**ANL 488 PROJECT PROPOSAL**

**Using Machine Learning to Uncover Insights**

**related to safety incidents (Mining Industry)**



**Submitted by Connie Tan @ Zann Tan**

**SCHOOL OF BUSINESS**

**Singapore University of Social Sciences**

**Presented to Singapore University of Social**

**Sciences in partial fulfilment of the**

**requirements for the**

**Degree of Bachelor of Science**

**in Business Analytics**

**2023**

# **ABSTRACT**

Safety is a top priority for all organizations in the world, however, safety is an issue, especially in the mining industry, where high injuries and fatalities rates were observed. This paper adopts the CRISP-DM framework to apply supervised machine learning methods like Decision Trees, Random Forests, Neural Networks, Support Vector Machines and Logistic Regression to uncover insights related to mining incidents. An interesting point to note is that this paper uses a multi-class target.

The analysis was performed with seven datasets provided by the mining organization which had mining operations in Australia and Canada. Data understanding and preparations were conducted to ensure the integrity of the data and the usefulness of modelling results. Additionally, statistical analysis like the Chi-square test and Cramer’s V has been conducted to ensure variables used were statistically significant and associated.

The champion model was identified as a Decision Tree with SMOTE which was evaluated through seven metrics like training and testing accuracy, precision, recall, F1 score, ROC curve and Kappa score. Therefore, the findings of this paper would provide useful recommendations for the mining organization to improve safety within the work environment through actionable safety controls to achieve zero injuries at work.

Table of Contents

[ABSTRACT 1](#_Toc149751087)

[Chapter 1: Business Understanding 3](#_Toc149751088)

[Chapter 1.1: Introduction 3](#_Toc149751089)

[Chapter 1.2: Business Problem & Objective 4](#_Toc149751090)

[Chapter 1.3: Data Mining Goal 5](#_Toc149751091)

[Chapter 1.4: Software 6](#_Toc149751092)

[Chapter 2: Literature Review 6](#_Toc149751093)

[Chapter 3: Data Understanding & Preparation 13](#_Toc149751106)

[Chapter 3.1: Data Understanding 13](#_Toc149751107)

[Chapter 3.1.1: New Data 14](#_Toc149751108)

[Chapter 3.2: Data Preparation 16](#_Toc149751109)

[Chapter 3.2.1: Data Cleaning 16](#_Toc149751110)

[Chapter 3.2.2: Data Selection 17](#_Toc149751111)

[Chapter 3.3: Data Specification 19](#_Toc149751112)

[Chapter 4: Modelling and Evaluation 21](#_Toc149751113)

[Chapter 4.1: Modelling 21](#_Toc149751114)

[Chapter 4.2: Evaluation 29](#_Toc149751115)

[Chapter 5: Recommendations, Deployment and Conclusions 30](#_Toc149751116)

[Chapter 5.1: Recommendations and Deployment 30](#_Toc149751117)

[Chapter 5.2: Conclusions 31](#_Toc149751118)

[References 32](#_Toc149751119)

[Appendix 38](#_Toc149751120)

# **Chapter 1: Business Understanding**

## Chapter 1.1: Introduction

|  |  |
| --- | --- |
| Mining is used to extract minerals from Earth, supplying the world with essential resources and creating employment opportunities. For example, coal mining created 6.5 million jobs worldwide (World Coal Association, 2023). Although mining offers employment advantages, miners often work in harsh work environments. Safe Work Australia (2022) recorded 11.2 serious claims per thousand workers and a total of 2,806 serious claims resulting in a median compensation of $15,072 per claim in the mining industry during 2021. Despite the hazardous environment in the mining sector, Manjunatha (2023) discovered only 2 out of 109 papers presented findings on occupational safety research suggesting safety is not widely studied. Alkaissy et al. (2023) cited insufficient injury data due to under-reporting, different recording methods and definitions (Safe Work Australia, 2013) including poor data quality were reasons for the lack of safety research.  *Figure 1: Location of mining sites in Australia and Canada* | |
|  |  |

For this study, datasets from a mining organization operating in Australia and Canada were used, illustrated in Figure 1. Additionally, Australian and Canadian weather data sourced from Kaggle were incorporated into the study to determine if environmental factors would result in a safety incident.

With insufficient safety knowledge from research, miners are constantly at risk of experiencing a safety incident, leading to detrimental effects on physical and mental well-being. Hence, this paper deploys machine learning alongside the CRISP-DM framework to investigate why safety incidents result in an injury to enhance safety performances in the mining industry through effective safety controls.

## Chapter 1.2: Business Problem & Objective

Miners are constantly exposed to different hazards at work ranging from chemical, ergonomic, health, physical, psychosocial, and safety (Government of Canada, 2023). For example, gas hazards, musculoskeletal disorders, lung diseases, heat stress, fatigue, confined spaces, and work at height (Queensland Government, 2023). To mitigate the risk of exposure to hazards, the government enacts legislation to protect miners. For instance, in Canada, health and safety is governed by the “Occupational Health and Safety Act” in Ontario (Ontario’s Regulatory Registry, 2021). In Australia, health and safety is governed by the “Mining and Quarrying Safety and Health Act 1999” in Queensland (Resources Safety & Health Queensland, 2023) and the “Mines Safety and Inspection Act 1994” in Western Australia (Government of Western Australia, 2022).

Despite legislation, audits through inspection checks, best practices, and guidelines in Australia and Canada, mine operators struggle to achieve zero injuries in the workplace. For example, Figure 2 illustrates injury rates in Australia and Canada during 2022. Notably, Ontario and Australia reported 1616 (Workplace Safety North, 2022) and 2806 (Safe Work Australia, 2022 injuries respectively. Failing to achieve zero injury rates highlights the importance of predictive safety being an iterative process to improve safety for miners to ensure a constant supply of minerals. Hence, the business problem centres around addressing injuries in mines resulting in lost-time injuries (LTI). The business objective is to enhance the safety of miners in Australia and Canada by striving to attain zero injuries at work.

*Figure 2: Number of injuries in Australia and Canada in 2022*

A graph of a number of injuries

Description automatically generated

## Chapter 1.3: Data Mining Goal

The data mining goal is to predict factors likely to cause injuries in mines using the given dataset. With new insights identified through machine learning beforehand, mining companies can take proactive actions to prevent injuries, thereby, achieving zero injuries, and keeping the workplace safe for everyone.

## Chapter 1.4: Software

This study utilizes four software to facilitate the research. Mainly, Excel and IBM SPSS for initial data exploration, Tableau for illustration, and Python for data preparation, modelling, and evaluation to assess the suitability of datasets for analysis. Jupiter notebooks containing Python codes are split into 3 files namely “Data Cleaning Codes.ipynb” for data preparation, “Canada Modelling Codes.ipynb” and “Australia Modelling Codes.ipynb” for modelling and evaluation.

# **Chapter 2: Literature Review**

Due to the complexity of different hazards present, human factors, tasks, and parties involved in safety incidents, Lee et al. (2020) deployed a three-step framework to refine variables in the final dataset for modelling incident prediction.

Firstly, two methods, latent class cluster analysis (LCCA) and the chi-square test were applied for comparison. LCCA, a clustering technique, suitable for categorical, numeric, and binary was applied to determine the optimum number of variables through an iterative process of experimenting with different values of “K” (latent class) by assessing the goodness of fit like Akiake information criteria”, “Bayesian information criteria”, “consistent AIC” and “R-squared” (Lee et al., 2020). Thereafter, Lee et al. (2020) applied the chi-square test to validate the findings presented by LCCA. For example, using LCCA, 130 variables in the original dataset were reduced to 7 variables in the final dataset, signifying a reduction of 94% of the variables from the original dataset. In this paper, 61 variables were documented within the seven datasets provided. After applying the 4-step data cleaning framework in Chapter 3.2.1, the chi-square test was conducted to ascertain the relevance of the variables to be used for modelling. Within the Australian dataset, 10 variables were statistically significant while 1 was not statistically significant. Whereas, within the Canada dataset, 9 variables were statistically significant while 2 were not statistically significant, as illustrated in Chapter 3.2.2. Consequently, achieving a reduction of 81% in variables.

Secondly, Lee et al. (2020) applied a support vector machine (SVM) and decision trees (DT) with an ensemble to evaluate the number of categories within each variable through predictive performance. The research concluded ten categories per variable was the optimum number to analyse safety datasets. Following Lee et al. (2020)’s findings, this paper adheres to a maximum of ten categories per variable. Type of Occurrence Classification System (TOOCS) established by the Australian Safety and Compensation Council (2004) is incorporated to reduce the number of categories in the following variables: “AgencyOfInjury”, “BodyPart”, “MechanismOfInjury” and “NatureOfInjury”. Table 1 describes a detailed explanation of the different classification systems used.

*Table 1: Detailed explanation of TOOCS*

|  |  |
| --- | --- |
| TOOCS classification | Explanation |
| Agency of injury | Recognizes the primary cause (object, substance, or circumstances) associated with the safety incident and respective causes directly responsible for the injury or disease |
| Body Location | Precisely identifies the injured body part |
| Mechanism of injury | Pinpoint the situation leading to the injury or disease |
| Nature of injury | Determine the injury or disease experienced by the worker including mental illness |

Lastly, Lee et al. (2020) performed correlation analysis and principal component analysis to prepare the final dataset. In this paper, Cramer’s v will be applied to categorical and numerical variables in the final dataset to measure the strength of association between the variables and target respectively, presented in Chapter 3.2.2. Table 2 provides a summary of factors influencing safety incidents from seven literature reviews. Also, comparisons with the given datasets were conducted to determine if similar data were available for this study.

*Table 2: Summary of factors influencing safety incidents from literature reviews*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variables/  Paper | Datasets provided | Stemn and Krampah (2022) *“Injury severity and influence factors in surface mines: A correspondence analysis”* | *Butani (1988)*  *“Relative risk analysis of injuries in coal mining by age and experience at present company”* | Asare-Doku et al. (2022) *“Mental health and workplace factors: comparison of the Ghanaian and Australian mining industry”* | Stemn and Benyarku (2023)  *“Mineworkers’ perspective of fatigue: A study of the Ghanaian mining industry”* | Muzaffar et al. (2013)  *“Factors associated with fatal mining injuries among contractors and operators”* | Dumrak et al. (2013)  *“Factors associated with the severity of construction accidents: The case of South Australia”* | Vlachos (2019)  *“INTERRELATION BETWEEN OCCUPATIONAL HEALTH & SAFETY LEADING AND LAGGING INDICATORS IN MINING INDUSTRY. AN EMPIRICAL STUDY”* |
| Working hours | ✓ | ✓ |  |  |  | ✓ |  |  |
| Days away from work | ✓ | ✓ |  |  |  |  |  |  |
| Time of accidents |  | ✓ |  |  |  |  |  |  |
| Work experience | ✓ | ✓ | ✓ |  |  | ✓ | ✓ |  |
| Mental health status |  |  |  | ✓ | ✓ |  |  |  |
| Age |  | ✓ | ✓ |  |  |  | ✓ |  |
| Employment Type | ✓ |  |  |  |  | ✓ |  |  |
| Mine type |  |  |  |  |  | ✓ |  |  |
| Gender |  |  |  |  |  |  | ✓ |  |
| Size of organization |  |  |  |  |  |  | ✓ |  |
| Project size |  |  |  |  |  |  | ✓ |  |
| Mechanism of incident | ✓ |  |  |  |  |  | ✓ |  |
| Injury location | ✓ |  |  |  |  |  | ✓ |  |
| Nature of injury | ✓ |  |  |  |  |  | ✓ |  |
| Native language |  |  |  |  |  |  | ✓ |  |
| Safety indicators |  |  |  |  |  |  |  | ✓ |

To begin, Stemn and Krampah (2022) investigated the association between injury severity and accident factors (who, what, where, and why) because most studies focus either on fatalities or injuries without understanding how different accidents result in different injury severity. The research revealed influential injury severity factors including work experience, days away from work, and timing of accidents. Namely, young, and less experienced miners and night shifts were hidden causes of safety incidents. Indeed, similar findings were also derived by Butani (1988). Furthermore, frequent occurrences of safety incidents were during morning shifts, specifically between 12 pm to 5.59 pm, the second 6 hours of a 12-hour shift. Hence, understanding the relationship between incident factors and different injury severity helps to formulate effective safety strategies to reduce safety incidents. Additionally, another research by Dumrak et al. (2013) emphasizes factors including demographics (age, gender, native languages spoken), work experience, and environments that influence injury severity. For example, the proportion of injuries increases with age and experience. New and young workers were more prone to suffer from severe accidents due to lack of experience while older and experienced workers were more prone to fatal accidents due to overconfidence. Moreover, workers who did not speak English were also at higher risk of critical injuries due to the inability to understand safety training and manuals. Besides, the study explains how the mechanism of injury and injury location could help to further explain injury severity, leading to more effective use of tools, equipment, and materials.

Next, Asare-Doku et al. (2022) highlight mineworkers’ mental and social well-being are important factors in mitigating the risk of safety incidents. In the research, poor mental health status was attributed to long work hours, labour-intensive tasks, and environmental factors, resulting in fatigue and poor judgement, leading to safety incidents. Indeed, Stemn and Benyarku (2023) support the research that fatigue is a contributing factor to increased safety incidents. Besides, Muzaffar et al. (2013) explored factors associated with fatal incidents across employment types and concluded work experience, working hours, and mine type (surface or underground) were crucial to determining fatalities. For instance, statistically significant associations were observed between fatalities and contractors, lesser work experience, and more than 8 hours of work at surface mines.

Finally, Vlachos (2019) challenges the accuracy of safety indicators used by mining organisations. The research reveals most safety indicators used by mining organizations serve as lagging indicators measuring past events instead of leading indicators which measure potential safety issues or areas requiring attention. Leading and lagging indicators are equally important because lagging indicator helps to assess the effectiveness of leading indicators in measuring safety performances. In particular, maintenance data serves as a leading indicator, to predict when a machine malfunctions and issues "warning alarms" when servicing is required (Benson, 2023). Consequently, Towsey (2011) questions the relevance of parameters used by organizations to represent safety. An example is the common use of lost-time injury frequency rates (LTIFR) to evaluate safety performance. Nevertheless, absolute measures of fatalities are preferred because analysis could be skewed to believe safety has improved due to fatalities being a rare event. Likewise, Safe Work Australia (2013) argues the use of LTI correlates negatively with influencing work injury and illness factors, resulting in the usage of ineffective measure safety performances. To illustrate, LTI with less severe injuries are more frequent than LTI with more severe injuries, potentially leading to a misleading impression of improved safety. Despite the challenges associated with LTI, organizations continue to adopt LTIFR to determine safety performance. Hence, Safe Work Australia (2013) recommends a severity framework that comprises injuries based on the mineworker’s impact.

*Figure 3: Injury severity framework*

Especially in this study, the framework to determine the severity of injury consists of lost time and injury classification as seen in Figure 3. Lost time is categorized as either “True” or” False” and injury classification comprising of “First Aid Injury”, “Lost Time Injury”, “Medical Treatment Injury”, “No Treatment”, “Non-Work”, “Occupational Injury/Illness” and “Restricted Work Injury’. Figure 4 presents a breakdown of severity. It is noted that “Critical”, “Severe” and “Serious” accounts for 2.21%, 4.82% and 6.53% respectively are considered rare events.

Overall, based on the reviewed literature, a combination of the methodology illustrated in Figure 5 and machine learning techniques explains why safety incidents result in an injury.

*Figure 4: Proportion of severity*

A screenshot of a computer

Description automatically generated

*Figure 5: Methodology created based on reviewed literature*

# **Chapter 3: Data Understanding & Preparation**

## Chapter 3.1: Data Understanding

*Table 3: Summary of given datasets*

|  |  |  |  |
| --- | --- | --- | --- |
| Filename | Description of dataset | Rows | Columns |
| safety\_events.csv | Significant incident and reportable safety incidents from 2000 to 2022 | 6970 | 35 |
| production\_data.csv | Daily production data (tonnes) for each mine site from 2016 to 2023 | 7047 | 6 |
| emp\_start\_end\_dates.csv | Employment history from 2016 to 2022 | 1,048,575 | 3 |
| person\_workgroup.csv | Workgroup an employee belongs to during employment from 2020 to 2022 | 28856 | 4 |
| employee\_roster.csv | Roster and leave records of employees from 2019 to 2023 | 1,048,575 | 6 |
| labour\_hours\_worked.csv | Monthly labour hours worked by contractor/staff from 2011 to 2022 | 2078 | 4 |
| site\_location.csv | Location of mine sites | 10 | 3 |
|  | Total |  | 61 |

Seven datasets were provided, comprising 61 columns with more than 2 million rows of data, ranging from years between 2000 to 2023, illustrated in Table 3. During preliminary checks, data exploration was conducted using Excel and IBM SPSS, to identify any irrelevant columns, invalid data, and records of missing or duplicate values. To illustrate, invalid columns represent columns with duplicate information like “LTIDays” and “LTI” which conveyed similar information in different formats, numeric and flag respectively. Also, invalid data contains an invalid date format like “2999-01-01T00:00:00Z”. A detailed observation of data quality issues during initial data exploration is in Appendix, Table 12. Datasets are also combined to gather more information about employees. For example, invalid start dates such as “1990-01-01” in “person\_workgroup.csv” are replaced with “start\_dates” from “emp\_start\_end\_dates.csv” using “name\_hash” to determine the employee’s actual start date in the workgroup. Lastly, an illustration in Figure 6 aids in comprehending the relationship among the seven datasets provided. For example, merging “safety\_events.csv” and “site\_location.csv” using “site\_key\_hashed” provides the location of the mining site.

*Figure 6: An illustration of how the seven datasets are inter-linked*

A screenshot of a computer screen

Description automatically generated

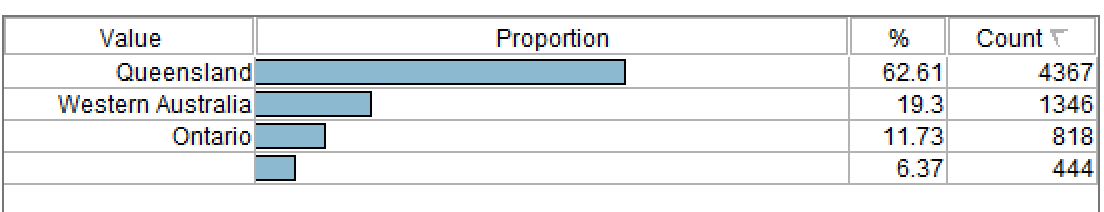
### Chapter 3.1.1: New Data

Using Kaggle, two new datasets on Australian weather data from Young & Young (2020), Canadian weather data from Turner (2020) and coordinates of mining locations in Ontario from the Ontario Ministry of Mines (2023) were incorporated to assess the impact of weather conditions on safety incidents. Firstly, the weather data retrieved includes daily records of rainfall and temperature alongside the respective cities in Australia and Canada. To illustrate in Figure 7, a methodical approach was used to select the nearest weather station in Perth, Brisbane, Townsville, and Ottawa to provide a better representation of the weather conditions experienced by mineworkers.

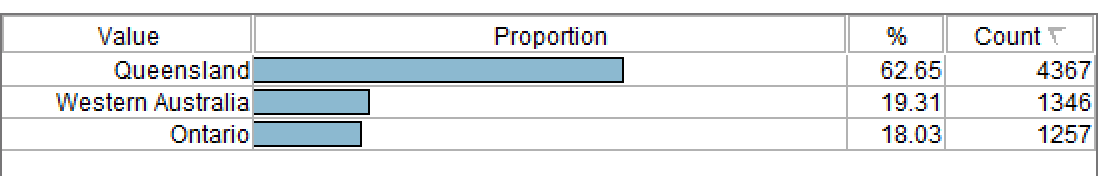
|  |  |
| --- | --- |
| *Figure 7: Location of mining sites in Australia and Canada* | |
|  |  |

Secondly, weather data obtained from Kaggle was merged into “site\_location.csv” through Python, with the unique key: “Region”. Subsequently, “site\_location.csv” was merged again with “safety\_events.csv” using “site\_key\_hashed” to derive an accurate representation of the actual weather conditions during the day of the incident. Lastly, coordinates of additional mining locations in “site\_location.csv” are merged using “site\_key\_hashed”. From Figure 8, 6.37% of missing values were replaced with Ontario mining location to balance the final dataset because Ontario contains only 11% of records, while Western Australia and Queensland contain 19.3% and 62.62% of records respectively. After merging, Ontario’s records contain 18.03% illustrated in Figure 9.

*Figure 8: Proportion of Ontario’s records (Before merging)*



*Figure 9: Proportion of Ontario’s records (After merging)*



# **Chapter 3.2: Data Preparation**

## Chapter 3.2.1: Data Cleaning

In this study, data cleaning adopts a four-step framework presented in Figure 10 to ensure data quality. Firstly, based on preliminary checks, irrelevant columns were removed. For example, ‘AgencyOfInjuryId’, ‘AgencyOfInjuryDescription’ and ‘InjuryTypeCode’.

Secondly, data cleaning was performed using Python on nine datasets including seven provided datasets and one weather dataset from Kaggle. The objective was to rectify invalid dates, eliminate duplicates, replace missing values, and remap the following variables: “AgencyOfInjury”, “BodyPart”, “Injury”, “MechanismOfInjury”, “NatureOfInjury”, “leave\_type” to a maximum of ten categories. For example, 36 categories within leave\_type from “employee\_roster.csv” were remapped to five categories. Also, columns containing “Not yet assessed” are removed from the dataset because it does not provide any information on the injury sustained from the safety event.

*Figure 10: Data preparation framework*

Thirdly, merging was executed using unique keys to incorporate human and organisational factors which could explain an occurrence of the safety incident. However, after merging the respective datasets, more missing (N/A) values were present. For instance, a merge using “name\_hash” between “employee\_roster.csv” and “person\_workgroup.csv” to understand the number of employees on leave within a workgroup during the time of the incident resulted in more missing values than the original dataset, depicted in Appendix, Table 13. Subsequently, similar variables like “Annual leave” containing 99.78% of N/A values are non-informative for this study were removed to prepare the final dataset.

Fourthly, a series of new columns including “Incident\_time\_period”, “Same\_Date\_Reporting” and “Severity” was derived. For example, “Incident\_time\_period” is derived from the time of the incident to determine if there were significant findings relating to safety incidents. Also, “Severity” was calculated from a combination of “LostTime” and“Injury”, presented in Table 4. Also, ‘Critical’, ’Severe’ and ‘Serious’ belongs to a minority class illustrated in Figure 4, hence, data is merged into 1 column, “Serious”.

*Table 4: Classification of severity*

|  |  |  |
| --- | --- | --- |
| LostTime | Injury | Severity |
| False | * Non-work * No Treatment | Minor |
| False | * Occupational Injury/illness * First Aid Injury | Moderate |
| True | * Medical Treatment Injury * Restricted Work Injury * Lost Time Injury | Serious |

## Chapter 3.2.2: Data Selection

The chi-square test and Cramer’s V are applied to the cleaned dataset comprising 11 variables to determine variables exhibiting significant correlation for modelling. Both the chi-square test and Cramer’s V are statistical techniques to assess if any significant association exists (Watson, 2022) and the strength of association respectively between categorical variables (Santos, 2023).

Chi-square test

The chi-square test is applied to the independent variables (X) against the dependent variable (Y), “Severity” to assess if there were any significant associations between the variables using hypothesis testing with an alpha of 0.05. For example, the null hypothesis, H0: AgencyOfInjury and “Severity” is independent and the alternate hypothesis, H1: AgencyOfInjury and “Severity” is not independent is defined. To determine if the null hypothesis is rejected, the p-value is used, and the null hypothesis is rejected if the p-value is less than alpha. For instance, AgencyOfInjury has a p-value of 8.857403318006178e-17 which is lesser than alpha, hence, the null hypothesis is rejected. Therefore, there is a significant association between the AgencyOfInjury and Severity. To summarise in Table 5, all independent variables are significantly associated with “Severity” except for “StaffContractor” within the Australia dataset and “Same\_Date\_Reporting” and “Incident\_time\_period” within the Canada dataset which will be dropped from the final dataset.

*Table 5: Chi-square test results*

|  |  |  |
| --- | --- | --- |
| Variables | Chi-square results  (Australia dataset) | Chi-square results  (Canada dataset) |
| AgencyOfInjury | Statistically significant | Statistically significant |
| BodyPart | Statistically significant | Statistically significant |
| Reportable | Statistically significant | Statistically significant |
| Significant | Statistically significant | Statistically significant |
| MechanismOfInjury | Statistically significant | Statistically significant |
| NatureOfInjury | Statistically significant | Statistically significant |
| Same\_Date\_Reporting | Statistically significant | Not Statistically significant |
| Incident\_time\_period | Statistically significant | Not Statistically significant |
| Daily\_Mean\_Temp\_degrees | Statistically significant | Statistically significant |
| Daily\_Rainfall\_mm | Statistically significant | Statistically significant |
| StaffContractor | Not statistically significant | Statistically significant |

Cramer’s V

Next, Cramer’s V is applied to evaluate the strength of association between independent variables (X) against the dependent variable (Y), “Severity”. The strength of association and degrees of freedom categorisation is retrieved from Peter Statistics (2020). For example, “AgencyOfInjury” and “Severity” reveal a significant but small association of 0.1218. From Table 6, in the Australia dataset, 8 variables have small associations, and 1 variable has negligible and large associations respectively with severity. Whereas, in the Canada dataset, 5 variables have a small association, 3 variables had a medium association and 1 indicates a large association with severity.

Detailed results of the chi-square test and Cramer’s V are presented in the Appendix, Table 14.

*Table 6: Cramer’s V association results*

|  |  |  |
| --- | --- | --- |
| Variable | Cramer’s V results  (Australia dataset) | Cramer’s V results  (Canada dataset) |
| AgencyOfInjury | Small | Medium |
| BodyPart | Small | Medium |
| Reportable | Large | Large |
| Significant | Small | Small |
| MechanismOfInjury | Small | Small |
| NatureOfInjury | Small | Small |
| Same\_Date\_Reporting | Small | Drop from dataset |
| Incident\_time\_period | Small | Drop from dataset |
| Daily\_Mean\_Temp\_degrees | Small | Medium |
| Daily\_Rainfall\_mm | Negligible | Small |
| StaffContractor | Drop from dataset | Small |

# **Chapter 3.3: Data Specification**

*Table 7: Summary of final dataset for modelling*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Name | Type | Values | Australia | Canada |
| Reportable | Categorical | True  False | ✓ | ✓ |
| Significant | Categorical | True  False | ✓ | ✓ |
| AgencyOfInjury | Categorical | Animal, Human and Biological Agencies  Environmental Agencies  Machinery and Equipment  Materials and Substances  Other and Unspecified Agencies | ✓ | ✓ |
| BodyPart | Categorical | Lower Body  Multiple or Unspecified Locations  Upper Body | ✓ | ✓ |
| StaffContractor | Categorical | Contractor  Staff | ✗ | ✓ |
| MechanismOfInjury | Categorical | Biological Factors  Direct Impact  Environmental Factors  Falls and Trips  Other or Unspecified Mechanisms  Stress (Physical/Mental) | ✓ | ✓ |
| Severity | Categorical | Minor  Moderate  Serious | ✓ | ✓ |
| NatureOfInjury | Categorical | Burn  Other or Unspecified  Traumatic Injuries  Diseases and Mental Health | ✓ | ✓ |
| Same\_Date\_Reporting | Categorical | Yes  No | ✓ | ✗ |
| Incident\_time\_period | Categorical | Evening  Midnight  Morning  Afternoon | ✓ | ✗ |
| Daily\_Rainfall\_mm | Categorical | Heavy  Light  No Rain | ✓ | ✓ |
| Daily\_Mean\_Temp\_degrees | Categorical | Cool  Hot  Very Hot  Warm | ✓ | ✓ |
| Monthly Contractor Hours per Mine | Numeric |  | ✓ | ✓ |
| Monthly Staff Hours per Mine | Numeric |  | ✓ | ✓ |
|  |  | Total count of rows: | 3155 | 620 |

Table 7 presents a summary of the final dataset consisting of the variable’s names, total count of rows, variables present in the Australian and Canadian datasets, data type, and values.

# **Chapter 4: Modelling and Evaluation**

## Chapter 4.1: Modelling

Before modelling begins, the dataset is partitioned into 70% training and 30% testing with a fixed random seed of 42 to prevent modelling results from changing. Also, One-Hot encoding is performed to transform categorical variables into 0,1 before applying machine learning models. The target variable is a multi-class target with 3 categories: Minor, Moderate and Serious. Since the data mining goal is to predict factors likely to cause injuries in mines, predictive modelling is chosen as the modelling technique. Hence, for multi-class modelling, Dataman (2023) recommends 7 supervised models consisting of a decision tree, decision tree with SMOTE, decision tree with boosting, support vector machine (OneVsRest approach), neural network, random forest, and logistic regression (OneVsRest approach). OneVsRest approach creates an individual classifier for each target class by marking the chosen class as positive and the rest as negative. To illustrate 3 target classes in severity, 3 classifiers are created as follows:

Classifier 1: Minor vs. Moderate, Serious (Minor, not minor)

Classifier 2: Moderate vs. Minor, Serious (Moderate, not moderate)

Classifier 3: Serious vs. Minor, Moderate (Serious, not serious)

Additionally, to prevent overfitting, pruning parameters in the decision tree including maximum depth, minimum samples required to split and minimum samples in a leaf node were implemented. To illustrate, the Decision Tree with SMOTE contains pruning parameters like a maximum depth of 5, minimum sample split of 100 and minimum sample leaves of 50.

Furthermore, model performance is assessed through 7 classification metrics comprising of training and testing accuracy, precision (hit rate), recall (accuracy), F1 score, Receiver Operating Characteristic (ROC) curve and Kappa score to determine the champion model for the Australia and Canada dataset.

Firstly, training and testing accuracy measures how well the model performs using train data and how well the model performs on unseen data (Smolic, 2022). Also, it is an important element to determine if the model is overfitting or underfitting. For instance, underfitting occurs when the model has both high training and testing accuracy with little improvements with more training. Whereas overfitting happens when the model has a significant difference between the training and testing accuracy (Castagno, 2023).

Secondly, precision (hit rate) and recall (accuracy) imply the proportion of predicted positives is actually positive and the proportion of events and non-events predicted correctly respectively (Shmueli, 2021). For example, the hit rate indicates the proportion of moderate predicted correctly are actually moderate while accuracy measures the proportion of moderate and non-moderate class predicted correctly.

Thirdly, the F1 score does not equate to a higher-performing model. However, the F1 score computes an average of precision and recall (Shmueli, 2023). Fourth, the ROC curve is a measure to determine the model’s performance between each target class, more suitable for balanced targets (T, 2023).

Lastly, the kappa score measures the probability by chance between actual and predicted values where a score of 1 implies a perfect agreement while 0 implies a chance agreement (Shmueli, 2021b). The classification metrics discussed above is applied to 7 supervised models to determine the champion model for the Australian and Canadian dataset, illustrated in Table 8 and 9.

Australian dataset

*Table 8: Summary of modelling results for the Australian dataset*

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Decision tree | | | Decision tree with SMOTE | | | Decision tree with boosting | | | Support vector machine (OneVsRest approach) | | |
| Target | Minor | Moderate | Serious | Minor | Moderate | Serious | Minor | Moderate | Serious | Minor | Moderate | Serious |
| Training  accuracy | 70.15% | | | 66.89% | | | 80.61% | | | 63.49% | | |
| Testing  accuracy | 70.96% | | | 66.73% | | | 66.84% | | | 65.78% | | |
| Hit Rate  (Precision) | 0.58 | 0.75 | 0.00 | 0.54 | 0.74 | 0.25 | 0.61 | 0.75 | 0.26 | 0.55 | 0.66 | 0.00 |
| Accuracy  (Recall) | 0.67 | 0.88 | 0.00 | 0.66 | 0.80 | 0.08 | 0.62 | 0.80 | 0.18 | 0.15 | 0.97 | 0.00 |
| F1-score | 0.62 | 0.81 | 0.00 | 0.59 | 0.77 | 0.12 | 0.61 | 0.77 | 0.21 | 0.24 | 0.79 | 0.00 |
| ROC Curve | 0.86 | 0.72 | 0.60 | 0.81 | 0.70 | 0.58 | 0.85 | 0.68 | 0.66 | 0.58 | 0.58 | 0.47 |
| Kappa Score | 0.36 | | | 0.32 | | | 0.33 | | | 0.10 | | |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Neural network | | | Random forest | | | Logistic Regression  (OneVsRest approach) | | |
| Target | Minor | Moderate | Serious | Minor | Moderate | Serious | Minor | Moderate | Serious |
| Training  accuracy | 48.35% | | | 99.59% | | | 62.36% | | |
| Testing  accuracy | 51.84% | | | 72.12% | | | 64.63% | | |
| Hit Rate  (Precision) | 0.29 | 0.68 | 0.00 | 0.66 | 0.75 | 0.45 | 1.00 | 0.65 | 0.00 |
| Accuracy  (Recall) | 0.60 | 0.61 | 0.00 | 0.62 | 0.90 | 0.10 | 0.00 | 1.00 | 0.00 |
| F1-score | 0.39 | 0.64 | 0.00 | 0.64 | 0.81 | 0.16 |  |  |  |
| ROC Curve | 0.63 | 0.59 | 0.46 | 0.63 | 0.59 | 0.46 | 0.27 | 0.35 | 0.54 |
| Kappa Score | 0.10 | | | 0.38 | | | 0.00 | | |

From Table 8, the Decision Tree with boosting and random forest exhibited overfitting issues based on significantly higher training and lower testing accuracy. Among the remaining models, Neural Networks presented the lowest training and testing accuracy of 48.35% and 51.84% respectively. Therefore, to determine the champion model, metrics such as precision, recall, f1 score, roc curve and Kappa score were assessed. Based on the kappa score metric, both the Decision Tree and the Decision Tree with SMOTE demonstrated fair agreement of 0.36 and 0.32 respectively (McHugh, 2012). Conversely, Logistic Regression and Support Vector Machine models yielded poor and slight agreement of 0.00 and 0.10 respectively (McHugh, 2012). Further evaluation through precision, recall, f1 score and roc curve revealed Decision Tree with SMOTE achieved a better precision, recall, f1 score and roc curve across all 3 targets compared to the Decision Tree illustrated in Table 8. Moreover, it was observed that the Decision Tree failed to predict any correct instances for the target class “Serious”. Hence, the champion model for the Australian dataset is the Decision Tree with SMOTE. Using SMOTE, the minority target like ‘Serious’ and ‘Minor’ has been oversampled to 1000 records respectively. Feature importance and confusion matrix for the champion model are presented in Figures 11 and 12 respectively.

It is observed that “Monthly staff hours per Mine”, “Incident\_Time\_Period\_Morning”, “Monthly contractor hours per Mine”, “BodyPart\_Upper Body” and “AgencyOfInjury\_Other and Unspecified Agencies” were the top five factors resulting in a safety incident which managers could focus on to prevent injuries within mines. Based on the champion model, a total of 28 rules were generated comprising 10, 12 and 6 rules for minor, moderate, and serious respectively presented in the Appendix, Table 15. Table 10 illustrates 1 set of decision tree rules from each target class for discussion.

*Figure 11: Feature importance of the champion model*

A blue and white bar chart

Description automatically generated

*Figure 12: Confusion Matrix for the champion model*

A screenshot of a computer

Description automatically generated

*Table 10: 1 of the Decision tree rules generated from the Decision Tree with SMOTE*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Minor | Moderate | Serious |
| 1 | Rule:  Monthly Staff Hours per Mine > 37239.0 & Monthly Contractor Hours per Mine > 23134.6630859375 & Monthly Staff Hours per Mine <= 55351.0 & Incident\_time\_period\_Morning > 0.919118732213974 & MechanismOfInjury\_Direct Impact > 0.5551130771636963 &  Probability of 'Minor': 0.6515  Probability of 'Moderate': 0.3182  Probability of 'Serious': 0.0303 | Rule:  Monthly Staff Hours per Mine > 37239.0 & Monthly Contractor Hours per Mine <= 23134.6630859375 & AgencyOfInjury\_Other and Unspecified Agencies <= 0.007798909675329924 & Monthly Contractor Hours per Mine <= 20169.0 & AgencyOfInjury\_Animal, Human and Biological Agencies <= 0.0037671816535294056 &  Probability of 'Minor': 0.0635  Probability of 'Moderate': 0.5476  Probability of 'Serious': 0.3889 | Rule:  Monthly Staff Hours per Mine > 37239.0 & Monthly Contractor Hours per Mine <= 23134.6630859375 & AgencyOfInjury\_Other and Unspecified Agencies > 0.007798909675329924 & Monthly Contractor Hours per Mine > 16191.3173828125 & Daily\_Mean\_Temp\_degrees\_Warm > 0.04577472060918808 &  Probability of 'Minor': 0.0750  Probability of 'Moderate': 0.0250  Probability of 'Serious': 0.9000 |

For instance, the likelihood of a safety incident predicted as “Minor” severity is 65.15% given the following conditions “Monthly Staff hours per Mine” ranging between 37,239 to 55,351 hours and “Monthly Contractor hours per Mine” exceeds 23,134 hours, the safety incident period occurring during the morning period and the mechanism of injury is identified as direct impact.

Next, the likelihood of a safety incident predicted as “Moderate” severity is 54.7% given that “Monthly Staff hours per Mine” is more than 37239 hours and “Monthly Contractor hours per Mine” ranging between 20169 to 23134 hours, and the agency of injury is classified as either “Other and Unspecified Agencies” or “Animal, Human and Biological Agencies”. Lastly, the probability of a safety incident resulting in “Serious” severity is 90 % when “Monthly Staff hours per Mine” exceeds 37,239 hours and “Monthly Contractor hours per Mine” ranging from 16191 to 23134 hours, the agency of Injury is Other and Unspecified Agencies and the ambient temperature for the day is warm.

Canadian dataset

*Table 9: Summary of modelling results for the Canadian dataset*

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Decision tree | | | Decision tree with SMOTE | | | Decision tree with boosting | | | Support vector machine (OneVsRest approach) | | |
| Target | Minor | Moderate | Serious | Minor | Moderate | Serious | Minor | Moderate | Serious | Minor | Moderate | Serious |
| Training  accuracy | 63.36% | | | 62.17% | | | 76.96% | | | 65.20% | | |
| Testing  accuracy | 66.66% | | | 62.90% | | | 60.21% | | | 63.44% | | |
| Hit Rate  (Precision) | 0.38 | 0.69 | 0.00 | 0.38 | 0.74 | 0.38 | 0.38 | 0.70 | 0.27 | 0.29 | 0.70 | 0.00 |
| Accuracy  (Recall) | 0.18 | 0.94 | 0.00 | 0.50 | 0.75 | 0.25 | 0.21 | 0.78 | 0.25 | 0.32 | 0.87 | 0.00 |
| F1-score | 0.24 | 0.80 | 0.00 | 0.43 | 0.75 | 0.30 | 0.27 | 0.74 | 0.26 | 0.31 | 0.78 | 0.00 |
| ROC Curve | 0.77 | 0.67 | 0.58 | 0.81 | 0.66 | 0.63 | 0.76 | 0.62 | 0.63 | 0.73 | 0.60 | 0.63 |
| Kappa Score | 0.07 | | | 0.23 | | | 0.11 | | | 0.10 | | |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Neural network | | | Random forest | | | Logistic Regression  (OneVsRest approach) | | |
| Target | Minor | Moderate | Serious | Minor | Moderate | Serious | Minor | Moderate | Serious |
| Training  accuracy | 45.24% | | | 98.38% | | | 60.82% | | |
|  |  | | |  | | |  | | |
| Testing  accuracy | 67.74% | | | 63.44% | | | 67.20% | | |
| Hit Rate  (Precision) | 0.00 | 0.68 | 0.00 | 0.43 | 0.71 | 0.27 | 0.00 | 0.68 | 0.25 |
| Accuracy  (Recall) | 0.00 | 1.00 | 0.00 | 0.21 | 0.84 | 0.19 | 0.00 | 0.98 | 0.03 |
| F1-score | 0.00 | 0.81 | 0.00 | 0.29 | 0.77 | 0.22 | 0.00 | 0.81 | 0.06 |
| ROC Curve | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.47 | 0.51 | 0.55 |
| Kappa Score | 0.0 | | | 0.13 | | | 0.01 | | |

From Table 9, the Decision Tree with boosting, Support Vector Machine (OneVsRest approach), and Random Forest exhibited overfitting issues based on significantly higher training and lower testing accuracy. Whereas Decision Trees, Neural Networks, and Logistic Regression (OneVsRest approach) revealed underfitting issues and generalisation issues where the training accuracy is lower than the testing accuracy. Therefore, based on training and testing accuracy, the champion model is the Decision Tree with SMOTE. Using SMOTE, the minority target like ‘Serious’ and ‘Minor’ has been oversampled to 100 records respectively. Additionally, based on the kappa score metric, the decision tree with SMOTE displayed modest agreement, which is also the best agreement among all models. Feature importance and confusion matrix for the champion model are presented in Figures 13 and 14 respectively.

*Figure 13: Feature Importance of the Champion Model*

A blue and white bar chart

Description automatically generated

*Figure 14: Confusion Matrix for the champion model*

A screenshot of a computer

Description automatically generated

It is observed that “Monthly staff hours per Mine”, “BodyPart\_Upper Body”, “NatureOfInjury\_Others or unspecified”, “AgencyOfInjury\_Environmental Agencies” and “Daily\_Mean\_Temp\_Degrees\_Warm” were the top five factors resulting in a safety incident which managers could focus on to prevent injuries within mines. Based on the champion model, a total of 7 rules were generated comprising 2, 4 and 1 rules for minor, moderate, and serious respectively presented in the Appendix, Table 16. Table 11 illustrates 1 set of decision tree rules from each target class for discussion.

*Table 11: 1 of the Decision tree rules generated from the Decision Tree with SMOTE*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Minor | Moderate | Serious |
| 1 | Rule:  Monthly Staff Hours per Mine <= 79996.0 & NatureOfInjury\_Other or Unspecified > 0.04530321806669235 & BodyPart\_Upper Body <= 0.9884687960147858 &  Probability of 'Minor': 0.5273  Probability of 'Moderate': 0.4545  Probability of 'Serious': 0.0182 | Rule:  Monthly Staff Hours per Mine <= 79996.0 & NatureOfInjury\_Other or Unspecified <= 0.04530321806669235 & Daily\_Mean\_Temp\_degrees\_Warm <= 0.13900060951709747 &  Probability of 'Minor': 0.0215  Probability of 'Moderate': 0.8387  Probability of 'Serious': 0.1398 | Rule:  Monthly Staff Hours per Mine > 79996.0 & AgencyOfInjury\_Environmental Agencies <= 0.00838455744087696 & BodyPart\_Upper Body > 0.5 &  Probability of 'Minor': 0.2115  Probability of 'Moderate': 0.3846  Probability of 'Serious': 0.4038 |

For example, the likelihood of a safety incident predicted as “Minor” severity is 52.73% given the following conditions “Monthly Staff hours per Mine” less than 79,996 hours, the nature of injury is either categorised as others or unspecified and the affected body part is upper body.

Next, the likelihood of a safety incident predicted as “Moderate” severity is 83.87% given that “Monthly Staff hours per Mine” are less than 79,996 hours, the nature of the injury is also either others or unspecified and the ambient temperature for the day is warm.

Last, the probability of a safety incident resulting in “Serious” severity is 40% when “Monthly Staff hours per Mine” is less than 79,996 hours, the agency of injury is environmental, and the compromised body part is the upper body.

## Chapter 4.2: Evaluation

This paper aims to evaluate the modelling results against the data mining objective of predicting factors likely to cause injuries in mines using the given dataset for actionable safety insights to enhance safety performances in mines. From the Australian dataset, the Decision Tree with SMOTE presented a training and testing accuracy of 66.89% and 66.73% respectively which doesn’t constitute a high accuracy model. Additionally, based on the confusion matrix in Figure 12, the model has only correctly classified 129, 492 and 11 records out of 3155 records for minor, moderate and serious injury severity respectively. Hence, the model might not be useful for the mining organization to predict a serious injury severity.

Similarly, for the Canadian dataset, the Decision Tree with SMOTE doesn’t result in a high-accuracy model based on the training and testing accuracy of 62.17% and 62.90% respectively. Also, based on the confusion matrix in Figure 14, the model has only correctly predicted 14, 95 and 8 records out of 620 records correctly for minor, moderate and serious injury severity respectively. The above evaluation suggests that future work could include collecting more data, especially for the minority target classes to help the model generalise better to reduce overfitting (S, 2023).

# **Chapter 5: Recommendations, Deployment and Conclusions**

## Chapter 5.1: Recommendations and Deployment

Based on Decision Tree rules generated for the Australian and Canadian datasets, the following safety recommendations can be considered by the management to enhance safety and reduce or prevent injuries within the organisation.

* Monthly working hours: Mines with monthly staff hours greater than 37,239 hours and contractor hours greater than 23,134 hours presented a higher occurrence of a safety event, hence, to mitigate this, it is recommended that the organization monitors the total monthly hours worked per mine by contractors and staff to avoid clocking more than a total of 37,239 hours and 23,134 hours for staff and contractors respectively and also reduce adjust the miners shift hours to prevent fatigue which is the primary cause of accidents.
* Shift Timing: safety incidents frequently occur during either morning or midnight. To minimize this occurrence, the organization could implement safety measures such as ensuring closer supervision during these shift hours and ensuring proper lighting, especially during midnight shifts.
* Environmental conditions: warm temperatures are associated with serious injuries while cool temperatures are associated with moderate injuries. Thus, it is recommended to ensure hydration, take frequent water breaks, monitor for signs of heat stress during warm temperatures and ensure proper thermal wear during cool temperatures.
* Agency of injury: “Others or unspecified agencies” and “Environmental Agencies” constitute the top 5 features importance for the primary cause associated with the safety incident. Hence, the recommendations to reduce this injury could entail offering training sessions to manage this hazard more effectively. Additionally, it would be more effective for the organization to deep dive into the root cause of “Others or unspecified agencies” as this categorization is too generalized.
* Mechanism of injury and Body Part (Upper Body): Direct impact emerges as one of the most significant feature importance within the Decision Tree rules. Additionally, the affected body part was localized at the upper body. Consequently, to reduce the occurrence of direct impact injuries to the miners, it is recommended that mining organizations enforce personal protective equipment like hard hats, safety glasses, ear plugs, safety gloves, protective clothing, and metatarsal boots (anbusafety, 2023).

## Chapter 5.2: Conclusions

To conclude, adopting the CRISP-DM framework, the business problem was defined to address injuries in mines resulting in lost-time injuries (LTI) with the business objective to enhance the safety of miners in Australia and Canada by striving to attain zero injuries at work and lastly, the data mining goal of predicting factors likely to cause injuries in mines.

To begin, a total of 61 columns based on the given dataset which was reduced to 14 columns after applying the chi-square test and Cramer’sV. Next, data cleaning was performed to prepare the datasets for modelling. Also, the cleaned dataset was encoded using OneHotEncoder and partitioned into 70% training and 30% testing before modelling. Then, the Decision Tree with SMOTE was selected as the champion model based on the modelling evaluation metrics. Additionally, the model has met the business objectives and data mining goal of predicting factors likely to cause injuries in mines.

Future works could include gathering more datasets based on miners' demographics like age, work experience, project size (Dumrak et al., 2013), and mine type (Muzaffar et al., 2013) to determine if these factors contribute to the likelihood of a safety incident.

Lastly, the mining organization could implement the recommendations generated based on the Decision Tree rules to enhance safety performance within the organization.

Total number of words: 6651 words

# **References**

Alkaissy, M., Arashpour, M., Golafshani, E. M., Hosseini, M. R., Khanmohammadi, S., Bai, Y., & Feng, H. (2023). Enhancing construction safety: Machine learning-based classification of injury types. *Safety Science*, *162*, 106102. https://doi.org/10.1016/j.ssci.2023.106102

anbusafety. (2023). *Complete Guide to Mining Personal Protective Equipment (PPE)*. Anbusafety. Retrieved October 31, 2023, from https://www.anbusafety.com/mining-ppe-2/

Asare-Doku, W., Jane, R. L., Kelly, B., Amponsah-Tawiah, K., & James, C. (2022). Mental health and workplace factors: comparison of the Ghanaian and Australian mining industry. *BMC Health Services Research*, *22*(1). https://doi.org/10.1186/s12913-022-07712-0

Benson, C. (2023). Indicators as early warning signal performance to solve underlying safety problem before they emerge as. . . *ResearchGate*. https://www.researchgate.net/publication/369562500\_Indicators\_as\_early\_warning\_signal\_performance\_to\_solve\_underlying\_safety\_problem\_before\_they\_emerge\_as\_accident\_risks

Bock, T. (2022, April 25). *How to interpret correspondence analysis plots (It probably isn’t the way you think) - Displayr*. Displayr. https://www.displayr.com/interpret-correspondence-analysis-plots-probably-isnt-way-think/

Butani, S. (1988). Relative risk analysis of injuries in coal mining by age and experience at present company. *Journal of Occupational Accidents*, *10*(3), 209–216. https://doi.org/10.1016/0376-6349(88)90014-4

Castagno, P. (2023, October 1). How to determine if your model has Overfitting or Underfitting. *Medium*. https://patriziacastagnod.medium.com/how-to-determine-if-your-model-has-overfitting-or-underfitting-9c9dc4860270

Dataman, C. K. (2023, January 30). A wide variety of models for multi-class classification. *Medium*. https://medium.com/dataman-in-ai/a-wide-choice-for-modeling-multi-class-classifications-d97073ff4ec8

Dumrak, J., Mostafa, S., Kamardeen, I., & Rameezdeen, R. (2013). Factors associated with the severity of construction accidents: The case of South Australia. *Construction Economics and Building*, *13*(4), 32–49. https://doi.org/10.5130/ajceb.v13i4.3620

Government of Canada. (2023, July 13). *CCOHS: Hazards*. Retrieved August 12, 2023, from https://www.ccohs.ca/topics/hazards/#ctgt

Government of Western Australia. (2022, March 31). *WALW - Mines Safety and Inspection Act 1994 - home page*. Western Australian Legislation. Retrieved August 15, 2023, from https://www.legislation.wa.gov.au/legislation/statutes.nsf/main\_mrtitle\_599\_homepage.html

Lee, J., Yoon, Y., Oh, T. K., Park, S., & Ryu, S. I. (2020). A study on Data Pre-Processing and Accident Prediction Modelling for occupational accident analysis in the construction industry. *Applied Sciences*, *10*(21), 7949. https://doi.org/10.3390/app10217949

Manjunatha, A. (2023, May 12). *Injury Prediction in Mining Industry through Applied Machine Learning Approaches  - NORMA@NCI Library*. National College of Ireland. Retrieved August 15, 2023, from https://norma.ncirl.ie/6559/

McHugh, M. L. (2012, October 1). *Interrater reliability: the kappa statistic*. PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/#:~:text=Cohen%20suggested%20the%20Kappa%20result,1.00%20as%20almost%20perfect%20agreement.

Muzaffar, S., Cummings, K. J., Hobbs, G. R., Allison, P. D., & Kreiss, K. (2013). Factors associated with fatal mining injuries among contractors and operators. *Journal of Occupational and Environmental Medicine*, *55*(11), 1337–1344. https://doi.org/10.1097/jom.0b013e3182a2a5a2

Ontario Ministry of Mines. (2023). *Map*. Retrieved August 16, 2023, from https://oma.on.ca/en/ontario-mining/Map.aspx

Ontario’s Regulatory Registry. (2021, July 28). *Mining Health and Safety Regulatory Amendment Proposal*. © King’s Printer for Ontario, 2022. Retrieved August 15, 2023, from https://www.ontariocanada.com/registry/view.do?postingId=37907

Peter Statistics. (2020, August 5). *Nominal vs. Nominal - Part 3c: Effect size (Cramer’s V)*. Retrieved October 15, 2023, from https://peterstatistics.com/CrashCourse/3-TwoVarUnpair/NomNom/NomNom-2c-Effect-Size.html

Queensland Government. (2023, June 9). *Mining hazards database*. Business Queensland. Retrieved August 15, 2023, from https://www.business.qld.gov.au/industries/mining-energy-water/resources/safety-health/mining/hazards/hazards

Resources Safety & Health Queensland. (2023, July 4). *What we do*. Resources Safety and Health Queensland. Retrieved August 15, 2023, from https://www.rshq.qld.gov.au/about-us/what-we-do

S, R. (2023, August 26). Machine Learning — overfitting and underfitting - Rahul S - Medium. *Medium*. https://ogre51.medium.com/machine-learning-overfitting-and-underfitting-bb88e6548676

Safe Work Australia. (2013). ISSUES IN THE MEASUREMENT AND  REPORTING OF WORK HEALTH AND  SAFETY PERFORMANCE: A REVIEW. In *Safeworkaustralia*. Retrieved September 5, 2023, from https://www.safeworkaustralia.gov.au/system/files/documents/1703/issues-measurement-reporting-whs-performance.pdf

Safe Work Australia. (2022, November 7). *Key work health and safety statistics Australia 2022*. Retrieved August 15, 2023, from https://www.safeworkaustralia.gov.au/doc/key-work-health-and-safety-statistics-australia-2022

Santos, G. (2023, March 1). How strongly associated are your variables? - Towards Data Science. *Medium*. https://towardsdatascience.com/how-strongly-associated-are-your-variables-80493127b3a2

Shmueli, B. (2021a, December 10). Multi-Class Metrics Made Simple, Part I: Precision and recall. *Medium*. https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2

Shmueli, B. (2021b, December 13). Multi-Class Metrics Made Simple, Part III: the Kappa Score (aka Cohen’s Kappa Coefficient). *Medium*. https://towardsdatascience.com/multi-class-metrics-made-simple-the-kappa-score-aka-cohens-kappa-coefficient-bdea137af09c

Shmueli, B. (2023, March 23). Multi-Class Metrics Made Simple, Part II: the F1-score. *Medium*. https://towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-ebe8b2c2ca1

Smolic, H. (2022, September 22). Training Data vs. Test Data in Machine Learning — Essential Guide. *Medium*. https://pub.towardsai.net/training-data-vs-test-data-in-machine-learning-essential-guide-c58404849cea

Stemn, E., & Benyarku, C. A. (2023). Mineworkers’ perspective of fatigue: A study of the Ghanaian mining industry. *Safety Science*, *162*, 106095. https://doi.org/10.1016/j.ssci.2023.106095

Stemn, E., & Krampah, F. (2022). Injury severity and influence factors in surface mines: A correspondence analysis. *Safety Science*, *145*, 105495. https://doi.org/10.1016/j.ssci.2021.105495

T, B. (2023, April 8). Comprehensive Guide to Multiclass Classification with Sklearn | Towards Data Science. *Medium*. https://towardsdatascience.com/comprehensive-guide-to-multiclass-classification-with-sklearn-127cc500f362

Towsey, C. (2011). Proactive Measures for fatality prevention in the mining industry — Why fatalities persist while lost time. . . *ResearchGate*. https://www.researchgate.net/publication/268048780\_Proactive\_Measures\_for\_Fatality\_Prevention\_in\_the\_Mining\_Industry\_-\_Why\_Fatalities\_Persist\_While\_Lost\_Time\_Injuries\_Decline

Turner, A. (2020, January 27). *Eighty years of Canadian climate data*. Kaggle. Retrieved September 6, 2023, from https://www.kaggle.com/datasets/aturner374/eighty-years-of-canadian-climate-data?resource=download

Vlachos, T. (2019). INTERRELATION BETWEEN OCCUPATIONAL HEALTH & SAFETY LEADING AND LAGGING INDICATORS IN MINING INDUSTRY. AN. . . *ResearchGate*. https://www.researchgate.net/publication/338264721\_INTERRELATION\_BETWEEN\_OCCUPATIONAL\_HEALTH\_SAFETY\_LEADING\_AND\_LAGGING\_INDICATORS\_IN\_MINING\_INDUSTRY\_AN\_EMPIRICAL\_STUDY

Watson, J. H. (2022, January 4). Contingency Tables, Chi-Squared and Cramer’s V - towards Data Science. *Medium*. https://towardsdatascience.com/contingency-tables-chi-squared-and-cramers-v-ada4f93ec3fd

Workplace Safety North. (2022). *Mining Statistics | Workplace Safety North*. Retrieved October 14, 2023, from https://www.workplacesafetynorth.ca/resources/mining-statistics

World Coal Association. (2023, March 20). *Coal’s contribution - World Coal Association*. Retrieved August 15, 2023, from https://www.worldcoal.org/coal-facts/coals-contribution/

Young, J., & Young, A. (2020, December 11). *Rain in Australia*. Kaggle. Retrieved September 6, 2023, from https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package

# **Appendix**

*Table 12: Detailed observation of data quality issues during initial data exploration*

|  |  |
| --- | --- |
| Dataset | Observation |
| person\_workgroup.csv | Using Excel, the following observations were made:  **Invalid date format**: from\_date, to\_date |
| employee\_roster.csv | No data quality issues were observed |
| labour\_hours\_worked.csv | No data quality issues were observed using Excel, but outliers were detected using IBM SPSS data audit node.  A white background with black text  Description automatically generated |
| site\_location.csv | site\_key\_hashed in safety\_events.csv revealed missing 5 locations when mapped using Excel VLOOKUP with site\_location.csv. |
| emp\_start\_end\_dates.csv | Using Excel, the following observations were made:   1. **Invalid date format**: start\_date, end\_date 2. **NA values:** start\_date, end\_date 3. **Duplicate values**: 334 duplicate records in name\_hash (each row represents a specific employee, hence, should be unique)   A screenshot of a computer  Description automatically generated |
| safety\_events.csv | Using Excel, the following observations were made:   1. **Irrelevant columns:** day, event\_time, event\_reported\_time, EventId, AgencyOfInjuryId, BodyPartId, InjuryTypeCode,   MechanismOfInjuryId, shift\_commenced\_day,  shift\_commenced\_time, shift\_end\_day, shift\_end\_time,  AgencyOfInjuryDescription, BodyPartDescription,  NatureOfInjuryDecription, MechanismOfInjuryDecription   1. **Invalid date format:** event\_dt, event\_reported\_dt,   derived\_shift\_start\_dt, derived\_shift\_end\_dt   1. **NA values:** PersonName\_hashed, StaffContractor,   OrganisationName\_hashed, TimeBand  Using IBM SPSS data audit node, the following findings were revealed:   1. Absence of missing values except for outliers for LTIDays A screenshot of a computer     Description automatically generated 2. Disproportionate proportion of “TRUE” and “FALSE” values for “Reportable” and “Significant” as seen below:A screenshot of a graph     Description automatically generated   Using IBM SPSS distribution plot (under Graph),   1. More than 10 categories were observed in AgencyOfInjury, BodyPart, Injury, MechanismOfInjury, NatureOfInjury   A graph of a number of different species  Description automatically generated with medium confidence  A graph with blue bars  Description automatically generated  A graph with blue rectangles and black text  Description automatically generatedA graph with blue and black bars  Description automatically generated  A graph with blue bars  Description automatically generated |
| production\_data.csv | Using Excel, the following observations were made:   1. **Invalid date format in respective columns**: date, acutal\_tonnes\_moved,budgeted\_tonnes\_moved,short\_range\_forecast\_tonnes\_moved, half\_2\_forecast\_tonnes\_moved contains values with different decimal places.   Using IBM SPSS data audit node, the following findings were revealed:   1. No missing values except for outliers for   acutal\_tonnes\_moved,budgeted\_tonnes\_moved,short\_range\_forecast\_tonnes\_moved, half\_2\_forecast\_tonnes\_moved  A screenshot of a computer  Description automatically generated |

*Table 13: Summary of new columns created after merging and percentage of N/A values*

|  |  |  |  |
| --- | --- | --- | --- |
| Files merged with safety\_events.csv | Unique keys | New column | Percentage of N/A values |
| production\_data.csv | site\_key\_hashed | ‘met\_target\_budgeted\_tones?, acutal\_tonnes\_moved, budgeted\_tonnes\_moved | A screenshot of a computer  Description automatically generated |
| labour\_hours\_worked.csv | site\_key\_hashed | ‘scheduled\_work\_hours | A screenshot of a computer  Description automatically generated |
| ‘employee\_roster.csv | ‘name\_hash | Annual leave,  Compassionate Leave, Injury Leave, Maternity Leave, Medical Leave, Off Day, Parental Leave | A screenshot of a computer  Description automatically generated  A screenshot of a computer  Description automatically generated |
| ‘person\_workgroup.csv | ‘name\_hash | ‘switching\_roles, switching\_roles\_count | A screenshot of a computer  Description automatically generated  A screenshot of a computer  Description automatically generated |
| ‘employee\_roster.csv | ‘name\_hash | Total number of employees on leave | A screenshot of a computer  Description automatically generated |
| ‘emp\_start\_end\_dates.csv | ‘name\_hash | employee\_count,  employee\_count\_difference,  length\_of\_employment | A screenshot of a computer  Description automatically generated  A screenshot of a computer  Description automatically generated |

*Table 14: Summary of chi-square test and Cramer’s V among independent and dependent variables*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | Chi-square Test | | Cramer’s V | |
| Dataset | Independent variable (X) | Dependent Variable (Y) | P-Value  (alpha = 0.05) | Significantly associated? | Cramer’s V | Strength of association |
| Australia | AgencyOfInjury | Severity | 8.857403318006178e-17 | Yes | 0.1218 | small |
| BodyPart | Severity | 4.393770568268758e-13 | Yes | 0.1006 | small |
| Reportable | Severity | 0.0 | Yes | 0.9948 | large |
| Significant | Severity | 4.630086433809787e-14 | Yes | 0.1395 | small |
| MechanismOfInjury | Severity | 1.802635368883181e-11 | Yes | 0.1068 | small |
| NatureOfInjury | Severity | 8.539475343494162e-11 | Yes | 0.0964 | small |
| Region | Severity | 1.1379134924781799e-60 | Yes | 0.2958 | medium |
| Weather\_Station | Severity | 1.1379134924782122e-60 | Yes | 0.2958 | medium |
| Same\_Date\_Reporting | Severity | 0.00015488091844225246 | Yes | 0.0746 | small |
| Incident\_time\_period | Severity | 7.167867577752338e-21 | Yes | 0.1305 | small |
| Daily\_Mean\_Temp\_degrees | Severity | 5.380523438309088e-07 | Yes | 0.2096 | small |
| Daily\_Rainfall\_mm | Severity | 0.01607393343242117 | Yes | 0.0439 | megligible |
| StaffContractor |  | 0.5904234423087857 | No | 0.0183 | negligible |
| Canada | AgencyOfInjury | Severity | 3.1550444155524957e-09 | Yes | 0.212 | medium |
| BodyPart | Severity | 1.2683413931305667e-08 | Yes | 0.1853 | medium |
| Reportable | Severity | 2.3372792850071602e-135 | Yes | 1.0 | large |
| Significant | Severity | 0.00016772257663415967 | Yes | 0.1675 | small |
| MechanismOfInjury | Severity | 0.0006879541902699265 | Yes | 0.157] | small |
| NatureOfInjury | Severity | 0.0019730914298895564 | Yes | 0.1296 | small |
| Region | Severity | 1.0 | No | 0.0 | negligible |
| Weather\_Station | Severity | 1.0 | No | 0.0 | negligible |
| Same\_Date\_Reporting | Severity | 0.1179815234596341 | No | 0.083 | small |
| Incident\_time\_period | Severity | 0.053592873983086355 | No | 0.1 | small |
| Daily\_Mean\_Temp\_degrees | Severity | 5.859708973371329e-10 | Yes | 0.2096 | medium |
| Daily\_Rainfall\_mm | Severity | 0.001238607478004469 | Yes | 0.1205 | small |
| StaffContractor | Severity | 0.0005599187064551688 | Yes | 0.1554 | small |

*Table 15: Full list of decision tree rules for the Australian Dataset*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Minor | Moderate | Serious |
| 1 | Rule:  Monthly Staff Hours per Mine > 37239.0 & Monthly Contractor Hours per Mine > 23134.6630859375 & Monthly Staff Hours per Mine <= 55351.0 & Incident\_time\_period\_Morning > 0.919118732213974 & MechanismOfInjury\_Direct Impact > 0.5551130771636963 &  Probability of 'Minor': 0.6515  Probability of 'Moderate': 0.3182  Probability of 'Serious': 0.0303 | Rule:  Monthly Staff Hours per Mine > 37239.0 & Monthly Contractor Hours per Mine > 23134.6630859375 & Monthly Staff Hours per Mine > 55351.0 & Daily\_Mean\_Temp\_degrees\_Warm <= 0.16371102631092072 & Incident\_time\_period\_Midnight <= 0.018503816798329353 &  Probability of 'Minor': 0.1220  Probability of 'Moderate': 0.6707  Probability of 'Serious': 0.2073 | Rule:  Monthly Staff Hours per Mine > 37239.0 & Monthly Contractor Hours per Mine > 23134.6630859375 & Monthly Staff Hours per Mine > 55351.0 & Daily\_Mean\_Temp\_degrees\_Warm > 0.16371102631092072 &  Probability of 'Minor': 0.1034  Probability of 'Moderate': 0.0345  Probability of 'Serious': 0.8621 |
| 2 | Rule:  Monthly Staff Hours per Mine > 37239.0 & Monthly Contractor Hours per Mine > 23134.6630859375 & Monthly Staff Hours per Mine <= 55351.0 & Incident\_time\_period\_Morning > 0.919118732213974 & MechanismOfInjury\_Direct Impact <= 0.5551130771636963 &  Probability of 'Minor': 0.4012  Probability of 'Moderate': 0.2778  Probability of 'Serious': 0.3210 | Rule:  Monthly Staff Hours per Mine > 37239.0 & Monthly Contractor Hours per Mine <= 23134.6630859375 & AgencyOfInjury\_Other and Unspecified Agencies > 0.007798909675329924 & Monthly Contractor Hours per Mine > 16191.3173828125 & Daily\_Mean\_Temp\_degrees\_Warm <= 0.04577472060918808 &  Probability of 'Minor': 0.2973  Probability of 'Moderate': 0.4054  Probability of 'Serious': 0.2973 | Rule:  Monthly Staff Hours per Mine > 37239.0 & Monthly Contractor Hours per Mine > 23134.6630859375 & Monthly Staff Hours per Mine > 55351.0 & Daily\_Mean\_Temp\_degrees\_Warm <= 0.16371102631092072 & Incident\_time\_period\_Midnight > 0.018503816798329353 &  Probability of 'Minor': 0.3611  Probability of 'Moderate': 0.2222  Probability of 'Serious': 0.4167 |
| 3 | Rule:  Monthly Staff Hours per Mine > 37239.0 & Monthly Contractor Hours per Mine > 23134.6630859375 & Monthly Staff Hours per Mine <= 55351.0 & Incident\_time\_period\_Morning <= 0.919118732213974 & Incident\_time\_period\_Midnight > 0.9843259751796722 &  Probability of 'Minor': 0.6619  Probability of 'Moderate': 0.3237  Probability of 'Serious': 0.0144 | Rule:  Monthly Staff Hours per Mine > 37239.0 & Monthly Contractor Hours per Mine <= 23134.6630859375 & AgencyOfInjury\_Other and Unspecified Agencies <= 0.007798909675329924 & Monthly Contractor Hours per Mine > 20169.0 & Daily\_Mean\_Temp\_degrees\_Cool > 0.6361536681652069 &  Probability of 'Minor': 0.0263  Probability of 'Moderate': 0.8421  Probability of 'Serious': 0.1316 | Rule:  Monthly Staff Hours per Mine > 37239.0 & Monthly Contractor Hours per Mine <= 23134.6630859375 & AgencyOfInjury\_Other and Unspecified Agencies > 0.007798909675329924 & Monthly Contractor Hours per Mine > 16191.3173828125 & Daily\_Mean\_Temp\_degrees\_Warm > 0.04577472060918808 &  Probability of 'Minor': 0.0750  Probability of 'Moderate': 0.0250  Probability of 'Serious': 0.9000 |
| 4 | Rule:  Monthly Staff Hours per Mine > 37239.0 & Monthly Contractor Hours per Mine > 23134.6630859375 & Monthly Staff Hours per Mine <= 55351.0 & Incident\_time\_period\_Morning <= 0.919118732213974 & Incident\_time\_period\_Midnight <= 0.9843259751796722 &  Probability of 'Minor': 0.8200  Probability of 'Moderate': 0.0886  Probability of 'Serious': 0.0914 | Rule:  Monthly Staff Hours per Mine > 37239.0 & Monthly Contractor Hours per Mine <= 23134.6630859375 & AgencyOfInjury\_Other and Unspecified Agencies <= 0.007798909675329924 & Monthly Contractor Hours per Mine <= 20169.0 & AgencyOfInjury\_Animal, Human and Biological Agencies <= 0.0037671816535294056 &  Probability of 'Minor': 0.0635  Probability of 'Moderate': 0.5476  Probability of 'Serious': 0.3889 | Rule:  Monthly Staff Hours per Mine <= 37239.0 & Incident\_time\_period\_Morning > 0.002650003181770444 & Incident\_time\_period\_Morning > 0.9984371364116669 & Monthly Contractor Hours per Mine > 20596.0 & Monthly Staff Hours per Mine <= 22380.0 &  Probability of 'Minor': 0.0079  Probability of 'Moderate': 0.3622  Probability of 'Serious': 0.6299 |
| 5 | Rule:  Monthly Staff Hours per Mine > 37239.0 & Monthly Contractor Hours per Mine <= 23134.6630859375 & AgencyOfInjury\_Other and Unspecified Agencies > 0.007798909675329924 & Monthly Contractor Hours per Mine <= 16191.3173828125 & Daily\_Rainfall\_mm\_No rain > 0.2527591288089752 &  Probability of 'Minor': 0.8404  Probability of 'Moderate': 0.0532  Probability of 'Serious': 0.1064 | Rule:  Monthly Staff Hours per Mine <= 37239.0 & Incident\_time\_period\_Morning > 0.002650003181770444 & Incident\_time\_period\_Morning > 0.9984371364116669 & Monthly Contractor Hours per Mine > 20596.0 & Monthly Staff Hours per Mine > 22380.0 &  Probability of 'Minor': 0.2324  Probability of 'Moderate': 0.4366  Probability of 'Serious': 0.3310 | Rule:  Monthly Staff Hours per Mine <= 37239.0 & Incident\_time\_period\_Morning > 0.002650003181770444 & Incident\_time\_period\_Morning <= 0.9984371364116669 & AgencyOfInjury\_Other and Unspecified Agencies <= 0.7225707173347473 & AgencyOfInjury\_Animal, Human and Biological Agencies <= 0.23405209928750992 &  Probability of 'Minor': 0.1020  Probability of 'Moderate': 0.0000  Probability of 'Serious': 0.8980 |
| 6 | Rule:  Monthly Staff Hours per Mine > 37239.0 & Monthly Contractor Hours per Mine <= 23134.6630859375 & AgencyOfInjury\_Other and Unspecified Agencies > 0.007798909675329924 & Monthly Contractor Hours per Mine <= 16191.3173828125 & Daily\_Rainfall\_mm\_No rain <= 0.2527591288089752 &  Probability of 'Minor': 0.5600  Probability of 'Moderate': 0.4000  Probability of 'Serious': 0.0400 | Rule:  Monthly Staff Hours per Mine <= 37239.0 & Incident\_time\_period\_Morning > 0.002650003181770444 & Incident\_time\_period\_Morning > 0.9984371364116669 & Monthly Contractor Hours per Mine <= 20596.0 & AgencyOfInjury\_Other and Unspecified Agencies > 0.015236075967550278 &  Probability of 'Minor': 0.3621  Probability of 'Moderate': 0.4655  Probability of 'Serious': 0.1724 | Rule:  Monthly Staff Hours per Mine <= 37239.0 & Incident\_time\_period\_Morning <= 0.002650003181770444 & BodyPart\_Upper Body <= 0.9938340187072754 & BodyPart\_Upper Body > 0.006742375437170267 &  Probability of 'Minor': 0.2113  Probability of 'Moderate': 0.0000  Probability of 'Serious': 0.7887 |
| 7 | Rule:  Monthly Staff Hours per Mine > 37239.0 & Monthly Contractor Hours per Mine <= 23134.6630859375 & AgencyOfInjury\_Other and Unspecified Agencies <= 0.007798909675329924 & Monthly Contractor Hours per Mine > 20169.0 & Daily\_Mean\_Temp\_degrees\_Cool <= 0.6361536681652069 &  Probability of 'Minor': 0.5748  Probability of 'Moderate': 0.1654  Probability of 'Serious': 0.2598 | Rule:  Monthly Staff Hours per Mine <= 37239.0 & Incident\_time\_period\_Morning > 0.002650003181770444 & Incident\_time\_period\_Morning > 0.9984371364116669 & Monthly Contractor Hours per Mine <= 20596.0 & AgencyOfInjury\_Other and Unspecified Agencies <= 0.015236075967550278 &  Probability of 'Minor': 0.0742  Probability of 'Moderate': 0.7161  Probability of 'Serious': 0.2097 |  |
| 8 | Rule:  Monthly Staff Hours per Mine > 37239.0 & Monthly Contractor Hours per Mine <= 23134.6630859375 & AgencyOfInjury\_Other and Unspecified Agencies <= 0.007798909675329924 & Monthly Contractor Hours per Mine <= 20169.0 & AgencyOfInjury\_Animal, Human and Biological Agencies > 0.0037671816535294056 &  Probability of 'Minor': 0.5455  Probability of 'Moderate': 0.4545  Probability of 'Serious': 0.0000 | Rule:  Monthly Staff Hours per Mine <= 37239.0 & Incident\_time\_period\_Morning <= 0.002650003181770444 & BodyPart\_Upper Body > 0.9938340187072754 & MechanismOfInjury\_Environmental Factors > 0.9702659249305725 &  Probability of 'Minor': 0.0517  Probability of 'Moderate': 0.9310  Probability of 'Serious': 0.0172 |  |
| 9 | Rule:  Monthly Staff Hours per Mine <= 37239.0 & Incident\_time\_period\_Morning > 0.002650003181770444 & Incident\_time\_period\_Morning <= 0.9984371364116669 & AgencyOfInjury\_Other and Unspecified Agencies > 0.7225707173347473 &  Probability of 'Minor': 0.5556  Probability of 'Moderate': 0.0000  Probability of 'Serious': 0.4444 | Rule:  Monthly Staff Hours per Mine <= 37239.0 & Incident\_time\_period\_Morning <= 0.002650003181770444 & BodyPart\_Upper Body > 0.9938340187072754 & MechanismOfInjury\_Environmental Factors <= 0.9702659249305725 & Daily\_Mean\_Temp\_degrees\_Cool > 0.00607723742723465 &  Probability of 'Minor': 0.1765  Probability of 'Moderate': 0.5686  Probability of 'Serious': 0.2549 |  |
| 10 | Rule:  Monthly Staff Hours per Mine <= 37239.0 & Incident\_time\_period\_Morning > 0.002650003181770444 & Incident\_time\_period\_Morning <= 0.9984371364116669 & AgencyOfInjury\_Other and Unspecified Agencies <= 0.7225707173347473 & AgencyOfInjury\_Animal, Human and Biological Agencies > 0.23405209928750992 &  Probability of 'Minor': 0.5357  Probability of 'Moderate': 0.0000  Probability of 'Serious': 0.4643 | Rule:  Monthly Staff Hours per Mine <= 37239.0 & Incident\_time\_period\_Morning <= 0.002650003181770444 & BodyPart\_Upper Body > 0.9938340187072754 & MechanismOfInjury\_Environmental Factors <= 0.9702659249305725 & Daily\_Mean\_Temp\_degrees\_Cool <= 0.00607723742723465 &  Probability of 'Minor': 0.0596  Probability of 'Moderate': 0.7660  Probability of 'Serious': 0.1745 |  |
| 11 |  | Rule:  Monthly Staff Hours per Mine <= 37239.0 & Incident\_time\_period\_Morning <= 0.002650003181770444 & BodyPart\_Upper Body <= 0.9938340187072754 & BodyPart\_Upper Body <= 0.006742375437170267 & AgencyOfInjury\_Other and Unspecified Agencies > 0.1743163913488388 &  Probability of 'Minor': 0.4156  Probability of 'Moderate': 0.4416  Probability of 'Serious': 0.1429 |  |
| 12 |  | Rule:  Monthly Staff Hours per Mine <= 37239.0 & Incident\_time\_period\_Morning <= 0.002650003181770444 & BodyPart\_Upper Body <= 0.9938340187072754 & BodyPart\_Upper Body <= 0.006742375437170267 & AgencyOfInjury\_Other and Unspecified Agencies <= 0.1743163913488388 &  Probability of 'Minor': 0.1423  Probability of 'Moderate': 0.7155  Probability of 'Serious': 0.1423 |  |

*Table 16: Full list of decision tree rules for the Canadian Dataset*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Minor | Moderate | Serious |
| 1 | Rule:  Monthly Staff Hours per Mine <= 79996.0 & NatureOfInjury\_Other or Unspecified > 0.04530321806669235 & BodyPart\_Upper Body <= 0.9884687960147858 &  Probability of 'Minor': 0.5273  Probability of 'Moderate': 0.4545  Probability of 'Serious': 0.0182 | Rule:  Monthly Staff Hours per Mine <= 79996.0 & NatureOfInjury\_Other or Unspecified <= 0.04530321806669235 & Daily\_Mean\_Temp\_degrees\_Warm <= 0.13900060951709747 &  Probability of 'Minor': 0.0215  Probability of 'Moderate': 0.8387  Probability of 'Serious': 0.1398 | Rule:  Monthly Staff Hours per Mine > 79996.0 & AgencyOfInjury\_Environmental Agencies <= 0.00838455744087696 & BodyPart\_Upper Body > 0.5 &  Probability of 'Minor': 0.2115  Probability of 'Moderate': 0.3846  Probability of 'Serious': 0.4038 |
| 2 | Rule:  Monthly Staff Hours per Mine > 79996.0 & AgencyOfInjury\_Environmental Agencies > 0.00838455744087696 &  Probability of 'Minor': 0.4648  Probability of 'Moderate': 0.2113  Probability of 'Serious': 0.3239 | Rule:  Monthly Staff Hours per Mine <= 79996.0 & NatureOfInjury\_Other or Unspecified <= 0.04530321806669235 & Daily\_Mean\_Temp\_degrees\_Warm > 0.13900060951709747 &  Probability of 'Minor': 0.0364  Probability of 'Moderate': 0.6364  Probability of 'Serious': 0.3273 |  |
| 3 |  | Rule:  Monthly Staff Hours per Mine <= 79996.0 & NatureOfInjury\_Other or Unspecified > 0.04530321806669235 & BodyPart\_Upper Body > 0.9884687960147858 &  Probability of 'Minor': 0.1000  Probability of 'Moderate': 0.7889  Probability of 'Serious': 0.1111 |  |
| 4 |  | Rule:  Monthly Staff Hours per Mine > 79996.0 & AgencyOfInjury\_Environmental Agencies <= 0.00838455744087696 & BodyPart\_Upper Body <= 0.5 &  Probability of 'Minor': 0.2692  Probability of 'Moderate': 0.4615  Probability of 'Serious': 0.2692 |  |