AIML

CAPSTONE PROJECT - INTERIM REPORT

NLP CHATBOT For Industry Safety

Team Mentor

Rohit Gupta

Team Members

Sarita Ghadge
Vikram Rai
Rohan Bhattacharya
Meetu Chandra

Problem Statement

Major Industries in Brazil are facing multiple workplace hazards leading to multiple accidents, some even leading to death. The aim of this project is to build an AI powered Chatbot for employees and stakeholders to understand the various reason for the accidents. This can help the employees take the necessary precautions and for the stakeholders to identify the root cause for the accidents and put a fix in place.

The Chatbot will be powered by an AI model that has been trained to classify the accident, given a summary of the various scenarios when the accident happened in past.

Approach

An NLP based AI model will be trained on industry data from **3 countries** and **12 plants.** The data is from industries operating majority in Metal and mining. Machine learning models using different strategies will be developed and compared to identify the best performing model.

Data Overview

We analysed data from one of the largest industries in Brazil, which consists of 425 accident records from January 2016 to September 2017 across 3 countries and 12 locations. The dataset includes key information provided in the table below

S. No.	Column Name	Description about the column
1	Unnamed	Index Column **Data type**: Integer
2	Data	Date of accident **Data type**: datetime **Range**: January 2016 to July 2017
3	Countries	which country the accident occurred (anonymised) **Data Type:** Object **Unique values**: Country_01 Country_02 Country_03
4	Local	The city where the manufacturing plant is located (anonymised) **Data Type**: Object **Unique values**: Local_01 to Local_12
5	Industry sector	Which sector the plant belongs to? **Data Type**: Object **Unique values**: Mining, Metas, Others
6	Accident level	From I to VI, it registers how severe was the accident (I means not severe but VI means very severe) **Data Type**: Object **Unique values**: I,II, III, IV, V
7	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) **Data Type**: Object **Unique values**: I,II, III, IV, V, VI
8	Gender	If the person is male of female **Data Type**: Object **Unique values**: Male, Female
9	Employee or Third Party	if the injured person is an employee or a third party **Data Type**: Object **Unique values**: Third Party Employee Third Party (Remote)
10	Critical Risk	Some description of the risk involved in the accident **Data Type**: Object **Unique values**: There are number of unique values like – could be a short description for the accident description.

11	Description	Detailed description of how the accident happened. **Data Type**: Object (textual data)

The below steps were performed

Exploratory Data Analysis (EDA)

Initial Analysis:

There were 425 records with 10 columns. None of the columns had missing values. All the columns except for Date and description were categorical.

Renaming columns and removal from initial analysis

Few of the column names had to be renamed to as per the data they were representing: -

- "Data" was renamed to Date
- "Genre" was renamed to Gender.
- "Unnamed" which was an index columns was dropped.

Check for Missing values

There were no missing values in the dataset

Check for Duplicates

- There were 7 records that were duplicate and the duplicate records were dropped. That left us with 418 records.
- Out of the remaining 418 records, 7 accident descriptions are repeated in the data.
- These are accidents which happened at the same time where a group was invovled and there are different records for each person.
- As this corresponds to different records they were not removed

Unique values of the categorical columns

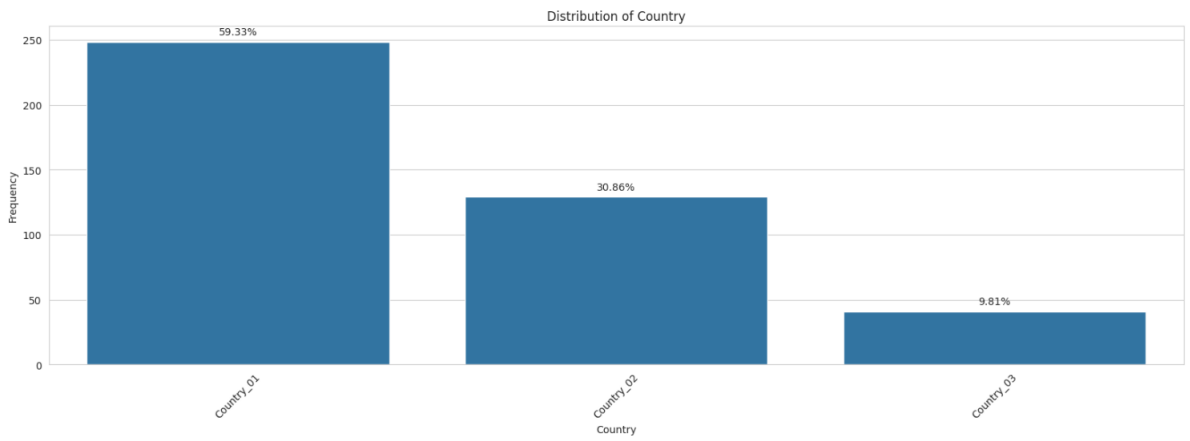
Unique values for the categorical columns were identified (The values are available in the above table).

Univariate Analysis

Count plots were used to understand the spread of data across each feature.

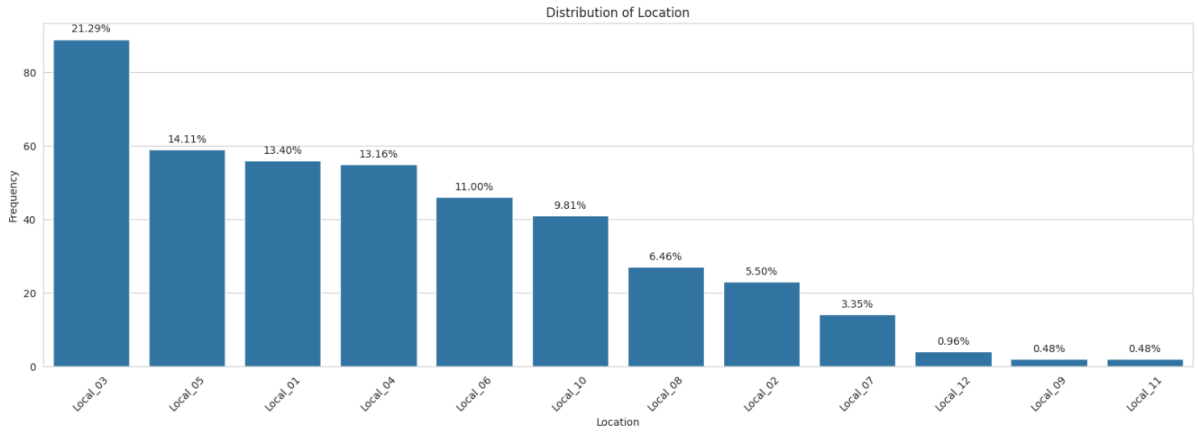
Country

There is data for 3 countries, 60% of the data is for Country_01, 31% for countrty_02 and only 10% for country_03. It is possible country_01 is more prone to accidents.



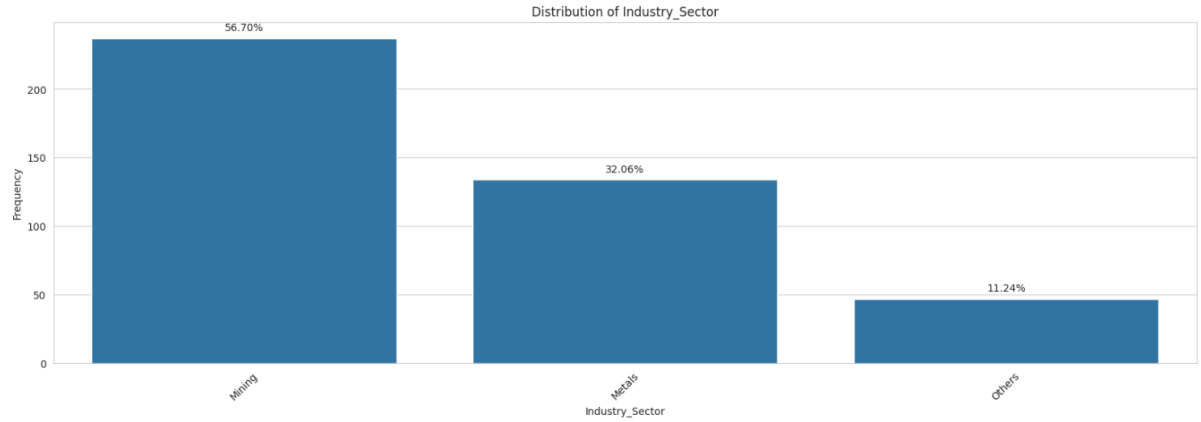
Location

There is data for 12 locations across 3 countries. Local_03 has seen the maximum number of accidents, which is around 20% of all the accident cases recorded.



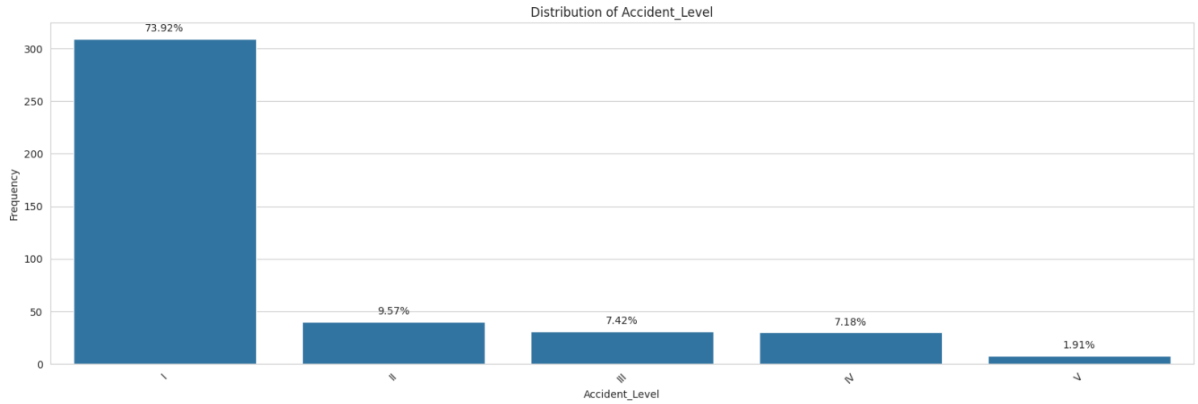
Industry Sector

Mining sector has the most accident cases than any other sector. Thus, we can say that jobs in the mining industry sector are riskier than metal or any other sector.



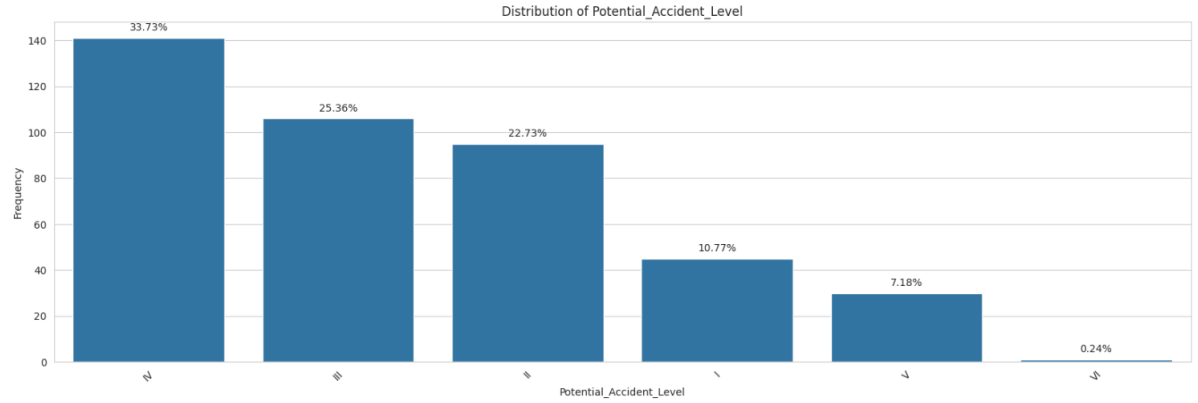
Accident Levels

Accident levels are mostly of Level 1 with 74% of the data, followed by 9.57% of level 2 and ~7% for levels 3 and 4 and ~2% for level 5. There have been no accidents of level 6 which is the highest level.



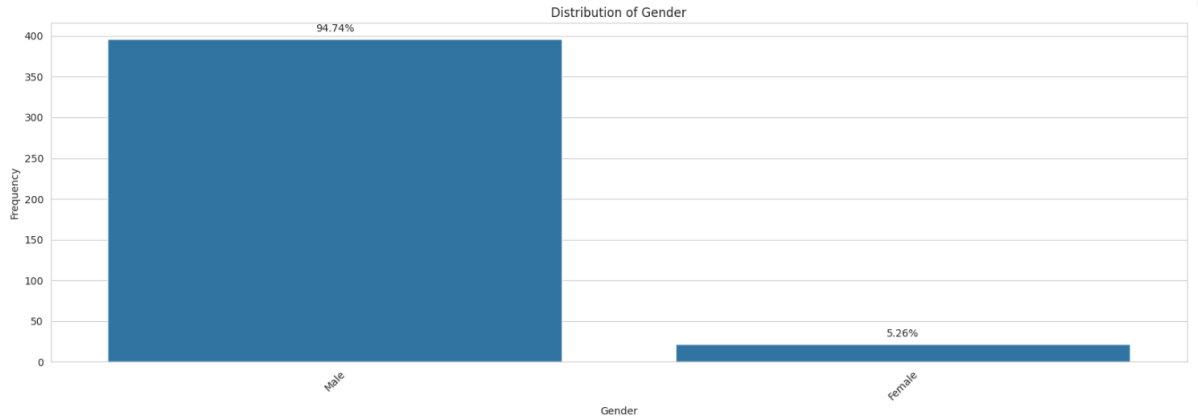
Potential Accident Level

Potential accident level indicates how severe the accident would have been due to other factors involved in the accidents. As per the graph, level IV has the highest count, which corresponds to moderate severity of accidents, followed by 25.3% of level 3. Also 0.24% chances of the most severe level 6.



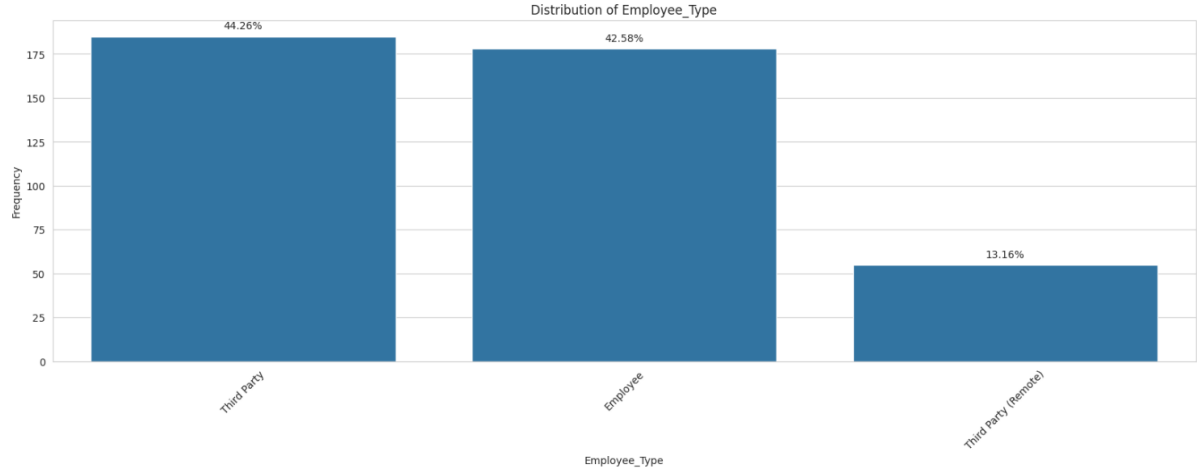
Gender

Dataset is more biased towards male employees; this is possible because Mining and Metal industries are more male dominant.



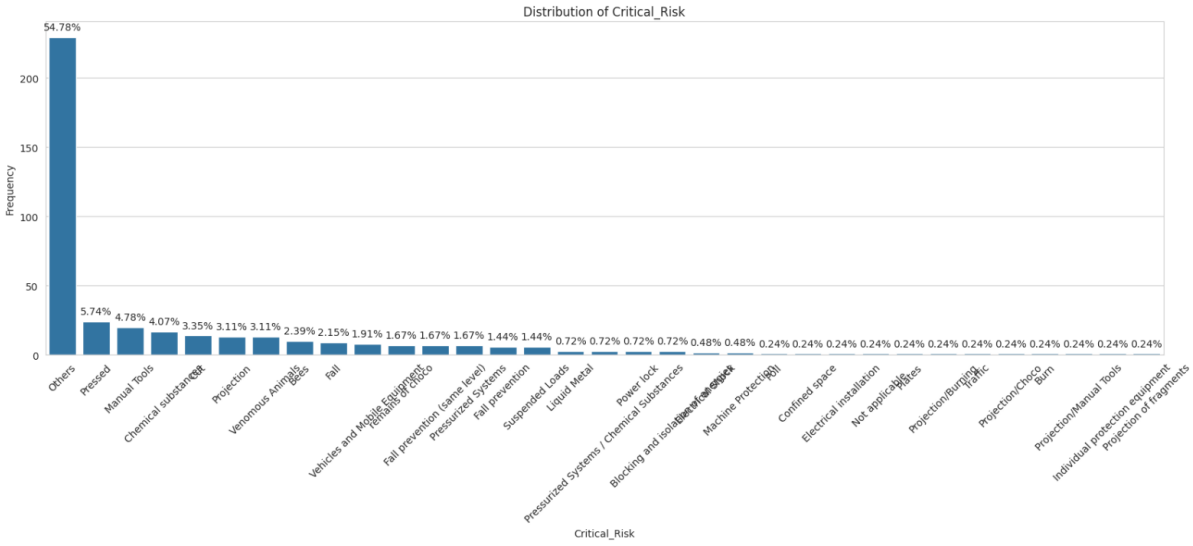
Employee Type

Total number of internal employees and Third-Party employees is more or less the same. But, we can also see that Third party remote employees are comparatively less in number.



Critical Risks

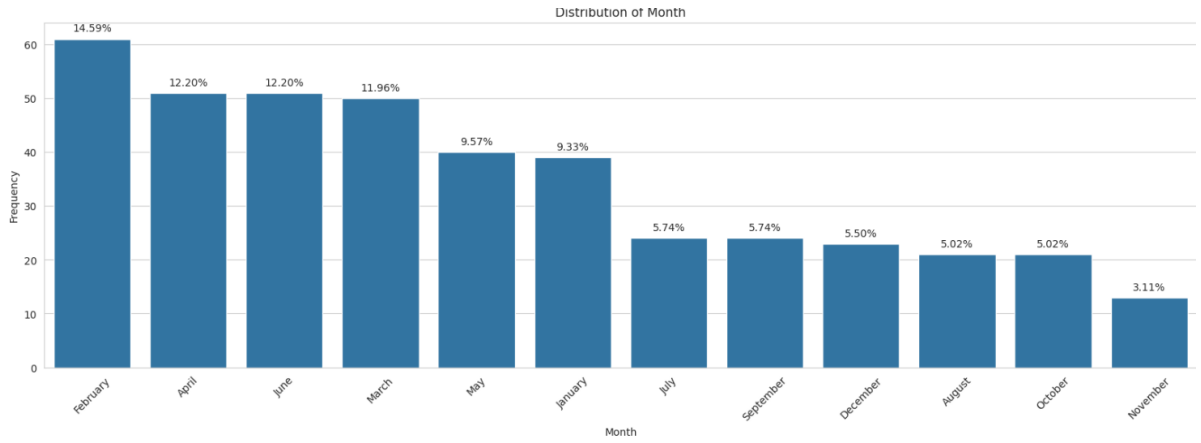
Most of the Critical Risks are classified as 'Others'. It holds around 55% of the total Critical Risks. It is followed by Pressed, Manual tools, Chemical substances, etc.



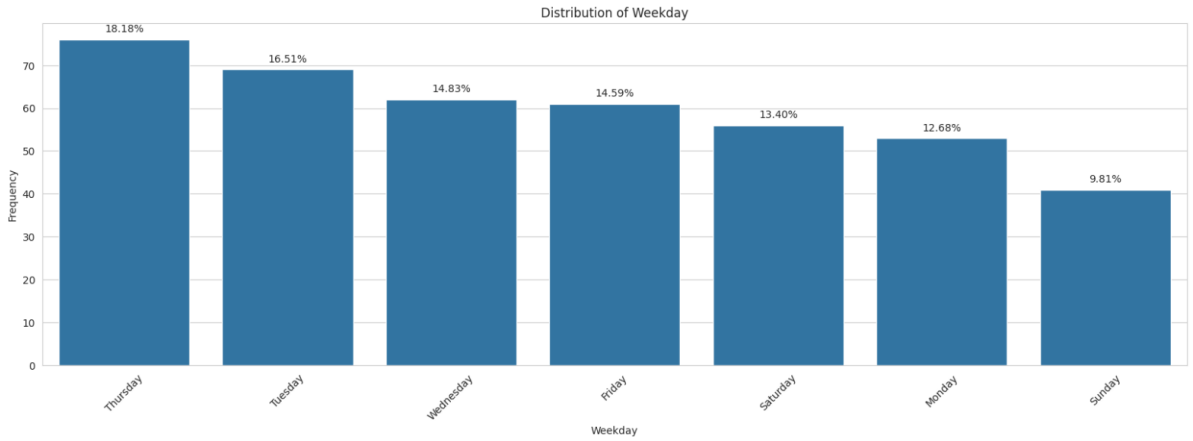
Date

The Date column was split into Year, Month, Date, Day of the week to check if there is any pattern.

- Data was available of 2 years 2016 and 2017.
- Year and Day column did not seem to any relevance for the analysis and were dropped.
- Accidents were more frequent in the first half of the year with maximum accidents in February at 14.59% followed April, June and March. November recorded the least number of potential accidents.



- Thursdays are more prone to accidents followed by Tuesday. Sunday has the least

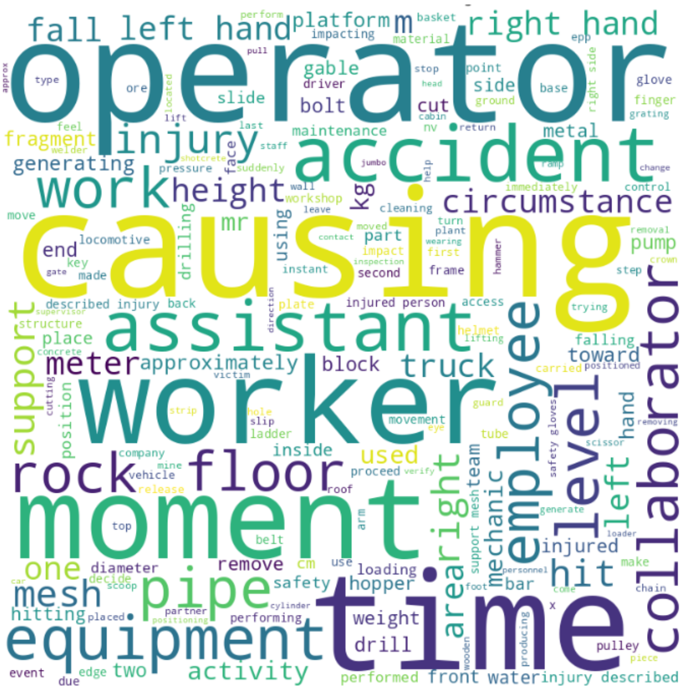


Description

Word Cloud were plotted to see which words were most frequent for the various industries and accident levels.

For Industry

Mining



Metals

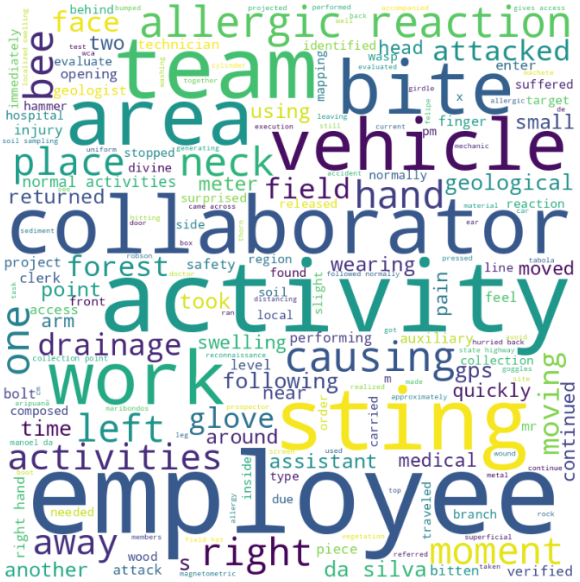
Others

Observations

Industry	10 most Frequent Keywords	Comments
Mining	causing, operator, time, worker, moment, accident, assistant, equipment, level, work	most of the accidents seem to be related to equipment and the impacted people were the equipment's operators.
Metals	employee, causing, operator, hit, activity, left hand, right, finger, left, area	most of the accidents seems to have been caused by some injury to the left hand or right finger where the operator was hit by something.
Others	employee, activity, team, causing, sting, bite, area, hand, left, right Looks like	most of the accidents seem to be due to bee bites on left or right area of the body.

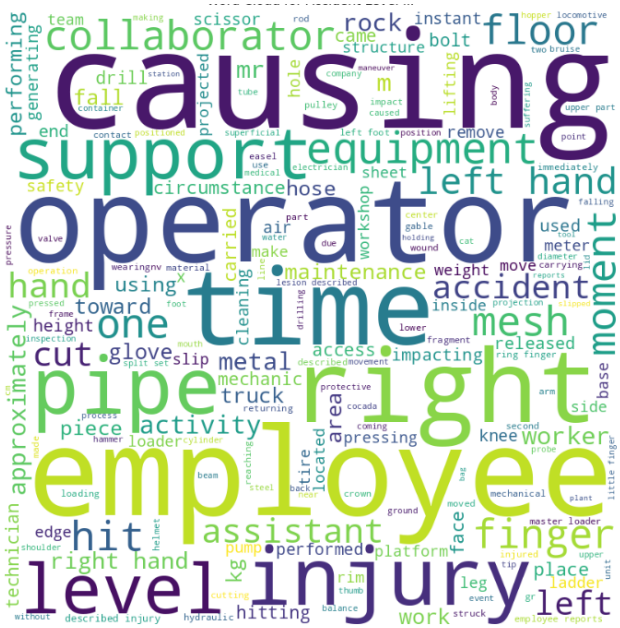
For Potential Accident Levels

Level 1



Level 2

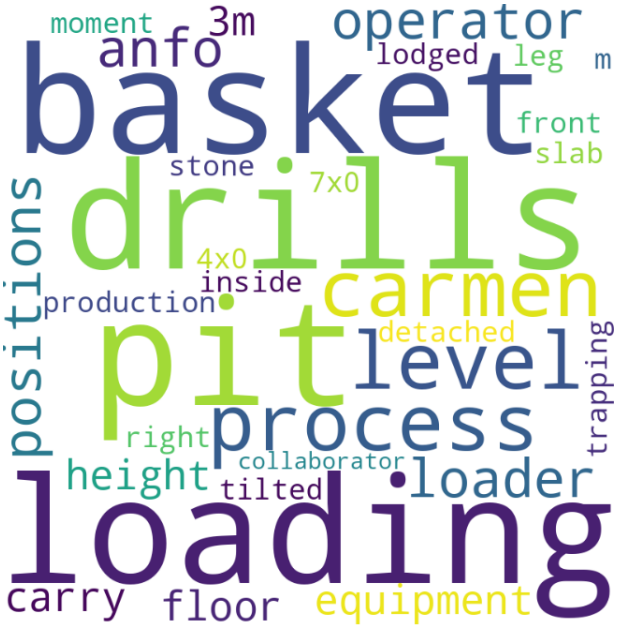
Level 3



Level 4

Level 5

Level 6



Observations

Potential Accident Level	10 most Frequent Keywords	Comments
Level 6	loading, drills, pit, basket, process, carmen, level, operator, positions, anfo	Major severity accidents seem to be forecasted at the drilling sites that could involve miners going down the cave in a basket or carmen using the vehicles on the site.

Level 5	operator, left, right, worker, side, equipment, m, work, part, causing	Severity 5 accidents are more prone with related to equipments were a part of the human body is injured.
Level 4	causing, operator, time, moment, employee, assistant, worker, accident, work, fall	Level 4 are more related to accidents that could happened due to falls.
Level 3	causing, employee, operator, time, right, injury, pipe, support, level, equipment	Level 3 could be injuries caused due to incorrect levels and burns due to pipe emissions
Level 2	causing, employee, right, worker, activity, mesh, left, hit, cut, time	Levels 2 could be minor cuts during working
Level 1	employee, activity, team, sting, collaborator, area, work, bite, vehicle, allergic reaction	Level 1 seem to be more related to bee stings and allergic reactions to some fumes/chemicals used in the industry.

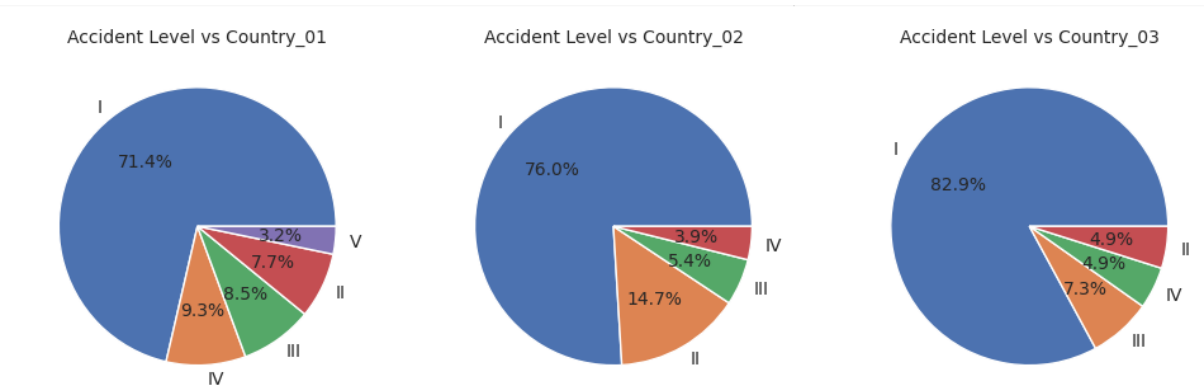
Bivariate Analysis

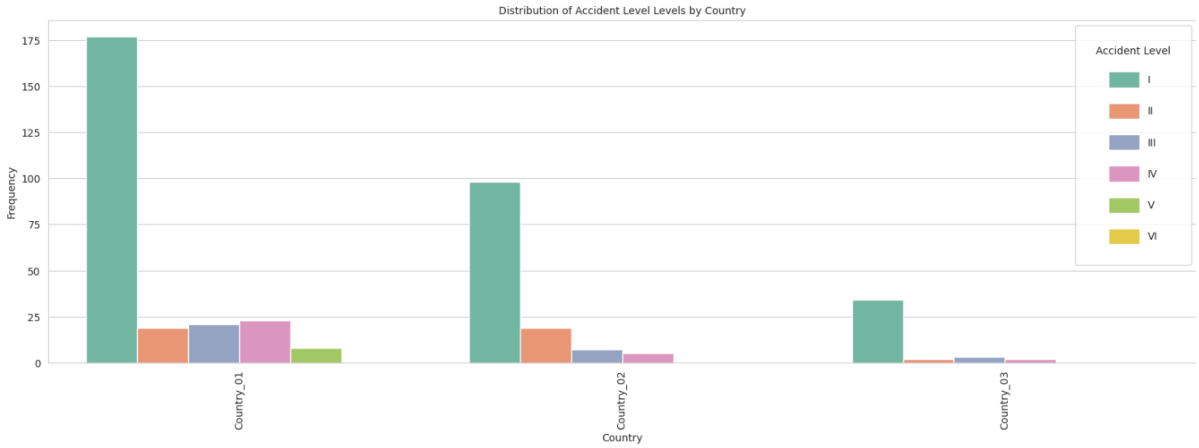
Bivariate analysis was then performed to see the patterns and spread of each feature with the accident levels and Potential accident levels. Below are the results and the observations

With Accident Levels

Accident Level/Country

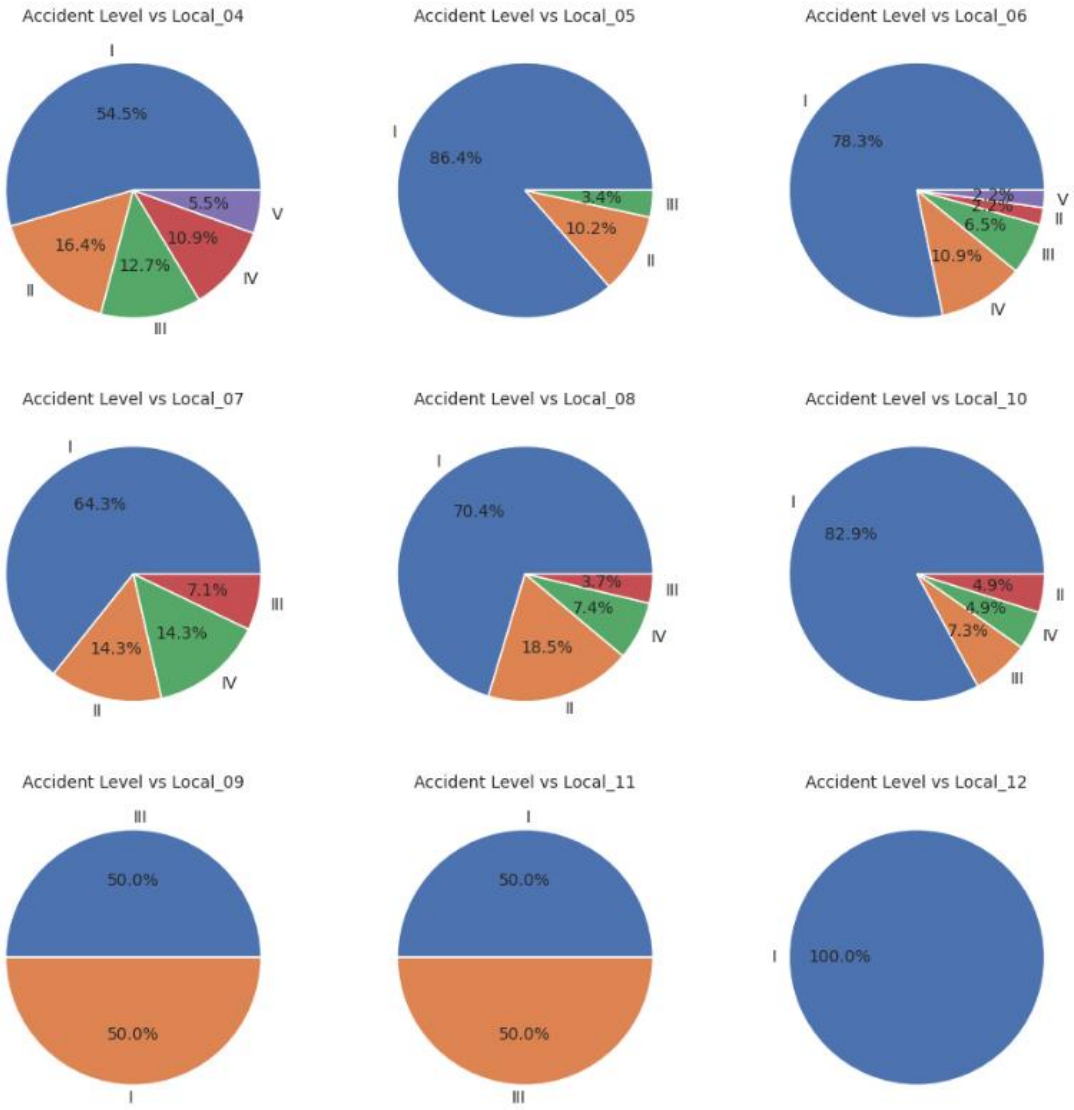
- More than 70% accidents in all three countries are Level 1 accidents.
- Severity of accidents are highest in count_01. All the Level 5 accidents occured only in country_01. It would be interesting to investigate further to know the reason for this. Why the most severe accidents are specific to country_01?

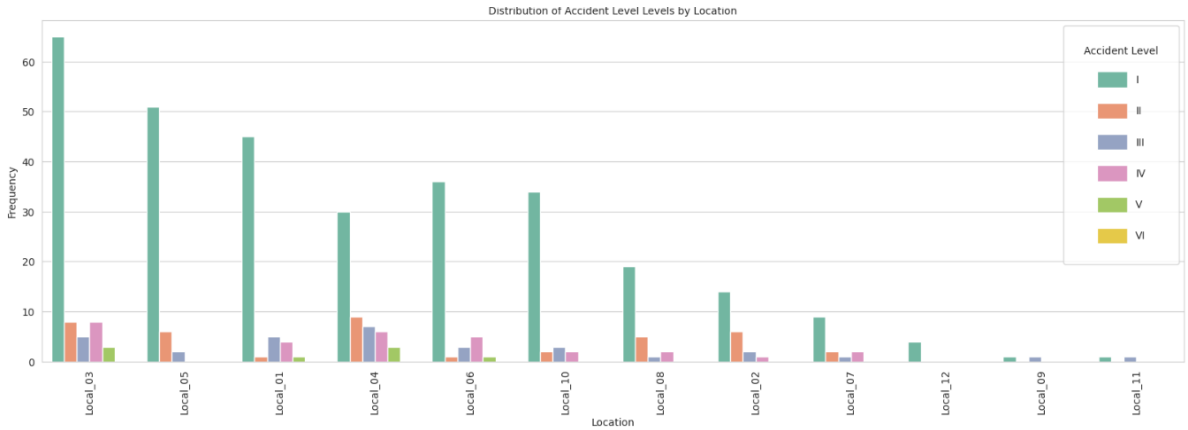




Accident Level/Location

1. Majority of accidents in all localities are level 1 accidents. (In fact, local_12 observed only level 1 accidents.)
2. Local_09 and local_11 observed only level 1 and 3 accidents, that too in equal proportion.
3. We can observe that the proportion of Level 4 and level 5 accidents is the highest in local_04.

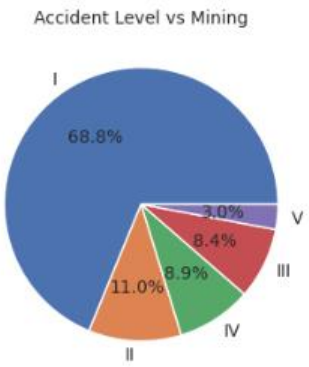


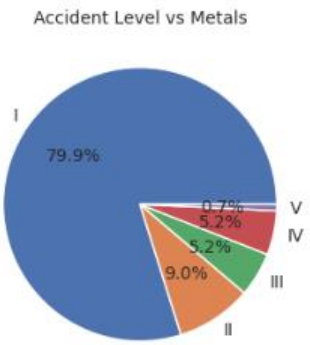


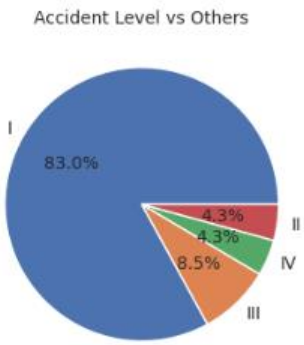
Accident Level/Industry Sector

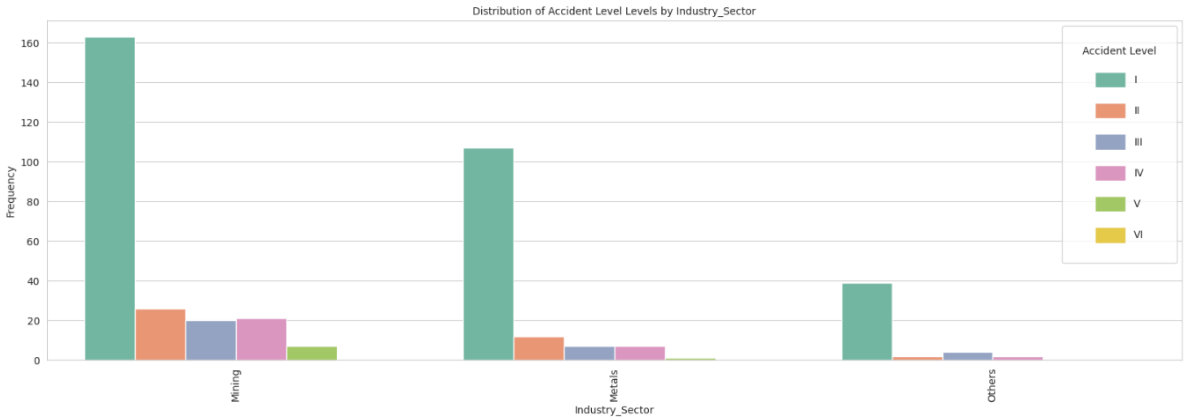
1. Mining sector has highest proportion of level 4 and level 5 accidents among all the sectors.

2. Proportion of level 1 accidents is highest in Others Sector.



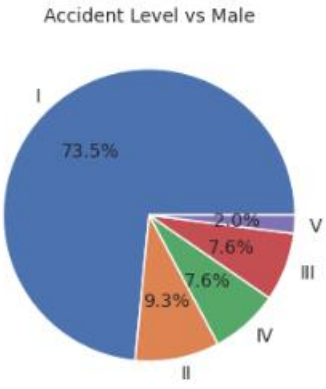


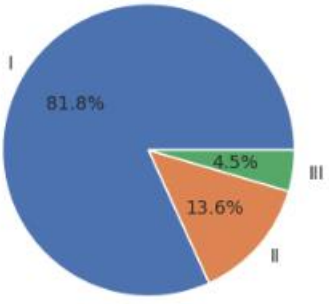


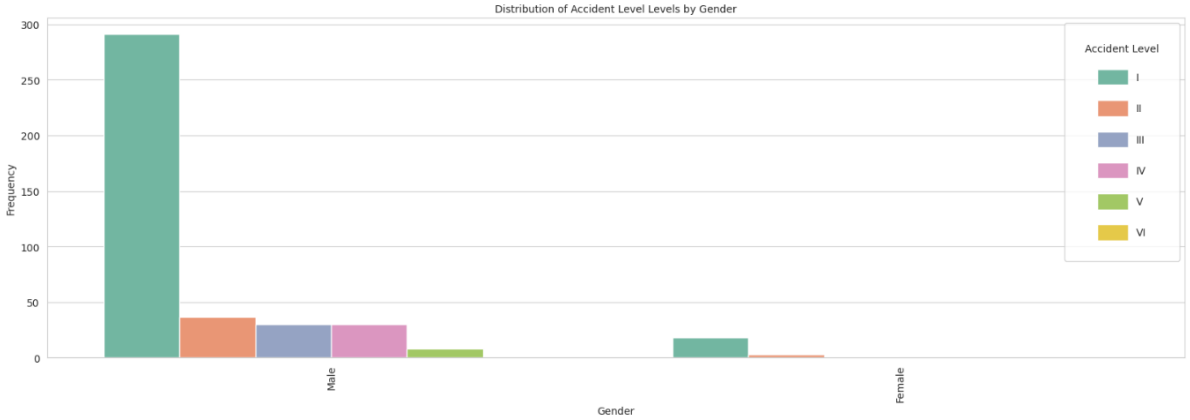


Accident Level/Gender

1. Men have faced more severe accidents than women.
2. Around 10% of overall accidents faced by men are level 4 and level 5 accidents, whereas women did not face any level 4 or 5 accident.
3. More than 80% of accidents faced by women are level 1 accidents.

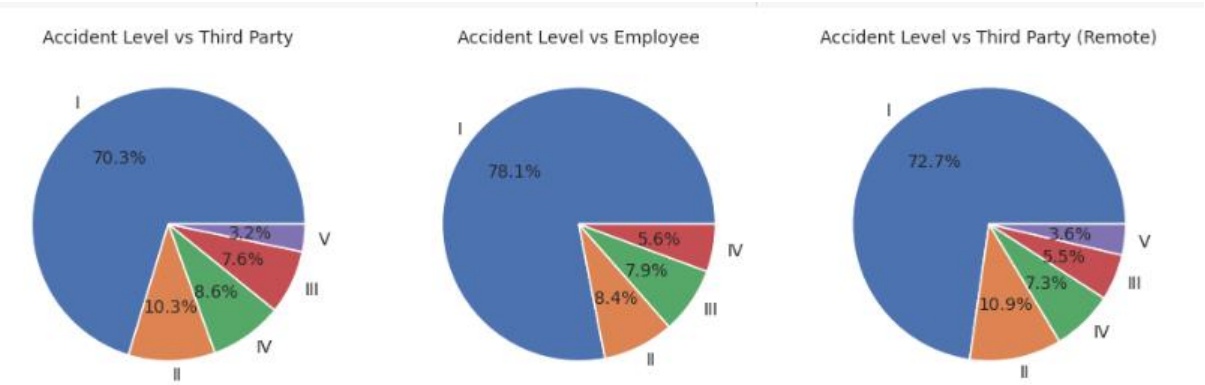


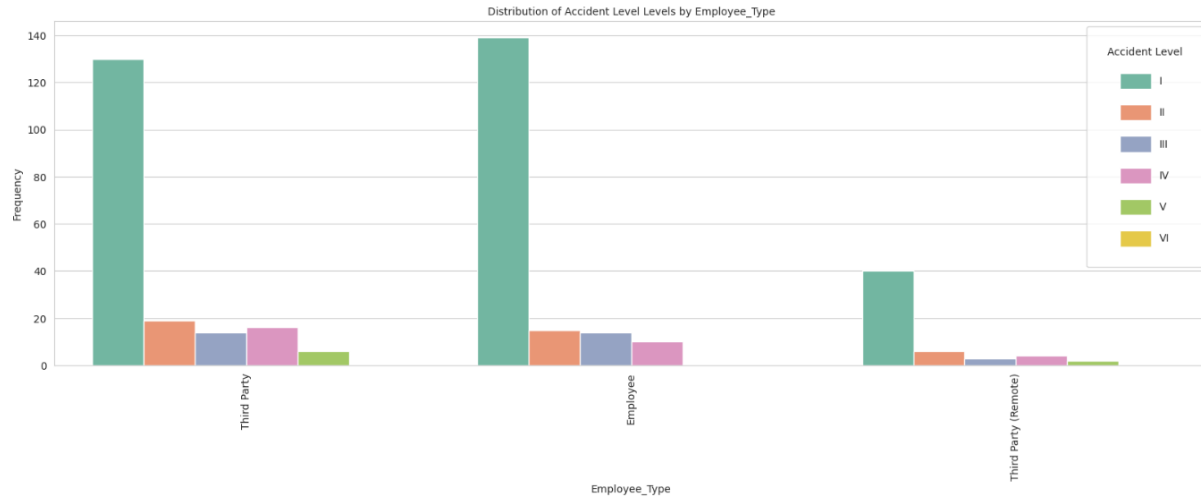




Accident Level/Employee Type

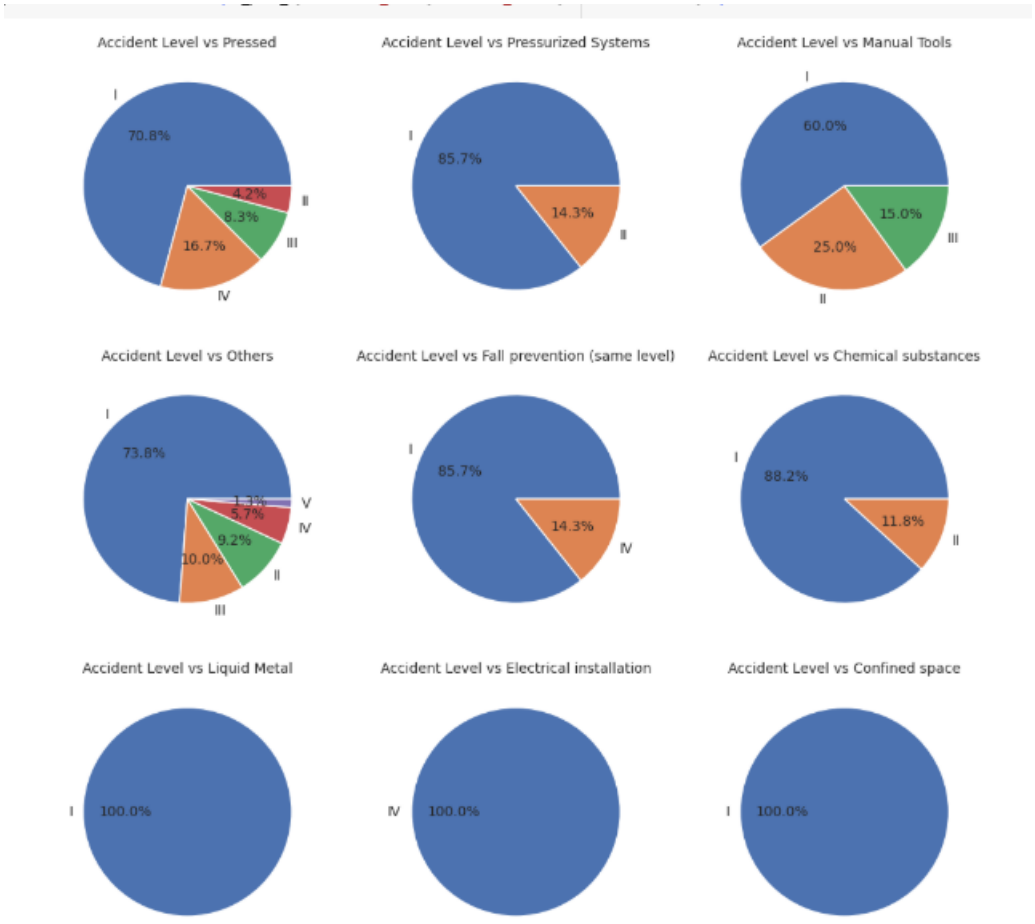
1. Level 1 accidents for all employee types is over 70%.
2. Internal employees did not face any level 5 accident, whereas proportion of level 5 accidents for both Third Party and Third Party remote is roughly the same.
3. This tells us that it is slightly more risky to be a Third party or remote employee than to be an internal employee.

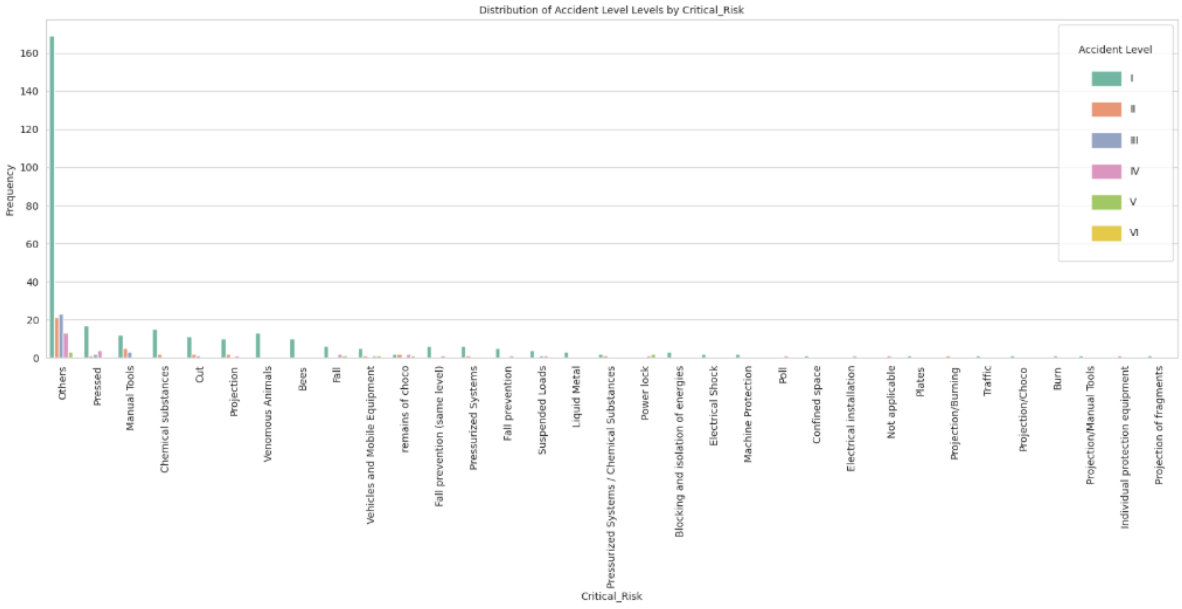




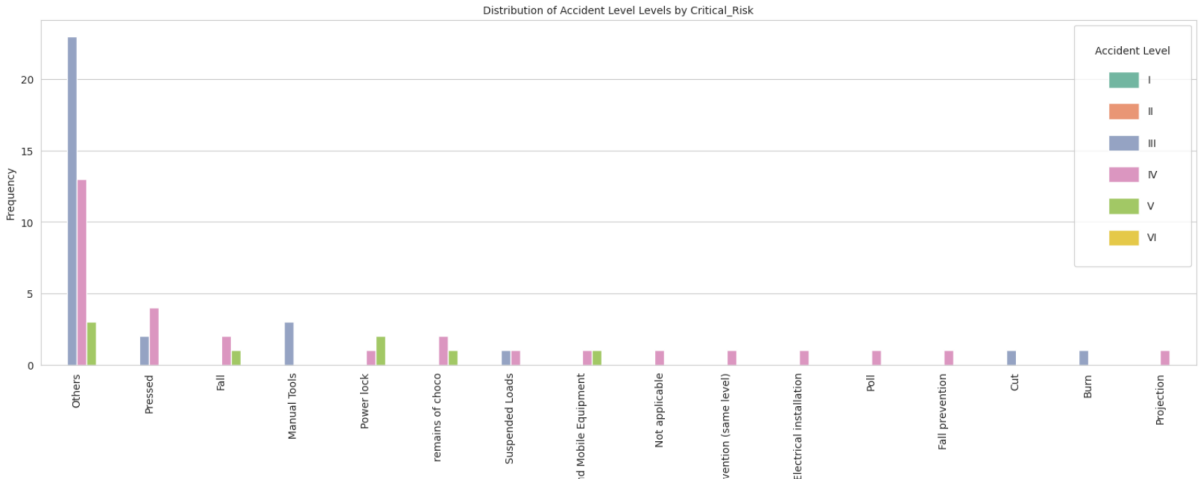
Accident Level/Critical Category

As there are several categories, displaying the pie-chart for a few of them below.



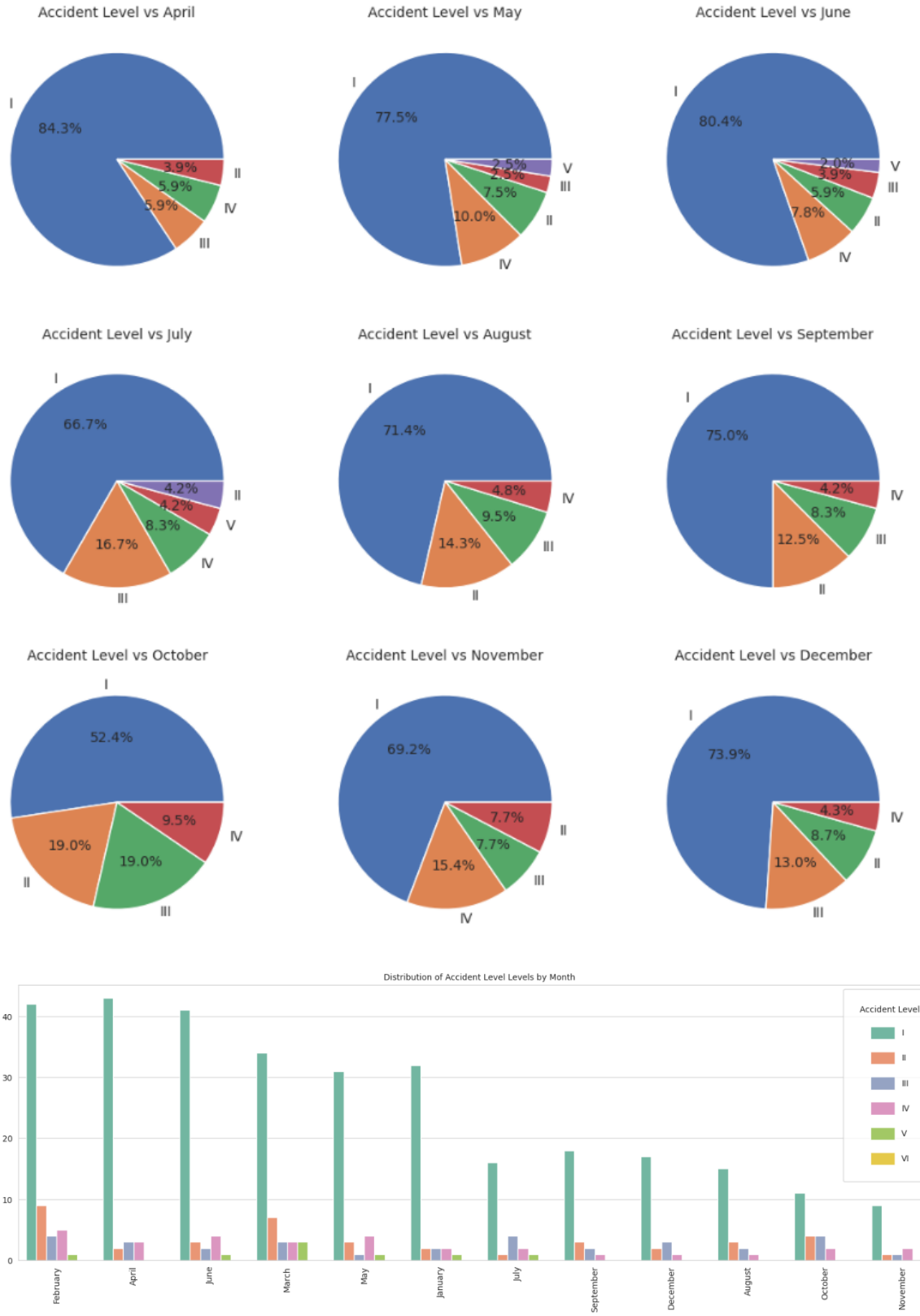


The most sever accidents for Levels 3,4 and 5 are for 16 critical Risk of Others, Pressed, falls, manual tools, power rack, remains of choco and others



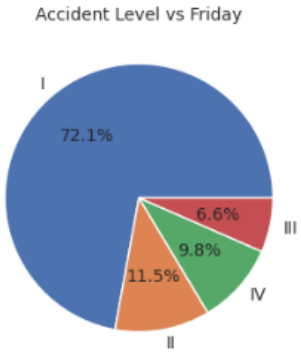
Accident Level/Month

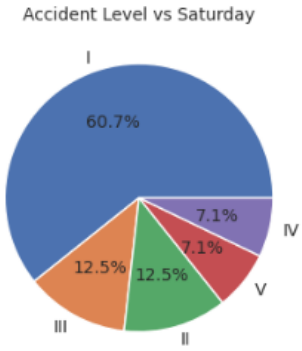
There have been severe accidents throughout the year. Severity 5 accidents all happened in May, June and July with levels 3 and 4 spread across the years. Every month there have been at-least 25-30% accidents with October being the most impacted month having almost 20% of accidents of level 3 and 4.

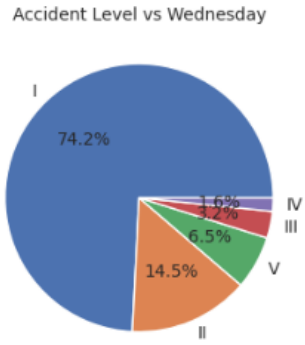


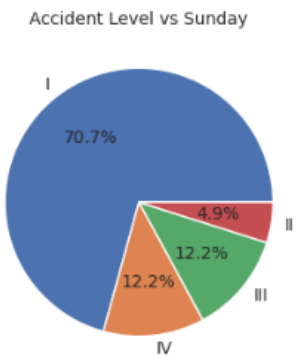
Accident Level/Weekday

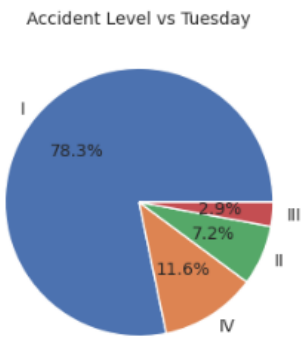
The ratio of severe accidents to the accidents that occur on a particular weekday is highest for Saturdays and Sundays. It is possible security measures are not being adhered to strictly on these days.

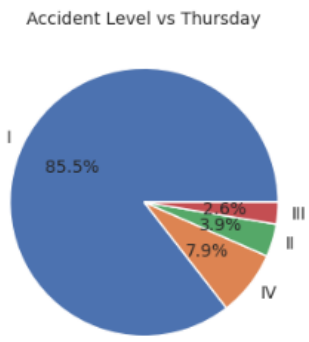

Accident Level vs Friday
I
72.1%
6.6%
III
9.8%
11.5%
IV
II

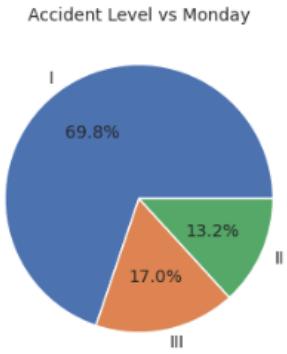

Accident Level vs Saturday
I
60.7%
7.1%
IV
7.1%
12.5%
12.5%
V
III
II

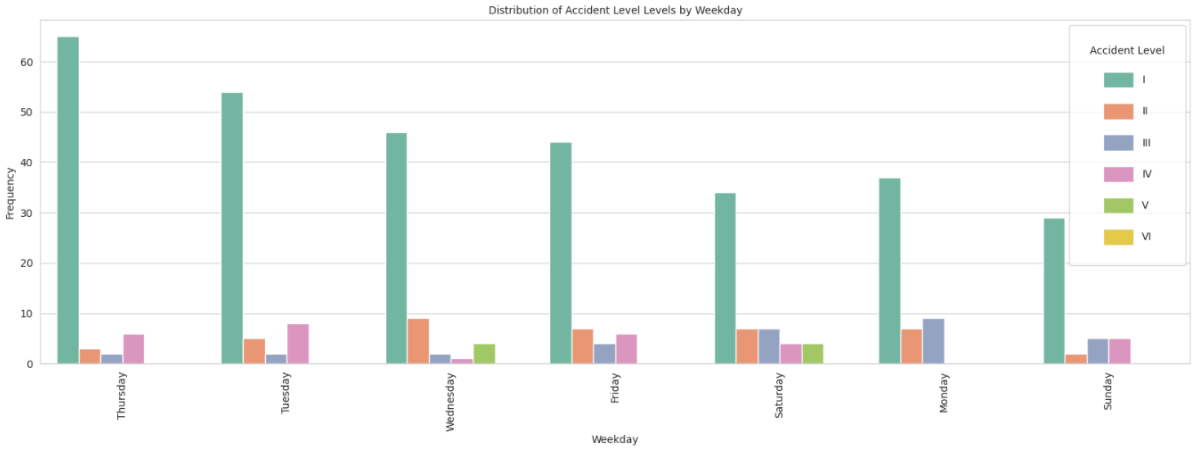

Accident Level vs Wednesday
I
74.2%
1.6%
IV
3.2%
III
6.5%
V
14.5%
II


Accident Level vs Sunday
I
70.7%
4.9%
II
12.2%
III
12.2%
IV


Accident Level vs Tuesday
I
78.3%
2.9%
III
7.2%
II
11.6%
IV


Accident Level vs Thursday
I
85.5%
2.6%
III
3.9%
II
7.9%
IV


Accident Level vs Monday
I
69.8%
13.2%
II
17.0%
III


Distribution of Accident Level Levels by Weekday
Frequency
0
10
20
30
40
50
60
Thursday
Tuesday
Wednesday
Friday
Saturday
Monday
Sunday
Weekday
Accident Level
I
II
III
IV
V
VI

With Potential Accident Levels

Accident Level/Country

Decide on the target column – analyse co-relation between Accident Level/Potential accident level

Correlation between accident level and potential accident levels was analysed to decide which column should be our target column.

Several techniques were used:-

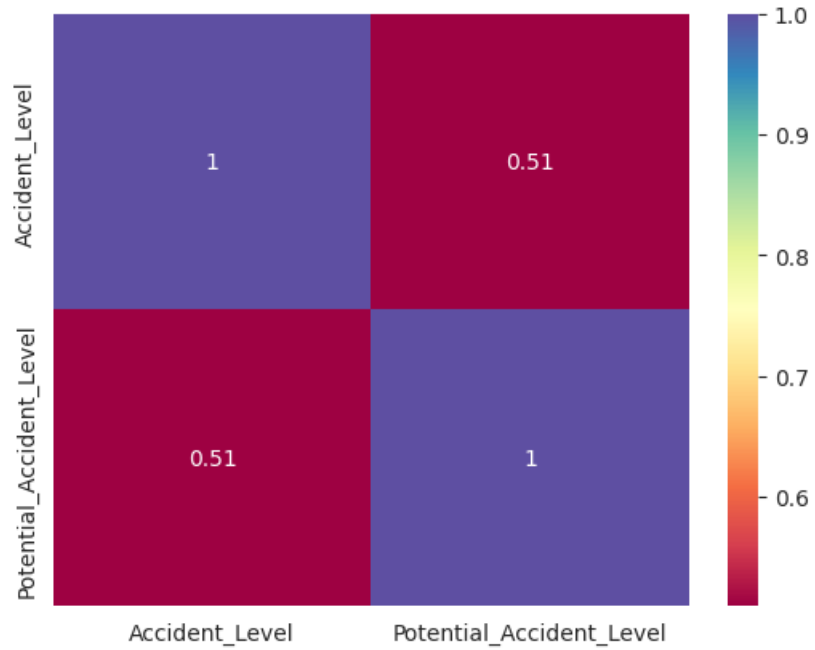
Pearson and Sperman coefficient

Between Accident Level and Potential Accident Level:

The Pearson correlation coefficient is 0.51

The Spearman correlation coefficient is 0.50

Heat Map



Cross tab

Though the potential levels are higher the actual accident levels are less severe.

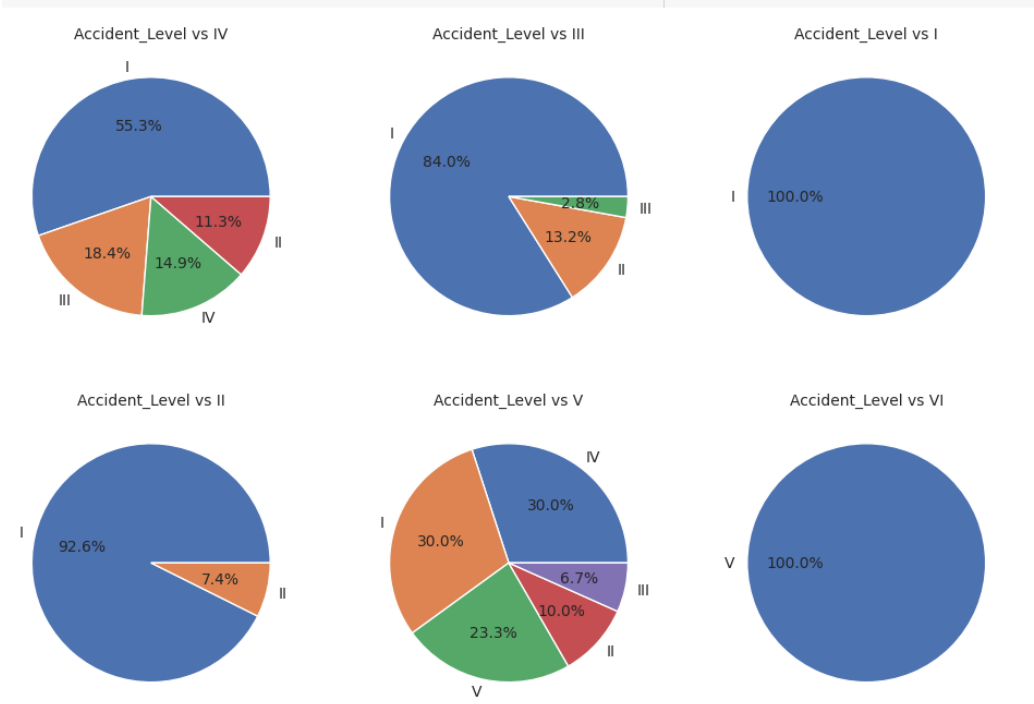
Potential_Accident_Level	1	2	3	4	5	6
Accident_Level						
1	45	88	89	78	9	0
2	0	7	14	16	3	0
3	0	0	3	26	2	0
4	0	0	0	21	9	0
5	0	0	0	0	7	1

Below graph show the percentage times when accident level was similar to what was the potential accident level for severity levels 3,4 and 5

Level V: Only 23.3% of the times accident level was of severity IV when the potential was also IV

Level IV: Only 14.9% of the times accident level was of severity IV when the potential was also IV

Level III: Only 2.8% of the times accident level was of severity IV when the potential was also IV



Observation and Inference:

- Both correlation coefficients (0.51 for Pearson and 0.50 for Spearman) suggest a ***moderate positive relationship*** between Accident_Level and Potential_Accident_Level. This indicates that higher levels of accidents are likely associated with higher potential levels of accidents.
- The similarity in values (0.51 for Pearson and 0.50 for Spearman) suggests that the relationship is reasonably ***consistent across both linear and monotonic assessments***. This consistency strengthens the reliability of the observed association.

Given the requirement for this project, we will take a pessimist approach and warn the user with higher potential accident that may be encountered. We have therefore consider Potential Accident Level as the target variable.

Model Building

Data Preprocessing

The below data pre-processing and clean up steps were performed:-

- As the number of severity VI potential accidents data is very small, the rows were merged as severity VI.
- Also, as the Potential Accident Level has inherent order representing increasing severity. Maintaining the numerical relationship is important for analysis, hence Ordinal Encoding was used.
- Country and critical risk (critical risk is just a classification of the description column.)

For the Basic Machine Learning Model

For the training a basic machine learning model the below data preprocessing steps were also performed:-

- The categorical columns were encoded using several techniques based on the nature of the data in that feature
    - **Location**: Frequency encoding

      Locations are anonymized but numbered sequentially from Local_01 to Local_12

      Some locations have significantly higher accident rates than others and the distribution is highly skewed.

    - **Industry sector** : One Hot encoding as there are only 3 categories and have no inherent ordering..
        - **Employee Type**: Employee Type has only 3 categories and have no inherent ordering. Hence One hot encoding was used.
    - **Gender**: Binary Encoding as it has only 2 values.
    - **Month and Weekday**: For Month and Weekday we would proceed with **Cyclical Encoding** as they represent time data. December is as close to January as November, and Sunday is as close to Monday as Saturday. Cyclical encoding would preserve this cyclical nature of data

Data Preprocessing on the Description column (for NLP based models)

Several text preprocessing was applied on the description column

- NER processing

    - The description contains names of person and places. Pretrained model from Hugging Face was used to extract and replace the named entities from the text.
- Lowercase conversion
- Timestamp conversion
    - We observe quite a few mentions of time in various formats. e.g. "9:45 am", "14:16", "04:50 p.m.", etc. Let us replace all these with period of the day - morning, afternoon, evening or night.
- Remove Numbers and Special characters
- Stop word removal
- Lemmatization
- Spell Checker

Data Up-sampling

For the Basic Machine Learning Model

SMOTE was used to up-sample the records for all the severity levels to handle the class imbalance.

For the NLP models

The below techniques were used and the comparison was performed using data from both these techniques

Synonym replacement technique

Criteria for Model Evaluation

Several machine learning models were developed with **a primary focus on optimizing Recall scores, particularly for higher accident severity levels (3, 4, 5, and 6).**

Since the goal is to predict whether an accident will occur and assess its severity, the emphasis was placed on achieving accurate predictions for severe accidents. This prioritization reflects the critical importance of identifying high-severity incidents, even if it means tolerating lower precision scores for lower-severity predictions.

The rationale for this approach is that failing to predict a severe accident or predicting a lower severity, has far graver consequences than overestimating the likelihood or severity of an incident. By focusing on higher Recall scores, the models aim to minimize the risk of undetected severe accidents while providing actionable insights for safety professionals.

Models Considered

The below models have been considered for Milestone1

Random Forests – As this model Works well for structured data representations like TF-IDF or word embeddings.

We used this model with various different types of inputs to get the best recall accuracy. All the models were Design and train a Random Forest Classifier model.

1. **Model 1 (Base Model (without Description))** - A base model was built using Random Forest Classifer with all the relevant features from the received dataset, except for the description
2. **Model 2 [Base Model without Description and best Features (from Model 1)]:** Second version was created using only the most important features returned from the base random forest model.
3. **Model 3 [Base Model with Hyper Tuning]:** Then the base model was hyper tuned using RandomSearchCV and GridSearchCV.

   We then used several techniques to create embedding and then built random forest. The below embedding techniques were used:-
    a. Word2Vec

   b. Glove
   c. Sentence Transformer (using 'sentence-transformers/all-MiniLM-L6-v2' model from hugging face)
4. **Model 4 [With Word2Vec embeddings]:** a base Random Forest classifier model was built using the word2Vec embeddings
5. **Model 5 [Word2Vec and class balancing]:** Random Forest classifier model with word2Vec embeddings and class balancing
6. **Model 6 [Tuned Model with Word2Vec and class balancing]:** Hyper tuned word2Vec Model with class balancing.
7. **Model 7 [With GloVe embeddings]:** a base Random Forest classifier model was built using the **GloVe** embeddings
8. **Model 8 [GloVe and class balancing]:** Random Forest classifier model with **GloVe** embeddings and class balancing
9. **Model 9 [Tuned Model with Word2Vec and class balancing]:** Hyper tuned **GloVe** Model with class balancing
10. **Model 10 [With transformer embeddings]:** a base Random Forest classifier model was built using the **transformer** embeddings
11. **Model 11 [transformer and class balancing]:** Random Forest classifier model with **transformer** embeddings and class balancing
12. **Model 12 [Tuned Model with transformer and class balancing]:** Hyper tuned **transformer** Model with class balancing

Model Performances

Comparing recall scores and accuracies:

We compared the models using their recall_scores for each accident level and weighted F1-Scores as shown in the following table [Recalls for severity level 5 (level 6 was added to level 5) etc.]

Model	Recall 5	Recall 4	Recall 3	Recall 2	Recall 1	Weighted F1-Score
Base Model (without Description)	0	0.2	0	0	0.8	0.58
Base Model without Description and best Features (from Model 1)	0	0.17	0.17	0.12	0.82	0.63
Base Model with Hyper Tuning	0	0.2	0	0.17	0.78	0.6
Base Model with Word2Vec	0	0.45	0.19	0.16	0.44	0.28
Base Model with Word2Vec and class balancing	0	0.45	0.29	0.16	0	0.26
Tuned Model with Word2Vec and class balancing	0	0.31	0.19	0.16	0.56	0.25
Base Model with GloVe	0.17	0.79	0.14	0.11	0.44	0.39
Base Model with GloVe and class balancing	0.17	0.86	0.19	0.16	0.44	0.44
Tuned Model with GloVe and class balancing	0	0.66	0.24	0.21	0.67	0.4
Base Model with Transformer	0.33	0.76	0.14	0.11	0.67	0.41
Base Model with Transformer and class balancing	0.33	0.9	0.24	0.05	0.67	0.47
Tuned Model with Transformer and class balancing	0	0.69	0.33	0.26	0.67	0.45

Best Performing Traditional ML Classification Model

Based on the criteria of recall scores, and the criteria of minimizing the underprediction of severity, we can select the best performing traditional ML classifier among all the classifiers built as the Random Forest class balanced model that has been training on Transformer embeddings as it is able to correctly predict more of the sever potential accidents than any of the other models.

Conclusion & Further Consideration for Improvement

More training data would have helped come up with more robust traditional Machine Learning classifiers.

Among all the embeddings transformer based embeddings are giving the best performance,.

However, performance of none of these model sufficient to convince us that these traditional Machine Learning models built on this small data set can be deployed for predictions in a real-time production environment.

Use advance models

In Milestone-2, we will consider more advanced models which are anticipated to perform better than the traditional ML models.

Some of these models can be:

- Artificial Neural Networks (ANN)
- Recurrent Neural Network (RNN)
- Long-Short Term Memory (LSTM)
- Large Language Model (LLM)

Use LLM for data augmentation

Also, we will use LLMS (LLAMA 3.2) model from hugging face to add more synthetic data. The below approach can be followed:-

1. Join a few descriptions randomly for a particular combination of features (potential accident level, industry, weekday, month etc.) to give as few-shot prompts to the model.
2. Use the most frequent words form the Word Cloud that was plotted again the potential accident levels and also industry-wise in the prompt

The process was tried using LLMA3.2 running on local machine using OLLAMA. We created the prompt by giving a few shot examples and also the most frequent keywords for that particular industry and severity level. As the inference in local is taking a lot of time due to lack of GPUs, we will using LLMA 3.2 or other models from hugging face in Google colab to generate more context aware synthetic data as part of MileStone 2.