finnish Workers' Compensation Center statistics of occupational accidents and diseases. For this reason, the data is not perfectly coherent throughout the years and requires some consideration before utilizing the data.

The most notable changes happened in 2005 when a new legislative reform took place. After the legislative reform, treatments done in public medical institutions were priced according to their actual cost instead of previously charged fixed customer payment. Before the legislative reform, the employees paid the customer payment by themselves.

After the legislative reform, the whole treatment is paid for by the employee's insurance company. Making treatment in public medical institutions free for employees naturally increased the number of claims from minor injuries that do not lead to lost income compensation. Because of this, we disregarded data before 2005.

A second significant change occurred in 2016 when the classification of occupations changed to correspond to the classification used by Statistics Finland. Unfortunately, the occupational classification standard used before 2016 is difficult to convert to be fully comparable with the new standard. Therefore, the variable describing occupation was discarded.

A third significant change happened in 2016 when a working environment variable replaced the workstation variable. Because of this neither variable covers the whole analysis period, and both variables describing the working environment were discarded.

The final data set consists of 9 predictive variables: working process, specific physical activity, deviation, contact, mode of injury, a material agent associated with the mode of injury, type of injury, part of the body injured, age, and gender. In addition, the target label was formed using the variable describing absence from work caused by workplace accidents.

After removing missing data, variable levels that had less than ten observations, and accidents that resulted in death, the data consisted of around 1.47 million accidents.

3.2. Distributions of the variable levels

This part examines the class distribution and the ratio between proportions of each class level for serious and non-serious accidents. The ratio tells if any class level is represented more among serious accidents than non-serious accidents and vice versa.

We also analyze the distribution of the variable describing the length of absence from work used to from the target labels.

Variables can be divided into three groups based on what stage of the accident they are describing: 1) variables that describe the outcome of the accident, including injury type and body variable 2) variables that describe the mechanism of the accident, including material agent associated with the mode of injury, deviation, contact mode of the injury and specific physical activity 3) variables describing the injured person before the accident including age, gender and working process.

Even though missing features (XX) are removed from the data when training the model, we include them when examining the variables.

3.2.1. Material agent associated with the mode of injury

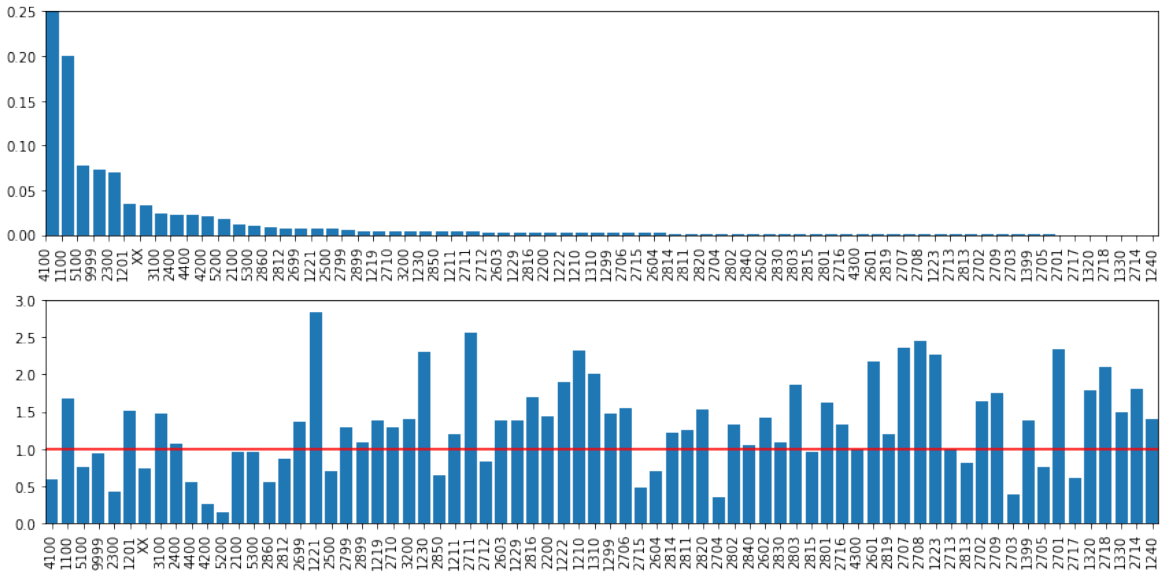

Figure 1. Material agent associated with the mode of injury. The upper plot describes the distribution of class levels where the x-axis represents class levels and y-axis proportions. The lower plot describes the ratio between level frequencies of serious and non-serious accidents where x-represents class levels and the y-axis is the ratio between proportions of serious and non-serious accidents. Feature descriptions can be found from Table A1.

Material agent associated with the mode of injury means the object, tool, or instrument with which the victim came into contact at the accident. We can see that the material agent variable is very unevenly distributed, with ten (out of 71) most frequent levels covering more than 80% of the cases (Fig. 1).

Materials, objects, products, machine or vehicle components, debris, or dust (4100), and buildings, structures, and surfaces at ground level (1100) are the most frequent material agents associated with the mode of injury, covering around 40 % of all the cases. However, this is not surprising since both classes cover many possible material agents. Buildings, structures, and ground-level surfaces (1100) are more common among serious accidents, and materials, objects, products, machine or vehicle components, debris, and dust (4100) are more common among non-serious accidents (Fig. 1).

Multiple material agents are substantially more common among serious accidents than non-serious accidents, but almost all of them appear very rarely. Mobile ladders (1221), temporary scaffolds (1230), fixed ladders (1210), mobile scaffolds (1223), fixed machine tools for sawing (2711), portable or mobile machines for extracting materials, working the ground, and civil engineering works (2601), fixed forming machines for pressing and crushing (2707), machines for calendering, rolling and cylinder pressing (2708) and fixed machines for extracting materials and working the ground (2701) are all twice as common among serious accidents than among non-serious accidents (Fig. 1). We can notice that these features describe places above ground level (1221, 1230, 1210, 1223) or machines (2701, 2707, 2601, 2711, 2708).

Bulk waste (5200), chemical, explosive, radioactive, biological substances (4200), and hand tools (2300) are substantially more common among non-serious accidents when considering more frequent material agents (Fig. 1).

The problem with the material agent associated with the mode of injury variable is that it has multiple features, but most of them appear infrequently. Some of these features

could share a similar accident mechanism and be combined. Mainly features describing materials above the ground level seem intuitively similar. However, in this research, we do not want to make any assumptions about the accident mechanism.

3.2.2. Working process

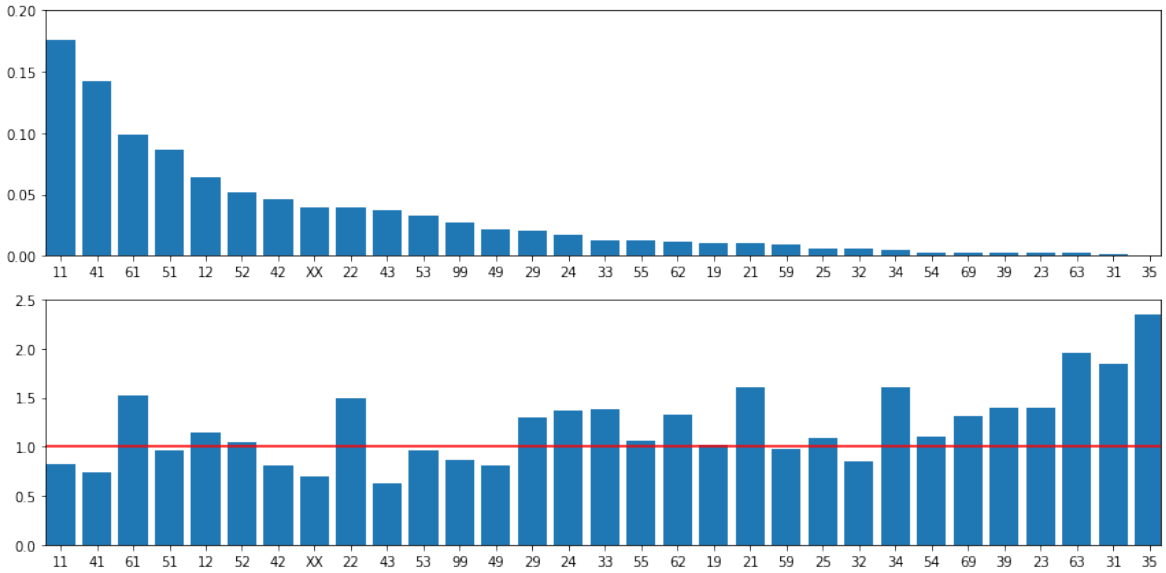

Figure 2. Working process. The upper plot describes the distribution of class levels where the x-axis represents class levels and y-axis proportions. The lower plot describes the ratio between level frequencies of serious and non-serious accidents where x-represents class levels and the y-axis is the ratio between proportions of serious and non-serious accidents. Feature descriptions can be found from Table A4.

The working process is the primary type of work or task (general activity) performed by the victim before the accident. It describes in broad terms the kind of work the victim was performing during a period ending at the instant of the accident.

Production, manufacturing, and processing (11) and service, care, or assistance to the general public (41) are the most common working processes (Fig. 2).

The differences between serious accidents and non-serious accidents are not generally significant. Sailoring (63), agriculture type of work (31), and fishing (35) are the only

working processes where the proportion of serious accidents can be considered quite large. However, those working processes also have the smallest number of observations (Fig. 2).

Among the more frequent classes, movement (61) and new construction buildings (22) are more common among serious accidents. On the other hand, buying, selling, and associated services (43) are considerably more common among non-serious accidents (Fig. 2).

3.2.3. Part of the body injured

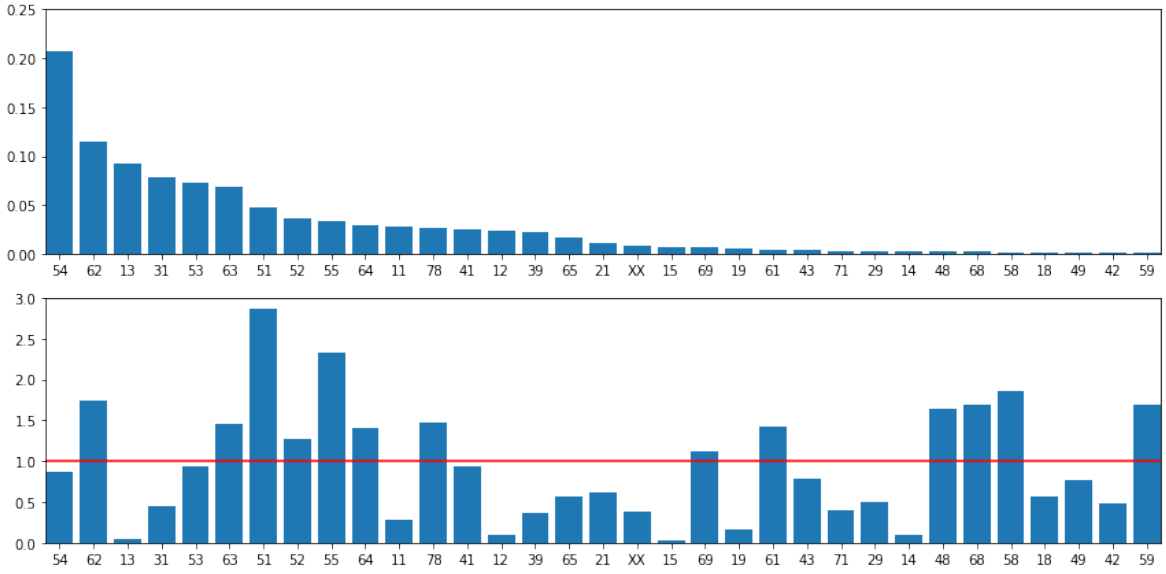

Figure 3. Part of the body that was injured. The upper plot describes the distribution of class levels where the x-axis represents class levels and y-axis proportions. The lower plot describes the ratio between level frequencies of serious and non-serious accidents where x-represents class levels and the y-axis is the ratio between proportions of serious and non-serious accidents. Feature descriptions can be found from Table A6.

Body variable is also unevenly distributed, with the ten most frequent levels covering almost 80% of the cases. Among these features, fingers (54) are the most injured body

part with 20% frequency and more common among non-serious accidents. At the same time, legs, and knee (62) injuries are the most common type of serious injuries (Fig. 3).

Shoulder and shoulder joints (51) and arm, including elbow (52), are more than twice as common among serious accidents than among non-serious accidents, and overall hand injuries seem to be more heavily represented among serious accidents (52, 53, 58, 59).

On the other hand, head injuries (11, 13, 12, 15, 19, 14) are several times more frequent among non-serious accidents than serious accidents (Fig. 3). Especially eye (13) and facial area (12) injuries are prevalent among non-serious accidents.

3.2.4. Deviation

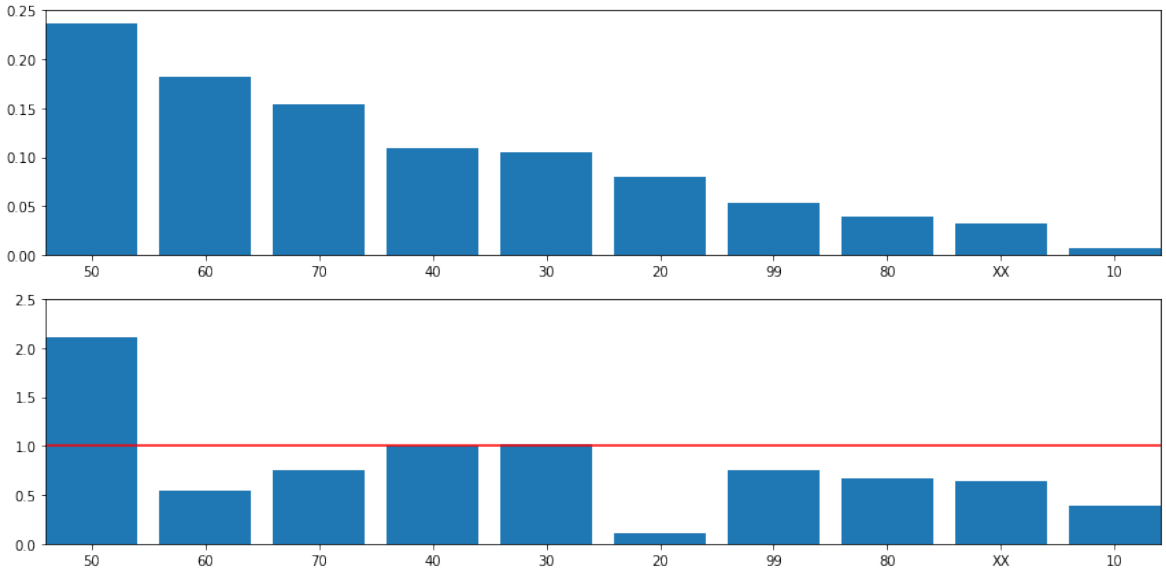

Figure 4. Deviation. The upper plot describes the distribution of class levels where the x-axis represents class levels and y-axis proportions. The lower plot describes the ratio between level frequencies of serious and non-serious accidents where x-represents class levels and the y-axis is the ratio between proportions of serious and non-serious accidents. Feature descriptions can be found from Table A5.

Deviation means the last event differing from the norm and leading to the accident. The deviation is the event that triggers the accident; however, the previous deviation is recorded if a chain of events leads to the accident.

Slipping, stumbling, and falling (50) is the most common deviation and the only type of deviation that is more common among serious accidents than non-serious accidents (Fig. 4). On the other hand, deviation by overflow, overturn, leak, flow, vaporization, emission (20) is several times as common among non-serious accidents than serious accidents. Body movement without physical stress (60) is also more than twice as common among non-serious accidents (Fig. 4).

The problem with the deviation variable is that its classes are very general, and the classes may encompass different accident types that could have other accident outcome implications.

3.2.5. Physical activity

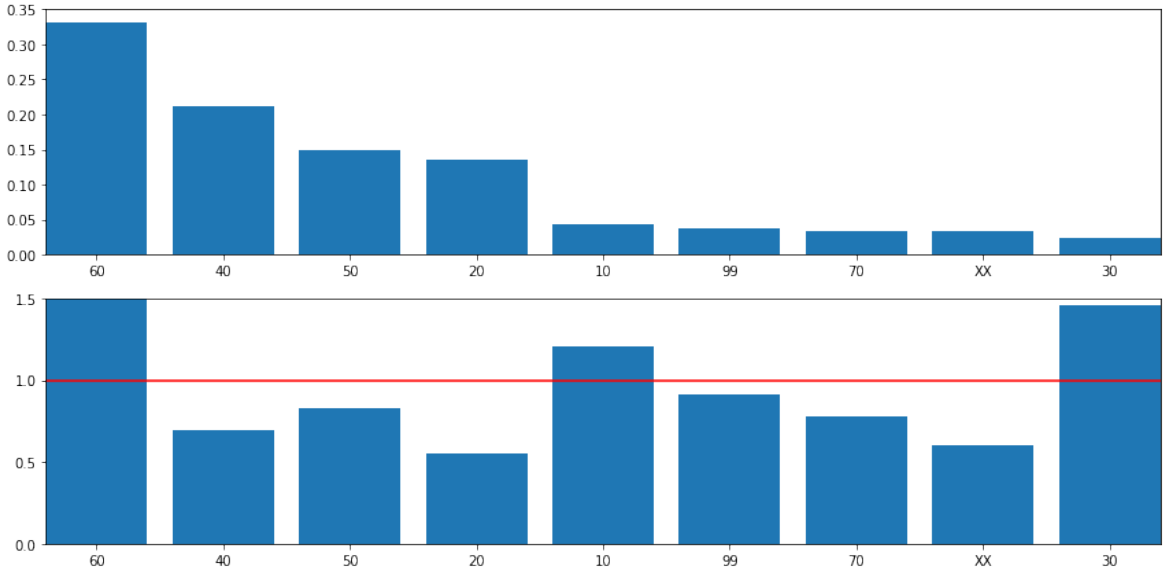

Figure 5. Specific physical activity. The upper plot describes the distribution of class levels where the x-axis represents class levels and y-axis proportions. The lower plot describes the ratio between level frequencies of serious and non-serious accidents where x-represents class levels

and the y-axis is the ratio between proportions of serious and non-serious accidents. Feature descriptions can be found from Table A3.

Specific physical activity refers to the action being performed by the victim at the exact time of the accident. It covers only a short period, while the working process describes a task performed over a substantial period. Specific physical activity is far more precise and can be isolated from the chain of events leading to the accident.

Movement (60) is the most common physical activity and the only variable with driving (30) that is more common among serious accidents than non-serious accidents (Fig. 5). On the other hand, handling objects (40) and working with hand-held tools (20) are the only variables significantly more common among non-serious accidents, the latter being twice as common among non-serious accidents (Fig. 5).

There are pretty slight differences between levels of specific physical activity-variable, indicating that it is not probably too relevant in predicting accident outcomes. One explanation is that the levels are too broad and do not accurately describe the physical activity. Another explanation is that physical activity is not an important variable when describing accident mechanisms.

3.2.6. Age

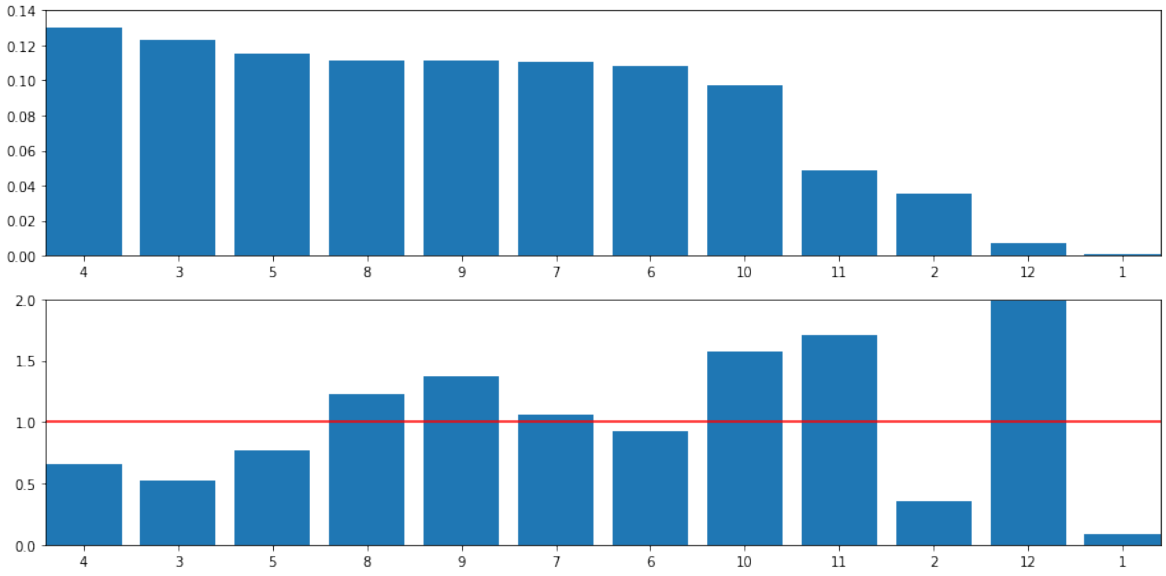

Figure 6. Age groups. The upper plot describes the distribution of class levels where the x-axis represents class levels and y-axis proportions. The lower plot describes the ratio between level frequencies of serious and non-serious accidents where x-represents class levels and the y-axis is the ratio between proportions of serious and non-serious accidents. Feature descriptions can be found from Table A8.

Age groups are ordered from youngest to oldest, with 1 being 0–14 years old and 12 over 64 years old.

Age groups are very evenly distributed apart from the two oldest and youngest age groups (1, 2, 11, 12). We can see clearly how the ratio between serious accidents and non-serious accidents raises with the increase of age, and for over 64 years old (12), serious accidents are twice as common. On the other hand, for the youngest groups (2, 1), non-serious accidents are several times more common (Fig. 6).

3.2.7. Contact mode of injury

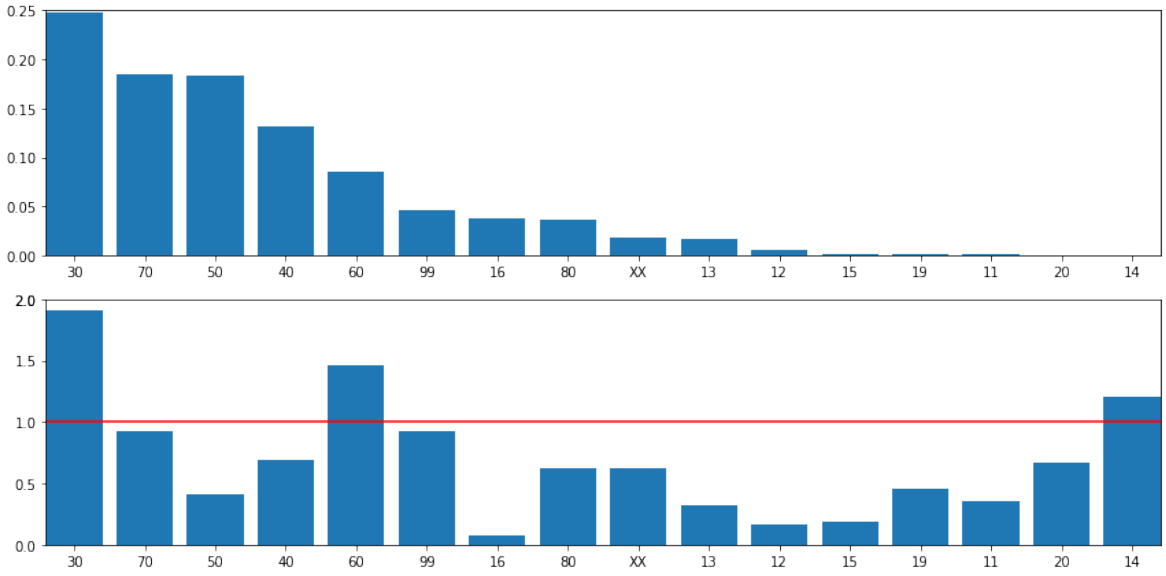

Figure 7. Contact mode of injury. The upper plot describes the distribution of class levels where the x-axis represents class levels and y-axis proportions. The lower plot describes the ratio between level frequencies of serious and non-serious accidents where x-represents class levels and the y-axis is the ratio between proportions of serious and non-serious accidents. Feature descriptions can be found from Table A7.

Contact mode of injury means the contact that injured the victim. It describes how the victim was hurt (physical or mental trauma) by the material agent that caused the injury. If there are several contact modes of the injury, the one causing the most serious injury is recorded.

Again, we notice a very uneven distribution of variable levels. Horizontal or vertical impact with or against a stationary object (30) is the most frequent contact mode of injury and the most common type of serious injury. Horizontal or vertical impact with or against a stationary object (30) and trapped or crushed (60) are the only variable levels that are somewhat more common among serious accidents (Fig. 7).

Contact with hazardous substances through skin or eyes (16), contact with hazardous substances by inhalation (15), contact with sharp, pointed, rough, or coarse material agent (50), direct contact with electricity (12), contact with naked flame or a hot or burning object or environment (13) and indirect contact with a welding arc, spark, lightning are all several times more common among non-serious accidents (Fig. 7).

3.2.8. Injury type

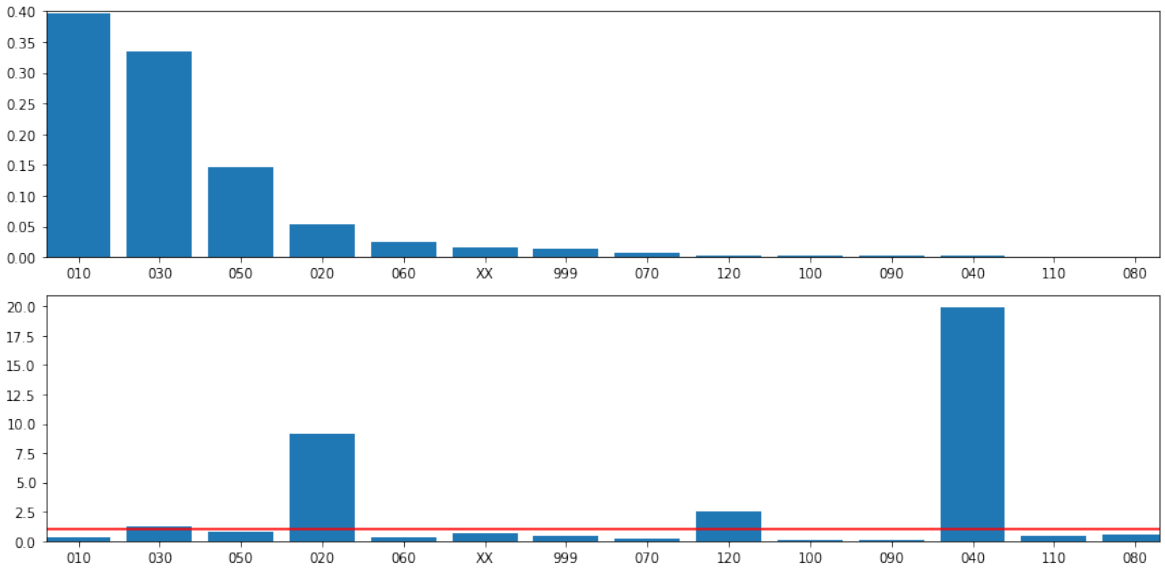

Figure 8. Injury type. The upper plot describes the distribution of class levels where the x-axis represents class levels and y-axis proportions. The lower plot describes the ratio between level frequencies of serious and non-serious accidents where the x-represents class levels and the y-axis is the ratio between proportions of serious and non-serious accidents. Feature descriptions can be found from Table A2.

Injury type describes physical consequences for the victim. In case of multiple injuries suffered in one accident where one of the injuries is more severe than the others, then this accident should be classified in the group corresponding to the nature of the more severe injury. Only in cases where the victim has contracted two or more injuries, and one of them cannot be said to be more serious than the other(s), the code 120 "multiple injuries" should be used.

Wounds and superficial injuries (010) and dislocations, sprains, and strains (030) are the most common type of injuries representing more than 70% of the accidents (Fig. 8).

Traumatic amputations (040) are about 20 times more common among serious accidents than non-serious accidents. In addition, bone fractures (20) are almost nine times more common among serious accidents than non-serious accidents. Also, cases with multiple injuries (120) are several times more common for serious accidents (Fig. 8).

Wounds and superficial injuries (010), effects of temperature extremes, light, and radiation (100), effects of sound, vibration, and pressure (090), and poisonings and infections (070) are all several times more common among non-serious accidents (Fig. 8).

The only features of the injury type variable that are not disproportionally distributed among serious and non-serious accidents are dislocations, sprains and strains (030), and concussion and internal injuries (050).

3.2.9. Gender

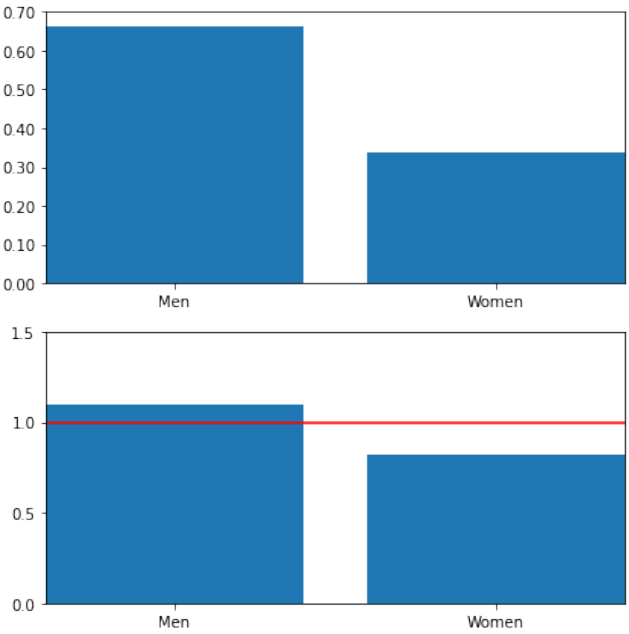

Figure 9. Gender. The upper plot describes the distribution of class levels where the x-axis represents class levels and the y-axis proportions. The lower plot describes the ratio between level frequencies of serious and non-serious accidents where x-represents class levels and the y-axis is the ratio between proportions of serious and non-serious accidents.

Men are represented around twice as frequently in the data compared to women. Men are also slightly more represented among serious accidents than women, likely because more men work in industries with higher accident frequency (Fig. 9).

3.2.10. Seriousness of the workplace accident

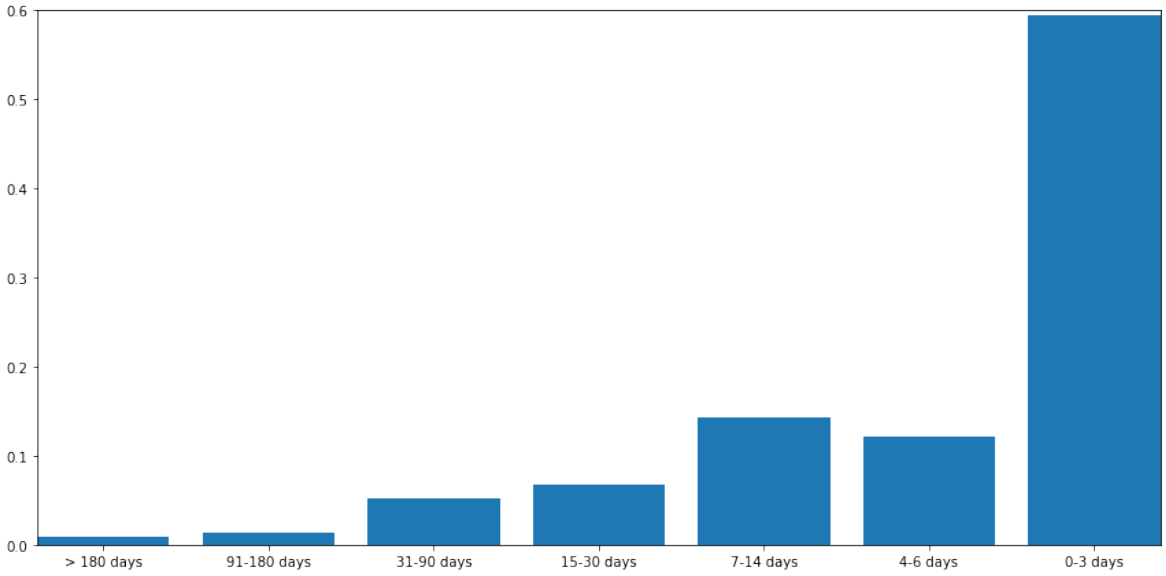

Figure 10. The length of absence from work. The y-axis describes frequency and x-axis day groups.

We can notice that accidents leading to 0-3 days absence from work are the most frequent, with almost 60% of the accidents belonging to that group. For this group, daily allowances are not paid. Interestingly accidents leading to 7-14 days absence are more common than accidents leading to 4-6 days absence from work (Fig. 10).

Around 8% of the accidents lead to an accident where the absence from work is more than 30 days and are considered serious from this research's perspective (Fig. 10).

4. Methods

In this part, we will explore methods used in this research. XGBoost is an implementation of gradient boosted decision trees. For this reason, we will first explain the basic ideas behind ensemble methods and decision trees and then proceed to an in-depth description of XGBoost to define the optimized parameters' purpose. After that, we cover hyperparameter optimization and relevant optimization algorithms concerning this research. Finally, we briefly discuss imbalance learning and its implications concerning this research.

4.1. Ensemble methods

Classification error is composed of two components: bias and variance. Often, these two components have a trade-off relationship where classifiers with low bias tend to have high variance and vice versa. The point of the ensemble system is to bypass this limitation by combing different classifiers' relatively fixed bias and then combining their outputs to reduce the variance. Ensemble systems work under the assumption that classifiers make different errors on each sample but generally agree on their correct classifications. Combining the classifier outputs reduce the error by averaging (some way) out the error components (Cha and Ma, 2012, p. 2-3).

There are two general paradigms of ensemble methods: boosting, where the weak learners are generated sequentially, and bagging, where the weak learners are generated in parallel. In this research, we are focusing on boosting.

The idea of boosting is to train a set of learners sequentially (vs. parallel) and combine them for prediction, where the later learners focus more on the mistakes of the earlier learners. Boosting algorithm begins with a simple high bias model initialized with a

constant value. It is then progressively made less biased by adding weak learners (Zhou and Zhi-Hua, 2012, p. 23).

Boosting for binary classification problems works by creating sets of three weak classifiers at a time. The first classifier, h_1, is trained on a random subset of the available training data. The second classifier, h_2, is trained with a subset of data correctly identified by h_1. The third classifier, h_3, is then trained with instances on which h_1 and h_2 disagree. These three classifiers are then combined through some method (Cha and Ma, 2012, p. 13).

There are many ways of combining ensemble members, but XGBoost uses a weighted sum over the ensemble members. Combining the classifier outputs does not necessarily lead to a better classification performance than the best classifier in the ensemble. However, it reduces our likelihood of choosing a classifier with poor performance (Zhou and Zhi-Hua, 2012, p. 68-69).

4.2. Decision trees

Decision trees are statistical models designed for supervised prediction problems where the model can be represented in a tree-like structure. Each internal node represents a split based on the values of one of the inputs, and the leaves represent the predicted target. At each node, one attribute is chosen to split the training data into distinct classes, and the new instance is classified by following a matching path to a leaf node. All cases reaching the same leaf are given the same predicted value or probability (Kotsiantis et al., 2006, p. 163). Lastly, to avoid overfitting, it is often desirable to prune the size of decision trees by removing sections of the tree that are non-critical and redundant to classify instances. Decision trees generally differentiate based on splitting variables and methods to prune the tree.

Decision trees have some advantages over some other supervised learning algorithms. Firstly, decision trees can be easily visualized and understood. Secondly, decision trees can handle missing values because they are part of the prediction rules, even though missing values are removed in this research. Thirdly, decision trees are fast compared to many other classification algorithms. Finally, decision trees can learn incrementally, which is useful when processing large data sets or streamed data (Aggarwal and Charu, 2015, Ch. 4.7).

Decision trees also have some significant weaknesses that should be considered during the model selection. Firstly, decision trees are volatile models, and minor changes in the training data set can cause substantial changes in the tree's structure even if the overall performance remains the same. The standard method to mitigate the instability problem is to create an ensemble of trees, where the prediction score is a sum of multiple decision trees (Pinheiro and Patetta, 2021, Ch. 3.2). Secondly, decision trees are prone to overfit, so an efficient way to prune the decision tree is needed (Kotsiantis et al., 2006, p. 163).

4.3. XGBoost

This part will cover XGBoost and its mathematical foundation to explain how the hyperparameters that are optimized in this research relate to the model.

4.3.1. Tree ensemble

The tree ensemble model in XGBoost consists of CARTs, where the prediction score is a sum of all these trees. CART is a non-parametric decision tree algorithm that considers all possible splits of the set into two disjoint and complementary subsets for discrete variables. In CART, each tree node is assigned the class label dominating within the node.

For a given data set $D = \{(x_i, y_i)\}$ ($|D| = n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R}$), where n is the number of data instances and m number of features the ensemble model can be expressed as following

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i) \;\; f_k \in \mathcal{F},$$

where $\mathcal{F} = \{f(x) = w_{q(x)}\}(q: \mathbb{R}^m \to T, w \in \mathbb{R}^T)$ is CART. Here q represents the structure of each tree that maps an example to the corresponding leaf index, T is the number of leaves in the tree and K is the number of weak learners, and specifically, in this case, trees. Each f_k corresponds to an independent tree structure q and leaf weights w (Chen and Guestrin, 2016, p. 2).

4.3.2. Objective function

To find the optimal solution for our machine learning problem, we need to measure the quality of this solution. An objective function does this. The objective function takes data and model parameters as arguments and returns a number representing solutions quality in terms of the objective function. The objective function provides a formal specification for the classification problem. In this case, the optimal parameters cannot be found exactly but can be approximated using an iterative algorithm.

To learn the set of function used in the model, we minimize the following regularized objective that consist of loss function and regularization term

$$\mathcal{L}(\phi) = \sum_{i}^{n} l(\hat{y}_i, y_i) + \sum_{k}^{K} \Omega(f_k),$$

where $l(\hat{y}_i, y_i)$ is the loss function and $\Omega(f_k)$ regularization term (Chen and Guestrin, 2016, p. 2).

4.3.3. Regularization

Regularization terms penalize the complexity of the model. The additional regularization term helps smooth the final learn weights to avoid over-fitting. When the regularization parameter is zero, the objective falls back to the traditional gradient tree boosting. Regularization term can be expressed as

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|w\|^2. \tag{1}$$

Parameters $\gamma \in [0,1]$ and $\lambda \in [0,1]$ are so-called called L1- and L2-regularization terms (Chen and Guestrin, 2016, p. 2). The previous equation (Eq. 1) shows that the L1-regularization term controls the model's complexity by penalizing models with a higher number of trees and the L2-regularization term the score distribution among leaves.

4.3.4. Loss function

The loss function describes the difference between models' prediction $\hat{y}_i \in \{0,1\}$ and actual target $y_i \in \{0,1\}$ when $i \in \{1, \dots, n\}$. In the context of gradient boosting, this loss function must be differentiable and convex. In this research we are going to use logarithmic loss function

$$l(\hat{y}_i, y_i) = -[y_i \, ln \, p_i + (1 - y_i) \, ln(1 - p_i)],$$

where y_i is a binary indicator of whether label is the correct classification for instance i and $p_i = \frac{1}{1+e^{-\hat{y}_i}}$ describes the model's probability assigning label to instance i. We can notice that in the logarithmic loss is just a logarithmic transformation of likelihood function of Bernoulli distribution and hence, by minimizing logarithmic loss function over a set of parameters we maximize the likelihood of the given observations (Painsky et al., 2020, p. 1659).

4.3.5. Additive training

The tree ensemble model includes functions as parameters. These functions contain the tree's structure and the leaf weights, making it impossible to optimize using traditional optimization methods in the Euclidean space. Instead, the model is trained in an additive manner. We can write the prediction value $\hat{y}_i^{(t)}$ at step t as following

$$\hat{y}_i^{(0)} = 0$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i)$$

$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i)$$

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i).$$

We want to choose a tree that minimizes our objective function at each step by adding f_t that most improves our model. In doing so, our objective function becomes the following

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i{}^{(t)}) + \sum_{i=1}^{t} \Omega(f_i)$$

$$= \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i{}^{(t-1)} + f_t(x_i) \right] + \Omega(f_t)$$

(Chen and Guestrin, 2016, p.3).

Since the evaluation of the objective function is expensive, we can use to second order Taylor approximation to optimize the loss function and we get the following expression for the loss function

$$l(y, \hat{y}) = l(y, \hat{y}^{t-1} + f_t(x))$$

$$= l(\hat{y}^{t-1}) + l'(\hat{y}^{t-1})(\hat{y} - \hat{y}^{t-1}) + \frac{1}{2} l''(\hat{y}^{t-1})(\hat{y} - \hat{y}^{t-1})^2$$

$$= l(y, \hat{y}^{t-1}) + l'(\hat{y}^{t-1}) f_t(x) + \frac{1}{2} l''(\hat{y}^{t-1}) \big(f_t(x)\big)^2.$$

Second partial derivatives of the loss function provide more information about the direction of gradients and how to get to the minimum of our loss function. In practice, we take steps in the opposite direction of the gradient to find the global minimum of the objective function.

We can achieve a more general expression for the loss function by substituting $g = l'(\hat{y}^{t-1})$ and $h = l''(\hat{y}^{t-1})$ and summing over all samples to get an approximation of our objective function

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} \left[l\big(y_i, \hat{y}_i{}^{(t-1)}\big) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t).$$

We can also remove the constant term, $\mathrm{l}(\mathrm{y}, \hat{\mathrm{y}}^{\mathrm{t}-1})$, to obtain the following simplified objective at step t

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t). \tag{2}$$

The previous formula (Eq. 2) becomes our optimization goal for the new tree. A critical advantage of this definition is that the value of the objective function only depends on g and h so we can optimize every any function using g and h as inputs (Chen and Guestrin, 2016, p.3).

When the data is imbalanced, we can apply weights on the loss function, specifically on g and h to make the model biased towards the minority class. In this research, we are using the ratio between majority class data instances and minority class instances as a weight. We would then define the objective function as

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n}\left[w_{pos\ weight}\cdot g_i f_t(x_i) + \frac{1}{2} w_{pos\ weight}\cdot h_i f_t^2(x_i)\right] + \Omega(f_t), \tag{3}$$

where $w_{pos\ weight} = \frac{Number\ of\ majoriy\ class\ data\ instances}{Number\ of\ minority\ class\ data\ instances}$. For simplicity, we are not going to include these weights in the later calculation and since these weights are constant, we can consider them to be a part of g and h.

4.3.6. The Structure Score

To define the leaf scores, we must also define the optimal weights for each leaf. After re-formulating the tree model, we can write the objective function as following

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2$$

$$= \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right) w_j^2\right] + \gamma T,$$

where $I_j = \{i|q(x_i) = j\}$ is the instance set of leaf j. After the re-formulation of the loss function, we can calculate the optimal weight w_j^* of leaf j by

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}. \tag{4}$$

To further avoid overfitting, we can scale the weight, w_j^*, by a parameter $\varepsilon \in]0,1]$ that effectively controls the learning rate of the process. Parameter ε reduces the influence of each tree and leaves space for future trees to improve the model. Sometimes the learning rate parameter is not enough to avoid overfitting, in which cases we can set an upper limit to the weight of a node.

After determining the optimal weights for each leaf, we can calculate the corresponding optimal value of the loss function by

$$\bar{\mathcal{L}}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T.$$

$\bar{\mathcal{L}}^{(t)}(q)$ can be used as a scoring function to measure the quality of a tree structure q. This is like the impurity score (for example Gini-index) for evaluating decision trees, except it is derived for a wider range of objective functions (Chen and Guestrin, 2016, p.3).

4.3.7. Learning the tree structure

Now that we can measure how good a tree is, we would like to enumerate all possible trees and pick the best one. However, this is intractable in practice since constriction of a decision tree is an NP-complete problem (Kotsiantis et al., 2006, p.163). So instead, we will optimize one level of the tree at a time by a greedy algorithm that starts from a single leaf and iteratively adds branches to the tree.

In this research, we use categorical data, so the number of possible splits is considerably small, and we can use a greedy algorithm to find the optimal split points. However, in the case of continuous variables, the number of candidate split points can grow so large that the greedy algorithm is no longer possible to use. XGBoost solves this problem using

weighted quantile sketch and sparsity-aware split finding methods (Chen and Guestrin, 2016, p. 4).

We can evaluate the split by comparing the loss reduction achieved by each split that is commonly referred as gain. Let I_L and I_R be instance sets of left and right nodes after the split and $I = I_L \cup I_R$ then the loss reduction after the split is given by

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma. \tag{5}$$

revious formula is usually used in practice for evaluating the split candidates. It can be decomposed in four parts: $\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i+\lambda}$ is the score on the new left leaf, $\frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i+\lambda}$ is the score on the new right leaf, $\frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i+\lambda}$ is the score on the original leaf and γ is the regularization on the additional leaf (Chen and Guestrin, 2016, p.3). We can see that if the gain is smaller than γ, we should not add that branch. Most of the tree-based models use this pruning technique.

There are two ways to add branches to the tree: leaf-wise and depth-wise methods. The leaf-wise method will choose the leaf that maximizes the loss reduction. On the other hand, the depth-wise method will finish the leaf growth at the same level for all leaves. As a result, the leaf-wise method tends to achieve lower loss and converge much faster, but it will also be more likely to overfit.

4.3.8. Pruning

Pruning is a data compression technique that reduces the size of decision trees by removing non-critical and redundant sections to classify instances. Pruning reduces the complexity of the final classifier, improving predictive accuracy by reducing overfitting.

XGBoost uses three different pruning techniques to control the size of the decision tree. Firstly, by controlling the loss reduction needed (Eq.5) for a new split. Secondly, by stopping the splitting process once the node reaches a certain level of purity, and thirdly, by stopping the splitting process once the depth of the tree reaches a certain level.

In addition to previously the introduced γ parameter (Eqs. 4 and 1), XGBoost indirectly regulates the tree depth by controlling the minimum sum of weights needed in a child node by stopping splitting the nodes if the sum of weights is smaller than a certain threshold that we can control as a parameter. In practice, this means not splitting the node once it reaches a specific purity level. We can calculate the sum of instance weights in the node by summing the second partial derivatives over all points in the node.

The third method is straightforward, and it works by setting a parameter that stops the splitting process once the depth of the tree grows too large. The other two methods indirectly control the tree depth, but they do not guarantee a limit for the depth of the tree. More complex trees allow more precise modeling of the data but, on the other hand, increase the risk of overfitting.

4.3.9. Time complexity

XGBoost can be considered computationally inexpensive since its time complexity is only $O(td\|x\|_0 \cdot logn)$, when d is the maximum depth and $\|x\|_0$ is the number of non-missing entries in the data (Chen and Guestrin, 2016, p. 6). For comparison, training time for SVM is between $O(n^2 \cdot m)$ and $O(n^3 \cdot m)$ depending on the training set and the set of hyperparameters, where m is the number of features

4.4. Hyperparameter optimization

Hyperparameter optimization is the problem of optimizing an objective function over some graph-structured configuration space. More formally, hyperparameter tuning problem can be given as follows

$$X^* = arg_{x \in S}\, max[f(x)],$$

where X^* denotes the set of hyper-parameters that yield the highest value of the objective score (for example AUC), $x \subset \mathbb{R}^d$ denotes the candidate set where d is the number of parameters and $f(x)$ is the objective score to minimize (Stuke et al., 2021, p. 15). In general, hyperparameter tuning is essential to obtain the best prediction performance.

This research will focus on sequential model-based optimization (SMBO) and tree-structured parzen estimator (TPE) that are part of the Bayesian optimization framework.

4.4.1. Bayesian optimization

In an application where the true objective function is costly to evaluate, Bayesian methods can be used to approximate the objective function with a surrogate function that is cheaper to evaluate. The surrogate function is a probabilistic model approximating the true objective function based on given hyperparameters and their associated output values. Bayesian models select the next set of hyperparameters based on the acquisition function using the past evaluations of the surrogate function. Bayesian optimization methods can be differentiated at a high level by their surrogate - and acquisition functions (Brochu et al., 2009, p. 2).

In addition to these methods, non-Bayesian alternatives for performing hyperparameter optimization include grid search, random search (Bergstra and Bengio, 2012), and particle swarm optimization (Kennedy and Eberhart, 1995). The problem with these approaches is that they are either computationally expensive or require good specification of the parameter space. In contrast, Bayesian methods are generally time-efficient and do not require a good specification of the parameter space. Concerning this research both aspects are important since the data is large and the model has multiple parameters that are difficult to define.

4.4.2. Sequential model-based optimization (SMBO)

Sequential model-based optimization (SMBO) is a class of optimization algorithms that iterate between building a model of some unknown objective function and using the information from that model to query the next point in the domain of that function. This approach offers the prospects of interpolating performance between observed parameter settings and extrapolating to previously unseen regions of parameter space.

Sequential model-based global optimization (SMBO) is a formalism of Bayesian optimization. The basic idea of SMBO to construct a surrogate function of the loss function and then use the subsequently obtained information to continuously optimize the alternative probability model to make it close to the actual distribution. SMBO algorithms keep track of past evaluation results which they use to form a probabilistic model mapping hyperparameters to a probability of a score on the objective function $p(y|x)$ where x represents hyperparameters and y the associated quality score. Each time the algorithm proposes a new set of candidate hyperparameters, it evaluates them with the actual objective function and records the result in a pair of score and hyperparameters (Hutter et al., 2011, p. 508-509).

The surrogate function is a probabilistic model that approximates the true function based on given hyperparameter values and associated output values (Brochu et al., 2010, p. 3).

Various surrogate functions can be used in the SMBO context (Eggensperger et al., 2013), but the model must define a predictive distribution $p(y|x)$. This distribution captures the uncertainty in the surrogate reconstruction of the objective function. In this research, we use a Tree-structured parzen estimator (TPE) as our surrogate function.

Acquisition functions are mathematical techniques that guide how the parameter space should be explored during optimization. Acquisition functions balance the trade-off of exploiting a known high-performing result and exploring uncertain locations in the hyperparameter space (Eduardo et al., 2019, p. 51). The acquisition function used in this research is expected improvement. It defines the non-negative expected improvement over the best previously observed objective value at a given location. If we consider some model M that is defined as $f: X \rightarrow \mathbb{R}^N$, expected improvement can be formally expressed as following

$$EI_{y^*}(x) = \int_{-inf}^{+inf} max(y-y^*, 0)p_M(y|x)dy.$$

Here y^* is a threshold value of the objective function, x is the proposed set of hyperparameters, $y \in \{0,1\}$ is the actual value of the objective function using hyperparameters $x \subset \mathbb{R}^d$, and $p(y|x)$ is the surrogate probability model expressing the probability of y given x (Jones et al., 1998, p. 471-472). Basically, expected improvement is the expectation that $f(x)$ will exceed some threshold y^*.

4.4.3. Tree-structured Parzen Estimator (TPE)

Tree-structured parzen estimator is a surrogate model for sequential model-based optimization. Compared to other Bayesian models that assume the form of the predictive distribution $p(y|x)$, TPE models it by $p(x|y)$ and $p(y)$. TPE was selected as a surrogate function because it supports a wide variety of variables in parameter search space and is highly time-efficient, having a time complexity of $O(N)$.

TPE models $p(y|x)$ by replacing the distributions of the configuration prior with non-parametric densities. As stated before, instead of directly representing $p(y|x)$ it models it by

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$

Using different observations $\{x_1, \cdots, x_k\}$ in the non-parametric densities TPE can produce a variety of densities over the configuration space X. The TPE defines $p(x|y)$ using two such densities

$$p(x|y) = \begin{cases} l(x) \; if \; y < y^* \\ g(x) \; if \; y \geq y^* \end{cases}$$

, where $l(x)$ is the density formed by using observations $\{x^{(i)}\}$ which corresponding loss $f(x^{(i)})$ was less than y^* and $g(x)$ is the density using remaining observations. TPE depends on y^* that is larger than the best-observed $f(x)$ so that some points can be used to form $l(x)$. The TPE algorithm chooses y^* to be some quantile γ of the observed *y* values, so that $p(y < y^*) = \gamma$ but no specific model for $p(y)$ is necessary.

Based on the definition of $p(x|y)$ we want to draw values from *x* that belong to $l(x)$ since this distribution consists only of values of *x* that yielded lower scores than the threshold. Next, we will present an alternative formalization of the expected improvement function in the context of TPE. By using the definition of $p(y|x)$ we can express the expected improvement as follows

$$EI_{y^*}(x) = \int_{-inf}^{y^*} (y^* - y)p(y|x)dy$$
$$= \int_{-inf}^{y^*} (y^* - y)\frac{p(x|y)p(y)}{p(x)}dy.$$

Following the previous definition of γ, g(x) and $l(x)$ we can define expected improvement in two parts where the prior density of scores (normalizing constant) is separated from the function:

$$p(x) = \int_{\mathbb{R}}^{inf} p(x|y)p(y)dy = \gamma l(x) + (1-\gamma)g(x),$$

and

$$\int_{-inf}^{y^*} (y^* - y)p(x|y)p(y)dy$$

$$= l(x)\int_{-inf}^{y^*} (y^* - y)p(y)dy = \gamma y^* l(x) - l(x)\int_{-inf}^{y^*} p(y)dy.$$

Finally combining both parts we the get an alternative formalization for expected improvement as:

$$EI_{y^*}(x) = \frac{y^* l(x)(\gamma - 1)\int_{-inf}^{y^*} p(y)dy}{\gamma l(x) + (1-\gamma)g(x)} \propto \left(\gamma + \frac{g(x)}{l(x)}(1-\gamma)\right)^{-1}.$$

From the final expression, we can see that to maximize expected improvement, we need *x* with a high probability under $l(x)$ and low probability under $g(x)$ because this distribution is based only on values of x that yielded lower scores than the threshold. The tree structured form of $l(x)$ and $g(x)$ makes it easy to draw many candidates according to $l(x)$ and evaluate them according to the proportion of $g(x)$ and $l(x)$. On each iteration, the algorithm returns the candidate $x^*$with the biggest expected improvement (Bergstra et al., 2011, p. 2550).

4.4.4. XGBoost parameters

XGBoost has multiple hyperparameters that must be set manually to build the classification model. This part will cover all the parameters we optimize or change from their default values during our research.

Table 2: XGBoost parameters

Parameter name	Search space	Value type	Reference	Explanation
eta	$]0,1]$	Float	Eq. 4, p. 32	Step size shrinkage used in update to prevents overfitting
gamma	$]0,1]$	Float	Eq. 5, p. 33–34	Minimum loss reduction required to make a further partition on a leaf node of the tree
max_depth	$[1, 15]$	Integer	p. 34	Maximum depth of a tree
min_child_weight	$]0,1]$	Float	p. 34	Minimum sum of weights needed in a child node
lambda	$]0,1]$	Float	Eq. 1. p. 29	L2 regularization term on weights
alpha	$]0,1]$	Float	Eq. 1, p. 29	L1 regularization term on weights
max_delta_step	$]0,1]$	Float	Eq. 4, p. 32	Maximum weight of a tree node
grow_policy	Loss guide, depthwise	Categorical	p. 34	Controls the way new nodes are added to the tree
scale_pos_weight	12.3	Float	Eq. 3, p. 32	Control the balance of positive and negative weights

4.5. Imbalance learning

An imbalanced classification problem appears when the target class has an uneven distribution of observations. Imbalanced classification poses a challenge for predictive modeling as most of the classification algorithms are designed to assume an equal number of examples for each class, resulting in models with good predictive power over the majority class but with the expense of poor predictive power over the minority class. To avoid this problem, we need to construct classifiers that are biased toward the minority class without harming the accuracy of the majority class.

Many techniques have been developed to answer the problem of imbalance classification. These techniques can be categorized into four main groups, depending on how they deal with the problem.

Algorithm level approaches try to adapt existing classifier learning algorithms to be biased toward the minority class. To perform the adaptation a special knowledge of both the corresponding classifier and the application domain is required to comprehend why the classifier fails when the class distribution is uneven.

Data level approaches aim at rebalancing the class distribution by resampling the data. The modification of the learning algorithm is avoided since the effect caused by imbalance is decreased with a preprocessing step. However, some commonly used rebalancing algorithms, such as SMOTE (Chawla et al., 2002), are computationally expensive, highlighted in hyperparameter optimization.

Cost-sensitive learning framework falls between data and algorithm level approaches. The classifier is biased toward the minority class by assuming higher misclassification costs for the minority class and seeking to minimize the total cost errors of both classes. Cost-sensitive learning is computationally inexpensive, making it attractive in

hyperparameter optimization and dealing with large data quantities. Also, cost-sensitive learning does not require further assumptions about the data or our prediction algorithm.

Ensemble-based methods usually consist of a combination between an ensemble learning algorithm and one of the techniques above. When using cost-sensitive ensembles, the base classifier is modified to accept costs in the learning process (Fernández et al., 2018, p. 22). In this research, we are using the cost-sensitive ensemble-based method since our model is an ensemble model, but the reasoning is based on the cost-sensitive approach.

4.5.1. Metrics for imbalance learning

This part will cover some of the most common metrics for binary imbalanced classification problems since evaluating the imbalance classification results require specific consideration for the metrics being used. For example, the performance of machine learning algorithms is typically evaluated using predictive accuracy. However, accuracy is not an appropriate metric when the data is imbalanced because the model can achieve high accuracy by only classifying all data instances as belonging to the majority class.

4.5.1.1. Confusion matrix

The confusion matrix is a useful tool when evaluating a discrete classifier. In the binary case, we denote the number of correctly classified positive and negative instances by true positives (TP) and true negatives (TN). False negatives (FN) stand for the number of instances predicted to be non-serious but are serious accidents, while the false positives (FP) are the number of instances falsely classified as serious accidents. The confusion matrix reveals how well the classifier predicts each accident outcome (Townsend, 1971, p. 41).

4.5.1.2. ROC-curves and AUC

The Receiver Operating Characteristic (ROC) curve is a standard technique for summarizing classifier performance over a range of trade-offs between true positive and false positive error rates (TPR and FPR). Probabilistic classifiers need a threshold to make a final decision for each class. Therefore, a different discrete classifier is obtained for each possible threshold, with a different TPR and FPR. ROC curve is obtained by considering all these possible thresholds, putting their corresponding classifiers in the ROC space, and drawing a line through them.

TPR is also known as recall and can defined as

$$TPR = \frac{TP}{TP + FN}.$$

FPR can be defined together with precision that is defined as $\frac{TP}{TP+FP}$, so then

$$FPR = 1 - \frac{TP}{TP + FP}.$$

ROC curve essentially tells what level of FPR is required to achieve a certain TPR. ROC values are independent of the class distribution, making them helpful in evaluating unbalanced problems. The area under the curve (AUC) is a traditional performance metric for a ROC curve, and it measures the area between the x-axis and the ROC curve. Classifiers that require a lower FPR for each threshold of TPR acquire a higher AUC number (Fan, 2006, p. 19-20). AUC is used in this research to compare different hyperparameter configurations' performance.

5. Results

In this part, we will explain the model selection process, hyperparameter optimization results, and the results of our final model.

5.1. Model selection

Since hyperparameter optimization is computationally expensive, we completed our initial analysis with only 100 000 data instances (70:30 training test split) and decided to use the best-performing model with the complite data.

The data consist of mostly categorical variables that are converted to a numeric format before applying the classification algorithm. Therefore, categorical variables are converted to dummy variables, and other categorical variables are label-encoded. All the variables in the data set are categorical variables except age, an ordinal variable. After conversion, we had 174 variables.

The fundamental consideration for model selection was the capacity to handle categorical data and the capability for rule-based pattern extraction. We also only considered algorithms with a solid theoretical grounding that can represent any level of complexity and are flexible in that no structures are imposed a priori that might conceal the actual underlying structure of the data. Requirement for the model's capability for rule-based pattern extraction ruled out the use of the artificial neural network (ANN) approaches even though ANN has enjoyed some success in previous research.

Based on our model requirements, we chose three models for the initial analysis: Suppor vector machine (SVM) (Cortes et al., 1995), complement naive Bayes (CNB) (Rennie et al., 2003), and XGBoost. Firstly, we decided to use SVM since it has been successfully applied in previous research of accident outcome prediction. Secondly, we chose

complement naive Bayes (CNB) because it works with unbalanced data and serves as a good baseline model for our approach to the problem of unbalanced data. Lastly, we considered XGBoost, since its success in numerous machine learning competitions (Qingyun et al., 2016), and it has not yet been applied to accident outcome prediction. Finally, candidate models were compared based on ROC curves.

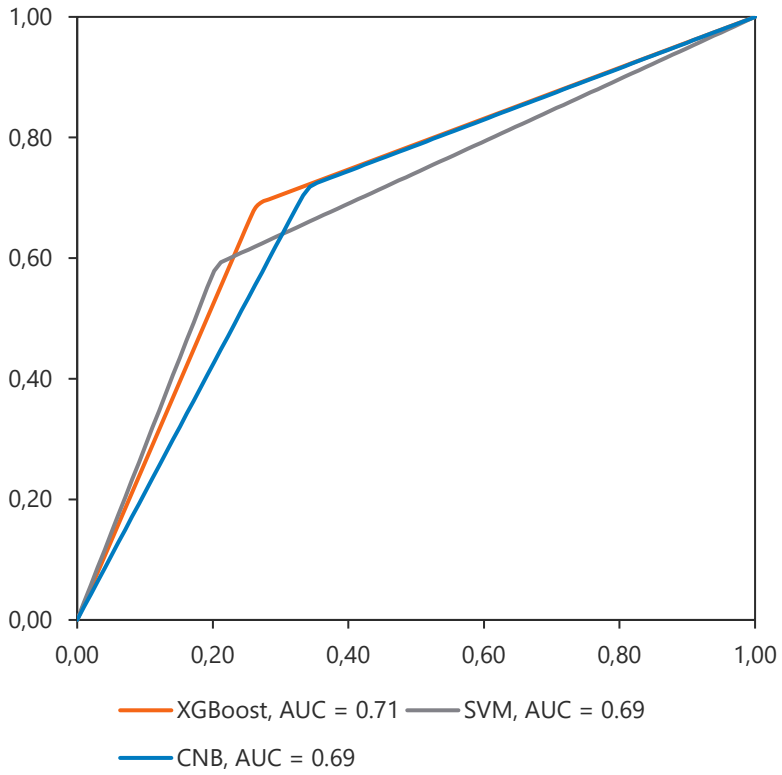

Figure 11. ROC curves of the candidate models. On the x-axis, FPR and on the y-axis, TPR.

All models have similar performance, but XGBoost slightly overperforms other models based on AUC. XGBoost practically dominates CNB and SVM has a slightly better performance in predicting non-serious accidents (TN). In addition, XGBoost is the most balanced model in predicting majority and minority class instances (Fig. 11). Based on these results and the fact that the training time for XGBoost is several times shorter than with SVM, we choose XGBoost as our final model.

5.2. Training the model

Data was divided into two parts of the training set and the test set by proportions of 7:3, respectively. The training set is used to train the model, and the test set is used to validate the model's results.

The model was trained with the training set using hyperparameter optimization with 200 iterations. After each iteration, the performance was measured using the average AUC between stratified 10-fold cross-validation splits. After finding the best model, the results were again validated using stratified 10-fold cross-validation. Stratified n-fold cross-validation (Scikit-learn, 2022) begins by shuffling the test set randomly and splitting it into n subsets containing approximately the same data instances. Each of these subsets was held as a test set while the remaining were used as a training set. This process was repeated with all the subsets, and models' evaluation metrics were calculated as averages between all subsets. Cross-validation helps to mitigate the problem of randomness in the validation results.

5.2.1. Hyperparameter optimization

This part will cover hyperparameter optimization results and the parameter values that resulted in the best performance in terms of AUC.

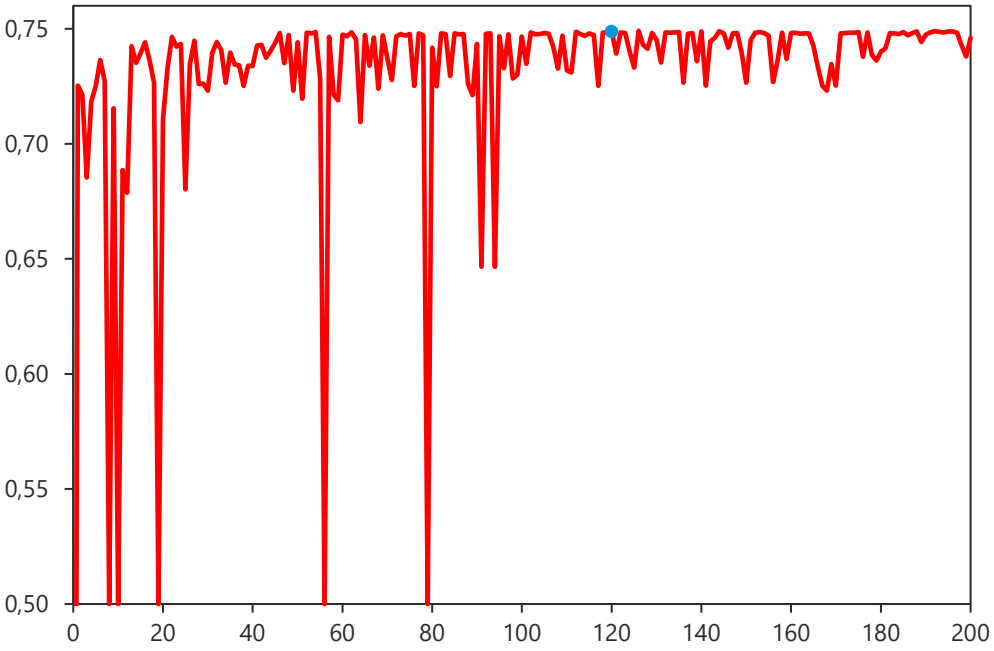

Figure 12. Hyperparameter optimization results. On the y-axis AUC and on the x-axis iterations. The Red line represents the results of different iterations, and the blue dot represents the best result among all iterations.

As we can notice, most hyperparameter combinations generate results generally close to each other, implying that the model is quite robust for different parametrization. However, some iterations deviate greatly form the best results caused by an extreme value of some of the pruning parameters (Fig. 12).

Furthermore, the results seem to stabilize with further trials indicating results reaching the optimal parametrization. The best result was AUC of 0.7491 and the ceiling of the models seems to be AUC of around 0.75.

Table 3: Parameter values

Parameter	Value of the model	Default value
eta	0.91390	0.3
gamma	0.00864	0
max_depth	11	6
min_child_weight	0.12736	1
lambda	0.01239	1
alpha	0	0
max_delta_step	0.25630	0
grow_policy	Loss guide	Depthwise
scale_pos_weight	12.3	1

We notice that all the parameters deviate from the default values except alpha. Especially eta, min child weight and max depth parameter differ significantly from the default values (Table 3). Generally, we can say that parameters controlling overfitting are most affected, probably caused by the depth of the decision tree that deviates significantly from the default value.

5.3. Evaluation of the results

This part will display the results of the predictive model in terms of evaluation metrics, feature importance, decision rules and wrongly classified non-serious accidents.

5.3.1. Evaluation metrics

In this research we use AUC, accuracy, recall, precision, and confusion matrix for evaluating the results.

Table 4: Stratified 10-fold validation results

Metrics	AUC	Accuracy	Recall	Precision
Split 1	0.7460	0.7664	0.7221	0.2030
Split 2	0.7447	0.7686	0.7165	0.2038
Split 3	0.7452	0.7699	0.7162	0.2047
Split 4	0.7447	0.7668	0.7186	0.2027
Split 5	0.7506	0.7700	0.7278	0.2067
Split 6	0.7497	0.7661	0.7305	0.2042
Split 7	0.7529	0.7720	0.7305	0.2087
Split 8	0.7567	0.7721	0.7385	0.2102
Split 9	0.7573	0.7722	0.7397	0.2104
Split 10	0.7454	0.7684	0.7183	0.2039
SD	0.0047	0.0022	0.0084	0.0028
Mean	0.7493	0.7693	0.7259	0.2058

We notice that the AUC of our training model is close to the validation results. Furthermore, the validation results have a low variation implying that the model is not overfitting (Table 4). Even though the mean accuracy of 77% cannot be considered excellent, these results are similar to previous research on accident outcome prediction using structural data (Anurag et al., 2020).

Table 5: Confusion matrix

Label	Non-serious (Predicted label)	Serious (Predicted label)
Non-serious (True label)	0.7728	0.2272
Serious (True label)	0.2741	0.7259

The model predicts non-serious accidents more accurately than serious accidents, but not much considering the unbalanced nature of the data. For example, we notice that the model predicts decently non-serious accidents with an accuracy of around 77% but serious accidents with less than 73% accuracy (Table 5).

5.3.2. Feature importance

Feature importance was calculated as the sum of loss reduction across all splits the feature is used. Here we refer to loss reduction as gain.

It seems features that have a significant disproportion between serious -and non-serious accidents are more likely to appear among more important features (Fig. 13). On the other hand, features that rarely appear in the data have lesser importance regardless of the distribution of serious -and non-serious accidents. It is also clear that features describing the outcome of the accident are the most prevalent since 21 out of 27 most important features belong to either injury type or body variable, and features of those variables cover 84% of all gains (Table 6).

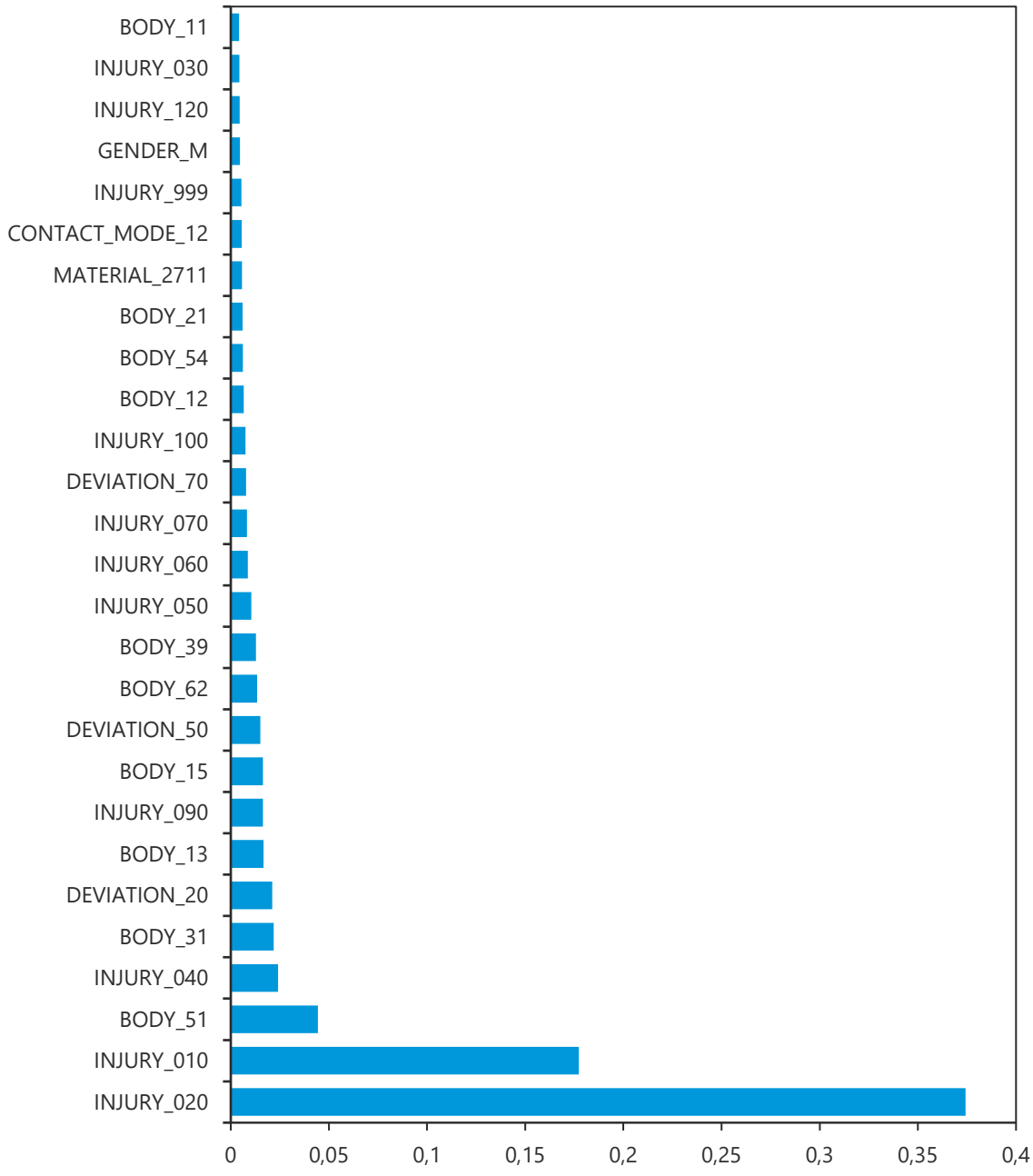

Figure 13. Normalized feature importance's of features that cover 85% of the total gains. On the x-axis proportions and on the y-axis features.

Table 6: Proportional sum of gains of all features

Variable	Proportion of gains
Injury type	0.64
Body part	0.2
Deviation	0.05
Material agent	0.05
Working process	0.03
Contact mode of injury	0.02
Physical activity	0.01
Gender	0
Age	0

Injury type is by far the most important variable (Table 6), and all its features are included among the most important features except shock (INJURY_100). Wounds and superficial injuries (INJURY_010) and bone fractures (INJURY_020) are by far the most important features among all the features (Fig. 13). In most cases, wounds and superficial injuries implicate a non-serious accident and bone fractures a serious accident. Even though traumatic amputations (INJURY_040) are over 20 times more common among serious than non-serious accidents, they appear so rarely (Fig. 8) that their importance is significantly lesser than wounds, superficial injuries, and bone fractures (Fig. 13).

The body part is the most represented variable among the most important features and the second most important variable (Table 6). Eye (BODY_13), shoulder and shoulder joints (BODY_51), spine, and vertebra (BODY_31) are the most important features belonging injured body part variable (Fig. 13). Eye injuries implicate a non-serious

accident, and shoulder and shoulder joints and back, spine, and vertebra in the back injuries' serious accidents.

The deviation-variable's most important feature is overflow, overturn, leak, flow, vaporization, or emission (DEVIATION_20). Other features of the deviation variable are slipping, stumbling, and falling (DEVIATION_50) and body movement under or with physical stress (DEVIATION_70). Body movement under or with physical stress implies a non-serious accident and slipping, stumbling, falling and overflow, overturn, leak, flow, vaporization, or emission a serious accident (Fig. 13).

Material agent associated with the mode of injury-variable has only one feature, fixed machine tools for sawing (MATERIAL_2711), representing it even though around 1/3 of the overall features belong to that variable (Fig. 13). Fixed machine tools for sawing as a material agent implicate a serious accident. Because of the large number of features, the material agent associated with the mode of injury variable has the third-highest total gains even though individual features are insignificant (Table 6).

Direct contact with electricity receipt of electrical charge in the body (CONTACT_MODE_12) is the only feature representing the contact mode of the injury (Fig. 13). Direct contact with electricity receipt of electrical charge in the body implicates a non-serious accident. Contact mode of injury is one of the least important variables, with its features totaling only 2% of all gains (Table 6). Physical activity is also one of the least important variables and is not even represented among important features. Accordingly, features of physical activity amount to only 1% of the total gains (Table 6). The low importance of physical activity variable corresponds to our earlier speculation about its low importance (Fig. 5).

Among variables describing the injured person, age (AGE) and male gender (GENDER_M) were included (Fig. 13), but the working process was not. Age and gender variables amount to only around 1% of the total gains (Table 6); however, age is a nominal variable,

and gender has only one feature. On the other hand, the working process variable amounts to more than 3% of the total gains, but mostly because it has multiple features.

5.3.3. Decision rules

One of the strengths of decision tree-based models is the capability to visualize the tree model and the decision rules behind the prediction. Leaf nodes represent the predicted value reached by following the decision rules defined by the model. Leaf node values can be converted to a probability score using the logistic function. The leaves with a probability score of at least 0.5 are classified as serious accidents and others as non-serious accidents.

Features with higher feature importance generally appear higher in the decision tree, but the feature importance does not entirely decide the relevance among the decision rules. For example, even though variables working process and physical activity are do not appear important in terms of loss reduction, they seem pretty frequently in the decision tree although on the lower level of the tree. The same applies to the variable describing the material agent associated with the injury; however, this variable has multiple times more features than other variables.

Since the decision tree is quite complex, we are only focusing on the higher levels of the tree, and simple decision rules in our analysis, mainly concerning features belonging to injury type or injured body part variable.

The wounds and superficial injuries feature is on the highest level of the decision tree. In most cases, the wounds and superficial injuries indicate a non-serious accident except when the material agent associated with the injury is fixed machine tools for sawing. When the material agent associated with the injury is something else, only very complex decision rules lead to the precision of a serious accident.

In the absence of wounds and superficial injuries, most other injury types indicate a serious accident, including dislocations, sprains, or strains, traumatic amputations, broken bones, shock, effects of sound, vibration, or pressure, drowning and asphyxiation and multiple injuries. Accidents with other injuries require further conditions to be predicted as serious accidents. For example, concussion or internal injuries are considered serious when the material agent associated with the injury is mobile ladders and burns, scalds, or frostbites when multiple sites of lower extremities of the body are affected.

Body part-variable also appears often on the higher levels of the decision tree. When injury type is not wounds, and superficial injuries spine and vertebra injuries, implicate a non-serious accident in most cases except when appearing together with a bone fracture. The same applies when other back parts are injured. Also, shoulder or shoulder joint injuries imply non-serious accidents except when age is over 35 years, and the contact mode is a collision with an object in motion.

On the lower level of the decision tree, some of the decision rules do not appear sensible. For example, when the injury type is poisonings or infection, injured body part is not leg or knee, spine or vertebra or other back part, and the material agent associated with the injury is a land vehicle the accident is classified as a serious accident. These results may imply that a simpler decision tree could describe the phenomena of workplace accidents better, even with the loss of prediction accuracy.

5.3.4. Wrongly classified non-serious accidents

This part will analyze the distribution of absence from work of wrongly classified non-serious accidents compared to the validation set distribution (Fig. 10). We are interested in knowing how much wrongly classified non-serious accidents deviate from the definition of serious accidents in terms of absence from work, since it could have implication on the model's financial benefits. We are only concerned about the wrongly

classified non-serious accident since an incorrect prediction of a serious accident does not affect the normal insurance claim process.

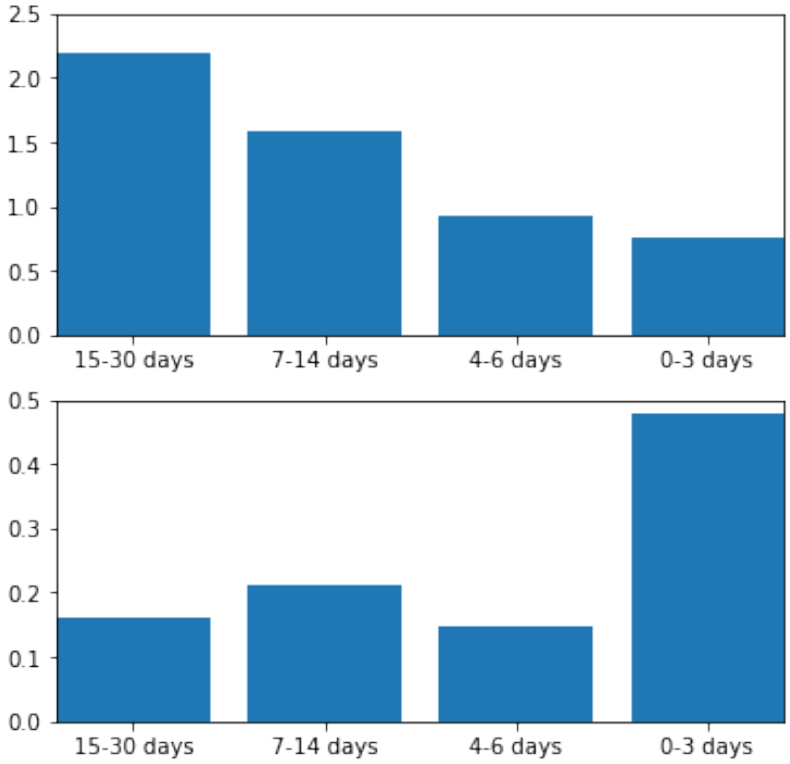

Figure 14. Distribution of wrongly classified non-serious accidents in terms of absence from work. The upper plot describes data instances predicted to be serious, but that are non-serious accidents, proportional to the distribution of the validation set where the y-axis represents proportions and x-axis absence from work. The lower plot describes the distribution of data instances predicted to be serious, but that are non-serious accidents, where the y-axis represents frequency and x-axis absence from work.

We would hope that the distribution of wrongly classified data instances would be skewed towards the threshold with non-serious and serious accidents more than they appear in the validation set. For example, the wrong prediction of a serious accident can be useful in preventing prolonged absence from work even if the actual accident would be only 28 days, thus lowering the misclassification cost of the model.

With wrongly classified serious accidents, we see that the 15–30-day category is more than twice as common among the wrongly classified data instances (Fig. 14) compared to the whole data (Fig. 10) and for 7-14 days more than 1.5 times. Even though these results are encouraging, the 0–3-day category is still the most frequent, and the 15–30-day category represents only 16% of the wrongly classified non-serious accidents. Therefore, even if all the wrongly classified non-serious accidents belonging to the 15–30-day category would lead to the same financial benefits as correctly classified serious accidents, the expected financial improvement per accident would only be $0.16 \cdot FPR \cdot administrative\ cost = 0.036 \cdot administrative\ cost$. Thus, we will not analyze the distribution more precisely (day level) or use this information when calculating the financial benefits of the model.

6. Conclusion

This part will present the research summary, practical implications of the results, limitations of the research, and suggestions for further research.

6.1. Research summary

In this research, we used the TPE method to optimize the XGBoost model to predict workplace accident outcomes based on the accident notices delivered to the insurance companies in Finland. Accident outcomes were divided into serious and non-serious accidents based on the absence from work resulting from the accident. Cases where the absence from work was more than 30 days, were considered serious.

Wounds and superficial injuries and bone fractures were found to be the most important features predicting the workplace accident outcome with wounds and superficial injuries, implying in most cases a non-serious accident and bone fractures serious accident. Overall injury variable was found to be most important ESAW-variable together with body

The model could predict serious accidents with an accuracy of 73% and non-serious accidents with an accuracy of 77%.

6.2. Practical implications

This part will display the models' potential financial benefits for the insurance company. We will not approximate financial benefits concerning all workplace accidents since there are too many unknown variables. Instead, we define break-even points when the model is profitable for individual cases.

The financial benefits of the model are affected by how many days it can advance the workers returning to work, the amount of daily allowance paid to the injured person, and the cost of the administrative process. Earlier return to work will also decrease other costs associated with workplace accidents but are challenging to define, so we don't consider them.

When determining the cost savings of the model, we are only concerned about the prediction accuracy of serious accidents since a correct prediction of a non-serious accident does not affect the normal insurance claim process. The expected savings of the model can be calculated as follows

$$E[savings] = TPR \cdot (A \cdot D - CA) - FPR \cdot CA,$$

where $A = daily\ allowance$, $D = earlier\ return\ to\ work\ in\ days$ and $CA = cost\ of\ the\ administrative\ process$.

Model shortens the administrative process by helping focus on cases where the risk of prolonged absence from work is elevated, but it is impossible to say precisely how many days this could shorten the absence from work of the injured person. For this reason, we examine different cost scenarios depending on different daily allowances and recovery times.

The amount of daily allowance depends on the injured person's income, so it makes sense to examine the benefits of the model for different income groups. Therefore, we defined three different income groups: 1) the lowest-earning 25% with a daily allowance of 70 euros, 2) the median with the daily allowance of 99 euros, and 3) the highest-earning 25% with a daily allowance of 138 euros (Statistic Finland, 2021).

Cost of administrative process incurs when the insurance company contacts the injured person and medical service providers to accelerate the diagnosis -and treatment process.

We do not know the exact cost of the administrative process, but we can define the thresholds when the model is profitable.

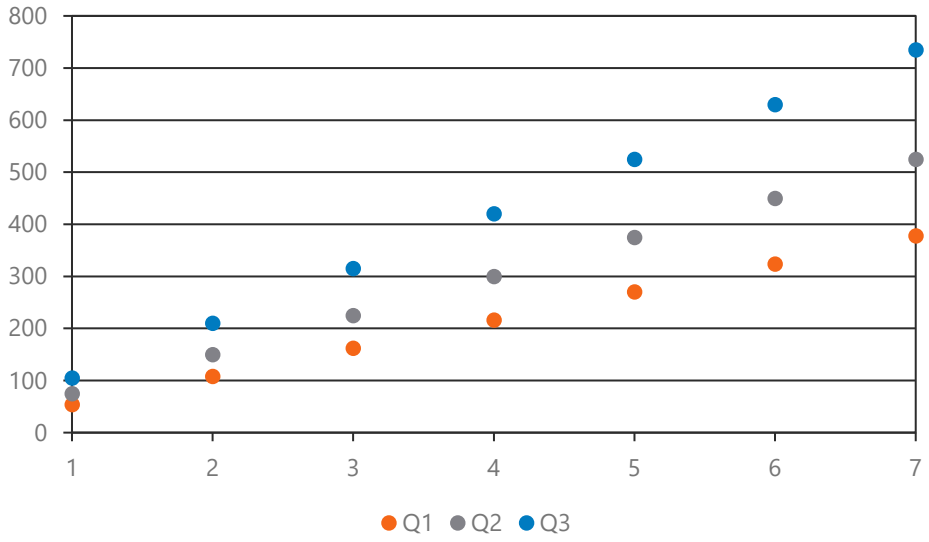

Figure 15. Break-even points for the administrative cost. Y-axis represents administrative cost in euros and x-axis earlier return to work in days.

When the return to work is advanced by only one day, administrative cost needs to be under 100 euros for the model to be profitable, and the difference between income groups is relatively small. However, with the increase of days of earlier return, the differences between different income groups become more evident. For example, when the cost of the administrative process is 200 euros model is profitable for the highest income group if the injured person returns to work two days earlier, for the median income group three days, and for the lowest income group four days (Fig. 15).

6.3. Limitations of the research

This research has some limitations concerning the chosen methodology, especially hyperparameter optimization, imbalance learning, and approach to serious and non-serious accidents.

The Bayesian approach to hyperparameter optimization cannot guarantee that the parameter values obtained after the optimization process would lead to the best possible performance of the model. In other words, it is possible that with further iterations, more optimal parameters values could be discovered.

It is also difficult to evaluate the effectiveness of our approach to the problem of imbalanced data compared to other methods. However, even though the data level approach is infeasible with the computer recourses available for this study, it could yield better results.

The definition of a serious accident in this research is arbitrary and based on the time frame for the administrative process set by the Finnish Accidents at Work and Occupational Diseases Act. Therefore, the definition used in this research does not necessarily reflect the optimal way to divide accidents into serious and non-serious accidents to advance the injured persons' recovery.

6.4. Suggestions for further research

Suggestions for future research consist of ways to binary classification model developed in this research, other approaches to the problem of binary classification, and other approaches to the problem of predicting workplace accident outcomes.

The model could be more useful if capable of predicting absence from work on a more detailed level. However, since the prediction accuracy is quite poor even with the binary classification, it is unlikely that increasing the predicted classes or predicting the length of the absence from work in days would lead to satisfactory results. Based on the results, significant improvements cannot be made to the prediction accuracy by relying on the ESAW variables even with other predictive algorithms. However, earlier research (Anurag et al., 2020) has shown that applying natural language processing (NLP) techniques on

the written accident descriptions could yield better results than structural data and thus enable these approaches.

However, some possible improvements are possible with the structural data. These improvements concern combining features and, on the other hand, including additional variables.

Several features of the material agent associated with the mode of injury-variable that represent structures above ground level are significantly more common among serious accidents than non-serious accidents. However, only two of those features appear in the decision rules even though the accident mechanism is intuitively quite similar. Thus, some of these features could be combined to reduce the features leading to a simpler decision tree with more sensible decision rules.

It could be argued that the occupation affects the seriousness of the injury. For example, a broken finger for a violist has more serious consequences than a broken finger for a truck driver. The working process variable describes the occupation on some level but maybe not well enough. Variables describing the occupation more precisely were discarded since it limited the available data considerably, but possible improvement could be achieved by including variables that describe the occupation more accurately.

References:

Aggarwal C. (2015). Data Classification: Algorithms and Applications. 35. Philadelphia, PA: Chapman and Hall/CRC. ISBN 9781466586741.

Anurag Y, Fatemeh K, and Ali J. (2020). Predictive Modeling for Occupational Safety Outcomes and Days Away from Work Analysis in Mining Operations. International Journal of Environmental Research and Public Health, 17(19): 7054. doi:10.3390/ijerph17197054.

Bergestra J, Bardenet R, Bengio Y, and Kégl B. (2011). Algorithms for hyperparameter optimization. In Proceedings of the 24th International Conference on Neural Information Processing Systems: 2546-2554. Curran Associates Inc., Red Hook, NY, USA.

Brochu E, Cora, M, Freitas N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. ArXiv, abs/1012.259.

Chawla N.V, Bowyer K.W, Hall L.O, Kegelmeyer W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16: 321–357. doi: 10.1613/jair.953.

Chen T and Guestrin C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 785–794. doi: 10.1145/2939672.2939785.

Cortes C, Vapnik V. Support-vector networks. (1995). Machine Learning, 20(3): 273–297. doi: 10.1007/BF00994018. S2CID 206787478.

Eggensperger K, Feurer M, Hutter F, Bergstra J, Snoek J, Hoos H, and Leyton-Brown K. (2013). Towards an empirical foundation for assessing Bayesian optimization of hyperparameters. In NIPS workshop on Bayesian Optimization in Theory and Practice.

European union commission regulation No 349/2011. Administrated in Brussel 11.4.2011. Available: https://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:097:0003:0008:EN:PDF.

European Commission. (2012). European Statistics on Accidents at Work (ESAW) – Summary methodology. Luxembourg: Publications Office of the European Union / Eurostat Methodologies & Working papers. doi: 10.2785/2509.

Fernández A, Garcia S, Galar M, Prati C.R, Krawczyk B. (2018). Learning from Imbalanced Data Sets. doi:10.1007/978-3-319-98074-4.

Finnish Accidents at Work and Occupational Diseases Act 459/2015. Administrated in Helsinki 24.4.2015. Available: https://www.finlex.fi/fi/laki/ajantasa/2015/20150459#O9L35.

Finnish Medical Society Duodecim. (2014). Rotator cutoff: Current Care Guidelines. Available: https://www.kaypahoito.fi/hoi50099. Cited: 12.12.2021.

Finnish Workers' Compensation Center. (2021). Statistics on accidents at work. Available: https://www.tvk.fi/tilastot-ja-julkaisusarjat/tilastot/tyotapaturmatilastot/. Cited: 31.02.2022.

Finnish Workers' Compensation Center. (2021). Statistics on paid claims. Available:

https://www.tvk.fi/en/statistics-and-publications/statistics/statistics-on-paid-claims/. Cited: 10.10.2021.

Finnish Workers' Compensation Center. (2019). Workers' compensation in numbers. Available: https://www.tvk.fi/document/172687/5A3029A1720AC8F37C459993C576A6662EC7F4886220D644CEBCC5BBF0EEAB5F. Cited: 10.10.2021.

Garrido-Merchán EC, Hernández-Lobato D. (2019). Predictive entropy search for multi-objective Bayesian optimization with constraints. Neurocomputing, 361:50-68. doi: 10.1016/j.neucom.2019.06.025.

Hantes M, Karidakis G, Vlychou M. (2011). A comparison of early versus delayed repair of traumatic rotator cuff tears. Knee Surgery, Sports Traumatology, Arthroscopy, 19: 1766–1770.

Hohmann E, Glatt V, Tetsworth K. (2017). Early and delayed reconstruction in multi-ligament knee injuries: A systematic review and meta-analysis. The Knee Journal, 24(5): 909–916.

Hutter F, Hoos H.H, Leyton-Brown K. (2011). Sequential Model-Based Optimization for General Algorithm Configuration. In: Coello C.A.C. Learning and Intelligent Optimization. LION 2011. Lecture Notes in Computer Science, 6683. https://doi.org/10.1007/978-3-642-25566-3_40.

Jerome F, Upadhye S, Worster A. (2006). Understanding Receiver Operating Characteristic (ROC) Curves. Canadian journal of emergency medicine 8(1): 19–20. doi:10.1017/S1481803500013336.

Jones D.R, Schonlau M, Welch W.J. (1998). Efficient global optimization of expensive black-box functions. Journal of Global optimization, 13(4): 455–492. doi:10.1023/A:1008306431147.

Kennedy J and Eberhart R. (1995). Particle swarm optimization. In Proceedings of the 1995 IEEE International Conference on Neural Networks, 4: 1942–1948. doi: 10.1109/ICNN.1995.488968.

Kotsiantis S.B, Zaharakis I.D, Pintelas P.E. (2006). Machine Learning: Review of Classification and Combining Techniques. The Artificial intelligence review 26(3): 159–190. doi: 10.1007/s10462-007-9052-3.

Lin SH, Wang TC, Lai CF, Tsai RY, Yang CP, Wong CS. (2017). Association of anterior cruciate ligament injury with knee osteoarthritis and total knee replacement: A retrospective cohort study from the Taiwan National Health Insurance Database. PLoS One 12(5). doi: 10.1371/journal.pone.0178292.

Matias J, Rivas T, Martin J, Taboada J. (2008). A machine learning methodology for the analysis of workplace accidents. International Journal of Computer Mathematics, 85(3–4): 559-578. doi: 10.1080/00207160701297346.

Molinero E, Pitarque S, Fondevila-McDonald Y, Martin-Bustamante M. (2015). How reliable and valid is the coding of the variables of the European Statistics on Accidents at Work (ESAW)? A need to improve preventive public policies. Safety Science, 79. doi: 10.1016/j.ssci.2015.05.005.

Nenonen N. (2011). Occupational accidents in the Finnish local government sector: Utilization of national statistics. International Journal of Injury Control and Safety Promotion, 18(4): 321-329

Occupational Health Care Act 1326/2010. Administrated in Helsinki 30.12.2010. Available: https://www.finlex.fi/fi/laki/ajantasa/2010/20101326.

Painsky A and Wornell G.W. (2020). Bregman Divergence Bounds and Universality Properties of the Logarithmic Loss. IEEE transactions on information theory 66(3): 1658–1673. doi: 10.1109/TIT.2019.2958705.

Petersen S and Murphy T. (2011). The timing of rotator cuff repair for the restoration of function. Journal of Shoulder and Elbow Surgery, 20(1): 62–68.

Pietilä, J, Tolonen J, Helander E. (2018). Työtapaturmaisten Olka- Ja Polvivammojen Hoitotoimenpiteet Ja -Kustannukset Sekä Hoidon Ja Sairauslomien Kesto Vakuutusyhtiön Rekisteriaineistoon Perustuen. Finnish Journal of eHealth and eWelfare 10(1): 113–132.

Pinheiro C.R, Patetta M. (2021). Introduction to Statistical and Machine Learning Methods for Data Science. 2021. SAS Institute. ISBN: 9781953329622.

Qingyun W. (2016). XGBoost. GitHub repository, https://github.com/dmlc/xgboost/blob/master/demo/README.md.

Rennie J. D, Shih, L, Teevan J, Karger D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. Proceedings of the Twentieth International Conference on International Conference on Machine Learning (ICML'03). AAAI Press: 616–623.

Rissanen M, Kaseva E. (2014). Cost of lost labor input. Ministry of Social Affairs and Health. Available: https://stm.fi/documents/1271139/1332445/Menetetyn+ty%C3%B6panoksen+kustannus+2+%282%29+%282%29.pdf/63af9909-0232-474d-bf2e-aa4c50936c33. Cited: 10.10.2021.

Ristiniemi J. 2018. Polven vammat. Lääkärin käsikirja. Available: www.terveysportti.fi. Cited: 12.12.2021.

Rojas M.M, Cabello A.T, Ferreira M.C.P, Romero J.C.R. (2020). An Approach to Explore Historical Construction Accident Data Using Data Mining Techniques. Advances in Engineering Networks. ICIEOM 2018. Lecture Notes in Management and Industrial Engineering. doi: 10.1007/978-3-030-44530-0_15

Salo K. (2015). Työtapaturma ja ammattitauti. Vantaa: Hansaprint Oy.

Sinnott P. (2009). Administrative Delays and Chronic Disability in Patients with Acute Occupational Low Back Injury. Journal of Occupational and Environmental Medicine, 51 (6): 690-699.

Sobhan S, Sammangi V, Rahul J. Maiti, Mitra. (2019). Application of optimized machine learning techniques for prediction of occupational accidents. Computers & Operations Research, 106: 210-224.

Statistic Finland. (2021). Labor force survey. Available: https://www.stat.fi/til/tyti/2021/06/tyti_2021_06_2021-07-27_tie_001_en.html

Statistic Finland. (2021). Wage statistics. Available: https://www.stat.fi/tup/kokeelliset-tilastot/tulorekisterin_palkat_ja_palkkiot/2021-toukokuu/index.html. Cited: 2021.12.12.

Stover B, Wickizer T, Zimmerman F, Fulton-Kehoe D & Franklin G. (2007). Prognostic Factors of Long-Term Disability in a Workers' Compensation System. Journal of Occupational and Environmental Medicine, 49(1): 31-40.

Stuke A, Rinke P, Todorovic M. (2021). Efficient Hyperparameter Tuning for Kernel Ridge Regression with Bayesian Optimization. 2021. Machine Learning: Science and Technology, 2(3). doi: 10.1088/2632-2153/abee59.

Townsend J.T. (1971). Theoretical analysis of an alphabetic confusion matrix. Perception & Psychophysics, 9: 40–50. doi: 10.3758/BF03213026.

Vastamäki M. (2002). Olkanivelen kiertäjäkalvosimen repeämä– älä viivyttele hoidossa!.. Suomen lääkärilehti, 17(57): 1915–1920. Available: https://bulevardinklinikka.fi/wp-content/uploads/2013/04/Olkanivelen-kiertajakalvosimen-repeama.pdf.

Zhang C and Ma Y. (2012). Ensemble Machine Learning: Methods and Applications. New York: Springer. doi: 10.1007/978-1-4419-9326-7.

Zhou ZH. (2012). Ensemble Methods: Foundations and Algorithms. London: Chapman and Hall/CRC. doi: 10.1201/b12207.

Appendix A: Feature descriptions of the variables

Table A1: Feature description of the Material agent associated with the mode of injury variable

Code	Material agent
1100	Building components including doors, walls, and intentional obstacles (windows, etc.)
1201	Fixed stairs, roofs, terraces, doors, windows, and quays
1210	Fixed ladders
1219	Other fixed structures above ground level
1221	Mobile ladders
1222	Temporary supports
1223	Mobile scaffolding
1229	Other structures above ground level
1230	Temporary structures including scaffolding and hanging racks
1240	Drilling platforms scaffolding or barges
1229	Other structures above ground level
1310	Excavations, trenches, wells, pits, escarpments, or garage pits
1320	Underground areas or tunnels
1330	Underwater environments
1399	Others structures below ground level
2100	Pipe networks for the supply and distribution of materials
2200	Motors or power generators (thermal, electric, radiation)
2300	Powered hand tools
2400	Not powered hand tools
2500	Hand tool (source of power unknown)
2601	Portable or mobile machines for extracting materials or working the ground
2602	Portable or mobile machines for working the ground in farming
2603	Portable or mobile machines for extracting materials or working the ground
2604	Floor cleaning machines
2699	Other portable or mobile machines or equipment

2701	Fixed machines for extracting materials or working the ground
2702	Fixed machines for crushing, pulverising, filtering, separating, mixing, blending or materials
2703	Fixed machines for chemical processes
2704	Fixed machines for hot processes (ovens, driers, kilns)
2705	Machines for cold processes
2706	Other machines for preparing raw materials
2707	Fixed forming machines for pressing or crushing
2708	Fixed forming machines for calendering, rolling or cylinder presses
2709	Fixed forming machines by injection, extrusion, blowing, spinning, moulding, melting, or casting
2710	Fixed machine tools for planning, milling, surface treatment, grinding, polishing, turning, or drilling
2711	Fixed machine tools for sawing
2712	Fixed machine tools for cutting, splitting, or clipping
2713	Fixed machines for cleaning, washing, drying, painting, or printing
2714	Fixed machines for galvanising or electrolytic surface treatment
2715	Assembling machines for welding, gluing, nailing, screwing, riveting, spinning, wiring, sewing, or stapling
2716	Packing machines
2717	Miscellaneous monitoring or testing machines
2718	Other machines used in farming
2799	Other known fixed or mobile machines and equipment
2801	Lifting equipment including slings, hooks, and ropes
2802	Elevators, lifts, hoists, bucket elevators and jacks
2803	Fixed belts, escalators, cableways, and conveyors
2811	Mobile machines for conveying but not lifting
2812	Pushcarts
2813	Mobile handling devices, handling trucks, barrows, or pallet trucks
2814	Containers with wheels
2815	Pallet jack
2816	Forklift

2819	Other conveying, transport, or storage systems
2830	Mobile silos, tanks, vats, and containers
2840	Handling trucks, barrows, and pallet trucks
2850	Shelves
2860	Miscellaneous small and medium-sized packaging
2899	Other known conveying, transport, and storage systems
3100	Land vehicles
3200	Other vehicles
4100	Materials, objects, products, machine, vehicle components, debris, or dust
4200	Chemical, explosive, radioactive or biological substances
4300	Safety devices and equipment
4400	Office equipment, personal equipment, sports equipment, weapons, or domestic appliances
5100	Human-beings, animals, or plants
5200	Bulk waste
5300	Physical phenomena and natural elements
9999	Other material agents
XX	Missing values

Table A2: Feature description of the Injury type variable

Code	Injury type
010	Wounds and superficial injuries
020	Bone fractures
030	Dislocations, sprains, and strains
040	Traumatic amputations
050	Concussion and internal injuries
060	Burns, scalds, and frostbites
070	Poisonings and infections
080	Drowning and asphyxiation
090	Effects of sound, vibration, and pressure
100	Effects of temperature extremes, light and radiation
110	Shock
120	Multiple injuries
999	Other specified injuries
XX	Missing values

Table A3: Feature description of the Physical activity variable

Code	Physical activity
10	Operating machine
20	Working with hand-held tools
30	Driving/being on board a means of transport or handling equipment
40	Handling of objects
50	Carrying by hand
60	Movement
70	Presence
99	Other specific physical activities
XX	Missing values

Table A4: Feature description the of Working process variable

Code	Working process
11	Production, manufacturing, or processing
12	Storing
19	Other production, manufacturing, processing, or storing
21	Excavation
22	New construction of buildings
23	New construction of civil engineering, infrastructures, roads, bridges, dams, or ports
24	Remodelling, repairing, extending, or building maintenance
25	Demolition
29	Other group of excavation, construction, repair, or demolition
31	Agricultural type of work working the land
32	Agricultural type work with vegetables or horticultural
33	Agricultural type work with live animals
34	Forestry type work
35	Fish farming or fishing
39	Other agricultural type work, forestry, horticulture, fish farming or work with live animals
41	Service, care, or assistance
42	Intellectual work including teaching, training, data processing, office work, organising or managing
43	Commercial activity including buying, selling and associated services
49	Other service provided to enterprise and/or to the public
51	Setting up, preparation, installation, mounting, disassembling, or dismantling
52	Maintenance, repair, tuning or adjustment
53	Cleaning working areas
54	Waste management, disposal, waste treatment of all kinds
55	Monitoring or inspection of manufacturing procedures
59	Other work

61	Movement, including aboard means of transport
62	Sport or artistic activity
63	Sailoring
69	Other movement, sport, or artistic activity
99	Other working processes
XX	Missing values

Table A5: Feature description the of Deviation variable

Code	Deviation
10	Operating a machine
20	Deviation by overflow, overturn, leak, flow, vaporization, or emission
30	Breakage, bursting, splitting, slipping, fall or collapse of material agent
40	Loss of control of machine, means of transport, handling equipment, hand-held tool, object, or animal
50	Slipping, stumbling, and falling
60	Body movement without any physical stress
70	Body movement under or with physical stress
80	Shock, fright, violence, aggression, threat, or presence
99	Other deviations
XX	Missing value

Table A6: Feature description of the Injured body part variable

Code	Injured body part
11	Head, brain and cranial nerves and vessels
12	Facial area
13	Eye(s)
14	Ear(s)
15	Teeth

18	Head, multiple sites affected
19	Other parts of head
21	Neck, inclusive spine, and vertebra in the neck
29	Other parts of neck
31	Back, including spine and vertebra in the back
39	Back, other parts not mentioned above
41	Rib cage, ribs including joints and shoulder blades
42	Chest area including organs
43	Pelvic and abdominal area including organs
48	Torso, multiple sites affected
49	Other parts of torso
51	Shoulder and shoulder joints
52	Arm, including elbow
53	Hand
54	Finger(s)
55	Wrist
58	Upper extremities, multiple sites affected
59	Other upper extremities
61	Hip and hip joint
62	Leg, including knee
63	Ankle
64	Foot
65	Toe(s)
68	Lower extremities, multiple sites affected
69	Other lower extremities
71	Whole body (Systemic effects)
78	Multiple sites of the body affected
XX	Missing value

Table A7: Feature description of the Contact mode of injury variable

Code	Contact mode
11	Indirect contact with a welding arc, spark, or lightning (passive)
12	Direct contact with electricity, receipt of electrical charge in the body
13	Contact with naked flame or a hot or burning object or environment
14	Contact with a cold or frozen object or environment
15	Contact with hazardous substances via inhalation
16	Contact with hazardous substances through skin or eyes
19	Contact with electrical voltage, temperature, or hazardous substances
20	Drowned, buried, or enveloped
30	Horizontal or vertical impact with or against a stationary object
40	Struck or collision with object in motion
50	Contact with sharp, pointed, rough or coarse material
60	Trapped or crushed
70	Physical or mental stress
80	Bite, kick, etc. (animal or human)
99	Other contact modes of injury
XX	Missing

Table A8: Feature description of the Age variable

Code	Age
1	under 15
2	15-19
3	20-25
4	25-29
5	30-34
6	35-39
7	40-44

8	45-49
9	50-54
10	55-59
11	60-64
12	over 64

Tapaturmavakuutuskeskus TVK, Itämerenkatu 11-13, 00180 Helsinki