INTERIM REPORT

CHOTBOT PROJECT

CAPSTONE PROJECT | NLP 2 GROUP 13

Table of Contents

Mentor: Mr. Abdul Manaf

Members:
Nagaraju Nayini
Shoaib Mohammed
Kirija A
Subhadip Chakraborty
Anshika

SUMMARY

There is an emerging need in the current industrial sector to find a solution for the increase in recent injuries/accidents in plants. There are cases of death which has also been registered.

The industry is in dire need of a solution for its employees to safeguard their interests in certain sectors.

Designing a ML/Dl based chatbot utility which can help the professionals to highlight the safety as per the incident description

PROBLEM STATEMENT

The database comes from one of the biggest industries in Brazil and in the world. It is an urgent need for industries/companies around the globe to understand why employees still suffer some injuries/accidents in plants. Sometimes they also die in such environment. #DATA DESCRIPTION: This The database is basically records of accidents from 12 different plants in 03 different countries which every line in the data is an occurrence of an accident. #Columns description:

1. Data: timestamp or time/date information
2. Countries: which country the accident occurred
3. Local: the city where the manufacturing plant is located
4. Industry sector: which sector the plant belongs to
5. Accident level: from I to V, it registers how severe was the accident (I means not severe but VI means very severe)
6. 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)
7. Genre: if the person is male of female
8. Employee or Third Party: if the injured person is an employee or a third party
9. Critical Risk: some description of the risk involved in the accident
10. Description: Detailed description of how the accident happened. Link to download the dataset: https://www.kaggle.com/ihmstefanini/industrial-safety-and-health-analytics-database

SUMMARY

INSIGHTS

On inspection of the dataset, it appears that:.

- The dataset is limited and consists of four hundred and twenty-five records only so training the models with high accuracy could be a challenge
- The dataset is imbalanced on certain variables like potential accident level and accident level, this means that we may not get consistent results unless the dataset is treated to reduce imbalance.
- Minor accidents are more common than major accidents, this looks like real world situations.
- There is data from three countries.
- There are twelve locals or cities from which the data is taken.
- There are two industry sectors - mining, metals and third all others grouped together as others.
- There are five accident levels.
- There are six potential accident levels.
- There are employees, third parties and remote third parties involved in the accidents.
- There are thirty-three different types of critical risk one of which has been assigned to an accident incident.
- The accident description is highly unclean and so it will require a considerable amount of effort to clean it to produce results.
- The dataset consists of data from January 2016 to July 2017.
- Males are more involved than females in accidents, this too looks similar to real world situations as there are considerably lower number of females working in industrial environments.

OBJECTIVE

GOAL

- To create an industrial conversational bot
- The chatbot will be used as an guide for the customer, employee and management to access the potential risk which might be involved for a specific sphere of work.

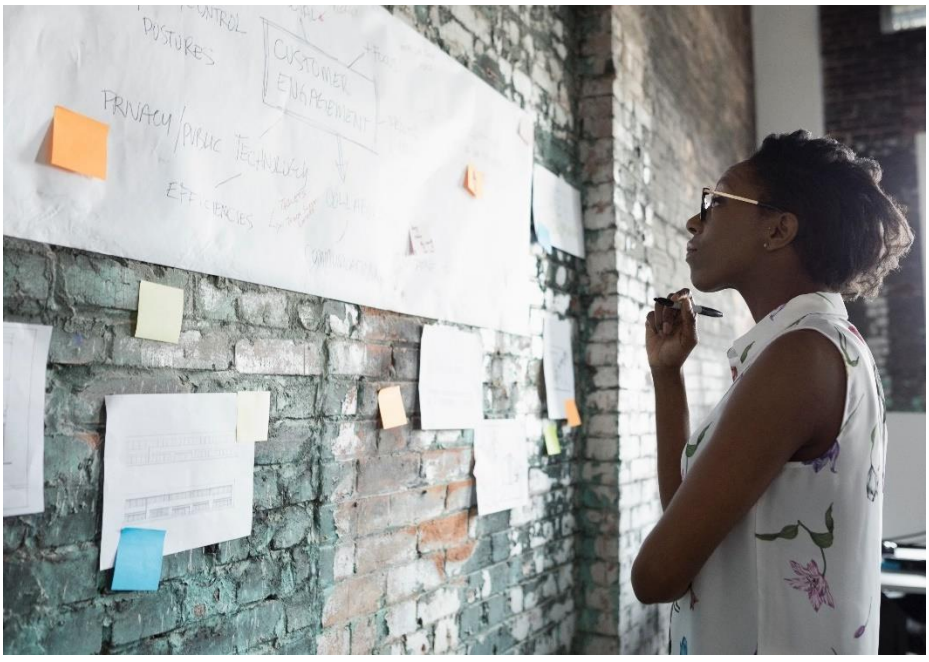
HIGHLEVEL FINDINGS

- Many body-related actions and accidents has been found.
- A lot of equipment related accident has been mentioned in the dataset.
- Poor features map found with lack of access to quality data found.

HIGHLEVEL IMPLICATIONS

- The main causes of the accidents are hand related operation and time base factor.
- More strict safety standard needs to be maintained to reduce accidents.
- Equipment based safety standards needs to be defined.

The broader version is evaluated further in the EDA.



.

EDA

APPROACH

The approach was initially to remove the stop words, use lemmatization.
We started using N Gram, Univariate m Bivariate and time series analysis to decide on the type, trend and pattern of the accident causes.

- Use Pre-processing technique
  - Time-related feature extraction
  - NLP pre-processing (special characters removal, removing stop words)
- Practice EDA technique
- Practice visualizing technique.
- Practice feature engineering technique
- Practice modeling technique
- Causal analysis skill

PRE-PROCESSING

- We noticed that except a 'date' column all other columns are categorical columns.
- We observed that there are records of accidents from 1st Jan 2016 to 9th July 2017 in every month. So there are no outliers in the 'Date' column.
- There are 12 Local cities where manufacturing plant is located and it's types are in sequence so there are no outliers in 'Local' column.
- There are only three Industry Sector types which are in sequence so there are no outliers in 'Industry Sector' column.
- There are only five Accident Level types which are in sequence so there are no outliers in 'Accident Level' column.
- There are only six Potential Accident Level types which are in sequence so there are no outliers in 'Potential Accident Level' column.
- There are only two Gender types in the provided data so there are no outliers in 'Gender' column.
- There are only three Employee types in the provided data so there are no outliers in 'Employee type' column.
- There are quite a lot of Critical risk descriptions, and we don't see any outliers but with the help of SME we can decide whether this column has outliers or not.

INSIGHTS

- Though the staffs of the manufacturing plants are mostly males, EDA shows that males are likely involved in accidents.
- And males tend to get involved in accidents with higher risk levels than females.
- Comparing employee's accidents count with third parties' accident count, EDA shows that third parties are likely involved in accidents (58%)

Deciding Models and Model Building

Objective:

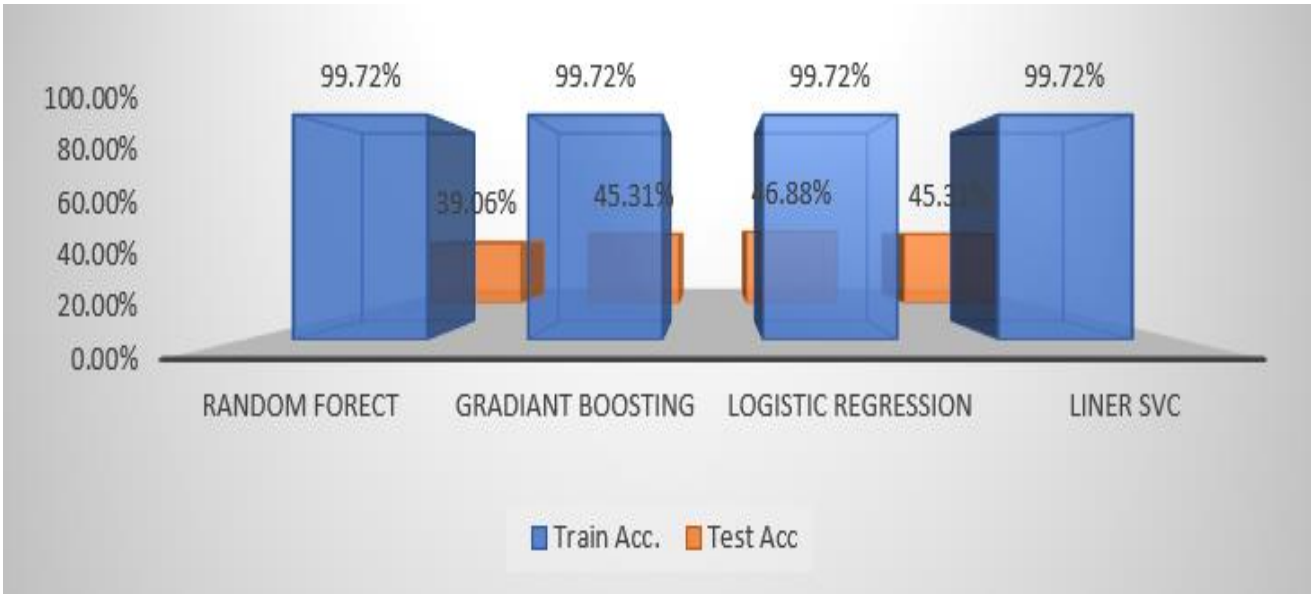
- Presumption of cause of accidents
- Surveying a factor that increases severity of accidents

Building the model which classify the severity of accidents, we can understand the factor related to the causality of accidents.
So, Machine Learning classifier models were built based on those cases below.

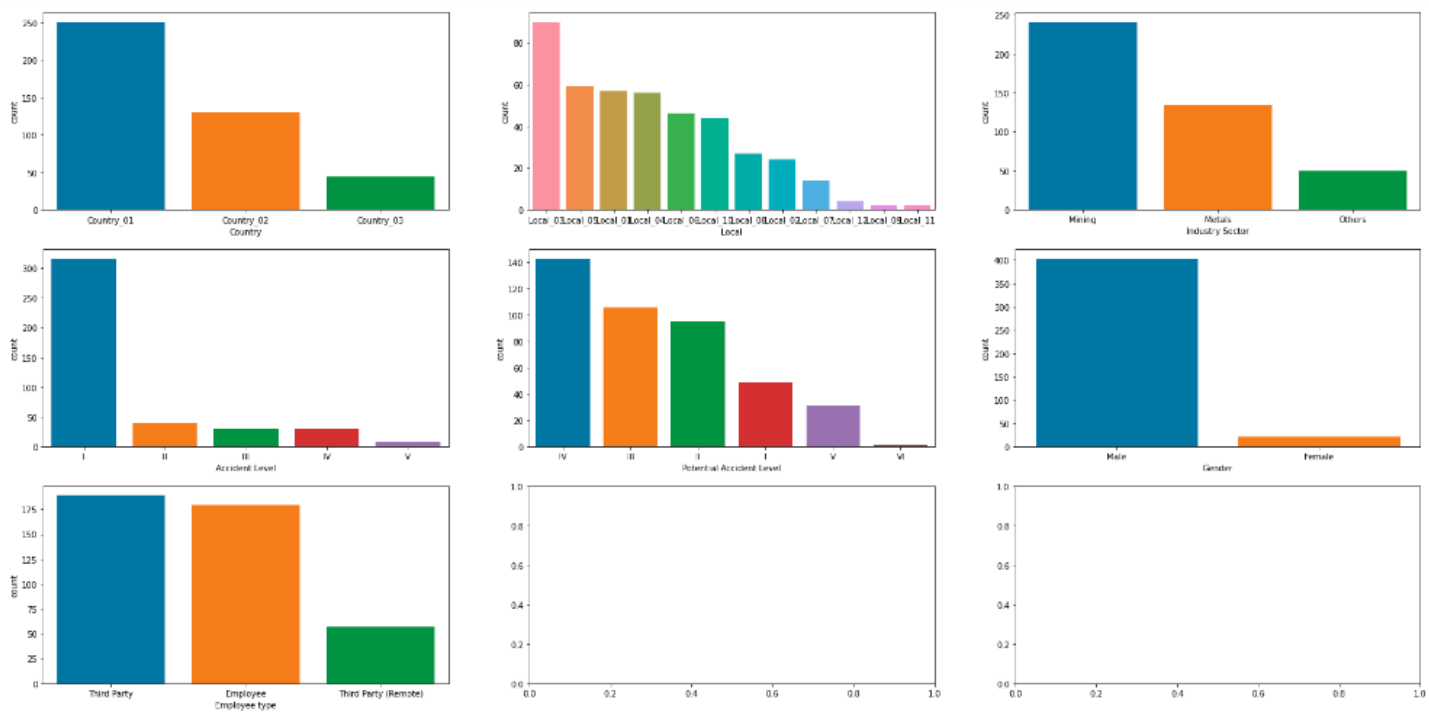
- Accident Level.

The model that has been used are below:

- Ridge Classifier
- Logistic Regression Classifier
- XGB Classifier
- K Neighbors Classifier
- SVC
- Decision Tree Classifier
- Random Forest Classifier
- Bagging Classifier
- ADA Boost Classifier
- Gradient Boost Classifier
- LGBM Classifier

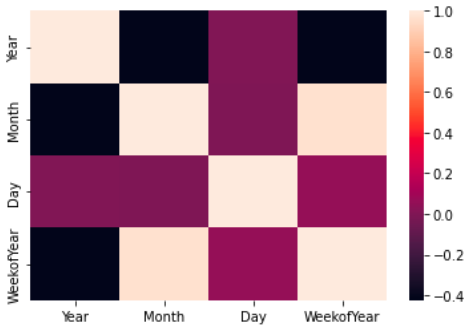


VISUAL ANALYSIS



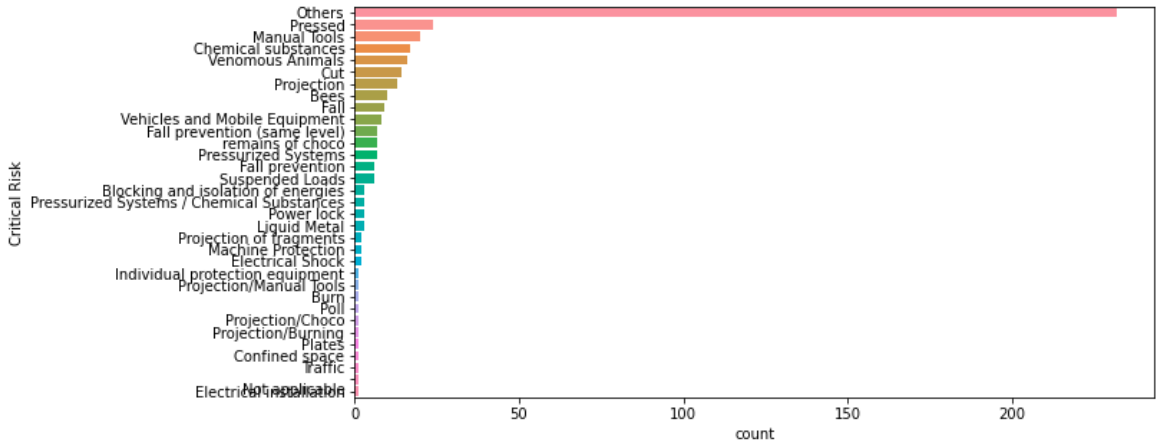
Observations:

1. Country 1 has highest accidents. country 3 has lowest
2. Highest manufacturing plants are in Local_03 city. Lowest manufacturing plants are in Local_11 city
3. Highest Industry Sector are Mining
4. Accident Level I has more counts and V has less counts
5. Potential Accident level IV has more counts and VI has less
6. Male counts are more than Female counts
7. Employee type Third Party are more affected.



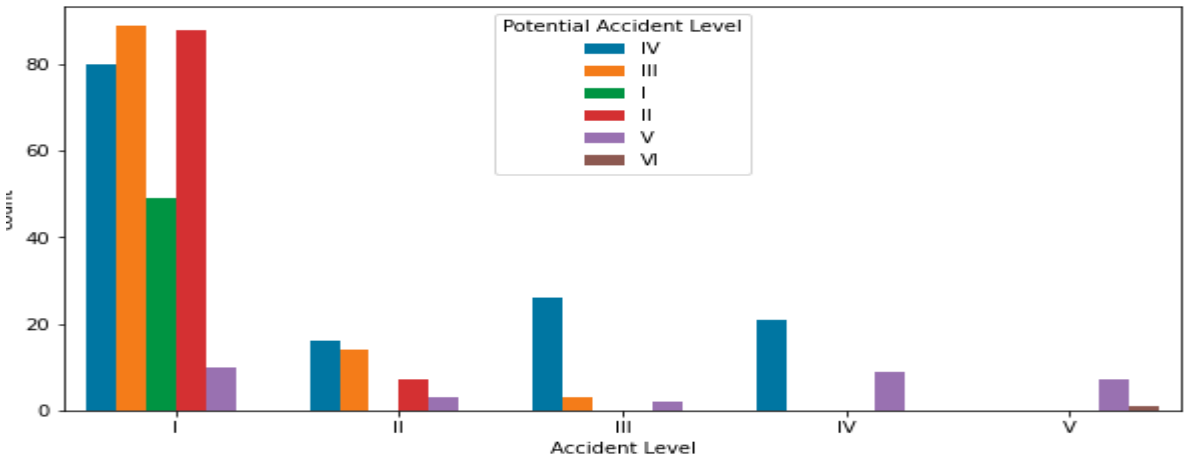
Observation

Weakofyear has high correlation with Month



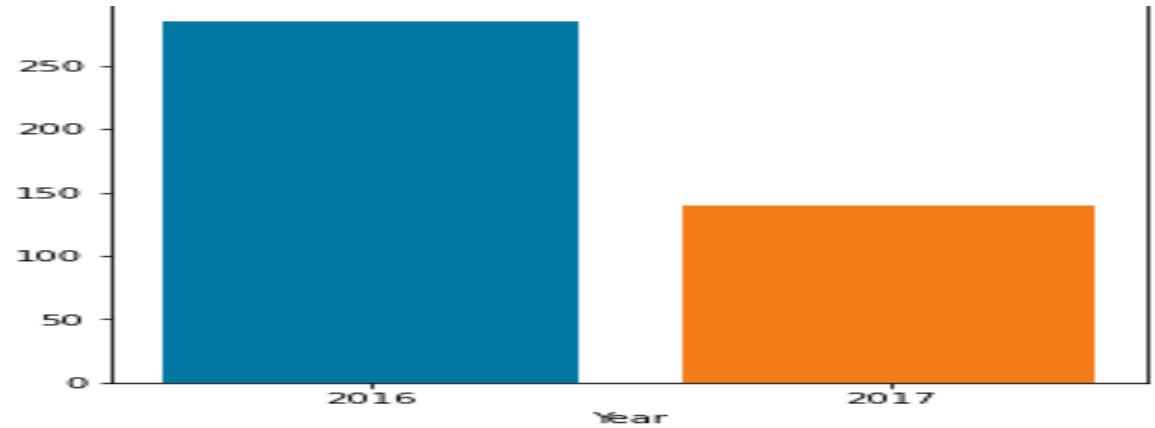
Observation

Most part of the critical risks are classified as 'Others'



Observation:

Potential of accidents level increases as the level of accident is I



Accidents in 2016 year are more than 2017

WORD CLOUD REPRESENTATION

<u>**Observations**</u>

There are many body-related, employee related, movement-related, equipment-related and accident-related words.

1. Words which are related to human body like Hand, Head, Foot, Leg, Finger
2. Measurement words like Cm, Height, Kg
3. Type of employee related words like Employee Collaborator, assistant, Operator
4. Movement related words like, Left, Right, Fall, Hit, slip
5. Equipment's related words like equipment, pump, meter, drill, truck
6. Accident-related words accident, activity, safety, injury, causing

- ✓ Mining sector is the most effected sector and most of the classes of Critical Risk comes from this sector.
- ✓ Most of the classes of Potential Accident Level are from other class of Critical Risk which is 232 in No.
- ✓ The severity of the Potential Accident Level is from the class Fall, Electrical installation, Vehicles, Projection, Pressed and Mobile equipment.

MODEL TUNNING

MACHINE LEARNING CLASSIFIER

Approach:

- Preprocessing
- Feature Engineering
- Encoding
- TIDF Vectorization
- Up Sampling
- Training & testing Model
- Classification Report for all the machine learning models

MACHINE LERANING MODELS SUMMARY REPORT

With Sampling

	Method	Train_Accuracy	Test_Accuracy	Precision score	Recall	F1score
1	Logistic regression	0.381377	0.261905	0.578291	0.261905	0.345706
2	RidgeClassifier	0.800000	0.392857	0.705034	0.392857	0.489311
3	KNeighborsClassifier	0.951417	0.404762	0.619183	0.404762	0.475959
4	SVC	0.289069	0.071429	0.025661	0.071429	0.033149
5	DecisionTreeClassifier	0.998381	0.630952	0.632864	0.630952	0.629707
6	RandomForestClassifier	0.998381	0.761905	0.700000	0.761905	0.711525
7	BaggingClassifier	0.998381	0.726190	0.673964	0.726190	0.698871
8	AdaBoostClassifier	0.365992	0.690476	0.559988	0.690476	0.618043
9	GradientBoostingClassifier	0.968421	0.523810	0.636409	0.523810	0.565801
10	LGBMClassifier	0.998381	0.714286	0.639456	0.714286	0.674315
11	XGBClassifier	0.995142	0.619048	0.624642	0.619048	0.621409

By comparing the results from all above methods,RandomForestClassifier is performing better with F1 score of 71.1% but overfit in training set

With Original Data

	Method	Train_Accuracy	Test_Accuracy	Precision score	Recall	F1score
1	Logistic regression	0.739521	0.738095	0.544785	0.738095	0.626875
2	RidgeClassifier	0.757485	0.726190	0.555850	0.726190	0.629704
3	KNeighborsClassifier	0.754491	0.690476	0.567177	0.690476	0.622429
4	SVC	0.739521	0.738095	0.544785	0.738095	0.626875
5	DecisionTreeClassifier	0.994012	0.654762	0.638983	0.654762	0.646408
6	RandomForestClassifier	0.967066	0.726190	0.562798	0.726190	0.634138
7	BaggingClassifier	0.961078	0.714286	0.546737	0.714286	0.619381
8	AdaBoostClassifier	0.736527	0.714286	0.627396	0.714286	0.650885
9	GradientBoostingClassifier	0.934132	0.714286	0.658175	0.714286	0.658399
10	LGBMClassifier	0.991018	0.750000	0.650000	0.750000	0.690951
11	XGBClassifier	0.826347	0.750000	0.632784	0.750000	0.676386

By comparing the results from all above methods, we can select the best method as AdaBoost classifier with f1-score 65.08% , some of them are overfitting in training set


END
OF
INTERM
REPORT