Utilize NLP to Analyze Industrial Safety and Health

**Interim Report**

A Capstone Project

*Submitted by:*

Ajay Tarachand Parja

Nikhil Terwadkar

Rahul Kumar Ray

Rajiv Nair

In fulfilment of the requirements for the award of

Post Graduate Program in

Artificial Intelligence & Machine Learning

from:

Great Learning

&

TEXAS McCombs

The University of Texas at Austin

Mentor:

Mr. Aniket Chhabra

February 2025

[1. Summary 4](#_Toc189840559)

[2. Introduction 4](#_Toc189840560)

[1.1 Purpose and Goals 4](#_Toc189840561)

[1.2 Data Sources 4](#_Toc189840562)

[1.3 Introduction to NLP 4](#_Toc189840563)

[1.3.1 Generic Architecture of NLP 5](#_Toc189840564)

[1.3.2 Applications for NLP Models 5](#_Toc189840566)

[1.3.2.1 The Role of Natural Language Processing in Workplace Safety 5](#_Toc189840567)

[1.4 Methodology 7](#_Toc189840568)

[1.4.1 Data Collection and Preprocessing 7](#_Toc189840569)

[1.4.1.1 Importing the data 7](#_Toc189840570)

[1.4.1.1.1 Display loaded data and a few records 8](#_Toc189840571)

[1.4.1.2 Data Exploration and Cleansing to handle inconsistencies 8](#_Toc189840572)

[1.4.1.3 Feature Engineering 9](#_Toc189840573)

[1.4.1.4 Exploratory Data Analysis (EDA) 9](#_Toc189840574)

[1.4.1.5 Univariate Analysis 10](#_Toc189840575)

[1.4.1.5.1 Sector/Industry, Region and Personnel Analysis 11](#_Toc189840576)

[1.4.1.5.2 Other Key Value Analysis 12](#_Toc189840577)

[1.4.1.6 Bivariate Analysis 13](#_Toc189840578)

[1.4.1.6.1 Accident Level and Country 13](#_Toc189840579)

[1.4.1.6.2 Accident Level and Sector 14](#_Toc189840580)

[1.4.1.6.3 Accident Level and Gender 14](#_Toc189840581)

[1.4.1.6.4 Accident Level and Employee Type 15](#_Toc189840582)

[1.4.1.7 Time Series Analysis 16](#_Toc189840583)

[1.4.2 Model Building 18](#_Toc189840584)

[1.4.2.1 Data Preprocessing (NLP Preprocessing techniques) 19](#_Toc189840585)

[1.4.2.1.1 Crete a copy of Data frame and One Hot Encoding 19](#_Toc189840586)

[1.4.2.2 Apply comprehensive Text preprocessing 20](#_Toc189840587)

[1.4.2.3 Word cloud to identify recurring word themes 21](#_Toc189840588)

[1.4.2.4 N-Grams 23](#_Toc189840589)

[1.4.2.4.1 Unigrams 23](#_Toc189840590)

[1.4.3 Data Preparation 25](#_Toc189840591)

[1.4.3.1 Bag of Words 25](#_Toc189840592)

[1.4.3.2 TF-IDF 26](#_Toc189840593)

[1.4.3.3 Store Cleaned Data into CSV 27](#_Toc189840594)

[1.4.4 Design train and test basic machine learning classifiers 27](#_Toc189840595)

[1.4.4.1 Train with Bag of Words (Base Model) 27](#_Toc189840596)

[1.4.4.1.1 Utilising SMOTE to address class imbalances 29](#_Toc189840597)

[1.4.4.1.2 Confusion Matrix 30](#_Toc189840598)

[1.4.4.2 Train with TF-IDF (Base Model) 31](#_Toc189840599)

[1.4.4.2.1 Utilising SMOTE 32](#_Toc189840600)

[1.4.4.2.2 Confusion Matrix 33](#_Toc189840601)

[1.4.4.3 Principal Component Analysis 33](#_Toc189840602)

[1.4.4.4 Hyperparameter Tuning 34](#_Toc189840603)

[1.5 Future Recommendations 38](#_Toc189840604)

[1.6 References 38](#_Toc189840605)

[2 Appendix 39](#_Toc189840606)

[2.1 Problem Statement 39](#_Toc189840607)

[2.2 Context: 39](#_Toc189840608)

[2.3 Data Description: 39](#_Toc189840609)

[2.4 Project Objective: 39](#_Toc189840610)

[2.4.1 Project Task: 40](#_Toc189840611)

[**2.5** Acronyms and Definitions 40](#_Toc189840612)

# Summary

This report presents the findings of a capstone project that leverages Natural Language Processing (NLP) to analyze industrial safety and health data. The project's primary goal is to identify potential safety risks based on incident descriptions and provide actionable insights to improve workplace safety. The analysis encompassed data from multiple manufacturing plants across three countries, detailing occurrences of accidents, their severity, and contributing factors. The report highlights the methodologies used, the findings, and future recommendations to enhance the study's accuracy and applicability.

# Introduction

## Purpose and Goals

The purpose of this capstone project is to utilize NLP techniques to analyze industrial safety and health data and identify potential safety risks. The project aims to design a machine learning model that can process incident descriptions and highlight safety risks, providing valuable insights to industry professionals to prevent future accidents.

The detailed Problem Statement, Project Objective and Tasks are available in the Appendix section of the document.

## Data Sources

The dataset used for this project originates from a major industry in Brazil and includes records of accidents from 12 different plants in three countries. Each record details the time, location, severity, and description of the accident, among other factors.

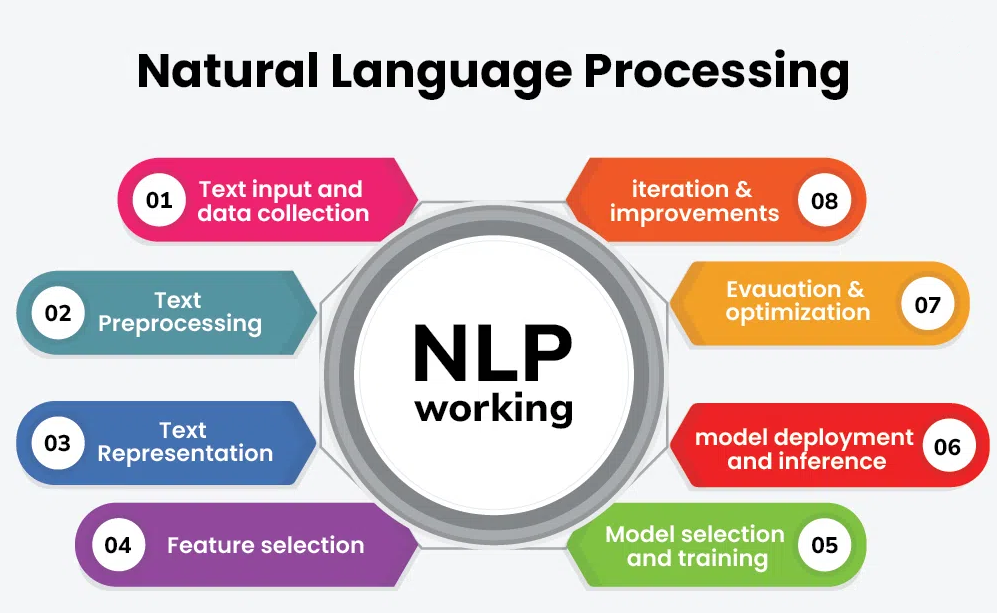
## Introduction to NLP

Natural Language Processing (NLP) is a branch of artificial intelligence that focuses on the interaction between computers and human language. By enabling machines to understand, interpret, and generate human language, NLP facilitates numerous applications across various fields. In the realm of industrial safety, NLP plays a crucial role in parsing and analyzing vast amounts of textual data related to incident reports and safety logs.

The importance of NLP in this context cannot be overstated, as it allows for the automation of data analysis, uncovering patterns and insights that might be overlooked by manual inspection. With its ability to process and analyze unstructured data, NLP can identify potential safety risks from incident descriptions, categorize them by severity, and highlight contributing factors. This capability is vital for developing proactive safety measures, as it provides industry professionals with actionable insights to prevent future accidents and enhance overall workplace safety.

### Generic Architecture of NLP

The architecture of an NLP system generally consists of several key components, each playing a specific role in the processing pipeline. Below is a diagram illustrating the generic architecture of an NLP system:



### *NLP Image sourced from © GeeksforGeeks*

### Applications for NLP Models

NLP models have many applications in various domains and industries, such as search engines, chatbots, voice assistants, social media analysis, text mining, information extraction, natural language generation, machine translation, speech recognition, text summarization, question answering, sentiment analysis, and more.

#### The Role of Natural Language Processing in Workplace Safety

In safety, employees in many works settings face great risk when there is an accident or hazard at work. Conventional safety management depends on written reports and is always a response action which means most problems have already happened when addressed.

|  |  |  |
| --- | --- | --- |
| # | NLP Application Scenarios | NLP can help by: - |
| 1 | Smarter incident reporting | * Automatically extracting key details like the event, location, and severity * Classifying incidents like slips, equipment failures, and exposure to hazardous materials. * Patterns in reports can be spotted to highlight recurring safety issues before they become larger problems. |
| 2 | Predicting risks before they happen | Analyze thousands of past accidents reports to help predict where the next hazard might occur. NLP can look at past incidents and environmental conditions too:   * Identify early warning signs of safety risks. * Flag anomalies in how safety procedures or equipment are used. * Provide recommendations to prevent future incidents. |
| 3 | Keeping an eye on workplace communication | Safety issues are usually raised in emails, logs, or team discussions, but can easily be missed. NLP can analyze these communications to:   * Detect potential safety concerns employees talk about. * Identify shifts in how people feel about workplace safety. * Alert managers to possible safety violations before they escalate. |
| 4 | Simplifying compliance with Safety Regulations | Workplaces must adhere to very strict safety rules, but it is not easy to keep up with changing regulations:   * Automatic checking of safety documents against statutory requirements. * Highlighting inconsistencies in policies and procedures. * Condensing compliance issues for easy access by the managers to address them. |
| 5 | Making safety training more interactive | The safety training must be undergone for long hours by employees. NLP can make this easier by:   * Converting thick safety books into conversational chatbots or voice assistants. * Condensing vital policies for quick reference by workers. * Immediate answers to questions relating to safety. |
| 6 | Safety alerts with Real-Time Effect | When combined with IoT sensors and other monitoring tools, NLP can help workplaces stay ahead of potential dangers by:   * Sending real-time alerts based on safety data and reports. * Offering instant risk assessments for hazardous areas. * Providing safety recommendations tailored to individual workers. |

.

## Methodology

### Data Collection and Preprocessing

The dataset comprises various columns, including timestamps, locations, industry sectors, accident levels, potential accident levels, gender, employee status, critical risks, and descriptions of accidents.

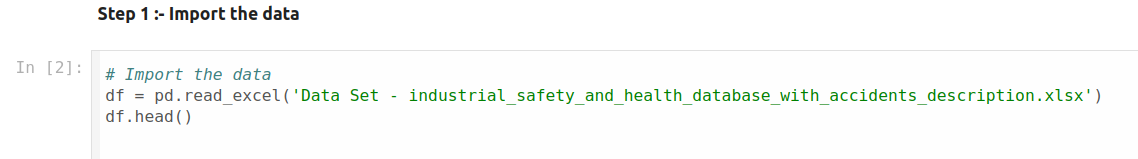
#### Importing the data

The data is imported from Google Drive.

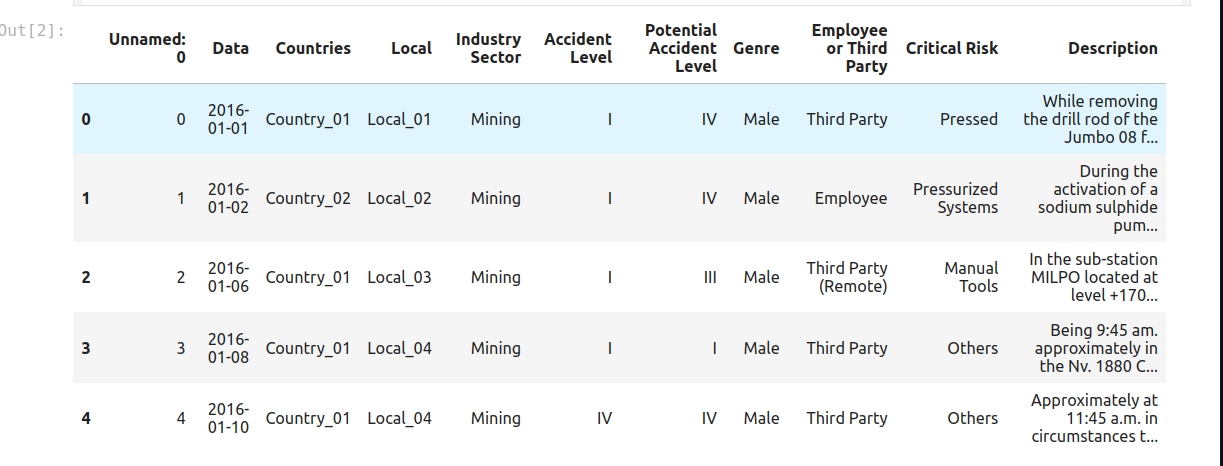
*# Mount google drive*

*from google.colab import drive*

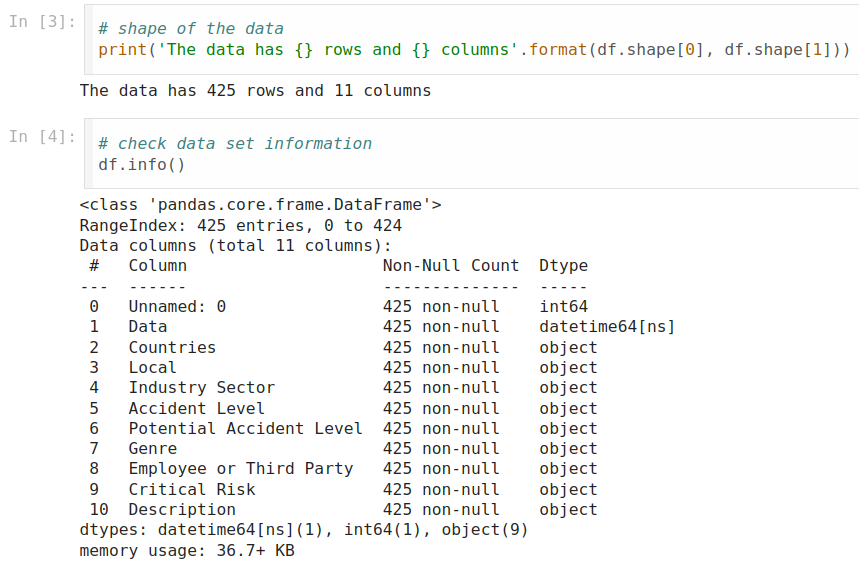
*drive.mount('/content/drive')*



##### Display loaded data and a few records



#### Data Exploration and Cleansing to handle inconsistencies



**Column to be deleted**

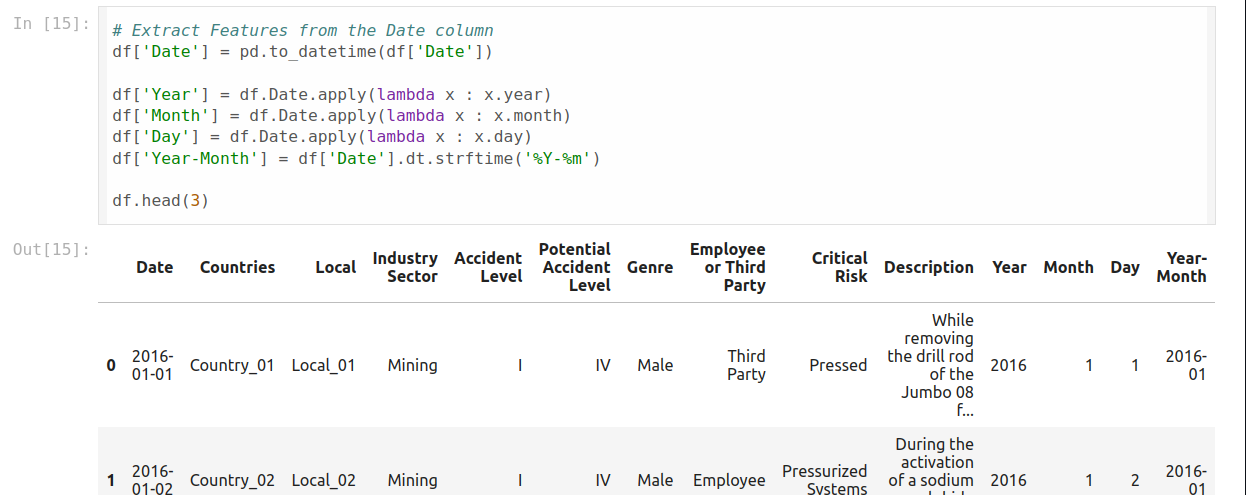
The dataset has no null values in any column. The 'Data' column is in datetime format. The first column appears to represent the record numbers and can be dropped as it does not contribute to the dataset features. The other columns are categorical or object types.

**Checks and remediations carried out include: -**

The analysis led to finding of **True Duplicate** records. These records were deleted to improve data consistency which may have led to redundant data, potentially leading to overfitting, inaccurate predictions and a distorted understanding of the relationships within the data.

|  |  |  |
| --- | --- | --- |
| # | Checks | Remediations |
| 1 | Duplicate Records | Duplicate records were deleted including |
| 2 | Identifying True Duplicates or identical duplicates | True Duplicates were identified and deleted |
| 3 | Check for (NA) Not available values | No records with NA values were found |
| 4 | Check for empty or blanks | None were found |
| 5 | Renaming columns for relevance and apt representation | 1. Data to Date 2. Genre to Gender 3. Countries to Country 4. Employee or Third Party to Employee Type |

#### Feature Engineering

Splitting of Date into Year, Month and Day was carried out to extract meaningful temporal features to significantly improve the model's ability to capture seasonal trends, cyclical patterns and time related relationships.

#### Exploratory Data Analysis (EDA)

To improve accuracy and identify data patterns a series of analysis was conducted. These included identifying unique values.

The following are the broad details of the unique value analysis

* There are records of accidents from 1st Jan 2016 to 9th July 2017.
* The plant is located at 12 cities which belong to 3 countries.
* Data available is related to 3 Industry Sectors like Metals, Mining and others.
* Each accident can be classified into any one of the five Accident Level. Higher the Accident Level Higher the severity.
* Accidents are related to both Male and Female.
* Employees are classified as into three categories such as Employee, Third Party and Third Party (Remote).
* Critical Risk feature contains most unique values and maybe it can explain the Accident Level along with the Description.

#### Univariate Analysis

Univariate analysis is a statistical method that examines a single variable in a data set

A screenshot of a computer

Description automatically generated

A group of pie charts

Description automatically generated

##### Sector/Industry, Region and Personnel Analysis

A graph of different colored squares

AI-generated content may be incorrect.

##### Other Key Value Analysis

A screen shot of a computer

AI-generated content may be incorrect.

A graph with different colored squares

AI-generated content may be incorrect.

Insights

* Most Accident occurred in Country - 01.
* The Mining industry sector is prone to the Accidents as 57% of accidents are related to Mining Industry
* Nearly 95% of the accidents were caused by Male Employee. However, as data is related to Mining and Metal Industry this percentage seems ok as most of workers will be Male.
* Employees on the payroll and Third-Party Employee equally contribute to the number of accidents. Very few accidents were caused by the Remote Employees.
* Least number of accidents occurred in local-9 and local 11 which is 2.
* For more than half of the Accident the Critical Risk is mentioned as others. Which means for most of the accident critical risk involved is too large or cannot be determined. Here SME can help the industry to assess the critical risk involved in the accident.
* Out of 424 accidents 315 accidents are of accident level 1 which means nearly 75% of the accidents are of least severity
* From Machine Learning perspective clearly, there is significant class imbalance in the target variable.

#### Bivariate Analysis

Bivariate analysis is a statistical method used to investigate the relationship between two variables.

##### Accident Level and Country

A screenshot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

##### Accident Level and Sector

A colorful chart with text

AI-generated content may be incorrect.

##### Accident Level and Gender

A graph with different colored squares

AI-generated content may be incorrect.

##### Accident Level and Employee Type

A graph of different colored squares

AI-generated content may be incorrect.

Insights

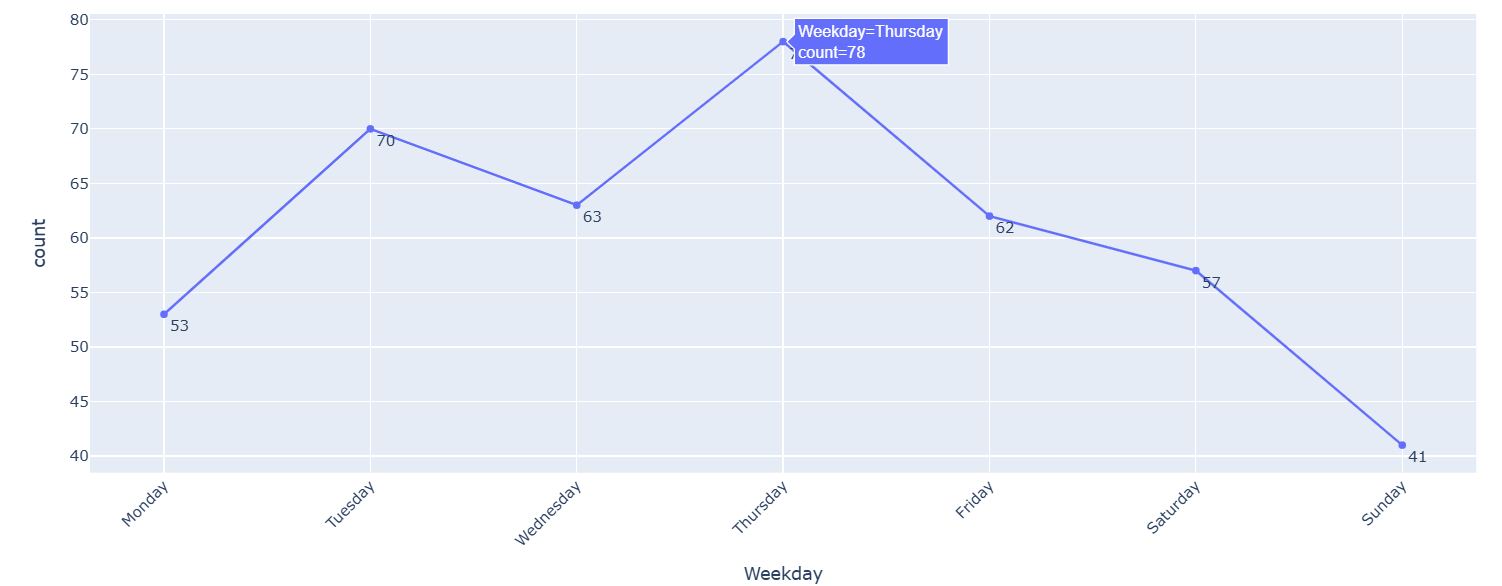
* Across most categories (Countries, Local, Industry Sector, Potential Accident Level, Genre, Employee or Third Party), Accident Level I is the most frequent, indicating a higher occurrence of less severe accidents. This suggests that most safety measures are effective in preventing major incidents.
* Country 2 exhibits a higher proportion of Level I, II and III accidents. This highlights potential areas for focused safety interventions and risk mitigation strategies.
* Country 1 exhibits a higher proportion of among all level of accidents compared to other countries. This highlights potential areas for focused safety interventions and risk mitigation strategies.
* Mining Sector Has Most Accidents, Metals Sector Shows Higher Severity: The 'Mining' sector has the highest number of accidents across all levels, indicating a higher risk environment. The 'Metals' sector, while having fewer accidents overall, has a relatively higher proportion of Level IV accidents, suggesting a higher potential for severe incidents.
* Potential Accident Level Strongly Correlated with Actual Accident Level: There's a strong correlation between 'Potential Accident Level' and 'Accident Level', meaning that higher potential accident levels are associated with higher actual accident levels. This emphasizes the importance of proactive risk assessment and addressing potential hazards before they escalate.
* Males and Employees Involved in Most Accidents: Most accidents involve males and employees. While this could be influenced by workforce demographics, it is important to ensure safety measures are inclusive and address the specific needs of all employee groups.
* Third Parties Involved in More Severe Accidents: Level IV and V accidents have a greater proportion involving third parties compared to other levels. This suggests the need for safety protocols and training to address risks involving external personnel.

#### Time Series Analysis

Time series analysis is a statistical method that uses data points collected at regular intervals to identify patterns and trends

A screenshot of a computer code

AI-generated content may be incorrect.

A graph with a line

AI-generated content may be incorrect.

Insights

* The source data timeline ranges from Jan 2016 to Jul 2017.
* Majority of the accidents are type level 1.
* In the Year 2016 Number of Accidents increased till Jun 2016 and then suddenly dropped in Jul 2016.
* Maximum accidents of 30 occurred in the month Jun 2016. In 2017 started with 21 accident which was maximum in that year.
* In 2017, starting from Jan 2017 accidents dropped and again increased in the subsequent month till Jul 2017.
* The most frequent work-related events that have occurred on Monday, Tuesday, and Thursday are:
  1. Cleaning (20 incidents) – Highest occurrence.
  2. Maintenance (14 incidents) – Regular maintenance activities reported.
  3. Cutting (12 incidents) – Cutting operations leading to incidents.
  4. Lifting (8 incidents) – Issues related to lifting heavy objects.
  5. Drilling (8 incidents) – Drilling-related incidents.
* This suggests that cleaning and maintenance tasks are the most accident-prone these days.
  + Workload and Scheduling Patterns

Vicious Circle Entrapment

* + Fatigue and Attention Levels
    - Monday: Employees may still be adjusting to work routines after the weekend, leading to lower concentration and increased accidents.
    - Thursday: Fatigue from a long workweek may begin setting in, affecting alertness and safety adherence.
  + Operational Priorities
    - Cleaning and maintenance often happen at the beginning and middle of the week to ensure smooth production workflows.
    - Drilling, cutting, and lifting operations might be scheduled early in the week when critical raw materials and setups are required.
  + Shift and Work Cycle Effects
* Possibly, different industry sectors might have higher activity levels on certain days, leading to more frequent incidents. For instance, loading/unloading tasks may be high on Monday (post-weekend material deliveries) and Thursday (end-of-week shipments).
  + Safety Protocol Lapses
* If safety briefings or equipment checks are skipped due to urgency in getting started on Monday, issues may arise.
* By Thursday, workers may become complacent, thinking the weekend is near, leading to procedural lapses.

**Recommendations based on EDA**

* Focus on Preventive Measures: Given the predominance of Accident Level I, prioritize preventive measures to maintain a low rate of less severe accidents.
* Targeted Interventions: Implement targeted safety interventions in Country 1 and 2 to address the higher proportion of Level III and IV accidents.
* Mining and Metals Sector Safety: Enhance safety protocols and training in the 'Mining' and 'Metals' sectors, focusing on risk assessment and hazard identification.
* Proactive Risk Management: Prioritize proactive risk management and address potential hazards before they escalate into actual accidents, as indicated by the correlation between 'Potential Accident Level' and 'Accident Level'.
* The combination of workload distribution, fatigue, and operational cycles contributes to the observed pattern of incidents. Companies could strengthen safety monitoring and awareness programs on these high-risk days to mitigate accidents
* Inclusive Safety: Ensure safety measures are inclusive and address the specific needs of all employee groups, including males and females.
* Third-Party Safety: Develop safety protocols and training to address risks involving third parties, particularly in higher-risk situations.

### Model Building

The project utilized machine learning classifiers to analyze the data. The models were trained to predict safety risks based on accident descriptions. The steps involved in model building included:

* Designing and training basic machine learning classifiers
* Evaluating model performance using metrics such as accuracy and loss
* Fine-tuning models to improve performance

**Target Variable**: - Accident Level

**Feature**: - Date, Countries, Local, Industry sector, Potential Accident Level, Gender, Employee Type, Critical Risk, Description

#### Data Preprocessing (NLP Preprocessing techniques)

##### Crete a copy of Data frame and One Hot Encoding



**One Hot Encoding**

Various Machine Learning models do not work with categorical data and to fit this data into the machine learning model it needs to be converted into numerical data. For example, suppose a dataset has a *Gender*column with categorical elements like *Male and Female*. One Hot Encoding in machine learning transforms categorical data into a numerical format that machine learning algorithms can process without imposing any ordinal relationships. (See Image below)

A screenshot of a computer

Description automatically generated

One Hot Encoded

#### Apply comprehensive Text preprocessing

A computer screen shot of a computer code

Description automatically generated

Sample Output

A screenshot of a computer

Description automatically generated

#### Word cloud to identify recurring word themes

A word cloud helps by visually representing the frequency of words within a text, allowing you to quickly identify the most important themes, concepts, or keywords in a large body of data briefly, making it easier to understand the overall sentiment or key points of a document or set of information.

A computer screen shot of words

Description automatically generated

Insights

Key Themes:

1. Safety and Incidents: Words like "injury," "hit," "fall," "hurt," and "accident" suggest a focus on workplace injuries and safety incidents.
   * Employee Involvement: Terms such as "employee," "worker," and "collaborator" indicate that the data involves reports and descriptions provided by employees or remote workers.
   * Equipment and Tools: Words like "equipment," "tool," "machine," and "vehicle" highlight the involvement of machinery and tools in the incidents.
2. Common Incidents:
   * Falls and Impact: Words like "fall," "hit," and "impact" suggest that falls and being struck by objects are common types of incidents.
   * Body Parts Affected: Terms such as "hand," "finger," "head," and "face" indicate the body parts most frequently injured.
3. Causes and Actions:
   * Causative Factors: Words like "cause," "move," "use," and "perform" point to actions or movements that may have led to the incidents.
   * Preventive Measures: Terms like "assist," "support," and "protect" suggest efforts to prevent or mitigate incidents.
4. Reporting and Documentation:
   * Descriptive Language: Words like "describe," "report," and "moment" indicate that the data includes detailed descriptions of incidents, possibly for reporting or analysis purposes.
5. Work Environment:
   * Industrial Setting: Terms like "platform," "pipe," and "ladder" suggest an industrial or construction environment where these incidents occur.

Overall, the word cloud provides a snapshot of the key concerns and themes related to workplace safety, highlighting the types of incidents, affected body parts, and the equipment involved.

#### N-Grams

N-grams are used in natural language processing (NLP) tasks like spell-checking, text prediction, and language modeling.

A screenshot of a computer program

AI-generated content may be incorrect.

##### Unigrams

A screen shot of a computer code

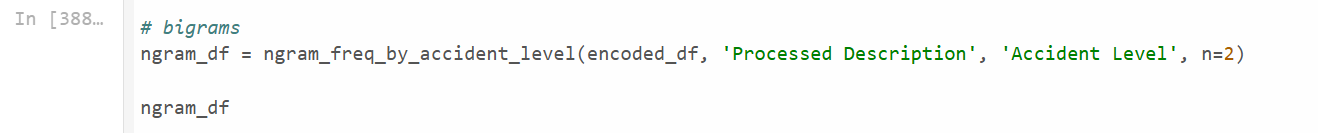
AI-generated content may be incorrect.

Sample Output

A screenshot of a computer

Description automatically generated

Bigrams

Sample Output

A screenshot of a phone

AI-generated content may be incorrect.

Insights

* When we look at the count of unigrams or bigrams with respect to Accident Level, we can see that Accident Levels 1-3 are related to employee and mostly involved the right/left hand.
* When it comes to sever accidents where accident level 4-5 mostly an operator is involved and particularly for accident level 5 mixer truck is involved.

### Data Preparation

#### Bag of Words

A method to convert text into numerical data for natural language processing.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

#### TF-IDF

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

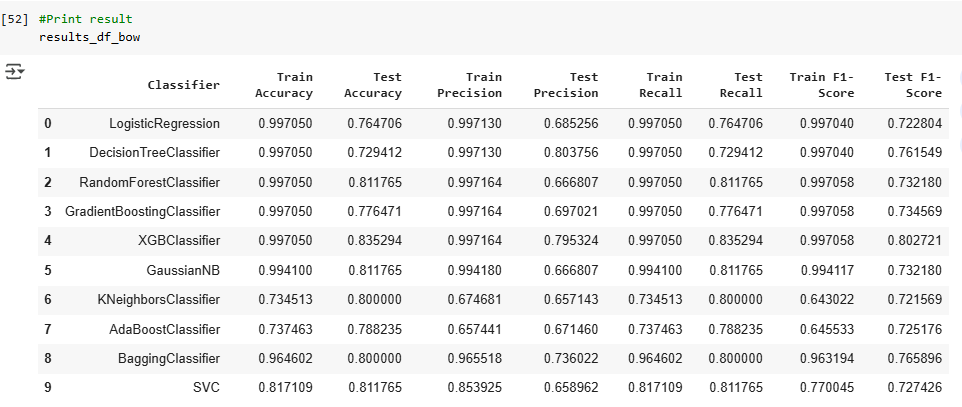
#### Store Cleaned Data into CSV

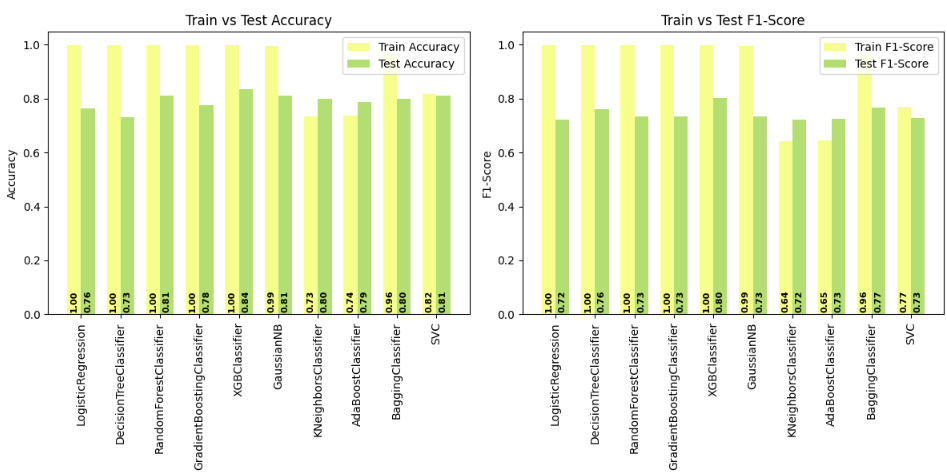
A screenshot of a computer

AI-generated content may be incorrect.

### Design train and test basic machine learning classifiers

#### Train with Bag of Words (Base Model)





##### Utilising SMOTE to address class imbalances

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

* When oversampling is applied using SMOTE, we did not see the substantial improvement in any of the base model performance.
* Using Bag of Words, the Adaboost Classifier achieved an F1-Score of 72.51% in both training and testing phases.

##### Confusion Matrix

We created a confusion matrix to thoroughly evaluate the performance of a classification model for a detailed breakdown of how many predictions were correct and incorrect, allowing us to identify specific areas where the model is struggling and calculate important metrics like precision and recall, which are not captured by simply looking at overall accuracy alone.

A screenshot of a graph

AI-generated content may be incorrect.

* Though Adaboost Classifier performed well in both Train and Test with respect to F1-Score from the Confusion Matrix above it is evident that it misclassifies the minority class into majority class.

#### Train with TF-IDF (Base Model)

A screenshot of a graph

AI-generated content may be incorrect.

A comparison of a graph

AI-generated content may be incorrect.

##### Utilising SMOTE

A table with numbers and letters

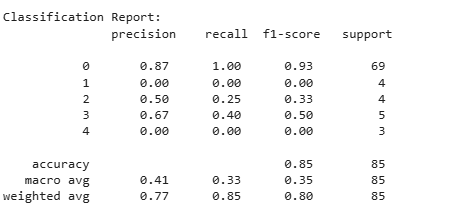
AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

* When oversampling is applied using SMOTE, we did not see the substantial improvement in any of the base model performance.
* So, with TF-IDF **Logistic Regression** performed well both in Train and Test with **highest F1-Score approx. 80%\*\***

##### Confusion Matrix



A diagram of a confusion matrix

AI-generated content may be incorrect.

* Though Logistic Regressor performed well in both Train and Test with respect to F1-Score from the Confusion Matrix above it is evident that it misclassifies the minority class into majority class.

#### Principal Component Analysis

Principal Component Analysis (PCA) is an essential technique not only for managing large datasets but also for highlighting significant patterns and trends in smaller datasets. By reducing the dimensionality of the data, even if the dataset isn't excessively large, hence we would use PCA in identifying the most influential variables, thus simplifying the model and potentially enhancing its performance and interpretability.

A table with numbers and letters

AI-generated content may be incorrect.

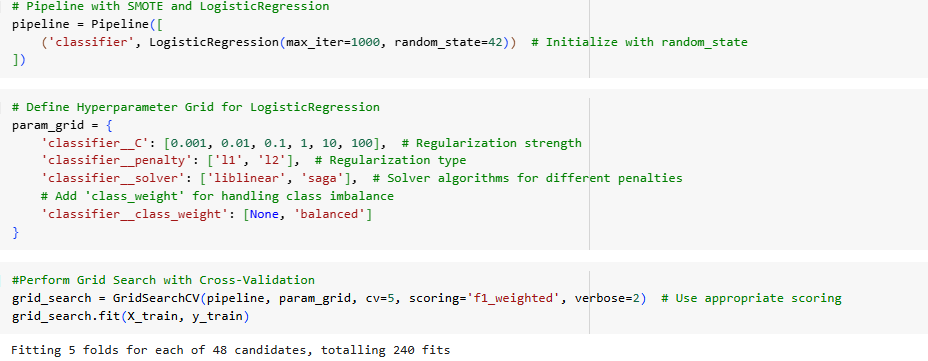
A graph of different colored bars

AI-generated content may be incorrect.

* PCA doesn't seems improving balance.
* Since there is insignificant improvement, it is a prudent measure to carry out hyperparameter tuning

#### Hyperparameter Tuning

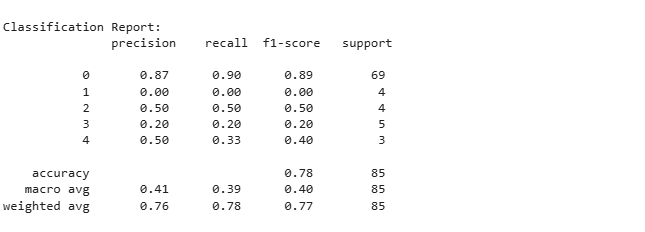
Hyperparameter tuning is the process of optimizing a model's parameters to improve its performance. Given we have seen LogisticRegression shows better accuracy, using logistic regression specific hyperparameter and Grid Search tuning method, we will evaluate the performance improvement.



Best Hyperparameter with associated classification report and confusion matrix: -

A screenshot of a computer code

AI-generated content may be incorrect.



A diagram of a confusion matrix

AI-generated content may be incorrect.

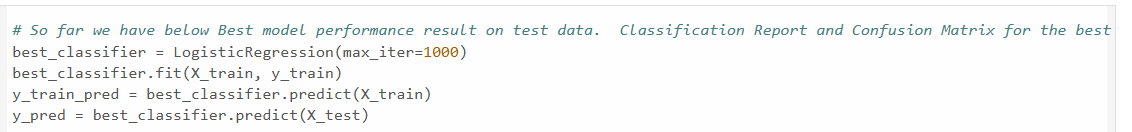
Insights

* With hyperparameter tuning also, we are not seeing performance improved compared to base model.
* Considering dataset has imbalance class, the F1-score is more informative. It considers both precision and recall, giving a better picture of the model's performance on the minority class too.
* Logistic Regression without oversampling so far is having the best F1-Score

**Best Model - Logistic Regression**

We conducted a thorough comparison using multiple performance metrics, including accuracy, precision, recall, and F1-score.

After an extensive evaluation and comparative analysis of various classification models, it has been conclusively determined that the Logistic Regression model outperforms other contenders and stands out as the most effective.



|  |  |
| --- | --- |
| classification\_report(y\_train, y\_train\_pred)) | classification\_report(y\_test, y\_pred)) |
|  |  |

Train Confusion Matrix Insights

* High Accuracy on Class 'I': The model is highly accurate in predicting class 'I' (Accident Level I - least severe), as indicated by the large number of correct predictions (True Positives) along the diagonal. This is expected since class 'I' is the most frequent class in the dataset.
* Misclassifications: The model has some misclassifications, particularly with classes 'II', 'III', 'IV', and 'V'. This is evident from the non-zero values in the off-diagonal cells. For instance, some samples belonging to class 'II' are misclassified as 'I' or 'III'.
* Class Imbalance Effect: The confusion matrix reflects the class imbalance in the dataset, with a larger number of predictions skewed towards class 'I'.

Test Confusion Matrix Insights

* Generalization to Unseen Data: The test confusion matrix shows how the model generalises unseen data. The patterns are generally like the train confusion matrix, suggesting that the model is not overfitting too severely.
* Slight Drop in Accuracy: There might be a slight drop in accuracy for some classes on the test data compared to the training data. This is expected as the model is evaluated on unseen samples.
* Challenges with Minority Classes: The model might still struggle with predicting the minority classes (II, III, IV, V) accurately, as indicated by the relatively lower number of true positives for these classes. Overall Insights and Recommendations:

Model Performance:

* Given the comprehensive analysis and the robust performance metrics, it is unequivocal that Logistic Regression is the optimal model for our classification task. Its balance of simplicity, interpretability, and superior performance metrics makes it an ideal choice. As a result, we recommend adopting Logistic Regression as the primary model for deployment and further fine-tuning to achieve the best possible outcomes.
* The model seems to perform well overall, particularly in predicting the majority class ('I'). However, there's room for improvement in predicting the minority classes.

## Future Recommendations

Operational recommendation

* Given the predominance of Accident Level I, prioritize preventive measures to maintain a low rate of less severe accidents.
* Implement targeted safety interventions in Country 1 and 2 to address the higher proportion of Level III and IV accidents.
* Enhance safety protocols and training in the 'Mining' and 'Metals' sectors, focusing on risk assessment and hazard identification.
* Prioritize proactive risk management and address potential hazards before they escalate into actual accidents, as indicated by the correlation between 'Potential Accident Level' and 'Accident Level'.
* The combination of workload distribution, fatigue, and operational cycles contributes to the observed pattern of incidents. Companies could strengthen safety monitoring and awareness programs on these high-risk days to mitigate accidents
* Ensure safety measures are inclusive and address the specific needs of all employee groups, including males and females.
* Develop safety protocols and training to address risks involving third parties, particularly in higher-risk situations.
* Enhance hand protection measures and training on safe equipment operation.
* Review and improve floor safety, housekeeping practices, and fall protection.
* Investigate the root causes of accidents and implement corrective actions.
* Provide comprehensive training on safe work practices and equipment use.
* Emphasize clear communication and coordination during collaborative tasks.

Model Recommendation

* Logistic Regression as the primary model for deployment given its outperforming result.
* Increase the dataset size by incorporating more records from additional sources
* Implement real-time data processing to provide immediate safety risk assessments
* To improve the model performance over the minority class we can try synthetic data generation.
* Along with synthetic data generation instead of BOW or TF-IDF we can try LLM's

## References

* Data Source: https://www.kaggle.com/ihmstefanini/industrial-safety-and-health-analytics-database

# Appendix

## Problem Statement

Domain:

Industrial safety, NLP-based Chatbot.

## Context:

The database comes from one of the biggest industries in Brazil and in the world. It is an urgent need for industries and companies around the globe to understand why employees still suffer from injuries and accidents in plants. Sometimes, they also die in such environments.

## Data Description:

* The database contains records of accidents from 12 different plants in 3 different countries. Each line in the data represents the occurrence of an accident.
* Data: Timestamp or time/date information
* Countries: The country where the accident occurred (anonymized)
* Local: The city where the manufacturing plant is located (anonymized)
* Industry sector: The sector to which the plant belongs
* Accident Level: from I to VI, it registers how severe the accident was (I means not severe but VI means very severe)
* Potential Accident Level: Depending on the Accident Level, the database also registers how severe the accident could have been (due to other factors involved in the accident)
* Genre: if the person is male or female
* Employee or Third Party: if the injured person is an employee or a third party
* Critical Risk: some description of the risk involved in the accident
* Description: Detailed description of how the accident happened.

## Project Objective:

* Design a ML/DL based chatbot utility which can help the professionals to highlight the safety risk as per the incident description.

### Project Task:

[Score: 100 points]

1. **Milestone 1: [Score: 40 points]**
   1. Input: Context and Dataset
   2. Process:
      1. Step 1: Import the data [3 points]
      2. Step 2: Data cleansing [5 points]
      3. Step 3: Data preprocessing (NLP Preprocessing techniques) [7 points]
      4. Step 4: Data preparation - Cleansed data in .xlsx or .csv file [5 points]
      5. Step 5: Design train and test basic machine learning classifiers [10 Points]
      6. Step 6: Interim report [10 points]
      7. Submission: Interim report, Jupyter Notebook with all the steps in Milestone-1
2. **Milestone 2: [Score: 60 points]**
   1. Input: Preprocessed output from Milestone-1
   2. Process:
      1. Step 1: Design, train and test Neural networks classifiers [5 points]
      2. Step 2: Design, train and test RNN or LSTM classifiers [10 points]
      3. Step 3: Choose the best performing classifier and pickle it. [5 points]
      4. Step 4: Final Report [40 Points]
      5. Submission: Final report, Jupyter Notebook with all the steps in Milestone-1 and Milestone-2
3. **Milestone 3: [Optional]**
   1. Process:
      1. Step 1: Design a clickable UI based chatbot interface
   2. Submission: Final report, Jupyter Notebook with the addition of clickable UI based interface\

## Acronyms and Definitions

* NLP – Natural Language Processing
* EDA – Exploratory Data Analysis
* BOW – Bag of words
* SMOTE – Synthetic Minority Oversampling Technique
* PCA – Principal Component Analysis
* TF-IDF – Term Frequency-Inverse Document Frequency

Team collaboration model