Injury Prediction in Mining Industry through Applied Machine Learning Approaches

MSc Research Project
Data Analytics

Akash Manjunatha
Student ID: x21141797

School of Computing
National College of Ireland

Supervisor: Christian Horn

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Akash Manjunatha
Student ID:	x21141797
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Christian Horn
Submission Due Date:	01/02/2023
Project Title:	Injury Prediction in Mining Industry through Applied Machine Learning Approaches
Word Count:	8200
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:	Akash Manjunatha
Date:	28th January 2023

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):	

Injury Prediction in Mining Industry through Applied Machine Learning Approaches

Akash Manjunatha
x21141797

Abstract

The mining industry is a significant contributor to the American economy, but it is also one of the most dangerous industries to work in due to the complex and risky nature of mining operations. To protect workers and reduce fatalities and accidents, the government created the Occupational Safety and Health Administration (OSHA) and the Mine Safety and Health Administration (MSHA). These organizations set safety regulations and penalties for companies that violate them. Despite the implementation of these safety measures, there are still unacceptable risks for workers in the mining industry. MSHA requires companies to record all workplace accidents and offers resources to help mine operators comply with safety regulations. Employers who violate safety regulations face steep fines from either OSHA or MSHA. Use of technology in the mining industry especially in health and safety is very minimal, given the volume of data that has become available over the years, this industry needs technology. In this research, five machine learning algorithms and one deep learning algorithm are employed to categorize the degree of injury in the mining industry. Three case studies were undertaken in order to address the research issue. Case studies 2 and 3 put the presumptions from case study 1 into practice. XGboost, decision trees, and artificial neural networks all performed admirably in case study 2's prediction of whether or not a worker will take a day off due to injury, with an accuracy rate of about 92%. The results of case study 3 showed that multi-classification with XGboost outperformed other algorithms by accurately identifying the degree of injury with 91% in all matrices. The outcome of this research can be used in organizations to improve their health and safety practices, and more effectively forecast, and prevent injuries or accidents altogether.

1 Introduction

1.1 Background

Any workplace is unsafe due to risky behavior, a risky environment, a risky situation, and negligence. Of all industries, the mining sector is said to be the most hazardous due to its numerous complex operations, risky mechanisms, complex processes, and enormous machinery. It also represents one of the most significant economic sectors in the United States. Practically every industry and consumer goods is supported by mining, which provides the country's homes and businesses with reliable, affordable fuel as well as essential power and minerals. Currently, there are more than 88 minerals that are thought

to be crucial to the American economy and national security [1]. Mining supports around two million well-paying employment and benefits the economy in every state[2]. There are approximately 12567 operating mines, which produced 505.89 million employee hours and 577 million tons of coal in 2021 alone, according to statistics from the National Institute for Occupational Safety and Health (NIOSH).
Given the enormous amount of labour and capital invested in the mining industry, as well as the fact that the current Ukraine conflict has increased demand for energy and, in turn, the mining industry's risk of accidents, it is crucial to safeguard the workers in this sector, as a result, the OSHA and MSHA were established. Almost every area of mine safety and health was impacted by the enactment of the Federal Mine Safety and Health Act in 1977. [3] Fewer fatalities and mishaps have occurred because of following the safety guidelines outlined in those statutes. MSHA requires companies to keep track of all workplace accidents in the mining sector and offers services and resources to help mine operators comply with safety and health rules and regulations[4]. OSHA or MSHA will impose severe fines on employers who violate any of the safety regulations outlined in the act.[5] However, even though dangers and working circumstances have drastically altered over time, Americans still face unacceptable risks at work.

1.2 Research Question and Objectives

In light of the analysis above, a research question for this study has been developed

- "How effective are machine learning and deep learning algorithms at accurately predicting the degree of injuries in the mining industry and how can these predictions be used to improve health and safety measures?"

Any company's top priority should be keeping its employees safe, creating a workplace climate where they won't feel unsafe, and minimizing accidents. However, according to the MSHA website, 29 fatalities were reported in 2020, but that number rose to 37 in 2022. An increase in fatalities not only puts businesses in a terrible position but also raises employee anxiety and casts doubt on the company's dedication to workplace health and safety. To pinpoint the disaster's root cause and lower subsequent casualties, technology is required. After reading several research articles and journals, it is noticed that most of them, particularly those that focus on the mining industry, use a qualitative approach, such as safety officers' experience and intuition, surveys, and more statistical analysis to determine what caused the incidents. authors of this research (Jung and Choi; 2021) examined 109 papers published since 2018 on the use of machine learning in the mining sector. Out of 15 mine safety studies, only two papers on occupational safety were released, emphasizing the need for more mining safety research. Accidents do not happen randomly; rather, they have underlying patterns and trends that can be examined and documented using machine learning approaches. By predicting injuries and their causes and preventing companies from having to pay costly fines, this project helps the company and government while also significantly improving the health and

[1] https://www.usgs.gov/news/national-news-release/us-mines-produced-estimated-823-billion-minerals-during-2020

[2] https://nma.org/

[3] https://www.dol.gov/general/aboutdol/history/carter-msha

[4] https://www.msha.gov/compliance-enforcement/compliance-assistance

[5] https://www.msha.gov/compliance-enforcement/penalty-assessments-payments

safety of mine workers. The data sets are acquired from the MSHA website. The dataset will go through different data processing phases, such as data cleaning, pre-processing, initial visualization, etc. A model-ready version of the data will be created. The artificial neural network, a deep learning technique, and machine learning models include XGboost, Random Forest, Decision Tree K-Nearest Neighbors, and Stochastic Gradient Descent. Evaluation matrices, including the Confusion matrix, F1 score, precision, and recall will be used to gauge the modeling performance. The findings of a hyper-tuning operation will be presented when the results have been fine-tuned.

1.3 Document Structure

The research document is broken down into seven sections, each of which offers details on a distinct aspect of the investigation. The second section, which is divided into three subsections, summarizes what prior research has been done and highlights the unique characteristics of this project before closing; in the third segment, the technique utilized in the study is detailed; and in the fourth section, the design specifications, it provides information on the procedures and methods used and the project's key performance indicators. Section 5 will demonstrate how the technical solution was used for the research, Section 6 provides the depth of case studies and how the assessment procedure helped the research attain its goals. Finally, in section 7 the research paper will wrap up the topic with the relevant findings and discuss potential future studies.

2 Related Work

Recent academic research has focused on the prediction of injuries in mining and related industries as well as the factors influencing health and safety listed below. The segment concludes with a discussion of the section's limitations and key ideas, and the sub sections 1,2 and 3 provide a quick summary of several strategies.

2.1 Analytical Techniques for Measuring the Health and Safety Risks in the Mining Industry

Every industry strives to have zero accidents, and in the mining sector safety is a need, which translates to more research being done on the topic to lower the occupational injury rate. The relationship between risk variables and occupational injuries is investigated in this research(Ajith et al.; 2020). The primary objective of this research is to identify the factors associated with the number of injuries sustained by small-scale mining employees in Migori, Kenya. The study involved collecting data from a sample of 74 injured miners and 162 unaffected miners. After conducting a logistic regression analysis, the study found that single workers, lower-risk drug users, and disgruntled workers were more likely to experience multiple injuries. The study used descriptive statistics to identify the variables associated with the number of injuries sustained.

This research (Shekarian et al.; 2021) aimed to identify the causes of pneumoconiosis in coal mine workers. The data were divided into three groups: surface, subterranean, and other. To analyze the relationship between pneumoconiosis and different factors, the study used a multivariate regression model called the Generalized Estimating Equation (GEE) model. Additional analyses were conducted to identify the most important predictors. The study found that all of the hypotheses were supported by the GEE regression, and common causes for pneumoconiosis were identified for the surface and subterranean groups.

This research (Ganguli et al.; 2021) involved using a random classifier to categorize accidents into one of 45 categories based on their narrative description. The data were split in half, and 9 models were created to analyze different types of accidents. When compared to non-MSHA data, the NLP-based machine learning system performed well, with an evaluation accuracy of 96%.

This research paper(Amoako et al.; 2021) aimed to assess the risks associated with different classes of injury. The target variables were narrowed down to six, and four were eliminated because of their low presence in the data. Multi-class logistic regression was used for the study, and Person correlation was used to identify the seven variables that were most important for the degree of injury. This research also focused (Gendler and Prokhorova; 2021) on determining the best options for improving workplace safety and reducing injury rates. The independent variable was analyzed at different levels of significance, and a correlation study was conducted on the region's climate conditions to determine the overall injury rate in the area.

According to this article (Matloob et al.; 2021), deep learning and support vector machines are the best options for improving injury prediction in the mining industry. The author recommends using RMSE and the coefficient of determination to evaluate the performance of these models. The article (Hyder et al.; 2019)also notes that roof collapses and explosions from hazardous gases are the most common causes of accidents in underground mining. The authors provide a critical evaluation of AI and ML in the mining industry, highlighting areas where these technologies may not be effective in the future. Overall, the article emphasizes the need for a greater focus on safety in the industry's future growth.

2.2 Forecasting Workplace Injuries by the Application of Machine Learning Techniques

The research study that follows investigates how to forecast health and safety concerns across several industry sectors. This research paper (Ajayi et al.; 2020) evaluates the effectiveness of decision trees (DT), random forests (RF), and gradient boosting (GBM) in predicting safety outcomes in the mining industry. The authors use chi-square tests to analyze the relationship between different variables and identify strategies for minimizing potential safety risks. They also discuss the challenges of using pre-made machine learning algorithms to remove redundant variables from large datasets. The study (KALE and Baradan; 2020) also uses particle swarm optimization for feature selection and gradient boosting adjustment. The authors compare their results to those of previous studies, which have primarily relied on the survey and descriptive statistics rather than inferential statistics. To create a categorically identifiable dataset, the authors examined and reorganized more than 2000 accident report forms, incorporating the injury severity score concept. Statistical techniques such as univariate frequency analysis, cross-tabulation, and logistic regression were used to analyze the data and identify explanatory variables.

This research study (Choi et al.; 2020) aims to develop a predictive model for identifying the likelihood of catastrophic incidents on construction sites. The study uses four algorithms (RF, logistic regression, AdaBoost, and DT) for comparison and evaluation, and applies random oversampling to balance the class imbalance in the data. The RF algorithm outperforms the other models, achieving an AUROC curve of 91%. Another research (AlMamlook et al.; 2019) used similar algorithms, including naive Bayes, to identify the components and classify the severity of injuries on construction sites. The RF algorithm performed better than the other algorithms in this research study as well.

This research study (Poh et al.; 2018) uses prediction models to identify fatal accidents and safety indicators and to classify different types of injuries. The study used the feature selection technique to choose six project-related features and seven safety-related features from a total of 13 features. SMOTE was used to balance the class imbalance, and KNN, SVM, DT, LR, and RF machine learning techniques were applied. The study used a maximum depth of 6 to assess all of the data and performed model parameter tuning to prevent overfitting. The random forest algorithm outperformed the other techniques, achieving an accuracy of 78%. The random forest algorithm is the only model employed in this research investigation (Kang and Ryu; 2019). Under-sampling used, which minimizes the size of the data, balances the effects of pre-processing. To find the best effect to target among 55 variables, feature selection was used, and it is interesting to note that weather information is linked to determining the accident type.

2.3 Deep Learning and Machine Learning Techniques' Dependability in Answering the Research Question

The following research talks about using deep learning and machine learning to forecast injuries, their repercussions, and how to prevent them. This research study(Sarkar et al.; 2020) uses combination analysis before and after data collection to forecast injuries. Six algorithms (C5.0, SVM, RF, NB, CART, and KNN) were used to classify the severity of injuries, with 10-fold cross-validation. The study also used oversampling techniques such as KMSMOTE, BLSMOTE, SMOTE, and MWMOTE to address class imbalance. Another research (Sequeira et al.; 2021) used similar algorithms, including Catboost and XGBoost, to predict occupant injuries in the mining industry. The distance-based algorithms performed better and achieved more than 90% accuracy.

This research study (Ma et al.; 2021) uses a deep-learning model called the stacked sparse autoencoder to provide a detailed analytical framework for predicting injury severity. Cat boost is used to accelerate computation by identifying important features, and K-mean clustering is used to categorize accident data. Another study (Cuenca et al.; 2018) uses GBM, deep learning, and NB to analyze accident categorization and severity prediction. The study uses 2 hidden layers, 10 epochs, and TanH activation functions for deep learning to test the precision, F1 score, and accuracy of the algorithms. The deep learning model with 10 epoch Tanh performed better than the other algorithms. this research study (Sarkar et al.; 2019) uses ANN and SVM. The chi-square test is used to establish the importance of each feature, and random forest imputation is used to handle missing values. Pso-SVM was discovered to be the most reliable classifier, outperforming other classifiers in terms of prediction accuracy.

This research study(Zhu et al.; 2021) uses eight machine learning algorithms (NB, AutoML, LR, RF, SVM, Multi-Layer Perceptron, KNN, and DT) to predict consequences in the construction industry, with SMOTE and 10-fold cross-validation. The study found that Auto ML was 84% more accurate in classifying the data. The most important variables were found to be safety accident type, emergency, and reporting Another study (Sabet et al.; 2021) used a decision support system along with CNN and RNN-LSTM algorithms to predict accidents in the construction industry. The RNN-LSTM model performed better than the other models.

statistical comparison is performed between observed and forecasted data using a deep feedforward neural network in this research study (Oyedele et al.; 2021) the model generated the best forecast accuracy of 0.95 and Cohen Kappa 0.95. The study also conducted a sensitivity analysis on the top three models to examine how changes in settings would affect their performance. The study found that the model performed worse on the test dataset but better on the training dataset when additional neurons were added.

2.4 Limitations and Takeaway from the Literature

Section 1 research is limited to statistical analysis and researchers fail to consider all the variables that need to be analyzed. Some outcome categories are also ignored, resulting in the loss of data. Journals indicate that key challenges in the industry include rapid environmental change and a lack of confidence in technology. In the second subsection, the study only used data from one organization, and using different algorithms and evaluation procedures, as well as under-sampling, could have improved the accuracy of the predictions. The study in subsection 3 only looked at one construction company and had a small amount of data; deep learning is not necessary for small data. Overall, this section's key lesson is that there are certain machine learning assessments and approaches for more accurate prediction.

There haven't been any attempts to use deep learning and machine learning, particularly for the MSHA mining data. This research will use a variety of pre-processing techniques, feature selection, and feature engineering to search for patterns. This research attempts to explore how these predictions might be implemented in the mining industry to improve health and safety processes by classifying the degree of injury. No one has yet connected two data sets to study the injury.

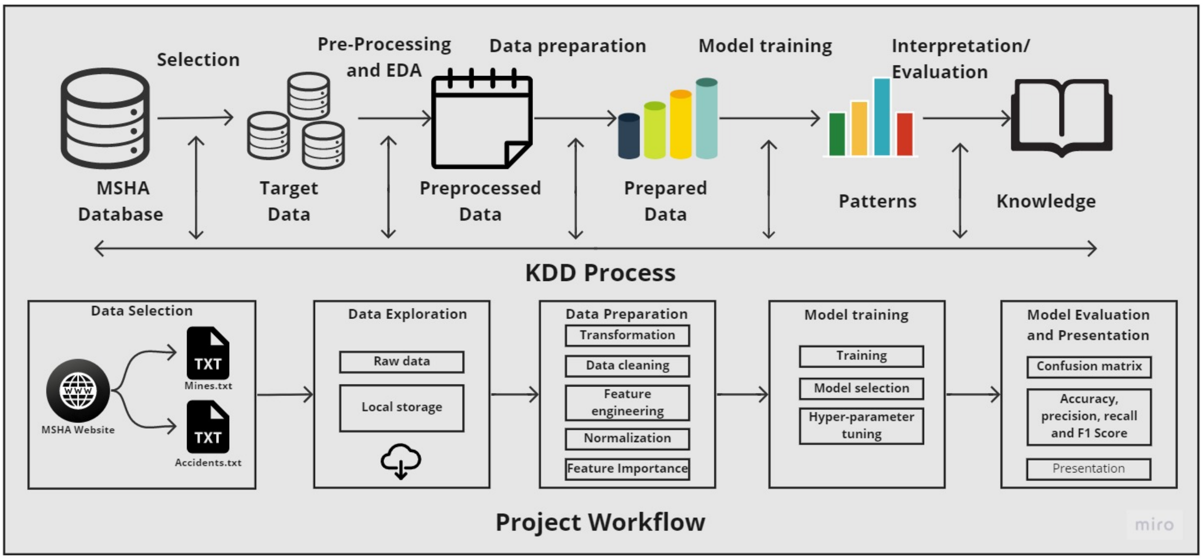


Figure 1: Project Workflow

3 Methodology

In this section, the research methodology will be covered in detail. Additionally, it will give more details on each step needed to implement the idea successfully and a technical explanation of the methodology's steps will also be given. Since our main objective is to identify the pattern of injury and its source, knowledge discovery in the database (KDD) is used. It is a useful step strategy to employ because it enables the routine discovery of legitimate, significant, and understandable patterns in vast and complex data sets (Plotnikova et al.; 2020). Figure 1 shows the steps of the same, Data selection and exploration, data pre-processing and preparation, model training, and model evaluation

3.1 Data Selection

Exploring the data and thoroughly understanding the domain are crucial steps before offering any solutions to real-world problems. Understanding the details of the domain will help to comprehend the issue and come up with a solution because every domain is unique. The main problem facing the mining industry is the high rate of injuries among workers, which was considered in this research. The degree of injuries was found to be a major concern when this was used as a starting point and explored in the injury data of the mining industry, which had text data with different numbers of columns. This helped the research to understand the requirements and to frame the research question or problem statement based on those requirements.

3.2 Exploration of the Data

In this stage, finding the relevant data according to the requirement of the process followed in the project. After finding the data next step is to identify that it is suitable to import into any programming language, For this research a zip file containing. .TXT files is fetched from the Mine Safety and Health Administration (MSHA) government website [6] two files that are suitable for the project fetched are the Accident Injury and the Mine dataset.

- Accident data: It includes data on all accidents, illnesses, and diseases that mine operators and contractors have reported from January 1, 2000, to till date.
- Mine data: All coal and metal/non-metal mines that have fallen under MSHA's purview since January 1, 1970, are listed in the Mine dataset. It contains details on the present state of each mine, the operating firm, etc. The unique key for this data is Mine ID.

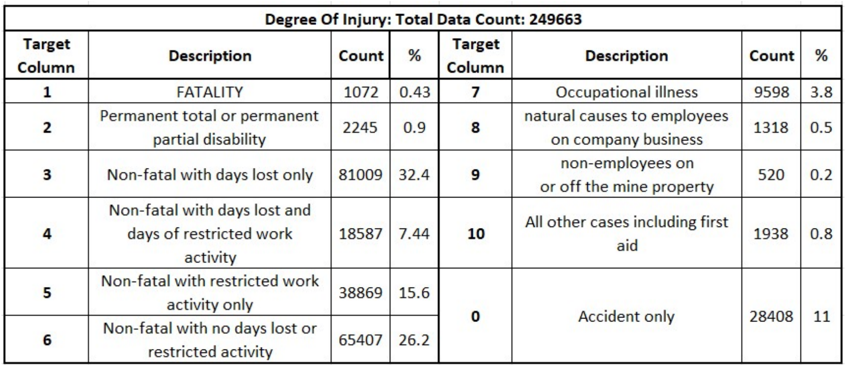
3.3 Data Preparation

The accident (Main data) contains 57 columns and the mine data contains 59 columns, the initial exploration of the data was performed using excel, and the same was visualized using Tableau software to illustrate the dataset's different variables. this was done to determine which columns are necessary for the problem statement and which columns need to be joined with what was all observed and studied in detail before loading the data into the main data frame. Since the data was obtained from a website and values from

[6] https://arlweb.msha.gov/OpenGovernmentData/OGIMSHA.asp

many organizations were combined and clubbed together to create it, there are many missing values, and the data from the year 2001 to the present would have undergone numerous variable changes and alterations, necessitating the careful study and collection of the raw data for processing. Data is fetched into the data frame using Pyspark. Apache Spark is a free and open-source platform for distributed computing that consists of several tools for real-time, big data processing. It functions essentially as a computational engine for processing huge amounts of data in parallel and batches. 29 columns from the accident dataset and 6 columns from the mine dataset were taken into consideration for the analysis. Some of the duplicate columns, some of the hugely missing values, some of the columns with IDs are neglected and only a few columns from the mine data were taken into consideration based on prior research (Amoako et al.; 2021). Accident data joined with mine data using left join.

3.3.1 Data Cleaning



Degree Of Injury: Total Data Count: 249663

Target Column	Description	Count	%	Target Column	Description	Count	%
1	FATALITY	1072	0.43	7	Occupational illness	9598	3.8
2	Permanent total or permanent partial disability	2245	0.9	8	natural causes to employees on company business	1318	0.5
3	Non-fatal with days lost only	81009	32.4	9	non-employees on or off the mine property	520	0.2
4	Non-fatal with days lost and days of restricted work activity	18587	7.44	10	All other cases including first aid	1938	0.8
5	Non-fatal with restricted work activity only	38869	15.6	0	Accident only	28408	11
6	Non-fatal with no days lost or restricted activity	65407	26.2				

Figure 2: Detail Description of Target Column

Data is further analyzed after being loaded into the data frame to ensure that the continuous column, categorical columns, and statistical values of each of the columns were observed, which allowed the data summary to be comprehended. Null values are checked and there are lots of special symbols in the data that were removed and replaced with null values for further analysis, any duplicate presence in the data is checked. Before moving to the further step, the target categories are studied thoroughly Figure 2 shows the detailed description of each category, the value zero value 0 = description says Accident only but in the PC7014 Report [7] it information is not explicitly mentioned, the Query was raised to MSHA website for the detail description of the same, as per the mail 0 signifies no accidents, so for further analysis, it is removed the categories 7,8,9, and 10 have a percentage presence in the overall statistics; they discuss illness and accidents unrelated to businesses rather than injuries, those categories are removed for further analysis. Further reading the document description it is found out that MSHA subunit code = 99 consists of injury related to the officers as the research is mainly focused on the mine labours the rows of the subunit codes are removed.

[7] https://arlweb.msha.gov/stats/part50/rptonpart50.pdf

3.3.2 Feature Engineering

The shift begins time is listed in 24-hour format, with the most frequent times being 7 a.m., 3 p.m., and 11 p.m. The outliers, such as times that are longer than 24 hours, are removed, and values that are close to the three frequent times are converted to the same to make it a useful categorical predictor variable for further analysis. The same procedure was used for the accident time, the total experience has been converted into seven categories 0-1, 1-3, 3-6, 6-10, 10-20, 20-30, and 30+ years experience to make it a useful predictor same followed by the mine experience and job experience.

Exploratory Data Analysis: The feature-engineered variables that are done above are plotted against the target variables; the pattern of the data is examined. Through the use of the Tableau and Python Visualization libraries, various additional variables are also taken into account and shown to determine the primary factor influencing the degree of injury.

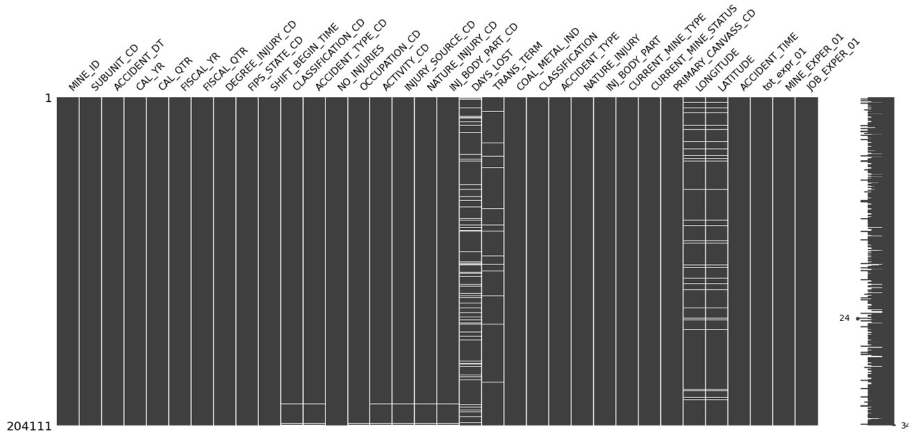


Figure 3: Missing Values Heat Map

Handling the Missing Values: Further process the data has been converted from PySpark to pandas for ease of analysis in Figure 3 shows the missing values heat map, According to the visualization, the following columns show continuous missing values. Classification, Accident Type, Employment, Activity, Injury Source, Nature Injury, and Inj Body Part. Rows are eliminated because there are only a small number of missing values in these columns (around 800). The missing data in COALMETALIND and PRIMARYCANVASS total roughly 20, those rows are removed, since most machine learning algorithms need numerical input, and missing values can cause problems, and the columns DAYS LOST and TRANS TERM have more missing values, it is not a good idea to remove those rows because doing so will cause considerable data loss. Imputing is the process of identifying missing values and filling them in. The KNN imputer is one of these imputers, which uses the training set to impute mean values from the sample's missing values' nearest neighbors. Based on the neighbors' closeness to one another, the attributes are evenly weighted or averaged. [8]. To impute those two variables, it is employed.

[8] https://scikit-learn.org/stable/modules/impute.html

Outlier Analysis: Uni-variate outlier analysis is carried out. Each column's statistical description is taken, and outliers might be present if there is a noticeable difference between 99 percent and the maximum. The columns NO INJURIES, INJ BODY PART CD, DAYS LOST, and CURRENT MINE STATUS are box-plotted, and only those with a substantial difference are considered. The outliers in mine statuses 5, 6, and 7 identify new mines, non-producing activities, and temporary idleness, respectively which have no bearing on problem statements are removed. Keeping an eye on data loss, only extreme outliers are purged. The rows are then re-indexed for further investigations.

Encoding: Now the data has been properly transformed to the proper data type, if model is given the feature variables directly, it won't be able to understand them. All independent and dependent variables, or input and output characteristics, must be numerical to be used with machines. Before we can fit our data into the model, we must first turn any categorical variables or text data in our data into numbers. The columns like TRANS TERM, COAL METAL IND, CURRENT MINE TYPE, CURRENT MINE STATUS, and experience columns are all encoded. Some of the categorical columns like OCCUPATION CD, ACTIVITY CD, and INJURY SOURCE CD have more than 100 unique values which may result in the creation of 100 to 300 extra columns resulting in a curse of dimensionality, OCCUPATION CD has 251 unique values which tells the job of the labours which can be dropped as it has many random definitions likewise activity, injury source also are dropped for further analysis, the column CLASSIFICATION, ACCIDENT TYPE, NATURE INJURY and INJ BODY PART with about 30 unique values, the top ten most frequent categories are considered and encoded, without losing any data or rows. The remaining categories will be given a zero because their presence in the data is minimal, and we also avoid losing any significant data by leaving the columns in place despite the presence of numerous unique categories. This results in the addition of 40 columns this encoding technique is inspired by the KDD cup challenge who won for doing the same [9], rest columns with a smaller number of unique values are all one hot encoded and the first column of each was dropped.

3.3.3 Normalization

The goal of the normalizing approach is to scale down the numerical columns of the dataset without distorting the range of values differences. Certain algorithms demand normalization, now that there are 89 columns in the data frame, only four of them—CAL YR, FISCAL YR, NO INJURIES, and DAYS LOST—are continuous variables, and both the financial year and the calendar year have values that are equally distributed across the previous 10 years. The number of injuries and days lost columns, however, had a sizable skewness that needed to be normalized before moving on. The box-cox method produced the best normalization curve of all the normalization techniques when plotted.

3.3.4 Feature Selection

Feature selection, which reduces the number of input variables by deleting superfluous or irrelevant features and then reducing the remaining features to only those that are essential for the machine learning algorithms, is an important stage in predictive analysis.

[9] http://proceedings.mlr.press/v7/niculescu09/niculescu09.pdf

The advantages of feature importance include a reduction in overfitting, an increase in accuracy, and a reduction in training time.

- Step 1: This is to check for the Constance feature present in the data which is not important for solving the problem statement for this variance threshold has been set to zero and run through each column, there was no Constance feature present in the data.
- Step 2: Feature selection through the Pearson correlation was performed because the data was large, and it had 88 independent columns for the target variable degree of injury. The data was split into train and test groups by 70:30, and the transformed train data was run through the correlation against the degree of injury. Because there are 89 columns and most of them are categorical, finding the important features against the degree of injury was challenging.
- Step 3: It's crucial to consider how the independent variables and dependent variables are related while doing statistical analysis. It is necessary to incorporate the independent variables in the modeling process if there is a linear relationship between these variables to produce precise predictions about the dependent variable. Eliminating multicollinearity between predictor variables is also crucial since it can diminish the model's statistical power and accuracy in how its computed coefficients are calculated. To check for multicollinearity, the correlation coefficient between each independent variable is calculated, and variables with a coefficient above 0.5 are checked, it is found that PRIMARY CANVASS CD 2 and COAL METAL IND 1, CAL QTR, and FISCAL QTR, tot expr 01 and MINE EXPER 01 are correlating more than 0.9 are removed, The Variance Inflation Factor (VIF), which determines how much a variable contributes to the standard error of the model, is also used to evaluate multicollinearity, columns with more than threshold 5 are iteratively eliminated.
- Step 4: Recursive Feature Elimination (RFE) is a feature selection technique that is used to identify the most important features in a dataset. This technique works by iteratively removing features from the dataset, building a model using the remaining features, and then ranking the features according to their significance. This process is repeated until the desired number of features is reached (Darst et al.; 2018). To improve prediction, top features are iteratively taken and incorporated into the model.

3.4 Model Training

In this called application phase, the process of deciding which model or algorithm is best suited for a given problem is known as the "model selection phase." To do this, many models and algorithms must be tested against the dataset to determine which one performs best in terms of metrics. The selected model can then be adjusted and trained again on the dataset to enhance its performance. At first, data is explored and the research question is studied thoroughly, the target variable chosen here is a degree of injury which has multiple categories, hence it's a multi-classification dataset, the algorithms that should be chosen for this should be a classifier algorithm, there are many algorithms to select for the analysis the popular and upon surveying the best algorithm in the previous research work, the algorithm random forest, XGBoost, Decision tree, K-nearest neighbor, SGD classifier and Artificial neural network with different hyperparameter tuning used for this research.

3.5 Model Evaluation and Presentation

During the machine learning phase, it is important to assess the performance of a trained model on a test dataset. This is typically done using a set of metrics that are suitable for the type of problem being tackled, such as classification. The output evaluation phase is crucial because it allows us to analyze the model's performance on unseen data and determine its ability to generalize to new situations. Additionally, it enables us to compare the effectiveness of different models and choose the most appropriate one for the current problem. Some of the common metrics used for classification problems include accuracy, precision, F1 score, recall, and confusion matrix.
The summary of all outputs, the depiction of the research's success, and a visual summary of the steps taken to answer the research questions will all be done during the conclusion phase. In tabular form, the modeling and evaluation will be shown, and an explanation of how the best result was reached will follow.

4 Design Specification

The design specification, which details the requirements, limitations, and objectives of a machine learning system, is one of the initial stages in product management. It often includes details on the methods and algorithms that will be applied as well as the system's expected performance metrics for the project are illustrated here. The execution phase of the procedure is extensively covered in this stage. The two processes in the modeling analysis are selecting the best model and then using it on the practice data. The evaluation matrices for each model are chosen in accordance with the answers to the research question.

4.1 Modelling Technique

- Random Forest Classifier(RF): An ensemble learning technique called a random forest classifier combines the predictions of multiple decision tree models to provide more accurate and reliable predictions, particularly for multiclass classification problems (Sarkar et al.; 2020) with many classes or features. Random forests offer the benefits of flexibility, feature importance, and reduced overfitting, as each decision tree is trained on a different random subset of the data, which improves generalization.

- Decision Tree Classifier(DT): The most well-known supervised machine learning method for classification issues it divides the data into the best attributes using attribute selection measures, generates a decision node for that attribute, and divides the dataset into smaller portions. Until one of the following conditions is met, there are no more attributes, all the attributes have the same value and no new occurrences. this process is repeated iteratively for each offspring: it can handle high dimension data with good accuracy; it takes less time to train than a neural network(Ajayi et al.; 2020).

- K-nearest neighbors(KNN): It is based on the idea of identifying a sample's k closest neighbors and assigning the class with the highest prevalence among them as the sample's class. The simplicity of implementation and interpretation of KNN for multiclass classification is one of its benefits. Additionally, it includes a manageable number of hyperparameters, making tuning simple (Poh et al.; 2018).

- Extreme Gradient Boosting(XGBoost): XGBoost is a widely-used ensemble learning method for multi-class classification trees. It is effective for multi-class classification (Oyedele et al.; 2021) problems with many classes and features, and offers good scalability and efficiency. XGBoost is able to train models quickly using parallel processing, automatically find the best splits in trees using a gradient-based optimization method, regularize the model to avoid overfitting, and handle large datasets with millions of samples and thousands of attributes.

- Stochastic Gradient Descent classifier(SGD): It is based on the notion of gradually altering the model's parameters in the direction of the gradient's negative sign of the loss function, which reduces the disparity between the predicted and actual labels (Riemer et al.; 2018). Rapid model training on enormous datasets is possible using the computationally effective SGD approach, even on a single system. Furthermore, it is simple to parallelize, allowing you to train models faster by using additional CPU or GPU cores.

- Artificial Neural Network(ANN): It is based on the structure and operation of the brain and can recognize intricate patterns in data by modifying the weights of the connections between the neurons(Sarkar et al.; 2020). The use of ANNs for multiclass classification has several benefits, including its capacity to handle highly skewed and high-dimensional datasets, which are common in many real-world applications. By automatically detecting the most important elements and patterns in the data, ANNs can improve in accuracy and performance. additional capability for finding nonlinear correlations in the data. Intricate patterns and structures that are difficult to model using current techniques can now be captured.

4.2 Evaluation Technique

- Accuracy: Out of all forecasts, how many were accurate is shown. In percentage form, it is derived by dividing the number of accurate forecasts by the total number of forecasts, the performance of the model cannot, however, be fully depended on for accuracy.

- Confusion matrix: The number of predictions for each class is shown in this table, which can be used to determine how well the model did for each class.

- Precision and recall: While recall is the percentage of positive samples that were anticipated to be positive, precision is the percentage of projected positive samples that were truly positive These are frequently employed in binary classification but can also be utilized for multi-classification.

- F1score: It is widely used as a single statistic to assess a classifier's performance, and it is the harmonic mean of precision and recall.

5 Implementation

There are several procedures that must be carefully planned and carried out to develop a machine learning model. To make sure the model works properly, can be deployed, and can be utilized successfully in a real-world situation, it is crucial to carefully develop and implement each stage.

5.1 Tools Used

Excel is utilized for preliminary analysis, and Tableau is used to visualize the root cause of the incidents. This research uses the Python programming language, which offers a large library of tools for everything from modelling to visualization, making it simple to program, analyse, and interpret the results.

5.2 Data Selection

In the methodology section and the draft Excel sheet, the specific data description and selection are explained. The Mines and Accident data are taken from the MSHA website, which is accessible to the public. Since neither the website nor the data-related stakeholders have requested any special authorization, this research does not contravene any moral or ethical standards. The loaded join data includes 252085 rows and 39 columns initially, with degree of injury being the target column. It has 11 categories, which are reduced to 5 to analyse only injury-related data, per the problem statement.

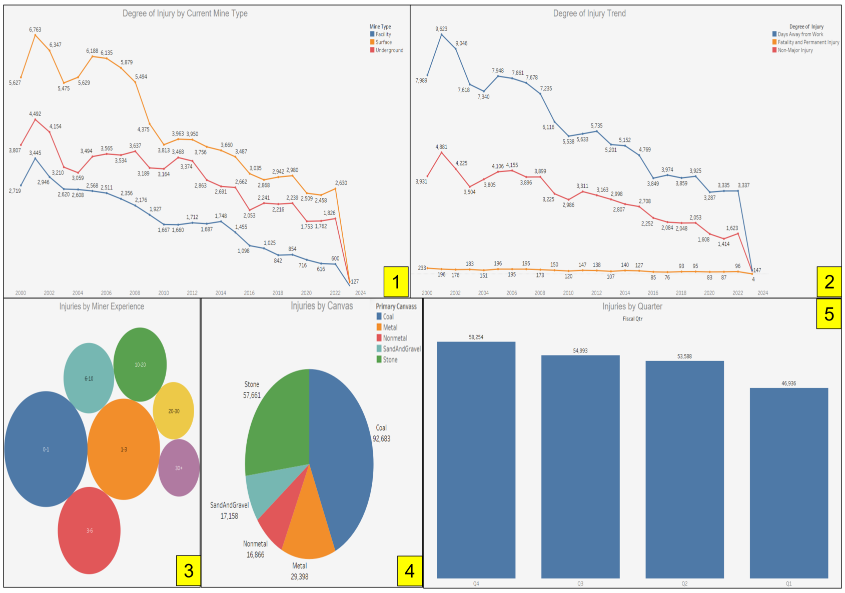


Figure 4: Dashboard of the Raw data

5.3 Exploratory Data Analysis

To find the pattern and identify the main cause of the incidents visualization is necessary, Figure 4 is drawn from the raw data, using tableau, mainly focused on the prominent variables, and plotted against the degrees of injury and number of injury. Label 1 illustrates how the injury counts have decreased over time, from the highest in the year 2001 to the lowest around 2020, by plotting the degree of injury against the various mine types from the year 2000 to the present. The fact that there are more occurrences on the surface than underground or in the facility, and that they have all remained constant across time, is another significant discovery. The levels of injuries each year are shown in label 2, where categories 1 and 2 are merged and classified as fatalities and permanent injuries, days off from work are defined as categories 3 and 4, and non-major injuries are designated as categories 5 through 7. Injury has reduced over the course of the year, while three categories have remained uniform. However, categories 3 and 4 have a larger prevalence and need extra care. Injury rates are plotted against experience in Label 3, which makes it obvious that those with less experience are more likely to be in accidents. Label 4: Injury rates are higher for workers in the coal, stone, and metal industries. According to higher injury rates at year's end in the fourth quarter, label 5 injury numbers are plotted against financial quarters.

5.4 Data Cleaning

- Handling null values: Null values are checked using IsNull() and sum() function , duplicates are check using duplicated() function.
- Handling Missing Values: The columns with a high percentage of missing data are deleted, and the KNN imputer is used to impute missing values.
- Outliers: Box plots are used to check for outliers, and only the most severe outliers are removed while keeping an eye on data loss.
- Normalization: Following graphing, box cox is used since it is more effective at normalizing skewed data than min-max scalar.
- Feature selection: It involves four steps a check for the presence of Constance features, a Pearson correlation analysis, statistical analysis, and recursive feature elimination.

5.5 Hyper parameter Tuning

The process of selecting the ideal hyperparameter values for a machine-learning model is known as hyperparameter tuning. Hyperparameters, which are parameters chosen before training, control the performance and behavior of the model. Since the values of the hyperparameters could have an effect on the model's performance, hyperparameter tuning is essential in multiclass classification. In this project, the following hyper-parameters are used.

- Random forest: GridSearchCV- n-estimators, max-depth, min samples split, min samples leaf, bootstrap.
- XG Boost: GridSearchCV- learning rate, max depth, reg lambda, n estimators.
- Decision Tree classifier: GridSearchCV- max depth, min samples leaf.
- KNN: GridSearchCV - n neighbors, weights, metric.

- ANN: The network is composed of three layers: the input layer, the hidden layer, which has 32 units and used the rectified linear unit (ReLU) activation function, and the output layer, which has the same number of units as the classes in the training data and used the softmax activation function. Utilizing the categorical cross-entropy loss function and the Adam optimizer, the model is created. Then, using a 64-batch size, it is trained for 100 epochs.

6 Evaluation

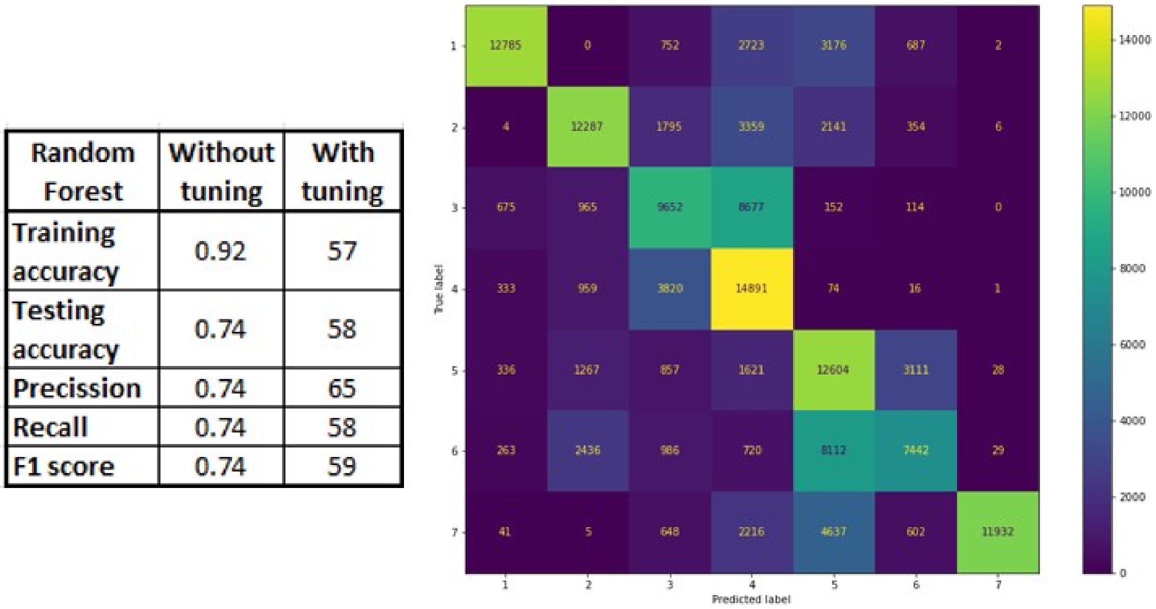
In the machine learning pipeline, evaluation is a crucial phase that helps gauge the model's effectiveness and confirm that it is operating as intended, it is crucial to carefully select the relevant assessment metrics and evaluate the model's performance.

- Class imbalance problem: Since the uneven data could have a substantial impact and could skew the modeling results, it is essential to treat the imbalanced classes while analyzing the category. Because of this, machine learning algorithms may favor the dominant class while having difficulty learning the minority class. This research makes use of a number of techniques, such as oversampling and class weighting, to deal with uneven data.

- Train and Test Split: As the last step before modelling, the dataset must be divided into test and training sets. The dataset was split into test and train sets in a 70:30 by using Sklearn package.

6.1 Case Study 1: Injury Prediction with all the Degrees of Injury Categories

After the initial preparation of the data and execution of each machine learning step, it is necessary to balance the data before moving on because there is a significant amount of imbalance in the target categories. To do this, SMOTE oversampling is used to balance the data. The model is then run through without any parameter tuning, and the training accuracy comes out to be an exceptional 92%, but when it is run through test data, the accuracy drops to 74%. This might be because of overfitting, which occurs when a model is trained too closely on the training data and finds it difficult to generalize to new, unexpected data. This can be avoided by fine-tuning the parameters involved in grid search cv tuning, which entails creating and testing a model for each combination of hyperparameters defined in a grid. The best parameter was then added to the model, matching the training and testing outcomes by decreasing 57% and 58% of accuracy, respectively, following class weight and other algorithms were performed the outcome was not satisfactory. To evaluate the reason for this a confusion matrix is plotted on the test set the Figure 5 shows, According to the confusion matrix, there is "confusion" between the numbers 3 and 4, 5 and 6, and 1 and 2 are frequently read as 3 or 4 or 6. This might be because of the resemblance between injury categories. 3 and 4, both of which result in day-loss accidents. The main distinction is that in 4 restricted activity is added, and a similar case in 5 and 6 as well. Since this is real-world raw data, and it incorporates self-enrolment data from the mining industry, there may be issues with misinterpretation in some areas. Perhaps the person in charge of reporting in one location chose how to encode, and perhaps the person in charge of reporting in another area chose to do the opposite.



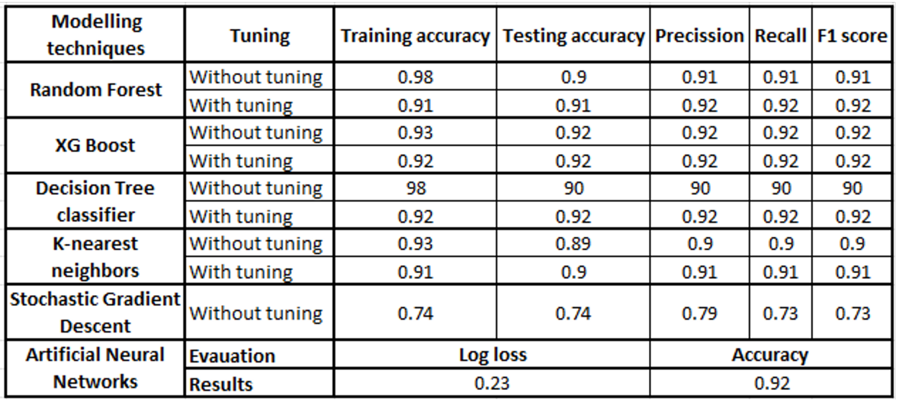
Random Forest	Without tuning	With tuning
Training accuracy	0.92	57
Testing accuracy	0.74	58
Precission	0.74	65
Recall	0.74	58
F1 score	0.74	59

True label \ Predicted label	1	2	3	4	5	6	7
1	12785	0	752	2723	3176	687	2
2	4	12287	1795	3359	2141	354	6
3	675	965	9652	8677	152	114	0
4	333	959	3820	14891	74	16	1
5	336	1267	857	1621	12604	3111	28
6	263	2436	986	720	8112	7442	29
7	41	5	648	2216	4637	602	11932

Figure 5: Model Results and Tuned confusion-Matrix of the Model

6.2 Case Study 2: Injury Prediction by Merging the two categories

By maintaining the previous analysis's supposition, categories 3 and 4 are combined, encoded to 0, and given the title "non-fatal accident with only days missed," while categories 5 and 6 are combined, encoded to 1, and given the title "non-fatal with restricted work." The analysis's findings provide the company with the information necessary to determine whether or not the workers will take a day off work as a result of their injuries. The top 20 predictor variables are taken into consideration for the analysis by means of the RFE. The target variable is now determined to be quite balanced with the value count being 1 = 94957 and 0 = 83460. and fed into the models without tuning the random forest model produced a training accuracy of 98% and testing accuracy of 90%, which is due to overfitting as it is evident that there is a slight imbalance in the target variable. This is reduced by tuning the model by using grid search CV, where the best parameters are fed into the model, resulting in a balanced outcome of test and train accuracy of 91% and 92% each in precision, recall, and recall accuracy. Like the same followed in the XG boost, Decision tree, and KNN classifiers, with the parameter tuning they yielded about 92%, 92%, and 90% of test matrix respectively, with SGD yielding the least among all results are shown in Figure 6, then for an experimental approach, the data is fed into the deep learning algorithm artificial neuron network with three layers: an input layer, a hidden layer with 32 units and ReLU activation, and an output layer with softmax activation. It is trained using categorical cross-entropy loss and the Adam optimizer with a batch size of 64 for 100 epochs produces results with 92% accuracy and the least log loss of 0.23. In comparison to other algorithms, XGboost, decision tree, and deep learning methods fared better overall. To evaluate the effectiveness of XG boost and decision tree in choosing the optimal matrix, the confusion matrix is plotted. The



Modelling techniques	Tuning	Training accuracy	Testing accuracy	Precission	Recall	F1 score
Random Forest	Without tuning	0.98	0.9	0.91	0.91	0.91
	With tuning	0.91	0.91	0.92	0.92	0.92
XG Boost	Without tuning	0.93	0.92	0.92	0.92	0.92
	With tuning	0.92	0.92	0.92	0.92	0.92
Decision Tree classifier	Without tuning	98	90	90	90	90
	With tuning	0.92	0.92	0.92	0.92	0.92
K-nearest neighbors	Without tuning	0.93	0.89	0.9	0.9	0.9
	With tuning	0.91	0.9	0.91	0.91	0.91
Stochastic Gradient Descent	Without tuning	0.74	0.74	0.79	0.73	0.73
Artificial Neural Networks	Evauation	Log loss		Accuracy		
	Results	0.23		0.92		

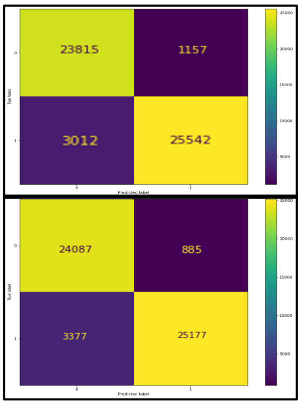
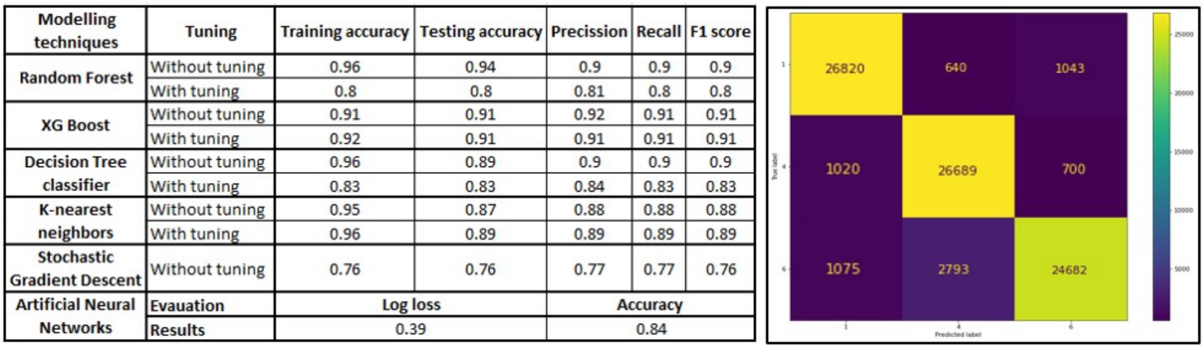


Figure 6: Model Results and Tuned confusion-Matrix of the Model XG Boost is Shown on top right and Decision tree on the bottom

project in question primarily focuses on both categories, and the AUC curve showed that XG boost outperformed the decision tree by 1%.

6.3 Case Study 3: Injury Prediction by Merging the three categories

Since the problem statement calls for a fatality prediction, the analysis was done so far, from the company's point of view, however, there is more to the analysis than determining the likelihood of a day off, which in the light of the analysis above the category 1 and 2 are merged and labeled as a fatality or permanent damage is added to the case study 2 merged target variables together now the target variables has 3 types 1,2= 1,3,4 = 4 and 5,6 = 6 with the value count of 1 =2934,4 =83460 and 6 = 94957 using the count It is evident that there is a significant imbalance in the data, and SMOTE, an oversampling technique, is employed to balance it, top 20 features are selected from the RFE, and data split in to test and run through the models, With a training accuracy of 96% and testing accuracy of 94%, the random forest model without tuning was overfitting due to an unbalanced target variable. The model's balance was enhanced using grid search CV tuning, which led to a test and train accuracy of 80% in each matrix. Due to overfitting, the decision tree classifier and KNN's training accuracy was about 95%; nevertheless, even after tuning, there was still a discrepancy of roughly 5 percent between training and testing accuracy, and SGD continued to be the least accurate predictor. Without tuning, XG boost achieved scores of 91% in both testing and training accuracy and in all other metrics. However, after tuning, the training accuracy went up by 1%, making it the best predictor overall in Case Study 3. The confusion matrix of the tuned XGBoost model shows that it correctly predicted 26995 examples of ”fatality”, 26708 examples of ”small injury and day off”, and 24784 examples of ”small injury no day off”. However, it also made some incorrect predictions, such as incorrectly predicting 528 examples of ”small injury and day off” as ”fatality” and 2713 examples of ”small injury no day off” as ”small injury and day off”.



Modelling techniques	Tuning	Training accuracy	Testing accuracy	Precission	Recall	F1 score
Random Forest	Without tuning	0.96	0.94	0.9	0.9	0.9
	With tuning	0.8	0.8	0.81	0.8	0.8
XG Boost	Without tuning	0.91	0.91	0.92	0.91	0.91
	With tuning	0.92	0.91	0.91	0.91	0.91
Decision Tree classifier	Without tuning	0.96	0.89	0.9	0.9	0.9
	With tuning	0.83	0.83	0.84	0.83	0.83
K-nearest neighbors	Without tuning	0.95	0.87	0.88	0.88	0.88
	With tuning	0.96	0.89	0.89	0.89	0.89
Stochastic Gradient Descent	Without tuning	0.76	0.76	0.77	0.77	0.76
Artificial Neural Networks	Evauation	Log loss		Accuracy		
	Results	0.39		0.84		

	1	4	6
1	26820	640	1043
4	1020	26689	700
6	1075	2793	24682

Figure 7: Model Results and Confusion-Matrix of the Best Model XG Boost.

6.4 Discussion

In this research, five machine learning algorithms and one deep learning algorithm were used to classify the degree of injury in the mining industry. In the first case study, the problem of class imbalance was encountered, and various iterations were performed to address this issue, including oversampling, class weight, normalizing, and different feature selection and hyperparameter techniques, but these did not produce satisfactory results. Upon cross-verifying the confusion matrix, it was found that there was a pattern of misclassification between similar categories, potentially due to incorrect entries in the dataset. Therefore, in the second case study, similar categories were merged, and only binary classification was performed, resulting in producing almost 92% accuracy using XGBoost, decision tree, and ANN algorithms. In the third case study, the fatality and permanent damage categories were merged with the case two categories and SMOTE and hyperparameter tuning was used to address the class imbalance and overfitting, resulting in producing 91% accuracy using the XGBoost algorithm. Overall, the project was successful in addressing the research question and predicting the degree of injury in the mining industry.

In this project, it was assumed that there may be incorrect entries in the dataset due to its nature as a collection of self-reported incidents from across the mining industry in the United States. This assumption was based on the results of the first case study, where a pattern of misclassification between similar categories was observed. The results of this research showed that multi-classification using machine learning and deep learning algorithms can accurately predict the degree of injury in the mining industry. This information can be used by companies to better predict and prevent injuries and to improve their health and safety measures. Additionally, by automating the target column and avoiding incorrect entries, companies can more accurately assess their potential fines and develop appropriate mitigation plans.

7 Conclusion and Future Work

The aim of this research is to use machine learning and deep learning algorithms to predict the degree of injury in the mining industry and explore potential applications for improving health and safety measures. The data was collected from the Mine Safety and Health Administration (MSHA) website from the year 2000 to the present. For the preliminary analysis, Excel and Tableau are utilized, and only important columns are taken into consideration. Apache pyspark was used for the initial data loading and pre-processing, followed by pandas for further pre-processing, such as handling null values, imputing missing values, removing outliers, encoding multi categories variables, and performing statistical analysis. Recursive feature selection was used to select important variables and the data was normalized using min-max and box-cox scaling before being fed to the model. Five machine learning algorithms and a deep learning algorithm were used to predict the degree of injury. Several approaches were used in the first case study to address the issue of class imbalance, but the outcomes were unsatisfactory. Further investigation revealed that the data had a pattern of misclassification between related categories. In the second case study, similar categories were combined, binary classification was carried out, and XGboost, decision tree, and ANN algorithms produced results with about 92% accuracy. In the third case study, additional death and permanent injury were combined and added to the case study two categories. To solve class imbalance and overfitting, approaches including SMOTE and hyperparameter tweaking were applied, which led to 91% accuracy using the XGBoost algorithm and helped the research reach its goals.

In future research, this work can be expanded by exploring different machine learning and deep learning techniques on the MSHA dataset. This could include adding additional variables by combining different datasets available on the website, as well as experimenting with different ways of handling missing values and encoding multi-category variables. Additionally, using tools like Tableau for data visualization can help to further analyze the data and identify the root causes of incidents. By taking these steps, it may be possible to achieve even better results and gain a deeper understanding of the data.

8 Acknowledgement

I am extremely thankful to Professor Christian Horn for his valuable expertise and guidance during my research. His insights were crucial in shaping the direction of the work, I also want to express my appreciation to the National College of Ireland for providing the opportunity to conduct this research. I am thankful to my cousin Jagruthi, my family, and my friends for their support and encouragement during my pursuit of a master's degree. Their support has been instrumental in helping me complete this research project.

References

Ajayi, A., Oyedele, L., Akinade, O., Bilal, M., Owolabi, H., Akanbi, L. and Delgado, J. M. D. (2020). Optimised big data analytics for health and safety hazards prediction in power infrastructure operations, *Safety science* **125**: 104656.

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.

AlMamlook, R. E., Kwayu, K. M., Alkasisbeh, M. R. and Frefer, A. A. (2019). Comparison of machine learning algorithms for predicting traffic accident severity, *2019 IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT)*, IEEE, pp. 272–276.

Amoako, R., Buaba, J. and Brickey, A. (2021). Identifying risk factors from msha accidents and injury data using logistic regression, *Mining, Metallurgy & Exploration* **38**(1): 509–527.

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.

Cuenca, L. G., Puertas, E., Aliane, N. and Andres, J. F. (2018). Traffic accidents classification and injury severity prediction, *2018 3rd IEEE International Conference on Intelligent Transportation Engineering (ICITE)*, IEEE, pp. 52–57.

Darst, B. F., Malecki, K. C. and Engelman, C. D. (2018). Using recursive feature elimination in random forest to account for correlated variables in high dimensional data, *BMC genetics* **19**(1): 1–6.

Ganguli, R., Miller, P. and Pothina, R. (2021). Effectiveness of natural language processing based machine learning in analyzing incident narratives at a mine, *Minerals* **11**(7): 776.

Gendler, S. and Prokhorova, E. (2021). Risk-based methodology for determining priority directions for improving occupational safety in the mining industry of the arctic zone, *Resources* **10**(3): 20.

Hyder, Z., Siau, K. and Nah, F. (2019). Artificial intelligence, machine learning, and autonomous technologies in mining industry, *Journal of Database Management (JDM)* **30**(2): 67–79.

Jung, D. and Choi, Y. (2021). Systematic review of machine learning applications in mining: Exploration, exploitation, and reclamation, *Minerals* **11**(2): 148.

KALE, Ö. A. and Baradan, S. (2020). Identifying factors that contribute to severity of construction injuries using logistic regression model, *Teknik Dergi* **31**(2): 9919–9940.

Kang, K. and Ryu, H. (2019). Predicting types of occupational accidents at construction sites in korea using random forest model, *Safety Science* **120**: 226–236.

Ma, Z., Mei, G. and Cuomo, S. (2021). An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors, *Accident Analysis & Prevention* **160**: 106322.

Matloob, S., Li, Y. and Khan, K. Z. (2021). Safety measurements and risk assessment of coal mining industry using artificial intelligence and machine learning, *Open Journal of Business and Management* **9**(3): 1198–1209.

Oyedele, A., Ajayi, A., Oyedele, L. O., Delgado, J. M. D., Akanbi, L., Akinade, O., Owolabi, H. and Bilal, M. (2021). Deep learning and boosted trees for injuries prediction in power infrastructure projects, *Applied Soft Computing* **110**: 107587.

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

Poh, C. Q., Ubeynarayana, C. U. and Goh, Y. M. (2018). Safety leading indicators for construction sites: A machine learning approach, *Automation in construction* **93**: 375–386.

Riemer, M., Cases, I., Ajemian, R., Liu, M., Rish, I., Tu, Y. and Tesauro, G. (2018). Learning to learn without forgetting by maximizing transfer and minimizing interference, *arXiv preprint arXiv:1810.11910* .

Sabet, M. F. A., Dahroug, A. and Hegazy, A. F. (2021). A proposed model for field workers injuries' prevention based on machine learning, *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*, IEEE, pp. 383–388.

Sarkar, S., Pramanik, A., Maiti, J. and Reniers, G. (2020). Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data, *Safety science* **125**: 104616.

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.

Sequeira, G. J., Lugner, R., Brandrneier, T., Elnagdy, E., Danapal, G. and Jumar, U. (2021). Investigation of different classification algorithms for predicting occupant injury criterion to decide the required restraint strategy, *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, IEEE, pp. 204–210.

Shekarian, Y., Rahimi, E., Shekarian, N., Rezaee, M. and Roghanchi, P. (2021). An analysis of contributing mining factors in coal workers' pneumoconiosis prevalence in the united states coal mines, 1986–2018, *International Journal of Coal Science & Technology* **8**(6): 1227–1237.

Zhu, R., Hu, X., Hou, J. and Li, X. (2021). Application of machine learning techniques for predicting the consequences of construction accidents in china, *Process Safety and Environmental Protection* **145**: 293–302.