SafeOne Machine Learning model to predict industrial incidents in Chemical and Gas Industries

Ganapathy Subramaniam Balasubramanian[a], and Dr. Ramaprabha Thangamani[b]

[A] Research Scholar, PG and Research Department of Computer Science and Applications, Vivekanandha College of Arts and Sciences for Women (Autonomous), Elayampalayam, Tamilnadu, India.

[b]Professor, PG and Research Department of Computer Science and Applications, Vivekanandha College of Arts and Sciences for Women (Autonomous), Elayampalayam, Tamilnadu, India.

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: Understanding activity incidents is one of the necessary measures in workplace safety strategy. Analyzing the trends of the activity incident information helps to spot the potential pain points and helps to scale back the loss. Optimizing the Machine Learning algorithms may be a comparatively new trend to suit the prediction model and algorithms within the right place to support human helpful factors. This research aims to make a prediction model spot the activity incidents in chemical and gas industries. This paper describes the design and approach of building and implementing the prediction model to predict the reason behind the incident which may be used as a key index for achieving industrial safety specific to chemical and gas industries. The implementation of the grading algorithmic program including the prediction model ought to bring unbiased information to get a logical conclusion. The prediction model has been trained against incident information that has 25700 chemical industrial incidents with accident descriptions for the last decade. Inspection information and incident logs ought to be chomped high of the trained dataset to verify and validate the implementation. The result of the implementation provides insight towards the understanding of the patterns, classifications, associated conjointly contributes to an increased understanding of quantitative and qualitative analytics. Innovative cloud-based technology discloses the gate to method the continual in-streaming information, method it, and output the required end in a period. The first technology stack utilized in this design is Apache Kafka, Apache Spark, KSQL, Data frames, and AWS Lambda functions. Lambda functions are accustomed implement the grading algorithmic program and prediction algorithmic program to put in writing out the results back to AWS S3 buckets. Proof of conception implementation of the prediction model helps the industries to examine through the incidents and can layout the bottom platform for the assorted protective implementations that continuously advantage the workplace's name, growth, and have less attrition in human resources.

Keywords: Occupational incidents, Prediction Model, Machine Learning, Occupational Safety.

1. Introduction

All workman who leaves their home for the work ought to return to home safe and sound. Thinking of the state of affairs otherwise, forever showing emotional sensitivity. Particularly within the field of chemical and gas industries, the incidents not solely affect the individual also the environment terribly. The impact would be there for years, typically decades. Generic machine learning algorithms, most of the time, demands a lot of parameters and have shortfalls to implement the precise want that doesn't work for all specific industries and organizations to supply the expected leads to a given timeline. As well as the assorted industry-specific factors into machine learning algorithms will offer advantageous impact for chemical and gas industries by reduced expenses, exaggerated productivity, improved work strategies. Analysis of business incidental safety measures seems to be the weakest part of the economic safety management system.

The Categorial Scoring Model and SafeOne Prediction Model based on Support Vector Machines (SVM) developed for prediction of incidents, positively want a design to urge through the suitable implementation. Inflow information ought to perpetually monitor to work out the precise score and supply the expected output. Developing the proof of construct (POC) can facilitate the organization to see-through the potential outcome of the answer and additionally helps to spot the gaps in it. It will additionally offer the stakeholders to internally measure the promising resolution that helps to scale back the gratuitous risk. Design expectations and potential timeline can even be determined before the all-out implementation. Applying an outlined algorithmic program is not a simple task. As a section of POC, it is necessary to create a visual interface to check the most effective attainable results. The approach is needed to be quantitative so that to describe the usefulness of the measurement rates towards the calculation of precision and accuracy. The accuracy score focuses on the outcome of the measurement rates to help the organizations in decision-making and also paves the path to eliminate occupational incidents.

The remaining of the paper is organized as, Section 2 lists out the review of the key literature work done by researchers and scholars in the field of workplace safety. Section 3 defines the research methodology and Section 4 explains the development and implementation process of the work. Section 5 discusses the results and compares the performance of the model against the other models and concludes the paper by describing the summary and directions of the future work.

2. Review of Literature

The internal-external locus of control theory, developed by Rotter in 1966, was one of the first psychological constructs examined as a possible predictor of accident potential. There has been much success in using this construct as a means for predicting involvement in accidents. Christopher A. Janicak, in 1996 published a study about predicting the accidents at work with measures of locus of control and job hazards [33]. The study analyzed the accident locus of control scale items which are very useful to measure the level of a job hazard. Christopher A Janicak resulted in his findings through the locus of control score combined with the level of job hazard score which produced 89% accuracy on accidents and 70% on non-accidents. When using the level of job hazard only as a parameter, it produced 79% accuracy. As same, when using locus of control score only, the model produced 86% accuracy on accidents and 43% on non-accidents. Ronza A. et al, in 2003, developed a methodology to describe a frequency value to the sequence scenarios, by multiplying the probability of occurrence by the frequency of the root event. The reliability of this procedure is proved by a wide range of historically documented accidents [56].

Martine Reurings and Theo Janssen, in 2006, developed a project Infrastructure and Road Safety aimed to find(mathematical) relations between characteristics of the Dutch road infrastructure and road safety. Research describes the relations between characteristics of the Dutch road infrastructure on the one hand and road safety, on the other hand, using risk and exposure measures. Models are the subject of Work package 2 of RIPCORD-ISEREST, which started with making an overview of the state-of-the-art on accident prediction models and road safety impact assessments [54]. Dipo T. Akomolafe and Akinbola Olutayo, in 2012, explained the use of the data mining technique to predict the cause of the accident and accident-prone locations on highways. Experiments were done using decision tree algorithms Id3 and FT (Function Tree) to determine the cause. From the detailed accuracy by class and confusion matrix, Id3 attained an accuracy rate of 0.777 and FT attained an accuracy rate of 0.703 [66].

Jan K. Wachter and Patrick L. Yorio, in 2013, theoretically and empirically develop the ideas around a system of safety management practices, ten practices were elaborated to test their relationship with objective safety statistics such as accident rates, and to explore how these practices work to achieve positive safety results which are accident prevention through worker engagement. Results indicated that there is a significant negative relationship between the presence of ten individual safety management practices, as well as the composite of these practices, with accident rates; there is a significant negative relationship between the level of safety-focused worker emotional and cognitive engagement with accident rates; safety management systems and worker engagement levels can be used individually to predict accident rates; safety management systems can be used to predict worker engagement levels, and worker engagement levels act as mediators between the safety management system and safety performance outcomes [69].

The study of the related papers provided clarity on the work and approaches that have been done earlier not specifically on the computation field but also in the various fields including psychology, mathematics, civil, human studies, reveals that there are limited works in extracting classified knowledge of incidents from the semi-structured inspection data. Reviewing the literature provides insight on the incidents but statistically did not contribute much to determine the prediction model. This research attempts to bridge the gaps of using semi-structured, multi-variate inspecting data by leveraging a vector-based classification model inspired by the principles of grid-search as a powerful tool to determine the prediction model for the given inspection data. This research intends to contribute to occupational safety by determining workplace safety to predict workplace incidents.

3. Research Methodology

This paper describes the methodology of predicting workplace incidents through the three processes in general. Data refinement, Categorial Safeness Score, and Prediction Model.

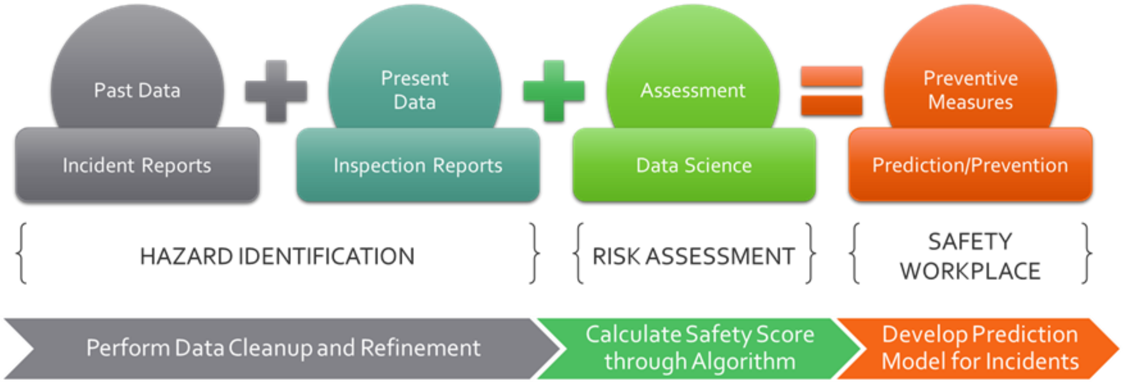


Figure 1. Research methodology to predict workplace incidents

Lack of an effective data analytics procedure from industrial historical incidents data, inspection/auditing dataset, and risk assessment dataset, leads to an unsafe workplace. Since data analytics and machine learning algorithms are too linear and too sparse where most of the time makes the algorithm overfits or underfits the requirement. To prevent fatal/non-fatal incidents, an efficient implementation of predictive analytics is expected. Predictive analytics not only includes the optimized data mining algorithm but also the identification of the right factors to make the precise prediction.

3.1. Data Refinement Model

Problem solutions should provide a viable course of action and form the basis for implementing and achieving the objectives and control measures. Employees and workers who are too close to the sequences may no longer perceive and recognize the hazards, or perhaps judge the incidents as trivial because to their knowledge no one has been harmed. The aim of the problem-solution should be that everyone tackles the scenario with a fresh pair of eyes and a questioning approach.

Main data cleansing is required to determine the cause of the incident for the incidents which are defined as unknown causes. Text Mining would be the best candidate as a problem solution to figure out the best match for predicting the cause of the incident. A multi-factor classification algorithm should be implemented to determine the weightage against the possible causes of the incident and the same can be used to decide upon the cause of the incident. Occurrences of the past data with the factors and defined cause serve as a base and the model will be trained with the same to determine the result. Cause factors are classified, and weightage will be calculated for each cause. Update the learning parameter at every calculation of the incident. Predict the cause of the incident based on the weight calculated using the learning parameter.

3.2. Categorial Safeness Score

Methodology to determine the safeness score involves the analysis of the historical safety incident data. The analysis crawls through the data and determines the score of the category which will be used as a key parameter in the prediction model. An improvised vector-based data mining algorithm inspired by the principles of co-ordinate descent can be implemented for each category of the safety measure to determine the score of the safety category. The score should be calculated in percentage with a value between 0 and 100. Scores are calculated based on the data collected for the past 30 days. Scores are calculated in a specified frequency (once a day, off-peak hours) as determined by the organization at their requirements and convenience. Overall Score is a weighted average of every category identified for the industry. Safe Score can be defined and classified as per the following ranges: 86 – 100: Overall Safe (Green), 60 – 85: Situation/Category requires attention (Amber), Below 59: Potential Safety Issue (Red).

3.3. SafeOne Prediction Model

The primary objective and the ultimate goal of this research to derive the prediction model to attain the unsafe percentage of the department or the industry with the use of the safety scores being transformed from information to knowledge. Analyzing the trends of the occupational incident data helps to identify the potential pain points and helps to reduce the loss. Optimizing the grid search-based machine learning model will fit in the right place to support human beneficial factors. Implementation of new algorithms and new models have a similar step-up process to verify and validate the real-time scenarios. Understanding the mathematical calculations of measurement rates helps to place the model that the industry demands. Incident rates, Lost time cases rate, Severity rate, and lost workday rate are important calculations to make sure the model complies with. Chemical and gas industries worldwide have got the potential risk of occupational hazards which leads to the incidents. A model to predict the incidents based on the inspection and incident data helps the industry by eliminating and/or at least reduce the incident rates.

The SafeOne prediction model has been constructed to predict the potential unsafe percentage value of the occupational incidents in the industry. The trained model takes care of applying the verification and validation of the model to write the prediction value. Safety scores are the inputs to this model and are weighed in different categories. Impact factors are applied to predict the Unsafe Percentage. SafeOne prediction model delivers the prediction for the next 3 to 7 days according to the industry under implementation through an expert assessment.

$$f(x_1) = \frac{w_{p1} \cdot |A_i\{i \leq score_{min}\}| \cdot 10^2 - \sum A_i\{i \leq score_{min}\}}{10^1}$$

$$f(x_2) = \frac{w_{p2} \cdot |A_i\{i \leq score_{max}\}| \cdot 10^2 - \sum A_i\{i \leq score_{max}\}}{10^2}$$

$$f(x_3) = \frac{w_{p3} \cdot |A_i\{score_{min} \leq i \leq score_{max}\}| \cdot 10^2 - \sum A_i\{score_{min} \leq i \leq score_{max}\}}{10^3}$$

$$f(y_1) = sup\ \{0.2, |A_i\{i \leq score_{min}\}|\}$$

$$f(SafeOne) = \frac{f(x_1) + f(x_2) + f(x_3)}{f(y_1)}$$

where wp1, wp2, wp3 represent the weights defaults to 9, 3, 1 respectively. A represents the data set, the score is an algorithm applied value to the data points determined using the Scoring algorithm. The SafeOne model predicts the "Unsafe" percentage concerning all the factors considered during the machine learning process.

SafeOne prediction model is being continuously trained using data streams which are from the industrial inspection data, observation data from reports feed. The factors considered for the classifications and models are given for the manipulation of data. Based on the data, the SafeOne prediction model has been trained to predict the potential value of occupational incidents in the industry. Once the model is trained, a test set of data has been applied for verification and validation of the model to write the prediction value. The outcome of the Prediction model should be validated against the known results to verify against the obtained result in the past. The machine learning model has been trained in such a way to produce the result from the scores through the learning parameters for the number of the specified next days.

4. Implementation

4.1. Working Model Architecture

The architecture of the working model, as shown in Figure 2, detailed the components involved for a better outcome. The integration between these components is aligned in such a way to establish a scalable solution for the future data load. Starting from the data stream through getting the outcome of the prediction model, cloud infrastructure helps to deliver a reliable solution.

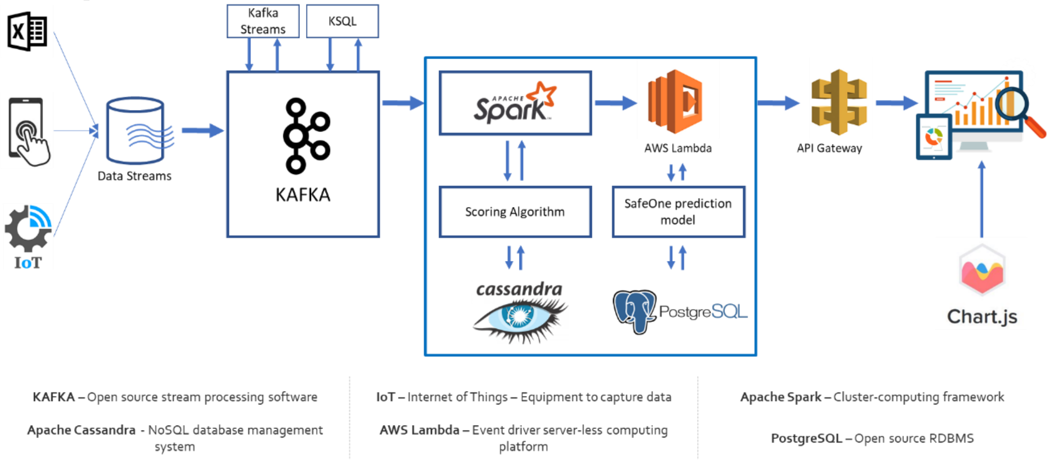

Kafka Streams
KSQL
IoT
Data Streams
KAFKA
Spark
AWS Lambda
Scoring Algorithm
SafeOne prediction model
cassandra
PostgreSQL
API Gateway
Chart.js
KAFKA – Open source stream processing software
Apache Cassandra - NoSQL database management system
IoT – Internet of Things – Equipment to capture data
AWS Lambda – Event driver server-less computing platform
Apache Spark – Cluster-computing framework
PostgreSQL – Open source RDBMS

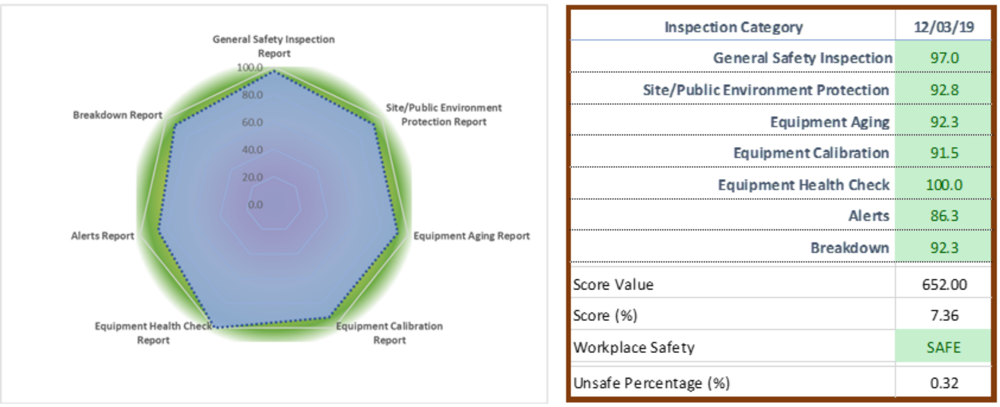
Figure 2. High-level architecture

Inspection data, sensor data from the gas detection instrument, and historical incidents should be streamed through Kafka. Data Streaming is a method of posting a continuous stream of data that can be processed through the algorithms to obtain structural data. Multiple sources can send the data simultaneously to meet the requirements of real-time data analytics. The continuous stream of data is put in a bucket called a topic. Topics in Kafka can be subscribed by the consumer program to stream for processing. These topics are partitioned based on the size and volume of speed and scalability. Data are sent by various data sources to topics and subscribed consumer application takes care of relaying it. Each partition is assigned to a Kafka Broker for parallel processing. Messages are typically key-value pairs to construct the structural data. The stream is divided into RDDs (Resilient Distributed Datasets) which is a fundamental data structure of Spark. RDDs are divided into partitions which consist of tuples. The worker node takes care of processing the data in the Spark. Kafka-Spark connector allows mapping partitions between RDD and Kafka topic.

Processing of the data takes place through Spark Jobs. Spark Jobs is the small set of programs that cleans up, manipulates, and applies the specific algorithm to the data streamed and stored into the data lake. A data lake is a collection of data frames stored in the storage bucket. Spark Jobs written using Scala language in Notebook executes the Scoring algorithm to refine and restructure the data which should be used as an input for the SafeOne prediction model. Lambda functions serve the purpose of executing the logic using the structured data to provide the expected outcome. The approach of incremental algorithms can be used to manipulate the history data and real-time data. Heatmap representation of the data can be generated from the algorithm to visualize the results. The data dashboard displays the required heatmap and also keeps the data live through push notifications.

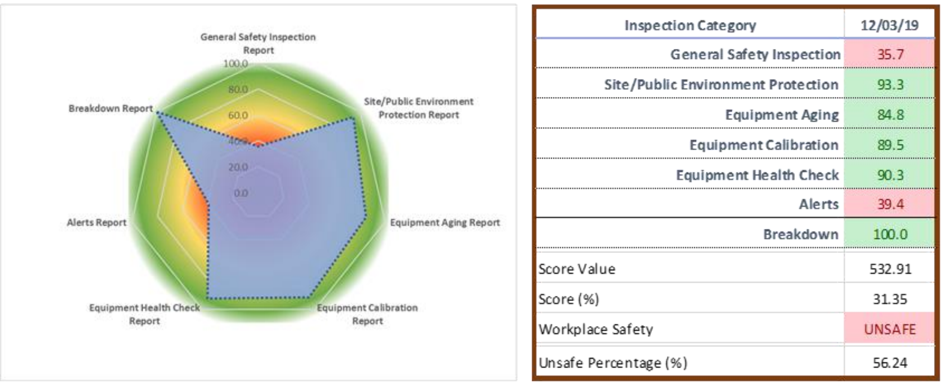
5. Results and Discussions

Graphing methods vary according to the scales of measurements and presentation. Evaluation of the categorized inspection score determines the safe score from the model where the radar graph, also called a spider web graph, is used to plot the scores of each inspection type. Value 0 is the safest zone and value 100 being the unsafe zone on the radar. A “Safe” and “Unsafe” radar graph representations of the processed inspection data with its score value obtained from the SafeOne prediction model are shown in Figure 8.3 and Figure 8.4 respectively along with table data including the unsafe percentage.



Inspection Category	12/03/19
General Safety Inspection	97.0
Site/Public Environment Protection	92.8
Equipment Aging	92.3
Equipment Calibration	91.5
Equipment Health Check	100.0
Alerts	86.3
Breakdown	92.3
Score Value	652.00
Score (%)	7.36
Workplace Safety	SAFE
Unsafe Percentage (%)	0.32

Figure 3. “Safe” radar representation of the workplace with table data



Inspection Category	12/03/19
General Safety Inspection	35.7
Site/Public Environment Protection	93.3
Equipment Aging	84.8
Equipment Calibration	89.5
Equipment Health Check	90.3
Alerts	39.4
Breakdown	100.0
Score Value	532.91
Score (%)	31.35
Workplace Safety	UNSAFE
Unsafe Percentage (%)	56.24

Figure 4. “UnSafe” radar representation of the workplace with table data

Training data was split into 70% which is 754,660 rows of training set data and 229,000 rows for the test set. An overall error has also been calculated from the results as a part of the prediction model calculation. These results are significant and provide a positive look forward solution to practice safety in organizations to prevent potential incidents. This working model provides the basic idea of visualizing the results in real-time by setting the base platform to smoothly walk through the workflow from raw data to prediction results. The outcome of the Prediction model is validated against the known results to verify against the obtained result in the past. Historical data of the prediction score is plotted to visualize the Safe and Unsafe values for the organization as shown in Figure 5. This has been leveraged to extrapolate the trend.

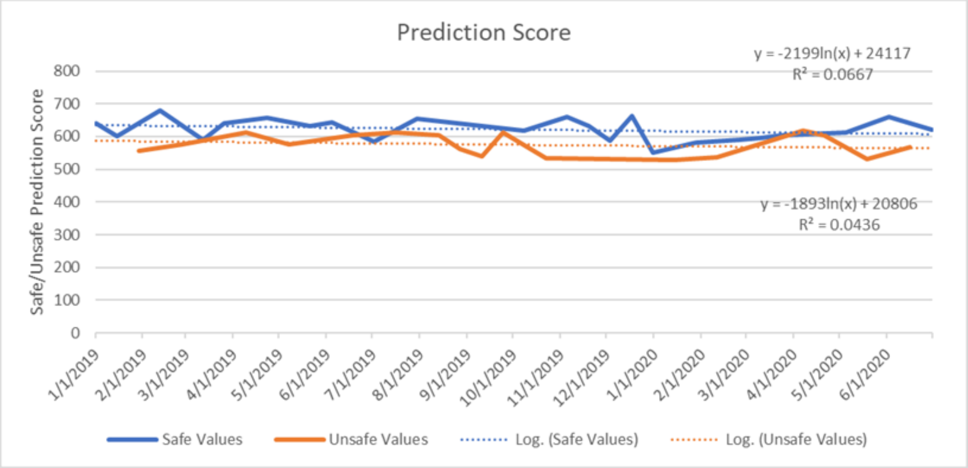


Figure 5. Prediction Score with Trend value

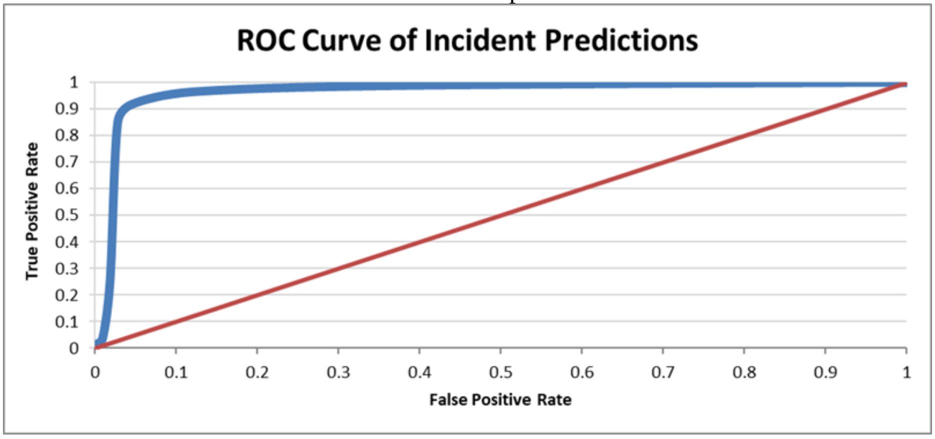
Evaluation of the model needs to compare against the existing models for appraising the proposed model. The results should be compared against the previous or existing models to prove the point of quality this research has achieved. The predictive analysis study has been done to compare the results against the Random Forest (RF) and Stochastic Gradient Descent (SGD) against the proposed classifier model. Comparison of performance measures Precision, Recall, F1 Score, and AUC (Area Under Curve) gives a clear evaluation of the model under validation and verification. The proposed Categorial Safeness Scoring model resulted in a dominant state across all the performance measures comparison.

	Fail (Actual)	Pass (Actual)	Total
Fail (Prediction)	104	13	117

Pass (Prediction)	3	74	77
Total	107	87	194
Accuracy	97%	85%	92%

Table 1. Accuracy Score

Summing across rows yields the number of total incidents with an actual positive state while summing in a column yields the total number of times that the corresponding decision was made. High results indicate the occurrence of the incident but the distributions of test result values in UNSAFE and SAFE incidents overlap then increasing the threshold value will make both false positive and true positive predictions less frequent. But the model will consider both true negative and false negative predictions more frequent. ROC curve appears much closer to the upper left corner means that the model has a highly efficient accuracy score. Table 1 shows the accuracy score obtained from the measurement done from the prediction model.



A threshold value should be determined to yield a compromise among these trade-offs which is between SAFE and UNSAFE incident results. Results closely match the objective of bringing up the accuracy value to predict occupational incidents in the chemical and gas industries. ROC Curve looks as shown in figure 6 which is an expected accuracy score of 92% as determined from the prediction model.

6. Conclusion

Contributions of the research, progressively, predicting the cause of the incident which helps the industries to have the cleaned-up data for the better placement of the score. The scoring algorithm helps to identify the safe score in each category and the SafeOne prediction model takes care of calculating the overall Safe Score based on the category scores. A score set of the inspection data has been obtained through the Scoring algorithm which serves as an input for the prediction model. Result visualization has been depicted using the heatmap, scattergram, and radar graph along with the prediction data defined to determine the safe and unsafe percentage. Inspection and Incidents are tightly coupled data chunk which brings the prediction mechanism to the spotlight. The objective of this study is to eliminate occupational incidents by providing an efficient and robust prediction model with high accuracy. Successful implementation of the accuracy score into the SafeOne prediction model proves that the prediction model works well in predicting incidents. Comparison study upon several models determines the accuracy score between them and put the SafeOne at the top.

The outcome of the research provides insight towards the understanding of the patterns, classifications, and also contributes to an enhanced understanding of quantitative and qualitative analytics. Cutting edge cloud-based technology opens the gate to process the continuous in-streaming data, process it, and output the desired result in real-time. The research contributes helps the industries to see through the incidents and will layout the base platform for the various safety-related implementations which always benefits the workplace's reputation, growth, and have less attrition in human resources. The number of auditing/inspection reports is directly proportional to workplace safety. Organizations that make employees in identifying unsafe workplaces have fewer incidents. The results of this research are to determine the best-suited algorithm, applied to workplace safety, which intends to send every employee home safe, at the end of every day. After all, if workplace incidents can be predicted, they can be prevented.

References

1. *Alkheder, S., Taamneh, M., & Taamneh, S. (2017). Severity Prediction of Traffic Accident Using an Artificial Neural Network. Journal of Forecasting, 36(1), 100–108. https://doi.org/10.1002/for.2425*
2. *Aloysius, G., &Binu, D. (2013). An approach to product placement in supermarkets using PrefixSpan algorithm. Journal of King Saud University - Computer and Information Sciences, 25(1), 77–87. https://doi.org/10.1016/j.jksuci.2012.07.001*
3. *Altman, N., &Krzywinski, M. (2015). Points of Significance: Simple linear regression. Nature Methods, 12(11), 999–1000. https://doi.org/10.1038/nmeth.3627*

4. *Amit Jha, ZakiHaidar Zaidi, Sridhar Ramaswamy, Sumeet Gupta, Ankit Gupta, B. S. (2017). 'Information& Cyber Insecurity' are the biggest risks in business operations. FICCI – Pinkerton India Risk Survey 2017.*
5. *Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., Pfau, D., Schaul, T., & de Freitas, N. (2016). Learning to learn by gradient descent by gradient descent. NIPS, 1–16. https://doi.org/10.1007/s10115-008-0151-5*
6. *Angra, A., & Gardner, S. M. (2018). The graph rubric: Development of a teaching, learning, and research tool. CBE Life Sciences Education, 17(4), 1–18. https://doi.org/10.1187/cbe.18-01-0007*
7. *Baranyi, J., & Buss da Silva, N. (2017). The use of predictive models to optimize risk of decisions. International Journal of Food Microbiology, 240, 19–23. https://doi.org/10.1016/j.ijfoodmicro.2016.10.016*
8. *Barnes, J. (2015). Getting started with Azure Machine Learning. Azure Machine Learning Microsoft Azure Essentials, 25–37. https://doi.org/10.1111/j.2041-210X.2010.00056.x*
9. *Ben-David, S., & Shalev-Shwartz, S. (2014). Understanding Machine Learning: From Theory to Algorithms. In Understanding Machine Learning: From Theory to Algorithms. https://doi.org/10.1017/CBO9781107298019*
10. *Bertke, S. J., Meyers, A. R., Wurzelbacher, S. J., Measure, A., Lampl, M. P., & Robins, D. (2016). Comparison of methods for auto-coding causation of injury narratives. Accident Analysis & Prevention, 88, 117–123. https://doi.org/10.1016/ J.AAP.2015.12.006*
11. *Beshah, T., & Hill, S. (2010). Mining road traffic accident data to improve safety: Role of road-related factors on accident severity in Ethiopia. AAAI Spring Symposium - Technical Report, SS-10-01(1997), 14–19.*
12. *Bhulai, S. (2015). Nearest neighbour algorithms for forecasting call arrivals in call centers. Smart Innovation, Systems and Technologies, 39, 77–87. https://doi.org/10.1007/978-3-319-19857-6_8*
13. *Bishop, C. M. (2013). Pattern Recognition and Machine Learning. In Journal of Chemical Information and Modeling (Vol. 53). https://doi.org/10.1117/1.2819119*
14. *Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., &Scuse, D. (2014). WEKA Manual for Version 3-7-12. 1–327.*
15. *Bureau of Indian Standards. (2007). Occupational Health and Safety Management Systems. 2000, 1–28. https://doi.org/10.1002/0471435139.hyg049.pub2*
16. *Cessna, J., Colburn, C., & Bewley, T. (2008). EnVE : A new estimation algorithm for weather forecasting and flow control. (June).*
17. *Chelly, S. M., & Denis, C. (2001). Applying Unsupervised Learning. Medicine and Science in Sports and Exercise, 33(2), 326–333. https://doi.org/10.1111/j.2041-210X.2010.00056.x*
18. *Chuahan, N. (2013). Safety and Health Management System in Oil and Gas Industry. 12.*
19. *Cichosz, P. (2015). Data Mining Algorithms. 1–5. https://doi.org/10.1002/ 9781118950951*
20. *Danjuma, K., & Osofisan, A. O. (2015). Evaluation of Predictive Data Mining Algorithms in Erythemato-Squamous Disease Diagnosis. (Cvd), 10.*
21. *Davoudi Kakhki, F., Freeman, S. A., & Mosher, G. A. (2019). Evaluating machine learning performance in predicting injury severity in agribusiness industries. Safety Science, 117(July 2018), 257–262. https://doi.org/10.1016/j.ssci.2019.04.026*
22. *De Saulles, M. (2016). The Internet of Things and Business. The Internet of Things and Business, 1–99. https://doi.org/10.4324/9781315537849*
23. *Dietterich, T. G. (2009). Machine learning in ecosystem informatics and sustainability. In IJCAI International Joint Conference on Artificial Intelligence. https://doi.org/10.1007/978-3-540-75488-6_2*
24. *Drymonitis, A. (2015). Introduction to Arduino. In Digital Electronics for Musicians. https://doi.org/10.1007/978-1-4842-1583-8_2*
25. *Eitrich, T., & Lang, B. (2006). Efficient optimization of support vector machine learning parameters for unbalanced datasets. Journal of Computational and Applied Mathematics, 196(2), 425–436. https://doi.org/10.1016/j.cam.2005.09.009*
26. *Gueniche, T., Fournier-viger, P., Raman, R., & Tseng, V. S. (2015). CPT + : A Compact Model for Accurate Sequence Prediction.*
27. *Gupta, M., & Han, J. (2012). Approaches for Pattern Discovery Using Sequential Data Mining. Pattern Discovery Using Sequence Data, 1–20. https://doi.org/10.4018/978-1-4666-2455-9.ch095*
28. *Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. In San Francisco, CA, itd: Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00001-0*
29. *Helvert, M. van. (2014). Data visualizations in popular Dutch media.*
30. *Hirate, Y., &Yamana, H. (2006). Generalized sequential pattern mining with item intervals. Journal of Computers, 1(3), 51–60. https://doi.org/10.4304/jcp.1.3.51-60*
31. *Hrymak, V., & Pérezgonzález, J. D. (2007). The costs and effects of workplace accidents Twenty case studies from Ireland. Health and Safety Authority Research Series 02/2007, (March), 1–138.*
32. *Iba, T., Miyake, T., Naruse, M., &Yotsumoto, N. (2009). Learning Patterns: A Pattern Language for Active Learners. Proceedings of the 16th Conference on Pattern Languages of Programs, PLoP'09.*

33. *Janicak, Christopher. A. (1996). Predicting accidents at work. 115–121.*
34. *Kaur, P., Singh, M., &Josan, G. S. (2015). Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector. Procedia Computer Science, 57, 500–508. https://doi.org/10.1016/j.procs.2015.07.372*
35. *Khan, F. I., Abbasi, S. ., Mpp, C. De, & European Union. (2006). Techniques and methodologies for risk analysis in chemical process industries. In Journal of Loss Prevention in the Process Industries (Vol. 11). https://doi.org/10.2790/73321*
36. *Kisan, M., Sangathan, S., Nehru, J., & Pitroda, S. G. (2007). Occupational Health and Safety Management Systems--Requirements with Guidance for Use. ICS 13.100(Bureau of Indian Standards).*
37. *Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal Of Statistical Software, 28(5), 1–26. https://doi.org/10.1053/j.sodo.2009.03.002*
38. *Laska, M., Herle, S., Klamma, R., &Blankenbach, J. (2018). A Scalable Architecture for Real-Time Stream Processing of Spatiotemporal IoT Stream Data—Performance Analysis on the Example of Map Matching. ISPRS International Journal of Geo-Information, 7(7), 238. https://doi.org/10.3390/ijgi7070238*
39. *Lord, D., & Persaud, B. N. (2000). Accident Prediction Models With and Without Trend: Application of the Generalized Estimating Equations (GEE) Procedure. Transportation Research Board Annual Meeting, (00), 1–19. https://doi.org/10.3141/1717-13*
40. *Marucci-Wellman, H. R., Lehto, M. R., & Corns, H. L. (2015). A practical tool for public health surveillance: Semi-automated coding of short injury narratives from large administrative databases using Naïve Bayes algorithms. Accident Analysis & Prevention, 84, 165–176. https://doi.org/10.1016/J.AAP.2015.06.014*
41. *Metz, C. E. (1978). Basic principles of ROC analysis. Seminars in Nuclear Medicine, 8(4), 283–298. https://doi.org/10.1016/S0001-2998(78)80014-2*
42. *Mitchell, T. (2006). Human and Machine Learning. Machine Learning.*
43. *Mohan, D. (2009). Road Accidents in India. IATSS Research, 33(1), 75–79. https://doi.org/10.1016/S0386-1112(14)60239-9*
44. *Mooney, C. H., & Roddick, J. F. (2013). Sequential pattern mining -- approaches and algorithms. ACM Computing Surveys, 45(2), 1–39. https://doi.org/10.1145/ 2431211.2431218*
45. *Prasad, A. V. K., & Rama, K. S. (2010). Data Mining for Secure Software Engineering – Source Code Management Tool Case Study. International Journal of Engineering Science and Technology, 2(7), 2667–2677.*
46. *Qadah, E., Mock, M., Alevizos, E., & Fuchs, G. (2018). A distributed online learning approach for pattern prediction over movement event streams with apache flink. CEUR Workshop Proceedings, 2083, 109–116.*
47. *Reurings, M., & Janssen, T. (2006). Accident prediction models for urban and rural carriage ways. 81.*
48. *Ronza, A., Félez, S., Darbra, R. M., Carol, S., Vílchez, J. A., & Casal, J. (2003). Predicting the frequency of accidents in port areas by developing event trees from historical analysis. Journal of Loss Prevention in the Process Industries, 16(6), 551–560. https://doi.org/10.1016/j.jlp.2003.08.010*
49. *Russell, S. J., &Norvig, P. (1995). Artificial Intelligence: A Modern Approach. In Neurocomputing (Vol. 9). https://doi.org/10.1016/0925-2312(95)90020-9*
50. *Sharma, M. (2014). Data Mining: A Literature Survey. International Journal of Emerging Research in Management &Technology, 9359(2), 1–4.*
51. *Slutsky, D. (2014). The Effective Use of Graphs. Journal of Wrist Surgery, 03(02), 067–068. https://doi.org/10.1055/s-0034-1375704*
52. *Srikant, R., & Agrawal, E. (1996). Mining Sequential Patterns: Generalization and Performance Improvements. 5th International Conference on Extending Database Technology (EDBT '96), 3–17. https://doi.org/10.1109/ICDE.1995.380415*
53. *Verma, S., & Chaudhari, S. (2017). Safety of Workers in Indian Mines: Study, Analysis, and Prediction. Safety and Health at Work, 8(3), 267–275. https://doi.org/10.1016/j.shaw.2017.01.001*
54. *Verslycke, T., Reid, K., Bowers, T., Thakali, S., Lewis, A., Sanders, J., & Tuck, D. (2014). The Chemistry Scoring Index (CSI): A hazard-based scoring and ranking tool for chemicals and products used in the oil and gas industry. Sustainability (Switzerland), 6(7), 3993–4009. https://doi.org/10.3390/su6073993*
55. *Wachter, J. K., &Yorio, P. L. (2014). A system of safety management practices and worker engagement for reducing and preventing accidents: An empirical and theoretical investigation. Accident Analysis and Prevention, 68, 117–130. https://doi.org/10.1016/j.aap.2013.07.029*
56. *Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., Steinberg, D. (2008). Top 10 algorithms in data mining. In Knowledge and Information Systems (Vol. 14). https://doi.org/10.1007/s10115-007-0114-2*
57. *Yael Gavish. (2017). Developing a Machine Learning Model from Start to Finish. Retrieved March 3, 2019.*

58. *Yannis, G., Dragomanovits, A., Laiou, A., Richter, T., Ruhl, S., La Torre, F., (2016). Use of Accident Prediction Models in Road Safety Management - An International Inquiry. Transportation Research Procedia, 14, 4257–4266. https://doi.org/10.1016/j.trpro.2016.05.397*

59. *Zaki, M. J. (2001). SPADE: An efficient algorithm for mining frequent sequences. Machine Learning, 42(1–2), 31–60. https://doi.org/10.1023/A:1007652502315.*