UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Conversion Master in IT

Mario Käära

Application of Machine Learning Techniques to Ensure Safer Work Environments in Estonia

Master's Thesis (15 ECTS)

Supervisor: Roshni Chakraborty, PhD

Tartu 2023

Application of Machine Learning Techniques to Ensure Safer Work Environments in Estonia

Abstract:

Occupational accidents are a major global concern which results in significant human and economic losses. In Estonia, over $4,000$ work-related accidents are recorded annually, and 428 fatalities were reported between 2001 and 2021. For example, work-related accidents led to a loss of $141,000$ workdays and approximately €5.3 million in 2021. Several studies across different countries have recently proposed automated data analytic tools and machine learning based models to understand occupational hazards and predict the likelihood and severity of accidents. These applications can identify high-risk workers and ensure robust safety management systems across various industries, such as construction and manufacturing. However, these proposed models are not directly applicable to Estonia, and no specific tools can handle the local settings. Through this Thesis, we aim to develop automated models based on machine learning techniques to predict the severity of occupational accidents in Estonia. We also identify critical factors for different industries contributing to these accidents. Our dataset consists of $82,641$ work-related accidents, featuring 37 variables, and spans the period from 2002 to 2022. The Thesis demonstrates that the best-performing models, including Support Vector Machine and Logistic Regression, can predict accident severity and identify crucial factors for targeted prevention strategies. The primary outcomes include critical insights into the important factors and the development of tailored machine learning models for occupations in specific economic sectors. Therefore, we propose accurate and efficient automated tools that can handle the inherent data challenges and ensure the significance of targeted modelling in accident prevention. The Thesis illustrates the potential of understanding the data patterns, developing specific data analytic tools and machine learning algorithms to improve decision-making in workplace safety and developing cost-effective prevention strategies.

Keywords:

machine learning, occupational accidents, extreme gradient-boosting, light gradient-boosting, logistic regression, support vector machine, random forest, random oversampling.

CERCS:

P176 - Artificial intelligence

Masinõppe meetodite rakendamine turvalisema töökeskkonna tagamiseks Eestis

Lühikokkuvõte:

Tööõnnetused on oluline ülemaailmne probleem, mis põhjustab märkimisväärset inim- ja majanduskahju. Aastatel 2001 kuni 2021 registreeriti Eestis keskmiselt ligikaudu $4,000$ tööõnnetust aastas ja hukkus kokku 428 inimest. Näiteks aastal 2021 olid töötajad tööõnnetuste tõttu $141,000$ päeva ajutiselt töövõimetud, mis tõi endaga kaasa rahalise hüvitise suuruses €5.3 miljonit. Hiljutised uuringud on rakendanud masinõppe algoritme erinevates tööstusharudes, sealhulgas ehituses ja tootmises, pakkudes välja andmeanalüüsi tehnikaid selliste probleemide lahendamiseks nagu kõrge riskiga töötajate tuvastamine ja ohutusjuhtimissüsteemide loomine. Küll aga nendes uuringutes väljapakutud mudelid ei ole Eestis otseselt rakendatavad ning puuduvad konkreetsed vahendid, mis kohalikes oludes hästi toimiksid. Selle tööga soovime välja töötada masinõppe meetoditel põhinevaid automatiseeritud mudeleid, mis ennustaksid tööõnnetuste tõsidust Eestis. Samuti tuvastame eri tööstusharude jaoks olulisi tegureid, mis raskeid õnnetusi põhjustavad. Meie andmestikus on $82,641$ tööõnnetust, millest igaühel on 37 muutujat ning see hõlmab ajavahemikku aastast 2002 kuni aastani 2022. See töö näitab, et kõige paremini toimivad mudelid, sealhulgas tugivektormasin ja logistiline regressioon, suudavad ennustada raskeid tööõnnetusi ja tuvastada sihipärase ennetusstrateegia jaoks olulisi tegureid. Töö põhitulemused rõhutavad, et keskmisi ennustamistäpsusi on võimalik saavutada ka ilma mudelite või tasakaalustamistehnikate hüperparameetrite häälestamiseta, kasutades ettevõtete tegevusalade ja ametite klasside kombinatsioonidele kohandatud masinõppemudeleid, millel on käsitsi valitud sõltumatud muutujad. See leid toetub olemasolevatele teadmistele, rõhutades sihipärase modelleerimise olulisust raskete tööõnnetuste ennetamisel, pakkudes täpsemat ja tõhusamat lähenemisviisi tööõnnetuste ja nende raskusastmete vähendamiseks erinevatel ettevõtete tegevusaladel. See töö toob esile masinõppe algoritmide potentsiaali tööohutuse alaste otsuste tegemise parandamisel ja kulutõhusate tööohutuse ennetusstrateegiate väljatöötamisel.

Võtmesõnad:

masinõpe, tööõnnetused, äärmuslik gradiendi hoogustamine, kerge gradiendi hoogustamine, logistiline regressioon, tugivektormasin, otsustusmets, juhuslik ülevalimine.

CERCS:

P176 - Tehisintellekt

Contents

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Roshni, for her invaluable guidance, support, and teaching throughout this process. Her dedication and expertise have been instrumental in shaping my research and helping me achieve my research goals.

I would also like to thank the Estonian Labour Inspectorate for their unconditional support and open communication, which has been crucial to the success of this research.

To my wife, Keiu, thank you for your unwavering support and encouragement throughout my master's studies. I promise to make it up to our lovely boys, Markus and Mattias, who have made the most significant sacrifices for my educational aspirations at the University of Tartu.

I would like to express my deep appreciation to the University of Tartu, especially the Institute of Computer Science, for allowing me to embark on this incredibly transformative journey.

Finally, I would like to thank my fellow students from TÜÜSAM for their camaraderie and support during our time together in Tartu. Through the coronavirus crisis, we have spent countless hours together on Zoom or Teams, and I am grateful for the memories we have made.

1 Introduction

Occupational accidents pose a significant threat to the global workforce, with more than 337 million work-related accidents occurring each year, resulting in 2.3 million deaths, according to the International Labour Organisation (ILO) report[1]. This results in the loss of 4% of global GDP and more than 15% of the national GDP in some countries if economic losses would take into account involuntary early retirement [27]. In Estonia, around $4,000$ reports of work-related accidents[2] are recorded yearly, equivalent to six incidents per $1,000$ workers. In 2021, this resulted in a loss of $140,000$ workdays and €5.3 million financially. Fatal work-related accidents have taken the lives of 428 workers in Estonia between 2001 and 2021.

To address these issues, leaders worldwide acknowledge that improving workplace safety and reducing occupational accidents and diseases can bring significant benefits[3]. The most effective approach to make decisions about health and safety interventions is through understanding of real data of the occupational accidents.

Employers are required to maintain a record of occupational accidents that result in an employee being unable to work for more than three days as part of EU Directive $89/391$/EEC. The European Statistics on Accidents at Work launch a project in 1990 to harmonise data on accidents at work for all accidents resulting in more than three days' absence from work. In 2001, Eurostat and European Commission detail the ESAW methodology[4] (ESAW) used to collect and analyse data on workplace accidents. The Estonian Occupational Health and Safety Act § 24 also requires employers to report occupational accidents resulting in temporary incapacity for work or death. Aligning reported data with the ESAW provides valuable empirical data that supports compliance with legal regulations and facilitates informed decision-making to improve health and safety in the workplace in Estonia.

Recently, several approaches based on machine learning (ML) techniques, such as decision trees (DT), random forests (RF), and Artificial Neural Networks (ANN), are proposed which work on real datasets of occupational accidents to predict the likelihood and severity of an accident. By analyzing historical accident data, these algorithms can identify patterns and relationships to accurately predict future accidents' probability and severity [5, 8, 25, 24, 30, 18, 20]. These techniques are instrumental in occupational safety, as they can identify high-risk areas and facilitate targeted safety interventions in the workplace.

[1] https://www.ilo.org/global/publications/world-of-work-magazine/articles/WCMS_099050/lang--en/index.htm

[2] https://www.ti.ee/media/391/download

[3] https://www.ilo.org/wcmsp5/groups/public/---ed_protect/---protrav/---safework/documents/publication/wcms_214163.pdf

[4] https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-ra-12-102

Through this research, we propose automated ML techniques which can improve accident response and prevention efforts in organizations in Estonia. We intend to identify critical factors contributing to occupational accidents and further predict the severity of occupational accidents. Although there are few existing research works in other countries, this research is the pioneering effort to investigate occupational accidents specifically in the context of Estonia automatically through machine learning approaches. Through this research, we use Estonia's national occupational accident dataset, which has never been studied before in ML settings. Estonia's national occupational accident dataset provides detailed information on work-related accidents and their outcomes. However, the dataset has several challenges, like high variance in the information, colossal information overload, and extensive missing values. Therefore, we propose comprehensive data analytic tools for specific pre-processing techniques concerning different features. Additionally, we explore several ML algorithms, such as RF, Support Vector Machine (SVM), Logistic Regression (LR), and Gradient Boosting frameworks such as Extreme Gradient Boosting Machine (XGBoost) and Light Gradient Boosting Machine (LightGBM) and their effectiveness in prediction. Furthermore, we propose specific models for some occupations in different economic sectors. We have studied severe accidents in more detail for *crafts and related trades workers* in the *retail and repair of the motor vehicle and motorcycles sector*, *professionals* in the *health care and social welfare sector* and *elementary occupations* in the *construction sector*. Based on our performance measurements, the proposed model can achieve an average F1 score of up to 0.73. Additionally, we determine *important features*, which helps in understanding what to focus on to ensure specific information related to an organization. For instance, in the *retail trade and repair of motor vehicles and motorcycles sector*, it is crucial to pay attention to workers' age and ensure equipment compliance for *craft and related trade workers*. In the *health care and welfare sector*, *professionals* should be mindful of human and animal contact as well as public places. On the other hand, in the *construction sector*, it is important to consider contact modes of injuries such as being trapped, crushed, or coming in contact with sharp objects for *elementary occupations*. By addressing these aspects, organisations can effectively mitigate workplace accidents in their respective sectors. The main contributions are summarized as follows:

- We conduct extensive data cleaning and pre-processing on the raw dataset to address the inherent challenges, such as significant noise, missing values, and substantial feature variance, while preserving relevant information.
- Our research investigates the impact of different resampling techniques on model performance and selects the optimal technique for each scenario to enhance the models' predictive power.
- We evaluate the performance of various ML models and feature sets to determine the best-performing models for each sector and occupation combination.

- We compare the performance of generic models for predicting occupational accident severity. Additionally, we explore sector- and occupation-specific models to address better the unique challenges faced by accidents with occupations within specific economic sectors.
- We identify and analyze the top-performing combinations of economic sectors and occupation classes, which provide a detailed examination of the *important* features that contribute to the severity of occupational accidents within the specific groups.

2 Literature Review

As ML algorithms can effectively analyze large and complex data to identify the relevant patterns and provide accurate predictions, they have been recently used in various economic sectors, such as the healthcare [13] and transportation [28], to predict future outcomes. Similarly, recent research has proposed different data analysis techniques and ML approaches in occupational safety. We segregate these works into groups covering ML models, balancing techniques and feature selection used in related works.

Kakhki *et al.* [8] focuses on predicting occupational injuries in the non-farm agricultural industry by different ML algorithms, such as DT, RF, and gradient boosting to develop a predictive model for the binary classification of incident severity. The study found that SVM with the Radial Basis Function (RBF) kernel outperformed the other models with high F1 score, recall, and accuracy values. Binary tree and Naïve Bayes (NB) models were also used to determine the most influential factors in predicting injury severity. Sarkar *et al.* [25] demonstrates the potential use of ML algorithms and empirical data analysis to ensure occupational safety. They use RF to impute missing values and the Chi-squared (χ^2) test [19] for feature selection. Given a custom feature set, their proposed study indicates that SVM outperforms the ANN for occupational accident prediction. However, they require extensive manual efforts for data pre-processing. Additionally, as the dataset was limited, they could not test the model in different economic sectors to validate its effectiveness. However, none of these approaches handles the high imbalance in data.

Several recently proposed approaches have addressed the data imbalance issues and proposed specific ML techniques and features applicable. For example, Choi *et al.* [5] argues that class imbalance in datasets can affect the performance of ML models, and to address this, they propose preprocessing techniques such as Random Over-sampling (ROS) and Random Under-sampling (RUS). They suggest that ROS is preferred over RUS if the class distribution is highly skewed since RUS may deteriorate the negative class distribution. The study performed ROS by duplicating fatality objects to address the class imbalance, eliminating the possibility of researcher manipulation. Eventually, they observed that the RF algorithm preprocessed with RUS performs best. Zhu *et al.* [30] expanded on previous studies by using eight different prediction algorithms to address their dataset's class imbalance problem. The number of small accidents is nearly four times that of large accidents. They applied the Synthetic Minority Over-sampling Technique (SMOTE) to balance the dataset and argue that experimental results demonstrate that SMOTE effectively enhances the performance of most algorithms. Koc *et al.* [18] further explored the issue of imbalanced datasets and the impact of resampling methods. The study found that the Random Forest-Random Under-sampling (RF-RUS) model was the most effective prediction tool for different accident datasets. They identified significant features, suggested additional features and under-sampling methods to improve the model. The study also highlighted the limitations of the SMOTE

and recommended evaluating other resampling techniques for datasets with extreme values.

Oyedele *et al.* [20] suggests that the efficiency and performance of a data-driven model are heavily reliant on selecting relevant features as inputs. Given the multitude of variables in the dataset, the authors employ Recursive Feature Elimination (RFE) to rank features based on their importance and use only the most significant ones as inputs to the ML models. They proposed deep learning models to address the limitations of various conventional ML techniques. They observed that the deep learning model was more reliable than traditional ML models in predicting lost-time injury. Choi *et al.* [5] determined the significance of features with the help of the LR model based on features' contribution to predicting the likelihood of the occurrence of specific events. They applied the LR model to categorical data. The correlation between items was demonstrated using a linear combination of independent variables as a probability model, and the factors were ranked based on their statistical significance. The factors that were found to be significant were sex, employer scale, length of service, month, and day of the week, while age and construction type were found to be insignificant.

While these studies have contributed significantly to the field, they have also revealed limitations and areas for further exploration. As observed from the previous studies, there is significant variance in the proposed models. For example, no generalized prediction model is applicable irrespective of occupation and location. Therefore, applying an ML model and understanding the specific features depends on the industry and location. We investigate and explore the applicability of features and models concerning data within industry sectors. This will provide an in-depth analysis of the factors contributing to accidents in different occupations. We tackle the data imbalance issue highlighted by Choi *et al.* [5] and Koc *et al.* [18] by extensive evaluation and studying of resampling techniques. Additionally, we explore the factors related to both observable factors as proposed by Kakhki *et al.* [8] and integrate new features not explored in any of the existing works. Furthermore, we can ensure that the proposed models do not require manual intervention. Therefore, based on the limitations of the existing research works, we can ensure that our research will contribute to a more nuanced understanding of the factors of occupational accidents irrespective of the industry for Estonia. We will discuss next our problem statement.

3 Motivation and Problem Statement

In this Section, we initially discuss our research goals, followed by the dataset discussion and the challenges. We propose automated approaches to increase the effectiveness of accident prevention in workplaces. For this, we thoroughly examine Estonia's national occupational accident records to understand the implicit and explicit factors that can cause accidents in the workplace. Our dataset comprises $82,641$ records which range from 2002 to 2022. Therefore, we propose exploring data analysis techniques to handle this huge amount of information to understand the important aspects. Additionally, based on this processed information from the data, we propose automated ML algorithms tuned to the specific application for effectiveness. We highlight our research goals next:

- Determine the specific ML algorithm for an application
- Identify the important factors that lead to severe workplace accidents
- Provide insights to develop targeted intervention strategies for improving workplace safety

3.1 Data Collection and Sampling

Although the ILO report[5] highlights that recording every incident can improve the usefulness of statistics, this method often needs adjustments due to the resources and time it takes to report the data. The report, citing Germany's experience, recommends establishing a reporting threshold for accidents that cause over three days of absence. This approach balances ensuring data comprehensiveness and maintaining practical resource allocation. As a result, both ILO guidelines[6] and the ESAW [7] determine that data should be collected for fatal occupational accidents and those that cause more than three days of absence.

The Estonian Labor Inspectorate publishes yearly summaries of work-related accidents on their website[8], using simple statistics. But when we asked for more detailed information, they provided us with a dataset covering occupational accidents in Estonia from 2002 to 2022. This dataset has $82,641$ observations and 37 features, with only 6 of these features being numeric. During the entire collection period of this dataset, there

[5] https://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/meetingdocument/wcms_088373.pdf

[6] https://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/meetingdocument/wcms_088373.pdf

[7] https://ec.europa.eu/eurostat/documents/3859598/5926181/KS-RA-12-102-EN.PDF.pdf/56cd35ba-1e8a-4af3-9f9a-b3c47611ff1c?t=1414782641000

[8] https://www.ti.ee/asutus-uudised-ja-kontaktid/statistika

has been a requirement for occupational accidents in Estonia to be reported to the Labour Inspectorate if they cause at least one day of absence.

However, there are some challenges in using this data. For instance, the reporting guidelines changed in 2014 when the Estonian Labor Inspectorate started using the ESAW. This caused some inconsistencies in the data. The dataset also has missing values, making automating the analysis process challenging. For instance, there is a lot of missing data for some features, such as *posted workers* where over $57,000$ values are missing, and *traffic accidents* where over $53,000$ values are missing. This is because these features were not reported before 2014, and the overview of missing values and feature cardinality can be seen in Table 1. This makes automating the prediction model on the complete set of features challenging. We discuss the main challenges next in detail.

3.2 Challenges

In this Section, we address various challenges related to data handling. In Subsection 3.2.1, we discuss the issues of unprocessed and disorganized data and missing values. In Subsection 3.2.2, we delve into the topic of feature variances, and in Subsection 3.2.3, we explore the problem of imbalanced datasets.

3.2.1 Information Overload

As previously discussed, there is a huge amount of unprocessed unorganized data with noise and missing values in the dataset, which requires severe data pre-processing. Additionally, this requires selecting the specific approach that applies to a particular feature. Due to this high information overload, the main challenges include high cardinality in data and the presence of *unspecified* and *other* types of values. For example, the data has a high cardinality, such as $1,088$ distinct *economic activities*[9] and 556 *occupation types*[10]. Similarly, the employment status feature has 32 unique values, leading to a low count of rows for many individual categories in the feature. High cardinality is mainly caused by lengthy and detailed codes or codes combined with category names, which can result in slight variations in characters or spaces, producing distinct and separate categories. For the *unspecified* and *other* types of values, it is highly required to understand what it could be replaced with so that the information loss is minimum without increasing noise in that feature. We follow the ESAW methodology to determine these value types for every feature. Depending on the specific feature and coding scheme employed, *unspecified* values may be represented as *0*, *00*, or *000*. In contrast, *other* type values may be denoted by *9*, *99*, *999* or *900*. Notably, missing values are mainly indicated by

[9] https://emtak.rik.ee/EMTAK/pages/klassifikaatorOtsing.jspx

[10] https://klassifikaatorid.stat.ee/item/stat.ee/b8fdb2b9-8269-41ca-b29e-5454df555147/24

Table 1. Feature cardinality and NaN values in the original dataset of $82,641$ rows

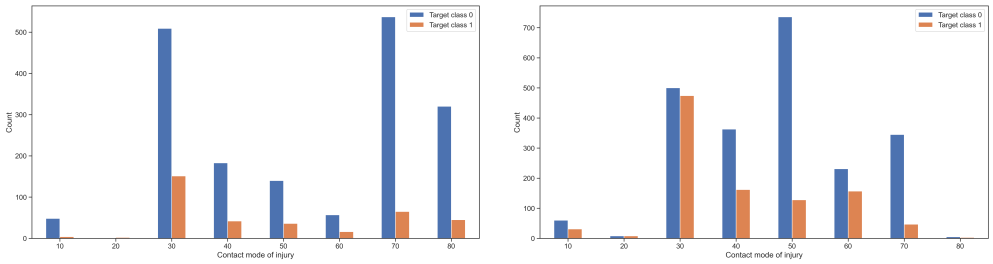
Feature	**Cardinality**	**NaN count**	**Datatype**
enterprise_ID	$14,973$	640	object
employees_in_enterprise	$2,026$	28	float64
employees_in_structural_unit	522	$73,773$	object
economic_activity	$1,088$	96	object
employee_ID	$68,958$	229	object
sex	4	86	object
age	74	605	float64
employment_status	31	$1,871$	object
employment_years	94	984	float64
is_posted_worker	2	$57,151$	object
date	$7,618$	0	object
time	$1,315$	803	object
full_hours_from_startofwork	41	86	float64
severity	3	75	object
location	44	$2,150$	object
causes	$1,023$	$5,008$	object
is_risk_assessment_done	2	$23,068$	object
are_risks_considered	2	$23,448$	object
under_investigation	2	$25,247$	object
causes_verified	490	$80,424$	object
age_group	7	$3,406$	object
enterprise_size	8	$6,587$	object
nationality	7	$5,525$	object
occupation_code	$5,56$	$8,623$	object
type_of_injury	65	$8,644$	object
injured_bodypart	52	$8,643$	object
lost_days	223	$18,957$	float64
workstation	9	$8,649$	object
working_environment	81	$8,649$	object
working_process	59	$8,647$	object
specific_physical_activity	48	$8,656$	object
material_agent_of_physical_act.	$1,259$	$8,756$	object
deviation	78	$8,654$	object
material_agent_of_deviation	$1,220$	$8,753$	object
contact_mode_of_injury	69	$8,657$	object
material_agent_of_contact_mode	$1,185$	$8,752$	object
is_traffic_accident	2	$53,851$	object

different string-type symbols like – or an empty string instead of the more ML-friendly NaN-type. Certain features, such as *full hours from the start of work* and *employment years*, contains numerous values of zero, which may not be considered true zeros and require appropriate handling.

3.2.2 Variance in Features

Understanding the variance in independent features concerning the target feature is crucial to building accurate and reliable ML models. The high feature variance can create noise and confusion in the model, making identifying patterns and relationships between the features and the target feature difficult. Our observations from the dataset indicate a high variance across features, such as the values of the feature's *contact mode of injury*, *deviation*, and *specific physical activity* varying greatly between different economic sectors and occupation classes.

For example, the *contact mode of injury* varies greatly between the *services and sales workers* in the *public administration and national defence sector* and *craft and related trades workers* in the *construction sector*, wherein for former, the most frequent *contact mode of injury* related to severe accidents the *horizontal or vertical impact with or against a stationary object* has the ratio of severe to non-severe accidents of 0.30. In contrast, it is 0.95 for the latter. Therefore, this indicates a three-fold higher importance for accident severity. The frequency of the *contact mode of injury* for severe and non-severe accidents for *services and sales workers* in the *public administration and national defence sector* is presented in Figure 1a and for the *craft and related trades workers* in the *construction sector* in Figure 1b.



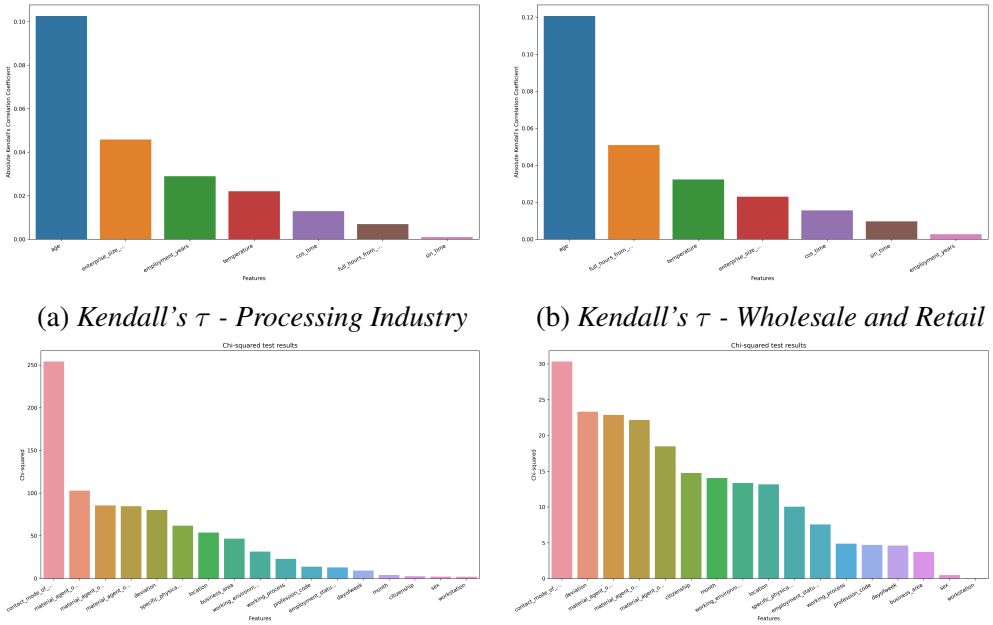
(a) *Public Administration Sector* (b) *Construction Sector*

Figure 1. Distribution of *contact mode of injury* by target classes for *services and sales workers* in *public administration and national defence sector* and for *crafts workers* in *construction sector* is shown

Therefore, one severe accident is reported for every non-severe accident among *craft workers* in *construction*. In comparison, only one severe accident is reported for every three non-severe accidents among *services and sales workers* in *public administration*.

Building general ML classification models across the entire dataset that distinguish between these occupations based on conflicting predictors could be challenging if different occupations in various sectors have distinct contributing factors to severe accidents. Contradictory data may create noise and lead to overfitting. We used statistical measures to calculate feature variances, such as the χ^2 test and Kendall's rank correlation test [17].

The relationship between features and the target feature differs, as shown by the χ^2 test and Kendall's τ coefficient values, which vary across economic sectors and occupation classes. For example, in Figures 2a and 2b, we see that the *statistically significant* features for *machine operators* in the *processing industry* are *age*, *enterprise size*, and *years of employment*. In contrast, the *statistically significant* features for those *machine operators* in the *wholesale and retail sector* are *age*, *hours since the start of work*, and *temperature*. χ^2 correlations show that *contact mode of injury* is the *statistically significant* feature for *machine operators* in the *processing industry*. At the same time, *deviation* and *specific physical activity* are 5th and 6th. However, for *machine operators* in the *wholesale and retail sector*, *deviation* is 2nd and *specific physical activity* is way back in 10th position. These findings can be seen in Figures 2c and 2d.



(a) *Kendall's τ - Processing Industry*

(b) *Kendall's τ - Wholesale and Retail*

(c) χ^2 - *Processing Industry*

(d) χ^2 - *Wholesale and Retail*

Figure 2. Kendall's τ and χ^2 test coefficients concerning target feature for *machine operators* in different industries are shown

3.2.3 Imbalanced Dataset

Imbalanced data refers to an unequal distribution of observations among different classes of the target feature. Imbalanced datasets can lead to biased ML models, as the model may tend to predict the majority class more accurately at the expense of the minority class [2]. This is a severe challenge, especially when the minority class is of greater importance.

Since we utilize the *severity* feature determined by medical practitioners following national guidelines[11] to create the labelled dataset and prediction of severe accidents is of higher importance in our research, we have only $12,541$ observations in the target Class 1 in contrast to $36,304$ observations in target Class 0. The balance of the target classes is shown in Figure 3. The dataset is imbalanced, with non-severe accidents being the majority class and severe accidents being the minority class. To address the issue of imbalanced data, we rely on existing techniques, such as oversampling the minority class, undersampling the majority class or generation of synthetic data for the minority class. We initially evaluate these methods on the dataset and select the most suitable approach for this specific dataset based on our initial experimental results. We further assess the impact of the chosen technique on the model's performance metrics, such as the F1 scores of both target classes and AUROC, to ensure that the final model is robust and accurate for both classes of the target feature.

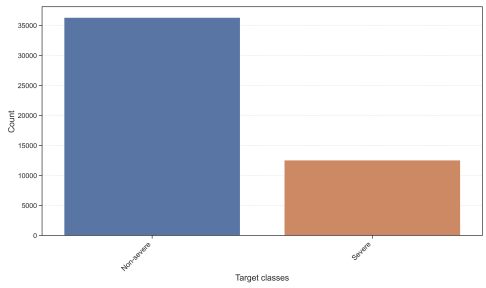


Figure 3. *Severity* data distribution

3.2.4 Summary of Insights

Therefore, the primary challenges of the dataset are unorganized data with noise and missing values, high cardinality in features, unspecified values, variance in features concerning the target feature, and imbalanced data. The combination of these challenges and many categorical predictors in the dataset resulted in a high Mean Absolute Error (MAE) for the initial regression models attempting to predict the continuous numeric

[11] `https://www.ti.ee/media/359/download`

target feature *lost days*. Consequently, this made the findings less meaningful and impractical for real-world application. To address this, we visualize the problem as a binary classification task to predict whether an accident is severe or non-severe (target feature is discussed in Subsection 4.1.8. We intend to propose an automated approach based on extensive data analysis approaches and ML algorithms to predict whether the accident is severe or non-severe, given the features of an accident at the workplace.

4 Proposed Methodology

In this Section, we discuss our proposed methodology in detail. We initially describe each of the available features in our dataset and our proposed methodology for pre-processing each feature in Section 4.1 followed by a brief overview of the ML algorithms used in our research in Section 4.3. We further discuss the sampling algorithms used and describe our reasoning behind selecting the respective sampling algorithm in Subsection 4.2. We, finally, discuss our proposed general model in the Subsection 5.2.1 and for different subsets of the dataset in Subsection 4.4.2.

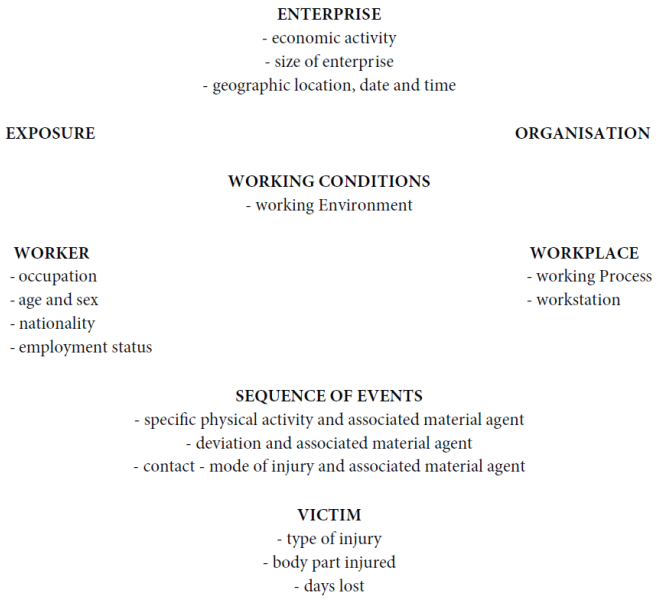


Figure 4. Grouping of the accident features according to ESAW

4.1 Feature Analysis

The dataset comprises information related to 37 different features for $82,641$ accidents. However, as shown in Figure 5, many of these accidents comprise one or more missing features. Therefore, to ensure we can consider the most accidents with the least number of features as missing, we set out the threshold for several features to have missing values as 8, which leads to a loss of $10,473$ rows, i.e., 12.7% of the data.

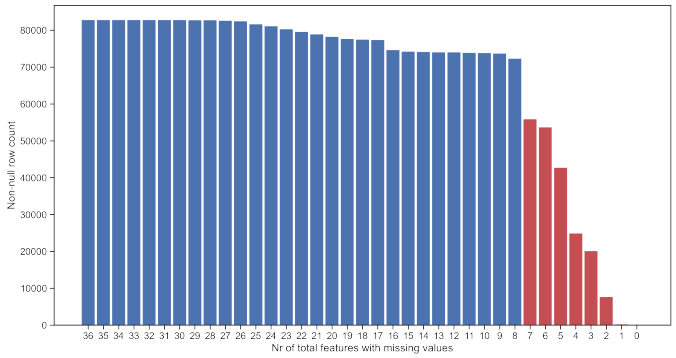


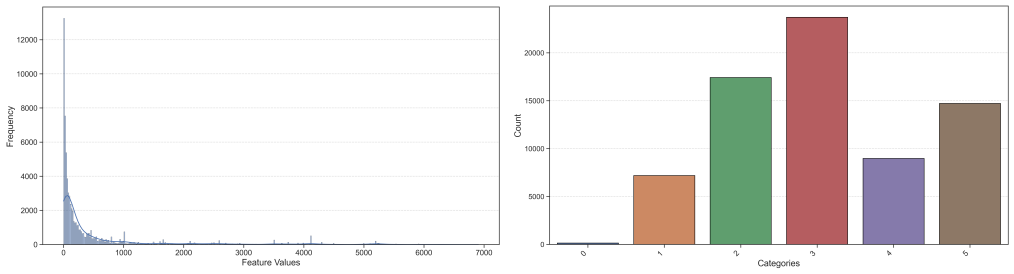
Figure 5. Cumulative count of accidents (y-axis) and the feature count with NaN values (x-axis)

4.1.1 Features of the Enterprise

In ESAW, the features of the enterprise are *economic activity* (activity in the economic sector), *size of the enterprise*, *geographic location*, *date* and *time*. Additionally, we derive *month* and *weekday* from the *date* feature and *sin time*, *cos time* from the *time* feature. We derive the economic sector from the *economic activity* feature, and weather data is added to complement the geographic location.

Enterprise size : We have two features that describe the size of an enterprise: a categorical feature called *enterprise size* and a numerical feature called *employees in the enterprise*. The former contains categories based on different employee count ranges, which can be easily aligned with the ESAW. The latter provides the actual number of employees in the enterprise. We drop the numerical feature and use the categorical feature to represent enterprise size, which follows normal data distribution. The number of zero values (no employees in enterprise) is lesser than in the numerical feature. The numerical feature *employees in the enterprise* and the categorical feature aligned with ESAW are shown in Figure 6a and Figure 6b. Additionally, we will apply Ordinal Encoding [9] to the categorical feature as the *enterprise size* categories have an inherent

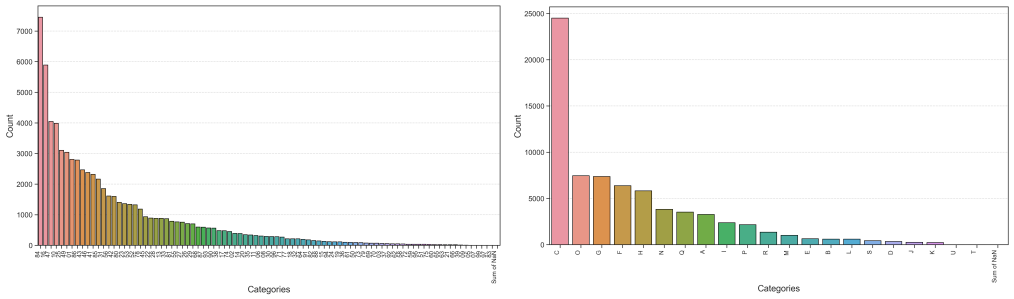
order.



(a) Numerical feature (b) Categorical feature aligned with ESAW

Figure 6. *Enterprise size* features are shown

***Economic activity* and economic sector** : *Economic activity* feature comprises more than $1,000$ unique categories, which makes it impossible to be handled efficiently. To generalize the categories and cover a broader range of *economic activity*, we group the sub-categories into a more general category, thereby reducing the total number of categories. We convert the 5-digit code to a 2-digit version representing the main *economic activity*. For example, we change the code *26121*, which represents the *production of circuit boards* to *26*, i.e., *production of computers, electronic and optical equipment*. We show this in Figure 7a. We introduce a new feature that assigns a letter value from *A* to *U* to identify the different economic sectors where *economic activities* occur (see Table 2). Figure 7b displays economic sectors. This feature helps to filter data by economic sector and is not used as a predictor.



(a) Activities (b) Sectors

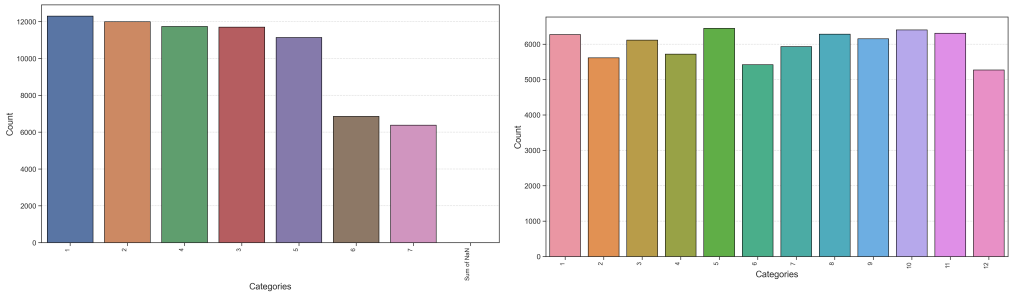
Figure 7. *Economic activities* and economic sectors are shown

Accident date : We create two new features from the *date* feature and remove the original *date* feature as the *date* feature has many unique values, which can lead to overfitting

Table 2. Economic sectors

Sector identifier	**Name of the economic sector**
A	Agriculture, forestry and fisheries
B	Mining industry
C	Processing industry
D	Supply of electricity, gas, steam and air conditioning
E	Water supply; Sewerage, waste and pollution management
F	Construction
G	Retail trade and repair of motor vehicles and motorcycles
H	Transportation and storage
I	Accommodation and catering
J	Information and communication
K	Financial and insurance activities
L	Real estate activity
M	Professional, scientific and technical activity
N	Administrative and support activities
O	Public administration and national defense
P	Education
Q	Health care and social welfare
R	Arts, entertainment and leisure
S	Other service activities
T	Activity of households as employers
U	Activities of extraterritorial organizations and units

in ML models. Therefore, we use the derived *month* and *weekday* features, which capture the temporal patterns and are more suitable for ML tasks. This approach simplifies the modelling process by reducing the dataset's dimensionality, making it computationally and logistically feasible while preserving the information in the original feature. We show our observations for *weekday* and *month* in Figure 8a and 8b, respectively.



(a) *Weekday* (b) *Month*

Figure 8. Accident *weekday* and *month* features data distribution is shown

Accident time : We extract the hour from the time feature of the accident as shown in Figure 9. Additionally, we create a new feature to represent the time feature concerning business hours called *is business hour*. If an accident occurs during business hours, this new feature is assigned a value of 1 or else 0. Furthermore, we observe that the *time* feature of accidents which occurred during the first and last hours of the day are next to each other instead of at opposite ends. To handle this, we create a circular representation of the accident *time* feature using sine and cosine functions instead of applying an Ordinal Encoder directly to the *time* feature. Figure 10 shows the resulting circular representation of the *time* feature. We drop the original *time* feature and the generated *hour of the accident* feature as they are highly correlated with the new *time sine* and *time cosine* features.

Full hours from the start of work : A significant portion of the accidents in the dataset occurred during the first working hour, as indicated by the $8,178$ rows with a zero value. Therefore, although this feature is not originally present in ESAW, we consider it a predictor. Figure 11a presents the original data distribution. We improve the data by replacing zero values with a constant of 0.5, ensuring that models will not confuse these values with other true zero or binary zero values, especially after scaling numerical features. However, several outliers are not making sense in the context of total working hours from the start of work. To address this, we used the interquartile (IRQ) method [11], and all 501 rows with values exceeding only the upper limit of the IRQ are removed. Figure 11b shows the resulting data distribution.

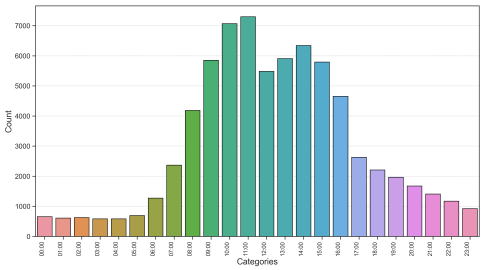


Figure 9. Categorical *time* feature distribution

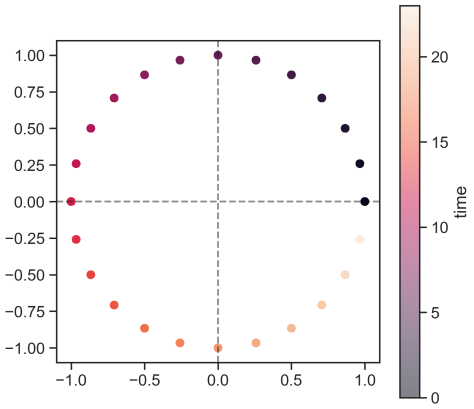
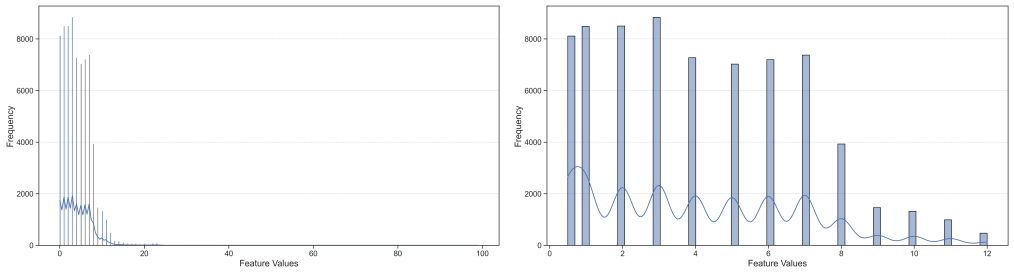


Figure 10. The relationship between the cosine and sine of a *time* feature is shown

Location : We observe that the location of the accident is either an Estonian county or a foreign country. As expected, the value count for a foreign country is very low. We further observed that most of the information belongs to the location *Tallinn* or the locations that belong to *Harju county*, where the capital Tallinn is also located, so we hereby merge the location *Tallinn* with its county. There are only 47 rows with missing values, which we eliminate and the updated data distribution is shown in Figure 12. All rows with the foreign country as a location will be dropped in the next paragraph since no weather data is associated with these accident locations.

Weather data : The Open-Meteo Historical Weather API[12] is the source of the collected weather data. The weather data is associated with the accident *date*, *time*, and *location* and reflects the prevailing weather conditions in the county capital of the accident location. Since the county capital has the highest number of employees and employers,

[12] `https://open-meteo.com/`



(a) Initial data distribution (b) Zeros replaced and outliers removed

Figure 11. *Full hours from the start of work* data distribution before and after cleaning is shown

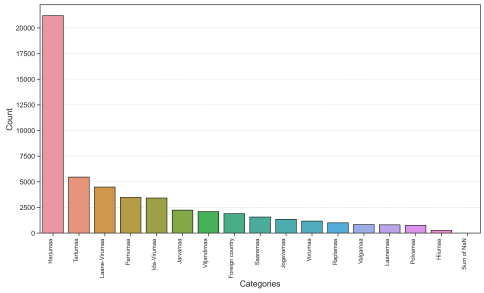
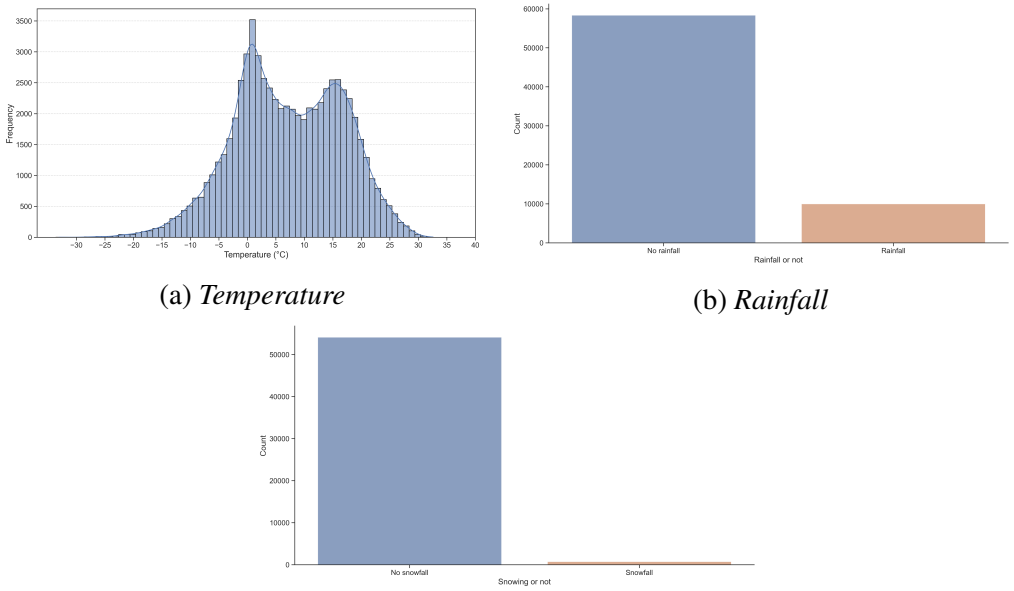


Figure 12. *Location* distribution after cleaning

the weather data provides relatively precise and realistic information regarding the weather conditions that could have contributed to the accident's probability and severity. Three new features were created for the weather data: *temperature* in Celsius, *rainfall* in millimetres and *snowfall* in millimetres. The *temperature* data conforms to a normal distribution. Transformative techniques like log and square root transformations do not produce better results than the original distribution. Figure 13a displays the original data distribution. All rows with no associated weather data are dropped, meaning all accidents outside Estonia will be dropped.

However, we observe that the *rainfall* feature exhibits a substantial number of true zero values, i.e., a total of 86% of the data. These values, if not handled correctly, could potentially diminish the predictive power of the feature. To mitigate this, we convert the *rainfall* feature into a binary feature, where 1 indicates *rainfall* occurrence and 0 indicates the absence. The updated data distribution after converting to binary is shown in Figure 13b. As we observe similarly with the *rainfall* feature, the true zero values in the *snowfall* feature make up 98.7% of the data. We change it into a binary feature with a 1, which indicates the occurrence of *snowfall*, and 0 otherwise. The updated data distribution after converting to binary is shown in Figure 13c.



(a) *Temperature*

(b) *Rainfall*

(c) *Snowfall*

Figure 13. Weather data distributions are shown

4.1.2 Feature of Working Conditions

Working environment : The only feature which characterizes the working conditions during an accident is a *working environment* feature. The *working environment* feature describes where the accident occurred, such as an *industrial site*, *construction site*, *farming*, etc. As 79 different categories for this feature make it difficult to visualize the data, we change them to make it consistent with ESAW guidelines and remove any unknown values, as shown in Figure 14. For example, the string value *025 Construction site - on/over water* was generalised as *020 Construction site, construction, opencast quarry, opencast mine, not specified*.

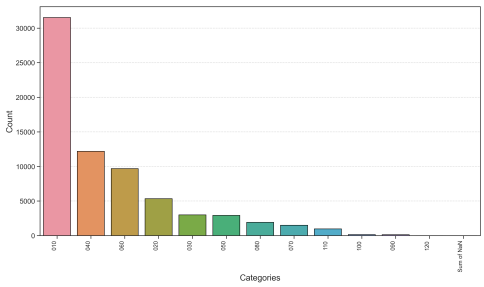
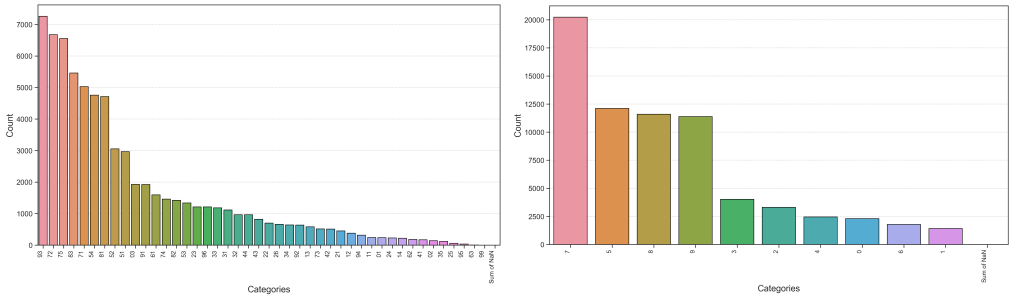


Figure 14. *Work environment* feature data distribution

4.1.3 Features of the Worker

In ESAW, the features of the worker are *occupation*, *nationality*, *sex*, *age*, and *employment status*. Next, we discuss each of them respectively.

Occupation types **and occupation classes** : The *occupation type* feature initially includes over 500 separate categories, making graphical representation difficult like for the *economic activity* feature. We generalize 4-digit codes into 2-digit *occupation type* representations. For example, we generalized the code *7411 - Construction electricians* to *74 - Electrical and electronics industry workers*. Figure 15a displays the result of this code generalisation. In addition, we add a new feature to the dataset to categorize occupations at a more comprehensive level, similar to *economic activities*. This feature assigns a number ranging from *0* to *9* to each observation, and resulting occupation classes are shown in Figure 15b. This feature, similar to the economic sector feature, helps to filter data by occupation class and is not used as a predictor.



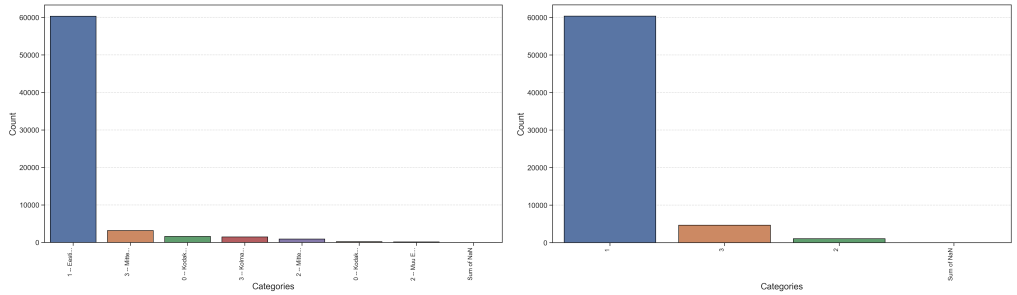
(a) *Occupation types* (b) Occupation classes

Figure 15. Generalised *occupation types* and occupation classes are shown

Table 3. Occupation classes

Class identifier	Name of the 0ccupation class
0	Armed forces occupations
1	Managers
2	Professionals
3	Technicians and associate professionals
4	Clerical support workers
5	Services and sales workers
6	Skilled agricultural, forestry and fishery workers
7	Craft and related trades workers
8	Plant and machine operators and assemblers
9	Elementary occupations

Nationality : Figure 16a shows the initial *nationality* distribution and Figure 16b shows the improved distribution aligned with ESAW. We removed any unknown values.



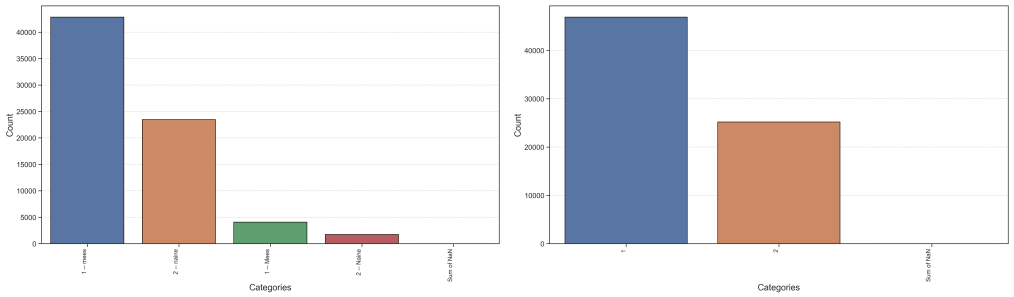
(a) Initial distribution (b) After cleaning

Figure 16. *Nationality* feature data distributions are shown

Sex : Figure 17a presents the initial appearance of the *sex* feature. The improvement according to ESAW is seen in Figure 17b. Male corresponds to *1* and female to *2*.

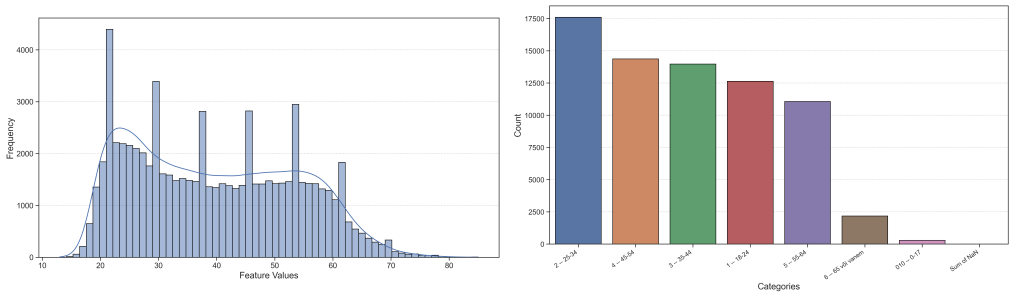
Age : Figure 18b presents the original categorical feature *age group*, and Figure 18a shows the numerical *age* feature. The distribution of the numerical feature indicates good quality because it resembles a normal distribution, and its initial skewness value [15] is only 0.26. We remove the highly correlated categorical analogue *age group* from the dataset.

Employment status : We show the initial *employment status* feature and the updated representation respectively in Figures 19a and 19b.



(a) Initial distribution (b) After cleaning

Figure 17. *Sex* feature distribution before and after cleaning is shown



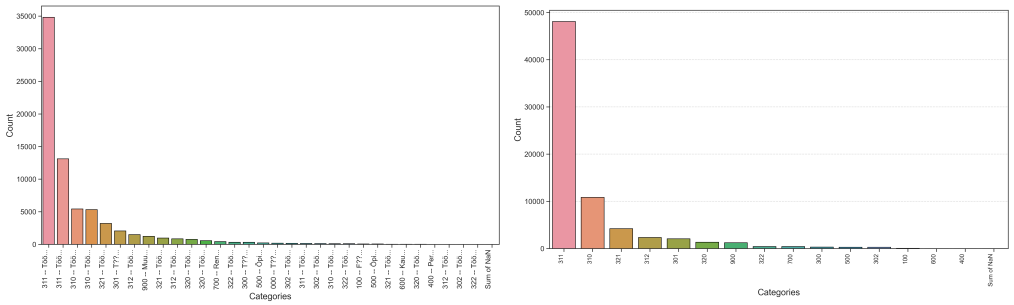
(a) *Age* (b) *Age group*

Figure 18. Numerical *age* and categorical *age group* features are shown

Employment years : The *employment years* feature exhibits several outliers with values equal to or exceeding $2,000$, which are erroneous. Moreover, 903 rows contain NaN-type values which suggest unknown values rather than missing data, and these are considered missing completely at random[13]. Therefore, $19,219$ rows have a value of 0, indicating that most accidents happen during employees' first year of work. The zeros in *employment years* are not precisely zero, so we use the median value of all values greater than 0 but less than 1 to replace the zero values. This means that the duration of employment years is not considered zero but as a median of those with a duration of employment years less than one year. These processing effects are shown in Figure 20.

Sauga et al. [26] explains that if the data follows a normal distribution, most data ranges between 1-3 standard deviations. If the model assumes a normal distribution and uses this assumption to make predictions, it produces more accurate predictions. We can apply transformations like a log, square root, and Box-Cox on *employment years* or *age* features to enhance their data distribution. However, in our case, lowering skewness values does not necessarily improve model performance and may sometimes worsen it, so we decided not to include data transformations in the preprocessing. We also tried the

[13] `https://medium.com/@kyawsawhtoon/a-guide-to-knn-imputation-95e2dc496e`



(a) Initial distribution (b) After cleaning

Figure 19. *Employment status* feature before and after cleaning is shown

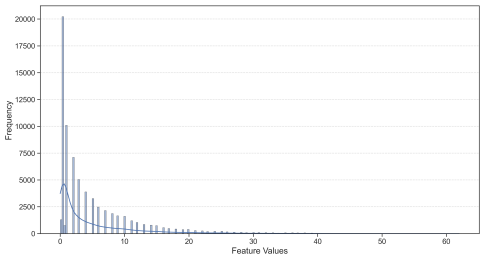


Figure 20. Numeric feature *employment years* is shown

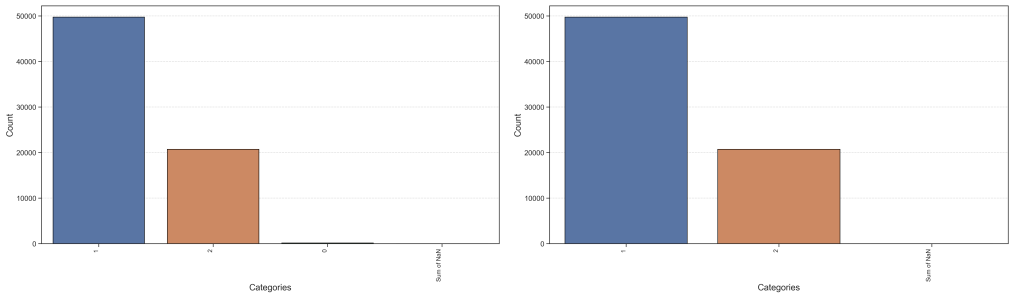
k-Nearest Neighbors (kNN) imputer[14] and applied it to the *employment years* feature. Still, its effect on models is similar to the data transformations, and we do not employ it in our final dataset.

4.1.4 Features of the Workplace

In ESAW, the workplace features are the *workstation* and *working process*.

Workstation : The *workstation* represents the victim's job post's usual or occasional nature at the time of the accident. After removing 147 rows with unknown values, the *workstation* feature divides into two categories: the traditional workstation in the usual work area, represented by the string value *1*, and the mobile workstation, represented by the string value *2*. Figure 21a illustrates the original data distribution of this feature, while Figure 21b presents the updated data distribution after aligning with the ESAW and removing unknown values.

[14] `https://machinelearningmastery.com/knn-imputation-for-missing-values-in-machine-learning/`



(a) Initial distribution (b) After cleaning

Figure 21. *Workstation* feature data distribution before and after cleaning is shown

Working process : The *working process* refers to the main type of work or task the victim carries out during the accident, such as *manufacturing*, *excavation*, or *intellectual activities*. Initially, the *working process* feature had 59 categories, making data visualization challenging. We generalised the codes to simplify and align the data with the ESAW. For instance, the string value *25 Demolition - all types of construction* was generalized as *20 Excavation, Construction, Repair, Demolition, not specified*. Figure 22 shows the resulting data distribution.

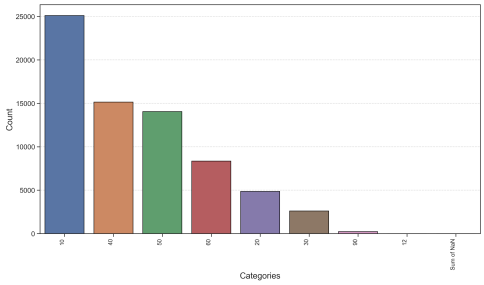
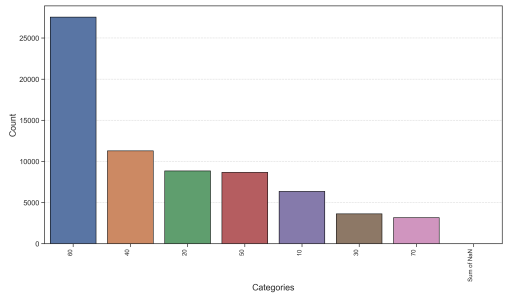


Figure 22. *Work process* feature data distribution

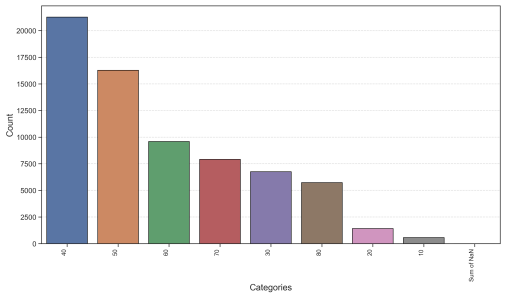
4.1.5 Features of the Sequence of Events and Associated Material Agents

In ESAW, the sequence of events features are the *specific physical activity*, *deviation* and *contact - mode of injury* with their associated *material agents*. The victim does a *specific physical activity* at the exact time of the accident. The last event differing from the norm leads to the accident, called the *deviation*. The *contact mode of injury* causes the injury. These three features have a specific coding scheme outlined in ESAW. The *deviation* and *contact mode of injury* features only have one missing value, which we remove during cleaning. We apply a generalisation of codes similar to the *working environment* and

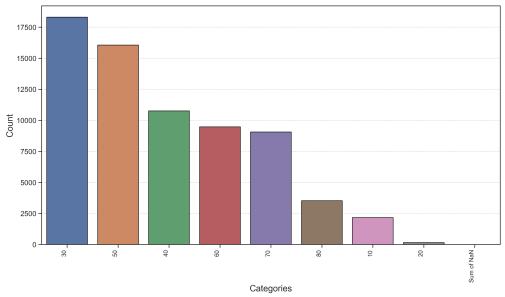
working process features. The resulting data distributions are shown in Figures 23a, 23b, and 23c.



(a) *Specific physical activity*



(b) *Deviation*



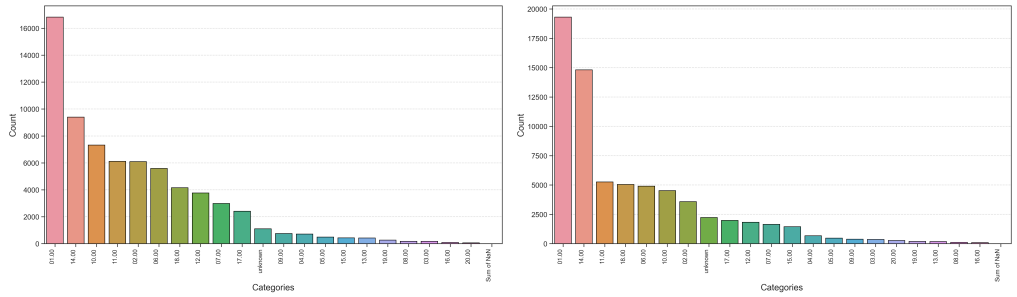
(c) *Contact-mode of injury*

Figure 23. Sequence of events features data distribution after cleaning and grouping is shown

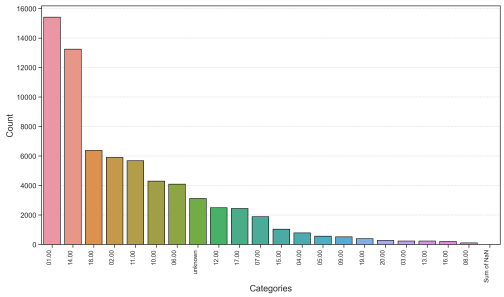
The accident dataset includes information on the *material agents* associated with *specific physical activity*, *deviation*, and *contact mode of injury*. *Machines*, *chemicals*, *hand tools*, etc. can be *material agents*. However, visualising these features is challenging due to the code format, such as *07.15.00.00*. Each feature initially has a cardinality value of approximately $1,200$. Since the ESAW addresses these features, we apply a coding scheme outlined so the coding generalisation is performed similarly to previous features. For example, the string value from the *deviation material agent 14.01.01.03 – Logs, support beams* is generalised as *14.00 - Materials, objects, products, machine or vehicle components, debris, dust, not specified*. We remove missing values during the cleaning process, and the resulting data distributions are seen in Figures 24a,24b, and 24c.

4.1.6 Feature of Accident Causes

Although this feature is not included in ESAW, we consider it useful to understand the accident's severity. The dataset provides a set of cause codes explaining reported causes of accidents. We organise codes into a list of unique items. For instance, the list



(a) *Material agents* of the *specific physical activity* (b) *Material agents* of the *contact mode of injury*



(c) *Material agents* of the *deviation*

Figure 24. *Material agents* of the sequence of the events after cleaning and generalizing are shown

representation of the cumbersome string value *,003,015, – 003 - Violation of occupational safety requirements by an employee 015 - Other reasons* is presented as *[003,015]*. We use dummy variables [16] to create new features generated from the lists of cause codes, resulting in 19 new features. If a particular observation lists one or more cause codes, the corresponding row values for new features are set to 1; otherwise, they are set to 0. Indicating the presence of a specific cause or causes of the accident. The dataset includes only rows that contain at least one cause code.

4.1.7 Features of the Victim

We have excluded features, such as *lost days*, *type of injury* and *injured body part*, which are highly related to the target feature, i.e., the severity of an accident. So, we leave out highly correlated and post-accident features to create more useful models for predicting and preventing accidents.

4.1.8 Severity as a Target

Figure 25 displays the *severity* feature, which has three categorical values and is chosen as a target. Professional medical practitioners primarily determine *severity* classifications based on national guidelines[15] for assessing severe health impairment. The string value *10* represents non-severe accidents, while *20* and *22* represent severe and fatal accidents, respectively. The target feature assigns a value of 0 for non-severe accidents and a value of 1 for severe or fatal accidents. Therefore, Class 0 has $36,304$ observations, and Class 1 has $12,541$ observations. Severity division into severe and non-severe accidents is seen in Figure 3. We cover the resulting dataset next in Subsection 4.1.9.

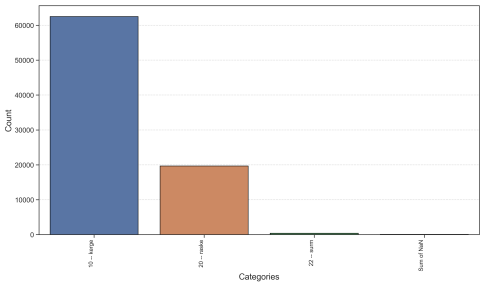


Figure 25. *Severity*'s' initial representation

4.1.9 Final Dataset

We only consider accidents that resulted in at least one lost day for the final dataset to ensure consistency with Estonian occupational accident reporting regulations. Our final dataset includes $48,845$ observations, which were then narrowed down to 47 columns after dropping the victim's features. Among these, 47 are the features and the target. The final dataset is shown in Tables 4 and 5, respectively.

[15] `https://www.ti.ee/media/359/download`

Table 4. Cardinality and NaN count of the final dataset - part 1

Feature	**Cardinality**	**NaN count**	**Datatype**
business_area	87	0	object
sex	2	0	object
age	73	0	float64
employment_status	15	0	object
employment_years	74	0	float64
full_hours_from_startofwork	13	0	float64
location	15	0	object
citizenship	3	0	object
profession_code	43	0	object
workstation	2	0	object
working_environment	12	0	object
working_process	7	0	object
specific_physical_activity	7	0	object
material_agent_of_physical_act	21	0	object
deviation	8	0	object
material_agent_of_deviation	21	0	object
contact_mode_of_injury	8	0	object
material_agent_of_contact_mode	21	0	object
enterprise_size_ordinal_enc	6	0	float64
dayofweek	7	0	object
month	12	0	object
sin_time	20	0	float64
cos_time	22	0	float64
is_business_hour	2	0	int64
temperature	570	0	float64

Table 5. Cardinality and NaN count of the final dataset - part 2

Feature	Cardinality	NaN count	Datatype
rain	2	0	int64
snowfall	2	0	int64
cause code 001	2	0	int64
cause code 002	2	0	int64
cause code 003	2	0	int64
cause code 004	2	0	int64
cause code 005	2	0	int64
cause code 006	2	0	int64
cause code 007	2	0	int64
cause code 008	2	0	int64
cause code 009	2	0	int64
cause code 010	2	0	int64
cause code 011	2	0	int64
cause code 012	2	0	int64
cause code 013	2	0	int64
cause code 014	2	0	int64
cause code 015	2	0	int64
cause code 017	2	0	int64
cause code 018	2	0	int64
cause code 019	2	0	int64
cause code 025	2	0	int64
target	2	0	int64

4.2 Imbalanced Datasets and Sampling Algorithms

As discussed in Subsection 3.2.3, there are $36,304$ observations for Class 0 and $12,541$ observations for Class 1. Additionally, as we are more interested in predicting Class 1, we must ensure that the proposed model can accurately predict Class 1. We discuss how we address the class imbalance challenge.

4.2.1 Random Under-sampling

Ali *et al.* [1], in their study about class imbalance problems, explains that RUS is the most basic and straightforward resampling method for imbalanced datasets. By randomly eliminating samples from the majority class, they balanced the distribution of classes for the learning process, with the disadvantage being the potential loss of valuable samples. To address this shortcoming, they propose oversampling techniques. We discuss our

application of RUS in Section 4.4.

4.2.2 Random Over-sampling

Ali *et al.* in [1] emphasizes that the simplest oversampling technique is ROS, which generates new samples in the minority class by selecting samples randomly. However, they point out that the generated samples may be too similar to the original ones, leading to overfitting. Therefore, they suggest that ROS is used where the number of samples in the minority class is significantly less than that of the majority class. We discuss our application of ROS in Section 4.4.

4.2.3 Synthetic Minority Oversampling Technique SMOTE

Chawla *et al.* in [3] analyzed the effects of oversampling with a replacement on minority class recognition and found that it does not significantly improve recognition. Instead, it leads to overfitting as the decision region becomes more specific without expanding into the majority class region. To address this, they proposed an over-sampling technique called Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic examples rather than replicating minority class samples. This method creates more extensive and less specific decision regions, allowing for better generalization and improved minority class recognition than oversampling with replacement. We discuss our experiment concerning SMOTE in Section 4.4.

4.3 Machine Learning Algorithms

In this Section, we describe the different ML models used in our research: Random Forest, Light Gradient Boosting Machine, Extreme Gradient Boosting, Support Vector Machine, and Logistic Regression.

4.3.1 Random Forest

RF is an ensemble learning method that combines multiple DTs to perform classification, regression and other tasks. The RF output for a classification task is the class that most trees select. As RF builds multiple DTs in randomly selected sub-spaces of the feature space, it can generalize well and reduce overfitting their classification in complementary ways. It can be combined to improve training and unseen data accuracy. Understanding the relationship between input features and output is more challenging for RF models. However, techniques such as feature importance can help provide insights into the model's decision-making process. RF does not handle categorical variables as naturally as numerical variables [10].

4.3.2 Light Gradient Boosting Machine

LightGBM is the Gradient-Boosting Decision Tree (GBDT) algorithm that addresses the limitations of existing GBDT implementations such as XGBoost and pGBRT, particularly concerning their efficiency and scalability in handling high-dimensional feature data and large datasets. LightGBM utilizes two innovative techniques: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), to overcome the limitations of traditional GBDT implementations. GOSS allows for the exclusion of data instances with small gradients, while EFB bundles mutually exclusive features to reduce the number of features and thus the algorithm's computational complexity [14].

4.3.3 Extreme Gradient Boosting

XGBoost is a highly popular ML library that consistently outperforms other algorithms in supervised learning tasks, making it a top choice for solving ML challenges and winning competitions. Designed for large-scale datasets, it combines efficiency and accuracy, making it a popular choice among data scientists. It can efficiently deal with datasets with many missing or zero-valued features. The algorithm is designed to split the data into different branches based on the feature values. The quantile sketch method allows the algorithm to find the best possible splits, even when using an approximate method. By optimizing cache access patterns, data compression, and sharding, XGBoost can solve real-world scale problems with minimal resources and outperform other tree-boosting methods on tasks like classification. However, it can overfit data with too many trees or too deep. It is essential to tune hyperparameters carefully and use techniques like cross-validation to avoid overfitting [4].

4.3.4 Support Vector Machine

SVM maps input vectors to a high-dimensional feature space and constructs a linear decision surface to separate the groups. The algorithm combines optimal hyperplanes, convolution of dot-product, and soft margins to allow for errors in the training set. SVMs have high generalization ability and can be extended to non-separable data through polynomial input transformations. SVMs have been shown to outperform other classical learning algorithms [6].

4.3.5 Logistic Regression

LR is a statistical method that models the relationship between a binary response variable and one or more predictor variables. It uses the logistic function to simplify mathematical theory and has applications in various fields such as medicine, biology, and ecology. The value of LR lies in simplifying mathematical theory, and it can be used in cases of estimation rather than significance testing [7]. It is widely used in predicting patient

outcomes, where predictors refer to independent factors being analyzed, and outcomes represent the dependent variable. By dealing with binary outcomes, the LR model calculates the odds of a particular outcome occurring and uses the odds ratio to measure the influence of each predictor on the outcome. However, the model's validity relies on the number and suitability of predictor variables, and collinearity can cause errors or uncertainty in the estimates. Additionally, the variables must maintain a constant magnitude of association across their range of values, which may not always be true. Furthermore, LR assumes that the value of another predictor does not influence the effect of one predictor [29].

4.4 Proposed Models

We discuss how we incorporate the specific features and pre-processing details to predict the severity of accidents using different ML models. We discuss this next in detail for a generalized model in Subsection 4.4.1 and specifically, in Subsections 5.2.2, 5.2.3 and 5.2.4.

4.4.1 Generic Proposed Model

We follow a Stratified K-Fold Cross Validation[16] where K is 5, i.e., the data is split into five parts such that each part has a similar proportion of target class. We train the proposed generic model on each of these parts. We also consider the same random seed for each model to ensure our results are consistent. We repeat this for the ML models, i.e., described in Section 4.3.

For the features, we investigate different scenarios to understand the applicability of features, i.e., a) we use all features (details in 4.1) and b) *statistically significant* features (we describe this later in this Subsection). In both scenarios, we consider the sampling techniques (as mentioned in Section 4.2). Our observations in Subsection 5.2.1 show that ROS and RUS perform better than the SMOTE technique, considering target Class 1, by approximately 40%. To ensure that the features are *statistically significant*, we follow different approaches for categorical and numerical features. For every categorical feature, we follow a Label Encoding technique [21], which ensures conversion into numerical data. We perform a χ^2 test to determine whether the feature is *statistically significant*. We consider a feature to be *statistically significant* if the p-value is below 0.05, i.e., the distribution of observed frequencies in each category with the expected frequencies, assuming no relationship (null hypothesis) between the feature and the target variable. Similarly, we consider Kendall's rank correlations for numerical features, which measure the association between the feature and the target variable. If the p-value of the test is below 0.05, we consider these features as *statistically significant*.

[16] `https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation`

We calculate the AUROC, F1 scores for both classes and average F1 score (details of the metrics are provided in Section 5.1). To understand the performance of the proposed model for each class, we calculate the F1 score for both target classes and the average F1 score, respectively. We identify the best model based on the performance of each of these ML models for different feature sets and sampling techniques. We discuss our results in detail in Subsection 5.2.1.

Since we aim to provide more targeted insights for different stakeholders in developing effective occupational health and safety strategies, we tried specific subsets of features as mentioned by ESAW guidelines, i.e., *enterprise*, *working conditions*, *worker*, *workplace*, *sequence of events*, and *victim*. Although we discuss our results in detail in Section 5, we briefly discuss them next. Our observations showed significantly low performance, i.e., Class 1 F1 score of 0.48 for *statistically significant* features and target classes balanced with ROS. Therefore, as previously highlighted, the development of a generic approach might not be able to cater to different economic sectors and occupation classes. This granular approach allows a better understanding of workers' unique risks and challenges in various roles and industries. Consequently, businesses in these sectors can benefit from tailored insights that help them design more effective occupational health and safety strategies. We discuss this specific scenario-based model next.

4.4.2 Specific Scenario-based Proposed Model

As previously discussed in Subsection 4.4.1, the high variance in features and huge dataset makes it impossible to propose a generic model for the entire dataset. Therefore, in this Subsection, we propose specific models for different scenarios, such that each scenario represents a particular economic sector and occupation class, namely, a) *craft and related trades workers* in the *retail trade and repair of motor vehicles and motorcycles sector* ($G7$), b) *professionals* in the *health care and social welfare sector* ($Q2$), and c) *elementary occupations* in the *construction sector* ($F9$), respectively. We could not consider all the possible combinations of economic sectors and occupation classes. The dataset size was insufficient for an ML model to train and might lead to overfitting. Additionally, we calculate the proportions of the positive target class in the test data and its 95% confidence interval [12]. We consider economic sector and occupation class combination only if the margin of error for the test set target size is 0.1 or less. We follow a similar setup as the general proposed model, which we discuss briefly next.

We follow a Stratified K-Fold Cross Validation where K is 5, i.e., the data is split into five parts such that each part has a similar proportion of the target class. We train the proposed general model on each of these parts. We also consider the same random seed for each model to ensure our results are consistent. We repeat this for the ML models described in Section 4.3. For the features, we consider different sets, such as a) all features, b) *statistically significant* features and c) *important* features. We follow the same procedure as in the generic proposed model to identify the *statistically significant*

features. For the categorical features, we initially perform Label Encoding and a χ^2 test to identify the *statistically significant* categorical features. Similarly, we perform Kendall's rank correlation tests to find *statistically significant* numerical features. Additionally, we ensure that each numerical value ranges between 0 and 1.

To identify the *important* features for $G7$, $Q2$, $F9$, we consider the best-performing model for the particular scenario. We, then, iteratively, remove a random feature, say R_f, retrain the best-performing model and consider R_f to be in the *important* features only if the F1 score for Class 1 decreased after removing. We repeat this process at least 50 times and name the best-performing feature set as an *important feature set*. The best-performing model differs for $G7$, $Q2$, and $F9$. While it's an LR model for the $G7$, $Q2$, SVM performs the best for $F9$. We calculate the AUROC, F1 scores for both the classes and the average F1 score (details of the metrics are provided in Section 5.1) for all the feature combinations. To understand the performance of the proposed model for each class, we calculate the F1 score for both classes and the average F1 score. We identify the best model based on the performance of each of these ML models for different feature sets. We discuss our results in detail in Section 5.

5 Experiments and Results

In this Section, we discuss our experiments and observations. We initially briefly describe the metrics used in Section 5.1 followed by a detailed discussion in the Subsection 5.2.1 about the general model and in Subsections 5.2.2, 5.2.3, 5.2.4 about the scenario-based models. We discuss the limitations of the proposed model in Subsection 5.3, and lastly, we describe implementation details in Subsection 5.4.

5.1 Metrics

We discuss the corresponding metrics used in this research next.

1. **Precision** : It is the proportion of correctly identified positive samples among all predicted positive samples [22].

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

2. **Recall** : It is the proportion of correctly identified positive samples among all actual positive samples [22].

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

3. **F1 score** : It is the harmonic mean of precision and recall. It is a single score that balances both metrics and is often used to evaluate the overall performance of a binary classification model [22].

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

Area Under Receiver Operating Curve AUROC : It measures the probability of a model correctly ranking a randomly chosen positive example higher than a randomly chosen negative example [23].

$$AUROC = \int_0^1 TPR(FPR^{-1}(t))\, dt \tag{4}$$

5.2 Results and Discussions

5.2.1 Generic Model

In the initial phase, we train our models on all available features and only on *statistically significant* features. We use three different re-sampling techniques.

Sampling Techniques : Our observations in Table 6 indicate that ROS and RUS perform similarly based on their F1 scores for Class 0 and Class 1. Specifically, for ROS, the F1 scores are 0.75 for Class 0 and 0.51 for Class 1, while for RUS, the F1 scores are 0.73 for Class 0 and 0.50 for Class 1. Even though SMOTE outperforms ROS and RUS on average by 0.09 for Class 0, it performs similarly to ROS with a difference of only 0.01 for Class 1. Since we focus on Class 1, we choose ROS as the sampling technique. Additionally, the dataset size decreases significantly when we consider specific scenarios of economic sectors and occupation classes. This is another reason we rely on ROS, as it ensures that the target class is adequately represented. We show the Receiver Operating Curves (ROC) of the best-performing ML models in Figure 26. We consider the LightGBM model balanced with ROS as our baseline model. This allows us to benchmark the specific scenario-based models.

Machine Learning Models : Table 7 presents our observations on different feature sets. All models have similar performance, based on their F1 scores for Class 1, with LightGBM trained on all features being marginally better than other models by 0.01. RF performs better than other models by 0.09 on average and equally in both feature sets. We can consider the LightGBM as our best-performing model on the set of all features. Most of the features may capture unique aspects of the problem, thus providing complementary information that helps improve model performance and therefore is considered *statistically significant* and the resulting dataset will be very similar to the actual unfiltered dataset.

However, as our results indicate, generic models might not address the specific needs of individual sectors and occupations. Therefore, it provides very little insight for the

companies to identify appropriate actions that can mitigate risks and ensure the safety of their employees. Therefore, we focus on combining economic sector and occupation class data to identify these models' effectiveness and specific factors on which the sector and class can focus. In the following Subsections 5.2.2, 5.2.3, and 5.2.4, we discuss the three top-performing combinations of the economic sector and occupation class in more detail.

Table 6. Comparison of best generic models using different sampling techniques

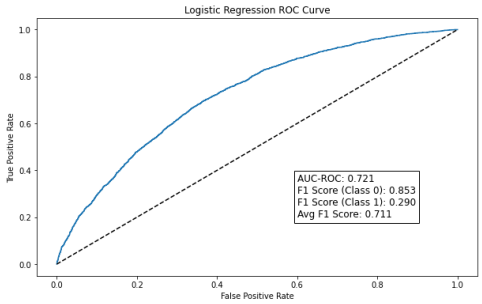
Balancing technique	**Best model**	**AUROC**	**AVG F1**	**Class 0 F1**	**Class 1 F1**
Unbalanced	LR	0.72	**0.71**	**0.85**	0.29
ROS	LightGBM	**0.73**	0.69	0.75	**0.51**
RUS	SVM	**0.73**	0.67	0.73	0.50
SMOTE	XGBoost	0.71	**0.71**	0.83	0.38

Table 7. Comparison of best generic models using different feature sets

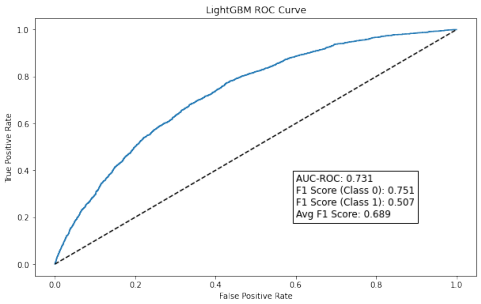
Feature set	**Best model**	**AUROC**	**AVG F1**	**Class 0 F1**	**Class 1 F1**
All features	RF	0.72	**0.72**	**0.84**	0.37
	LightGBM	**0.73**	0.69	0.75	**0.51**
	XGBoost	**0.73**	0.68	0.73	0.50
	SVM	**0.73**	0.70	0.77	0.50
	LR	0.72	0.68	0.74	0.50
Significant features	RF	0.72	**0.72**	**0.84**	0.36
	LightGBM	**0.73**	0.69	0.75	0.50
	XGBoost	**0.73**	0.68	0.73	0.50
	SVM	**0.73**	0.70	0.77	0.50
	LR	0.72	0.67	0.73	0.50

5.2.2 G7 - *Craft and Related Trades Workers* in the *Retail Trade and Repair of Motor Vehicles and Motorcycles*

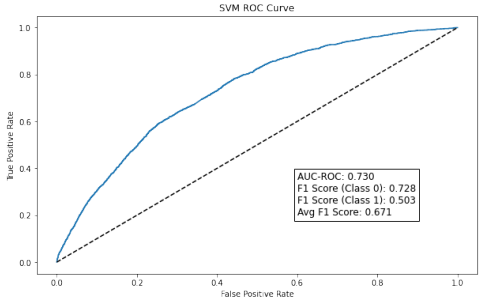
In this Subsection, we compare the performance of the *statistically significant* features and *important* features (discussed in Section 4.4.2) for all the ML models (discussed in Section 4.3) to determine a best-performing model for *craft and related trades workers*



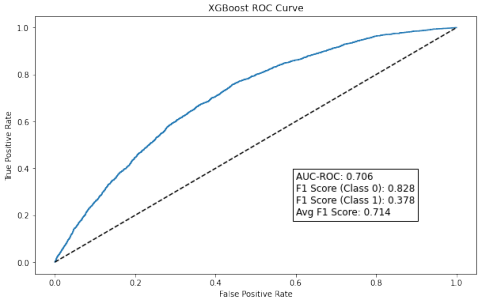
(a) Best model with unbalanced data



(b) Best model with ROS



(c) Best model with RUS



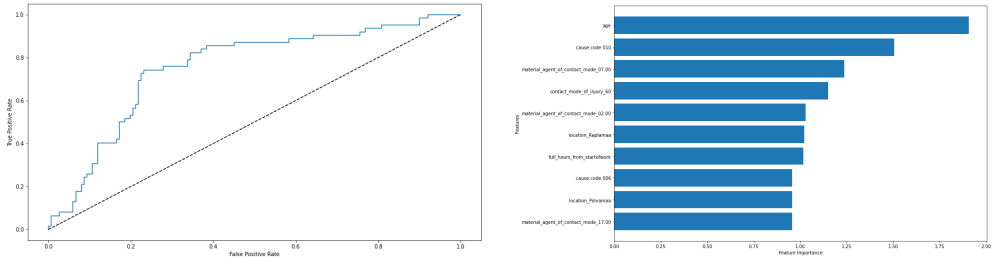
(d) Best model with SMOTE

Figure 26. ROC's of best generic models using different sampling techniques are shown

in the *retail trade and repair of motor vehicles and motorcycles* sector. Our observations, as shown in Table 8, indicate that using *important* features outperform the *statistically significant* features in terms of average F1 score (0.77 versus 0.73) and individual F1 scores for both Class 0 (0.85 versus 0.82) and Class 1 (0.64 versus 0.62). Additionally, we observe that the LR model performs best on *important* features with Class 0 and Class 1 F1 scores 0.81 and 0.64, respectively. Figure 27a displays the ROC for the best model with *important* features. We discuss next our analysis on *important* features next.

Table 8. Comparison of best models using different feature sets for $G7$

Feature set	**Best model**	**AUROC**	**AVG F1**	**Class 0 F1**	**Class 1 F1**
Significant features	RF	0.69	0.71	0.82	0.44
	LightGBM	0.69	0.69	0.79	0.45
	XGBoost	0.71	0.71	0.77	0.56
	SVM	0.74	0.73	0.80	0.58
	LR	0.76	0.73	0.78	0.61
Important features	RF	0.72	0.72	0.84	0.44
	LightGBM	0.76	0.72	**0.85**	0.58
	XGBoost	0.69	0.72	0.79	0.56
	SVM	0.74	0.77	0.83	0.62
	LR	**0.77**	**0.77**	0.81	**0.64**



(a) ROC

(b) *Important* features

Figure 27. Logistic Regression ROC and *important* features affecting the target Class 1 are shown

We identify *important* features by considering the top 10 features on the basis of their score of absolute coefficients as shown in Figure 27b such that these features can ensure a positive effect with respect to Class 1. Our observations highlight the importance of age, non-compliance of work equipment with safety requirements, lack of personal protective equipment, full hours from the beginning of the workday, and contact mode of injury involving trapping or crushing as key predictors of severe occupational accidents among *craft and related trades workers* in the *retail trade and repair of motor vehicles and motorcycles sector*. For example, age is an important predictor, indicating that the risk of severe accidents increases as workers age, highlighting the importance of considering age-specific interventions and training to enhance workplace safety for older workers. Furthermore, we identify two primary causes of severe accidents: non-compliance of work equipment with safety requirements and lack of personal protective equipment.

These findings emphasize the need for employers to ensure that work equipment adheres to safety standards and that workers have access to proper protective gear.

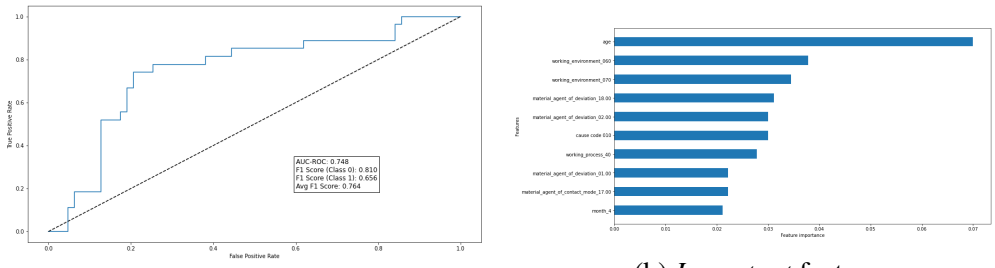
The number of full hours from the beginning of the workday or work shift also plays a crucial role in predicting severe accidents. With more hours from the start of work, the ratio of severe to non-severe accidents grows, suggesting that fatigue may be a contributing factor. Implementing measures such as regular breaks, ergonomic workstations, and adequate staffing can help mitigate the risk of accidents caused by fatigue. Similarly, the contact mode of injury involving trapping or crushing is identified as a critical factor in severe accidents. This finding underscores the importance of implementing safety measures, such as regular equipment inspections and proper machine guarding, to prevent such incidents. Therefore, we believe by addressing these factors, employers can significantly reduce the risk of severe accidents and improve workplace safety.

5.2.3 Q2 - *Professionals* in the *Health Care and Social Welfare Sector*

In this subsection, we assess the performance of both *statistically significant* features and *important* features (addressed in Section 4.4.2) across all machine learning models (outlined in Section 4.3) to establish the best-performing model for *professionals* within the *health care and social welfare* sector. As demonstrated in Table 9, our findings reveal that employing *important* features surpasses the use of *statistically significant* features in terms of average F1 score (0.79 compared to 0.65) and individual F1 scores for both Class 0 (0.86 compared to 0.79) and Class 1 (0.66 compared to 0.54). Moreover, the LR model best performs when utilizing *important* features, achieving F1 scores of 0.81 and 0.66 for Class 0 and Class 1, respectively. Figure 28a showcases the ROC for the top-performing model based on *important* features. The following section discusses our analysis of the *important* features.

Table 9. Comparison of best models using different feature sets for $Q2$

Feature set	Best model	AUROC	AVG F1	Class 0 F1	Class 1 F1
Significant features	RF	0.64	0.64	0.79	0.29
	LightGBM	0.59	0.63	0.74	0.35
	XGBoost	0.63	0.64	0.72	0.46
	SVM	0.71	0.64	0.70	0.50
	LR	0.68	0.65	0.70	0.54
Important features	RF	0.71	0.73	0.84	0.48
	LightGBM	0.69	0.78	**0.86**	0.60
	XGBoost	0.69	0.70	0.78	0.53
	SVM	0.72	**0.79**	0.85	0.63
	LR	**0.75**	0.76	0.81	**0.66**



(a) ROC

(b) *Important* features

Figure 28. LR model ROC curve and *important* features affecting the target Class 1 are shown

To ensure high performance concerning Class 1, we identify *important* features by considering the top 10 features based on their score of absolute coefficients as shown in Figure 28b. Our observations highlight the importance of age, public area and home or communal parts of the building as the places where the victim was present or working just before the accident and humans or buildings as material agents associated with the abnormal event leading to the accident.

Similar to the findings in Subsection 5.2.2, age is a crucial predictor highlighting the significance of implementing age-specific interventions and training programs to improve workplace safety for older workers. *Professionals* in the *healthcare and social welfare* sector are at risk of serious injuries from human contact or animal bites. These injuries may occur when healthcare workers attempt to restrain patients who are agitated or confused or when dealing with animals brought in for treatment. Healthcare organizations

must provide appropriate training, education, and equipment to ensure the safety of healthcare workers at risk of exposure to infectious agents and injuries from humans or other living organisms. Additionally, infection control protocols, proper handling of sharp objects, and workplace safety guidelines can help prevent these types of injuries and illnesses in the *healthcare and social welfare* sector.

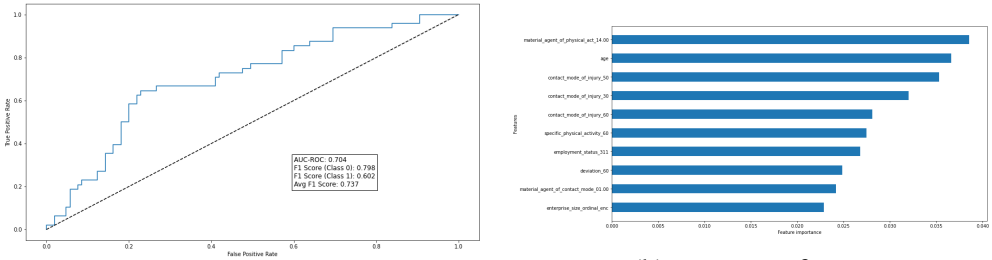
The importance of work environment variables, specifically public areas and home visits, highlights the need for additional safety measures for healthcare professionals in these settings. Public areas can be crowded and unpredictable, increasing the risk of accidents, especially for medical staff carrying equipment or working in a fast-paced, high-stress environment. Home visits also present unique risks, such as exposure to domestic animals, unstable environmental conditions, and violent or intoxicated patients or their family members. Therefore, it is crucial for healthcare organizations and institutions to provide adequate safety training and resources to medical staff who work in these environments. Additionally, healthcare workers could benefit from cooperating with law enforcement agencies to ensure their safety when working in public areas or during home visits. Hence, we believe employers can effectively mitigate the risk of severe accidents and enhance workplace safety by focusing on these factors.

5.2.4 F9 - *Elementary Occupations* in the *Construction Sector*

In this Subsection, we evaluate the performance of both *statistically significant* features and *important* features (discussed in Section 4.4.2) across all machine learning models (outlined in Section 4.3) to identify the best-performing model for *elementary occupations* within the *construction* sector. As illustrated in Table 10, our findings indicate that using *important* features yields better results than *statistically significant* features, as evidenced by a higher average F1 score (0.79 compared to 0.65) and individual F1 scores for both Class 0 (0.86 compared to 0.79) and Class 1 (0.66 compared to 0.54). Furthermore, the LR model performs best on *important* features, with F1 scores of 0.81 and 0.66 for Class 0 and Class 1, respectively. Figure 29a presents the ROC for the best model utilizing important features. We discuss our analysis of the *important* features in the following section.

Table 10. Comparison of best models using different feature sets for $F9$

Feature set	Best model	AUROC	AVG F1	Class 0 F1	Class 1 F1
Significant features	RF	0.68	0.66	0.75	0.46
	LightGBM	0.68	0.67	0.76	0.47
	XGBoost	0.68	0.63	0.69	0.48
	SVM	0.66	0.68	0.74	0.53
	LR	0.70	0.63	0.70	0.49
Important features	RF	**0.72**	0.70	**0.82**	0.44
	LightGBM	0.69	**0.74**	**0.82**	0.54
	XGBoost	0.68	0.66	0.79	0.49
	SVM	0.70	**0.74**	0.80	**0.60**
	LR	0.67	0.69	0.74	0.57



(a) ROC

(b) *Important* features

Figure 29. SVM model ROC and *important* features affecting the target Class 1 are shown

To ensure high performance concerning Class 1, we have identified *important* features by considering the top 10 features based on their score of absolute coefficients as shown in Figure 29b. In predicting severe occupational accidents for *elementary occupations* in the *construction* sector, we have identified the most important features as material agents associated with the specific physical activity, age of the victim, and various contact modes of injury. The material agents, which include materials, objects, products, machines, or vehicle components, play a crucial role in determining the severity of accidents. Ensuring that all workers receive adequate training on the safe handling, operation, and maintenance of materials, objects, products, machines, and vehicle components is vital. This includes emphasizing the importance of following safety protocols and guidelines and implementing a routine inspection and maintenance schedule for the equipment used at the construction sites. This helps to identify potential hazards and prevent accidents

caused by malfunctioning or faulty components. Furthermore, the contact modes of injury, including trapped or crushed incidents, interactions with sharp or rough agents, and impacts with stationary objects, are also important factors in predicting accident severity. These contact modes highlight the interactions between workers and their environment that are more likely to result in severe accidents, highlighting the need for proper personal protective equipment, such as gloves, safety goggles, and high-visibility clothing, to minimize the risk of contact-related injuries. Ensuring adequate training on safe handling, operation, and maintenance of material agents, along with proper personal protective equipment, helps minimize risks and prevent severe accidents in this high-risk sector.

5.3 Analysis of Limitations

We cover detailed discussions and propose different specific models that can handle the challenges in occupational hazard detection. We investigate more closely the limitations next.

- **Dataset Size** : As we previously discussed, we fail to propose a specific model for every combination of the economic sector and occupation class due to fewer data for information related to Class 1 incidents.
- **Location Specificity** : As we focus on workplace accidents in Estonia, our proposed data analytics, models, and observations may not apply directly to other countries. For example, factors unique to each country can significantly influence the accuracy of our models. Therefore, we can not ensure that the proposed model is generic for any location. As a future direction, we intend to explore the available datasets of different countries and propose modifications to the proposed model based on the dataset difference.

5.4 Implementation Details

This section outlines the various tools utilized throughout the research and writing process of this research. These tools were employed to streamline the workflow, enhance the readability and quality of the text, and assist in data analysis and model training. The tools were grouped into writing assistance, programming, and hardware.

Writing Assistance Tools : Two primary writing assistance tools were employed to improve the readability and quality of the text:

- **ChatGPT** was used to provide overviews of various sources and improve the readability of the sentences.

- **Grammarly** was used to ensure the text's grammatical accuracy, enhance its readability, and maintain a consistent tone throughout the thesis.

Programming Tools : Various programming tools and libraries were employed to facilitate coding, data analysis, and model training.

VSCode was the primary code editor for this research, providing a versatile and efficient platform for writing, debugging, and executing Python scripts. The following Python libraries were crucial to the data analysis and ML aspects of this research:

- Pandas: For efficient data manipulation and analysis.
- Scikit-learn: For ML algorithms and model evaluation.
- NumPy: For numerical computing and array manipulation.
- SciPy: For scientific computing and advanced mathematical operations.
- Matplotlib and Seaborn: For data visualization and plotting.
- Imbalanced-learn: For dealing with imbalanced datasets and providing resampling techniques.

Hardware : A Dell XPS $13\ 9370$ laptop with an Intel $\mathrm{i7-8550U}$ processor was used as the primary computing resource for training the ML models. This device provided sufficient processing power and moderate efficiency for handling the computational demands of the project. Generic model training times generally ranged between 650 to 750 minutes.

6 Conclusions and Future Works

In conclusion, the Thesis demonstrated the performance of various ML models in predicting accident severity for different occupational classes in specific economic sectors. Different feature sets and resampling techniques were utilized to identify the best-performing models and *important* features affecting accident severity. The custom feature selection approach proved more effective than training on *statistically significant* features based on correlations, allowing for the development of more accurate and targeted models to help organizations better understand and mitigate the factors contributing to severe accidents for occupations in their respective sectors. By adopting such a methodology, companies can make more informed decisions to improve worker safety and reduce severe accidents. Additionally, the observations highlighted the differences among the important factors that led to severe accidents in different sectors and occupations.

Valuable insights into the factors contributing to severe accidents were provided and guided the development of targeted prevention strategies to improve workplace safety.

As a future direction, we intend to include more features and explore deep learning techniques to increase prediction accuracy and extend to different industries and countries. Additionally, we intend to explore post-accident based features, such as *injury type* and *injured body part*, to develop injury-specific prevention strategies. We will develop a tool that utilizes the most effective models to minimize severe accidents. This can be achieved by integrating the results with access control or workforce management systems, similar to the one proposed by Choi *et al.* [5]. By using this tool, safety managers can predict potential accident risks and focus on areas that require additional safety measures.

References

[1] Haseeb Ali, Mohd Salleh, Kashif Hussain, Ayaz Ullah, Arshad Ahmad, and Rashid Naseem. A review on data preprocessing methods for class imbalance problem. *International Journal of Engineering & Technology*, 8(3):390–397, 10 2019.

[2] Nitesh V. Chawla. *Data Mining for Imbalanced Datasets: An Overview*, pages 875–886. Springer US, Boston, MA, 2010.

[3] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. pages 785–794, 08 2016.

[5] Jongko Choi, Bonsung Gu, Sangyoon Chin, and Jong-Seok Lee. Machine learning predictive model based on national data for fatal accidents of construction workers. *Automation in Construction*, 110:102974, 2020.

[6] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[7] DR Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242, 1958.

[8] Fatemeh Davoudi Kakhki, Steven A. Freeman, and Gretchen A. Mosher. Evaluating machine learning performance in predicting injury severity in agribusiness industries. *Safety Science*, 117:257–262, 2019.

[9] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, page 504. Springer Series in Statistics. Springer, 2 edition, 2009.

[10] Tin Kam Ho. Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1:278–282, 1995.

[11] Richard A. Johnson and Gouri K. Bhattacharyya. *Statistics: Principles and Methods*, page 55. John Wiley & Sons, 7 edition, 2014.

[12] Richard A. Johnson and Gouri K. Bhattacharyya. *Statistics: Principles and Methods*, page 310. John Wiley & Sons, 7 edition, 2014.

[13] Manjit Kaur, Hemant Kumar Gianey, Dilbag Singh, and Munish Sabharwal. Multi-objective differential evolution based random forest for e-health applications. *Modern Physics Letters B*, 33(05):1950022, 2019.

[14] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[15] John D. Kelleher, Brian Mac Namee, and Aoife D'Arcy. *Applied Predictive Modeling*, page 52. Springer, 2013.

[16] John D. Kelleher, Brian Mac Namee, and Aoife D'Arcy. *Applied Predictive Modeling*, page 56. Springer, 2013.

[17] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[18] Kerim Koc, Ömer Ekmekcioğlu, and Asli Gurgun. Prediction of construction accident outcomes based on an imbalanced dataset through integrated resampling techniques and machine learning methods. *Engineering, Construction and Architectural Management*, 06 2022.

[19] Gordon S. Linoff and Michael J. A. Berry. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, page 132. Wiley, Hoboken, NJ, 3 edition, 2011.

[20] Ahmed O. Oyedele, Anuoluwapo O. Ajayi, and Lukumon O. Oyedele. Machine learning predictions for lost time injuries in power transmission and distribution projects. *Machine Learning with Applications*, 6:100158, 2021.

[21] Sebastian Raschka and Vahid Mirjalili. *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow*, page 117. Packt Publishing, 2 edition, 2017.

[22] Sebastian Raschka and Vahid Mirjalili. *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow*, page 214. Packt Publishing, 2 edition, 2017.

[23] Sebastian Raschka and Vahid Mirjalili. *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow*, page 216. Packt Publishing, 2 edition, 2017.

[24] Sobhan Sarkar, Anima Pramanik, J Maiti, and Genserik Reniers. Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data. *Safety science*, 125:104616, 2020.

[25] Sobhan Sarkar, Sammangi Vinay, Rahul Raj, Jhareswar Maiti, and Pabitra Mitra. Application of optimized machine learning techniques for prediction of occupational accidents. *Computers & Operations Research*, 106:210–224, 2019.

[26] Ako Sauga. *Statistika*, page 210. Varrak, 2017.

[27] Jukka Takala, Päivi Hämäläinen, Kaija Leena Saarela, Loke Yoke Yun, Kathiresan Manickam, Tan Wee Jin, Peggy Heng, Caleb Tjong, Lim Guan Kheng, Samuel Lim, and Gan Siok Lin. Global estimates of the burden of injury and illness at work in 2012. *Journal of Occupational and Environmental Hygiene*, 11(5):326–337, 2014. PMID: 24219404.

[28] Jinjun Tang, Lanlan Zheng, Chunyang Han, Weiqi Yin, Yue Zhang, Yajie Zou, and Helai Huang. Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review. *Analytic Methods in Accident Research*, 27:100123, 2020.

[29] Juliana Tolles and William J. Meurer. Logistic Regression: Relating Patient Characteristics to Outcomes. *JAMA*, 316(5):533–534, August 2016.

[30] Rongchen Zhu, Xiaofeng Hu, Jiaqi Hou, and Xin Li. Application of machine learning techniques for predicting the consequences of construction accidents in china. *Process Safety and Environmental Protection*, 145:293–302, 2021.

Appendix

I. Access to the Code and Source Data

- Source code and reports available at: `bit.ly/3HDAhJq`
- Source data available at: `bit.ly/3nyHF1J`

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Mario Käära**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Application of Machine Learning Techniques to Ensure Safer Work Environments in Estonia,

   supervised by Roshni Chakraborty.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe on other persons' intellectual property rights or rights arising from the personal data protection legislation.

Mario Käära
09/05/2023