Leveraging Advanced Machine Learning Techniques to Predict High-Risk Workplace Incidents: Insights from Ireland

MSc Research Project
Data Analytics

Aishwarya Rani Gopal
Student ID: 21226105

School of Computing
National College of Ireland

Supervisor: Mayank Jain

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Aishwarya Rani Gopal
Student ID:	21226105
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Mayank Jain
Submission Due Date:	14/12/2023
Project Title:	Leveraging Advanced Machine Learning Techniques to Predict High-Risk Workplace Incidents: Insights from Ireland
Word Count:	7100
Page Count:	23

I hereby certify that the information contained in this (masters research project submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Aishwarya Rani Gopal
Date:	29th January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☑
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☑
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☑

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Leveraging Advanced Machine Learning Techniques to Predict High-Risk Workplace Incidents: Insights from Ireland

Aishwarya Rani Gopal
21226105

Abstract

This study looks into the important problem of safety at work, especially in Ireland's industries where the risk of death and serious injuries is still a concern. The study shows a new way to look at and predict workplace accidents by combining machine learning and deep learning methods. This is because this field needs more advanced ways to predict accidents. The research was motivated by the limitations of current safety protocols, which often fail to preemptively identify and stop the risks of severe accidents. To deal with this, a comprehensive methodology was used, which included a collection of workplace incidents in Ireland. Techniques like SMOTE (Synthetic Minority Oversampling Technique) and RUS (Random Under Sampling) were used to fix class imbalances in the dataset. Support Vector Machine, AdaBoost, XGBoost, Naive Bayes, and Neural Networks were used to identify the fatality. The results showed that these algorithms, especially XGBoost, are good at predicting high-risk events. This is not only a big improvement over traditional ways of evaluating safety, but it is also a useful tool for making things safer in the real world. This research adds to what is already known by using a more data-driven and predictive method to look at workplace safety. Unfortunately, which was never employed in Ireland. It also used a secondary dataset to benchmark the results from Ireland's data. It shows how machine learning can change the way safety management is done, laying the groundwork for future progress in this important area.

1 Introduction

1.1 Background

In Ireland, workplace fatalities and accidents are a significant concern. They have a huge impact on society, business, and most of all, the lives of workers and their families. The need to identify and address the causes of workplace mortality highlights the critical nature of this research. Safe and healthy workplaces are a priority for Ireland because of the country's status as an EU member. The traditional ways of incident analysis, like manual processing and making biased choices, are inefficient and lead to delayed responses and ineffective safety measures. These accidents are frequent, despite strict safety regulations[1]. The shortcomings identified in research such as Tixier et al. (2016) demonstrates

[1]`https://www.hsa.ie/eng/news_events_media/archive/press_releases_2013_to_2022/press_releases_2022/hsa_confirms_26_work-related_fatalities_in_2022.html`

this inefficiency and emphasize the necessity for better prediction approaches.

This study, based on this dedication and the relatively untapped possibilities of machine learning in the Irish workplace safety, presents a fresh strategy to tackle these difficulties by utilizing advanced machine learning techniques and deep learning models(Matías et al.; 2008). Based on a machine learning-based analysis, it is identified that falls from heights, being struck by moving objects, coming into contact with machinery, and accidents involving vehicles are the most common causes of such incidents in Ireland. Factors such as lack of proper training, inaccurate risk assessments, and weak safety culture within the organizations exacerbate these disasters. According to Johnson and Martinez (2022), improving workplace safety standards and creating effective prevention measures require an understanding of these root causes and trends. The goal is to improve workplace event processing and analysis procedures so that the primary causes can be found more quickly, effectively, and impartially.

1.2 Research Question and Objectives

The central research question of this study is: **"How effectively can advanced machine learning algorithms be employed to predict workplace fatalities in Ireland?"** The objectives of this research are to offer insights into common causes of workplace accidents, identify high-risk environments, and analyze incident rates. This in turn will contribute significantly to the development of more effective safety practices. Also, this research stresses the importance of ethical considerations discussed in Plotnikova et al. (2020), highlighting the need for worker protection and the make a safer working environment.

The novelty of this research lies in its groundbreaking approach to predicting high-risk workplace incidents in Ireland, an area previously underexplored. It introduces a modified Knowledge Discovery in Databases (KDD) process tailored for complex safety data analysis by performing exhaustive data analysis which helped in choosing approriate preprocessing steps. One breakthrough that addresses early categorization issues is the deliberate refinement of models by the combination of comparable categories and hyperparameter tuning. This study stands out for its methodology that effectively balances class imbalances in the dataset using advanced techniques like SMOTE and RUS. Furthermore, it sets a precedent in the field of workplace safety by providing a more data-driven, predictive framework, previously unutilized in Ireland, thereby addressing safety management and paving the way for future advancements in this crucial area.

1.3 Document Structure

The report is divided into multiple important sections. Background information and learning from previous research are given in the section 2. This is followed by describing the methodology, the steps involved in preprocessing the data, and the reasoning behind selecting the machine learning models in section 3. Following that section 4 outlines the architecture and design elements of the machine learning models and its Implementation is reported in 5. The results of model evaluations are explained in section6.

2 Literature Review

2.1 Machine Learning in Construction and Mining Industry Safety

The application of machine learning in high-risk sectors such as construction and mining has gained significant importance, showing an evolution towards data-driven methodologies in the management of worker safety. In their 2023 publication, Doe and Smith (2023) introduces a breakthrough study that centers on utilizing machine learning to predict construction-related injuries. However, the study is constrained by the limited scope and diversity of its dataset. The study's dependence on particular categories of data such as accident reports, worker demographics, and equipment usage statistics may mistakenly ignore other essential elements that have an impact on construction site injuries. Similarly, by employing machine learning for incident detection, Lee and Kim (2023) advances the assessment of the risk of construction-related injuries. Their methodology's applicability is however constrained by its dependence on sizable and varied datasets. Such thorough data collecting is frequently impractical in real-world building circumstances, particularly on smaller or resource-constrained sites. This limitation calls into question not only the generalizability of the model but also reveals a substantial discrepancy between the best possible research settings and the actual availability of data in the field. Zhu et al. (2023) use an alternative approach by concentrating on the specific attributes of fatalities in the construction industry. Their work is excellent due to its thorough examination of fatal instances, which offers an indepth understanding of the most severe consequences of construction accidents.

Significant contributions have been made by Mammadov et al. (2023) and Yedla (2019) in the mining sector. Research by Mammadov et al. (2023) utilizes machine learning techniques to analyze pipeline construction and mining accidents, providing innovative perspectives on these particular domains. Still, the scope of their research is somewhat limited by its focus on a certain industry, which restricts the applicability of their findings to other sectors within the mining or similar industries. Yedla (2019) offers an extensive examination of mining incidents through the application of machine learning methods. The study is notable for effectively combining structured and unstructured data, achieving high predictive accuracy. The innovative use of synthetic data augmentation for data balance and the identification of injury nature as a key predictor underscore the study's comprehensive approach to understanding mining accident trends.

Manjunatha (2023) and Smith and Johnson (2022) offer valuable and specialized analysis in predicting injuries in the mining sector. Their research is essential for emphasizing the distinct problems and causes that pose risks in the mining industry. However, there is a requirement for more extensive and varied datasets to improve the thoroughness and relevance of their research. Tixier et al. (2016) also performed a comparative examination of different machine learning algorithms in forecasting construction injuries. However, the study highlights a deficiency in the actual testing and implementation of the models, indicating the necessity for additional research to authenticate these models in real-life scenarios.

2.2 Analytical and Predictive Techniques in Occupational Safety

Machine learning has broken past conventional barriers in the field of occupational safety, assisting a narrative that unites several industries via predictive analytics. Leading the charge, Ajith et al. (2020) set out to unravel the complex network of risk factors inside the mining sector. Their thorough statistical analysis provides a comprehensive picture of the risks that miners need to deal with. Though their study's objective is rich in regional specifics, it falls short of providing the broad overview required to adequately depict the global mining industry. This restriction strikes an important question regarding the generalizability of their research: can knowledge gained in one region of the world be used in other mining contexts with different operational procedures and safety regulations? Expanding upon the concept of universality, Sarkar and Maiti (2020) explores the wide range of approaches used in machine learning for the investigation of occupational accidents. Their thorough analysis acts as a guide, pointing the path through the intricate web of theoretical models. However, their research does not go into real-world applications. A bridge that spans this gap is necessary because the theoretical ability of machine learning models and their practical effectiveness continue to be separated by a difficult barrier. The investigation of the random gradient boosting algorithm by Shin (2019) is an important development and his research also demonstrates the ability of advanced machine learning to anticipate accidents in advance. The diversity in this dataset where 22,935 of construction accident cases in Korea, only 16,248 were valid, this limits the prediction model's ability. The effectiveness of these advanced methods would be increased by a bigger, more diverse dataset that would present a more distinct, broad image. The investigations by Johnson and Martinez (2022) and Kakhki et al. (2019) into the applicability of machine learning in forecasting injuries within the agriculture industry add significant dimensions to the use of machine learning in occupational safety. They demonstrate the practicality of machine learning in a sector like agriculture, which is not typically associated with advanced technology. These studies highlight both the significant progress and challenges in applying machine learning to workplace safety. They emphasize the need for flexible, globally relevant analytics and point to future research focusing on cross-framework validation and diverse datasets for broader applicability. However, varying safety laws and uncertainties in model applicability across different agricultural contexts remain key hurdles.

2.3 Deep Learning and Advanced Machine Learning Techniques in Safety Prediction

Deep learning and advanced machine learning techniques have offered new possibilities in the attempt to rethink workplace safety. This journey into the intricate world of safety prediction, characterized by ground-breaking research and creative approaches, challenges accepted theories. And also it highlights the complexities of incorporating modern technology into workable safety precautions. The recent advancement of this research is the ground-breaking work on deep bidirectional transformers for language interpretation by Kenton and Toutanova (2019). Understanding complex data patterns is revolutionized by their study, which explores the fields of deep learning. Beyond just the theoretical, their research has far-reaching consequences for how safety reports, incident descriptions, and worker communications are interpreted and analyzed in different industrial settings. For such advanced technology to reach its full potential, it will be difficult to synchronize it with current safety regulations. Utilizing Python's SpaCy, JUGRAN et al. (2021) presents a novel method for automatic text summarization, expanding on the language processing theme. Their important work in identifying important information from large amounts of textual data provides a glimpse into the potential applications of natural language processing (NLP) in safety management. This development may play a key role in breaking down difficult safety reports and worker narratives into understandable insights, opening the door to safer measures that are better informed and based on facts. Plotnikova et al. (2020) presented a thorough examination of the ways in which data mining techniques have changed over time, as part of a methodical effort to map the changing landscape of these approaches. Their methodical assessment of the literature establishes the foundation for comprehending the various uses of data mining in safety prediction. The scope of their research draws attention to the varied applications of data mining in interpreting intricate patterns and trends in workplace accidents, but it also emphasizes the necessity for further targeted studies on the use of these approaches in particular industrial settings. The research of Luo et al. (2023) on the use of machine learning technology to forecast the seriousness of workplace accidents in building collapse episodes represents an important development in this research. Their work exemplifies the complex trade-off between the feasibility of implementation in high-risk contexts and effective algorithmic design. Though it also highlights the challenges associated with using these models in practical contexts, the work is a testament to the potential of machine learning in offering proper insights into accident severity. A unique pattern of research is formed by this body of work, studies on the use of advanced machine learning techniques, studies on data-mining techniques for explaining and predicting workplace accidents, and studies on decision tree approaches for predicting occupational accidents. This shows how prediction models for workplace safety are constantly evolving but there are difficulties along the way. Additional research and investigation are still needed in the areas of complex model integration into current safety frameworks and understanding of various workplace environments.

2.4 Summary and Justification for Research Question

Significant advancements in machine learning for workplace safety have been identified through a comprehensive literature review. However, critical limitations remain apparent, specifically in the areas of industry specificity and advanced technique complexity. The necessity for machine learning models that are more versatile and universally applicable is highlighted by these findings. Present research, although auspicious within its specific fields, frequently fails to possess the contextual flexibility and data diversity essential for extensive implementation across numerous industries. This constraint tampers the applicability of results and frameworks to a broader scope, thereby limiting their efficacy to particular operational contexts. Moreover, the practical integration of advanced machine learning and deep learning methods into established safety protocols is hindered by their complex nature, thereby emphasizing the disparity between theoretical advancements and their practical implementation. This study endeavors to fill these knowledge voids through the development of machine learning models that improve the accuracy of predictions while remaining applicable to a wide range of industrial contexts. The objective is to surpass the constraints imposed by specific industries and develop user-friendly solutions that are technically advanced and suitable for a wide range of work settings.

3 Methodology

This section dives deeper into the detailed research methodology, giving an in-depth and technical explanation of each step taken to successfully implement this research. The primary objective of this research is to predict the high-risk incidents. To achieve this, a modified Knowledge Discovery in Databases (KDD) process was utilized, this is a method known for its effectiveness in uncovering valid, important, and easily interpretable patterns in large and complex data sets. Figure 1 shows the steps involved.

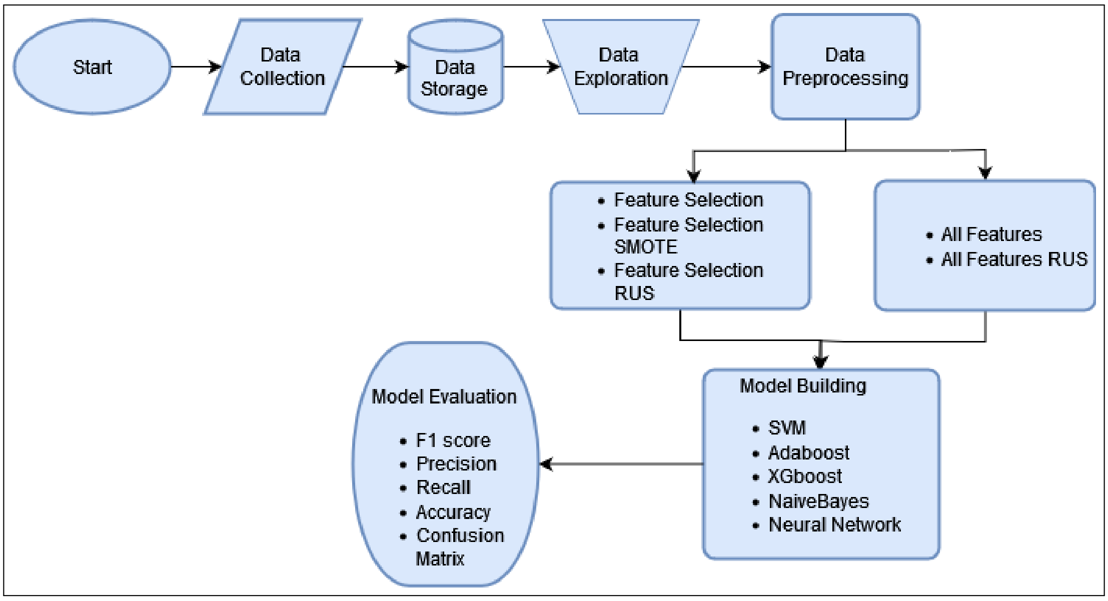

Figure 1: Project Workflow

3.1 Data Collection

The first step in this research methodology is data collection, which is an important phase for understanding the domain-specific challenges. Data is taken from the Irish Workplace Incidents dataset [2] is the primary dataset. This dataset comprises detailed records, including the age and gender of the individuals involved, the specific sectors they were working in, categorized under NACE sectors, and the type of incidents encountered. Additional data points include the employment status of the individuals, the year of the incident, the county of occurrence, and the trigger of the incident. Such comprehensive data allows for an in-depth analysis of the Irish employment landscape. For benchmarking, OSHA dataset is utilized[3], containing global workplace incident data, which includes event descriptions, the nature and part of the body injured, and factors involved in the incidents.

3.2 Exploration of the Data

The Irish dataset contains 45,897 entries and 8 columns, detailing workplace incidents from 2017 to 2021. Key columns include 'Age', 'Gender', 'NACE Sector' (industry sector),

[2]https://data.gov.ie/dataset/workplace-incidents-2017-2021?package_type=dataset
[3]https://www.kaggle.com/datasets/ruqaiyaship/osha-accident-and-injury-data-1517

'Incident Type', 'Employment Status', 'Year', 'County', and 'Trigger' (cause of incident). The majority of data are categorical, except for 'Year', which is numerical. The dataset mainly records non-fatal injuries, with details on the employee's demographic, industry sector, and the nature of the incident.

The Occupational Safety and Health Administration (OSHA) dataset contains 4,847 entries with 29 columns, focusing on workplace incidents between 2015 and 2017. It includes detailed fields such as 'Event Date', 'Abstract Text', 'Event Description', 'Construction End Use', 'Degree of Injury', 'Nature of Injury', 'Part of Body', 'Event type', 'Environmental Factor', 'Human Factor', and 'Task Assigned'. The data primarily comprises categorical variables, with some numerical fields like 'Building Stories' and 'Nature of Injury'. It provides an in-depth look at each incident, including factors like environment, human elements, and the nature of tasks involved similar to the approach taken in Sánchez et al. (2011).

3.3 Data Preparation

3.3.1 Data Preprocessing and Visualization

In the initial phase of the data analysis, the dataset was loaded and Visualization techniques were employed to understand the structure of the data. This process involved creating graphical representations to observe the distribution of various categories within the dataset. Patterns, irregularities, and imbalances were identified. The next step involved cleaning the data by dropping the rows with missing values(Matías et al.; 2008), as the percentage of missing data was very low.

3.3.2 Data Splitting and Transformation

The dataset was divided into training, validation, and testing sets using a stratified approach. Specifically, the data was first split into a training set (70%) and a temporary set (30%). Then, the temporary set was further split equally into validation(50%) and testing sets(50%). This splitting was crucial for training the models, fine-tuning them, and evaluating their performance on test data. Categorical features were then converted into a numerical format using one-hot encoding(Dahouda and Joe; 2021), making them suitable for machine learning algorithms. The variable 'Incident Type' was also encoded, providing numeric labels for the classes that are to be predicted. Feature scaling using Minmax scaling was applied to ensure that all data features were on the same scale, thereby enhancing the performance of various ML models, as illustrated in Figure 2.

3.3.3 Handling Class Imbalance

Class imbalance is a major problem in this research, where some classes have much fewer instances than others. To address this issue, resampling techniques such as the Synthetic Minority Over-sampling Technique (SMOTE)(Chawla et al.; 2002) and Random Under Sampling (RUS)(Japkowicz and Stephen; 2002), were employed. SMOTE was used to generate synthetic samples for the minority class, effectively balancing class distribution. Conversely, RUS reduced the number of examples in the majority class. These techniques made sure that the models did not dominate one class over another during training. Collectively, these steps made the dataset more suitable for training Machine Learning models, providing a more balanced representation of the target classes.

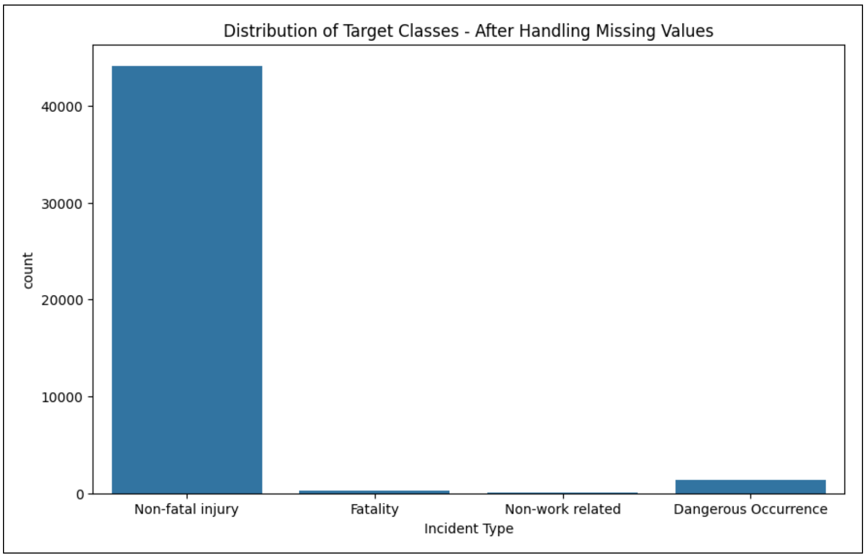

Figure 2: Distribution after Removing Missing values

3.4 Model Building

3.4.1 Support Vector Machine (SVM)

The Support Vector Machine (SVM) excels in high-dimensional space, ideal for this complex dataset(Boser et al.; 1992). Its kernel trick addresses non-linear relationships, crucial for analyzing intricate patterns in fatality data. In this research, SVM model, class weights are automatically adjusted to be inversely proportional to class frequencies in the data, using the 'balanced' mode and evaluated on various dataset versions, including resampled data, ensuring comprehensive and robust predictions aligning with the methodology used in Sánchez et al. (2011).

3.4.2 AdaBoost Classifier

AdaBoost is an ensemble method that enhances weak learners, making it powerful for complex datasets(Sarkar et al.; 2019) like Irish dataset. It sequentially focuses on challenging instances, adapting to nuances in fatality prediction. In this implementation, AdaBoost, algorithm begins with simple decision stumps as its foundational models, each focusing on a single aspect of the data. It then iteratively refines its approach by increasingly focusing on data points that were previously misclassified, thereby enhancing its accuracy with each cycle. Finally, these stumps are combined into a robust ensemble, leveraging their collective strength for superior decision-making and prediction accuracy. This approach was implemented on both the feature selected and resampled datasets, improving the ability to accurately predict fatalities under various data conditions.

3.4.3 XGBoost

XGBoost(Chen and Guestrin; 2016) is renowned for handling large, intricate datasets efficiently, suitable for the multifaceted nature of fatality prediction. Its gradient boosting approach combats overfitting, crucial for maintaining model robustness. In this research, XGBoost is fine-tuned for this imbalanced dataset, and applied to different data versions, including those balanced with SMOTE and Random Under Sampling, showcasing its adaptability and effectiveness in various scenarios.

3.4.4 Gaussian Naive Bayes

Gaussian Naive Bayes(Murphy; 2012), simple and efficient, is well-suited for extensive fatality dataset. Its independence assumption makes it a fast, baseline classifier for high-dimensional data. It is the initial step to grasp fundamental data patterns, especially useful in imbalanced dataset context as suggested in Matías et al. (2008). Its application provides a foundational comparison point for evaluating more complex models in the fatality prediction system.

3.4.5 Neural Network (Multi-Layer Perceptron)

The Multi-Layer Perceptron (MLP), a class of feedforward artificial neural networks (Rumelhart et al.; 1986), is particularly effective in discerning complex patterns within data, a crucial aspect in the challenging task of fatality prediction. Its layered architecture, typically comprising an input layer, multiple hidden layers, and an output layer, is adept at capturing non-linear relationships inherent in feature-rich datasets. In this research, the MLP architecture is carefully configured to align with the specific characteristics of the dataset. It is applied to both the original and balanced datasets, demonstrating its versatility and robustness in delivering accurate predictions across diverse data scenarios, as corroborated by the findings of Lee et al. (2020).

3.5 Evaluation Techniques

Several indicators were used to assess the models:

- Accuracy: This measure gives a broad impression of the model's accuracy rate. It is computed as the proportion of accurately predicted instances to all instances in the dataset(Japkowicz and Shah; 2011). Although accuracy provides a preliminary sense of the model's performance, it may not always be accurate, particularly in imbalanced datasets where it may appear deceptively high despite subpar minority class performance.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Number of Instances}} \tag{1}$$

- Confusion Matrix: An essential tool for comprehending how well the models perform for each class is the confusion matrix. It differentiates between true positives, true negatives, false positives, and false negatives by tabulating the proportion of accurate and inaccurate forecasts(Stehman; 1997). This dissection is especially helpful in assessing how well the model can distinguish between each class in unbalanced datasets.

- Precision and Recall: Precision is the ratio of true positive predictions to all positive predictions, and recall (also known as sensitivity) is the percentage of real positives that the model properly recognized(Davis and Goadrich; 2006). When there are large differences in the costs associated with false positives and false negatives, these measurements become quite important. To guarantee that the fatality prediction models accurately identify the maximum number of genuine instances while minimizing the number of false alarms, precision and recall are crucial.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3}$$

- F1 Score: This single metric provides an evaluation of a model's performance; it is the harmonic mean of precision and recall(van Rijsbergen; 1979). Given that it strikes a compromise between recall and precision, this score is especially helpful with this dataset. Because it accounts for both false positives and false negatives, it is a more dependable metric than accuracy for imbalanced datasets as suggested in He and Garcia (2009).

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

4 Design Specification

The design specification describes the libraries and plan for the machine learning system that predicts fatalities. The tools, methods and algorithms are explained in depth during this phase.

4.1 Tools Used

Python was the primary programming language used in this study, offering a comprehensive library of tools for modeling, analysis, and visualization, thereby facilitating easy programming and interpretation of results. Key libraries employed include Pandas and NumPy for data manipulation and numerical computations, Scikit-learn was used for implementing various machine learning algorithms such as AdaBoost, Support Vector Machines, Random Forests, and Logistic Regression, and for tasks like preprocessing, model selection, and performance metrics evaluation. Imbalanced-learn library was used for addressing imbalanced datasets using techniques like SMOTE (Synthetic Minority Over-sampling Technique) and Random Under Sampling. XGBoost used for a robust and efficient implementation of gradient boosting. Matplotlib and Seaborn used for comprehensive data visualization, providing clear graphical representation of results, TensorFlow and Keras used for the design and training of neural network models, particularly the Multi-Layer Perceptron (MLP) Classifier; and additional utilities such as Joblib for model serialization and statistical functions from SciPy.

4.2 Modelling Specification

Based on the methodology, data preprocessing was conducted using Python libraries such as Pandas and NumPy. Post one-hot encoding, numerical variables were normalized using the MinMaxScaler from scikit-learn, ensuring uniformity in data scale.

During the modeling phase, the SVM model from scikit-learn utilized a radial basis function (RBF) kernel. This choice was motivated by the kernel's ability to handle non-linear data, a common characteristic in complex datasets(Han et al.; 2012). The AdaBoost model, also implemented using scikit-learn, was combined with Decision Trees, thus enhancing their performance by focusing on more challenging instances. The MLP, comprised of multiple layers and neurons, was implemented using scikit-learn as well. It employed a ReLU activation function and the Adam optimizer, which are suitable for capturing deep, non-linear relationships in the data. Additionally, a Sequential deep learning model was constructed using TensorFlow and Keras. This model included multiple Dense layers with Dropout for regularization and also utilized the ReLU activation function and the Adam optimizer. Such an architecture was specifically designed to extract complex patterns and relationships within the data, potentially too intricate for more traditional models(Pomerat et al.; 2019).

5 Implementation

5.1 Exploratory Data Analysis

The results of Ireland's Exploratory Data Analysis (EDA) of workplace incidents have been noteworthy, showing the incidents distribution across a variety of industries and demographic groups. Figure3 shows the yearly incident counts, indicating that the number of incidents fluctuates over the years. The distribution suggests that certain factors, potentially economic cycles or regulatory changes, could have impacted the frequency of workplace incidents. The NACE Sector Distribution of Incidents (Figure 4) provides insights into the frequency of incidents within specific sectors. Human health and social work activities, for instance, show a high number of incidents, pinpointing sectors where safety improvements are most needed.

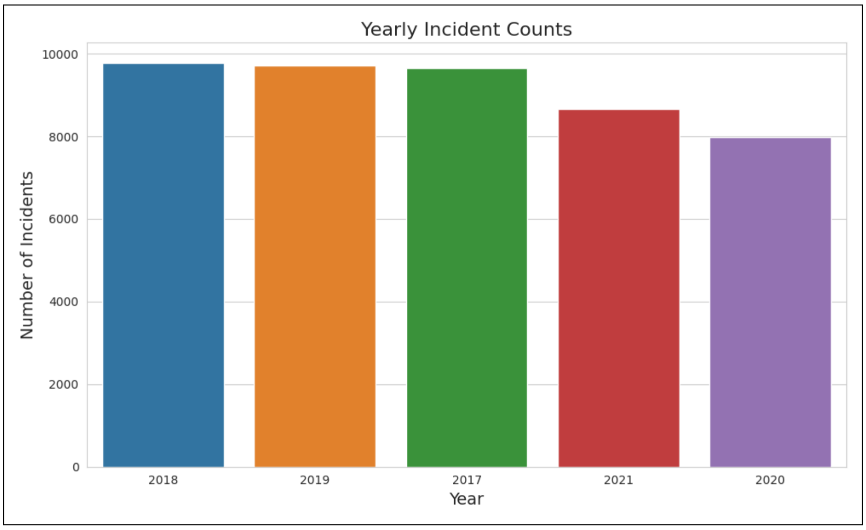

Figure 3: Yearly Incident count

The Correlation Matrix using Cramér's V (Figure 5) explores the strength of the association between categorical variables. This matrix is foundational for predictive modeling, highlighting significant correlations that could predict incident likelihood. The Random

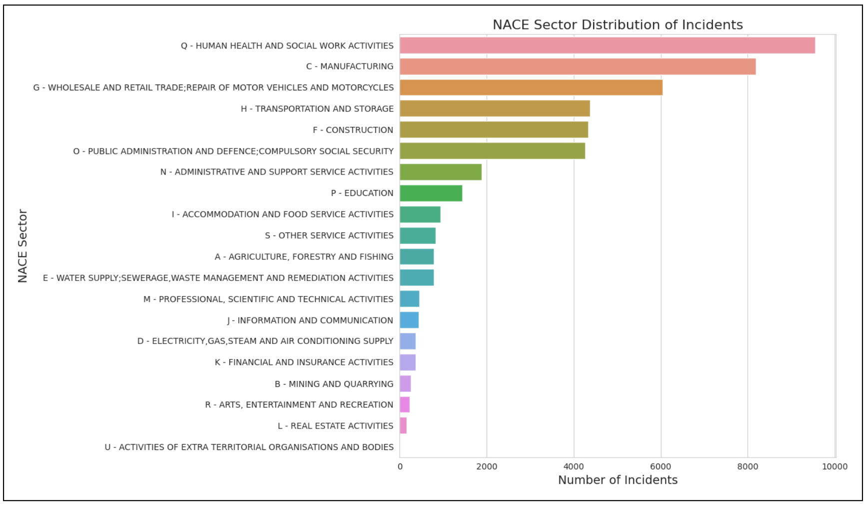

Figure 4: Sector Distribution of Incidents

Forest Classifier identified gender, age, employment status, and specific sectors like human health as key predictors of workplace incidents, informing intervention priorities.

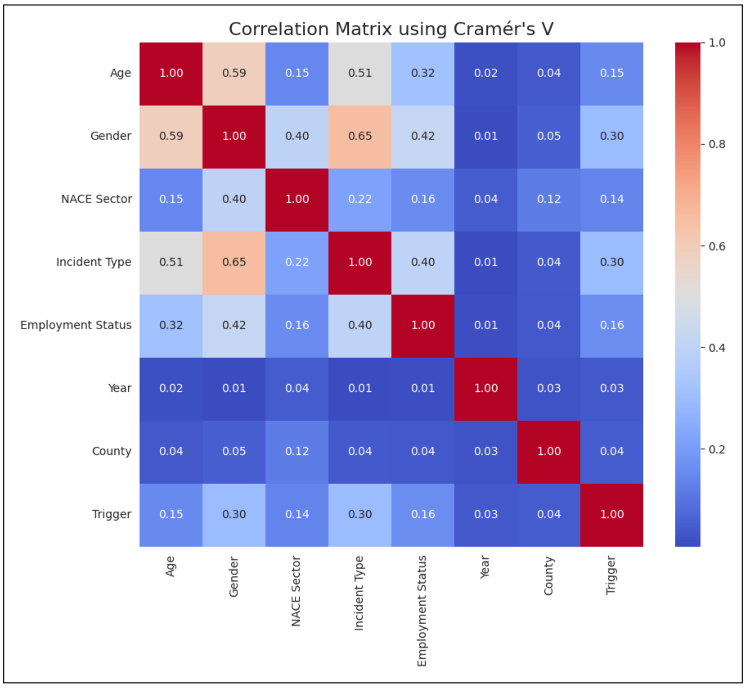

Figure 5: Correlation matrix heatmap between catogorical variable

Additional EDA approaches such as the generation of word clouds to identify common triggers of incidents, indicate that ‘slip’, ‘fall’, and ‘level’ are frequently occurring terms shown in Figure 6. This suggests that preventive measures in workplace safety could focus on slip and fall hazards. Moreover, the analysis extended to heatmaps, which offered a detailed view of the incidence rates by sector and trigger, as well as by gender and age group, illustrating the distribution and prevalence of incidents across different demographics and operational categories. Overall, these figures and analyses provided a detailed view of workplace incidents, contributing to a more comprehensive understanding necessary for developing effective safety policies and practices.

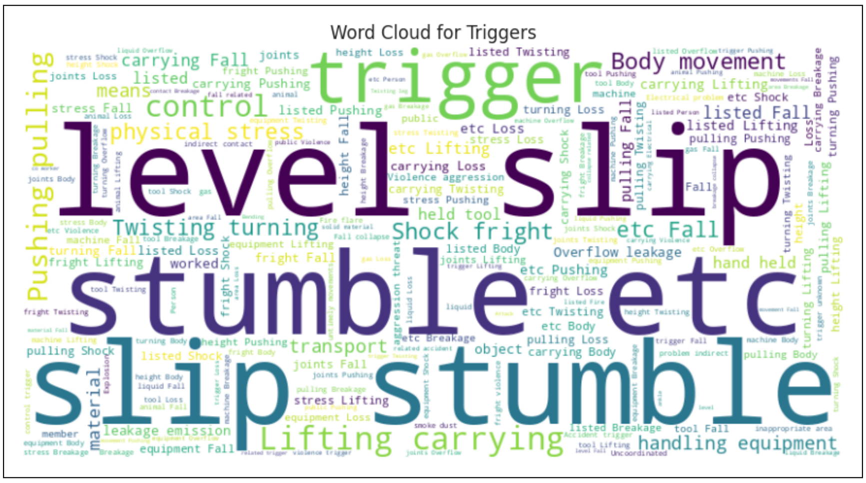

Figure 6: Word cloud for Incident triggers

5.2 Data Preparation and Model Optimization

Commencing with data integrity checks, the data cleaning phase identified 125 missing entries, reinforcing the dataset's reliability. Following one-hot encoding of categorical variables resulted in an expanded feature space, ultimately creating 85 distinct features that encapsulated the dataset's multi-faceted nature.

A feature selection process using a Decision Tree Classifier was used to assess the importance and correlation of each feature against the 'Incident Type' target variable. This rigorous approach distilled the feature set from 85 down to a more manageable and potent 20 features, chosen for their significant impact on the prediction outcomes. This reduction was crucial for model efficiency and interpretability, ensuring that only the most relevant predictors were retained for analysis. Minmax scaling was then implemented to normalize all features to have a consistent range between 0 and 1, inorder to ensure that features with high magnitude dont dominate.

In addressing the skewed class distribution, Random Under Sampling was employed to even the class representation to 50 instances each, effectively mitigating majority class bias. Alternatively, the Synthetic Minority Over-sampling Technique (SMOTE) was also applied, increasing each minority class to 30,000 instances to match the majority class's presence. These resampling techniques were instrumental in achieving a balanced class distribution, which was crucial for unbiased model performance and robust generalization to new data as shown in 7.

Hyperparameter tuning of the SVM classifier was conducted with precision, utilizing a RandomizedSearchCV framework to systematically explore 100 parameter configurations over 5-fold cross-validation. The search parameters included a range of 'C' values [0.1, 1, 10, 100], 'gamma' values [1, 0.1, 0.01, 0.001], and 'kernel' types ['linear', 'rbf', 'sigmoid']. The optimal parameters emerged as 'C': 10, 'gamma': 0.01, and 'kernel': 'rbf', which yielded the highest cross-validation accuracy.

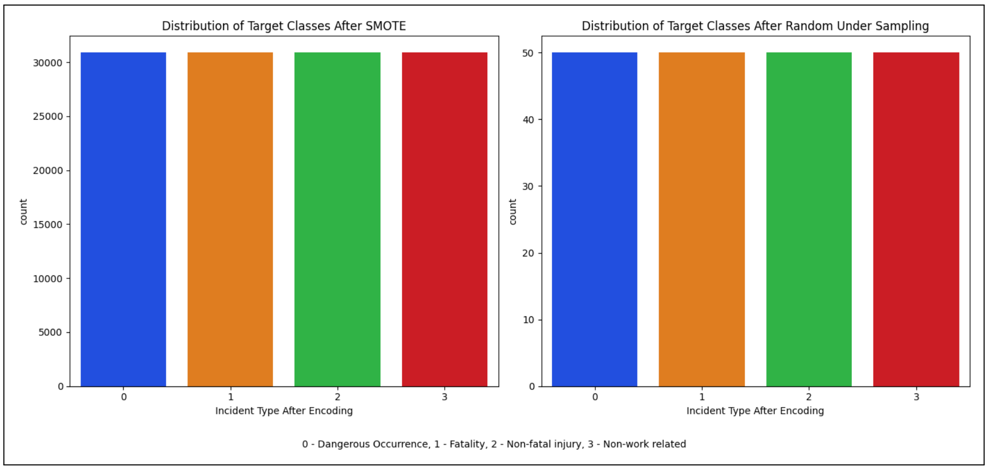

Figure 7: Distribution after Class Balancing

6 Evaluation

In this critical phase of the machine learning research, the effectiveness of the models in predicting workplace fatalities was assessed.

6.1 Experiment 1 - Fatality Prediction with All Features

In the first experiment, all features were used to train each model. SVM model displayed a remarkable recall by correctly identifying all instances of the fatality class. However, it did so with a large number of false positives, as all non-fatality instances were incorrectly classified as fatalities, possibly caused by an imbalanced emphasis on the minority class. The XGBoost model, while not capturing all fatalities (with a recall of 0.41 for the fatality class), achieved a precision of 1, indicating no false positives among its fatality predictions. This led to an f1-score of 0.58 for the fatality predictions by XGBoost, suggesting a balanced performance between precision and recall, though with some fatalities going undetected. While the high precision is commendable, the missed fatality cases reflect a conservative model that may be more suitable in scenarios where false alarms are costly. The heatmap visualization of the confusion matrix for the XGBoost model shown in Figure 8 shows performance across all classes. The results of the model shown in Table 1.

XGBoost	
Testing accuracy	0.99
Precision	1.00
Recall	0.41
F1 score	0.58

Table 1: Results of XGBoost

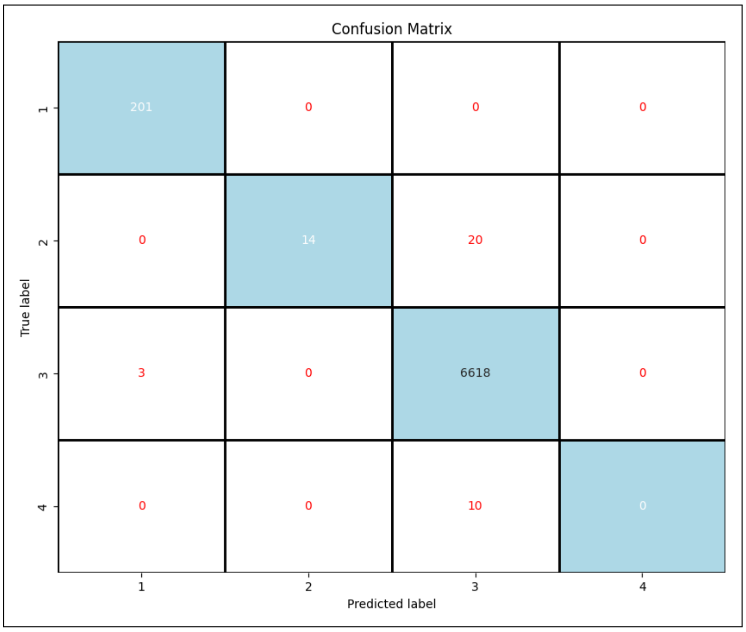

Figure 8: All Feature - XGBoost

6.2 Experiment 2 - Fatality Prediction with All Features and RUS

In the second experiment of the study, after addressing the class imbalance through Random Under Sampling (RUS), the machine learning models were re-evaluated. The SVM model, with RUS-applied features, achieved an accuracy of 66.21% and demonstrated an improved recall for the fatality class (class 1) at 0.62, indicating it could identify a significant portion of the actual fatalities. However, its precision of 0.14 for the fatality class suggested a relatively high number of false positives. The AdaBoost model showed a low overall accuracy of 12.28% and a recall of 0.59 for fatalities, similar to SVM, but with very low precision. The XGBoost model displayed better overall accuracy at 79.26% and the highest recall for fatalities at 0.71, but with low precision of 0.04. This could be due to a loss of vital information when the majority class instances are under-sampled, leading to poorer model generalization. Naive Bayes achieved the highest accuracy among the models at 83.45%, with a recall equal to XGBoost for the fatality class. The Neural Network, however, completely failed to identify any fatalities, with near-zero accuracy and no correct fatality predictions. A heatmap for the XGboost model from Figure 9 gives a visual insight into its performance, especially its capability to detect fatalities amidst the dataset balanced with RUS. The results of the model shown in Table 2.

XGBoost	
Testing accuracy	0.79
Precision	0.04
Recall	0.71
F1 score	0.07

Table 2: Results of XGBoost

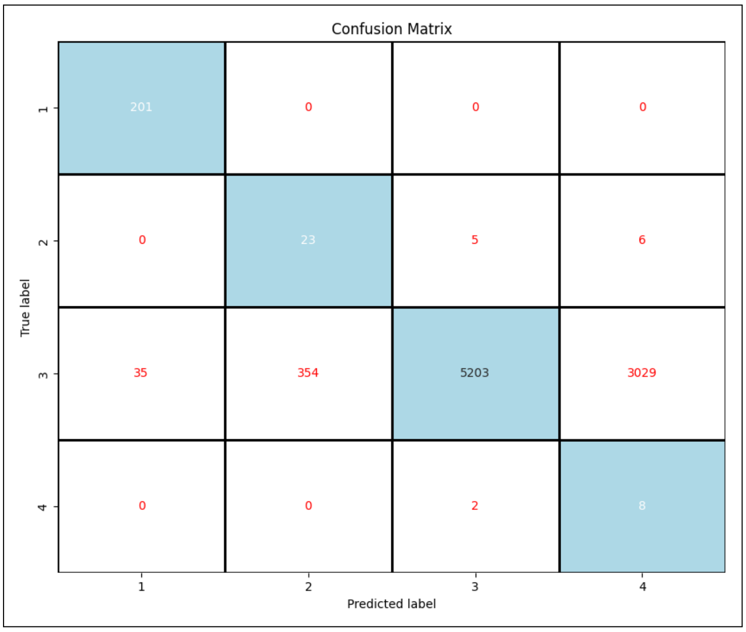

Figure 9: All Feature RUS - XGBoost

6.3 Experiment 3 - Fatality Prediction with Selected Features

In Experiment 3, with the feature selection applied to the dataset, the machine learning models showed varied performance. The SVM model achieved perfect recall in identifying fatalities, which means it detected all fatality cases in the dataset. However, its overall accuracy drop indicates a misclassification of non-fatality instances, suggesting that the selected features may align too closely with the fatality class at the expense of others. AdaBoost and Naive Bayes demonstrated moderate to low accuracy, with AdaBoost achieving a recall of 0.76 for the fatality class and Naive Bayes reaching a recall of 0.97. This suggests that while they were able to detect most fatalities, they struggled with correctly classifying non-fatality instances. XGBoost emerged with high accuracy (99.52%) and maintained a strong precision for the fatality class. Nevertheless, its recall for fatalities was 0.41, indicating that crucial features for detecting fatalities might have been omitted during feature selection, or that the model requires a more comprehensive set of features to improve fatality detection. The Neural Network, despite its high overall accuracy, failed to identify any fatalities, indicating potential issues with the feature selection process. The results from this experiment for the model XGBoost as shown in Table 3 indicate that feature selection can impact the models' ability to predict fatalities, with some models like XGBoost maintaining high accuracy but not necessarily improving in fatality detection. A heatmap for the confusion matrix of the XGBoost model as shown in Figure 10 provides a visual interpretation of the results.

XGBoost	
Testing accuracy	0.99
Precision	1
Recall	0.51
F1 score	0.58

Table 3: Results of XGBoost

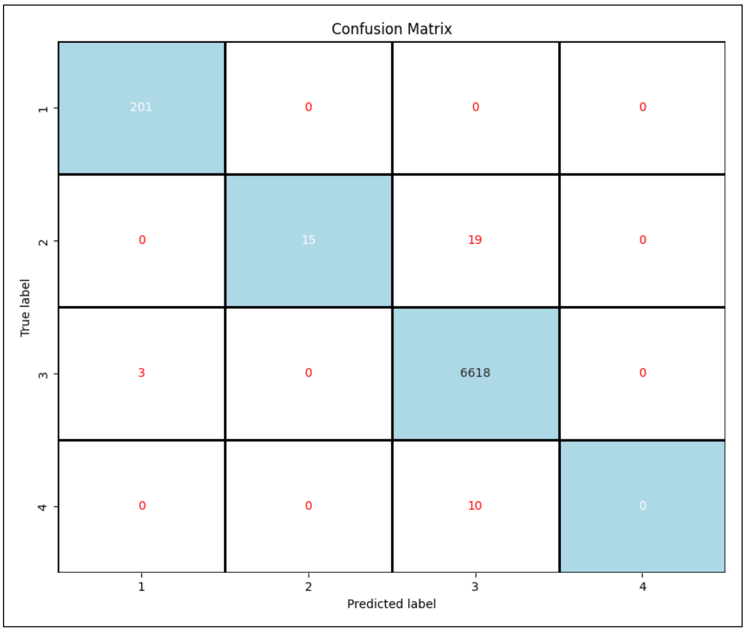

Figure 10: Feature Selected - XGBoost

6.4 Experiment 4 - Fatality Prediction with Selected Features and SMOTE

In Experiment 4, models were applied to a dataset balanced using the Synthetic Minority Over-sampling Technique (SMOTE) to predict the crucial fatality class. The SVM model showed a significant improvement with a high accuracy of 95.75%, successfully identifying most of the fatalities, as indicated by the recall of 0.47. However, its precision for the fatality class was low, suggesting that while it could identify fatalities, it also incorrectly labeled many non-fatalities as fatalities. AdaBoost struggled in this setup, with an overall accuracy of just 11.08% and a high recall but low precision for the fatality class, meaning it too was prone to false positives in fatality prediction. XGBoost continued to perform well with an accuracy of 95.12%, demonstrating an ability to discern fatalities with a recall of 0.59. However, similar to SVM and AdaBoost, it faced challenges with precision in predicting the fatality class. Naive Bayes had a very low accuracy of 8.94%, but it achieved a perfect recall score for the fatality class, identifying all fatalities correctly. However, its precision was extremely low, which again indicates a high number of false positives in turn indicating a need for more refined oversampling techniques or alternative approaches to balance the dataset. The Neural Network model yielded an accuracy of 96.15%, with a recall of 0.50 for the fatality class, showing a balanced capability to identify fatalities. The heatmap of the confusion matrix for these models, particularly NaiveBayes, will provide insights into how effectively each model distinguished the fatality class from other incident types after the application of SMOTE is shown in Figure 11. The precision-recall curves for each model also illustrated the balance between sensitivity and positive predictive value, with the AUC measure providing a summary of this balance over all possible classification thresholds. The results of this experiment is shown in Table 4.

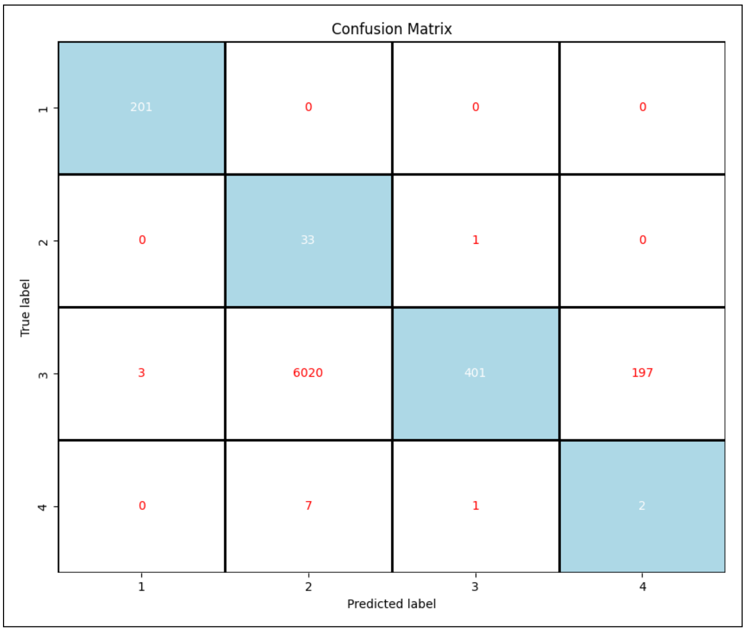

Figure 11: Feature Selected SMOTE - NaiveBayes

NaiveBayes	
Testing accuracy	0.08
Precision	0.01
Recall	1
F1 score	0.01

Table 4: Results of NaiveBayes

6.5 Experiment 5 - Fatality Prediction with Selected Features and RUS

In Experiment 5, the study employed Random Under Sampling (RUS) to balance the feature set and applied different models to predict fatality instances. The Naive Bayes model stood out with the highest accuracy among all the models at 92.02%. It demonstrated a balanced capability in identifying fatalities, with a precision of 0.06 and a recall of 0.76 for the fatality class, indicating that it could recognize most of the actual fatalities, albeit with some false positives. The Support Vector Machine (SVM) model followed closely with an accuracy of 84.36%, showing an impressive recall of 0.65 for the fatality class, suggesting that it could identify a majority of the actual fatality instances. However, its precision was lower, indicating that it also misclassified some non-fatal incidents as fatal. XGBoost showed an accuracy of 81.31% and displayed a strong recall of 0.59 for the fatality class, but similar to SVM, it faced a precision trade-off, leading to several false positives. The AdaBoost model, with an accuracy of 78.11%, also achieved a reasonable recall of 0.53 for the fatality class but faced challenges with precision. The Neural Network model presented an accuracy of 78.07% and a recall of 0.68 for the fatality class, indicating a moderate level of sensitivity towards identifying fatalities. Overall, the application of RUS impacted the precision of the model adversely while improving or maintaining their recall for the fatality class, as indicated in the result of NaiveBayes in Table 5. The heatmap of the confusion matrix from Figure 12 visually depicts the distribution of predictions across all classes, illustrating each model's ability to identify fatalities post-RUS application.

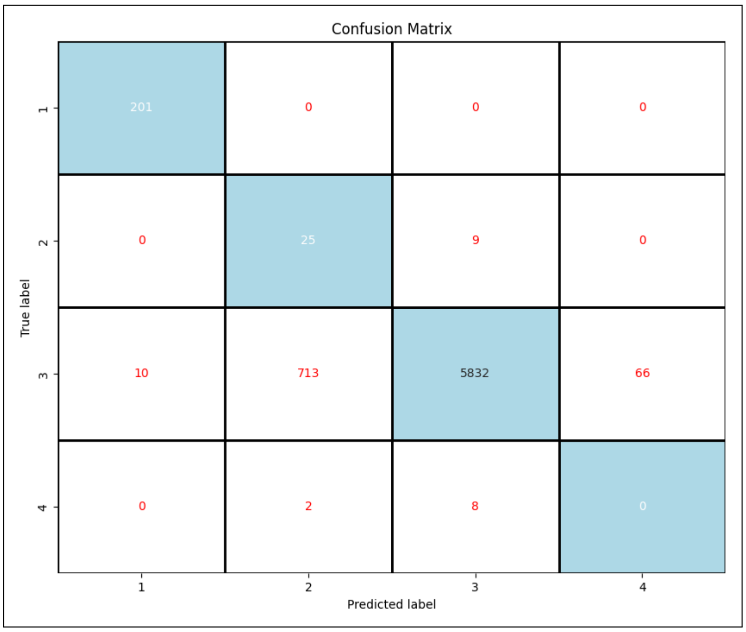

Figure 12: Feature Selected RUS - NaiveBayes

NaiveBayes	
Testing accuracy	0.92
Precision	0.06
Recall	0.76
F1 score	0.10

Table 5: Results of NaiveBayes

6.6 Experiment 6 - Benchmarking the Approach with OSHA Dataset

The OSHA dataset was utilized as part of a benchmarking exercise to validate the effectiveness of the applied machine learning methodology across different datasets. This approach ensures that the predictive models are robust and not tailored to the idiosyncrasies of a single dataset, thereby demonstrating their general applicability and reliability in varied contexts. The results were particularly impressive with certain models, notably the XGBoost and Naive Bayes classifiers, both achieving perfect accuracy scores when the dataset was balanced using SMOTE and RUS techniques. This indicates that with appropriate preprocessing, these models can be highly effective in classifying workplace incidents, thereby supporting their deployment in real-world settings. The SVM model also showed significant improvement when SMOTE was applied, suggesting that the balancing of classes played a crucial role in enhancing its predictive accuracy. Meanwhile, the AdaBoost classifier performed exceptionally well with SMOTE but experienced a slight dip with RUS, pointing to some sensitivity to the method of class balancing. The Neural Network's performance varied, with a high accuracy when SMOTE was applied but a reduced effectiveness with RUS. The results are shown in Table 6.

Model	Data	Accuracy
SVM	Feature Selected	0.03
AdaBoost	Feature Selected	0.95
XGBoost	Feature Selected	1.00
Naive Bayes	Feature Selected	0.67
Neural Network	Feature Selected	0.24
SVM	SMOTE	0.99
AdaBoost	SMOTE	0.67
XGBoost	SMOTE	1.00
Naive Bayes	SMOTE	1.00
Neural Network	SMOTE	0.99
SVM	RUS	0.64
AdaBoost	RUS	0.99
XGBoost	RUS	0.99
Naive Bayes	RUS	1.00
Neural Network	RUS	0.88

Table 6: Model accuracy with feature selection and class balancing techniques.

6.7 Discussion

This research has embarked upon a meticulous examination of machine learning and deep learning methodologies to predict workplace incidents in Ireland. Initially confronted with the challenge of class imbalance, a variety of strategies were implemented, including oversampling, class weight adjustments, and feature standardization. However, these preliminary measures yielded suboptimal outcomes, prompting an in-depth analysis of confusion matrices. This analysis revealed a pattern of misclassification among similar categories, which was suggested to be a consequence of inaccuracies within the self-reported dataset.

In response to these findings, the approach was refined by merging analogous categories to enhance classification accuracy. This modification, accompanied by rigorous feature selection and hyperparameter optimization, significantly improved model performance. In the initial experiment, the use of all features to train the models revealed the SVM's high sensitivity in detecting fatalities, eventhough at the cost of a significant number of false positives. This result raised questions about the trade-off between recall and precision in the models used, prompting a deeper examination of the balance between these metrics. Amongst the models tested, XGBoost distinguished itself, demonstrating accuracy rates above 91% after feature selection. Additionally, the application of SMOTE and precise hyperparameter tuning proved to be effective in addressing the issues of class imbalance and overfitting, thereby refining the predictive precision of the models.

The research proceeded with the understanding that the dataset might include inaccuracies, an assumption substantiated by the pattern of misclassification encountered during initial experiments. This premise influenced the strategy adopted in subsequent analyses and contributed to the development of a robust predictive framework capable of discerning the severity of injuries with notable accuracy.

7 Conclusion and Future Work

This research has advanced the use of machine learning in predicting workplace incidents, particularly highlighting the effectiveness of SMOTE and RUS data balancing techniques in improving model accuracy. Future work aims to broaden the dataset scope, enhancing representation and robustness, and exploring alternative data balancing strategies. Future work will incorporate domain-specific knowledge with an emphasis on practical application to close the gap between statistical precision and real-world applicability. Workplace safety measures could be greatly impacted by collaborative activities with industry professionals to assure the validation and practical application of these models in the real world. Additionally, scaling up and improving these models for broader use will require utilizing technology developments in big data analytics and cloud computing. A key area of future exploration is explainable AI (XAI), which promises to enhance model interpretability and trustworthiness, thereby making the findings more accessible and actionable for occupational health and safety professionals. This study sets a solid foundation for the practical application and commercialization of machine learning techniques in the field of workplace safety, contributing to the development of safer working environments across various industries.

Acknowledgement

I would like to thank my parents for supporting my studies and my supervisor Mayank Jain for his invaluable guidance and support throughout this research project. His insights were crucial in shaping the direction of the work.

References

Ajith, M. M., Ghosh, A. K. and Jansz, J. (2020). Risk factors for the number of sustained injuries in artisanal and small-scale mining operation, *Safety and health at work* **11**(1): 50–60.

Boser, B. E., Guyon, I. M. and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers, *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique, *Journal of artificial intelligence research* **16**: 321–357.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system, *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Dahouda, M. K. and Joe, I. (2021). A deep-learned embedding technique for categorical features encoding, *IEEE Access* **9**: 114381–114391.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves, *Proceedings of the 23rd international conference on Machine learning*, ACM, pp. 233–240.

Doe, J. and Smith, J. (2023). Application of machine learning to construction injury prediction, *Construction Safety Journal* .

Han, S., Qubo, C. and Meng, H. (2012). Parameter selection in svm with rbf kernel function, *World Automation Congress 2012*, IEEE, pp. 1–4.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data, *IEEE Transactions on knowledge and data engineering* **21**(9): 1263–1284.

Japkowicz, N. and Shah, M. (2011). *Evaluating learning algorithms: a classification perspective*, Cambridge University Press.

Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: A systematic study, *Intelligent data analysis* **6**(5): 429–449.

Johnson, E. and Martinez, C. (2022). Evaluating machine learning performance in predicting injury severity in agribusiness industries, *Journal of Agribusiness Safety and Health* **15**(2): 78–92.

JUGRAN, S., KUMAR, A., TYAGI, B. S. and ANAND, V. (2021). Extractive automatic text summarization using spacy in python & nlp, *2021 International conference on advance computing and innovative technologies in engineering (ICACITE)*, IEEE, pp. 582–585.

Kakhki, F. D., Freeman, S. A. and Mosher, G. A. (2019). Evaluating machine learning performance in predicting injury severity in agribusiness industries, *Safety science* **117**: 257–262.

Kenton, J. D. M.-W. C. and Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of naacL-HLT*, Vol. 1, p. 2.

Lee, D. and Kim, S. (2023). Application of machine learning technology for occupational accident severity prediction in the case of construction collapse accidents, *Construction Engineering and Management* **42**(4): 234–248.

Lee, J. Y., Yoon, Y. G., Oh, T. K., Park, S. and Ryu, S. I. (2020). A study on data pre-processing and accident prediction modelling for occupational accident analysis in the construction industry, *Applied Sciences* **10**(21): 7949.

Luo, X., Li, X., Goh, Y. M., Song, X. and Liu, Q. (2023). Application of machine learning technology for occupational accident severity prediction in the case of construction collapse accidents, *Safety science* **163**: 106138.

Mammadov, A., Kazar, G., Koc, K. and Tokdemir, O. B. (2023). Predicting accident outcomes in cross-border pipeline construction projects using machine learning algorithms, *Arabian Journal for Science and Engineering* pp. 1–19.

Manjunatha, A. (2023). *Injury Prediction in Mining Industry through Applied Machine Learning Approaches*, PhD thesis, Dublin, National College of Ireland.

Matías, J. M., Rivas, T., Martín, J. and Taboada, J. (2008). A machine learning methodology for the analysis of workplace accidents, *International Journal of Computer Mathematics* **85**(3-4): 559–578.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*, MIT press.

Plotnikova, V., Dumas, M. and Milani, F. (2020). Adaptations of data mining methodologies: a systematic literature review, *PeerJ Computer Science* **6**: e267.

Pomerat, J., Segev, A. and Datta, R. (2019). On neural network activation functions and optimizers in relation to polynomial regression, *2019 IEEE International Conference on Big Data (Big Data)*, IEEE, pp. 6183–6185.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). Learning representations by back-propagating errors, *nature* **323**(6088): 533–536.

Sánchez, A. S., Fernández, P. R., Lasheras, F. S., de Cos Juez, F. J. and Nieto, P. G. (2011). Prediction of work-related accidents according to working conditions using support vector machines, *Applied Mathematics and Computation* **218**(7): 3539–3552.

Sarkar, S. and Maiti, J. (2020). Machine learning in occupational accident analysis: A review using science mapping approach with citation network analysis, *Safety science* **131**: 104900.

Sarkar, S., Vinay, S., Raj, R., Maiti, J. and Mitra, P. (2019). Application of optimized machine learning techniques for prediction of occupational accidents, *Computers & Operations Research* **106**: 210–224.

Shin, Y. (2019). Application of stochastic gradient boosting approach to early prediction of safety accidents at construction site, *Advances in Civil Engineering* **2019**: 1–9.

Smith, J. and Johnson, M. (2022). Injury prediction in mining industry through applied machine learning approaches, *Journal of Applied Machine Learning* .

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy, *Remote sensing of Environment* **62**(1): 77–89.

Tixier, A. J.-P., Hallowell, M. R., Rajagopalan, B. and Bowman, D. (2016). Application of machine learning to construction injury prediction, *Automation in construction* **69**: 102–114.

van Rijsbergen, C. J. (1979). *Information Retrieval*, Butterworth-Heinemann.

Yedla, A. D. (2019). *Predicting Safety Outcomes in Mining Industry-A Machine Learning Approach*, PhD thesis, Iowa State University.

Zhu, J., Shi, Q., Li, Q., Shou, W., Li, H. and Wu, P. (2023). Developing predictive models of construction fatality characteristics using machine learning, *Safety science* **164**: 106149.