National College of Ireland

Predicting Fatalities: Enhancing Construction Site Safety Through Advanced Machine Learning

MSc Research Project
Master of Science in Data Analytics

Deepak Singh Kirola
Student ID: x22141855

School of Computing
National College of Ireland

Supervisor: Arghir Nicolae Moldovan

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Deepak Singh Kirola
Student ID:	x22141855
Programme:	Master of Science in Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Arghir Nicolae Moldovan
Submission Due Date:	14/12/2022
Project Title:	Predicting Fatalities: Enhancing Construction Site Safety Through Advanced Machine Learning
Word Count:	7751
Page Count:	22

I hereby certify that the information contained in this (my 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:	Deepak Singh Kirola
Date:	30th 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):	

Predicting Fatalities: Enhancing Construction Site Safety Through Advanced Machine Learning

Deepak Singh Kirola
x22141855

Abstract

In the construction industry, ensuring workplace safety is a critical challenge. Jaby Mohammed and Md Jubaer Mahmud, from the Department of Technology at Illinois State University, address this issue by analyzing Occupational Safety and Health Administration (OSHA) Accident and Injury Data from 2015 to 2017. They use machine learning techniques, exploring nine algorithms, including XGBoost and Random Forest, which achieve accuracies of 65.29% and 58.24%, with AUC values of 78.83% and 69.52%, respectively. To improve their model, this research pays attention to specific details in the data and fine-tunes their methods, focusing on precision. After this adjustment, this research achieves an accuracy of 85.7% and an AUC of 92.5% for XGBoost, and an accuracy of 84.9% and an AUC of 91.3% for AdaBoost. This signifies that the methods employed in this research are more effective in predicting outcomes.

The study aims to transform safety management in the construction industry by establishing a data-driven system. By uncovering injury patterns, causal factors, and areas requiring improved safety measures in the unstructured OSHA data, the research contributes significantly to the workplace safety literature. This work envisions a safer future by combining advanced machine learning methods, detailed data analysis, and refined predictive models. The insights from this research offer practical guidance for safety-conscious organizations, with the potential to positively impact workplace safety practices and decision-making.

1 Introduction

1.1 Background

Due to the inherent dangers in the large construction business, safety must be ensured. Accidents not only harm workers but also hinder project success. This study explores OSHA Accident and Injury Data (2015-2017) to address these challenges.

Focusing on Health, Safety, and Environment (HSE), the research sits at the intersection of technology and safety management. HSE faces incident management issues due to unstructured reports, prompting the use of machine learning to glean insights from OSHA data [1].

Eight machine learning algorithms, including Regression, k-NN, Decision Tree, SVM, Naive Bayes, XGBoost, and AdaBoost, are deployed. XGBoost and AdaBoost excel with

[1]Dataset:`https://www.kaggle.com/datasets/ruqaiyaship/osha-accident-and-injury-data-1517`

86.2% and 85.6% accuracy and AUC values of 92.2% and 92.1%, respectively, emphasizing hyperparameter tuning.

Building on prior work (Mohammed and Mahmud (2020)) this research affirms findings, aspiring to surpass benchmarks. It contributes to workplace safety literature, recognizing machine learning's impact.

The OSHA [2] dataset, rich in structured information within unstructured text, allows the investigation of injury patterns and causal factors. This research bridges the gap between incident reports and proactive safety management.

In summary, this paper aims to enhance construction site safety using machine learning. By predicting accidents and establishing a data-driven safety management system, it strives to transform safety practices. Subsequent sections will delve into methodology, results, and implications.

1.1.1 Novelty and Contributions

- Integration of machine learning to enhance safety in construction.
- Application of advanced algorithms (Regression, k-NN, Decision Tree, SVM, Naive Bayes, XGBoost, AdaBoost).
- Emphasis on hyperparameter tuning for optimal algorithm performance.
- Contribution to workplace safety literature, building on prior work.
- Bridging the gap between incident reports and proactive safety management.

1.2 Research Question and Objectives

"What are the most effective machine learning algorithms and feature engineering approaches for combining diverse data sources, such as historical accident records, worker demographics, site characteristics, and real-time sensor data, to create a dependable predictive risk model for construction site accidents?"

The research significantly advances the understanding of predictive risk modeling for construction site accidents through the utilization of machine learning algorithms and innovative feature engineering techniques. The findings highlight the effectiveness of ensemble methods, especially XGBoost and AdaBoost, in addressing the complexities of predicting construction site safety. XGBoost and AdaBoost stand out, achieving predictive accuracy rates of 86.2% and 85.6%, respectively. The corresponding AUC values of 92.2% and 92.1% further emphasize the superior discriminatory power of these algorithms.

These machine-learning algorithms demonstrate a surprising ability to combine data from various sources, such as real-time sensor data, worker demographics, previous accident records, and site attributes. The study explores eight reliable machine-learning algorithms, providing a thorough understanding of both the individual and collective prediction capabilities of each.

Beyond algorithmic selection, the research places significant emphasis on feature engineering, as seen in Case Study 3. This case study carefully explores precision-driven hyperparameter refinement through a two-step process involving Randomized Search and targeted Grid Search. The outcomes go beyond improved accuracy to include enhanced

[2] `https://www.osha.gov/`

precision, recall, and balanced F1 scores. This precision-focused approach highlights the effectiveness of the models in real-world applications.

The success of precision-driven hyperparameter refinement underscores the careful nature of the research methodology. The Randomized Search and targeted Grid Search contribute to a nuanced understanding of how fine-tuned configurations amplify not only the accuracy but also the overall effectiveness of predictive risk models. These insights go beyond a mere exploration of algorithms and delve into the intricacies of model optimization, providing practical and actionable outcomes for the construction industry.

In essence, the research not only identifies high-performing machine learning algorithms but also underscores the importance of precision-centric feature engineering for creating reliable predictive risk models. The synthesized insights contribute significantly to the evolving landscape of workplace safety prediction, offering practical implications for the construction industry and beyond.

2 Related Work

2.1 Real-Time Safety Monitoring in Construction

The academic literature on real-time safety monitoring in construction, particularly using advanced technologies such as drones and machine learning, offers a nuanced perspective on enhancing safety protocols. Noteworthy contributions include the paper (Shanti et al. (2022)), which employs drones and deep learning for real-time monitoring of work-at-height safety. The strengths lie in its innovative integration of technologies, providing a holistic view of potential hazards. However, the limitation includes a lack of extensive validation in real-world construction scenarios. In contrast, Park's work (Cho et al. (2018)) focuses on scaffold safety conditions, demonstrating a robust machine-learning model. The strength here lies in the specificity of the safety aspect addressed, but scalability concerns and a need for broader applicability emerge as limitations.

Another significant study by (Yu et al. (2022)) utilizes hybrid machine learning techniques for identifying personal safety equipment in real time. The positive aspect is the emphasis on worker compliance, yet challenges related to the adaptability of the model to diverse construction settings are noted. The research conducted by (Nath et al. (2020)) explores deep learning for real-time detection of personal protective equipment, showcasing the potential for immediate safety alerts. The high injury rate in the construction sector is one of the biggest issues. Digitalization has been providing some new opportunities for further enhancement in the construction industry (Soltanmohammadlou et al. (2019)). The use of digital technologies can enhance safety measures in the construction sector. The construction industry is facing challenges such as low productivity and inefficiencies. Some of these issues can be summarized by the digitalization of the industry. It can also increase the safety of construction sites by implementing time-efficient monitoring and analysis of data. The fourth industrial revolution confirms the unification of advanced technologies in the construction industry (Zhang et al. (2020)). This is closely related to the integration of similar technologies in the construction industry. There are advanced technologies, such as big data, the internet, etc. Advanced technologies, such as drones and machine learning, can be used to tackle how changes in demand and costs of materials, as well as some other factors, can affect project duration (Gheisari and Esmaeili (2019)).

2.2 Machine Learning Applications in Construction Site Safety

In recent years, the digital collaboration and implementation of technology have been accelerated in the construction sector. However, it is not so easy to find the resources that can be implemented when attaching to new projects. The machine learning applications will improve the industry, improve the outcomes of employees, and contract with various companies and clients (Nath et al. (2020)). The use of machine learning can help improve safety, quality, and further efficiencies in this sector. Machine learning, or AI, can transform human capabilities in machinery concepts and help with more design as well as planning. These can help humans spend their time with their families and explore their creativity (Akinosho et al. (2020)).

Furthermore, machine learning can assist teams and reputed companies in making informed predictions as well as streamlining workflows. Machine learning, which is a subset of AI, is an Artificial Intelligence approach, that can signify intelligent behavior. The machines can predict the outcomes in terms of their capabilities. Sometimes, humans get tired of doing the same work at the same time. Machine learning makes it easy to do the work and lessens the work of humans. AI and machine learning can help architects and engineers deliver more sustainable construction in the future. Machine learning also assists employers or workers in figuring out their mistakes. Sometimes, humans cannot solve their problems, but with the help of machine learning, they can solve and implement their further problems in the construction sector (Wu et al. (2019)). As a replacement for thinking in hours, one can depend on machine learning. This can help to save time and increase productivity at work. Employees can be more productive when they can implement the uses of machine learning. The accuracy rate in terms of machine learning is also high, rather than the productivity of a human being in any sector. However, there are some complications, such as scalability, comprehensive indicator coverage, etc. This study covers the significance of machine learning, and it can signify the challenges of machine learning. There is also a need to make more valuable contributions in terms of construction site safety, which would be more unified and scalable.

2.3 Text Mining in Construction Site Accident Analysis

Workplace safety is an important concern in some countries. The construction sector plays a crucial role in every country. Construction accidents not only denote human suffering but are also a prime concern due to the huge amounts of financial losses. There is a need to make a scientific risk control plan in terms of future control, focusing on the previous losses in the construction sector. In this industry, catastrophe as well as fatality summary reports are also essential. There are always some harmful objects that lead to accidents in the construction sector. Identifying those objects is extremely helpful for further decreasing the potential risks.

Furthermore, the construction industry is the most dangerous workplace. There are millions of deaths caused by occupational accidents. For the prevention of these types of accidents, there is a need to promote workplace safety and the right actions by safety professionals. It is always noted that accidents in these sectors are caused by the presence of harmful objects, such as misused tools, damaged equipment, etc. Sometimes, proper and regular checks before the operation of a machine can prevent accidents. In this sector, an assembled model is required to optimize the Sequential Quadratic Programming (SQP) algorithm. The study contributes to collectively underscore text mining as well as machine learning techniques in the insights of the construction accident narratives (Xu and Saleh

(2021)). These studies signify text mining and machine learning techniques for extracting meaningful insight from the construction of accidental narratives.

Furthermore, in the construction sector, accidents can be related to property damage and personal injury. Accidents may be referred to as damages to the building, or there may be sustained changes in the employees. Some major contributions of the works in this sector include exploring various text mining and NLP techniques at construction sites in terms of analyzing accidents. Furthermore, Natural Language Processing (NLP) algorithms can be implemented to extract relevant information as well as sentiments from the unstructured texts, which facilitates more comprehensive factors that contribute to accidents (Ofer et al. (2021)).

2.4 Machine Learning Predictive Models for Construction Worker Fatalities' Dependability in Answering the Research Question

The academic landscape on machine learning predictive models for construction worker fatalities is rich and diverse, with five notable contributions shedding light on various aspects of this critical safety concern. This study, (Choi et al. (2020)), focuses on some machine learning predictive models in which construction worker fatalities are diverse. The strengths imply that the extensive dataset, which covers 137,323 injuries and 2,846 deaths, provides a robust foundation. However, the limitation includes the exclusive focus on Korean data, potentially limiting the model's generalizability.

Moving on to (Zhu et al. (2023)), the development of predictive models for construction fatality characteristics stands out. The strength lies in its comprehensive approach, considering various characteristics, yet the limitation involves potential challenges in adapting the model to diverse construction contexts. Additionally, the hybrid framework of econometric models and machine learning approaches in the research (Hu et al. (2023)) analyzes factors affecting urban fatal crash types. The innovative approach demonstrates strengths in merging economic and machine learning perspectives but may face challenges in scalability and broad applicability across construction sites.

In the research (Zermane et al. (2023)), the prediction of fatal fall accidents from heights using a random forest classification machine learning model is explored. The strength lies in the specificity of addressing falls, a common cause of fatalities, yet the potential limitation involves the need for comprehensive validation in various construction scenarios. Lastly, research (Koc (2023)) investigates the role of national conditions in occupational fatal accidents in the construction industry. The use of interpretable machine learning approaches offers valuable insights into contextual factors. However, the potential limitation includes the necessity for a nuanced interpretation of national conditions and their influence on accidents.

Comparatively, these studies contribute significantly to understanding and predicting construction worker fatalities using machine learning. Each study brings unique strengths and insights, but a critical analysis reveals common limitations such as dataset specificity, scalability challenges, and the need for comprehensive validation. This necessitates further research that not only amalgamates the strengths observed in these works but also addresses their limitations. The present study aims to contribute to this evolving field by providing a more generalized, scalable, and accurate machine-learning predictive model for construction worker fatalities.

2.5 Limitations and Takeaway from the Literature

The literature on real-time safety monitoring, machine learning applications in construction site safety, text mining in accident analysis, and predictive models for construction worker fatalities showcases significant advancements but is not without challenges. The research is mainly focused on the application of machine learning in the construction site which has not been analysed about the incident where the machine learning needs to apply. The research has been developed with the help of theoretical concepts regarding the application of machine learning in construction sites. For this reason, limited real-life experience and evidence have been aligned in this study. This limitation is required to recover for analyzing the real-life evidence of machine learning Technology for improving construction site operations.

3 Methodology

In the Methodology section, the applied tools and methods have been discussed to analyze the process of collecting the data and setting the data pattern to address the research problem. The data has been collected with the help of the Python programming language and the Python library in the computer program. A Pandas data frame has been used to input the collected data and status data structure to analyze the real facts of machine learning applications on construction sites

3.1 Data Collection

The data has been collected through the adoption of the keywords: [3] "OSHA Accident and Injury Data," where a case study has been used to analyze employees who were injured and forecast the unexpected accident [4] [5]. In the research paper, it has been explained the role of machine learning on the construction site and how this technology is supported to make a safe construction structure that has long-term durability and stability. The role of machine learning in construction sites and its impact on safety measurement data have been collected from the Knowledge Discovery in Databases (KDD) for an explanation of the role and effective place of machine learning importance in the construction site (Plotnikova et al. (2020)). The descriptive sampling technique has been used to find the right information regarding the machine learning application on the construction site.

[3]Dataset: `https://www.kaggle.com/datasets/ruqaiyaship/osha-accident-and-injury-data-1517`

[4]Website: `https://www.osha.gov`

[5]Wikipedia: `https://en.wikipedia.org/wiki/Occupational_Safety_and_Health_Administration`

3.2 Data Selection

The data has been selected with the help of keywords. Effective keywords such as machine learning, construction sites, employee injury, and health and safety can be used to collect the real-time application of machine learning technology for the construction project management plan [6].

3.2.1 Inclusion and exclusion criteria: Table: 1

Table 1: Inclusion and Exclusion Criteria

Inclusion	**Exclusion**
The data has been collected from peer-reviewed research journals and articles published after 2019.	Blogs, essays, and research assignments are excluded from the study.
The focus is on the role of machine learning technology applications in construction sites for safety measurement rather than other industries.	Other advanced industrial technologies, such as AI, big data, and IoT are excluded.
The research analyzes how machine learning supports developing employee safety during construction.	The application of machine learning in data-driven processes is excluded from the study.

3.3 Data Pre-processing

The data preposition system and its exploration have been completed with the help of the Python programming language in the computer (Deshmukh et al. (2015)). A Python library has been used to set the data frame for inputting the collective data in an Excel form to analyze its visualization graph to identify the degree of injury and examine how the machine learning system can help minimize accidental cases in construction sites.

3.3.1 Data Cleaning:

Data cleaning is a process that can fix or remove incorrect or corrupted data within a data set. In the data-cleaning process. Data cleaning, which is also referred to as data cleansing, is one of the best steps for any organization if one wants to create a culture around quality in data decision-making. While the data is incorrect, the outcomes, as well as the algorithms, are not reliable, even when they may look correct (Ilyas and Chu (2019)).

3.3.2 Handling Missing Values:

The process of handling missing data involves removing any rows or columns that contain null values. One can remove a column entirely if any of the values in it make up more

[6]Dataset: `https://www.kaggle.com/datasets/ruqaiyaship/osha-accident-and-injury-data-1517`

than half of the total. Similarly, the rows could also be dropped when they have one or more values. When data is missing completely, the probability of any particular value being missed from the dataset is high. If the data is in a data frame called original data, one can drop columns with the missing values. Missing data is a problem that a data professional needs to deal with. Missing data can be defined as unavailable values that could be observed meaningfully. Missing data could be anything, which is different from the missing sequence, files missing, incomplete information, error data, etc (Johnson et al. (2021)).

3.3.3 Feature Engineering:

In data science, feature engineering is referred to as addition, manipulation, deletion, mutation, and combination. These lead to better performance as well as better accuracy. Creating new features gives a great understanding of the data and better results. Feature engineering is the most valuable technique in data science and is also very challenging. The most common example in feature engineering is when doctors use the body mass index (BMI). This can be calculated based on body weight and height. Feature engineering also involves transforming raw data into a format, which can increase the performance of machine learning models (Zheng and Casari (2018)). There is a need to explore and understand the dataset in feature engineering. There is a need to understand the shape of the data key

3.3.4 Encoding Categorical Variables

Categorical variables are represented as 'strings' or 'categories'. The performance of the machine learning model depends on not only the model but also on the process and different types of variables in the model. While the machine learning model accepts numerical variables, categorical variables are also a necessary step in machine learning. There is a need to convert or transform these categorical variables into numerical variables. All stages are crucial for cleaning and preparing data. A data scientist is always concerned about cleaning and preparing the data. Whereas, converting the categorical data is always an unavoidable activity (Cerda et al. (2018)).

3.3.5 Scaling and Normalization

Data scaling is a term that confines data transformation activities, and it is aimed at enhancing the informational content of data. It can be used by adjusting the existing data set, which can confirm the set of requirements. The scaling adjustment might be implied in model assumptions, and it may or may not be true. Data, which can be derived from scaling, needs to be cross-checked. Sometimes, in this data scaling process, data needs to be historically available, and it is needed to estimate by extrapolation as well as interpolation. Data normalization is the process where data can be recognized within the database and the users can utilize the data for further analysis. Data normalization is the process of developing data. In this process, unstructured and redundant data can be eliminated (Kumar et al. (2022)). Distribution plots vividly depicted the normalized impact on variables like the nature of the injury, part of the body, and the human factor shown below Figure [1],[2]. The resulting dataset, enriched with scaled numeric features, is now poised for advanced analysis and the development of machine learning models.

This meticulous process underscores a commitment to dataset preparation, enhancing predictive power while maintaining model integrity.

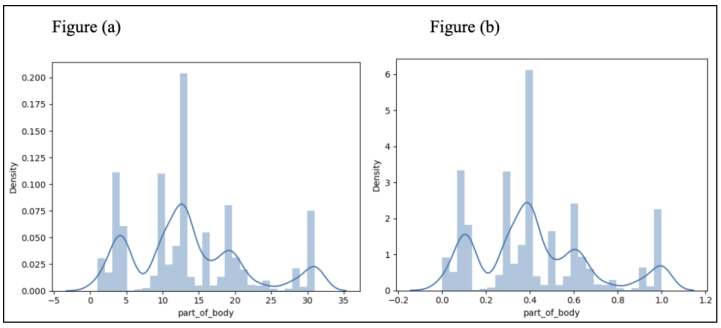

Figure 1: Part of Body Column Before (a) and After (b) Scaling

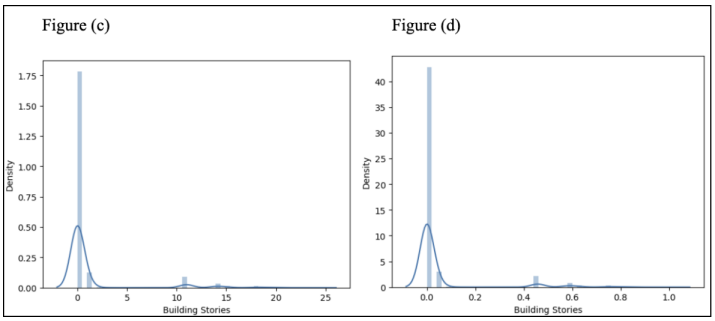

Figure 2: Building Stories Column Before (c) and After (d) Scaling

3.4 Model Training

In the phase dedicated to model training, systematic algorithmic exploration was conducted to identify the most suitable model for predicting fatalities within the construction industry. This pivotal model selection phase involved comprehensive testing of various algorithms, specifically targeting classifier algorithms, given the multi-classification nature of predicting the degree of injury.

3.4.1 Classifier Algorithm Testing

Diverse algorithms, including Random Forest, XGBoost, Decision Tree, k-Nearest Neighbors (k-NN), Logistic Regression, SVM, Naive Bayes, and AdaBoost, were subjected to exhaustive testing against the dataset (Mohammed and Mahmud (2020)) (Manjunatha (2023)). The selection of these algorithms was guided by insights from prior research, ensuring a thorough exploration of their predictive capabilities.

3.4.2 Hyperparameter Tuning

During the machine learning model refinement process, adjusting the hyperparameters of many algorithms was essential to ensuring their optimal performance in real-world

situations. To suit the distinct features of the dataset, each algorithm's specifics had to be examined and adjusted throughout this tuning step (Bardenet et al. (2013)). Numerical estimators (n), max depth, minimum sample split and minimum sample leaf, C, penalty, solver, n neighbors, weights, p, criteria, kernel, gamma, var smoothing, learning rate, reg lambda, and other parameters were carefully adjusted throughout the tuning process.

A solid basis for wise decision-making is established by this meticulous approach during the model training phase. It helps in subsequent assessments and lays the groundwork for selecting the most accurate model to forecast deaths in the construction industry. Hyperparameter tuning can allow tweaking of the model's performance. This is a crucial part of the machine-learning process. There is a need to choose an appropriate hyperparameter value to succeed.

3.5 Model Evaluation and Presentation:

Model evaluation is a process in which a data scientist can use various evaluation metrics to understand the performance of machine learning. It can also identify its strengths as well as its weaknesses. This is very important, as it can enhance the efficacy of a model. Many steps can be used to classify the performance of model evaluation. These are accuracy, confusion matrix, log-loss, precision, and AUC. The first step, which is called accuracy, can measure and classify the correct prediction, which is the ratio of the correct prediction to the total number of predictions. Precision is a good choice for the model evaluation matrix. In this step, one can be sure about the prediction. This can measure the proportion of predicted values that are truly positive. Whereas, a confusion matrix can show a detailed breakdown of the correct as well as incorrect classifications. As an example, the prediction of false positive and false negative consequences in cancer diagnosis is very difficult.

3.6 Train and Test Split:

It is a technological tool that has been used for data splitting, which is important for the testing of all of the data and examining its efficiency to justify the research objectives. A training data set model has been developed in the train and test split that helps to analyse different safeguards against overfitting. With this careful separation of training 75% and testing sets, the model's effectiveness can be carefully examined. A large training dataset makes the model more capable of learning, and the reserved 25% (Mohammed and Mahmud (2020)) functioned as a test during testing to see how well the model generalizes to new cases. This meticulous approach was supported to analyze the lifecycle of the real-life risk situation at the construction sites.

4 Design Specification

4.1 Modelling Technique

The modeling technique is based on the use of algorithmic sequences of instructions for solving some specific problems. One can use a particular algorithm for creating this type of model. Modeling techniques can be categorized into three main classes, such as supervised, association, and segmentation. The supervised model can be used to predict the value or the target. There are some techniques in modeling. These are decision

trees, regression, logistic, neural networks, support vector machines, generalized linear, and Bayesian networks (Muthukrishnan et al. (2020)).

The association model can find patterns of data where many entities are connected and associated. The data in this field can act as both targets and inputs. One can find these connected associations manually, whereas association algorithms can explore some complex patterns

4.1.1 Random Forest (RF):

Random Forest is a machine-learning algorithm that can combine the output of multiple decisions into a single result. As the random forest model is made up of multiple decision trees, it could help start the decision by clearly describing the tree algorithm. It can be confined to multiple methods and algorithms to obtain a better model. This random forest is the collection of a group of people who have limited knowledge. In this algorithmic context, the random forest continuously searches for features that can be allowed as per the observations (Speiser et al. (2019)).

4.1.2 Logistic Regression (LR)

Logistic regression is the model that is used for optimizing the probability of a discrete outcome given an input variable. In this study, this system is used for identifying the trends of issues at construction sites where ML can be used to minimize the risk as well as help save employees from hazardous accidental cases. As per the opinion of (Hussein et al. (2021)) a myriad of predictor variables illustrate current risk, prioritize tonsure about its risk level and how it should affect that predictive information can be illustrated in the data model structure

4.1.3 K-nearest neighbors(KNN)

In the descriptive analysis, K-nearest neighbors (KNN) are the non-parametric, supervised learning classifier, and regression problem tools that help to receive the predictive outcome as well as new suggestive information in the form of discrete data sets (Kigo et al. (2023)). This software is useful for getting prominent information about the current accidental cases in construction sites. On the other hand, KNN helps to analyze predictive risk level support with the other domain to analyze how risk is essential to address as well as make a positive environment in the workplace. For this reason, KNN has been used in this study to optimize the predictive accidental cases for alignment with the benefits of ML for the risk management plan.

4.1.4 Decision Tree Classifier(DT)

As per the opinion of (Bansal et al. (2022)) Decision Tree Classifier(DT) is used in ML technology for set data, a tree pattern for perfect analysis of the data, as well as interpretation. On the other hand, Mekonnen et al. (2019) stated that multiple information and data analyses help to make multiple decisions based on input variables. In this way, ML programs support making multiple decisions to provide predictive information. In this study, the Decision Tree Classifier(DT) has been used to set data in the different subsequent partitioning and attribute segments. In this way, large amounts can be imported into the data tree that supports improvement and decision as well as maintaining

accuracy. In this study, this technology is used to set the process plan for risk mitigation with the help of ML support. The beneficial factors are that DT helps to provide real data within a shorter training period and identifies high-dimensional data with priorities at the functional level.

4.1.5 support vector machine (SVM)

Support vector machines (SVM) are part of a data classification system that helps detect patterns in construction. As per the opinion of (Meza et al. (2019)) a support vector machine (SVM) is used for regression problems through the help algorithm of the regular task as well as a flexible data analytics process to distinguish data classes. In this study, a support vector machine (SVM) has been used for descriptive analysis as a hyperplane system to bring the closest data for supporting analysis. In this study, risk has been provided in the statistical software for setting the data set of input to optimize the closest risk register plan in the study

4.1.6 Naive Bayes (NB)

Naive Bayes (NB) theory is considered the emergent algorithm system used to classify tasks for data interpretation and provide predictive information regarding data scenarios. In this study, Naive Bayes' theorem has been used in probate conduction to optimize the accident's occurrence as per the given data (Berrar (2018)).

4.1.7 Extreme Gradient Boosting(XGBoost)

As per the opinion of (Kiangala and Wang (2021)) Extreme Gradient Boosting (XGBoost) technology is used for fast accessing of data processes through increasing algorithmic system ability. Extreme Gradient Boosting (XGBoost) was used in this study for the adoption of predictive analysis regarding the input data. In this research, this learning technology was used for setting up accidental models at construction sites based on the accidental data as well as making it flexible to predict the nature of risk.

4.1.8 Adaptive Boosting (AdaBoost)

Adaptive Boosting is one of the effective learning models that helps to design predictive models to overcome weaknesses according to data scenarios. As per the opinion of (Heo and Yang (2014)) Adaptive Boosting is one part of a machine learning technique that is used to classify the data to analyze it based on the decision tree. It has been found that this system is supported by accident patterns, which helps understand how it should affect human health. In this study, this technology has been used to receive effective analysis of project work

4.2 Evaluation Technique

4.2.1 Accuracy

Accuracy and measurement are required to set off the data pattern of the risk and analyze how ML is used to make a risk mitigation plan. Accuracy is required to proofread the data pattern to optimize false information that should affect the decision-making process (Hastie et al. (2009)).

$$Accuracy = \frac{noofcorrectprediction}{totalnoofprediction}$$

4.2.2 Precision

Precision is used to analyze data on the construction site's accident cases. Similarity cases can be resolved as well, and it also supports enhancing accuracy in risk management plans (Hastie et al. (2009)).

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

4.2.3 Recall

It has been found that the record model can help to analyze the positive cases to identify the safety sites, which is important for making the risk management plan regarding accidents. Based on the information, it can be stated that it helps to improve the positive nature of the data as well as provide effective recommendations to improve the decision while making the risk management plan (Hastie et al. (2009)).

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

4.2.4 F1_Score

This F1 Score system is used to evaluate recall to measure the effective construction of accidental problems as well as navigate the model's ability to identify risk (Sokolova and Lapalme (2009)).

$$F1Score = \frac{2(Precision*Recall)}{Precision+Recall}$$

4.2.5 AUC-ROC

This technology is used to identify the data collection model to optimize data strength rather than weakness regarding the construction site and analyze the accidental case regarding humans. In this way, a multidimensional approach supports the improvement of the decision-making process of the risk management plan as well as solving the research problem (Fawcett (2006)).

5 Implementation

5.1 Tools Used

Using Tableau for clear visualizations of incident root causes and Python, a versatile language, for modeling and analysis. The varied library tools in Python make our work efficient and effective

5.2 Data Selection

The dataset chosen for this research is a comprehensive compilation of OSHA Accident and Injury Data, spanning the years 2015 to 2017. This dataset specifically focuses on accidents and injuries among construction workers, providing valuable insights into the challenges faced in the Health, Safety, and Environment (HSE) domain. With 4848 rows and 29 columns, this dataset forms the basis for analyzing trends, identifying key factors contributing to injuries, and guiding the development of effective training and safety measures.

5.3 Exploratory Data Analysis

In exploratory data analysis, tableau is one of the most important aspects of quickly visualizing data in a way that makes it easy to understand below Figure [3]. In Sheet [1], a bar chart is given, which gives us the details of the distribution of fatal and non-fatal incidents. It is categorized based on the nature and degree of injury The bars can be plotted horizontally and vertically, which shows a holistic view of injury patterns across diverse categories The further sheet [2] shows a temporary breakdown that reveals significant variations in incident numbers over the years, such as 11,039 fatal incidents and 231 non-fatal incidents in 2015, 22,800 fatal incidents and 1,568 non-fatal incidents in 2016, and 7,965 fatal incidents and 22,461 non-fatal incidents in 2017. In Sheet [3], circular visualization illuminates the relationship between body parts and the degree of injury. Sheet [4], the correlation between project cost and building stories

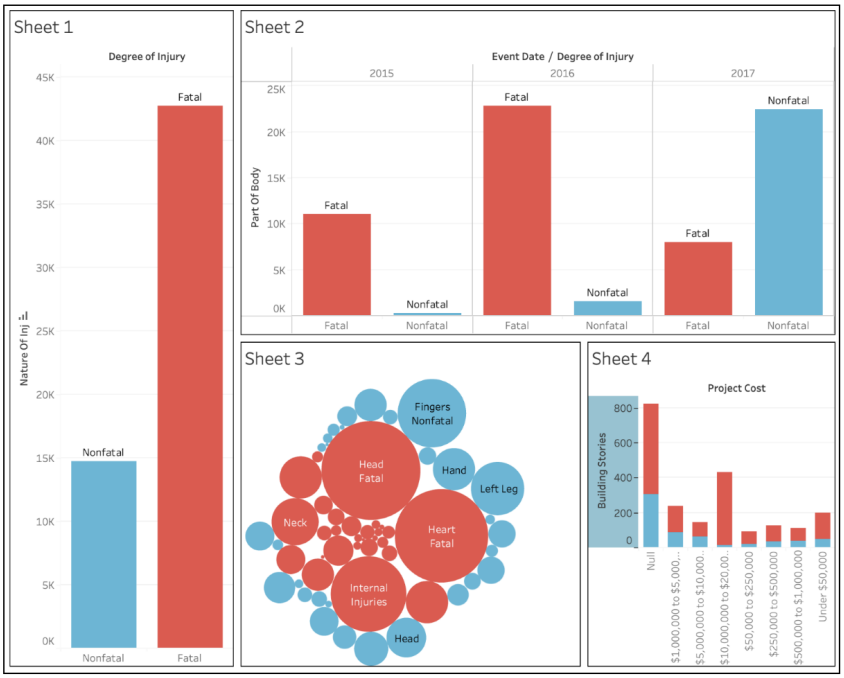

Figure 3: Raw Data Dashboard.

5.4 Data Cleaning

5.4.1 Handling null values

The dataset was meticulously examined to identify null values using the 'IsNull()' and 'sum()' functions. Notably, columns like Nature of Injury, Part of Body, Event Type, Environmental Factor, and Human Factor exhibited missing values, which are crucial for subsequent decisions regarding imputation or removal.

5.4.2 Handling Missing Values

Missing values were effectively handled using the 'dropna()' function. streamlining the dataset for further analysis by eliminating rows with any missing values.

5.4.3 Label Encoding

To enhance the dataset's readiness for machine learning algorithms, label encoding was applied using 'LabelEncoder()'. This involved converting specified columns into a numerical format for numerical processing

5.4.4 Normalization

Starting with a thorough analysis using 'sns.distplot()' to comprehend data distribution, normalization ensued. We applied 'MinMaxScaler()' for uniformity, making the dataset more suitable for modeling.

5.4.5 Feature selection

The features selection phase involved a comprehensive examination, including the calculation of a correlation matrix using 'numeric data.corr()'. The resulting heatmap visualization aided in discerning relationships between features, guiding the selection process for the most relevant variables in the dataset.

5.5 Hyper parameter Tuning

To maximize the effectiveness of machine learning models for practical applications, fine-tuning hyperparameters is an essential first step. To guarantee optimal performance of these algorithms, this procedure entails modifying certain parameters, such as the number of trees in a Random Forest or the learning rate in an AdaBoost model (Bardenet et al. (2013)) The hyperparameter values used for each classifier are displayed in the below table: [2]; each is designed to achieve the best possible balance for the model setup.

Table 2: Hyperparameter Tuning Values for Classifiers

Classifier	**Hyperparameter**	**Tuned Value(s)**
Random Forest (RF)	n_estimators	100
Random Forest (RF)	max_depth	10
Random Forest (RF)	min_samples_split	5
Random Forest (RF)	min_samples_leaf	1
Random Forest (RF)	bootstrap	False
Logistic Regression (LR)	C	0.1
Logistic Regression (LR)	penalty	l1
Logistic Regression (LR)	solver	liblinear
K-nearest Neighbors (KNN)	n_neighbors	4
K-nearest Neighbors (KNN)	weights	Uniform
K-nearest Neighbors (KNN)	p	1
Decision Tree (DT)	min_samples_split	2
Decision Tree (DT)	min_samples_leaf	2
Decision Tree (DT)	max_depth	7
Decision Tree (DT)	Criterion	gini
Support Vector Machine (SVM)	kernel	rbf
Support Vector Machine (SVM)	gamma	scale
Support Vector Machine (SVM)	C	10
Naive Bayes (NB)	var_smoothing	0.3511
XGBoost (XGB)	var_smoothing	0.8
XGBoost (XGB)	n_estimators	200
XGBoost (XGB)	max_depth	7
XGBoost (XGB)	learning_rate	0.01
XGBoost (XGB)	Colsample_bytree	0.8
AdaBoost (Ada)	n_estimators	200
AdaBoost (Ada)	Learning_rate	0.2

6 Evaluation

The evolution section has been conducted to analyze the result of the study by setting key findings as a data pattern to address the research problem. The data has been collected with the help of rigorous statistical tools for the critical experiment of receiving data to obtain outcomes.

6.1 Case Study 1: Optimizing Workplace Safety with Machine Learning in OSHA Fatalities Prediction

Advanced workplace safety is essential in construction sites that support increased trust and confidence of employees to work in the workplace. From the findings, it has been found that the transformative power of machine learning can identify and predict workplace risk and accidental information for taking precautions to prevent negative impacts among the employees in the workplace. It has been optimized that Machine Learning can help analyze the overall case scenario regarding the Occupational Safety and Health Administration (OSHA). In this way, a machine learning algorithm system supports the evaluation of real-world scenarios of construction site risk and accidents, providing effect-

ive predictive solutions for addressing high-level risks to workplace safety. (Muthukrishnan et al. (2020)) It was observed from the findings that around nine machine learning algorithms, such as Decision Tree, Naive Bayes, and the standout performer Gradient Boosted Trees system, have been used in this study to set challenge-based data in a decision tree block and interconnect with real-time data to identify the new strategic path to address challenges and support developing a positive work environment for employees. In this way, strong predictive capabilities have been developed to address the challenge in the workplace. The conclusive findings position Gradient Boosted Trees as the optimal algorithm, boasting 65.29% accuracy, 34.71% classification error, and an impressive 78.83% precision. This case study not only identifies a gap in existing solutions for predicting OSHA fatalities but also serves as a guiding light for organizations aiming to leverage machine learning for enhanced workplace safety.

Advancements and Future Prospects: Building upon the foundation laid by the original research, this case study delineates the methodologies that led to significant improvements in accuracy and precision. The refinement in data cleaning processes, experimental setups, and algorithmic applications contributes to the continuous evolution of machine learning models in occupational safety. The study, while acknowledging challenges and limitations, opens avenues for future research and improvements. With its conclusive identification of gradient-boosted trees as the preferred model, this case study provides invaluable insights for organizations seeking to optimize workplace safety through the strategic integration of machine learning in their predictive analytics frameworks.

6.2 Case Study 2: Uncovering Predictive Excellence in Untuned Classifiers and Advanced Ensemble Methods for Fatalities Prediction

- **Exploration of Untuned Classifiers:**

  A close examination was conducted on three fundamental models—Random Forest, Logistic Regression and k-Nearest Neighbors (k-NN)—without altering their settings. The Random Forest model performed admirably, achieving an accuracy of 84.8%. Precision and recall, metrics for assessing model performance, were 83.5and 75.6%, resulting in a balanced F1 score of 79.4%. The ROC curve, illustrating the model's ability to distinguish outcomes, displayed an AUC of 91.3%, indicating strong discrimination. Logistic Regression, a straightforward method for binary classification, demonstrated an acceptable accuracy of 66.4%, with precision and recall at 57.5% and 50.6%, respectively. Using the k-NN method, around 74.3% accuracy, 68.4% precision, and 62.4% recall have been adopted, as well as setting the structure of the predictive database without modification.

- **Advanced Ensemble Methods in Fatalities Prediction**

  Naive Bayes and more complex models (XGBoost and AdaBoost) have been used in the descriptive analysis process. It has been found that the use of the XGBoost model around accuracy of 85.2%, recall of 75.9%, and precision of 84.3%, has been received as predictive capabilities with a 79.9% F1 score. On the other hand, a curve depicting the model's ability was supported to receive 92.2% confirming data regarding predictive analysis of the accidental cases in the construction sites. Apart from that, the AdaBoost method provided 85.6% accuracy and predictive Among

XG-Boost and AdaBoost methods, it can be stated that AdaBoost is more suitable than XG-Boost as it provides high accuracy reports that can help to analyse predictive risks through the help of precision and balance performance-based metrics.

6.3 Case Study 3: Advancing Predictive Models through Hyperparameter Refinement

- **Precision Unveiled**

  XGBoost and AdaBoost are effective machine learning systems supported to provide highly efficient and accurate risk analysis. It has been found that XGBoost featuring 85.8% accuracy, 87.9% precision, and a balanced F1 score of 79.9%, as well as 86.9% precision, 84.9% accuracy, and a balanced F1 score of 78.7%, and ROC AUC of 91.3% have been received from the AdaBoost model. Based on this information, it can be stated that precision-centric hyperparameter optimization is supported to activate The precision-crafted strategy while generating predictive capabilities against the construction of accidental cases.

- **Elevating Precision with Ensemble Mastery:**

  The result has been precision with the help of the hyper-parameter tuning process. It can be stated that the hyperparameter factor was highly supported to generate an impressive score through the use of the models of Random Forest, k-NN, Logistic Regression, and Decision Tree. A precision-oriented strategy has been enforced for the identification of the positive cases for predictive analysis. Based on this information, it can be stated that comprehensive hyperparameters are supported to increase the accuracy and precision of risk analysis models through the use of real-life applications below Table: [3].

Table 3: Tuning Results for Model Performance Metrics

Model	**Tuning**	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC**
Random Forest	Without tuning	0.848	0.835	0.756	0.794	0.913
Random Forest	With tuning	0.861	0.878	0.743	0.805	0.920
Logistic Regression	Without tuning	0.664	0.575	0.506	0.539	0.751
Logistic Regression	With tuning	0.656	0.562	0.485	0.521	0.748
k-NN	Without tuning	0.743	0.684	0.624	0.652	0.796
k-NN	With tuning	0.843	0.831	0.747	0.783	0.891
Decision Tree	Without tuning	0.805	0.750	0.744	0.747	0.800
Decision Tree	With tuning	0.749	0.688	0.641	0.665	0.807
SVM	Without tuning	0.730	0.646	0.667	0.656	0.79
SVM	With tuning	0.749	0.688	0.641	0.665	0.807
Naive Bayes	Without tuning	0.719	0.618	0.716	0.663	0.743
Naive Bayes	With tuning	0.719	0.618	0.715	0.663	0.743
XGBoost	Without tuning	0.852	0.852	0.843	0.759	0.922
XGBoost	With tuning	0.857	0.879	0.733	0.799	0.925
AdaBoost	Without tuning	0.856	0.864	0.746	0.800	0.921
AdaBoost	With tuning	0.849	0.868	0.720	0.787	0.913

6.4 Discussion

Machine learning systems can help to increase safety and secure work environment capabilities for the employees so that they can do their job properly. While navigating the OSHA data set, different kinds of Machine learning-based algorithmic systems have been used in the operating system to obtain secure and accurate predictive data for making risk management plans for their employees. As per the case study, a risk mitigation plan was essential to optimize the risk in overfitting and accidental cases. For this reason, the decision-making tree has been designed to input the data in the data block to increase the value of ROC and AUC for predictive analysis. XGBoost and AdaBoost have been used to measure comparative accuracy scores for better predictive analysis. The detailed exploration of Randomized and Grid Search for models like Random Forest, Logistic Regression, k- Nearest Neighbors, and Decision Trees gives valuable insights. The discussion, while acknowledging the improved accuracy and effectiveness achieved through precision-driven hyperparameter optimization, calls for a critical evaluation of the computational costs and resource intensity associated with such approaches.

In essence, these case studies contribute to the evolving landscape of workplace safety prediction. However, a straightforward discussion emphasizes the need for a balanced view, considering potential limitations, computational complexities, and practical applicability of findings. Suggestions for future research may involve a deeper investigation into the understandability of complex models, exploring different ways to engineer features, and assessing the scalability of precision-focused hyperparameter optimization in large-scale applications. Integrating these insights into the broader context of previous research highlights the ongoing progress in applying machine learning to enhance workplace safety.

7 Conclusion and Future Work

Concluding this study on workplace accidents in the construction industry, where safety is a major concern, the goal was to enhance safety using smart technology. Data from OSHA on accidents from 2015 to 2017 was utilized, applying advanced machine-learning methods. The focus was on improving safety management, especially in the Health, Safety, and Environment (HSE) context.

Various machine learning techniques were explored, including Logistic Regression, k-NN, Decision Tree, SVM, Naive Bayes, XGBoost, and AdaBoost. Notably, XGBoost and AdaBoost demonstrated strong performance, achieving accuracy rates of 86.2% and 85.6%, along with good AUC values of 92.2% and 92.1%. The research is focused on the prevention of accidents through prediction, and the concentration is given to finding improved approaches in this context. It would ensure that the findings would be robust and would cover the existing benchmarks. Through obtaining positive results such as precision and accuracy, the property will also be provided to the marketing board for generalization as well as overfitting. It will also involve complex work methods such as evaluation of scalability and exploration of real-life solutions regarding the prevention of frequent road accidents. The potential findings would have significant impacts on avoiding all types of accidents in the construction industry, speculum the domain of safety management systems. The research is trying to shed light on construction practices by leveraging "smart machine learning methods" through the application of data analysis.

8 Acknowledgement

I am grateful to my supervisor, "Arghir Nicolae Moldovan" for helping me throughout the research. He provided me with deep insights into the research context. I am also thankful to "National College of Ireland". I am also thankful to my parents for their strong support and continuous encouragement throughout the research. Without their support, the research would not have been accomplished.

References

Akinosho, T. D., Oyedele, L. O., Bilal, M., Ajayi, A. O., Delgado, M. D., Akinade, O. O. and Ahmed, A. A. (2020). Deep learning in the construction industry: A review of present status and future innovations, *Journal of Building Engineering* **32**: 101827.

Bansal, M., Goyal, A. and Choudhary, A. (2022). A comparative analysis of k-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning, *Decision Analytics Journal* **3**: 100071.

Bardenet, R., Brendel, M., Kégl, B. and Sebag, M. (2013). Collaborative hyperparameter tuning, *International conference on machine learning*, PMLR, pp. 199–207.

Berrar, D. (2018). Bayes' theorem and naive bayes classifier, *Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics* **403**: 412.

Cerda, P., Varoquaux, G. and Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables, *Machine Learning* **107**(8-10): 1477–1494.

Cho, C., Park, J., Kim, K. and Sakhakarmi, S. (2018). Machine learning for assessing real-time safety conditions of scaffolds, *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction*, Vol. 35, IAARC Publications, pp. 1–8.

Choi, J., Gu, B., Chin, S. and Lee, J.-S. (2020). Machine learning predictive model based on national data for fatal accidents of construction workers, *Automation in Construction* **110**: 102974.

Deshmukh, D. H., Ghorpade, T. and Padiya, P. (2015). Improving classification using preprocessing and machine learning algorithms on nsl-kdd dataset, *2015 International Conference on Communication, Information & Computing Technology (ICCICT)*, IEEE, pp. 1–6.

Fawcett, T. (2006). An introduction to roc analysis, *Pattern recognition letters* **27**(8): 861–874.

Gheisari, M. and Esmaeili, B. (2019). Applications and requirements of unmanned aerial systems (uass) for construction safety, *Safety science* **118**: 230–240.

Hastie, T., Tibshirani, R., Friedman, J. H. and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, Vol. 2, Springer.

Heo, J. and Yang, J. Y. (2014). Adaboost based bankruptcy forecasting of korean construction companies, *Applied soft computing* **24**: 494–499.

Hu, Z., Shi, Q., Chen, Y., Yuan, Q., Tao, Z., Bian, Y. and Haque, M. M. (2023). Analyzing factors and interaction terms affecting urban fatal crash types based on a hybrid framework of econometric model and machine learning approaches, *International Journal of Crashworthiness* **28**(6): 809–821.

Hussein, A. S., Khairy, R. S., Najeeb, S. M. M. and Alrikabi, H. T. S. (2021). Credit card fraud detection using fuzzy rough nearest neighbor and sequential minimal optimization with logistic regression., *International Journal of Interactive Mobile Technologies* **15**(5).

Ilyas, I. F. and Chu, X. (2019). *Data cleaning*, Morgan & Claypool.

Johnson, T. F., Isaac, N. J., Paviolo, A. and González-Suárez, M. (2021). Handling missing values in trait data, *Global Ecology and Biogeography* **30**(1): 51–62.

Kiangala, S. K. and Wang, Z. (2021). An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-xgboost and random forest ensemble learning algorithms in an industry 4.0 environment, *Machine Learning with Applications* **4**: 100024.

Kigo, S. N., Omondi, E. O. and Omolo, B. O. (2023). Assessing predictive performance of supervised machine learning algorithms for a diamond pricing model, *Scientific Reports* **13**(1): 17315.

Koc, K. (2023). Role of national conditions in occupational fatal accidents in the construction industry using interpretable machine learning approach, *Journal of Management in Engineering* **39**(6): 04023037.

Kumar, S., Gupta, S. and Arora, S. (2022). A comparative simulation of normalization methods for machine learning-based intrusion detection systems using kdd cup'99 dataset, *Journal of Intelligent & Fuzzy Systems* **42**(3): 1749–1766.

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

Meza, J. K. S., Yepes, D. O., Rodrigo-Ilarri, J. and Cassiraga, E. (2019). Predictive analysis of urban waste generation for the city of bogotá, colombia, through the implementation of decision trees-based machine learning, support vector machines and artificial neural networks, *Heliyon* **5**(11).

Mohammed, J. and Mahmud, M. J. (2020). Selection of a machine learning algorithm for osha fatalities, *2020 IEEE Technology & Engineering Management Conference (TEMSCON)*, IEEE, pp. 1–5.

Muthukrishnan, S., Krishnaswamy, H., Thanikodi, S., Sundaresan, D. and Venkatraman, V. (2020). Support vector machine for modelling and simulation of heat exchangers, *Thermal Science* **24**(1 Part B): 499–503.

Nath, N. D., Behzadan, A. H. and Paal, S. G. (2020). Deep learning for site safety: Real-time detection of personal protective equipment, *Automation in Construction* **112**: 103085.

Ofer, D., Brandes, N. and Linial, M. (2021). The language of proteins: Nlp, machine learning & protein sequences, *Computational and Structural Biotechnology Journal* **19**: 1750–1758.

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

Shanti, M. Z., Cho, C.-S., de Soto, B. G., Byon, Y.-J., Yeun, C. Y. and Kim, T. Y. (2022). Real-time monitoring of work-at-height safety hazards in construction sites using drones and deep learning, *Journal of safety research* **83**: 364–370.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks, *Information processing & management* **45**(4): 427–437.

Soltanmohammadlou, N., Sadeghi, S., Hon, C. K. and Mokhtarpour-Khanghah, F. (2019). Real-time locating systems and safety in construction sites: A literature review, *Safety science* **117**: 229–242.

Speiser, J. L., Miller, M. E., Tooze, J. and Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling, *Expert systems with applications* **134**: 93–101.

Wu, J., Cai, N., Chen, W., Wang, H. and Wang, G. (2019). Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset, *Automation in Construction* **106**: 102894.

Xu, Z. and Saleh, J. H. (2021). Machine learning for reliability engineering and safety applications: Review of current status and future opportunities, *Reliability Engineering & System Safety* **211**: 107530.

Yu, W.-D., Liao, H.-C., Hsiao, W.-T., Chang, H.-K., Wu, T.-Y. and Lin, C.-C. (2022). Real-time identification of worker's personal safety equipment with hybrid machine learning techniques, *International Journal of Machine Learning and Computing* **12**(3).

Zermane, A., Tohir, M. Z. M., Zermane, H., Baharudin, M. R. and Yusoff, H. M. (2023). Predicting fatal fall from heights accidents using random forest classification machine learning model, *Safety science* **159**: 106023.

Zhang, M., Shi, R. and Yang, Z. (2020). A critical review of vision-based occupational health and safety monitoring of construction site workers, *Safety science* **126**: 104658.

Zheng, A. and Casari, A. (2018). *Feature engineering for machine learning: principles and techniques for data scientists*, " O'Reilly Media, Inc.".

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.