JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

GAS PLANT LEAKAGE DETECTION USING MACHINE LEARNING

[1]RAGAVI. P, [2] Dr. T. A. ALBINAA
[1]PG Research Scholar, [2] Assistant Professor
Department of Data Analytics (PG)

PSGR Krishnammal College for Women Coimbatore, India

Abstract: Gas plant Leak detection is an important and persistent problem in the Oil and Gas plant industry. This is very important as pipelines are the most common way of transporting natural gas. This research aims to study the ability of data-driven intelligent models to detect small leaks for a natural gas pipeline using basic operational parameters and then compare the models among themselves using existing performance metrics. This project applies the observer design technique to detect leaks in natural gas pipelines using a regression classification hierarchical model where an intelligent model acts as a Linear regression and logistic regression acts as a classifier. The result shows that while support vector machines and artificial neural networks are better at regression than the others, they do not provide the best results in leak detection due to their internal complexities and the volume of data used while prediction. The developed model was trained and tested using the sequence of concentration profiles generated using open-source simulated data. The model learned successfully to predict the gas leakage and classify its size.

Keywords: Gas leakage, pipeline, random forest algorithm regression classification.

1. INTRODUCTION

Over the ages, the transportation of goods has been one of the human basic needs. For the transportation of fluids, pipelines have proven since 400 BC to be a suitable method. For natural gas, pipelines are still the most efficient and cost-effective way of transporting medium to large volumes over short to medium distances. They are virtually everywhere. While they are mostly buried, their flexibility in terms of the range of terrains (under the sea, through a desert, across a swamp, etc.) makes them the most common means of transporting natural gas. However, over time, unintended harm can come from the use of pipelines due to leaks caused by corrosion, the environment, external parties, etc. Many of these harms may be severe in nature affecting the environment, damaging properties, causing injuries, and maybe even loss of lives. With about three million kilometers of pipelines running constantly worldwide, leak detection in pipelines is key to minimizing the effect of these harms. Work is continually ongoing to improve the accuracy of leak detection and location in order to allow for a swift and efficient response. The efficiency of these systems can be assessed based on their accuracy, reliability, robustness, and sensitivity. The Oil and Gas industry is said to be one of the largest generators of data in terms of volume after the likes of Google, Facebook, Amazon, etc. While these companies have found ways of learning from these data and making better and smarter decisions, the Oil and Gas industry still lags much behind. However, the industry is beginning to realize the need for adaptation as major players are beginning to form Big Data partnerships to see how their volume of data can yield better-informed decisions. As shown by the biggest technology companies in the world, data in the hands of intelligent models can work wonders in terms of improving the effectiveness and efficiency of processes.[10]

In the oil and gas industry, various problems and anomalies could damage oil and gas pipelines, which could ultimately result in human injuries and financial loss. A few examples of problems in gas plants include corrosion, leakage, rust, etc. Oil and gas plant leakage can be dangerous for people's health and the surrounding environment. Additionally, the leakage of gases such as isobutane and propane into the atmosphere is very harmful because of their effect on ozone depletion or global warming [5]. Recent advancements in artificial intelligence (AI) and data sensing have created new opportunities to solve challenging problems in environmental monitoring, such as solid waste, air, and wastewater pollution [3]. Artificial Intelligent is one of the most useful technologies in this age. It encompasses a wide array of technologies, including machine learning (ML) and deep learning (DL), which can be used in various applications such as industry, health, economies, etc. [13] Furthermore, AI plays a pivotal role in improving the oil and gas industry, and various ML- and DL based AI techniques have been used to detect anomalies in pipelines. In previous studies, several deep-learning models were implemented to detect oil and gas plant leakages. The authors aimed to reduce environmental pollution by developing an ML model to detect oil and gas leakage. The model resulted in an accuracy of 81.397% using linear Regression, 48.12% using Logistic Regression and KNN, and 48.837% using Random Forest Algorithm.

In this paper, the authors propose an ML model to detect oil and gas plant leakages. This work provides a comparative analysis of four ML models to detect Gas plant leakage using an industrial dataset. Additionally, several optimization techniques are applied to this model to attain the highest accuracy.

2. **Data Preprocessing**

The success of Machine Learning algorithms depends on various factors. Data Processing is, therefore, important to improve the overall data quality of duplicate or missing values may give an incorrect view of the overall statistics of data, and outliers and inconsistent data points often tend to disturb the model, leading to false predictions. The work in the training phase needs to have reliable data that does not contain any noisy or redundant values. Data preprocessing and filtering are important steps in processing ML problems. Data preprocessing includes data cleaning, feature normalization, and extraction. [1]

3. Related works

Melo et al. [4] introduced different techniques for the detection of natural gas leakage in oil facilities. Different CNNs were proposed to detect the leakage of natural gas. The dataset that they used contained 2980 images and was divided into two classes, namely, 'with leak' (980 images) and 'without leak' (2000 images). The performance of 27 different CNN models was evaluated to achieve the best accuracy. The model with the best performance had the following characteristics: SGDM optimization algorithm, 18 convolution layer architecture, and dropout regularisation technique, and it yielded an accuracy of 99.78% and a false-negative rate of 0%. In the future, the researchers plan to evaluate the generalization ability of the model on unseen images of different types.

Lu et al. [2] proposed a model that can extract the features of pipelines to detect leakage. The expansion of pipeline networks and the lack of research in the field of pipeline leakage recognition using leak features were the two main driving factors in this study. A combination of various algorithms and SVM was proposed to extract pipeline leakage characteristics. The researchers employed three types of kernel functions, namely the polynomial kernel, linear kernel, and radial basis function (RBF) kernel. The researchers found RBF to be the optimum kernel function with 96% accuracy, 92% specificity, and 100% sensitivity.

Sumayh S. Aljameel, Dorieh M. Alomari, Shatha Alismail, Fatimah Khawaher, Aljawharah A. Alkhudhair, Fatimah Aljubran, and Razan M. Alzannan [29] In this paper, one of the most prominent issues faced by most oil and gas companies is highlighted, which is the problem of oil and gas leakage inside pipelines. Several previous studies were reviewed to benefit from some proposed solutions to solve the leakage problem and identify which algorithms will prevent the risk. The appropriate dataset was found, several predictive models were built using several Machine Learning algorithms, and then a comparison was made between them, for choosing the best one in terms of performance. During the stage of evaluating models on the dataset of oil and gas pipeline leakage, two experiments were conducted, the first before parameter optimization and the second after that. The first experiment result showed that all the proposed models resulted in a good performance in anomaly detection, with a performance of more than 83% in all the evaluating matrices. In comparison, the SVM model has a high performance with an accuracy of 96.1%, followed by the RF model with an accuracy of 91.56%. In the second experiment of the optimized models, there was a significant improvement in the performance of all types of models. The SVM model is still considered the best among the rest of the models, with an accuracy of 97.43% and 97% in precision, recall, f1-score, and ROC-AUC. Next to the SVM followed by the RF model with an accuracy of 91.81% and 92% in all other matrices. According to these results, the proposed model achieved good performance in the industrial data, and the main aim goal of this study is to be used in the real world.

4. Methodology

The methodology that was followed during this study includes important steps to building an ML model. The first step is to collect the required dataset and a preprocessing phase. The second step is to train the proposed model and evaluate its performance. A detailed description of the methodology is included in this section. Fig:4.1 summarizes the methodology of this study.

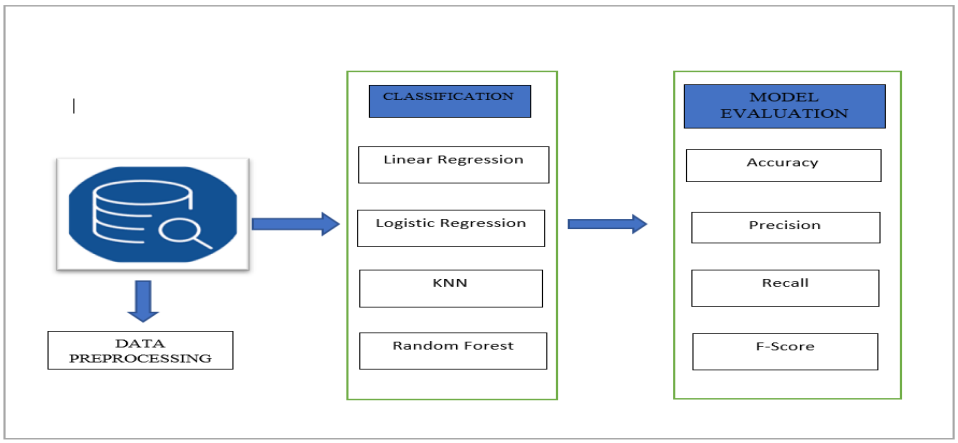

Fig: 4.1 Proposed Methodology

4.1 Data Collection

Data are collected with the help of the company. We have collected the data of different gas plant Operators with the sample data from 2010 to 2017. The dataset contains 48 features and 2795 instances, and it contains both categorical and numerical attributes. Additionally,

the dataset could be used for regression and classification problems, and it is split into training and testing sets. Table 4.1.1 describe its various features.

Features	Description
Accident Year	Year of accident happens in a gas plant
Accident Date/Time	Date and time of the accident at a gas plant
Pipeline Location	Location of the pipeline like underground, above ground, tank, etc.
Property Damage Costs	Property damage cost of gas plant
Lost Commodity Costs	The lost commodity cost of gas plant
Public/Private Property Damage Cost	Public/private property damage cost of gas plant
Emergency Response Costs	Emergency Response cost of gas plant
Environmental Remediation Costs	Environmental remediation cost of gas plant
Other Costs	Other costs of gas plant
All Costs	All costs of gas plant

Table 4.1.1 Features description

4.2 Classification and Model Design

The author-built ML models after preprocessing and cleaning the dataset. To build this model, the dataset was divided into samples to train and test the model. The model's performance was measured in terms of accuracy on the testing sample. [26] The most common approaches for splitting the dataset are 6:4 (training: testing). In the 6:4 approach, the dataset is divided into two samples, one for training and the other for testing. The training sample represents 60% of the dataset, and the testing sample is the remaining 40%. [27] The training sample is used to train the model and enhance its ability to learn the complexity behind the features of the dataset, whereas to measure the performance of the model on unseen data the testing sample is used.

4.2 Linear regression

Linear regression analysis is used to predict the variable based on another variable. The variable that you want to predict the value is called the dependent variable. The variable that you are using to predict the other variable is called the independent variable. This form of analysis estimates the coefficient of the linear equation, involving one or more independent variables that fit the best-fit line to predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the distance between predicted and actual output values. There is a simple linear regression calculator that uses a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).[15]

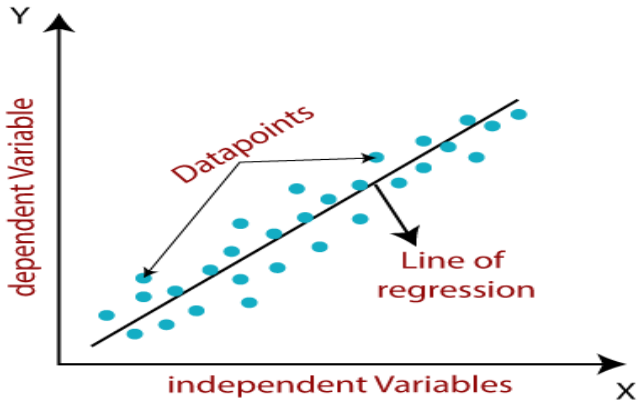

Fig:4.3.1 Linear Regression

4.4 Logistic Regression

This type of statistical model is often used for classification and predictive analytics. Logistic regression is also known logit model. Logistic regression estimates the probability of an event occurring, such as voting or did not vote, based on a given dataset of independent variables. Since the outcome is a probability of an event occurring, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied to the odds, that is the probability of success is divided by the probability of failure. This is also known as the log odds or the natural logarithm of odds, and this logistic function is represented by the following formulas:

$$\text{Logit}(pi) = 1/(1+ \exp(-pi))$$

$$\ln(pi/(1-pi)) = \text{Beta}_0 + \text{Beta}_1*X_1 + \ldots + B_k*K_k$$

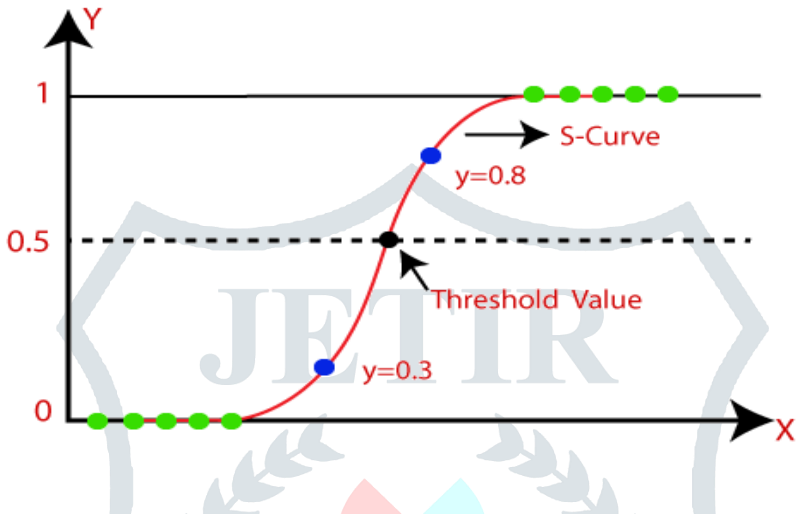

Fig:4.4.1 Logistic Regression

In this logistic regression equation, logit(pi) is the dependent or response variable and x is the independent variable. The beta parameter, or coefficient, in this model, is commonly estimated via maximum likelihood estimation (MLE). This method tests different values of beta through multiple iterations to optimize for the best fit of log odds. All of these iterations produce the log-likelihood function, and logistic regression seeks to maximize this function to find the best parameter estimate. Once the optimal coefficient (or coefficients if there is more than one independent variable) is found, the conditional probabilities for each observation can be calculated, logged, and summed together to yield a predicted probability. For binary classification, a probability less than .5 will predict 0 while a probability greater than 0 will predict 1. After the model has been computed, it's best practice to evaluate how well the model predicts the dependent variable, which is called goodness of fit. The Hosmer–Lemeshow test is a popular method to assess model fit**. [**16]

4.5 Random Forest

RF is one of the most widely used supervised ML algorithms. It is responsible for building an ensemble of decision trees and then training them using the bagging method; therefore, it is called a 'random forest'. Bagging is a concept that aims to integrate several learning models to improve the overall performance of the achieved result. In recent years, this algorithm has garnered popularity owing to its simplicity and versatility in being applied to both classification and regression models. Moreover, the isolated tree structure in the forest can predict the class, which is basically the class that obtains the highest number of votes within the model. [25]

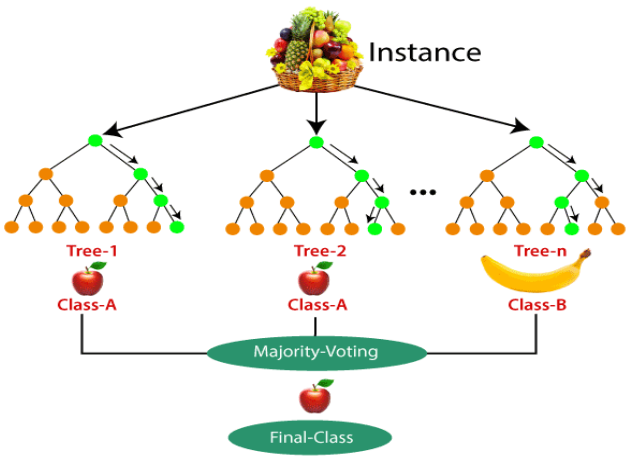

Fig:4.5.1 Random Forest

It is proven that multiple unlinked trees working together is more efficient than a single isolated tree [7]. Because of this, the trees tend to protect and shield each other from defects that may develop within the forest structure. This protection is maintained while they do not walk within the same path. An interesting mystery involves the method by which the RF algorithm ensures that the behavior of these individually isolated trees does not overly correlate with other tree structures within the model. Finally, random forest is a very useful and versatile algorithm that can be used for both regression and classification processes**.** [25]

4.6 K-Nearest Neighbor

KNN is one of the supervised algorithms used to solve both classification and regression problems. This algorithm is known as the lazy learning algorithm because it only stores data during the training phase without performing any arithmetic operations on it. This algorithm creates a predictive model that predicts the correct category of test data by finding the distance between it and the training data. The algorithm determines the k number of points closest to the test data. Next, it calculates the probability of the test data falling into the category k group, and finally, it chooses the category that achieves the highest probability. The parameter k represents the number of neighbors' relatives included in the voting process. The distance between point data and its nearest neighbor can be calculated as Euclidean distance, Manhattan distance, Hamming distance, Minkowski distance, etc. Among these distance metrics, Euclidean distance is the most widely used [21]. Figure 4.6.1 show the Euclidean distance of the KNN classifier.

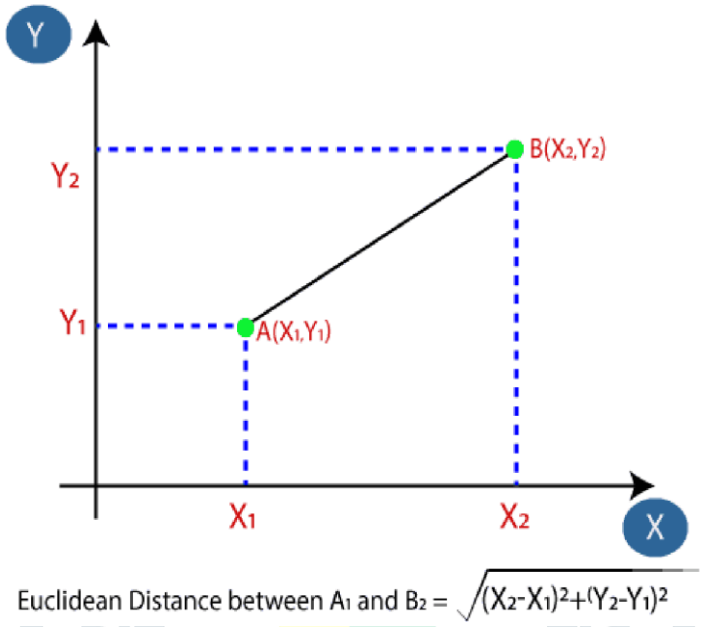

Fig:4.6.1 The Euclidean distance of the KNN classifier.

Fig:4.6.1 depicts a graph containing two classes of datasets A and B, and a new data point for which the class it might belong needs to be predicted. Using the Euclidean distance equation with a value of k equal to 5, the distance between the data points can be calculated to obtain the nearest neighbors [22]. Fig:4.6.2 shows the classification of KNN.

As shown in Fig:4.6.2, the three nearest neighbors are from class A, and the two nearest neighbors are from class B, so the new point belongs to class A [22].

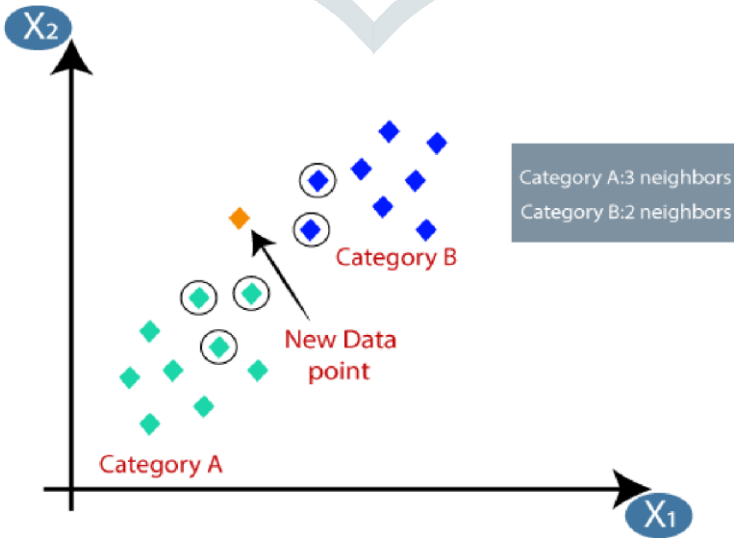

Fig:4.6.2 Three NN of KNN classifier

4.7. Evaluation Metrics

In addition to accuracy, other metrics, namely, the confusion matrix, precision, recall (Or sensitivity), specificity, F-score, and receiver operator characteristic–area under the curve (ROC-AUC) are used to measure the performance [24].
The confusion matrix measures the performance of the ML model by comparing the predicted values with the real values. Fig:4.7.1 shows a confusion matrix for binary problem classification [28].

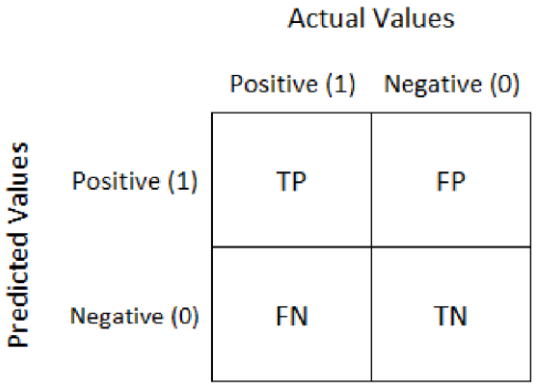

Fig:4.7.1. Confusion matrix

The symbols TP, TN, FP, and FN indicate true positive, true negative, false positive, and false negative, respectively [27].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

Accuracy represents the percentage of the truly predicted samples among all the samples in the testing set [23].

$$\text{Precision} = \frac{TP}{TP + FP}.$$

Precision represents the percentage of the truly predicted samples of the positive class among all the positive predictions [23]

$$\text{Recall} = \frac{TP}{TP + FN}.$$

Recall (also known as sensitivity) represents the percentage of the positive samples that were correctly predicted among all the real positive samples [23].

$$\text{Specificity} = \frac{TN}{TN + FP}.$$

Specificity represents the percentage of the negative samples that were correctly predicted among all the real negative samples [23].

$$F1 - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

The F1-score represents the average of the truly predicted samples of the positive class (precision) and the positive samples that are correctly predicted (recall). It is used to evaluate the balance of the model's predictions among the two classes [23,3]. ROC-AUC plots the probability of TP and FP at various thresholds. Thus, it shows the ability of the model to distinguish between the two classes [20].

5. Result and Discussion

Table 5.1 show the results of evaluating the models on the gas plant leakage dataset prior to parameter optimization. Here we have taken seven damages in a gas pipeline and calculated the classification report using linear regression, logistic regression, KNN, and Random Forest algorithm based on the damages and calculated the accuracy.

Logistic Regression

	precision	recall	f1-score	support
ALL OTHER CAUSES	0.00	0.00	0.00	50
CORROSION	0.00	0.00	0.00	238
EXCAVATION DAMAGE	0.00	0.00	0.00	41
INCORRECT OPERATION	0.09	0.03	0.04	155
MATERIAL/WELD/EQUIP FAILURE	0.50	0.95	0.66	560
NATURAL FORCE DAMAGE	0.00	0.00	0.00	53
OTHER OUTSIDE FORCES DAMAGE	0.00	0.00	0.00	21
accuracy			0.48	1118
macro avg	0.08	0.14	0.10	1118
weighted avg	0.26	0.48	0.33	1118

KNN

	precision	recall	f1-score	support
ALL OTHER CAUSES	0.00	0.00	0.00	50
CORROSION	0.00	0.00	0.00	238
EXCAVATION DAMAGE	0.00	0.00	0.00	41
INCORRECT OPERATION	0.09	0.03	0.04	155
MATERIAL/WELD/EQUIP FAILURE	0.50	0.95	0.66	560
NATURAL FORCE DAMAGE	0.00	0.00	0.00	53
OTHER OUTSIDE FORCES DAMAGE	0.00	0.00	0.00	21
accuracy			0.48	1118
macro avg	0.08	0.14	0.10	1118
weighted avg	0.26	0.48	0.33	1118

Random forest

	Precision	recall	f1-score	support
ALL OTHER CAUSES	0.54	0.14	0.22	50
CORROSION	0.38	0.29	0.33	238
EXCAVATION DAMAGE	0.31	0.12	0.18	41
INCORRECT OPERATION	0.20	0.06	0.09	155
MATERIAL/WELD/EQUIP FAILURE	0.54	0.82	0.65	560
NATURAL FORCE DAMAGE	0.14	0.02	0.03	53
OTHER OUTSIDE FORCES DAMAGE	0.20	0.05	0.08	21
accuracy			0.49	1118
macro avg	0.33	0.21	0.23	1118
weighted avg	0.43	0.49	0.43	1118

Algorithm	**Accuracy**
Linear Regression	**81.39%**
Logistic Regression	**47.94%**
Random Forest	**49.46%**
KNN	**47.94%**

Table 5.1

As evident from Table 5.1, the Linear Regression model resulted in the best performance with an accuracy of 81.39%. The other models resulted in accuracy below 50%. In figure 5 here we compared the accuracy of four algorithms using a graph. By using the gas plant leakage dataset Linear Regression has a high accuracy than the other three algorithms.

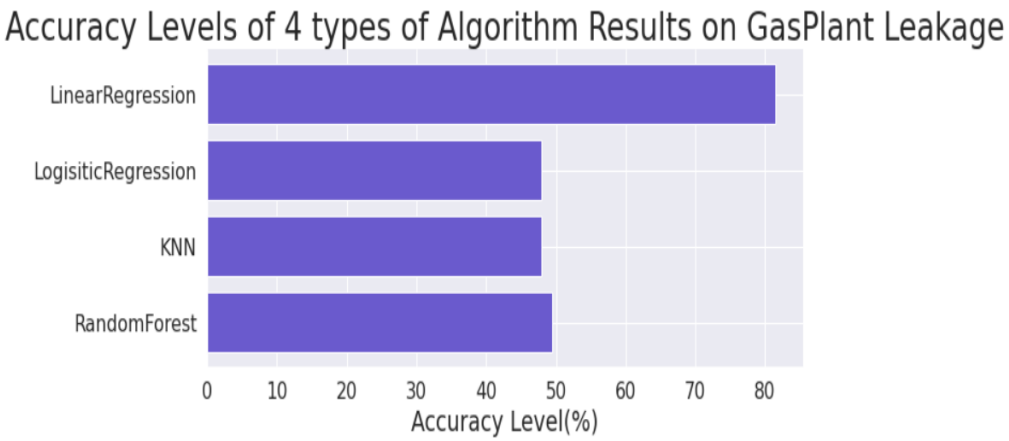

Fig:5 Result

6. Conclusion and Future Work

In this paper, one of the most prominent issues faced by most gas plant companies is highlighted, which is the problem of gas plant leakage inside pipelines. Several previous problem studies were reviewed to benefit from some proposed solutions to solve the leakage problem and identify which algorithms can be used. The appropriate dataset was found, several predictive models were built using ML algorithms, and then a comparison was made between them for choosing the best one in terms of performance. Here we used four types of algorithms like Linear Regression, Logistic Regression, KNN, and Random Forest. In this problem, we have taken seven types of damages that mostly occurred in gas plants and here we calculated accuracy, precision, recall, and f-score for all damages in different algorithms. Linear regression has an accuracy of 81.39%, Logistic Regression has an accuracy of 47.94%, Random Forest has an accuracy of 49.46 and KNN has an accuracy of 47.94%. By comparing all the algorithms only Linear regression has above 80% and all other algorithm has below 50%. Here Linear Regression has a high accuracy than the other three algorithms Logistic Regression, KNN, and Random Forest Algorithm. According to these results, the proposed model achieved good performance in the industrial data that was used, and the main aim goal of this study is to be used in the real world. Using the proposed models, it is possible to develop systems capable of effectively identifying the unusual event of gas plant leakage, thus facilitating the proper operation of the industry and avoiding any damage to the industrial companies and the surrounding environment.

Future Work

To make this project more comprehensive, the following could be used in future works pertaining to this project.

- To make the reduction of the concentration of the gas more efficient, an extractor should be used in place of the exhaust fans because it would then take the leaked gas out of the room thus reducing the concentration of the gas.
- A valve should be placed at the top of the cylinder to stop any further leakage.

REFERENCES

[1] Kotsiantis, S.B.; Kanellopoulos, D. Data preprocessing for supervised learning. Int. J. 60, 143–151, 2011.
[2] Lu, J.; Yue, J.; Jiang, C.; Liang, H.; Zhu, L. Feature extraction based on variational mode decomposition and support vector. Trans. Inst. Meas. Control 42, 759–769, 2020.
[3] Meribout, M.; Khezzar, L.; Azzi, A.; Ghendour, N. Leak detection systems in oil and gas fields: Present trends and future prospects. Flow Meas. Instrum. 75, 101772, 2020.
[4] Melo, R.O.; Costa, M.G.F.; Costa Filho, C.F.F. Applying convolutional neural networks to detect natural gas leaks in wellhead images. IEEE Access 8, 191775–191784, 2020.
[5] Morrow G., Dickerson P., 2014. Leak Sensitivity, Location Accuracy, and Robustness in Natural Gas Pipelines. The paper was prepared for presentation at the PSIG Annual Meeting in Baltimore, Maryland, 6 May – 9 May 2014.
[6] Nooralishahi, P.; López, F.; Maldague, X. A Drone-Enabled Approach for Gas Leak Detection Using Optical Flow Analysis. Appl. Sci. 11, 1412, 2021.
[7] Vairo, T., Pontiggia, M., Fabiano, B. Critical aspects of natural gas pipeline risk assessments. A case-study application on the buried layout. Process Saf. Environ. Protect. 149, 258–268, 2021.
[8] Wang S. & Carroll J.J., 2006. Leak Detection for Gas and Liquid Pipelines by Transient Modelling. The paper was prepared for presentation at the 2006 SPE International Oil and Gas Conference in China held in Beijing, 5 – 7 December 2006.

[9] Wang, F.; Liu, Z.; Zhou, X.; Li, S.; Yuan, X.; Zhang, Y.; Shao, L.; Zhang, X. (INVITED)Oil and Gas Pipeline Leakage Recognition Based on Distributed Vibration and Temperature Information Fusion. Results Opt. 5, 100131, 2021.
[10] Willey, R.J., Carter, T., Price, J., Zhang, B. Instruction of hazard analysis of methods for chemical process safety at the university level. J. Loss Prev. Process. Ind.63, 103961, 2020.
[11] Yang, X., Haugen, S., Paltrinieri, N. Clarifying the concept of operational risk assessment in the oil and gas industry. Saf. Sci. 108, 259–268, 2018.
[12] Yoo, B., Lee, Y.S. Designing an effective mitigation system based on the physical barrier for hazardous chemical leakage accidents. J. Ind. Eng. Chem. 80, 370–375, 2019.
[13] Zhang, B., Liu, Y., Qiao, S. A quantitative individual risk assessment method in process facilities with toxic gas release hazards: a combined scenario set and CFD, 2018.
[14] https://towardsdatascience.com/splitting-a-dataset-e328dab2760a
[15] https://developers.google.com/machine-learning/crash-course/descending-into-ml/linear-regression
[16] https://en.wikipedia.org/wiki/Logistic_regression
[17] https://www.ibm.com/in-en/topics/logistic-regression#:~:text=Resources-,What%20is%20logistic%20regression%3F,given%20dataset%20of%20independent%20variables
[18] https://www.researchgate.net/publication/332397437_Gas_Leakage_Detection_and_Alert_System_using_IoT
[19] https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
[20] https://medium.com/griddb/k-nearest-neighbor-algorithm-in-java-griddb-open-source-time-series-database-for-iot-6bf934eb8c05
[21] https://www.javatpoint.com/knearest-neighbor-algorithm-for-machine-learning
[22] https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b
[23] https://neptune.ai/blog/%20performance-metrics-in-machine-learning-complete-guide
[24] https://en.wikipedia.org/wiki/Random_forest
[25] https://machinelearningmastery.com/train-testsplit-for-evaluating-machine-learning-algorithms/
[26] https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62
[27] https://towardsdatascience.com/understanding-random-forest-58381e0602d2
[28] https://www.oracle.com/in/artificial-intelligence/what-is-ai/