Using Machine Learning and Keyword Analysis to Analyze Incident Reports and Reduce Risk in Oil Sands Operations

by

Daniel G. Kurian

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Engineering Management

Department of Mechanical Engineering
University of Alberta

ABSTRACT

Many companies maintain large databases of incident reports. Incidents that have severe consequences are analyzed in detail to prevent recurrence, while minor incidents are typically stored without any further evaluation. Especially with common incidents and those with lesser consequences, details that are necessary to understand the cause of the incident might be missing. Incidents that occur in the oil and gas industry can be reported more accurately and analyzed to provide value to companies maintaining databases, and to prevent and mitigate risks. Such information can be used to lower costs and improve safety culture.

The initial objective of this study was to create a risk matrix system for collectively analyzing incident reports, commensurate across companies, for increased reliability in reporting and enhanced analytical power across an industry. A supervised machine learning approach was applied in conjunction with this risk matrix to analyze incident reports and provide a risk score.

During this research project, 15,000 incident reports, including both process and occupational-type incidents, were analyzed from three oil sand companies across Alberta. The results were classified by incident type (determined by industry experts) and consequence type (using the risk matrix). Furthermore, potential and actual risk scores were evaluated for every incident using the risk matrix. This analysis built the foundation for a system to identify trends and leading indicators, and to design prevention and mitigation strategies across the entire industry.

The goals of this researched evolved to include the application of artificial intelligence and machine learning to create a digitalized system for efficiently reporting incidents that can be used to generate a risk matrix, trend report, prevention and mitigation strategies, and leading indicator identification for every incident report that is inputted.

Implementing this system was accomplished by utilizing a combination of supervised machine learning and keyword analysis. During this research project, the 15,000 incident reports were analyzed to build a customized library of keywords. These keywords were assigned to a list of statements that were generated using a company's safety guidelines, standard operating procedures, and asset management systems. The basic structure for generating outputs was demonstrated using a large incident database provided by collaborators of the project and some sample inputs. Three case studies of incident reports were also processed and presented using the proposed methodology, delivering practical outputs that could be used by workers and companies to improve safety and increase hazard awareness.

PREFACE

Chapter 2 of this thesis has been submitted for publishing as Kurian, D., Ma, Y., Lefsrud, L., and Sattari, F. “Seeing the Forest and the Trees: Using Machine Learning to Categorize and Analyze Incident Reports for Alberta Oil Sands Operators,” Journal of Loss Prevention in the Process Industries.

Chapter 3 of this research has been submitted for publishing as Kurian, D., Ma, Y., Lefsrud, L., and Sattari, F. “Using Machine Learning and Keyword Analysis to Analyze Incidents and Reduce Risk in Oil Sands Operations,” Journal of Safety Science.

I was responsible for the data analysis as well as the manuscript composition. Dr. Yongsheng Ma was the main supervisor for this project and contributed to manuscript edits. Dr. Lianne Lefsrud was the co-supervisor and was responsible for data collection and to edit manuscripts. Dr. Fereshteh Sattari was involved in providing research direction, editing manuscripts, and assisting with publishing procedures.

ACKNOWLEDGEMENTS

I would like to thank Dr. Yongsheng Ma and Dr. Lianne Lefsrud, my supervisor and co-supervisor respectively, for their on-going support of my research. They have encouraged and motivated me to complete my Master's degree for the past couple years, and I appreciate the opportunity to express my past year of work in writing this paper. I am thankful to Dr. Ma for his professional knowledge which helped to provide research direction through the completion of this degree. I am also immensely grateful to Dr. Lefsrud who was instrumental in the process of data collection, providing research funding, and creating opportunities to present my research.

I wish to express my sincere gratitude to post-doctoral fellow, Dr. Fereshteh Sattari, for her advice in guiding my research – she provided guidance for my research and helped me with editing manuscripts and revising presentations.

I would like to thank Rhonald John Rose who motivated me to learn Python, structured and composed the pseudo code to analyze incident data, and guided the development of my skills with Python and machine learning.

I would like to thank the various companies who have been so generous as to provide access to their incident databases. I also want to acknowledge Kathleen Baker who completed her graduate studies using the same incident report database and other sources of data – she provided me with guidance during the start of this project.

Finally, I want to thank my parents, Daince Kurian and Shyla Kurian-George, for supporting me throughout my Master’s program, financially and otherwise. Their love and support have helped me achieve all I have to this day.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1. BACKGROUND

The data analyzed in this research is the incident databases of several large companies involved in the energy sector in Alberta. By participating in this study, these companies were hoping to see overall mitigation and prevention of risk. With this in mind, different companies have different standards for the details involved in incident reporting. Furthermore, these incident reports are treated with varying levels of importance by various persons within the organization. My initial impression of the incident database was a large number of low consequence incidents that constantly recur with disproportionate number of high consequence incidents that rarely occur.

Incident reports are written to document the details of an incident and to provide information that has the potential to be useful in the future (Kane, 1985). This information can be used in a variety of different ways – for the purpose of investigation, as legal documentation, and to learn from the past to prevent similar occurrences in the future (Macrae, 2015; Christou and Konstantinidou, 2013). In order to fulfil these requirements, the reports need to accurately describe the incident, contain complete information, and include details to facilitate future investigation. An incident report should include location, time/date, and the names of the individuals and employers involved (Occupational Health and Safety Act [OHS], 2018).

While focusing on individual incident reports, it became more obvious that the incident data was not "clean." A number of incident reports were incomplete (it was impossible to determine the event that occurred with the information provided), spelling mistakes were very common, and it was difficult to determine the causes of incidents. Unlike high consequence incidents, it was difficult to trace root causes of low consequence incidents – there was no further pertaining to such incidents.

In addition to the raw incident data, the companies participating in this research also had their own risk matrices for analyzing incidents – these risk matrices were used to evaluate the risk involved with incidents. Some companies also provided other forms of analysis, for example, root cause analysis and risk scores using the risk matrices. The same mistakes plaguing the incident reports could also be found in these risk assessments – human error and incorrect calculations were quite common.

These incident databases created many opportunities for research. First, the system for reporting incidents can be modified to allow incidents to be reported more clearly. Second, there was an opportunity to evaluate the risk involved with each incident. Finally, it was possible to suggest improvements to prevent or reduce the consequences of the incidents that were occurring in these company databases.

For this research, 15,000 incident reports were analyzed using supervised machine learning and keyword analysis. These incident reports are used to generate a "Customized Library" to analyze the input of new incident reports and to provide actionable information. Risk is evaluated using a risk matrix and trends were analyzed to identify different types of incidents and their occurrences. Prevention and mitigation strategies are also suggested using a variety of engineering and operating standards. Finally, leading indicators are identified for the provided incident reports. Leading indicators can be arranged into three groups: operations-based (pertaining the functioning of an organization’s operations), systems-based (relevant to the management of environment, health, and safety), and behavior-based (referring to the relationships between people and groups) (Inouye, n.d.). This was used as a guideline for matching leading indicators to incident reports.

1.2. RESEARCH OBJECTIVES

Our research seeks to improve current methods of incident reporting by:

- Implementing machine learning and keyword analysis to categorize incidents and analyze risk;
- Automating the process of evaluating risk by utilizing machine learning to generate risk matrices and trend reports; and
- Providing actionable information to companies – prevention and mitigation strategies and the identification of leading indicators.

The premise of using machine learning and keyword analysis is to enhance the consistency and accuracy of the methods currently used in industry. By using these methods, it is possible to reduce bias and human error in the process of incident reporting. Companies can identify incidents that should be prevented and create safer environments for their workers. The benefits of this research are twofold: the costs of damages caused by incidents can be reduced and companies will gain the trust of their workers with the employees knowing that their company is looking after their wellbeing. Consequently, workers can be trained to identify hazards and improve the safety of their workplaces.

1.3. THESIS OUTLINE

This thesis includes four chapters. Chapter 2 uses machine learning to analyze and categorize incident reports to automate the process of predicting the frequency and consequence of an incident. Chapter 3 improves this machine learning algorithm by adding another computational layer for keyword analysis. Incident reports were analyzed, and four outputs were delivered: risk matrices, trend reports, prevention and mitigation strategies, and leading indicators. Chapter 4 summarizes the results of this research.

2. SEEING THE FOREST AND THE TREES: USING MACHINE LEARNING TO CATEGORIZE AND ANALYZE INCIDENT REPORTS FOR ALBERTA OIL SANDS OPERATORS

A version of this chapter was published as Kurian, D., Ma, Y., Lefsrud, L., and Sattari, F. "Seeing the Forest and the Trees: Using Machine Learning to Categorize and Analyze Incident Reports for Alberta Oil Sands Operators," Journal of Loss Prevention in the Process Industries.

2.1. INTRODUCTION

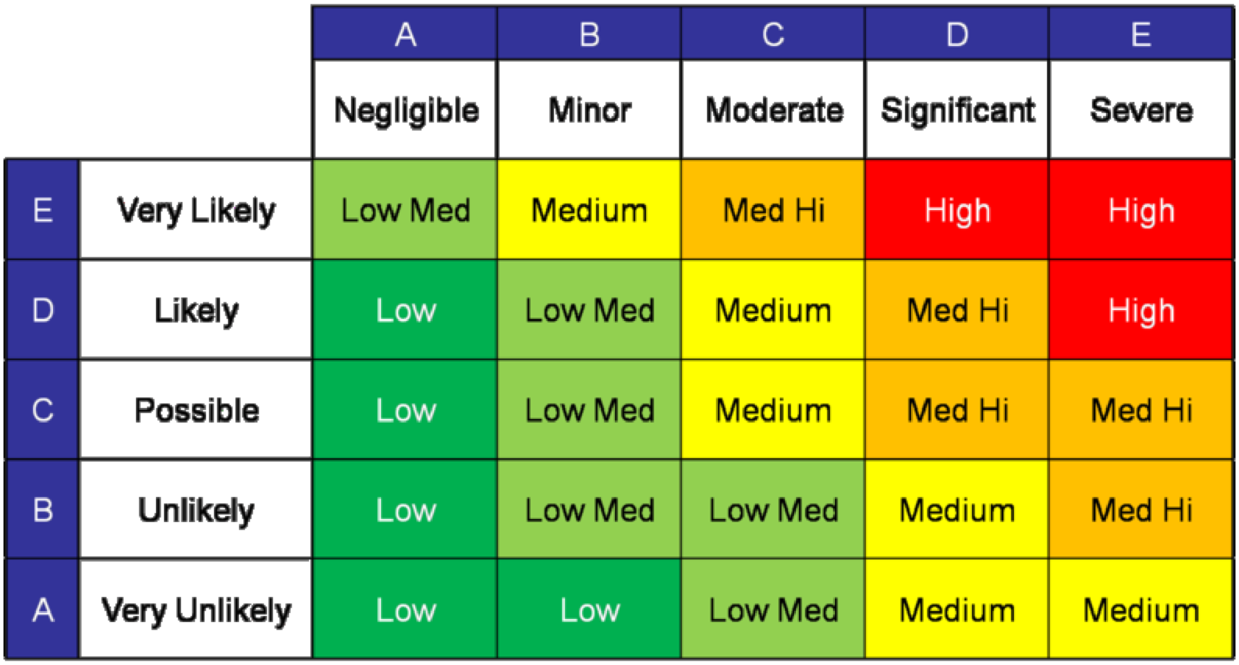
This research is motivated by an opportunity to analyze past incident reports across organizations with similar operations. In Canada, Occupational Health and Safety (OHS) is handled by provincial jurisdictions. Incident reporting is mandated by law in the province of Alberta (Occupational Health and Safety Act [OHS], 2018). Many incidents involving hazardous chemicals occur yearly, and there is always potential for loss of containment when these substances are not properly controlled. Process safety management seeks to manage hazards associated to process industries, and to reduce the risks involved with the release of hazardous chemicals (Occupational Safety and Health Administration [OSHA], 2000). When applied to the oil and gas industry, process safety incidents typically involve failure in a pipeline system or facility. Such incidents can result from small mistakes that lead to disastrous consequences, and it is important to learn from incidents to reduce the risk involved with these events (Ness, 2015).

In addition to the reporting requirement for loss incidents, many large companies have designed their own systems for the internal reporting of incidents. This includes seemingly meaningless or irrelevant incidents – low risk (either low frequency or low consequence) which simply remain in company data repositories without further analysis – and other incidents which could have had more serious consequences and, thus, trigger change in industry practice (Greenwell et al., 2003).

Risk is defined as the effect of uncertainty affecting objectives (International Organization for Standardization [ISO], 2018). Given this, a hazard may or may not lead to a loss incident, under slightly different circumstances. To evaluate the level of risk associated to an incident, companies often use a risk matrix (i.e., *A Guide to the Project Management Body of Knowledge [PMBOK]*,

2017) – a simple, yet powerful risk evaluation tool used for semi-quantitative risk analysis. In addition to the ISO (2018) definition, the PMBOK (2017) defines risk *(R)* as the product of probability *(P)* and consequence *(C)*: $R = P \times C$. This definition can be applied to the risk matrix where the total risk *(R_T)* is the sum of all risks in the system where n is the total number of risks: $R_T = \sum_{i=1}^{n} R_n$.

An example of a risk matrix can be seen in Figure 2-1. Risk matrices are often color-coordinated or zoned, usually based on the organizations' risk tolerability (Kletz, 2005; Markowski & Mannan, 2008). At some level, low risks are categorized as acceptable, some medium risks are considered to be tolerable, and high risks are categorized as intolerable and requiring reduction. This scale is often made more specific – for example, medium risks could potentially be further classified as tolerable acceptable or tolerable unacceptable. The goal of the risk acceptability principles is to understand the organization's tolerability level, assess all risks, and reduce these to an acceptable level by lowering the likelihood or minimizing the consequences (Kletz, 2005; Markowski & Mannan, 2008). The risk matrix is one of the most widely used tools for risk evaluation and prioritization – it is simple to implement, maintain, understand, and explain (Animah & Shafiee, 2019; Gul & Guneri, 2016; Landell, 2016).



		A	B	C	D	E
		Negligible	Minor	Moderate	Significant	Severe
E	Very Likely	Low Med	Medium	Med Hi	High	High
D	Likely	Low	Low Med	Medium	Med Hi	High
C	Possible	Low	Low Med	Medium	Med Hi	Med Hi
B	Unlikely	Low	Low Med	Low Med	Medium	Med Hi
A	Very Unlikely	Low	Low	Low Med	Medium	Medium

Figure 2-1. A sample risk matrix used in risk evaluation (PMBOK, 2017)

Typically, organizations file reports for all incidents that occur on-site, including process safety incidents, and categorize these into their risk matrix according to the consequences or potential consequences (health and safety of people, damage to the environment, financial loss, reputation, etc.).

While risk matrices have a variety of benefits, including assisting decision-making and prioritizing risks, they also have many weaknesses (Thomas et al., 2013). First, it is difficult to use a matrix to evaluate risk in a system with no high consequences risks. In such a case, it would be challenging to define the maximum and minimum boundaries of the consequence scale. Risk matrices are also limited by their two-dimensional precision outlook (Bjerga & Aven, 2015). Other technical weaknesses of risk matrices identified by Duijm (2015) and Thomas et al. (2013) include:

- the type of scale used to calculate risk scores (linear, logarithmic, exponential, etc.),
- the inability to classify multiple risks which may be associated to each other, or a single risk which might have multiple consequences,
- the assessment of very likely, moderate impact risks and possible, significant impact risks as being the same when they might be very different qualitatively, and
- the arbitrary nature of ranking (ascending or descending) which can play a significant role in the final stage of prioritizing and managing risk.

Another issue is the difference in risk matrices used by different companies. Even with the overall structures of the risk matrices remaining the same, varying scales demonstrate individual organizations' resilience to certain consequences. Companies will define catastrophic financial loss differently, creating radically dissimilar upper limits to high consequences items. Furthermore, in common practice, risk matrices use an open upper limit. This could mean that, theoretically, a tenfold factor between two financial losses could still be classified in an identical manner.

Aside from these "technical" issues, there are also other issues with the usage of such risk practices in industry: the bias and inconsistency of human reporting. Thomas et al. (2013) defines centering bias as the tendency of people to avoid reporting extreme values. When dealing with incident reports, skewness (i.e., low frequency, high consequence events) is more concerning than centering on 'normal' incidents. Additionally, with many individuals rating risks, it is likely that similar incidents might be rated quite differently by various people within the same company. It might even be possible for the same individual to evaluate the same incident differently at various points in time. Given this, it can be easy to conclude that reporting and assessing risks can vary significantly across companies, which prevents meaningful comparison.

Many of these issues can be resolved by introducing machine learning. Automating the process of using a rule-based matrix to evaluate risk will remove bias and allow for consistent risk ranking. Having a consistent system in place would allow for cross-company collaboration to better understand incidents and assess the associated risks. The benefits of this would extend when analyzing the data for trends and leading indicators, as a larger dataset would allow for greater statistical power when examining trends and identifying low-frequency, high-impact events. Taleb (2007) describes black swan events as events that are outliers, have a high impact, and are explainable only after they have occurred. By applying this logic to incident reports, combining several large incident databases allows companies to identify and develop strategies against hazards and latent causes found by other companies that have not yet been identified on their own sites. Using machine learning algorithms, it is also possible to create a system to rank and categorize incidents.

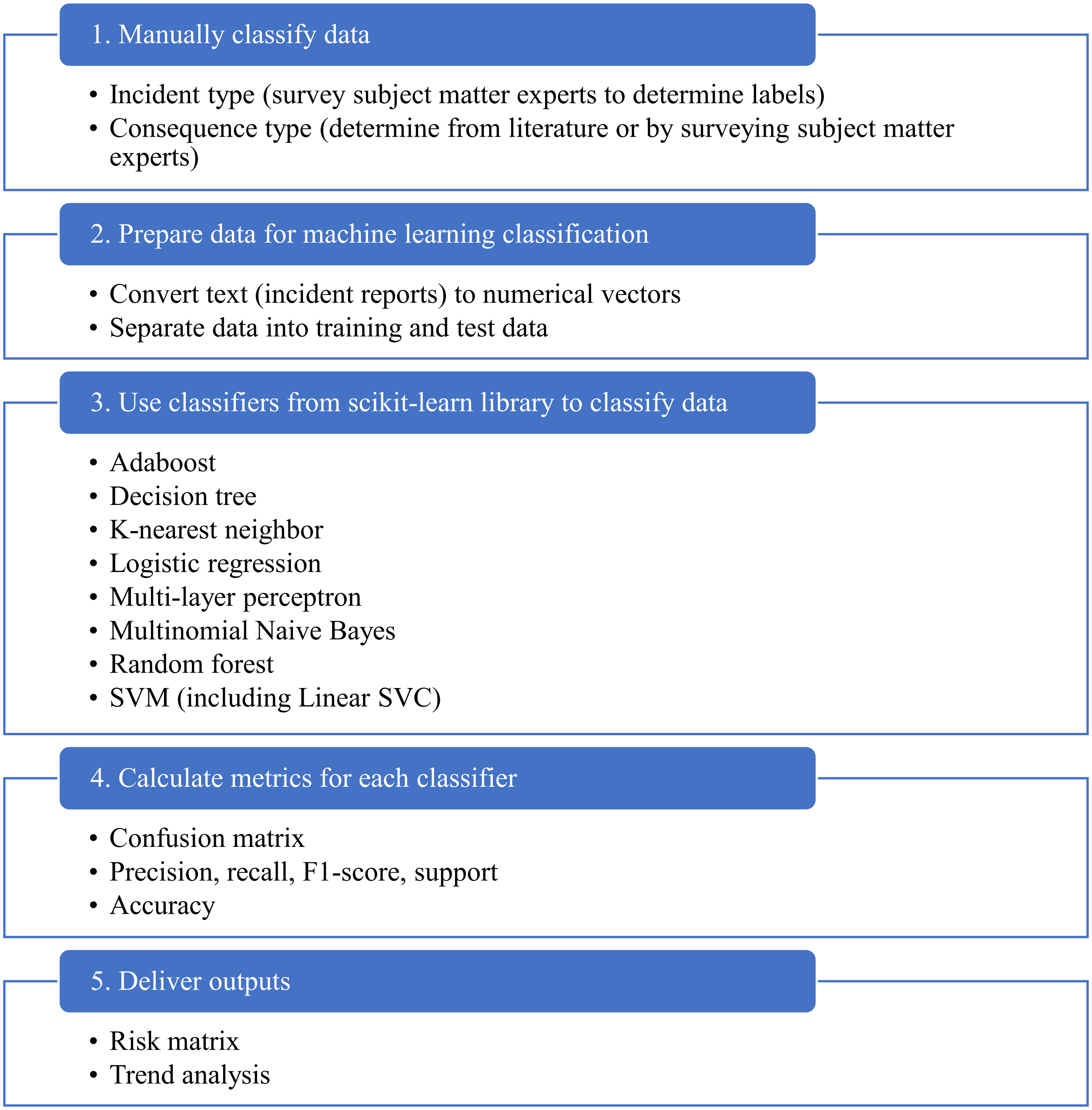
2.2. METHODOLOGY

For this research, several companies provided access to their incident databases and risk matrices to improve their current system of reporting incidents. This is accomplished by:

- creating a risk matrix with a consistent scale that can be used by all companies in Alberta’s oil sands sector and
- automating the process of classifying and categorizing risks by utilizing supervised machine learning.

The programming language, *Python*, is used to design the supervised machine learning algorithm to classify incident reports. Python is an open source project, it has many libraries that are readily available for download, and has a variety of online support through various forums. From a numerical perspective, Python has been rated the number one programming language for engineering and applied sciences (Cass, 2019).

Supervised machine learning functions by using predictor features to predict a class label (Kotsiantis, 2007). It differs from unsupervised learning in that the class labels are known when applying supervised learning. In this case, the predictor features are incident reports and class labels are assigned categories. For this research, several different class labels are applied: to provide a general category for incident reports and to score an incident report on a risk matrix (consequence type, actual risk score, and potential risk score). The process for applying supervised methodology to incident reports can be seen in Figure 2-2. In order to consider model accuracy, all incident reports used in this research must be classified manually; however, manually classifying such large amounts of data is not necessary to predict class labels for commercial use. For this research, 15,000 incident reports are used, randomly selected from the 54,000 incident reports available (provided by five oil sand companies operating in Alberta).



1. Manually classify data

- Incident type (survey subject matter experts to determine labels)
- Consequence type (determine from literature or by surveying subject matter experts)

2. Prepare data for machine learning classification

- Convert text (incident reports) to numerical vectors
- Separate data into training and test data

3. Use classifiers from scikit-learn library to classify data

- Adaboost
- Decision tree
- K-nearest neighbor
- Logistic regression
- Multi-layer perceptron
- Multinomial Naive Bayes
- Random forest
- SVM (including Linear SVC)

4. Calculate metrics for each classifier

- Confusion matrix
- Precision, recall, F1-score, support
- Accuracy

5. Deliver outputs

- Risk matrix
- Trend analysis

Figure 2-2. Process for classifying incident reports using supervised machine learning.

2.2.1 Manually Classify Data

By interviewing several industry experts, we decided to classify incidents into the following primary categories: communication, health/safety, leak/spill, miscellaneous, operation, and vehicle. These incidents can then be further classified into more specific sub-categories. The primary, secondary, and tertiary levels of classification can be seen in Figure 2-3. This form of categorization can also be used to determine the frequency of an incident occurring, which is one of the outputs necessary to calculate risk.

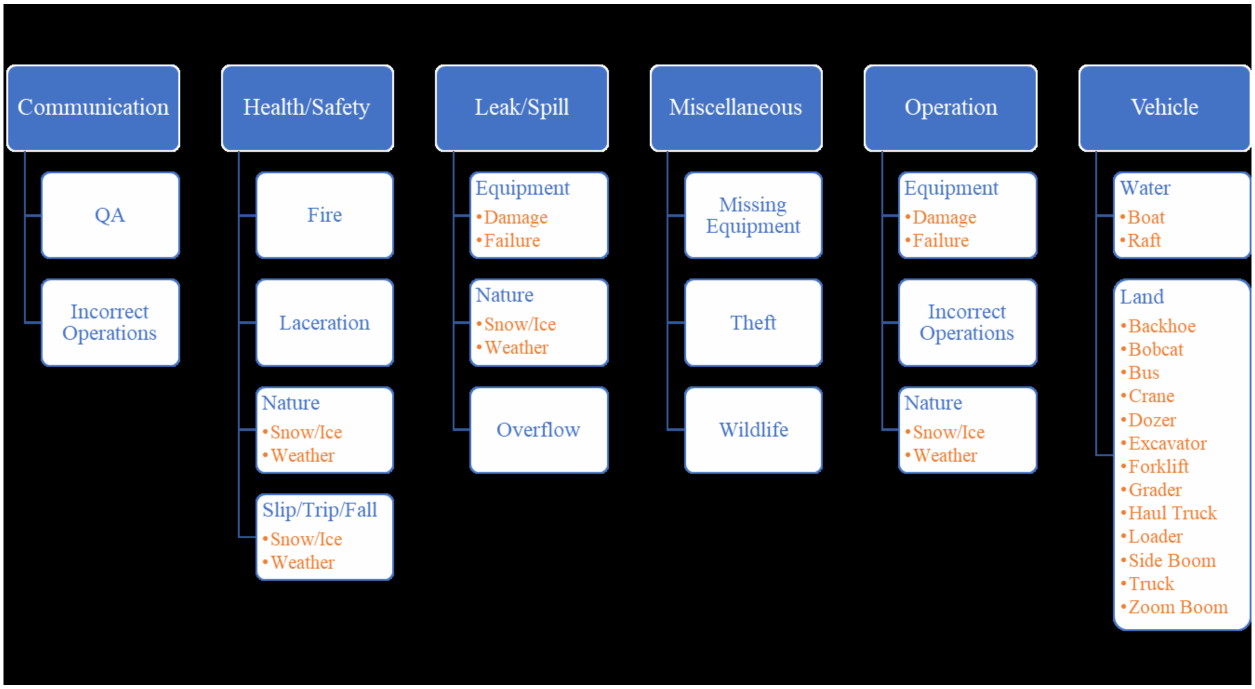


Figure 2-3. Primary (white), secondary (blue), and tertiary (orange) tiers for the classification of incident reports.

It is also necessary to apply labels to the incident reports in a manner such that an incident report can be evaluated by the risk matrix. When calculating consequence, we consider financial loss, environmental impact, damage to reputation, and worker health (Muhlbauer, 2004). Two types of scores are given to each incident report: an actual risk score, based on the consequences of the incident, and a potential risk score, based on possible "worst-case" scenarios. This potential risk score is important when analyzing near misses that have no actual consequences. To justify these risk scores, labels are applied for the type of consequence (health/safety, environment, finance, and reputation). Once a severity rating is given, a risk level can be calculated by evaluating the severity in conjunction with the frequency calculation.

Risk matrices were acquired from the companies participating in this study. The severity ratings of these companies were averaged to create a scale for consequence. Using average values in such a fashion, while uncommon, can be considered good practice (Cunha, 2016). Consequences are rated based on impact to health and safety, environment, company's reputation, and finance using a five-point scale.

2.2.2. Prepare Data for Machine Learning Classification

The typical classification-type problem is numerical in nature. When a dataset is expressed graphically, classification boundaries can be drawn to separate different numerical values. The content of an incident report consists mostly of text, which makes working with incident reports a challenge. To overcome this challenge, the Python *scikit-learn* library is used to transform the incident reports into a numerical form (Imani et al., 2018). The scikit-learn library includes a feature called the *TfidfVectorizer* to convert incident reports to numerical values that can then be used for classification (Garreta et al., 2017). *Tf-idf* stands for term frequency times inverse document frequency. Term frequency refers to the number of times a term appears in a document – in this case, the number of times a word appears in an incident report. The term frequency can be scaled (typically logarithmically) for document length. The inverse document frequency refers to a weight that is applied to give more value to words that are rare across multiple documents and to reduce the value of words that are common. Together, term frequency times inverse document frequency applies a high weight to a word (term) that appears multiple times in the same document but is rare in the collection of documents. In summary, this method builds a dictionary using the terms found in the documents, counts the occurrences of each term, and applies weights based on the occurrences. The final result is the transformation of text to a numerical vector.

In order to train a machine learning algorithm to rank and categorize risks, many incident reports must be classified to set a guideline for the program (Raschka & Mirjalili, 2017). An accepted method for accomplishing this is to separate the data into a training set and a test set, where the entirety of this data must be classified manually. Applying the train/test split method to this research, the 15,000 incident reports being analyzed will be divided into a training set and a test set: 10,500 incident reports (70% of the randomly selected data) are used as training data while the remaining 4,500 incident reports (30% of the randomly selected data) are used as test data. As its name implies, the training data are used to train the program in classifying incidents and the test data can then be used to judge the accuracy of different classifiers. The "predicted" values of the test data are then compared with the manually classified "true" values to determine the accuracy of the model. The feature used to split the data into training and test data is called *train_test_split* and can be found as part of the scikit-learn library.

2.2.3. Use Classifiers from scikit-learn Library to Classify Data

The scikit-learn library also has many built in classifiers that can be used. For this research, every classifier compatible with the data was used to identify the classifier with the highest accuracy. The vectors generated by the TfidfVectorizer are considered sparse matrices. This means that most numbers in each vector are 0. This sparseness can result in incompatibility with some classifiers. Given this, we use several classifiers: *Adaboost* classifier, *decision tree* classifier, *k-nearest* *neighbors*, *logistic regression*, *multi-layer perceptron* classifier, *multinomial Naïve Bayes* classifier, *random forest* classifier, and *support vector machine* classifier (including *linear support vector* classifier). Once the data from the vectorized incident reports are expressed graphically, these classifiers utilize different approaches to generate decision boundaries that can be used to categorize the data.

2.2.4. Calculate Metrics for Each Classifier

Metrics are calculated in conjunction with train/test split to deliver scores and performance metrics, also calculated using the Python library, scikit-learn (Garreta et al., 2017; Pedregosa et al., 2011). For this study, the metrics used are *confusion matrices*, *classification reports*, and *accuracy scores* to determine a model's feasibility. Once a model is fitted to a training data set, these tools can be used to analyze the model's accuracy on the corresponding test data set. Confusion matrices are used to evaluate the accuracy of a classification. A confusion matrix requires the true classification, the predicted classification, and the labels (optional). If labels are not provided, the confusion matrix will arrange all labels found in the training and test data sets alphabetically. The output is a matrix, $C_{i,j}$ where i is the true number of observations in a group and j is the predicted number. From this matrix, it is easy to count the number of true negatives, false negatives, true positives, and false positives. A classification report delivers a text report of precision, recall, F1-score (or F-measure), and support for each label found in the data. The input variables are the true classification, predicted classification, and labels (optional). Precision is the ratio of the number of true positives to the sum of true positives and false positives. Recall is the ratio of the number of true positives to the sum of true positives and false negatives. The F1-measure is the mean of precision and recall, and support is simply the number of true occurrences.

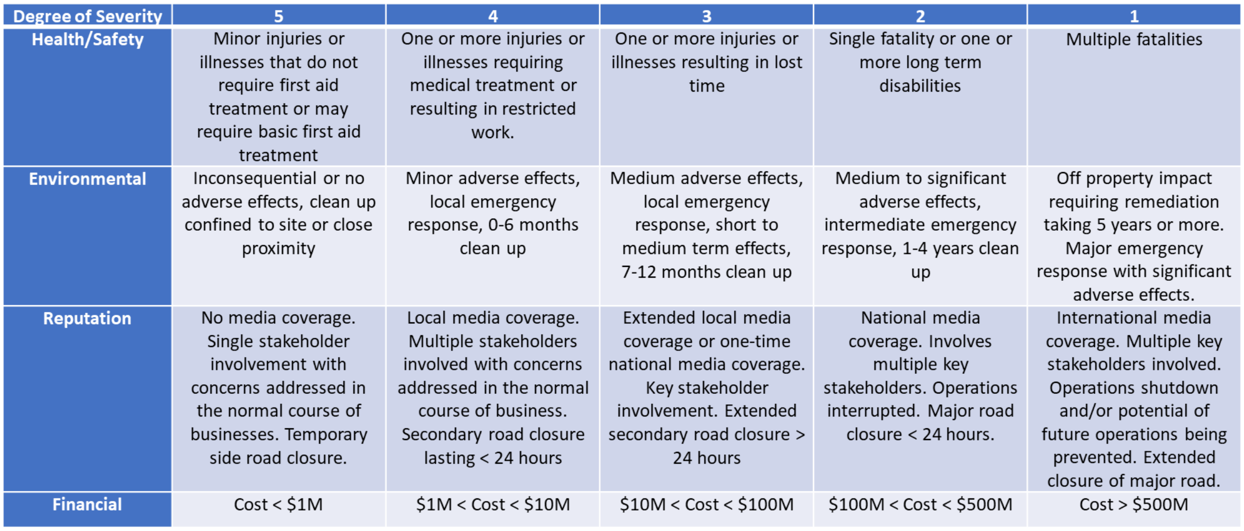
Precision and recall will have values between 0 (poor prediction) and 1 (good prediction). Accuracy score computes the percentage of predicted labels that exactly match the corresponding true labels. The input parameters are the true classification and the predicted classification labels, and the output is a single accuracy value as a percentage (100% = perfect accuracy).

2.2.5. Deliver Outputs

The typical process of determining a scale for likelihood is by allowing several experts to determine the probability of an incident occurring within a certain time period. For example, a low likelihood risk would occur once every 10,000 years while a high likelihood risk would occur once per year (Basu, 2017; Calixto, 2016). The incident databases supplied by companies contain many incident reports and the incident date and time, which allows us to calculate frequency of each incident type within a certain time period, and eliminates the need for a human to predict the likelihood. Applying this to a sample incident where a feed stream is taken offline due to a leak, one could tally the total number of incidents resulting in loss of production due to leaks over a period of time, and either interpolate or extrapolate information to determine likelihood. These likelihoods also allow comparisons such as the ratio of the total number of leaks in proportion to the total number of incidents reported, or the number of leaks resulting in feed streams being taken offline in proportion to the total amount of leaks. Such calculations are useful in providing strategical insight when planning preventative measures.

2.3. RESULTS AND DISCUSSION

To rate the severity of an incident, the values of the risk matrices from several collaborators were averaged to develop a 5-point scale (see Figure 2-4). This scale is used to manually classify both the actual and potential risk associated with incidents. And the supervised machine learning algorithm aims to predict consequence based on the same scale.



Degree of Severity	5	4	3	2	1
Health/Safety	Minor injuries or illnesses that do not require first aid treatment or may require basic first aid treatment	One or more injuries or illnesses requiring medical treatment or resulting in restricted work.	One or more injuries or illnesses resulting in lost time	Single fatality or one or more long term disabilities	Multiple fatalities
Environmental	Inconsequential or no adverse effects, clean up confined to site or close proximity	Minor adverse effects, local emergency response, 0-6 months clean up	Medium adverse effects, local emergency response, short to medium term effects, 7-12 months clean up	Medium to significant adverse effects, intermediate emergency response, 1-4 years clean up	Off property impact requiring remediation taking 5 years or more. Major emergency response with significant adverse effects.
Reputation	No media coverage. Single stakeholder involvement with concerns addressed in the normal course of businesses. Temporary side road closure.	Local media coverage. Multiple stakeholders involved with concerns addressed in the normal course of business. Secondary road closure lasting < 24 hours	Extended local media coverage or one-time national media coverage. Key stakeholder involvement. Extended secondary road closure > 24 hours	National media coverage. Involves multiple key stakeholders. Operations interrupted. Major road closure < 24 hours.	International media coverage. Multiple key stakeholders involved. Operations shutdown and/or potential of future operations being prevented. Extended closure of major road.
Financial	Cost < $1M	$1M < Cost < $10M	$10M < Cost < $100M	$100M < Cost < $500M	Cost > $500M

Figure 2-4. Consequence scale for incident reports using average values from multiple companies.

By using the methods discussed, we created a program which consistently classifies and analyzes incidents. Table 2-1 depicts the accuracies of the classifiers used in this study when applying primary labels to incident reports. The least accurate classifier was the basic Support Vector Classifier (SVC). Support Vector Machines (SVMs) were originally designed for "one-against-one" approaches or binary classification (Guenther & Schonlau, 2016). The SVM is a learning algorithm that maximizes the margin between classification boundaries. A hyperplane is created for each class label using support vectors – the training samples closest to the hyperplane – and the decision boundary attempts to maximize the distance between hyperplanes. The basic SVC can be seen to have the lowest accuracy of the selected classifiers at 56.98%. Unlike the SVC, the Linear Support Vector Classifier (Linear SVC) is a type of SVM that adopts the "one-vs-rest" or "one-vs-all" approach for classification (Milgram et al., 2006). It compares each class against every other class when drawing decision boundaries which leads to very accurate classification when dealing with data that is linearly separable. Based on this data, the most accurate classifier was the Linear SVC with an accuracy of 88.48% when predicting primary labels. For the remainder of this section, the focus will be on the results obtained using the Linear SVC as it was consistently the most accurate classifier for analyzing incident reports.

Table 2-1. Classification accuracy of primary labels for different classifiers

Classification Method	**Accuracy**
Support Vector Classifier (SVC)	56.98%
Adaboost	63.21%
Multinomial Naïve Bayes	66.76%
k-Nearest Neighbors	73.56%
Random Forest	75.80%
Decision Tree	75.95%
Logistic Regression	84.37%
MLP Classifier (Neural Network)	85.50%
Linear SVC	88.48%

Table 2-2 displays the confusion matrix associated to primary label classification for the Linear SVC. The confusion matrix shows the actual (manually classified) labels for the incident report on the y-axis and the predicted (machine learning classified) labels on the x-axis. This metric displays where the classifier is most accurate, and where it has misclassified the labels. The main diagonal of this matrix shows the number of actual labels that have been correctly identified by the classifier.

Table 2-2. Confusion matrix for Linear SVC used in primary label classification

		Predicted						
		Comm.	Health/S	Leak/Spill	Misc.	Operation	Uncat.	Vehicle
Actual	Comm.	62	2	0	1	9	0	0
	Health/S	0	709	4	0	110	0	20
	Leak/Spill	0	2	241	0	48	0	4
	Misc.	0	1	1	43	10	1	0
	Oper.	2	68	14	5	2571	4	51
	Uncat.	0	16	0	4	38	43	2
	Vehicle	0	31	7	2	91	1	547

Table 2-3 depicts the classification report for the Linear SVC when analyzing primary labels. These metrics display how the model classified the data and the accuracy to which the model is predicting each class label. As a general statement, it can be seen that as support increases, the accuracy of the supervised machine learning algorithm also increases. The algorithm was able to score the highest F1-score (0.92) when predicting the Operation label due to the massive amount of data available. Recall that the F1-score is the average of precision and recall and that the closer the values of precision, recall, and F1-score are to 1, the more accurate the class prediction. Support is the total number of actual labels assigned to each primary label.

Table 2-3. Classification report for Linear SVC used in primary label classification.

Primary Label	Precision	Recall	F1-score	Support
Comm.	0.97	0.84	0.90	74
Health/S	0.86	0.84	0.85	843
Leak/Spill	0.90	0.82	0.86	295
Misc.	0.78	0.77	0.77	56
Oper.	0.89	0.95	0.92	2715
Uncat.	0.88	0.42	0.57	103
Vehicle	0.88	0.81	0.84	679

The accuracies of the Linear SVC when predicting the consequence labels of Health and Safety, Environment, and Finance was calculated and reported in Table 2-4. It was difficult to assign a label for damage to a company's reputation due to the large number of incidents and the difficulty to estimate the severity of the consequence. In order to train a machine learning algorithm to assign a label for damage to a company's reputation, it would require attempting to find public records of every incident. Due to the intrinsic nature of high consequence incidents, it is assumed that these will garner media attention along with other consequences (i.e. health and safety, environment and/or finance). Thus, a consequence label of "Reputation" was attached to any incident with an actual risk score of 1 or 2.

Table 2-4. Classification accuracy of Health & Safety (H/S), Environment (E), and Finance (F) labels for different classifiers

Classification Method	Accuracy (H/S)	Accuracy (E)	Accuracy (F)
Adaboost	79.91%	93.87%	77.70%
Decision Tree	78.19%	93.14%	73.04%
Linear SVC	85.29%	94.85%	81.37%
k-Nearest Neighbors	79.66%	92.16%	78.19%
Logistic Regression	77.45%	89.22%	79.16%
MLP Classifier (Neural Network)	83.09%	90.44%	79.90%
Multinomial Naïve Bayes	74.75%	88.97%	77.21%
Random Forest	80.64%	91.42%	76.72%
Support Vector Classifier (SVC)	70.83%	88.97%	71.81%

The classification accuracies for actual and potential risk scores can be found in Table 2-5. Actual risk scores refer to the severity of an incident that has occurred while potential risk scores refer to near misses and possible consequences of incidents that could have been more severe given alternate circumstances. In most cases, for incidents that occurred, actual and potential risk scores are equal.

Table 2-5. Classification accuracy of Actual Risk and Potential Risk Scores for different classifiers.

Classification Method	Accuracy (Actual Risk Score)	Accuracy (Potential Risk Score)
Adaboost	82.11%	46.32%
Decision Tree	72.79%	53.68%
Linear SVC	82.35%	67.65%
k-Nearest Neighbors	76.72%	57.35%
Logistic Regression	78.43%	66.18%
MLP Classifier (Neural Network)	80.88%	64.46%
Multinomial Naïve Bayes	78.43%	62.01%
Random Forest	78.92%	59.56%
Support Vector Classifier (SVC)	78.43%	45.10%

Along with classification, another goal of this study was trend analysis of incidents. Figure 2-5 illustrations how descriptive labels can be used to analyze trends in incidents. Incidents are categorized based on their primary labels and counted based on month. The lowest number of incidents have occurred in January. This can be an example of how trend analysis can be used for risk analysis – it might be worthwhile to pursue the reason behind such a low incident count in January while the remaining months have roughly the same total amount of incidents. Possible reasons for this low incident count could be employees taking time off, an error when reporting, or very strict operating procedures for the first month of the year. The figure also shows that the month of March has the most incidents and that operational-type incidents are the most common, accounting for 56.98% of total incidents. Primary labels can also be analyzed in greater detail.

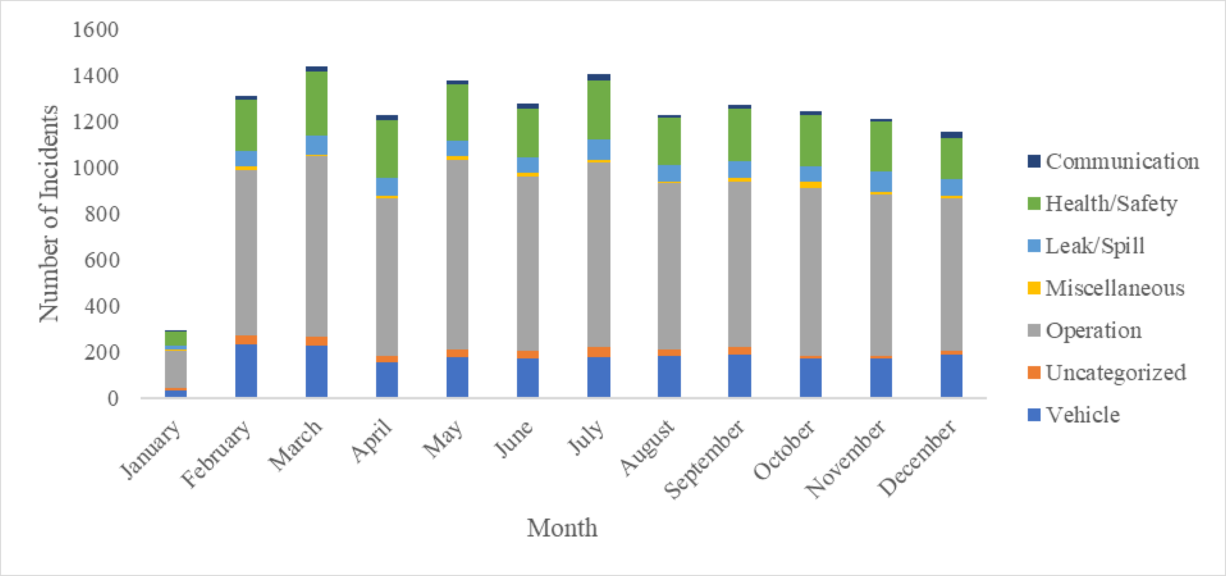


Figure 2-5. Number of incidents per month for a random sample of 15,000 incidents using primary classification labels.

Figure 2-6 demonstrations the subcategories of operation-type incidents per month. Operation-type incidents are subcategorized as follows: equipment, incorrect operations, nature, and uncategorized. Most operation-type incidents occur because of equipment. This information can be used to prioritize risks and avoid loss by designing appropriate prevention and mitigation strategies.

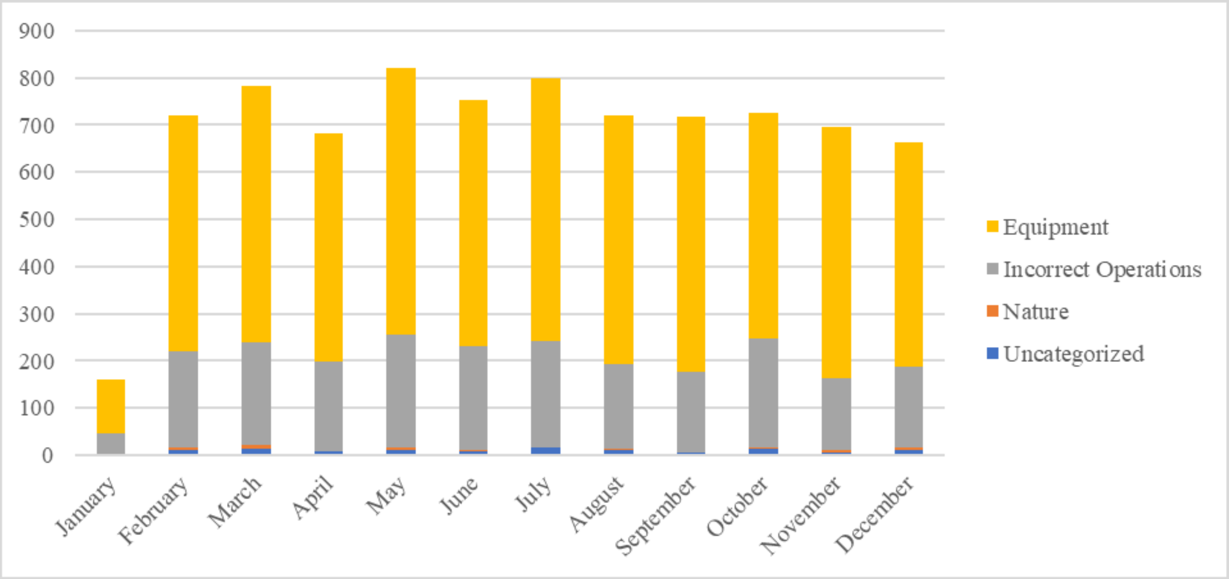


Figure 2-6. Number of operational-type incidents per month for a random sample of 15,000 incident reports using secondary classification labels.

Additional information that can be gleaned from Figures 2-5 and 2-6 is the number of incidents pertaining to process safety. Most incidents classified with the primary labels of "Operation" and "Leak/Spill" fall under process safety, accounting for approximately 9,450 of 15,000 incident reports, or 63% of the total number of incidents. This demonstrates the need to prioritize risk reduction in process safety. There are many ways this can be done. One such approach would be to isolate process safety incidents and identify safety indicators that can be further analyzed and acted upon to reduce the risk involved with such incidents (Swuste et al., 2016). This could also be used to identify correlations between different incidents that might otherwise remain unnoticed. Furthermore, with the cooperation of the companies involved in this research, it would also be possible to gather detailed information about incidents that are not included in typical incident reports. These details could be used to support prevention and mitigation strategies to target specific incident types that are often overlooked – many incidents that are high probability/low consequence are often ignored due to their “low risk” nature, when in fact, these traits might lead to severe loss over an extended timespan (Greenwell et al., 2003; Leistad & Bradley, 2009).

Incident report data are also useful in process safety education by providing a comprehensive understanding of latent root causes (Mkpat et al., 2018). Achieving excellence in process safety leads to fewer incidents, mitigating the consequences of incidents that occur, and providing optimal emergency response (Halim & Mannan, 2018). Education can be provided within a company, through educational institutions, and can also include government agencies and authorities. Such collaborations are already being used to increase worker awareness toward common hazards, and in the future, should be used to allow operators to identify precursory conditions of process safety incidents such as equipment failure (Halim & Mannan, 2018; Rowe & Francois, 2016).

The metrics calculated using Linear SVC for assigning consequence type and risk scores can be found in Appendix A. The assigned labels can be used to determine the type and risk level of incident reports. While the machine learning algorithm can operate with training and test data, it can also be used to make predictions about unclassified incident reports. The caveat to using machine learning to predict unclassified data is the probability of misclassification. With ~80% accuracy for the Linear SVC, approximately one out of five incidents will be misclassified.

2.4. CONCLUSIONS

By working with such a large, combined incident database, the goal of this research was to increase awareness of process safety incidents that occur in the oilsands sector and, potentially, the oil and gas industry more broadly. We used a supervised machine learning model to analyze the incident reports in order to simplify the reporting process, increase accuracy, and to reduce human bias.

Our analysis included predicting labels for incident type, consequence type, actual risk score, and potential risk score. While experimenting with different classifiers, the Linear SVC was found to have the greatest accuracy. Further, a supervised machine learning model implemented with the Linear SVC can analyze tens of thousands of incident reports in minutes, allowing for companies to develop prevention and mitigation strategies for incidents. Companies can target the most common and costly incidents, as well as the highest consequence incidents, in order to prevent or reduce the likelihood of such incidents from recurring. This study found that operation-type incidents were the most common from one sample of incident data, with most operational incidents being related to process safety. In this case, it is advisable to understand why such incidents are so common and to reduce both the consequences and the likelihood of these incidents. Properly managing these process safety issues would be beneficial to a company, especially when the areas for improvement are clearly delineated by quantitative studies.

The supervised machine learning model has much potential for future use. A supervised machine learning algorithm can be easily modified to accept new training/test data to predict class labels for a specific company based on its existing systems. This can allow a company to focus its efforts on preventing incidents that are causing the company losses, in addition to using the much larger aggregated database to properly increase hazard awareness for its workers. There is an opportunity to create a practical product for industry by using machine learning to improve trend analysis, design prevention and mitigation strategies, and to identify leading indicators.

2.5. REFERENCES

A Guide to the Project Management Body of Knowledge [PMBOK] (Sixth Edit). (2017). Newtown Square, PA, USA: Project Management Institute, Inc.

Animah, I., & Shafiee, M. (2019). Application of risk analysis in the liquefied natural gas (LNG) sector: An overview. *Journal of Loss Prevention in the Process Industries*, *63*(October 2019), 103980. https://doi.org/10.1016/j.jlp.2019.103980

Basu, S. (2017). Basics of Hazard, Risk Ranking, and Safety Systems. In *Plant Hazard Analysis and Safety Instrumentation Systems*. https://doi.org/10.1016/b978-0-12-803763-8.00001-7

Bjerga, T., & Aven, T. (2015). Adaptive risk management using new risk perspectives - An example from the oil and gas industry. *Reliability Engineering and System Safety*, *134*, 75–82. https://doi.org/10.1016/j.ress.2014.10.013

Calixto, E. (2016). Reliability and Safety Processes. In *Gas and Oil Reliability Engineering*. https://doi.org/10.1016/b978-0-12-805427-7.00006-3

Cass, S. (2019). The Top Programming Languages 2019. Retrieved from IEEE Spectrum website: https://spectrum.ieee.org/computing/software/the-top-programming-languages-2019

Cunha, S. B. da. (2016). A review of quantitative risk assessment of onshore pipelines. *Journal of Loss Prevention in the Process Industries*, *44*, 282–298. https://doi.org/10.1016/j.jlp.2016.09.016

Duijm, N. J. (2015). Recommendations on the use and design of risk matrices. *Safety Science*, *76*, 21–31. https://doi.org/10.1016/j.ssci.2015.02.014

Garreta, R., Hauck, T., & Hackeling, G. (2017). *Scikit-learn: machine learning simplified*. Birmingham, UK: Packt Publishing.

Greenwell, W. S., Knight, J. C., & Strunk, E. A. (2003). *Risk-based Classification of Incidents*. 39–50.

Guenther, N., & Schonlau, M. (2016). Support vector machines. *The Stata Journal*, *16*(4), 917–937. https://doi.org/10.1177/1536867X1601600407

Gul, M., & Guneri, A. F. (2016). A fuzzy multi criteria risk assessment based on decision matrix technique: A case study for aluminum industry. *Journal of Loss Prevention in the Process Industries*, *40*, 89–100. https://doi.org/10.1016/j.jlp.2015.11.023

Halim, S. Z., & Mannan, M. S. (2018). A journey to excellence in process safety management.

Journal of Loss Prevention in the Process Industries, *55*(April), 71–79. https://doi.org/10.1016/j.jlp.2018.06.002

Imani, A., Forman, J. E., & Amir, W. (2018). *A Clustering Analysis of Codes of Conduct and Ethics in the Practice of Chemistry*.

International Organization for Standardization [ISO]. (2018). *Risk Management - Guidelines* (ISO 31000:2018E)

Kletz, T. A. (2005). Looking beyond ALARP - Overcoming its limitations. *Institution of Chemical Engineers Symposium Series*, (150), 69–76. https://doi.org/10.1205/psep.04227

Kotsiantis, S. B. (2007). *Supervised machine learning: a review of classification techniques*. https://doi.org/10.1115/1.1559160

Landell, H. (2016). *The Risk Matrix as a tool for risk analysis*.

Leistad, G. H., & Bradley, A. R. (2009). Is the focus too low on issues that have a potential to lead to a major incident? *Society of Petroleum Engineers - Offshore Europe Oil and Gas Conference and Exhibition 2009, OE 2009*, *1*, 467–472. https://doi.org/10.2118/123861-ms

Markowski, A. S., & Mannan, M. S. (2008). Fuzzy risk matrix. *Journal of Hazardous Materials*, *159*(1), 152–157. https://doi.org/10.1016/j.jhazmat.2008.03.055

Mkpat, E., Reniers, G., & Cozzani, V. (2018). Process safety education: A literature review. *Journal of Loss Prevention in the Process Industries*, *54*(October 2017), 18–27. https://doi.org/10.1016/j.jlp.2018.02.003

Muhlbauer, W. K. (2004). *Pipeline Risk Management Manual - Ideas, Techniques, and Resources* (Third Edit). 200 Wheeler Road, Burlington, MA, USA: Gulf Professional Publishing (an imprint of Elsevier).

Ness, A. (2015). Lessons learned from recent process safety incidents. *Chemical Engineering Progress*, *111*(3), 23–29.

Occupational Health and Safety Act [OHS]. , Pub. L. No. Chapter O-2.1 (2018).

Occupational Safety and Health Administration [OSHA]. , U.S. Department of Labor § (2000).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. Retrieved from https://scikit-learn.org/stable/user_guide.html

Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning* (2nd Editio). Birmingham, UK: Packt Publishing.

Rowe, S., & Francois, J. M. (2016). Process safety data – The cornerstone of PSM and often it's undermining. *Journal of Loss Prevention in the Process Industries*, *43*, 736–740. https://doi.org/10.1016/j.jlp.2016.06.002

Swuste, P., Theunissen, J., Schmitz, P., Reniers, G., & Blokland, P. (2016). Process safety indicators, a review of literature. *Journal of Loss Prevention in the Process Industries*, *40*, 162–173. https://doi.org/10.1016/j.jlp.2015.12.020

Taleb, N. N. (2007). *The black swan: the impact of the highly improbable*. USA: Random House.

Thomas, P., Bratvold, R. B., & Bickel, J. E. (2013). The Risk of Using Risk Matrices Decision Analysis View project A Generalized Sampling Approach for Multilinear Utility Functions Given Partial Preference Information View project The Risk of Using Risk Matrices. *Article in SPE Economics and Management*, (April 2015). https://doi.org/10.2118/166269-MS

3. USING MACHINE LEARNING AND KEYWORD ANALYSIS TO ANALYZE INCIDENTS AND REDUCE RISK IN OIL SANDS OPERATIONS

A version of this chapter has been submitted to the Journal of Safety Science as Kurian, D., Ma, Y., Lefsrud, L., and Sattari, F. "Using Machine Learning and Keyword Analysis to Analyze Incidents and Reduce Risk in Oil Sands Operations."

3.1. INTRODUCTION

Risk, as defined by the International Organization for Standardization [ISO] (2018), is uncertainty of all types and sizes, both internal and external, which can affect an organization as it attempts to achieve its goals. According to the Project Management Institute [PMI] (2004), risk management processes need to be tailored specifically to each project. As such, organizations work to manage risk by identifying, analyzing, and evaluating risk, and then taking appropriate courses of action – planning responses, implementing changes, and continual monitoring. Generally, the process of identifying hazards and estimating risk is considered qualitative risk management and should be conducted first to identify and prioritize risks requiring detailed quantitative analysis.

An effective method of identifying hazards and estimating risk is to analyze historical data (Patriarca et al., 2018). In the oil and gas industry, such historical data can be found in incident reports (Nordlöf, 2015; Laberge et al., 2014). Incident reports contain many instances of past shortcomings or failures, which can be used as learning experiences to prevent similar incidents from reoccurring. Companies can use this knowledge to train their workers, and workers can study specific cases to identify hazards and to learn appropriate responses and countermeasures.

In Alberta, incident reports are required to contain the location of the incident, time and date, name of the employer involved, contact information of the site contact, and a general description of the incident (Government of Alberta, 2019). To build rapport, some companies add further details to incident investigation. Some measures might include root cause analysis, hazard and operability (HAZOP) studies, and basic risk ranking procedures such as risk matrices (Nordlöf, 2015; Pasquini et al., 2011). The risk matrix is a tool used to provide an estimation of the frequency and possible consequences of the incident (on two axes), identify the relative severity of the risk (mapping zones of low, medium, high, etc.), and to determine what course of action must be taken to prevent or mitigate future incidents of that type (Albery et al., 2016).

A risk matrix is easy to implement, maintain, understand, and explain – due to these benefits, the tool is commonly used by companies to assess risk (Thomas et al., 2014); however, there are many drawbacks to using a risk matrix for risk analysis, including human bias and inconsistencies when reporting (Goerlandt and Reniers, 2016; Duijm, 2015). To strengthen this existing system, we applied a supervised machine learning approach to accurately analyze and evaluate risk in incident reports in previous research. Artificial Intelligence and Machine Learning (AI/ML) hold great promise for enhancing process safety management by visualizing data and recognizing patterns across big datasets in real-time, determining the most effective leading indicators, especially how they may relate to low-frequency high-consequence events, and prioritizing improvements to safety processes. AI/ML have already been applied to established process safety tools like bowties (Khakzah, Khan & Amyotte, 2013), process hazard assessments (PHAs) and layers of protection analysis (LOPAs) (Xu et al., 2018), and hazard and operability studies (HAZOPs) (Zhao et al., 2009).

Yet, for many reasons, implementing AI/ML into companies' legacy process safety management systems has been slow. These databases create an overwhelming amount of (often dirty) data that are rarely analyzed in detail and effectively leveraged. There are several reasons for this. First, operators tend to only analyze incidents with severe consequences to prevent recurrence, while minor incidents are only stored without any further evaluation. Yet, high-frequency and low-consequence incidents often display leading indicators that are overlooked but would be useful to predict high-consequence incidents (Aven, 2011; Steen & Aven, 2011). Second, while detailed data is used to create HAZOPs, PHAs, LOPAs, and bowties, there are issues with the data itself. This data is often ‘dirty’ or incomplete, fragmented across data sources, proprietary with little incentive for sharing with others, or has uncertain or contested ownership (Dong et al., 2017; Ransbotham et al., 2017). Third, leading operators have invested in developing internal AI/ML skills through training or hiring, but many operators outsource their AI/ML services. Yet, operators are surprised by AI/ML researchers’ and suppliers’ requirement for large datasets to allow their algorithms to learn, which results in operators perceiving AI/ML as a high-effort, low-payoff venture (Ransbotham et al., 2017).

To address these barriers, researchers and consultants often aggregate data across operators to create more complete, ‘clean’, and larger datasets to enhance algorithm training and ‘detectability’ of leading indicators (Kurian et al., 2020). Yet, cross-organizational aggregation and collaboration introduces other barriers such as: differences in representativeness, context, and content that makes the data incommensurate (Zuboff, 2015; Kellogg, Valentine & Christin, 2020) and model overfitting that can lead to inaccurate predictions when the model is used on different or more general data (Bengio, Goodfellow & Courville, 2017).

We have recognized and begun to address these barriers in our previous research; a total of 15,000 incidents were manually classified: descriptive labels, actual and potential risk scores, and consequence labels (environment, finance, health/safety, and reputation) were applied to each incident. The incident reports were then divided into training and test data, and the machine learning algorithm used the training data to predict labels for the test data. The result of this research was a machine learning algorithm that could apply labels to incidents with 75-90% accuracy (depending on the label), and the outputs were used to develop risk matrices and to analyze trends in incidents.

The machine learning used in previous research was an attempt to remove human bias, and this method allowed for consistent reporting of incidents. However, many different variables (mentioned earlier) had to be manually analyzed and it was difficult to improve the accuracy beyond a certain level. Some incident reports lacked the detail required for classification and it was impossible to completely remove bias as using a supervised learning model implies manual training.

We continue to address these barriers with this research, by using machine learning to attach a basic label to describe an incident report. Furthermore, this research applies additional keyword analysis to increase the accuracy of machine learning classification. This research provides significant changes to the current system of incident reporting. By having the user select options from a standardized list that allow for detailed analysis of risk, the user is forced to accurately describe the risk involved in an incident. Additionally, due to the increased efficiency in reporting incidents, it becomes possible to provide practical outputs beyond typical risk evaluation: prevention and mitigation strategies, such as leading indicators to increase the awareness of hazards in the workplace. This information can be used to predict incidents and to train workers to prevent/mitigate risk from incidents that might occur in the future.

3.2. METHODOLOGY

The objectives of this research are to:

- strengthen the current incident reporting system by creating a customized library using artificial intelligence, machine learning, and statistics;
- support the design of more sensitive risk prevention and mitigation strategies, as well as leading factors; and
- enhance organizations learning from incidents and create opportunities to reduce losses.

Figure 3-1 shows an overview of the methodology used in this research. The process of reporting incidents is expanded to a three-step procedure with an intermediate step for user input. For the first step, a company is required to provide data pertaining to past incidents and safety requirements. The second step involves designing a customized library for analyzing future incidents that are reported. The final step provides a detailed analysis and suggestions that can be used to prevent incidents from occurring or to minimize the damage caused by such events.

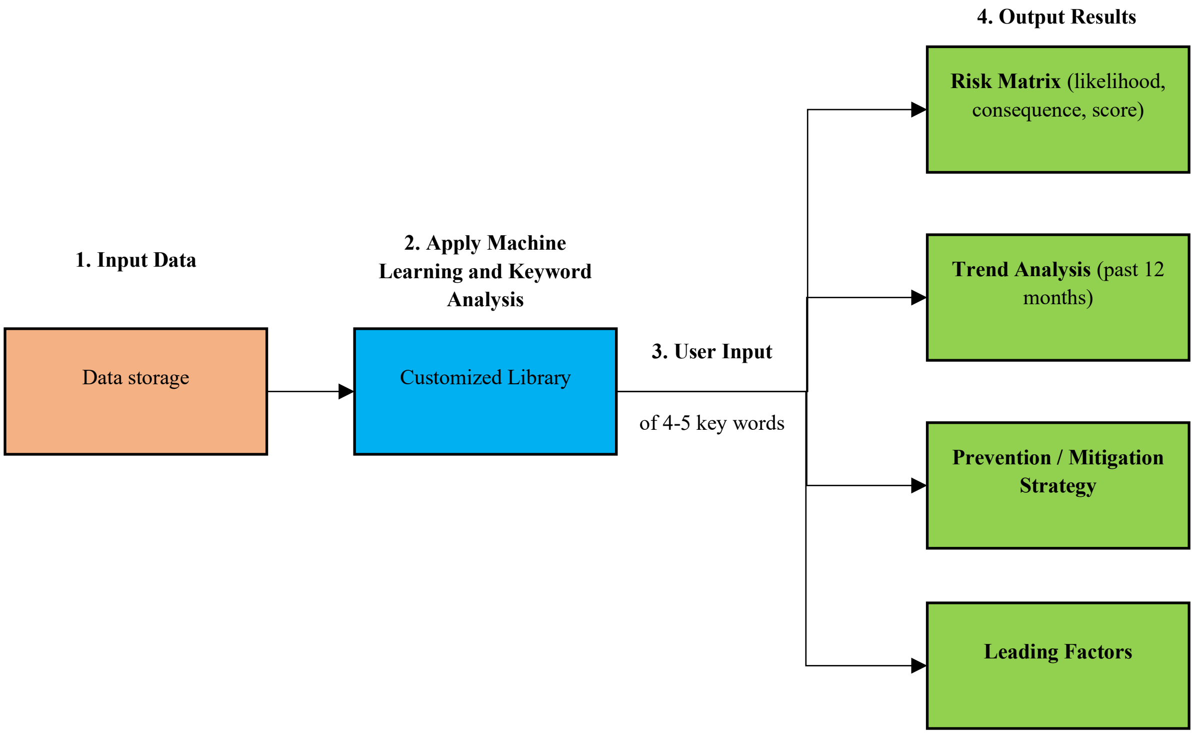
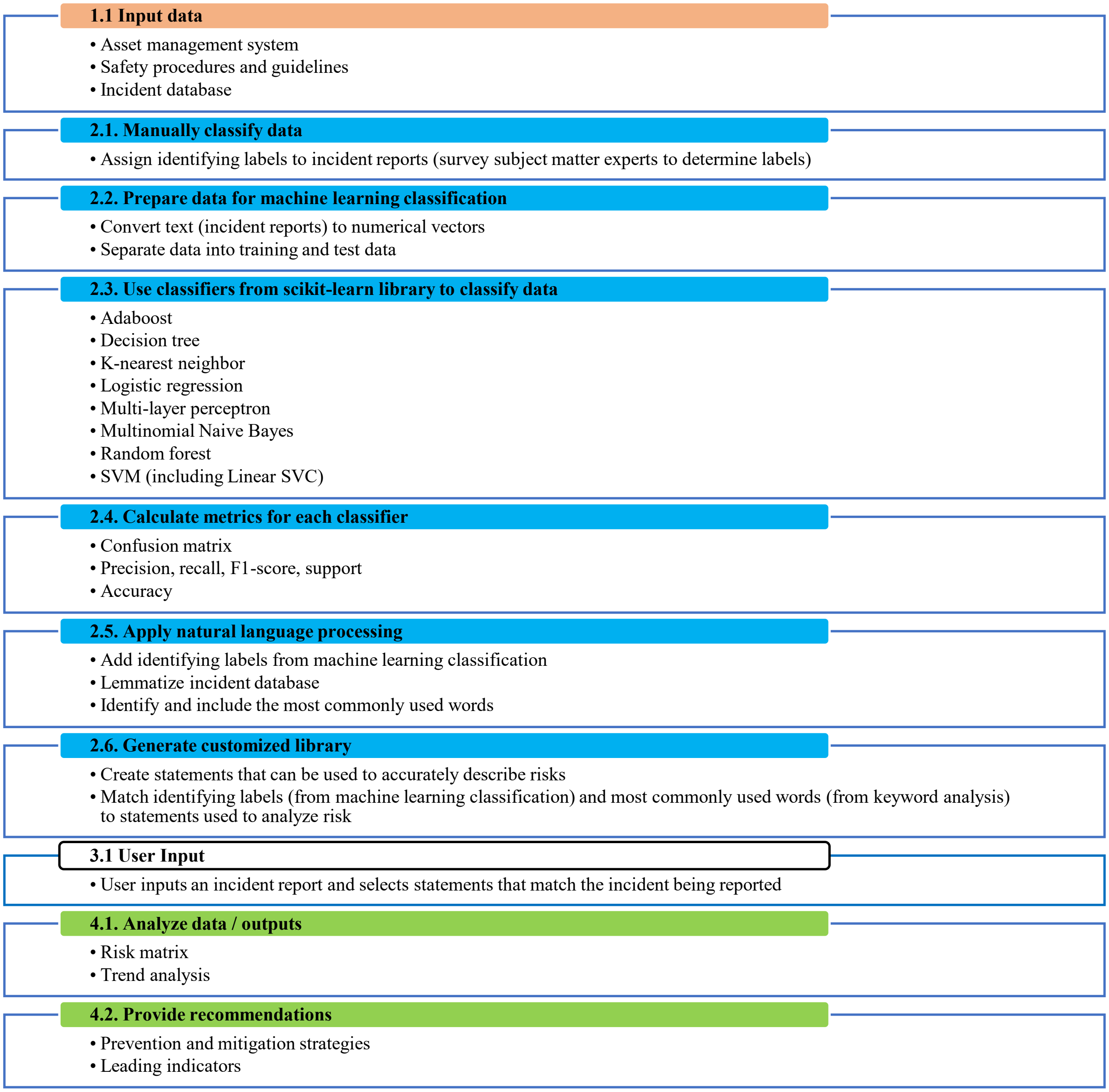

4. Output Results
Risk Matrix (likelihood, consequence, score)
1. Input Data
2. Apply Machine Learning and Keyword Analysis
Trend Analysis (past 12 months)
Data storage
Customized Library
3. User Input
of 4-5 key words
Prevention / Mitigation Strategy
Leading Factors

Figure 3-1. Overview of methodology

Figure 3-2 provides a more detailed description of the steps involved in designing the customized library and delivering outputs (with the corresponding steps coloured similarly). For this research, several collaborating companies provided access to their incident databases containing incident reports from 2013 to 2017, inclusive. The methods described in this research are applied to the data supplied from one of the participating companies including a total of approximately 15,000 incident reports. A customized library is generated from this data and output results are programmed given the input data.



1.1 Input data

- Asset management system
- Safety procedures and guidelines
- Incident database

2.1. Manually classify data

- Assign identifying labels to incident reports (survey subject matter experts to determine labels)

2.2. Prepare data for machine learning classification

- Convert text (incident reports) to numerical vectors
- Separate data into training and test data

2.3. Use classifiers from scikit-learn library to classify data

- Adaboost
- Decision tree
- K-nearest neighbor
- Logistic regression
- Multi-layer perceptron
- Multinomial Naive Bayes
- Random forest
- SVM (including Linear SVC)

2.4. Calculate metrics for each classifier

- Confusion matrix
- Precision, recall, F1-score, support
- Accuracy

2.5. Apply natural language processing

- Add identifying labels from machine learning classification
- Lemmatize incident database
- Identify and include the most commonly used words

2.6. Generate customized library

- Create statements that can be used to accurately describe risks
- Match identifying labels (from machine learning classification) and most commonly used words (from keyword analysis) to statements used to analyze risk

3.1 User Input

- User inputs an incident report and selects statements that match the incident being reported

4.1. Analyze data / outputs

- Risk matrix
- Trend analysis

4.2. Provide recommendations

- Prevention and mitigation strategies
- Leading indicators

Figure 3-2. Detailed description of methodology

3.2.1. Input Data

To design a customized library, there was a requirement for input data from the companies participating in the research. This data included an incident database containing incidents for the past five years, standard operating practices, safety procedures, and guidelines to develop proper responses to incidents and hazards, and asset management systems to better understand the different systems and equipment involved in incident reporting. Input data was stored securely and used to design a customized library of keywords for a company.

15,000 incident reports were selected from the provided incident reports, analyzed, and used to generate output results. These incident reports were used to train a machine learning algorithm to predict class labels for new incident reports that will be inputted. By using these class labels in conjunction with keyword analysis, it was possible to develop outputs for any incident report that shares similarities to other incidents in the incident database.

3.2.2. Apply Machine Learning and Keyword Analysis

Applying machine learning and keyword analysis to incidents reports is a multi-step process. For this research, a supervised machine learning algorithm was used to classify incident reports. Depending on the data being analyzed and the selected classifier, the total computational time of supervised machine learning algorithms can be very small compared to other approaches (Singh et al., 2017). Supervised machine learning operates by using predictor features to forecast class labels – it aims to categorize data by utilizing prior information (Kotsiantis, 2007). The first step to implementing supervised machine learning towards the classification of incident reports is to manually classify incident reports by labelling them with consistent identifiers (key descriptors, immediate and latent causes, contributing factors). By interviewing university professors and industry experts from participating companies, the following labels were selected to identify incidents: communication, health/safety, leak/spill, miscellaneous, operation, uncategorized, and vehicle. The label of "uncategorized" was assigned to incident reports that could not be classified.

Once the incident reports were manually classified, the data in the incident database was prepared for machine learning classification. The TfidfVectorizer feature was used from Python's scikit-learn library to transform each incident report into a numerical vector, and thus, the incident database is transformed into a matrix (Imani et al., 2018). Alternatively, the incident can be viewed as a dictionary with the individual incident reports being documents and the words found in the incident reports being terms. The occurrence of each term is counted, and weights are applied by comparing how often a term is found in a document versus the entire dictionary. The result is the transformation of text to a numerical vector. These manually classified incidents were then separated into training and test data sets, containing 70% and 30% of the data, respectively. The numerical vectors of the incident reports in the training set were expressed graphically, and a classifier was used to generate decision boundaries used to classify data. The numerical vectors representing the incident reports are considered sparse matrices – matrices in which most of the numbers are 0. The reason for this sparsity is because of the way that the dictionary was built using the TfidfVectorizer – every word (term) found in the incident database is added to the dictionary sequentially. For every word (term) found in an incident report (document), a count is applied to the position of the word in the incident database (dictionary). Subsequently, the terms in the dictionary that are not found in the document are assigned values of 0. Given the massive number of terms compiled in the dictionary, the vectors used to represent each incident report, and thus, the matrix used to represent the incident database, will be sparse. A number of classifiers from the scikit-learn library that are compatible with sparse matrices were used to classify the incident reports: Adaboost classifier, decision tree classifier, k-nearest neighbors, logistic regression, multi-layer perceptron classifier, multinomial Naïve Bayes classifier, random forest classifier, and support vector machine classifier (including linear support vector classifier). The supervised machine learning algorithm then attempted to identify features in the incident report that could be used to connect it to a given label, and metrics were calculated for different classifiers to identify the most suitable classifier for the data.

Previous research discovered that the most accurate classifier for categorizing incident reports from Alberta's oil and gas industry was the Linear Support Vector Classifier (Linear SVC), boasting accuracies close to 90% when predicting labels (Kurian et al., 2020). The metrics used were the confusion matrix, classification report, and accuracy score (Garreta et al., 2017). The confusion matrix was used to demonstrate how a classifier makes predictions for labels and requires the true and predicted classifications of the model. In a confusion matrix, the true label can be found on the y-axis and the predicted label on the x-axis. The classification report delivers precision, recall, F1-score, and support with inputs of the actual and predicted labels. These metrics can be described as follows:

$$Precision = \frac{true\ positives}{true\ positives + false\ positives}$$

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

$$F1 - score = \frac{precision + recall}{2}$$

Values for precision, recall, and F1-score will be between 0 and 1, where values closer to 1 represent a more robust model. Support is the count of true occurrences for each label. Finally, the accuracy score is the percentage of predicted labels that the model correctly identifies.

After determining accuracies from the machine learning classification, Natural Language Processing (NLP) was used to analyze keywords. NLP allows computers to interact with humans by processing and analyzing natural language data (Srinivasa-Desikan, 2018). Aside from the scikit-learn library which was used to convert incident reports into numerical vectors, there are two Python libraries that are commonly used for NLP: spaCy and Natural Language Toolkit (NLTK). The primary difference between these two libraries is that spaCy adopts an object-oriented approach while NLTK is used as a string processing library. Consequently, spaCy is more efficient when working with words, while NLTK performs better than spaCy when analyzing sentences (Malhotra, 2018). As such, spaCy was selected for keyword analysis in this research. SpaCy has many features that can be used to pre-process text data – it comes with tokenization and lemmatization features which were used to transform the words in the incident database to their canonical form (Srinivasa-Desikan, 2018). For instance, the words "run," "running," and "ran" would all be reverted to "run."

Keyword analysis was completed by lemmatizing all the words found in the incident database. A counter was then used to identify and tally the lemmatized words, and these words were then arranged from most frequent to least frequent. The keywords that could be used to classify incidents were then selected to include in the customized library (stop words, punctuation, names of individuals, etc. were removed).

The customized library was created with two variables: the identifying *labels* used to train the machine learning algorithm and the *keywords* identified using the spaCy library. The labels and keywords stored in the customized library were then matched to statements that could be used to analyze and evaluate risk. These statements were used to encompass varying levels of risk and restrict a user to select an option that could be used to accurately analyze the risk in an incident. The purpose of using both machine learning and a "manual" keyword approach was to increase accuracy and ensure that the generated statements could accurately describe any incident. To some extent, the keyword analysis was also used as a buffer to compensate for misclassification by the machine learning algorithm.

3.2.3. User Input

The labels and keywords found in the customized library were used to generate a list of statements. These statements were rule-based outputs developed in accordance to the inputted standard operating procedures, safety guidelines, and asset management systems provided by the company. When a user enters an incident into the system, there is a prompt to select statements applicable to the incident being reported. A list of statements is generated from which the user can select those relevant to the incident being reported. Parameters can be assigned to generate a specific number of statements and to restrict the maximum number of statements that can be selected. There was also an attempt to attach priority to the statements most likely to match the incident – the statements generated by the supervised machine learning algorithm appear first followed by the statements selected by the keyword analysis. In practice, the statements selected by the keyword analysis should have a wider range of selection and should include the statements selected by the machine learning algorithm. This is because machine learning predicts a single label to match the incident whereas the keyword analysis identifies every word that is common in the incident report and the customized library. In the case where both the machine learning algorithm and keyword analysis yields the same results, the duplicates are removed, and the statements selected by the machine learning algorithm retain priority in the listing. Finally, there is also a feedback loop that is designed in the user input stage. When a user selects statements to match the incident report, this information is recorded and used to improve machine learning accuracy for future incident report classification.

3.2.4. Output Results

When an incident report is inputted, four outputs are delivered (based on statements selected by the user): a risk matrix for the incident, trends of similar incidents, prevention and mitigation strategies to reduce the risk of the incident in the future, and leading indicators that can be identified by workers prior to the recurrence of a similar event.

A risk matrix is generated by calculating frequency and consequence (Ni et al., 2010). Frequency is a prediction of how likely it is for an incident to occur within a given time period. With access to a company's incident database, the actual count of incidents were used to calculate frequency as opposed to predicting the frequency of an incident. In Alberta's oil and gas industry, consequences can be categorized into four types: impact to worker health/safety, environmental damage, financial loss, and harm to a company's reputation (Muhlbauer, 2004). Based on the statement selected by a user, each incident is categorized into one or more of these consequence categories.

Another practical output that was delivered was trend reports. Trends were calculated by analyzing the statements selected by the user and the date of the incident. The total count of the selected statements was plotted by month to show incident trends and identify where and when improvements are needed and where safety measures are excelling.

Based on the inputted standard operating practices, safety, and asset management system, specific prevention and mitigation strategies were assigned to each of the statements that were selected. Additional statements were also programmed for specific groups of statements that were commonly selected together. This same process was applied to identify leading indicators for specific incidents. This type of output is based entirely on the input of an incident report, and provides actionable information to users as they enter incident reports and to companies as they seek to reduce risk in their work sites.

3.2.5 Summary of Methodology

To summarize the methodology, the first step is to input data, the second step is to process data, the third step is to input new incident reports, and the fourth step is to provide outputs. There is a feedback loop between steps 1 and 3 where new incident reports that are inputted will be analyzed and then added to the existing database.

Supervised machine learning was implemented in previous research to complete basic risk analysis and evaluation of incident reports in the form of risk matrix outputs, and this research was integrated into the current methods of analysis. For example, the trend reports currently generated are based on the outputs of machine learning from past research; however, the system is designed to create updated trend reports as new incident reports are inputted. As such, current analysis is based on the incident database that has already been provided by companies. In the future, as incident reports are added using the proposed methodology, outputs will become more specific to the newly inputted incident data as accuracy continues to improve.

3.3. RESULTS AND DISCUSSION

Identifying labels were used to manually classify 15,000 incident reports: communication, health/safety, leak/spill, miscellaneous, operation, uncategorized, and vehicle. As suggested by the previous study (Kurian et al., 2020), supervised machine learning was used with the Linear SVC classifier to predict labels for incidents since it provides the highest accuracy. Table 3-1 displays the confusion matrix for the Linear SVC classifier. The actual (manually classified) labels are shown on the y-axis while the predicted (machine learning classified) labels are shown on the x-axis. The main diagonal of this matrix demonstrates the number of true labels that the classifier accurately predicted.

Table 3-1. Confusion matrix for Linear SVC when predicting the identifying label.

		Predicted						
		Comm.	Health/S	Leak/Spill	Misc.	Operation	Uncat.	Vehicle
Actual	Comm.	62	2	0	1	9	0	0
	Health/S	0	709	4	0	110	0	20
	Leak/Spill	0	2	241	0	48	0	4
	Misc.	0	1	1	43	10	1	0
	Oper.	2	68	14	5	2571	4	51
	Uncat.	0	16	0	4	38	43	2

	Vehicle	0	31	7	2	91	1	547

Table 3-2 is the classification report for the Linear SVC when predicting the identifying labels for incident reports. From here, it can be seen how accurately each label is predicted by the supervised machine learning algorithm. The overall accuracy of the Linear SVC when predicting the identifying label is ~88.48%. F1-scores (average of precision and recall) closer to 1 signify better model accuracy while support is the number of true occurrences of each label. A low F1-score and support means the model requires more exposure to predictor features to become more accurate.

Table 3-2. Classification report for the Linear SVC when predicting the identifying label.

Identifying Label	Precision	Recall	F1-score	Support
Comm.	0.97	0.84	0.90	74
Health/S	0.86	0.84	0.85	843
Leak/Spill	0.90	0.82	0.86	295
Misc.	0.78	0.77	0.77	56
Oper.	0.89	0.95	0.92	2715
Uncat.	0.88	0.42	0.57	103
Vehicle	0.88	0.81	0.84	679

Prior research applied the supervised machine learning to the incident database classify incident reports. The supervised machine learning is now used to predict labels for incident reports that will be inputted in the future. Table 3-3 matches statements that can be used to accurately describe incidents to the identifying labels used in supervised machine learning classification. When the user inputs an incident report, the machine learning algorithm will predict a class label for the incident report. The statements found corresponding to the predicted label will then be made available for user selection. Note that some labels have only minor differences in syntax (e.g. 6 types of statements pertaining to equipment describing different types of risk and severity of consequences). These labels can play a strong role when determining outputs; further, selecting a specific statement from this list can help to distinguish a minor incident from a major incident.

Table 3-3. Statements generated for user selection based on labels selected by supervised machine learning algorithm.

Statements	Identifying Label
Equipment (Damage - cost <$1m)	Operation
Equipment (Failure - cost <$1m)	Operation
Equipment (General - cost <$1m)	Operation
Equipment (Damage - cost >$1m)	Operation
Equipment (Failure - cost >$1m)	Operation
Equipment (General - cost >$1m)	Operation
Fatality	Health/Safety, Vehicle
Fire (Damage)	Vehicle
Fire (Injury)	Health/Safety
Incorrect Operations	Communication, Operation
Injury	Health/Safety, Vehicle
Laceration/abrasion	Health/Safety

Leak/spill	Health/Safety, Leak/Spill
Minor Injury	Health/Safety, Vehicle
Miscellaneous	Miscellaneous
Miscommunication	Communication, Operation
Missing Equipment	Miscellaneous
Near Miss	Health/Safety, Vehicle
No Treatment Injury	Health/Safety, Vehicle
Property Damage	Miscellaneous
Quality Assurance	Communication
Severe Injury	Health/Safety
Slip/trip/fall	Health/Safety, Weather
Snow/ice	Weather
Sprain/strain	Health/Safety
Vehicle (heavy equipment)	Vehicle
Vehicle (light vehicle)	Vehicle
Vehicle collision (no injury)	Vehicle
Vehicle collision (with injury)	Vehicle
Weather	Weather
Wildlife	Miscellaneous

Table 3-4 shows how statements are matched to keywords. It is important to remember that the spaCy library lemmatizes the words in the incident reports. This means that keywords can be inputted in their canonical form without having to account for variations of a word (i.e. verb tense, singular vs plural, etc.). Here, the 15,000 incident reports were analyzed, and words found in the incident reports that could be used to classify incidents were matched with corresponding statements. One point to note is that abbreviations are also considered as keywords – the spaCy library ignores words that it does not recognize when lemmatizing the incident reports. Some common abbreviations found in the incident reports are: HT (haul truck), MOP (maximum allowable pressure), SOL (safe operating limit), QA (quality assurance), ROW (right of way), and STF (slip/trip/fall).

To summarize, both the *labels* used in supervised machine learning and the *keywords* found in the incident report are assigned to *statements*. When an incident report is inputted into the system, the machine learning algorithm predicts a label to describe the incident and the incident report is lemmatized for keyword analysis. A list of statements is then generated based on the predicted label and matching keywords. The user is required to select statements that accurately describe the incident.

Table 3-4. Statements generated for user selection based on keywords found in incident reports

Statements	Keywords
Equipment (Damage - cost <$1m)	damage, defective, equipment, exchanger, filter, hose, maintenance, not working, pump, seal, sump, valve, working
Equipment (Failure - cost <$1m)	defective, equipment, exchanger, failure, filter, hose, maintenance, not working, pump, seal, sump, valve, working

Equipment (General - cost <$1m)	defective, design, equipment, exchanger, filter, hose, maintenance, missing, not working, pump, seal, SOL, sump, trip, valve, venting, working
Equipment (Damage - cost >$1m)	damage, defective, equipment, exchanger, filter, hose, maintenance, not working, pump, seal, sump, valve, working
Equipment (Failure - cost >$1m)	defective, equipment, exchanger, failure, filter, hose, maintenance, not working, pump, seal, sump, valve, working
Equipment (General - cost >$1m)	defective, design, equipment, exchanger, filter, hose, maintenance, missing, not working, pump, seal, SOL, sump, trip, valve, venting, working
Fatality	fatality, fire, h2s, vehicle
Fire (Damage)	alarm, burn, burnt, fire, flame
Fire (Injury)	alarm, burn, burnt, fire, flame
Incorrect Operations	adequate, allowable, engineering, exceed, exceeded, improper, incorrect, incorrect, operations, knowledge, less, management, missing, missing, sign, missing, tag, MOP, performance, skill, SOL, unacceptable, unauthorized, verbal, wrong
Injury	abrasion, fall, finger, fire, h2s, illness, injure, injury, laceration, rest, slip, sprain, stf, strain, trip, vehicle
Laceration/abrasion	abrasion, bruise, cut, finger, laceration, paper, cut, papercut

Leak/spill	contaminate, drain, overflow, spill, leak, smell, seal
Major leak/spill	contaminate, drain, overflow, spill, leak, smell, seal
Minor Injury	abrasion, aid, fall, finger, fire, first, illness, injure, injury, laceration, slip, stf, treatment, trip, vehicle
Miscellaneous	missing, missing, equipment, theft
Miscommunication	communicate, communication, incorrect, management, miscommunicate, miscommunication, missing, missing, tags, operation, order, unacceptable, vendor, wrong, performance, less, adequate, verbal, skill
Missing Equipment	missing
Near Miss	miss, near, near, miss
No Treatment Injury	no, treatment, stf, treatment
Property Damage	drain, fire, leak, odor, odour, smell
Quality Assurance	assurance, document, documentation, incorrect, order, qa, quality, vendor, wrong
Severe Injury	fall, fire, h2s, illness, slip, sprain, stf, strain, trip, vehicle, rest, injury, injure, finger, disability
Slip/trip/fall	fall, fell, ice, injure, injury, oil, slip, snow, stf, trip, water
Snow/ice	ice, nature, poor, weather, snow, weather
Sprain/strain	back, finger, ice, injure, injury, lift, oil, slip, snow, treatment, water
Vehicle (heavy equipment)	accident, bulldozer, collision, dozer, haul, truck, ht, loader, loader, ROW, vehicle, zoom, boom, zoomboom

Vehicle (light vehicle)	accident, bus, car, collision, light, vehicle, lv, ROW, truck, vehicle
Vehicle collision (no injury)	accident, bulldozer, bus, collision, crane, dozer, excavator, fork, lift, forklift, haul, truck, ht, loader, truck, vehicle, zoom, boom, zoomboom
Vehicle collision (with injury)	accident, bulldozer, bus, collision, crane, dozer, excavator, fork, lift, forklift, haul, truck, ht, loader, truck, vehicle, zoom, boom, zoomboom
Weather	hail, ice, nature, poor, weather, rain, sleet, snow, weather, wind
Wildlife	animal, bird, fish, fox, wildlife, wolf

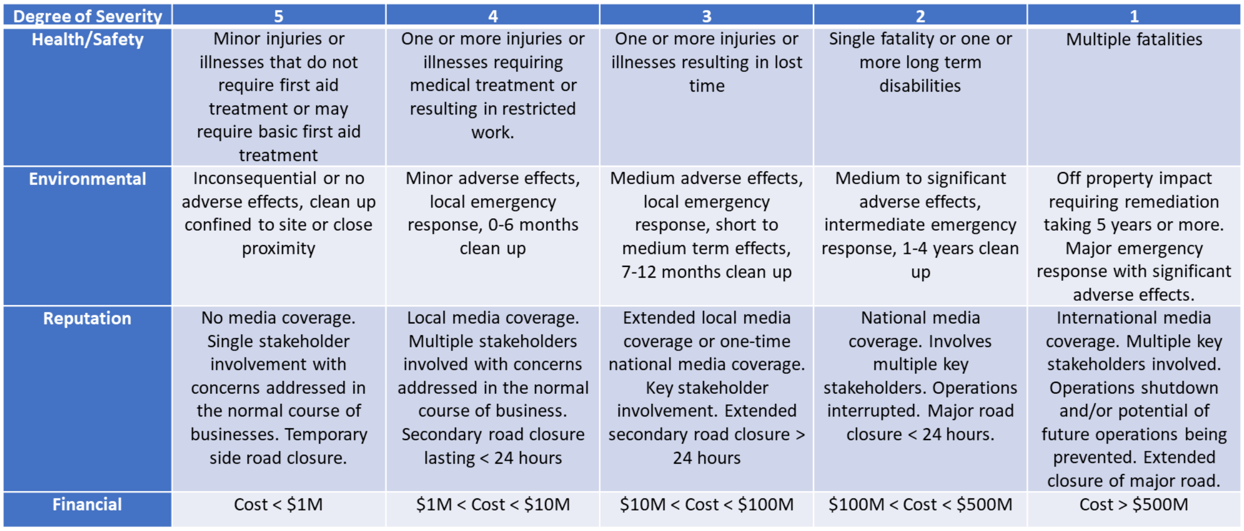
The statements are also matched to practical outputs that can be used by industry. Table 3-5 demonstrates how statements are categorized into the consequences categories used to generate a risk matrix. Several of these statements can fall into multiple categories.

Table 3-5. Statements categorized by consequence type.

Statement	Consequence
Leak/spill, Major leak/spill, Wildlife	Environment
Equipment (Damage - cost <$1m), Equipment (Failure - cost <$1m), Equipment (General - cost <$1m), Equipment (Damage - cost >$1m), Equipment (Failure - cost >$1m), Equipment (General - cost >$1m), Fire (Damage), Incorrect Operations, Leak/spill, Major leak/spill, Miscellaneous,	Finance

Miscommunication, Missing Equipment, Property Damage, Quality Assurance, Vehicle (heavy equipment), Vehicle (light vehicle), Vehicle collision (no injury), Weather, Wildlife	
Fatality, Fire (Injury), Injury, Laceration/abrasion, Minor Injury, Near Miss, No Treatment Injury, Severe Injury, Slip/trip/fall, Snow/ice, Sprain/strain, Vehicle collision (with injury), Weather	Health/Safety
Fatality, Severe Injury, Wildlife	Reputation

Our prior research generated a consequence scale to assign a numerical value denoting the severity of the risk, found in Figure 3-3 (Kurian et al., 2020). This consequence scale was created using the average values of consequences taken from the risk matrices of several companies collaborating with this research.



Degree of Severity	5	4	3	2	1
Health/Safety	Minor injuries or illnesses that do not require first aid treatment or may require basic first aid treatment	One or more injuries or illnesses requiring medical treatment or resulting in restricted work.	One or more injuries or illnesses resulting in lost time	Single fatality or one or more long term disabilities	Multiple fatalities
Environmental	Inconsequential or no adverse effects, clean up confined to site or close proximity	Minor adverse effects, local emergency response, 0-6 months clean up	Medium adverse effects, local emergency response, short to medium term effects, 7-12 months clean up	Medium to significant adverse effects, intermediate emergency response, 1-4 years clean up	Off property impact requiring remediation taking 5 years or more. Major emergency response with significant adverse effects.
Reputation	No media coverage. Single stakeholder involvement with concerns addressed in the normal course of businesses. Temporary side road closure.	Local media coverage. Multiple stakeholders involved with concerns addressed in the normal course of business. Secondary road closure lasting < 24 hours	Extended local media coverage or one-time national media coverage. Key stakeholder involvement. Extended secondary road closure > 24 hours	National media coverage. Involves multiple key stakeholders. Operations interrupted. Major road closure < 24 hours.	International media coverage. Multiple key stakeholders involved. Operations shutdown and/or potential of future operations being prevented. Extended closure of major road.
Financial	Cost < $1M	$1M < Cost < $10M	$10M < Cost < $100M	$100M < Cost < $500M	Cost > $500M

Figure 3-3. Consequence scale used to analyze the severity of incidents.

Table 3-6 uses the consequence scale (from Figure 3-3) to assign a severity rating to each statement and frequency scores were determined by using the tally of keywords in the incident database. Using the consequence and frequency scores, a risk score is generated. If multiple statements are selected, every category pertaining to the selected statements are represented on the risk matrix with their corresponding risk scores. If multiple selected statements have different consequence or frequency ratings within the same risk category, the greatest consequence value is selected to be represented on the risk matrix (along with its corresponding frequency).

Table 3-6. Consequence score assigned to statement using consequence scale in Figure 3.

	Consequence				**Frequency**			
Statement	Environment	Finance	Health/Safety	Reputation	Environment	Finance	Healthy/Safety	Reputation
Equipment (Damage - cost <$1m)		5				5		
Equipment (Failure - cost <$1m)		5				5		
Equipment (General - cost <$1m)		5				5		
Equipment (Damage - cost >$1m)		4				3		
Equipment (Failure - cost >$1m)		3				3		
Equipment (General - cost >$1m)		3				3		
Fatality			1	1			1	1
Fire (Damage)		4				3		
Fire (Injury)			3				3	
Incorrect Operations		5				5		
Injury			3				2	
Laceration/abrasion			4				4	
Leak/spill	4	4			3	3		
Major leak/spill	2	2		2	1	1		1
Minor Injury			4				5	
Miscellaneous		5				2		
Miscommunication		5				3		

Missing Equipment		5				2		
Near Miss			5				2	
No Treatment Injury			5				2	
Property Damage		5				1		
Quality Assurance		5				3		
Severe Injury			2	3			1	1
Slip/trip/fall			5				5	
Snow/ice		5	5			5	5	
Sprain/strain			3				3	
Vehicle (heavy equipment)		4				4		
Vehicle (light vehicle)		5				4		
Vehicle collision (no injury)		5				5		
Vehicle collision (with injury)		4	3			2	2	
Weather		5	5			5	5	
Wildlife	4	4		3	1	1		1

By counting the selected statements, and taking into account the date of the incident, it is also possible to plot trends of specific incident types by month. It is also possible to design prevention and mitigation strategies to match statements and combinations of statements. A similar process can also be used to identify leading indicators. These suggestions can be designed using a company's safety guidelines and procedures and asset management systems.

To illustrate our methodology, we present case studies with different consequences. We have received inputs from companies and generated the customized library. We assume that a user is inputting new incident reports, make assumptions about the statements selected by the user, and review the outputs created using the proposed methodology to demonstrate its practicality.

3.3.1 Case Study 1

The following sample incident taken from a company database is presented to demonstrate how this methodology is used to produce results: "*Hose-Traceability. Heat number on the elbows do not match with the heat number on the documents. Followed up with vendor to get appropriate heat numbers as they showed something different.*"

Given this incident report, the user will have to select from the following statements: Equipment (Damage - Cost <$1M), Equipment (Failure - Cost <$1M), Equipment (General - Cost <$1M), Equipment (Damage - Cost >$1M), Equipment (Failure - Cost >$1M), Equipment (General - Cost >$1M), Incorrect Operations, Miscommunication, Quality Assurance.

It was assumed that the user selects: Miscommunication, Quality Assurance, and Equipment (General – Cost <$1M).

With these statements, it was determined that the consequence is a financial risk with a consequence score of 5. By looking at the total number of occurrences of similar incidents, frequency was assigned a score of 3. This can be seen in the risk matrix found in Figure 3-4. The sample incident provided is a low consequence incident that occurs somewhat frequently. In most cases, missing quality assurance documents are simply inconvenient and may result in minor financial losses. Companies might decide to implement prevention methods or to encourage workers to be more methodical when filing such documents.

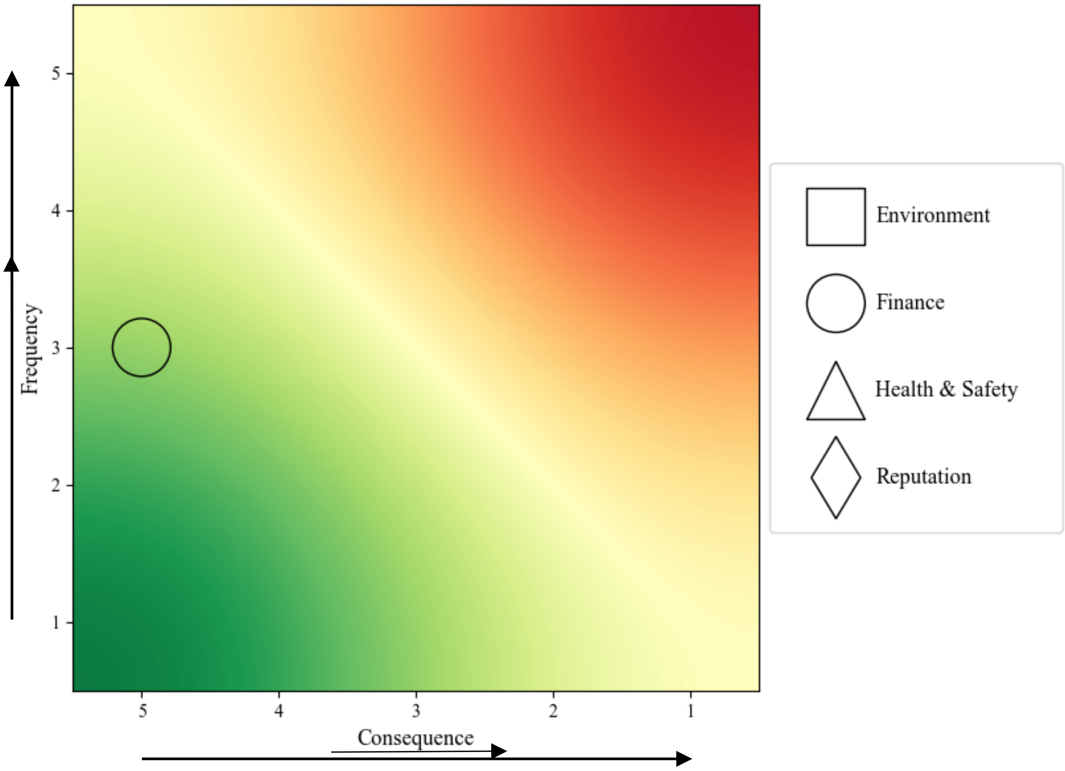


Figure 3-4. Risk matrix for Case Study 1

With respect to the formatting of the risk matrix, the axis labels (from least to greatest, 1-5 frequency scale and 5-1 consequence scale) was determined based on industry practice. The risk matrix is also given a gradient effect to show low impact risks as green and high impact risks as red.

A trend report, shown in Figure 3-5, is generated by creating a histogram of similar incidents by month. The algorithm counts the number of occurrences of incidents with the same statements selected and displays the trends of these incidents for the past year. In this example, it can be observed that it is common for incidents of this type to occur during the middle of the year. It might also be worth investigating why such incidents occurred frequently in December, but were quite rare in January.

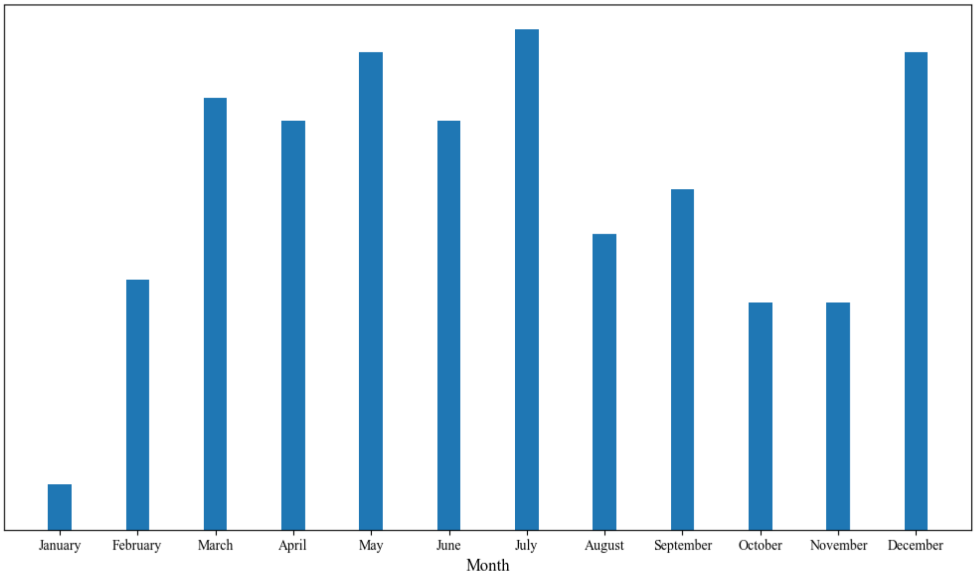


Figure 3-5. Example of trend analysis for Case Study 1 indicating number of incidents per month.

Based on the selected statements, it is also possible to design prevention and mitigation strategies. These suggestions can be modified in the future and tailored to match the safety guidelines and procedures of different companies. For this incident, the program has been designed to provide the following prevention and mitigation strategies: (1) verify that instructions are clearly received, (2) clarify any doubts with the individual(s) assigning the task, (3) ensure that the task at hand is logical without blindly completing the assigned work, (4) verify part number before and after the order, and (5) ensure that all documents are properly handled and stored. The leading indicators are identified as: (1) poor filing system and (2) inadequate training.

3.3.2. Case Study 2

Next, we present a different incident with multiple consequences to demonstrate the versatility of the methodology: "*Icy road conditions. Employee truck and 3rd party vehicle made driver side contact. Employee complained about minor whiplash.*"

This incident prompts the user to select from the following statements: Fatality, Fire (Damage), Fire (Injury), Injury, Minor Injury, Near Miss, No Treatment Injury, Severe Injury, Vehicle (Heavy Equipment), Vehicle (Light Vehicle), Vehicle Collision (No Injury), Vehicle Collision (With Injury).

Here, it was assumed that the user selects: Injury, Vehicle (Light Vehicle), and Vehicle Collision (With Injury).

Given these selections, it was determined that the consequence is a Health/Safety risk with a consequence score of 3 and a Financial risk with a consequence score of 5. By looking at the total number of occurrences of similar incidents, the frequency was assigned a score of 2 – the reports of vehicle collisions with injury are much fewer in number than those of vehicle collisions with no injury. The risk matrix for this incident can be seen in Figure 3-6. The incident can be considered a low-to-medium risk where the health and safety component of the incident has more ramifications than the financial loss. Also note that for incidents with multiple consequences, the risk matrix has been designed to allow symbols to overlap.

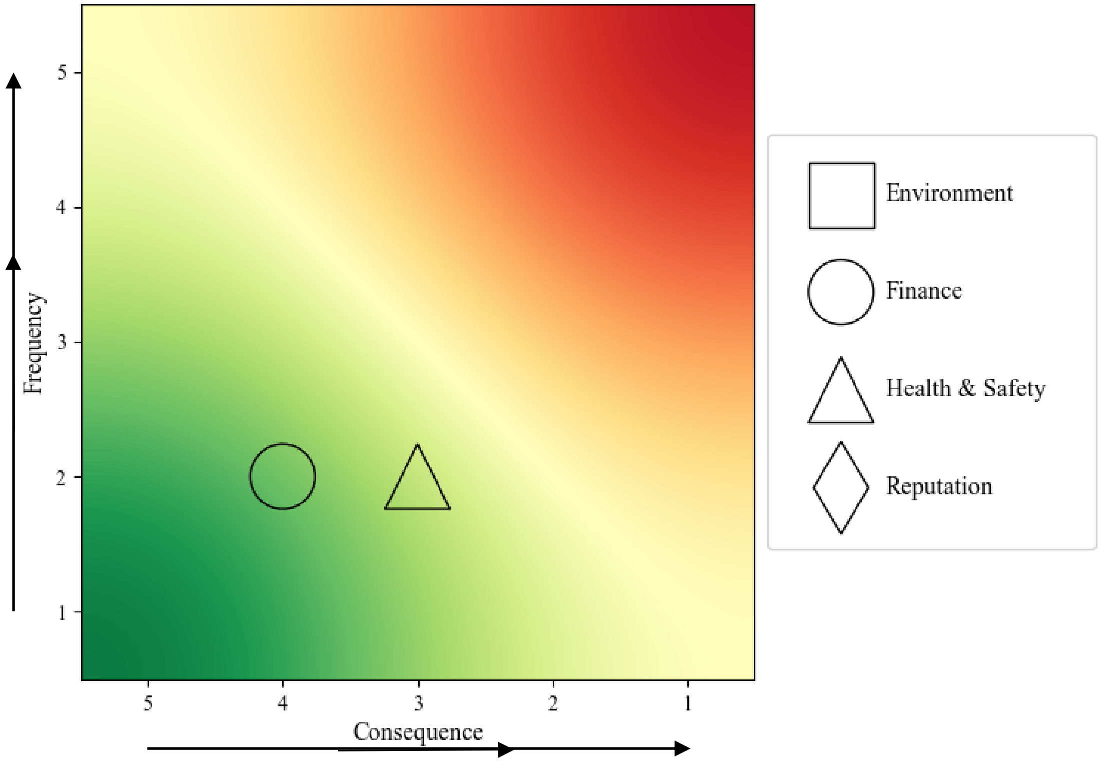


Figure 3-6. Risk matrix for Case Study 2.

Figure 3-7 shows the trends for vehicular incidents with injuries. As expected, vehicular incidents involving collisions occur more frequently in winter. It might be beneficial for the company to identify factors pertaining to the cause of similar incidents, such as the geographic location, time of day, existing traffic signs, driving conditions, etc., in order to determine methods for prevention and mitigation.

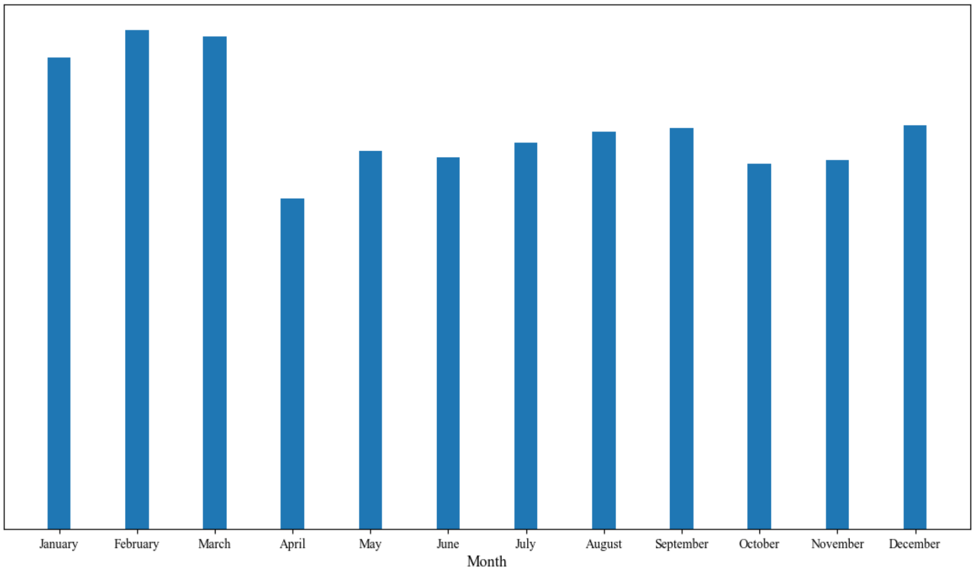


Figure 3-7. Example of trend analysis for Case Study 2 indicating number of incidents per month.

Based on the statements selected, the algorithm has been designed to provide the following prevention and mitigation strategies: (1) drive at a speed suitable to road conditions, (2) ensure that vehicle is properly equipped for winter weather (e.g. winter tires, first aid kit, etc.), (3) pay attention to other vehicles on the road (e.g. make sure other drivers are not distracted, maintain safe following distance, check blind spots), (4) make sure that the seat is properly adjusted to provide ample neck and lumbar support, and (5) provide training for workers to drive in winter road conditions. The leading indicators are identified as: (1) winter weather and (2) poor traction.

3.3.3. Case Study 3

The final case study is a very frequent incident found in the database: "*Worker slipped and fell in the parking lot. Employee took a shortcut between the middle and largest (left most) park after parking her vehicle in the middle parking lot.*"

The resulting statements are given for the user to select from: Fatality, Fire (Injury), Injury, Laceration / Abrasion, Leak / Spill, Major Leak / Spill, Minor Injury, Near Miss, No Treatment Injury, Severe Injury, Slip / Trip / Fall, Sprain / Strain, Vehicle (Heavy Equipment), Vehicle (Light Vehicle), Vehicle Collision (No Injury), Vehicle Collision (With Injury).

It was assumed that Minor Injury and Slip / Trip / Fall were selected.

With these selections, the risk has a health and safety consequence score of 4 and a frequency of 5. Such incidents are very common, especially in the winter season. The risk matrix for the incident is displayed in Figure 3-8.

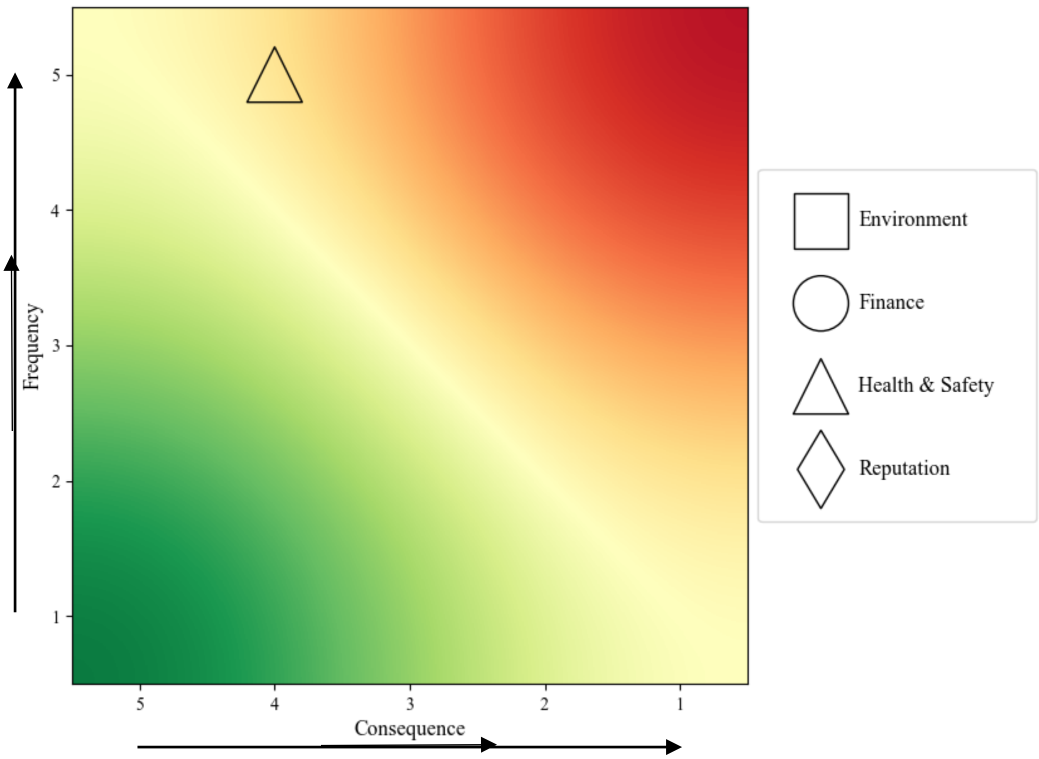


Figure 3-8. Risk matrix for Case Study 3.

An example of a trend report for slip/trip/fall incidents can be seen in Figure 3-9. Slip/trip/fall incidents are common, particularly due to snow or ice in winter. There are also other contributing factors in other seasons that might result in slippery surfaces. It might be beneficial for a company to impress upon workers the importance of proper footwear such as cleats and requesting signage or countermeasures (e.g. salt or gravel) at the source of a tripping hazard.

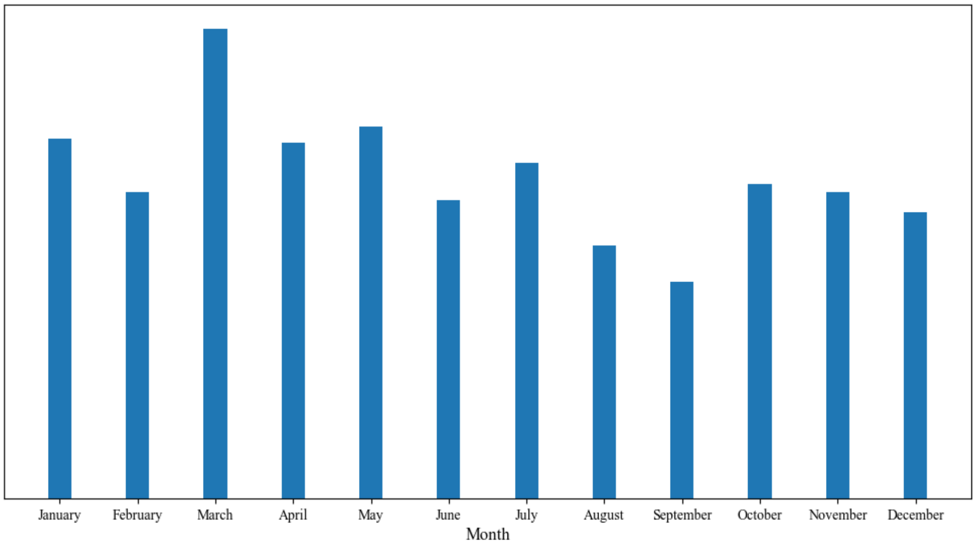


Figure 3-9. Example of trend analysis for Case Study 3 indicating number of incidents per month.

Based on the statements selected to describe the incident, the algorithm provides the following prevention and mitigation strategies: (1) walk slowly and carefully on snow and ice, (2) wear proper and weather-appropriate footwear outdoors, and (3) ensure that proper authorities are notified to apply salt/gravel to regular walkways and parking lots. The leading indicators are identified as: (1) snow and/or ice and (2) lack of salt/gravel.

As demonstrated by the case studies, applying such methods to incident reporting makes it possible to improve safety in a company at a foundational level. Having access to such information can allow companies to enhance their safety culture by providing timely prevention and mitigation strategies; giving feedback on safety performance versus historical trends; building a reputation with their employees for protecting occupational safety, process safety, and the environment; and reducing financial losses by focusing on higher priority risks that require attention.

3.4. CONCLUSION

We began with the observation that AI/ML hold great promise for enhancing companies' safety management systems. Yet, the adoption of AI/ML tools has been slow given an overwhelming quantity of dirty, incomplete, and fragmented data; a multitude of low-consequence incidents of unknown analytical value; and operators' resulting perception that this is a high-effort, low-payoff venture (Ransbotham et al., 2017). We have addressed these barriers by aggregated data across operators and using keyword analysis to create more complete, 'clean', and larger datasets to enhance algorithm training, avoid overfitting, and more sensitive 'detectability' of leading indicators (Kurian et al., 2020).

Our objectives for expanding on this research was to improve current incident reporting systems, provide practical and tailored outputs to prevent and mitigate risk, and to create opportunities to reduce losses due to incidents. By implementing a system that incorporates machine learning and keyword analysis with an intermediate step for user input, it was possible to accomplish these goals. With the methodology used in this research, we analyzed incident reports and generated a framework for evaluating and reducing risk.

The analysis of incident reports used a supervised machine learning algorithm to predict identifying labels incidents for incidents. Next, the spaCy library from Python was used to lemmatize the incident reports. The resulting words were tallied and the most common words, along with the identifying labels used for machine learning analysis, were used to generate a customized library. The words stored in the customized library were assigned to statements used to accurately describe risk involved in incidents. When a new incident report is inputted, this system runs the machine learning algorithm to predict an identifying label for the incident report, based on the newly expanded text corpora. The incident report is then lemmatized and cross referenced with the words stored in the customized library. Statements corresponding to the words in the customized library are then provided to the user. When the user selects statements matching the incident that occurred, a series of output results are provided. These outputs include a risk matrix, trends of similar incidents within the past year, suggested prevention and mitigation strategies, and any leading indicators that could be identified to prevent future occurrences of similar events. In this manner, the system is constantly learning from newly inputted incident data. In this manner, the system is constantly learning from newly inputted incident data. Likewise, experts can examine trends and revisit suggested prevention and mitigation strategies, to continually refine these.

Three case studies of incident report inputs were analyzed using the proposed methodology. These incidents included quality assurance, vehicular, and slip/trip/fall -type incidents. By analyzing the trends in these incidents, it was surprising to see the low number of quality assurance-type incidents in January (in comparison to the rest of the year). As expected, the highest number of vehicular and slip/trip/fall incidents occurred during the winter months. By analyzing the factors resulting in the trends of such incidents, it is possible for companies to develop plans for future incidents. For example, a company could attempt to identify the reason for the low number of quality assurance incidents in January and attempt to reproduce these results for the remainder of the year. Additional effort could also be focused on reducing vehicular and slip/trip/fall incidents to protect worker safety.

This research proposed new methods to report and analyze incident reports using artificial intelligence, machine learning, and keyword analysis. There is much potential for future use and implementation – many of the variables used in this study can be easily modified to match the varying needs of different companies. The reports generated by implementing this methodology can allow a company to focus its efforts on preventing those incidents that are causing the greatest losses and to identify strengths within their existing systems. Furthermore, the methodology described in this research can be applied by any industry seeking to reduce risk using incident reports.

3.5. REFERENCES

Albery, S., Borys, D., & Tepe, S. (2016). Advantages for risk assessment: Evaluating learnings from question sets inspired by the FRAM and the risk matrix in a manufacturing environment. *Safety Science*, *89*, 180–189. doi: 10.1016/j.ssci.2016.06.005

Aven, T. (2011). On some recent definitions and analysis frameworks for risk, vulnerability, and resilience. Risk Analysis: An International Journal, 31(4), 515-522.

Bengio, Y., Goodfellow, I., & Courville, A. (2017). Deep learning (Vol. 1). MIT press.

Dong, C., Dong, X., Gehman, J., & Lefsrud, L. (2017). Using BP neural networks to prioritize risk management approaches for China's unconventional shale gas industry. Sustainability, 9(6), 979.

Duijm, N. J. (2015). Recommendations on the use and design of risk matrices. *Safety Science, 76*, 21–31. https://doi.org/10.1016/j.ssci.2015.02.014

Garreta, R., Hauck, T., & Hackeling, G. (2017). *Scikit-learn: machine learning simplified*. Birmingham, UK: Packt Publishing.

Goerlandt, F., & Reniers, G. (2016). On the assessment of uncertainty in risk diagrams. *Safety Science*, *84*, 67–77. doi: 10.1016/j.ssci.2015.12.001

Government of Alberta (2019). *Reporting and investigating injuries and incidents – OHS information for employers, prime contractors and workers*.

Imani, A., Forman, J. E., & Amir, W. (2018). *A Clustering Analysis of Codes of Conduct and Ethics in the Practice of Chemistry*.

International Organization for Standardization [ISO]. (2018). *Risk Management - Guidelines* (ISO 31000:2018E)

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. Academy of Management Annals, 14(1), 366-410.

Khakzad, N., Khan, F., & Amyotte, P. (2013). Dynamic safety analysis of process systems by mapping bow-tie into Bayesian network. Process Safety and Environmental Protection, 91(1-2), 46-53.Kotsiantis, S. B. (2007). *Supervised machine learning: a review of classification techniques*. https://doi.org/10.1115/1.1559160

Kurian, D., Ma, Y., Lefsrud, L., & Sattari, F. (2020). *Seeing the Forest and the Trees: Using Machine Learning to Categorize and Analyze Incident Reports for Alberta Oil Sands Operators*. Manuscript in revision for publiciation.

Laberge, M., Maceachen, E., & Calvet, B. (2014). Why are occupational health and safety training approaches not effective? Understanding young worker learning processes using an ergonomic lens. *Safety Science*, *68*, 250–257. doi: 10.1016/j.ssci.2014.04.012

Malhotra, A. (2018). Introduction to Libraries of NLP in Python - NLTK vs. spaCy. Retrieved from https://medium.com/@akankshamalhotra24/introduction-to-libraries-of-nlp-in-python-nltk-vs-spacy-42d7b2f128f2.

Milgram, J., Cheriet, M., & Sabourin, R. (2006). "One Against One" or "One Against All": Which One is Better for Handwriting Recognition with SVMs? Tenth International Workshop on Frontiers in Handwriting Recognition, Université de Rennes 1, Oct 2006, La Baule (France). inria00103955. Retrieved from https://hal.inria.fr/inria-00103955

Muhlbauer, W. K. (2004). *Pipeline Risk Management Manual - Ideas, Techniques, and Resources* (Third Edit). 200 Wheeler Road, Burlington, MA, USA: Gulf Professional Publishing (an imprint of Elsevier).

Ni, H., Chen, A., & Chen, N. (2010). Some extensions on risk matrix approach. *Safety Science*, *48*(10), 1269–1278. doi: 10.1016/j.ssci.2010.04.005

Nordlöf, H., Wiitavaara, B., Winblad, U., Wijk, K., & Westerling, R. (2015). Safety culture and reasons for risk-taking at a large steel-manufacturing company: Investigating the worker perspective. *Safety Science*, *73*, 126–135. doi: 10.1016/j.ssci.2014.11.020

Pasquini, A., Pozzi, S., Save, L., & Sujan, M.-A. (2017). Requisites for Successful Incident Reporting in Resilient Organisations. *Resilience Engineering in Practice*, 237–256. doi: 10.1201/9781317065265-17

Patriarca, R., Bergström, J., Gravio, G. D., & Costantino, F. (2018). Resilience engineering: Current status of the research and future challenges. *Safety Science*, *102*, 79–100. doi: 10.1016/j.ssci.2017.10.005

Project Management Institute. (2004). A guide to the project management body of knowledge (PMBOK guide). Newtown Square, Pa: Project Management Institute.

Ransbotham, S., Kiron, D., Gerbert, P., & Reeves, M. (2017). Reshaping business with artificial intelligence: Closing the gap between ambition and action. MIT Sloan Management Review, 59(1).

Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms.

Srinivasa-Desikan, B. (2018). *Natural language processing and computational linguistics a practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Birmingham: Packt.

Steen, R., & Aven, T. (2011). A risk perspective suitable for resilience engineering. Safety Science, 49(2), 292-297.

Thomas, P., Bratvold, R. B., & Bickel, J. E. (2013). The Risk of Using Risk Matrices Decision Analysis View project A Generalized Sampling Approach for Multilinear Utility Functions Given Partial Preference Information View project The Risk of Using Risk Matrices. *Article in SPE Economics and Management*, (April 2015). https://doi.org/10.2118/166269-MS

Xu, Q., Xu, K., Li, L., & Yao, X. (2018). Safety assessment of petrochemical enterprise using the cloud model, PHA–LOPA and the bow-tie model. Royal Society open science, 5(7), 180212.

Zhao, J., Cui, L., Zhao, L., Qiu, T., & Chen, B. (2009). Learning HAZOP expert system by case-based reasoning and ontology. Computers & Chemical Engineering, 33(1), 371-378.

Zuboff, S. (2015). Big other: surveillance capitalism and the prospects of an information civilization. Journal of Information Technology, 30(1), 75-89.

4. CONCLUSIONS AND FUTURE RESEARCH

4.1. CONCLUSIONS

The objectives for completing this research were to improve current incident reporting systems, provide methods for companies to improve consistency and accuracy when evaluating risk, and deliver practical and tailored outputs to prevent and mitigate risk. The goal of consistently and accurately evaluating risk was accomplished by implementing machine learning. The machine learning system classified risks with accuracies >80%. Later research combined a portion of the machine learning algorithm (with accuracy of ~88.48%) with keyword analysis to further increase accuracy. By implementing a step for user input, it was possible to create a new method for reporting incidents with users selecting from a list of programmed statements pertaining to an incident. This improved system of reporting incidents allowed for specific outputs to be designated to statements, providing practical and tailored outputs for companies to use to prevent and mitigate risk. Identifying leading indicators was another target for this research. Leading indicators were identified using the categories of operations-based, systems-based, and behavior-based leading indicators as a guideline.

4.2. CONTRIBUTIONS AND FUTURE WORK

The greatest empirical contribution made by this research was the methodology employed – the application of machine learning and keyword analysis along with rule-based arguments to the reporting of incidents. While it is possible to find previous work exploring machine learning as an analysis tool for specific incident types, there is little research that has been completed to broadly analyze all types of incident reports (Kakhki, 2019). Most previous research also focuses on the aspect of predicting incidents while this research seeks to change how incidents are reported and to prevent incidents from occurring (without stopping at simply predicting incidents) (Sarkar et al., 2019). This research allows companies to realize the potential of and bring value to their incident databases. The identification of leading indicators, in particular, has created opportunities for companies to increase awareness of incidents that have already occurred and existing hazards. With this information, workers can be trained to deal with these hazards prior to beginning their work.

With this, my research also makes an empirical contribution, by suggesting enhancements to how companies input, analyze, and use incident data. It is possible to improve the accuracy of machine learning classification by focusing the efforts of study on a single company's incident database as opposed to the aggregated incident database. By doing this, it is possible to increase the accuracy while using the aggregated incident database to tailor the results for specific requirements. The outputs generated by the process of analyzing incident reports with machine learning and keyword analysis can be modified to fulfil different roles – these outputs can also be used to identify incidents that can be prevented to maximize safety and minimize losses. The methodology described in this research can be applied by any industry seeking to reduce risk using incident reports.

Further, by aggregating incidents across multiple companies, it is possible to identify the leading indicators for higher consequence, low frequency incidents. Aven (2015) describes black swans as surprising events that have extreme impact. Due to the relatively low number of high consequence incidents that occur, it is difficult to prepare for events that could be classified as black swans. While black swan events are considered outliers and difficult to predict (Aven, 2015; Taleb, 2007), combining the incident databases of several companies increases the predictive power of incident analysis by increasing the amount of information and experience available. To put it differently, companies can analyze the leading indicators of incidents that have been found by other companies to develop strategies to deal with such hazards. As a defining feature of black swan events, surprise is an element that must be removed; and by pooling resources and increasing strength of knowledge, it is possible to prepare for low frequency risks (Aven and Krohn, 2014).

Research is currently being undertaken to apply elements of Process Safety Management (PSM) from OSHA's standard to categorize the risks found within incident reports. This will further align the deliverables of the methodology described in this paper to an industry standard where the outputs can be analyzed by individuals seeking to properly investigate incidents and reduce risk. As can be seen from this example, many of the variables used in this research can be modified for a variety of purposes to benefit a company. By changing how incidents are classified, it is possible to search for different trends within the data. In fact, almost every variable can be changed to match different specifications. The risk matrix used to evaluate risk can be changed; different companies can input their own operating procedures and safety guidelines to design customized prevention and mitigation strategies; and even the incident database can be replaced to provide more accurate results specific to a single company.

There are two aspects of the methodology used in this research that are constantly evolving: the machine learning that predicts a label for the incident report and the trend analysis output. When a new incident report is inputted, the machine learning algorithm adds the incident report and its label to the existing incident database to further bolster the accuracy of predicting labels. The trend reports that are outputted also change constantly. The current system is designed to provide trends for incidents that have occurred in the past twelve months; however, this duration can be modified. As time passes, old incidents are removed from the system and new incidents are added to provide current trends within incident reports that can be used to develop up-to-date strategies to prevent or mitigate risk.

Lastly, this research contributes to resilience theories. For organizations to learn from their operations experience, subject matter experts, and previous incidents, they must be able to monitor their incidents, appropriately respond to negative trends, and address incidents with near miss potential (see Figure 4-1). However, companies are often overwhelmed by thousands of ‘messy’ (incomplete reports, rife with spelling mistakes, erroneously rated) incident reports and are unable to analyze the latent causes, consequences, and mitigations. As a result, their responses are often insufficient, which can result in a more consequential incident.

This research provides companies with a method to design inherently safe systems by automating the process of analyzing incident reports. By tracking the efficacy of mitigation measures over time, it is possible to evaluate system effectiveness. At the same time, tracking the efficacy of leading indicators can be used to evaluate monitoring effectiveness. Finally, effectiveness of operations can be evaluated by tracking the efficacy of inherently safer design efforts to eliminate hazards. Evaluating system, monitoring, and operations effectiveness is an ongoing process that will allow companies to constantly improve and design more resilient systems (Hollnagel et al., 2007). The methods described in this research have placed emphasis on constantly improving the accuracy of classification and providing up-to-date trends as new incident reports are added. In addition to this ability to evolve, the algorithm is able to handle large amounts of incident reports and constantly provide feedback. This can be used to take advantage of massive incident databases, that might otherwise remain misused, and use this data to provide valuable insight towards creating a resilient system.

As part of designing a system that is resilient, organizations must be willing to learn, anticipate, monitor, and respond to changes (Lefsrud, 2019). A key component of a resiliency is to constantly test whether ideas about risk match reality (Hollnager et al., 2007). As such, this research delivers outputs that can be used to design a resilient system. By designing systems that are resilient, an organization is better equipped to know what has happened, know what to expect, know what to do, and know what to look for.

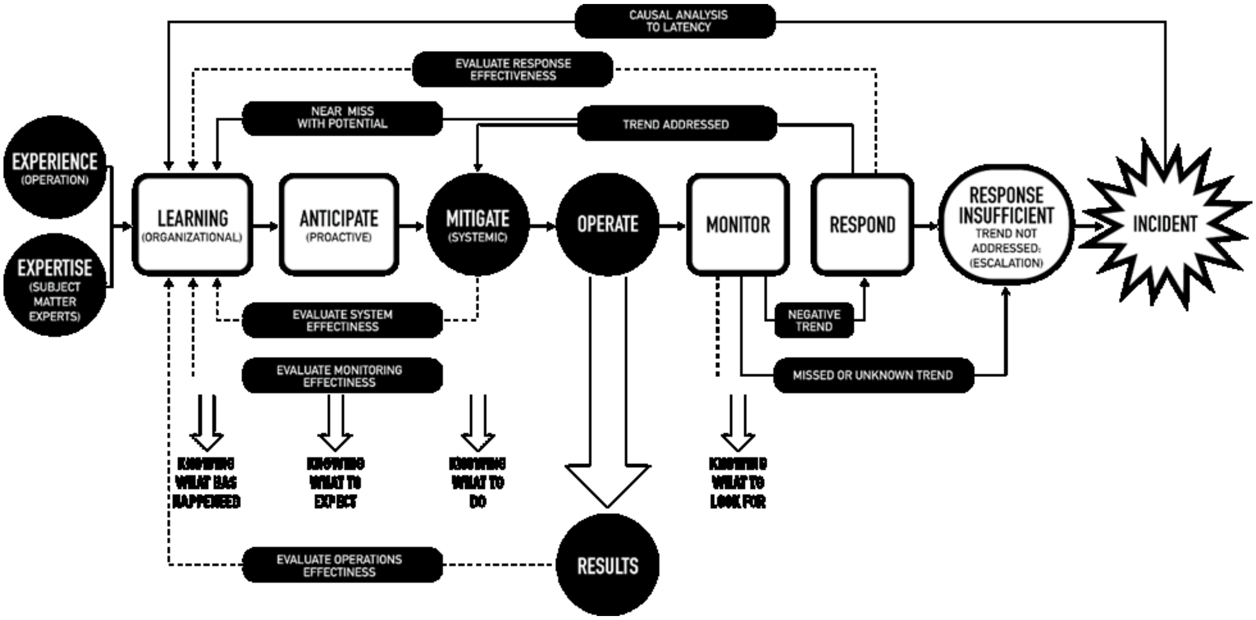

CAUSAL ANALYSIS TO LATENCY
EVALUATE RESPONSE EFFECTIVENESS
NEAR MISS WITH POTENTIAL
TREND ADDRESSED
EXPERIENCE (OPERATION)
EXPERTISE (SUBJECT MATTER EXPERTS)
LEARNING (ORGANIZATIONAL)
ANTICIPATE (PROACTIVE)
MITIGATE (SYSTEMIC)
OPERATE
MONITOR
RESPOND
RESPONSE INSUFFICIENT TREND NOT ADDRESSED: (ESCALATION)
INCIDENT
EVALUATE SYSTEM EFFECTINESS
NEGATIVE TREND
MISSED OR UNKNOWN TREND
EVALUATE MONITORING EFFECTINESS
KNOWING WHAT HAS HAPPENED
KNOWING WHAT TO EXPECT
KNOWING WHAT TO DO
KNOWING WHAT TO LOOK FOR
EVALUATE OPERATIONS EFFECTINESS
RESULTS

Figure 4-1. Safety Management System processes build resilience (Lefsrud, 2019; revised from Hollnagel et al., 2007)

4.3. LIMITATIONS

This paper has highlighted several areas for future research. At the same time, there were certain limitations due to the scope of research and the resources available. The “Statements” included in the customized library were chosen by surveying several professors and industry experts. Also, due to lack of corporate support, the suggested prevention and mitigation strategies and the identification of leading indicators were determined manually using resources available to students. These resources included provincial, federal, and global standards typically used for analyzing processes and risk. Future research could use a single company’s safety procedures, operating guidelines, and asset management systems to tailor the outputs for a specific company’s needs. The “Statements” could be determined using a machine learning algorithm to analyze the incident database. Furthermore, the relationships between the “Statements” and the corresponding outputs could also be determined using a combination of machine learning and principles of science and engineering. This would create a more robust system for both reporting incidents and analyzing the risk within incidents.

One point which was not mentioned previously deals with inputting an incident report which has not been previously reported into the system. In such a case, while the combination of machine learning and keyword analysis can still predict “Statements” to match the incident report, there is a very low chance that the algorithm will not be able to supply any outputs for a user to select. Future research could accommodate for this by allowing the user to classify the incident and then enabling the machine learning algorithm to teach itself how to deal with similar incidents in the future.

Another aspect for future research could focus on the predictive element of analyzing incident data. By teaching a machine learning algorithm to provide prevention and mitigation strategies and to identify the leading indicators of an incident, the same rhetoric would enable companies to train their workers in identifying hazards and preventing incidents from occurring. As incidents are reported for a work site, trend reports about the incidents occurring on site can be used to help workers prepare for such incidents, paying particular attention to the incidents that are occurring most frequently.

Finally, one final project for the future could include designing a user interface to process incident reports using methodology similar to that proposed in this paper. Screenshots of some work that has been started can be found in Appendix B. Research in this area could create a website or mobile application that could be accessed in a variety of scenarios. The potential for this is immense – workers could have access to incident databases and solutions for dealing with incidents in the palm of their hands.

REFERENCES

A Guide to the Project Management Body of Knowledge [PMBOK] (Sixth Edit). (2017). Newtown Square, PA, USA: Project Management Institute, Inc.

Albery, S., Borys, D., & Tepe, S. (2016). Advantages for risk assessment: Evaluating learnings from question sets inspired by the FRAM and the risk matrix in a manufacturing environment. *Safety Science*, *89*, 180–189. doi: 10.1016/j.ssci.2016.06.005

Animah, I., & Shafiee, M. (2019). Application of risk analysis in the liquefied natural gas (LNG) sector: An overview. *Journal of Loss Prevention in the Process Industries*, 63(October 2019), 103980. https://doi.org/10.1016/j.jlp.2019.103980

Aven, T. (2015). Implications of black swans to the foundations and practice of risk assessment and management. *Reliability Engineering & System Safety*, *134*, 83–91. doi: 10.1016/j.ress.2014.10.004

Aven, T., & Krohn, B. S. (2014). A new perspective on how to understand, assess and manage risk and the unforeseen. *Reliability Engineering & System Safety*, *121*, 1–10. doi: 10.1016/j.ress.2013.07.005

Aven, T. (2011). On some recent definitions and analysis frameworks for risk, vulnerability, and resilience. Risk Analysis: An International Journal, 31(4), 515-522.

Basu, S. (2017). Basics of Hazard, Risk Ranking, and Safety Systems. In *Plant Hazard Analysis and Safety Instrumentation Systems*. https://doi.org/10.1016/b978-0-12-803763-8.00001-7

Bengio, Y., Goodfellow, I., & Courville, A. (2017). Deep learning (Vol. 1). MIT press.

Bjerga, T., & Aven, T. (2015). Adaptive risk management using new risk perspectives - An example from the oil and gas industry. *Reliability Engineering and System Safety, 134,* 75–82. https://doi.org/10.1016/j.ress.2014.10.013

Calixto, E. (2016). Reliability and Safety Processes. In Gas and Oil Reliability Engineering. https://doi.org/10.1016/b978-0-12-805427-7.00006-3

Cass, S. (2019). The Top Programming Languages 2019. Retrieved from IEEE Spectrum website: https://spectrum.ieee.org/computing/software/the-top-programming-languages-2019

Christou, M., & Konstantinidou, M. (2013). Safety of offshore oil and gas operations Lessons from past accident analysis : ensuring EU hydrocarbon supply through better control of major hazards - Study. *JRC Scientific and Policy Reports*. doi: 10.2790/71887

Cunha, S. B. da. (2016). A review of quantitative risk assessment of onshore pipelines. *Journal of Loss Prevention in the Process Industries*, 44, 282–298. https://doi.org/10.1016/j.jlp.2016.09.016

Dong, C., Dong, X., Gehman, J., & Lefsrud, L. (2017). Using BP neural networks to prioritize risk management approaches for China's unconventional shale gas industry. Sustainability, 9(6), 979.

Duijm, N. J. (2015). Recommendations on the use and design of risk matrices. *Safety Science*, 76, 21–31. https://doi.org/10.1016/j.ssci.2015.02.014

Garreta, R., Hauck, T., & Hackeling, G. (2017). *Scikit-learn: machine learning simplified*. Birmingham, UK: Packt Publishing.

Goerlandt, F., & Reniers, G. (2016). On the assessment of uncertainty in risk diagrams. *Safety Science*, *84*, 67–77. doi: 10.1016/j.ssci.2015.12.001

Government of Alberta (2019). *Reporting and investigating injuries and incidents – OHS information for employers, prime contractors and workers*.

Greenwell, W. S., Knight, J. C., & Strunk, E. A. (2003). *Risk-based Classification of Incidents*. 39–50.

Guenther, N., & Schonlau, M. (2016). Support vector machines. *The Stata Journal*, *16*(4), 917–937. https://doi.org/10.1177/1536867X1601600407

Gul, M., & Guneri, A. F. (2016). A fuzzy multi criteria risk assessment based on decision matrix technique: A case study for aluminum industry. *Journal of Loss Prevention in the Process Industries*, *40*, 89–100. https://doi.org/10.1016/j.jlp.2015.11.023

Halim, S. Z., & Mannan, M. S. (2018). A journey to excellence in process safety management. *Journal of Loss Prevention in the Process Industries*, 55(April), 71–79. https://doi.org/10.1016/j.jlp.2018.06.002

Hollnagel, E., Woods, D. D., & Leveson, N. (2007). *Resilience engineering: Concepts and precepts*. Ashgate Publishing, Ltd.

Imani, A., Forman, J. E., & Amir, W. (2018). *A Clustering Analysis of Codes of Conduct and Ethics in the Practice of Chemistry*.

Inouye, J. (n.d.). Practical guide to leading indicators: Metrics, case studies & strategies. Retrieved from https://www.nsc.org/Portals/0/Documents/CambpellInstituteandAwardDocuments/WP-PracticalGuidetoLI.pdf

International Organization for Standardization [ISO]. (2018). *Risk Management - Guidelines* (ISO 31000:2018E)

Kakhki, F. D., Freeman, S. A., & Mosher, G. A. (2019). Evaluating machine learning performance in predicting injury severity in agribusiness industries. *Safety Science*, *117*, 257–262. doi: 10.1016/j.ssci.2019.04.026

Kane, S. (1985). Medical Risk Management: Preventive Legal Strategies for Health Care Providers. *AORN Journal*, *42*(6), 950. doi: 10.1016/s0001-2092(07)64436-6

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. Academy of Management Annals, 14(1), 366-410.

Khakzad, N., Khan, F., & Amyotte, P. (2013). Dynamic safety analysis of process systems by mapping bow-tie into Bayesian network. Process Safety and Environmental Protection, 91(1-2), 46-53.

Kletz, T. A. (2005). Looking beyond ALARP - Overcoming its limitations. *Institution of Chemical Engineers Symposium Series*, (150), 69–76. https://doi.org/10.1205/psep.04227

Kotsiantis, S. B. (2007). *Supervised machine learning: a review of classification techniques*. https://doi.org/10.1115/1.1559160

Kurian, D., Ma, Y., Lefsrud, L., & Sattari, F. (2020). Seeing the forest and the trees: Using machine learning to categorize and analyze incident reports for Alberta oil sands operators. *Journal of Loss Prevention in the Process Industries*, *64*, 104069. doi: 10.1016/j.jlp.2020.104069

Laberge, M., Maceachen, E., & Calvet, B. (2014). Why are occupational health and safety training approaches not effective? Understanding young worker learning processes using an ergonomic lens. *Safety Science*, *68*, 250–257. doi: 10.1016/j.ssci.2014.04.012

Landell, H. (2016). *The Risk Matrix as a tool for risk analysis*.

Lefsrud, L. M. (2019). *Safety Moment: Equipping your organizations to better see the unforeseen*. Presentation to Suncor Leadership Roundtable, 27 May 2019.

Leistad, G. H., & Bradley, A. R. (2009). Is the focus too low on issues that have a potential to lead to a major incident? *Society of Petroleum Engineers - Offshore Europe Oil and Gas Conference and Exhibition 2009, OE 2009*, 1, 467–472. https://doi.org/10.2118/123861-ms

Macrae, C. (2015). The problem with incident reporting: Table 1. BMJ Quality & Safety, 25(2), 71–75. https://doi.org/10.1136/bmjqs-2015-004732

Malhotra, A. (2018). Introduction to Libraries of NLP in Python - NLTK vs. spaCy. Retrieved from https://medium.com/@akankshamalhotra24/introduction-to-libraries-of-nlp-in-python-nltk-vs-spacy-42d7b2f128f2.

Markowski, A. S., & Mannan, M. S. (2008). Fuzzy risk matrix. *Journal of Hazardous Materials*, *159*(1), 152–157. https://doi.org/10.1016/j.jhazmat.2008.03.055

Milgram, J., Cheriet, M., & Sabourin, R. (2006). "One Against One" or "One Against All": Which One is Better for Handwriting Recognition with SVMs? Tenth International Workshop on Frontiers in Handwriting Recognition, Université de Rennes 1, Oct 2006, La Baule (France). inria00103955. Retrieved from https://hal.inria.fr/inria-00103955

Mkpat, E., Reniers, G., & Cozzani, V. (2018). Process safety education: A literature review. *Journal of Loss Prevention in the Process Industries*, *54*(October 2017), 18–27. https://doi.org/10.1016/j.jlp.2018.02.003

Muhlbauer, W. K. (2004). *Pipeline Risk Management Manual - Ideas, Techniques, and Resources* (Third Edit). 200 Wheeler Road, Burlington, MA, USA: Gulf Professional Publishing (an imprint of Elsevier).

Ness, A. (2015). Lessons learned from recent process safety incidents. *Chemical Engineering Progress*, *111*(3), 23–29.

Ni, H., Chen, A., & Chen, N. (2010). Some extensions on risk matrix approach. *Safety Science*, *48*(10), 1269–1278. doi: 10.1016/j.ssci.2010.04.005

Nordlöf, H., Wiitavaara, B., Winblad, U., Wijk, K., & Westerling, R. (2015). Safety culture and reasons for risk-taking at a large steel-manufacturing company: Investigating the worker perspective. *Safety Science*, *73*, 126–135. doi: 10.1016/j.ssci.2014.11.020

Occupational Health and Safety Act [OHS]. , Pub. L. No. Chapter O-2.1 (2018).

Occupational Safety and Health Administration [OSHA]. , U.S. Department of Labor § (2000).

Pasquini, A., Pozzi, S., Save, L., & Sujan, M.-A. (2017). Requisites for Successful Incident Reporting in Resilient Organisations. *Resilience Engineering in Practice*, 237–256. doi: 10.1201/9781317065265-17

Patriarca, R., Bergström, J., Gravio, G. D., & Costantino, F. (2018). Resilience engineering: Current status of the research and future challenges. *Safety Science*, *102*, 79–100. doi: 10.1016/j.ssci.2017.10.005

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. Retrieved from https://scikit-learn.org/stable/user_guide.html

Project Management Institute. (2004). A guide to the project management body of knowledge (PMBOK guide). Newtown Square, Pa: Project Management Institute.

Ransbotham, S., Kiron, D., Gerbert, P., & Reeves, M. (2017). Reshaping business with artificial intelligence: Closing the gap between ambition and action. MIT Sloan Management Review, 59(1).

Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning* (2nd Edition). Birmingham, UK: Packt Publishing.

Rowe, S., & Francois, J. M. (2016). Process safety data – The cornerstone of PSM and often it's undermining. *Journal of Loss Prevention in the Process Industries*, *43*, 736–740. https://doi.org/10.1016/j.jlp.2016.06.002

Sarkar, S., Vinay, S., Raj, R., Maiti, J., & Mitra, P. (2019). Application of optimized machine learning techniques for prediction of occupational accidents. *Computers & Operations Research*, *106*, 210–224. doi: 10.1016/j.cor.2018.02.021

Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms.

Srinivasa-Desikan, B. (2018). *Natural language processing and computational linguistics a practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Birmingham: Packt.

Steen, R., & Aven, T. (2011). A risk perspective suitable for resilience engineering. Safety Science, 49(2), 292-297.

Swuste, P., Theunissen, J., Schmitz, P., Reniers, G., & Blokland, P. (2016). Process safety indicators, a review of literature. *Journal of Loss Prevention in the Process Industries*, *40*, 162–173. https://doi.org/10.1016/j.jlp.2015.12.020

Taleb, N. N. (2007). *The black swan: the impact of the highly improbable*. USA: Random House.

Thomas, P., Bratvold, R. B., & Bickel, J. E. (2013). The Risk of Using Risk Matrices Decision Analysis View project A Generalized Sampling Approach for Multilinear Utility Functions Given Partial Preference Information View project The Risk of Using Risk Matrices. *Article in SPE Economics and Management*, (April 2015). https://doi.org/10.2118/166269-MS

Xu, Q., Xu, K., Li, L., & Yao, X. (2018). Safety assessment of petrochemical enterprise using the cloud model, PHA–LOPA and the bow-tie model. Royal Society open science, 5(7), 180212.

Zhao, J., Cui, L., Zhao, L., Qiu, T., & Chen, B. (2009). Learning HAZOP expert system by case-based reasoning and ontology. Computers & Chemical Engineering, 33(1), 371-378.

Zuboff, S. (2015). Big other: surveillance capitalism and the prospects of an information civilization. Journal of Information Technology, 30(1), 75-89.

APPENDIX A

Tables A-1 to A-6 show the confusion matrices and classification reports for the Linear SVC applied to consequence labels (Health/Safety, Environment, and Finance) and risk scores (actual and potential).

Table A- 1. Confusion matrices for Linear SVC used for predicting consequence labels of Health/Safety, Environment (Env), and Finance (Fin)

		Predicted					
		Health/S	Uncat (H/S)	Environment	Uncat (Env)	Finance	Uncat (Fin)
Actual	Health/S	829	561	0	0	0	0
	Uncat (H/S)	140	3235	0	0	0	0
	Env	0	0	292	233	0	0
	Uncat (Env)	0	0	12	4228	0	0
	Finance	0	0	0	0	3235	187
	Uncat (Fin)	0	0	0	0	701	642

Table A- 2. Classification reports for Linear SVC used for predicting consequence labels of Health/Safety, Environment (Env), and Finance (Fin)

Primary Label	Precision	Recall	F1-score	Support
Health/S.	0.86	0.60	0.7	1390
Uncat (H/S)	0.85	0.96	0.90	3375
Environment	0.96	0.56	0.70	526
Uncat.	0.95	1.00	0.97	4239
Finance	0.82	0.95	0.88	3422
Uncat.	0.77	0.48	0.59	1343

Table A- 3. Confusion matrix for Linear SVC used in actual risk score classification

		Predicted		
		0	4	5
Actual	0	12	0	339
	4	0	210	467
	5	23	12	3702

Table A- 4. Classification report for Linear SVC used in actual risk score classification

Actual Risk Score	Precision	Recall	F1-score	Support
0	0.33	0.03	0.06	351
4	0.95	0.31	0.47	677
5	0.82	0.99	0.90	3737

Table A- 5. Confusion matrix for Linear SVC used in potential risk score classification

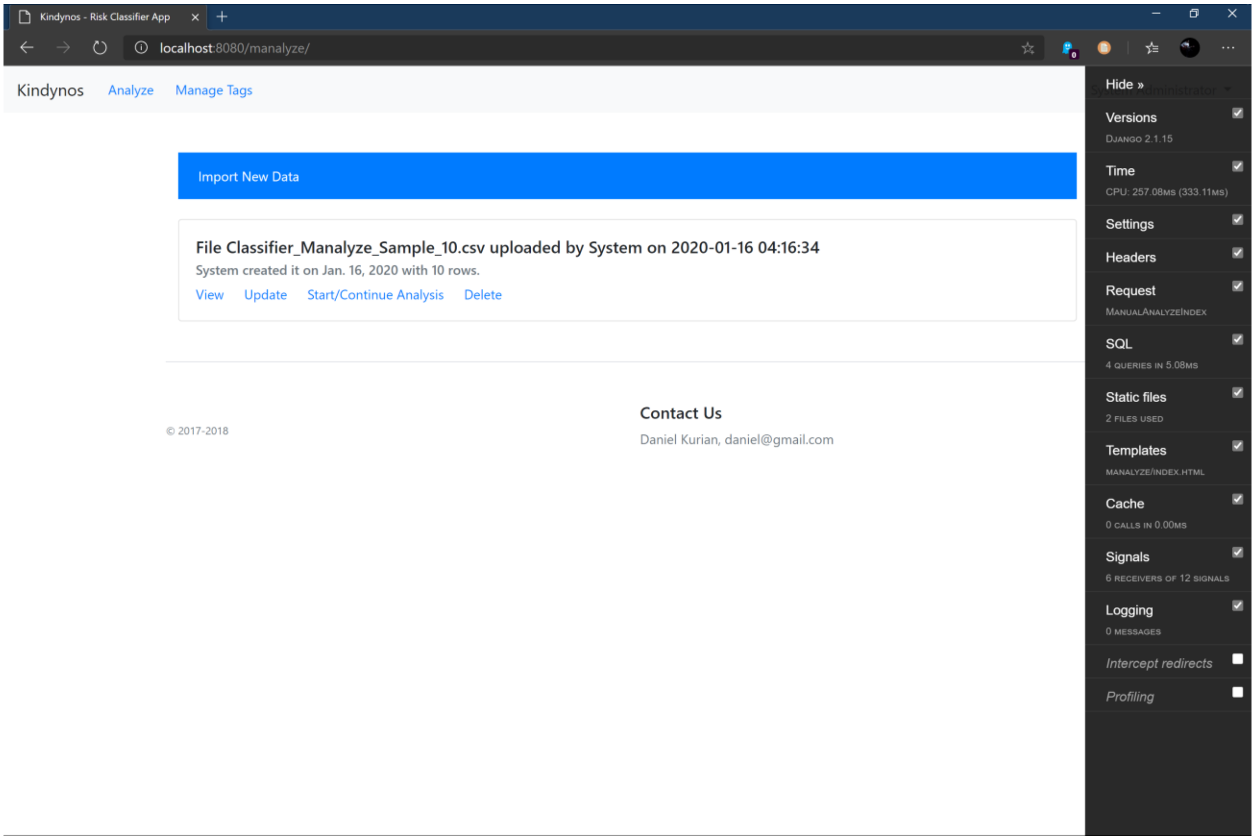
		Predicted				
		0	2	3	4	5
Actual	0	23	0	0	117	210
	2	0	12	0	0	12
	3	12	0	82	58	140
	4	12	0	0	1436	502
	5	0	0	12	467	1670

Table A- 6. Classification report for Linear SVC used in potential risk score classification

Potential Risk Score	Precision	Recall	F1-score	Support
0	0.50	0.07	0.12	350
2	1.00	0.50	0.67	24
3	0.88	0.28	0.42	292
4	0.69	0.74	0.71	1950
5	0.66	0.78	0.71	2149

APPENDIX B

Figures B-1 to B-4 show screenshots of how a user interface for reporting incidents could be designed.



File Classifier_Manalyze_Sample_10.csv uploaded by System on 2020-01-16 04:16:34

System created it on Jan. 16, 2020 with 10 rows.

© 2017-2018

Contact Us

Daniel Kurian, daniel@gmail.com

Versions ☑
Django 2.1.15

Time ☑
CPU: 257.08ms (333.11ms)

Settings ☑

Headers ☑

Request ☑
ManualAnalyzeIndex

SQL ☑
4 queries in 5.08ms

Static files ☑
2 files used

Templates ☑
manalyze/index.html

Cache ☑
0 calls in 0.00ms

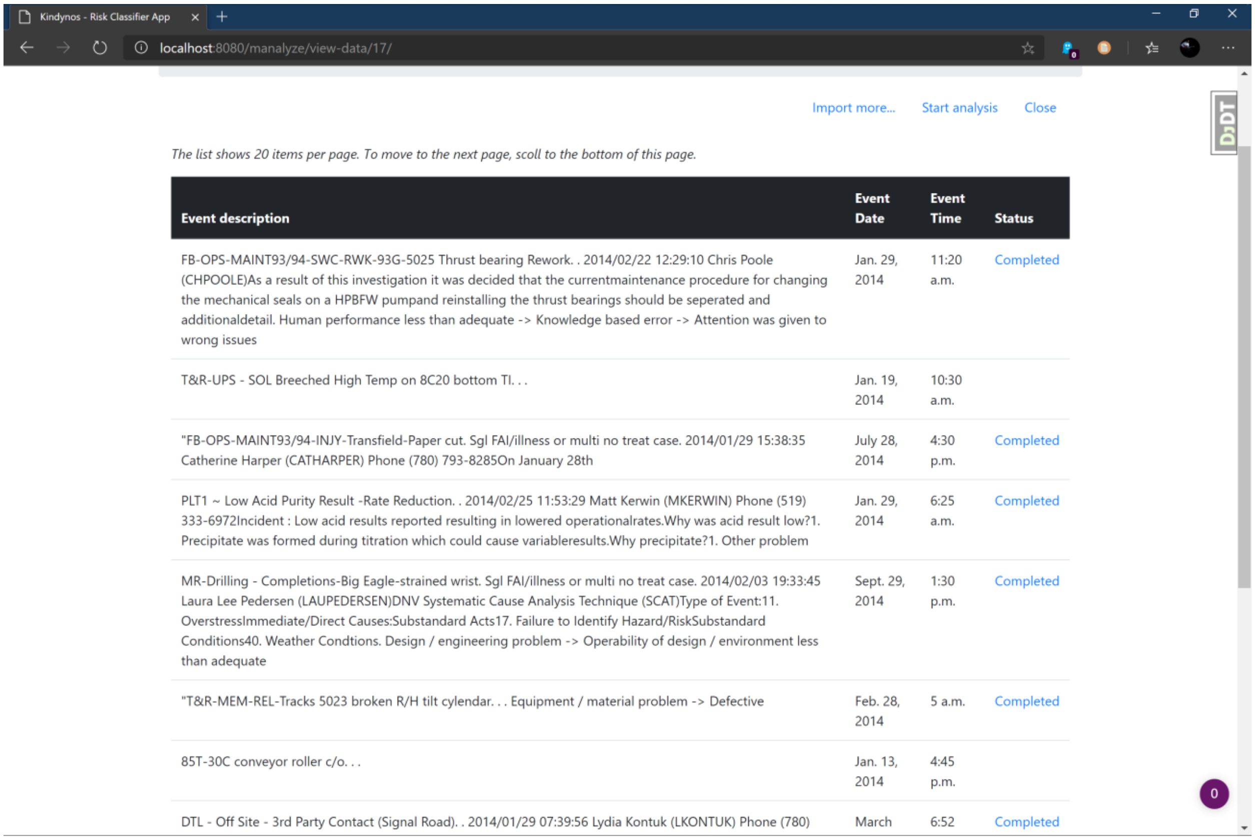
Signals ☑
6 receivers of 12 signals

Logging ☑
0 messages

Intercept redirects □

Profiling □

Figure B- 1. Screenshot of a user interface supporting the upload of incident reports in .csv format.

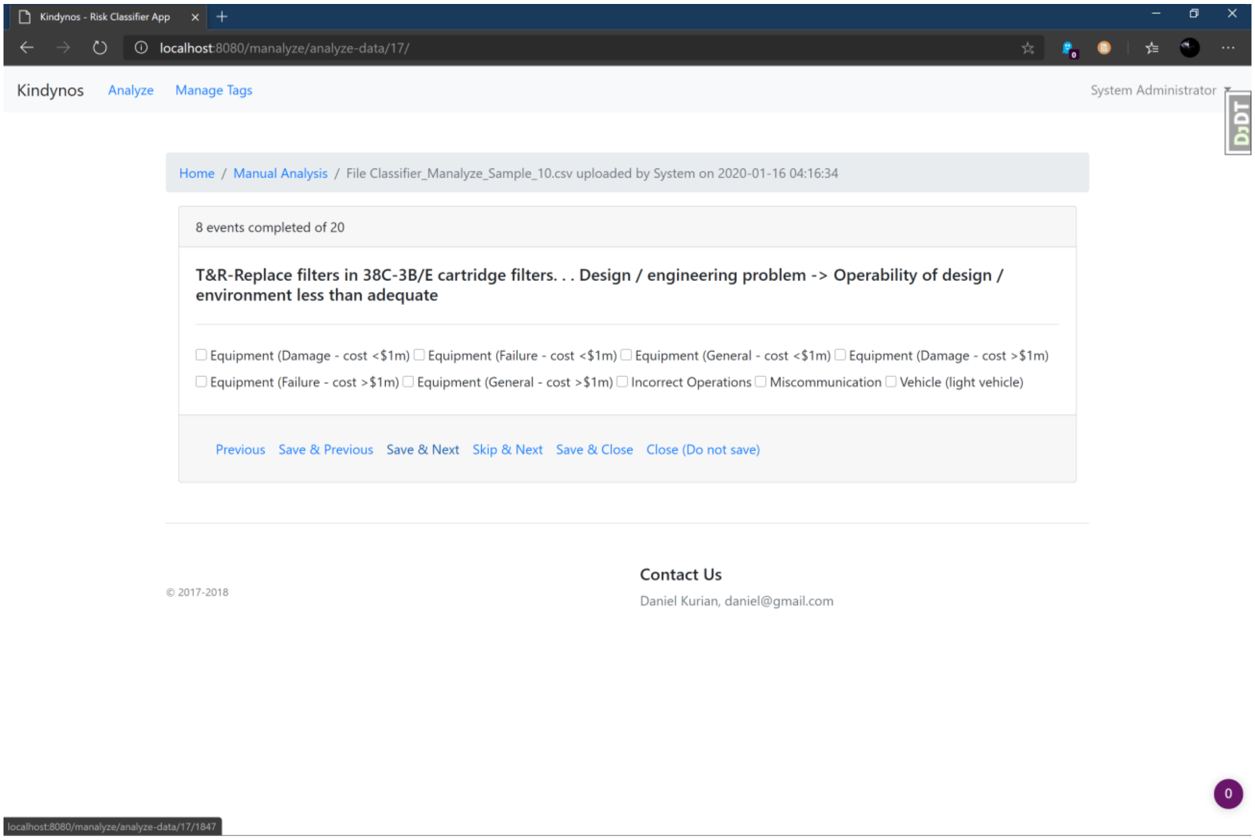


Kindynos - Risk Classifier App
localhost:8080/manalyze/view-data/17/
Import more...
Start analysis
Close
0

The list shows 20 items per page. To move to the next page, scoll to the bottom of this page.

Event description	Event Date	Event Time	Status
FB-OPS-MAINT93/94-SWC-RWK-93G-5025 Thrust bearing Rework. . 2014/02/22 12:29:10 Chris Poole (CHPOOLE)As a result of this investigation it was decided that the currentmaintenance procedure for changing the mechanical seals on a HPBFW pumpand reinstalling the thrust bearings should be seperated and additionaldetail. Human performance less than adequate -> Knowledge based error -> Attention was given to wrong issues	Jan. 29, 2014	11:20 a.m.	Completed
T&R-UPS - SOL Breeched High Temp on 8C20 bottom TI. . .	Jan. 19, 2014	10:30 a.m.	
"FB-OPS-MAINT93/94-INJY-Transfield-Paper cut. Sgl FAI/illness or multi no treat case. 2014/01/29 15:38:35 Catherine Harper (CATHARPER) Phone (780) 793-8285On January 28th	July 28, 2014	4:30 p.m.	Completed
PLT1 ~ Low Acid Purity Result -Rate Reduction. . 2014/02/25 11:53:29 Matt Kerwin (MKERWIN) Phone (519) 333-6972Incident : Low acid results reported resulting in lowered operationalrates.Why was acid result low?1. Precipitate was formed during titration which could cause variableresults.Why precipitate?1. Other problem	Jan. 29, 2014	6:25 a.m.	Completed
MR-Drilling - Completions-Big Eagle-strained wrist. Sgl FAI/illness or multi no treat case. 2014/02/03 19:33:45 Laura Lee Pedersen (LAUPEDERSEN)DNV Systematic Cause Analysis Technique (SCAT)Type of Event:11. OverstressImmediate/Direct Causes:Substandard Acts17. Failure to Identify Hazard/RiskSubstandard Conditions40. Weather Condtions. Design / engineering problem -> Operability of design / environment less than adequate	Sept. 29, 2014	1:30 p.m.	Completed
"T&R-MEM-REL-Tracks 5023 broken R/H tilt cylendar. . . Equipment / material problem -> Defective	Feb. 28, 2014	5 a.m.	Completed
85T-30C conveyor roller c/o. . .	Jan. 13, 2014	4:45 p.m.	
DTL - Off Site - 3rd Party Contact (Signal Road). . 2014/01/29 07:39:56 Lydia Kontuk (LKONTUK) Phone (780)	March	6:52	Completed

Figure B- 2. Screenshot of a user interface showing how uploaded incidents could look along with the status of analysis


Kindynos - Risk Classifier App
localhost:8080/manalyze/analyze-data/17/
Kindynos
Analyze
Manage Tags
System Administrator
Home / Manual Analysis / File Classifier_Manalyze_Sample_10.csv uploaded by System on 2020-01-16 04:16:34
Previous
Save & Previous
Save & Next
Skip & Next
Save & Close
Close (Do not save)
Contact Us
© 2017-2018
Daniel Kurian, daniel@gmail.com
0
localhost:8080/manalyze/analyze-data/17/1847

8 events completed of 20

T&R-Replace filters in 38C-3B/E cartridge filters. . . Design / engineering problem -> Operability of design / environment less than adequate

□ Equipment (Damage - cost <$1m) □ Equipment (Failure - cost <$1m) □ Equipment (General - cost <$1m) □ Equipment (Damage - cost >$1m)
□ Equipment (Failure - cost >$1m) □ Equipment (General - cost >$1m) □ Incorrect Operations □ Miscommunication □ Vehicle (light vehicle)

Figure B- 3. Screenshot of user interface for selecting statements that match an incident.

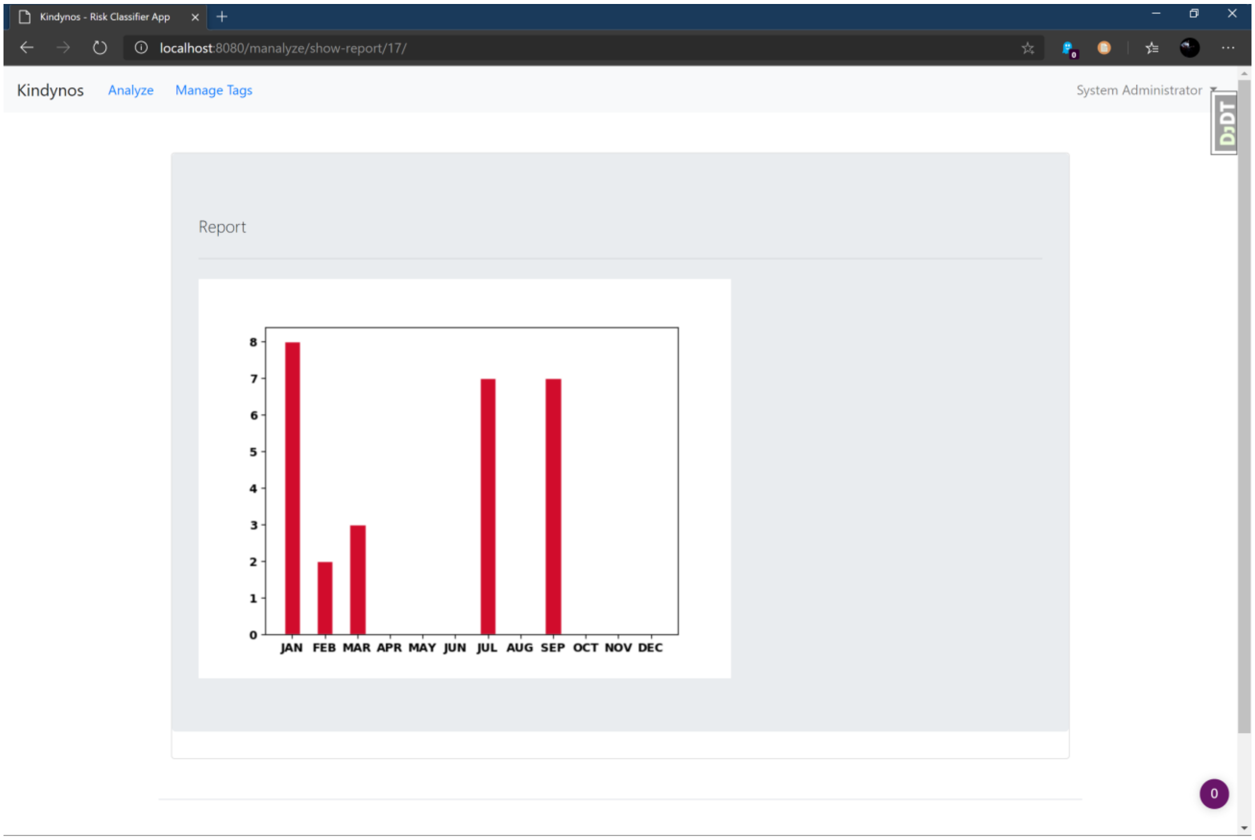


Figure B- 4. Screenshot of user interface displaying a trend report.