|  |
| --- |
| strath_main.jpg |
| Aspects of Oil Pipeline Accidents that Leads to Pipeline Shutdown |
| US Transportation Project |
|  |
|  |
| **2nd November 2022** |
|  |

|  |
| --- |
|  |

**Table of Contents**

**List of Tables ……………………………………………………………………………………………… 03**

**List of Figures .……………………………………………………………………………………………. 04**

1. **Introduction ………………………………………………………………………………… 05**
2. **Dataset (2010-2017 Oil Pipeline Accidents) ………………………………….. 06 - 19**
   1. Aim ………………………………………………………………………………… 06
   2. Source ………………………………………………………………………………… 06 - 07
   3. Analysis ………………………………………………………………………………… 07 - 19
3. **Unsupervised Analysis – Clustering ……………………………………………… 19 - 21**
   1. K-Means ………………………………………………………………………………… 20 - 21
4. **Supervised Analysis ……………………………………………………………………. 22 - 23**
   1. K-Neighbors Classifier ……………………….………………………………… 22 - 23
5. **Reflection ……………………………………………………………………………........... 24**
6. **Conclusion ……………………………………………………………………………………….. 25**

**Bibliography …………………………………………………………………………………………… 26 - 27**

**Appendix …………………………………………………………………………………………....... 28**

**List of Tables**

1. Feature Table Used for Analysis ……………………………………………………………. 07
2. Supervised Learning Output Table ………………………………………………………….. 23

**List of Figures**

1. Accident Count - Bar Chart …………………………………………………………………........ 08
2. Group by Pipeline Shutdown and Company Name - Bar Chart …………………. 08
3. Pipeline Type Accidents - Pie Chart …………………………………………………………... 09
4. Group by Pipeline Shutdown and Pipeline Type - Bar Chart ………………………… 10
5. Liquid Type Accidents - Pie Chart ………………………………………………………………. 11
6. Group by Pipeline Shutdown and Liquid Type - Bar Chart …………………………… 11
7. Cause Category Type Accidents - Pie Chart ………………………………………………… 12
8. Group by Cause Category and Pipeline Shutdown - Bar Chart ……………………… 13
9. Group by Pipeline Shutdown and Liquid Ignition - Bar Chart ……………………… 14
10. Group by Pipeline Shutdown and Liquid Explosion - Bar Chart …………………… 14
11. Emergency Response Cost vs. Pipeline Shutdown - Scatter Plot …………………. 15
12. All Cost vs. Pipeline Shutdown - Scatter Plot ……………………………………………… 16
13. Unintentional Barrel Release vs. Pipeline Shutdown - Scatter Plot ……………… 17
14. Liquid Recovery vs. Pipeline Shutdown - Scatter Plot …………………………………. 17
15. Net Loss vs. Pipeline Shutdown - Scatter Plot ……………………………………………. 18
16. All Variables - Heat Map ……………………………………………………………………………. 19
17. Scores vs. Clusters - Line Graph …………………………………………………………………... 21

**Introduction:**

Oil accidents are one of the vital points for environmental damage and climate change. Though not frequently publicized or talked about, 2018 Resource Watch report highlighted that approximately 11 accidents can occur per month (Cassidy, 2019). According to the National Research Council (NRC), on an average 1.3 million tons of oil was released in the sea as of 2002 with oil spills from accidents accounting to 12% (GPA, no date).

Although there are rules and regulations set by the US government (Bureau of Safety and Environmental Enforcement) on operations of the Oil Facilities, there are large scale accidents that can cause major disasters like ‘The Deepwater Horizon Disaster’ in 2010, where the Entire Deepwater Horizon burnt for a day and sank. It caused havoc on water, air and land habitats and disrupted the environment. Prime factors that led to this disaster were lack of maintenance, operator error and untimely closure of the pipeline itself (Arnold & Itkin, no date).

The majority of accidents in US occurs due to similar reasons. An article by NRDC conveys that only 5% of US Oil and Gas Operators follow the regulations subjected to Pipeline and Hazardous Materials Safety Administration (PHMSA). Also, according to the article, major accidents involve hazardous liquid pipelines (64%) and on the contrary new pipelines are more vulnerable to accidents compared to older ones (Mall, 2019).

Considering these multiple factors influence the oil pipeline accidents, if we can predict and timely shut the pipeline for minor incidences then the occurrences of major accidents can be prevented. This report therefore investigates 2010 to 2017 US Oil Accident data in order to understand the relationship between features of the pipeline and pipeline shutdown so that there is a less likelihood of pipeline accidents.

**Dataset** **(2010-2017 Oil Pipeline Accidents):**

**2.1. Aim:**

To establish relationship between Pipeline Features and Oil Pipeline Shutdown and to apply prediction models in order to prevent accidents by shutting down the pipeline at the right time.

**2.2. Source:**

The dataset used for this report was taken from Department of Transportation, 2016 (refer ***Appendix 3***) which was collected and published by the U.S. Department of Transportation (DOT).

The original dataset consists of 48 columns and 2795 rows, and it is multivariate in nature. The dataset reports oil pipeline leak or spills reported to DOT since 2010. In the dataset each row represents an accident occurred and each column represents features of the accidents, e.g.: Year associated to the accident, Operator Name, Location, Cause of accidents, Pipeline Type, Type of Hazardous Liquid and Cost associated with it.

The data was segregated, and redundant columns were identified and removed by analyzing the dataset subjected to the Output Column (Pipeline Shutdown). The redundant columns were identified by establishing a link with the Output Column and to achieve more precision with our modeling, columns with more than 50% of missing data were discarded. Also, to improve the analysis further, the missing rows of the dataset were cleaned as it had only 9% of missing data.

Finally, the cleaned dataset was stored in a 2-D Pandas dataset of dimension 12 columns and 2560 rows. Refer Table 2.1 for the list of the columns used for the dataset.

***Table 1:*** Feature Table Used for Analysis

|  |  |
| --- | --- |
| **Features** | **Output** |
| * Operator Name | Pipeline Shutdown (YES/NO) |
| * Pipeline Type |
| * Liquid Type |
| * Cause Category |
| * Cause Subcategory |
| * Liquid Ignition |
| * Liquid Explosion |
| * Emergency Response Costs |
| * All Costs |
| * Unintentional Barrel Release |
| * Liquid Recovery |
| * Net Loss |

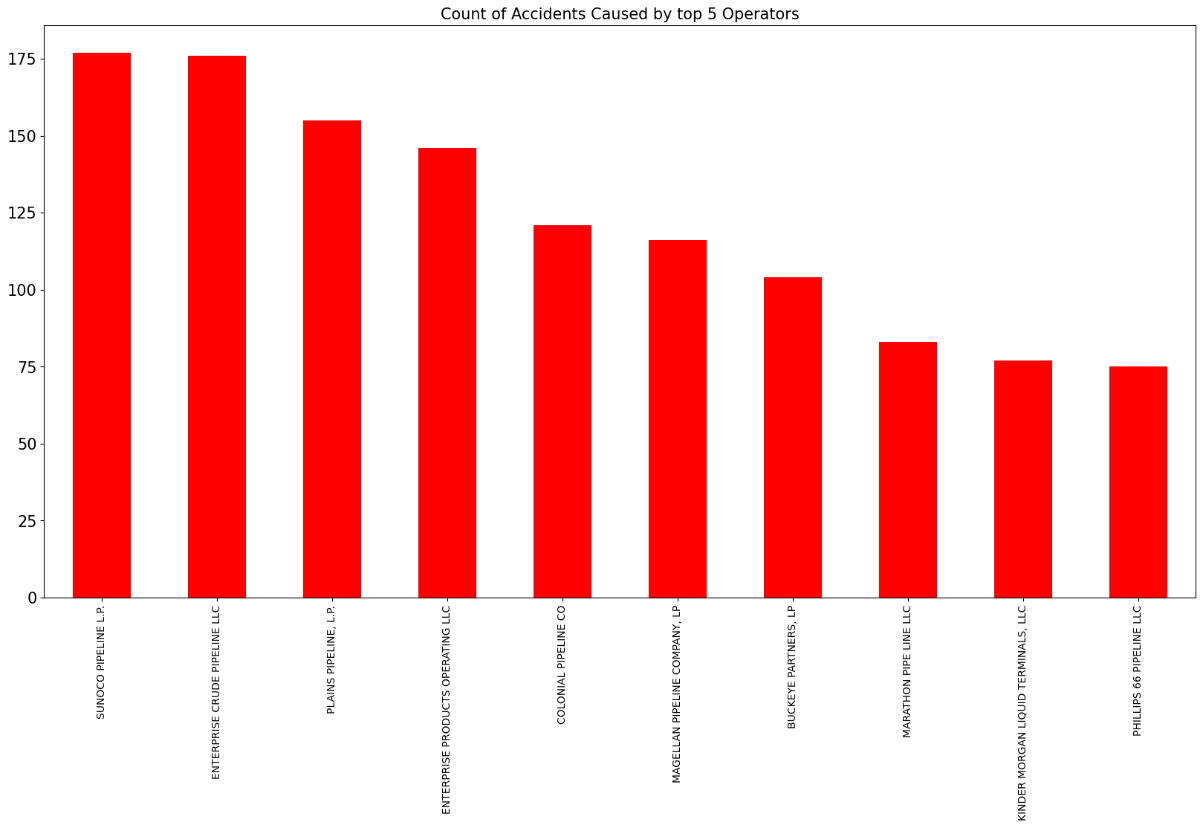
**2.3. Analysis:**

Analysis for each of the features listed in the ***Table 1*** above is performed so that a relation can established between the Features and the Output as a part of future predictions process.

* + 1. **Operator Name:**

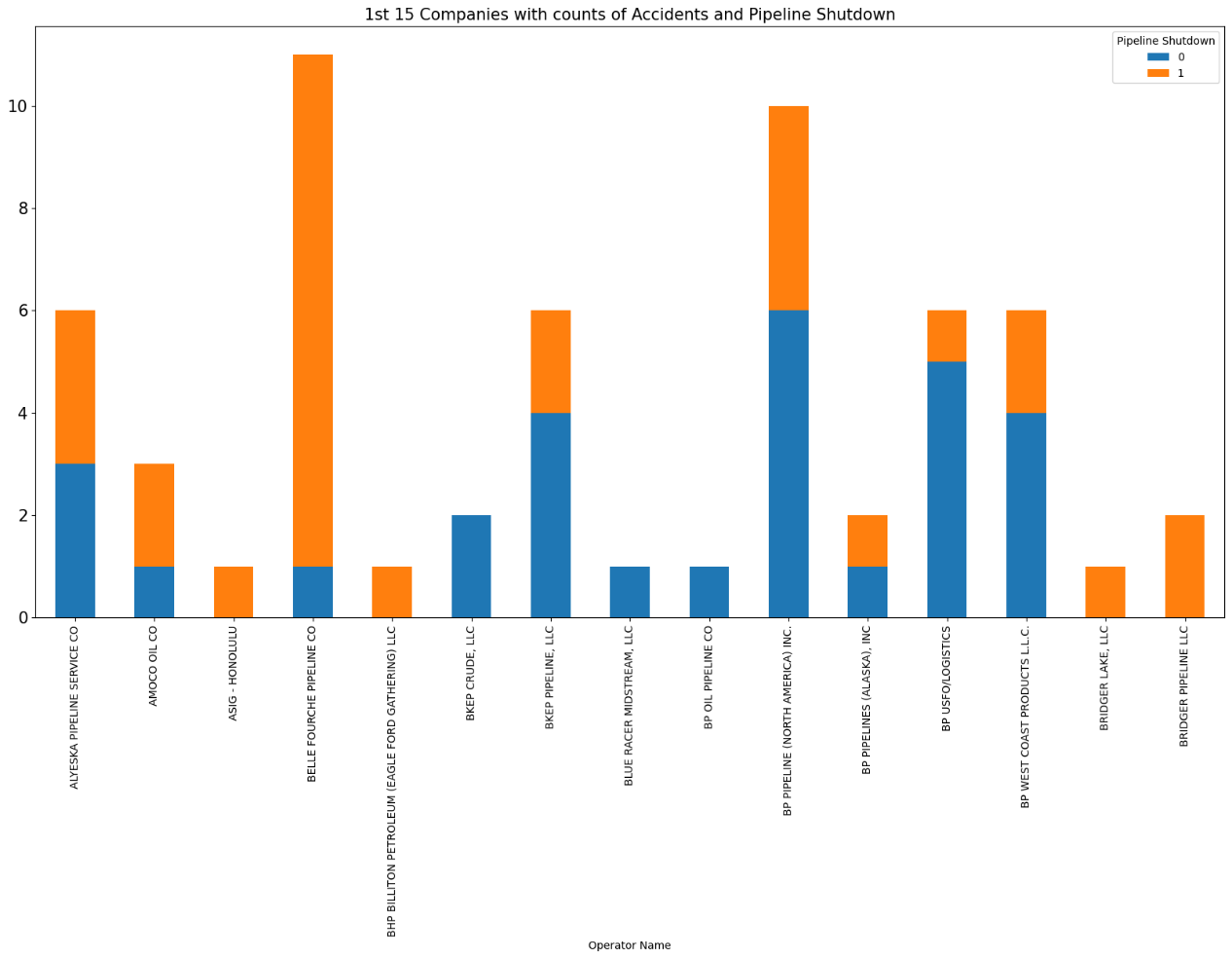
The Operator Name lists the companies responsible for the accidents. There are 220 Companies listed in ‘Operator Name’ Column (string value).The Operators run the Pipeline Facilities where all workers are appointed and trained.

***Figure 1*** below demonstrates the top 10 companies that are responsible for most accidents from 2011 to 2017, and as depicted there is a wide discrete range of accidents associated to each of the companies.



***Figure 1:*** Accident Count - Bar Chart

***Figure 2*** below depicts a diverse range of pipeline shutdown with company names and count (0 means ON, 1 means OFF), however it’s difficult to establish a relationship just by observing the bar chart. As there are both large scale and small-scale operators, we would not be able to differentiate or find a pattern if either of them is shutting off the pipeline or not.



***Figure 2:*** Group by Pipeline Shutdown and Company Name - Bar Chart

* + 1. **Pipeline Type:**

The pipeline type plays a key role in pipeline accidents in terms of observation of the pipeline to maintain it. ***Figure 3*** below shows the accidents related to the pipeline types.

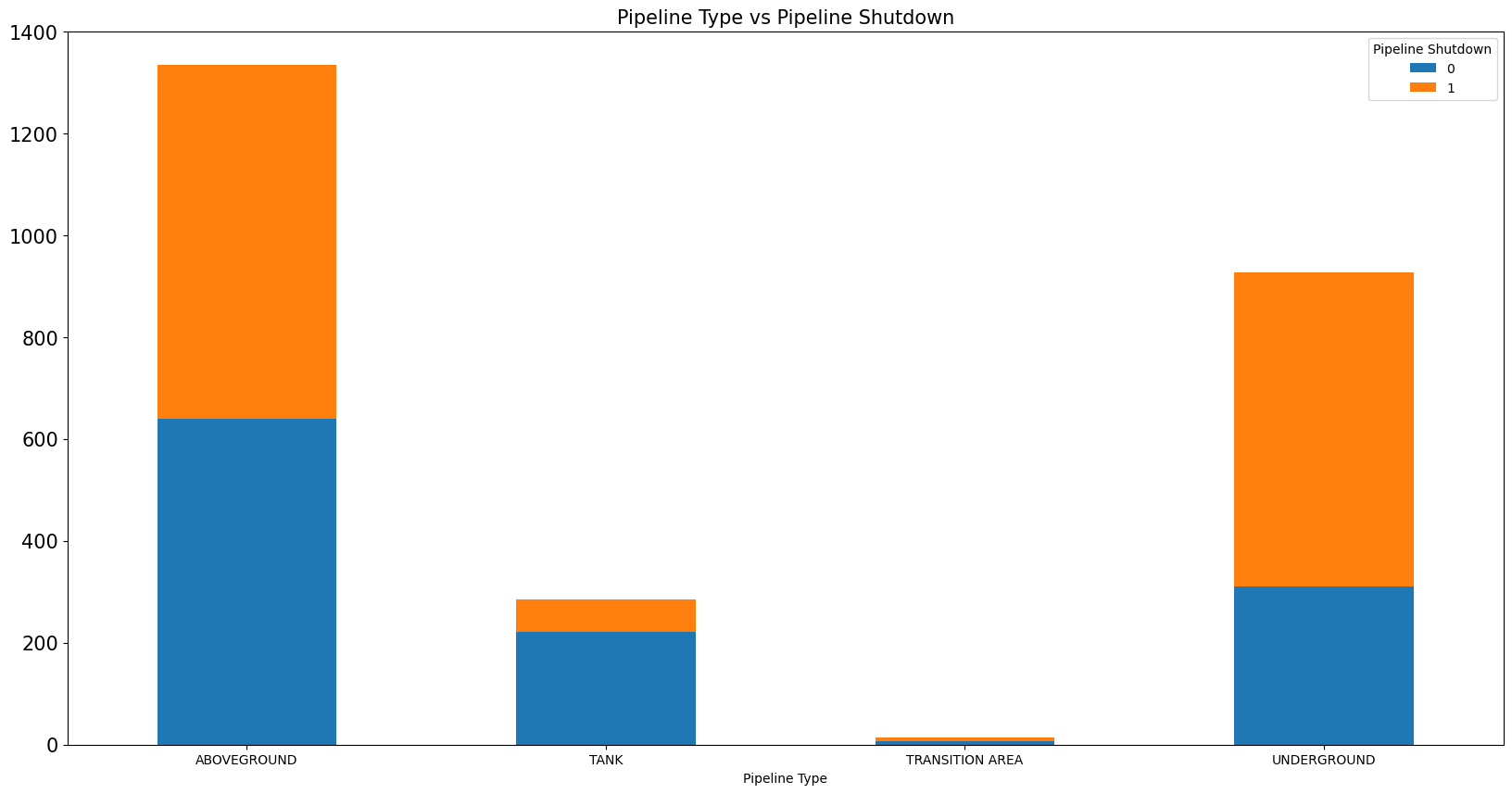
Chart, pie chart

Description automatically generated

***Figure 3:*** Pipeline Type Accidents - Pie Chart

There are 4 types of pipelines, and 52% of accidents occurred due to Aboveground pipelines, which is followed by Underground pipeline with 36.2% as these accidents are possibly caused due to corrosion, liquid flowing inside the line, and lack of maintenance. Pipeline associated with Tank and Transition area are under greater observation of the staff in and around and are thus more maintained.

***Figure 4*** below portrays that there is no definitive link between Pipeline shutdown and Pipeline Type, but we can say ‘Tank’ Type pipeline are more vulnerable to accidents as they are left ‘ON’ most of the times which makes them vulnerable for accidents even though they are observed regularly by their Staff Facility. Also, approximately 50% of Aboveground and 60% of Underground pipelines are left ‘ON’ making them vulnerable to major accidents.

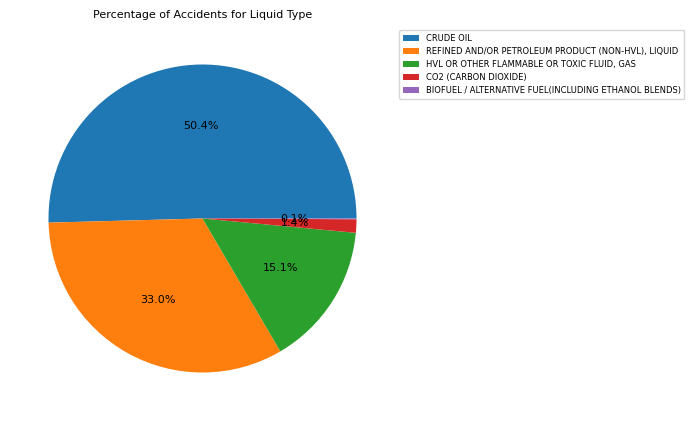


***Figure 4:*** Group by Pipeline Shutdown and Pipeline Type - Bar Chart

* + 1. **Liquid Type:**

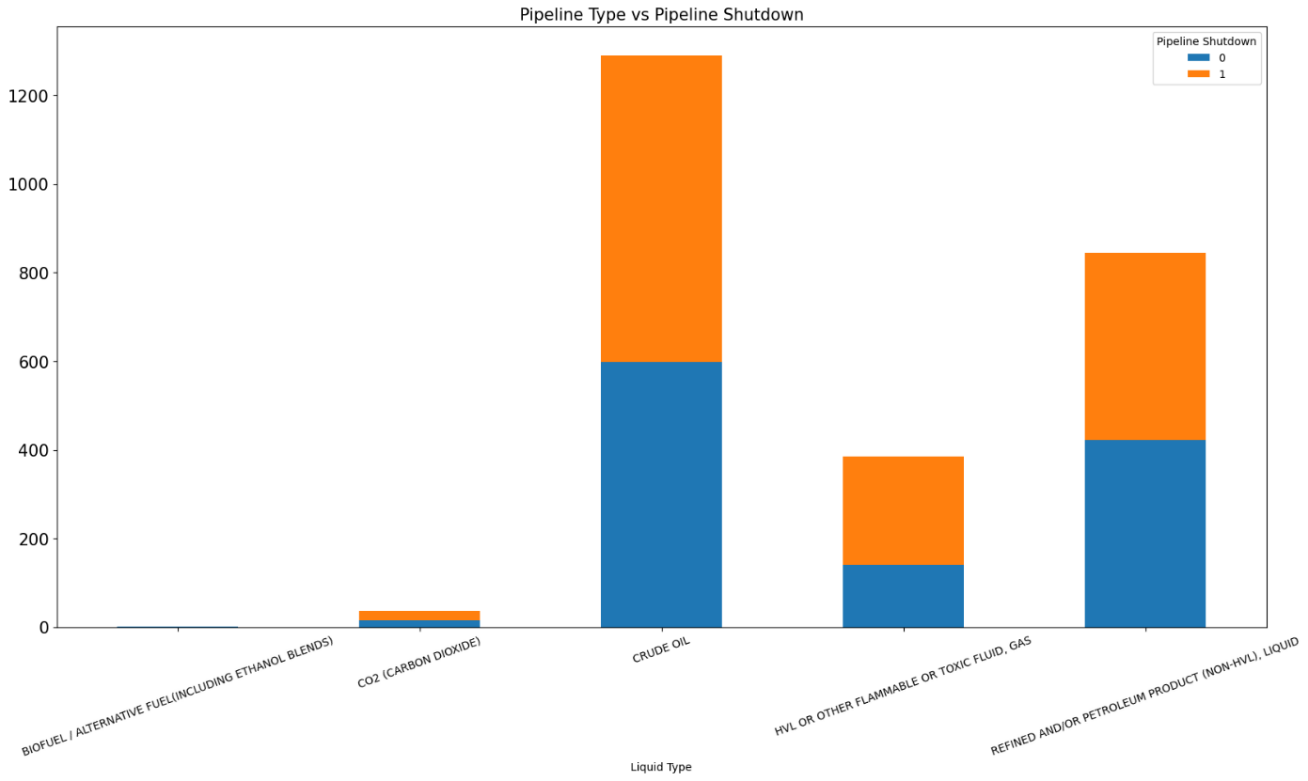
The Liquid Type indicates which liquid is passing from the pipeline and has an important role in pipeline accidents. Liquids such as Crude oil and Petroleum are highly flammable and majorly cause explosion and fire on the contrary HLV or Gas are highly toxic and poisonous that can cause severe damage when spread.

***Figure 5*** below shows the percentage of accidents. 50.4% and 33.0% of accidents are through crude oil, which further explains that highly flammable liquids cause more damages and accidents indeed. HLV and other Gases also cause a considerable number of accidents.

****

***Figure 5:*** Liquid Type Accidents - Pie Chart

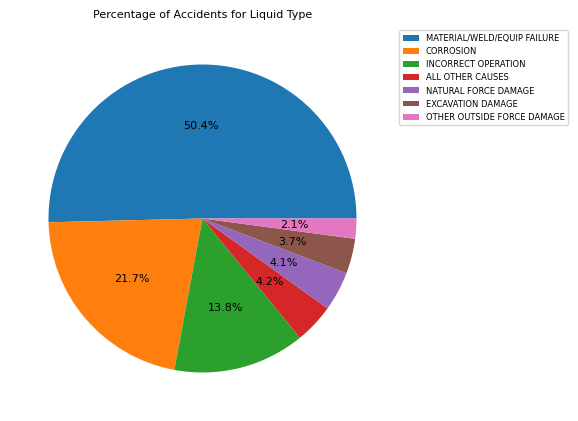
***Figure 6*** below points that even though there were accidents related to the Crude Oil, HLV and Petroleum, the operator closed the pipeline only approximately 50% of the time and this is an example of ignorance by the operator which can even cause severe damage.



***Figure 6:*** Group by Pipeline Shutdown and Liquid Type - Bar Chart

* + 1. **Cause Category:**

The Cause Category defines the cause by which the accident has occurred. The Cause Category can be a good factor to determine where the US Operators need to work so that they can prevent accident from happening. ***Figure 7*** below demonstrates that vast majority of the accidents are due to maintenance of the pipeline which is clearly the operator’s fault.

****

***Figure 7:***Cause Category Type Accidents - Pie Chart

In reference to ***Figure 8*** below, further investigation suggests that Cause Category and Pipeline Shutdown is not much corelated, but in case of ‘Excavation Damage’, the pipeline is mostly closed.

**Chart, bar chart

Description automatically generated**

***Figure 8:*** Group by Cause Category and Pipeline Shutdown - Bar Chart

* + 1. **Cause Subcategory:**

Cause subcategory is a part of cause category and normally describes the ‘Cause Category’ in detail. This Column is a supportive information of the ‘Cause Category’ column.

* + 1. **Liquid Ignition:**

The Liquid Ignition describes the possibility of fire in pipelines. ***Figure 9*** below highlights that even when there was a fire in the pipeline, the pipeline was not shut down completely. This is a strict violation of rules set by DOT and is completely the operator’s fault.

This could be because the operators are usually more interested in profits as they would generate more oil by simply keeping the pipeline ON.

**Chart, bar chart

Description automatically generated**

***Figure 9:*** Group by Pipeline Shutdown and Liquid Ignition - Bar Chart

* + 1. **Liquid Explosion:**

Liquid Explosion is explained with relation to pipeline shutdown in ***Figure 10*** below which shows that very few accidents result to Liquid Explosion.

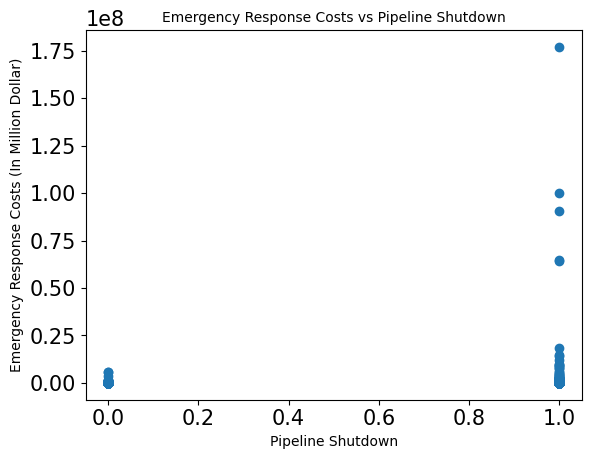
**Chart, bar chart

Description automatically generated**

***Figure 10:*** Group by Pipeline Shutdown and Liquid Explosion - Bar Chart

* + 1. **Emergency Response Costs:**

Emergency Response Cost is the cost bared by the operators to nullify the consequences of the accidents and rescue the workers and equipment. Emergency Response Cost is highly corelated with the pipeline shutdown and explains why the Operators do not close the pipeline very often during accidents as deduced by ***Figure 11*** below.

****

***Figure 11:*** Emergency Response Cost vs. Pipeline Shutdown – Scatter Plot

* + 1. **All Costs:**

All Costs is the total cost which was spent to overcome the accident damages and generally the major components include Emergency Response cost, and Environmental cost. Also, ***Figure 12*** below shows the same pattern as Emergency Response Cost pattern depicted in ***Figure 11*** aboveand explains the reason behind the pipeline’s ‘ON’ state.

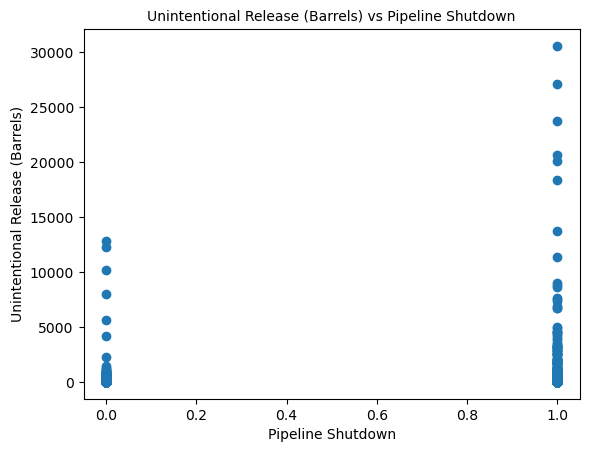
**Chart

Description automatically generated**

***Figure 12:*** All Cost vs. Pipeline Shutdown – Scatter Plot

* + 1. **Unintentional Barrel Release:**

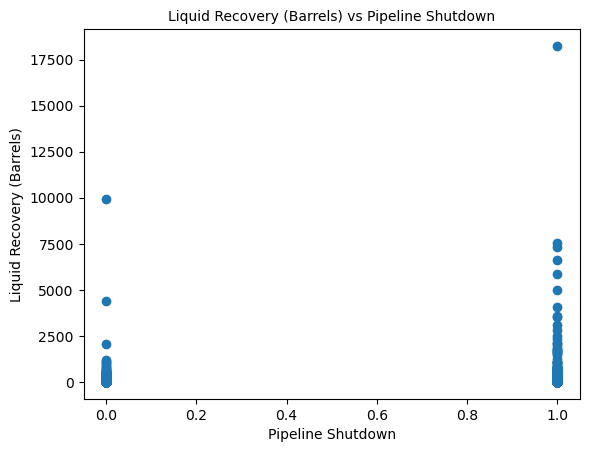
This column explains the Oil Spill in barrels due to unintentional activities.This phenomenon shows (refer ***Figure 13*** below) that there is a very unusual correlation between unintentional release and Pipeline shutdown. Because, when the pipeline is in ‘OFF’ state the barrel release is high and there can be other factors that can come in play.

****

***Figure 13:*** Unintentional Barrel Release vs. Pipeline Shutdown – Scatter Plot

* + 1. **Liquid Recovery:**

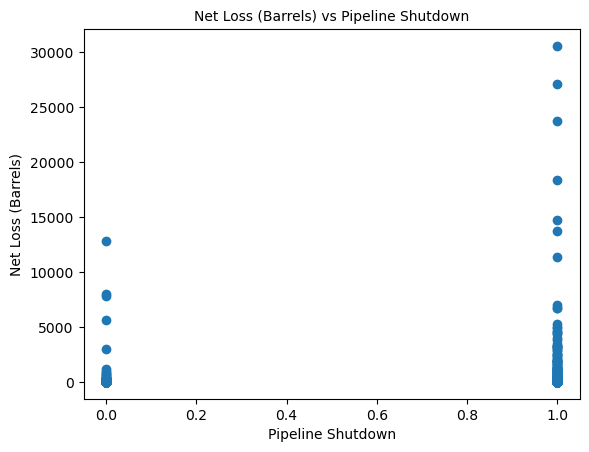
***Figure 14*** below indicates that there is a high amount of liquid recovered when the pipeline is closed but since the cost of the recovery is way more than generating new oil, operators might not always opt to close the pipeline.

****

***Figure 14:*** Liquid Recovery vs. Pipeline Shutdown – Scatter Plot

* + 1. **Net Loss:**

From ***Figure 15*** below we can infer that the Net Loss also has significant relationship with the pipeline’s shut state, and when net loss was high, the pipeline was shutdown most often.

****

***Figure 15:*** Net Loss vs. Pipeline Shutdown – Scatter Plot

* 1. **Synopsis:**

The entire analysis done above points that the Pipeline Type, Liquid Ignition, Liquid Explosion, Emergency Response Cost, All Costs, Unintentional Barrel Release, Liquid Recovery and Net Loss are significant factors that influence our output variable.

Also, ***Figure 16*** below establishes further correlation amongst the Input columns such as:

* Liquid Recovery has a very positive relationship with Emergency Response Cost, All Costs, and Unintentional Barrel Release.
* Liquid Ignition and Liquid Explosion has a very positive relation with Unintentional Barrel Release which resulted into increased Net Loss.
* Also, to some extent the Cause Category is positively linked to Liquid Type.

**Chart

Description automatically generated**

***Figure 16:*** All Variables – Heat Map

**Unsupervised Analysis – Clustering:**

Until now we have only analyzed the data through graphs and charts to find a link between our output variable and its features. This part aims to further analyze the data to find any hidden patterns or any data groupings that were missed from analyzing above. Considering the data is very discrete in nature and many of the features have strong relationship among them, it’s difficult to state whether Unsupervised learning would work or not.

For this part of analysis, since the dataset is multivariate, we must first convert the string value of the dataset into float value. To achieve this task, Ordinal Encoder is used to dedicate each string value a unique float digit (normally in terms of alphabetical order, e.g.: string starting with A will be given the value as 0, string starting with B will be given the value of 1, and so on) and paired them together as a column to complete the dataset.

So, different Columns have different number of float integers and therefore a wide range of value exist within different columns depending on the type of string values in it. The sole purpose of converting the string value to float value is because Unsupervised Analysis can only be done through float values.

**3.1. K-Means:**

K-means is an unsupervised data exploration model that plots different data point with similar features by often grouping them and placing into cluster. To form a cluster, the algorithm creates random centroids in the vector space of the points. Then, each data point gets a centroid that is near to it and a cluster is formed, however this process is recursive until all points are associated with the respective cluster and can stabilize the model so we can achieve least cumulative variation.

Chart, line chart

Description automatically generated

***Figure 17:***Scores vs. Clusters – Line Graph

As shown in ***Figure 17*** above, Number of Clusters (K) are selected as 2 to 15. The K-Means algorithm performed well in terms of silhouette score and as analyzed previously, the model has successfully demonstrated K=2 where silhouette score is 0.94, since there are only two values in the output 1 and 0 (1 – YES and 0 – NO).

However, the Completeness score and the Homogeneity Score ranges from 0 to 1 where 1 being similar types of items are placed together and 0 being no items of similar types in the cluster.

By observing the dataset closely, we realize that since there were fewer relations established between the variables themselves there is very less possibility of high homogeneity score.

**Supervised Approach:**

Supervised learning can be defined as algorithm that can predict or find patterns in the dataset when trained by the Input data (trained dataset). The purpose of this approach is to predict if our model can learn and train itself to predict given the training features at the input, if it is better to shutoff the pipeline or not after an accident so that we can prevent further damages.

Various supervised models like K-Neighbors, Decision Tree, and Gaussian NB were used to choose the best model that suits the data. As our problem is a classification type, for final modelling K Neighbors was applied since the algorithm performed better than the rest of the models as expected. But Decision Tree also delivers similar results while the Gaussian NB failed to perform.

**4.1. K-Neighbors Classifier:**

The K Neighbors applies non-parametric means where all the training data is generated by testing phase. It has many advantages over other models as this model train faster and is easy to implement. As the name suggests, the model assumes its neighboring point as a single class. The nearest neighboring class is calculated by calculating Euclidean Distance between the points.

***Table 2*** below shows the performance of K Neighbors, the F1-score (Weighted Average of Precision and Recall) and the Precision average (0.63). Also, the table indicates that the model predicts OFF value better than ON value, this can be explained as there might be more data and features linked with OFF value. Additionally, to improve the output, we have calibrated the random state of the algorithm to increase the precision.

However, adapting this model fully is difficult due to the result of 144 False Positive and 142 False Negative while predicting which is not suitable for the prediction. For further improvement we need more features and rows that can relate to the output.

***Table 2:***Supervised Learning Output Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1- Score** | **Support** |
| **0.0 (ON)** | 0.61 | 0.60 | 0.60 | 362 |
| **1.0 (OFF)** | 0.65 | 0.65 | 0.65 | 406 |
| **Accuracy** | - | - | 0.63 | 768 |
| **Macro Avg.** | 0.63 | 0.63 | 0.63 | 768 |
| **Weighted Avg.** | 0.63 | 0.63 | 0.63 | 768 |

**Conclusion:**

To conclude based on the observations made above; we would be able to deduce that there is a decent positive co-relation between the output and the features.

We had expected some features to be very influenced by the output, but it was not the case. For example, Cause Category and Liquid Type were not answerable.

The observations made above were justified by the Supervised and the Unsupervised models. Though the unsupervised algorithm was able to demonstrate clustering, but it would not be able to interpret any informative things, as the completeness and homogeneity is far too less for concrete analysis.

Therefore, we developed a supervised model that can predict if the pipeline should be shutdown or not in case of an accident with 63% accuracy, which is not bad accuracy considering the input data consisted of 2560 rows only.

Finally, I am very positive about the supervised model as it can be improved further by providing more rows and more sounding features.

**Reflections:**

Attending lectures, practicing coding and preparation of this report has helped me to dive deeper into a lot of concepts and techniques. Firstly, I realized that working with real life dataset can never have perfect data for analysis. Hence, understanding the dataset and the problems associated with it requires time and skills in order to figure out the core objective of the problem to solve it aptly.

Secondly to achieve precision of analysis, I figured that it’s best to remove many columns (features) that would have interlinked with the output while cleaning the dataset. Furthermore, I learnt that there are some features that may not directly exhibit to us that they would have a link, yet they serve its purpose such as ‘Unintentional Barrel Release.’

Finally, I can say that there are no perfect models for a dataset, all models serve its own purpose, and it depends on the application and the problem you are trying to solve. Despite this, each model can work as a guideline for others to refer to from time to time to work on their dataset problems.

**Bibliography**

Arnold & Itkin (no date). *Major Oil Rig Disasters: History’s Largest Oil Rig Explosions & Fires.* Available at: <https://www.arnolditkin.com/oil-rig-explosions/major-oil-rig-disasters/> [Accessed: 21 October 2022].

Britannica (no date). *Environmental costs: Deepwater Horizon oil spill: avian casualty.* Available at: <https://www.britannica.com/event/Deepwater-Horizon-oil-spill/Environmental-costs> [Accessed: 18 October 2022].

Bureau of Safety and Environmental Enforcement (2018). *Oil and Gas and Sulphur Operations on the Outer Continental Shelf-Oil and Gas Production Safety Systems.* Available at: <https://www.federalregister.gov/documents/2018/09/28/2018-21197/oil-and-gas-and-sulphur-operations-on-the-outer-continental-shelf-oil-and-gas-production-safety> [Accessed: 18 October 2022].

Cassidy, E. (2019). *There Were 137 Oil Spills in the US In 2018: See Where They Happened.* Available at: <https://blog.resourcewatch.org/2019/02/07/there-were-137-oil-spills-in-the-us-in-2018-see-where-they-happened/> [Accessed: 19 October 2022].

GPA (no date). *Global Marine Oil Pollution Information Gateway: Sources.* Available at: <http://oils.gpa.unep.org/facts/sources.htm#:~:text=discharges%20from%20consumption%20of%20oils,accidental%20spills%20from%20ships%3B%2012%25> [Accessed: 21 October 2022].

Mall, A. (2019). *Pipeline Incident Statistics Reveal Significant data.* Available at: <https://www.nrdc.org/experts/amy-mall/pipeline-incident-statistics-reveal-significant-dangers> [Accessed: 23 October 2022].

Pedregosa, F. *et al.* (2011). *Scikit-learn: Machine Learning in {P}ython}: sklearn.neighbors.KNeighborsClassifier.*Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html> [Accessed: 29 October 2022].

U.S. Department of Transportation (2022). *Distribution, Transmission & Gathering, LNG, and Liquid Accident and Incident Data.* Available at: <https://www.phmsa.dot.gov/data-and-statistics/pipeline/distribution-transmission-gathering-lng-and-liquid-accident-and-incident-data> [Accessed: 19 October 2022].

U.S. Department of Transportation (2022). *Incident Reporting.* Available at: <https://www.phmsa.dot.gov/incident-reporting> [Accessed: 19 October 2022].

Vairo, T. *et al.* (2017). *An Oil Pipeline Catastrophic Failure: Accident Scenario Modelling and Emergency Response Development.* Available at: <https://www.aidic.it/cet/17/57/063.pdf> [Accessed: 20 October 2022].

**Appendices**

1. Environment: Python version: 3.9.13
2. Jupyter Notebook version: 6.4.12
3. ***Dataset:***

Department of Transportation (2016). *Oil Pipeline Accidents, 2010-Present: Causes, injuries/fatalities, and costs of pipeline leaks and spills.* Available at: <https://www.kaggle.com/datasets/usdot/pipeline-accidents> [Accessed: 16 October 2022].