International Journal for Modern Trends in Science and Technology, 8(06): 496-502, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0806086
Available online at: http://www.ijmtst.com/vol8issue06.html

Detection of Deceitful Behavior in Water Consumption

Mugada.Sri Lakshmi Vani, B. Netaji, Ch. Lokesh, P. Sai Bharath

Department of Information Technology, Andhra Loyola Institute of Engineering and Technology,Vijayawada, AP, INDIA.

To Cite this Article
Mugada.Sri Lakshmi Vani, B. Netaji, Ch. Lokesh and P. Sai Bharath. Detection of Deceitful Behavior in Water Consumption. International Journal for Modern Trends in Science and Technology 2022, 8(06), pp. 496-502. https://doi.org/10.46501/IJMTST0806086

Article Info
Received: 20 May 2022; Accepted: 17 June 2022; Published: 22 June 2022.

ABSTRACT

Fraudulent behaviour in drinking water consumption is a significant problem facing water supplying companies and agencies. This behaviour results in a massive loss of income and forms the highest percentage of non-technical loss. Intelligent machine learning techniques can help water supplying companies to detect these fraudulent activities to reduce such losses. The system will help the companies to predict suspicious water customers to be inspected on site. In existing system, SVM and KNN machine learning algorithms were used but they yield different accuracy on different datasets. Based on a particular algorithm, we cannot decide that algorithm will yield high accuracy than other one. This project explores the use of two classification techniques (SVM and KNN) to detect suspicious fraud water customers. The SVM and KNN based approaches uses customer load profile attributes to expose abnormal behaviour that is known to be correlated with non-technical loss activities. In this, we are using Ensemble Voting Classifier which provides an optimal accuracy from the above algorithms to predict the fraudulent behaviour.

KEYWORDS: Machine learning, SVM, KNN, Classifier

1. INTRODUCTION

Water is the basic necessity of every living thing in the world. It is an essential element for the uses of households, industry, and agriculture. But, water scarcity is a major issue that is rising rapidly in modern world. The problem has become so severe that in many areas the ground water has almost dried up and people have to depend on water supply from other sources. Efforts of the ministry of Water and irrigation to improve water and sanitation services are faced by managerial, technical and financial determinants and the limited amount of renewable freshwater resources. To address these challenges, ministry of water and irrigation as in many other countries is striving, through the adoption of a long-term plan, to improve services provided to citizens through restructuring and rehabilitation of networks, reducing the non-revenue water rates, providing new sources and maximizing the efficient use of available sources. At the same time, the Ministry continues its efforts to regulate the water usage and to detect the loss of supplied water. Water supplying companies incur significant losses due to fraud operations in water consumption. The customers who tamper their water meter readings to avoid or reduce billing amount is called a fraud customer. In practice, there are two types of water loss: the first is called technical loss (TL) which is related to problems in the production system, the transmission of water through the network (i.e., leakage), and the network washout. The second type is called the non-technical loss (NTL) which is the amount of delivered water to customers but not billed, resulting in loss of revenue. A

prediction model which predicts whether a customer has done a fraud or not based on the given input is proposed using Machine Learning (ML). In this, Ensemble method is using machine learning algorithms such as Support Vector Machine (SVM), K-Nearest Neighbour (KNN) to predict the water fraud customer. Using the predictions from multiple classifiers, the voting classifier makes predictions based on the most frequent one and will produce whether a person is water fraud or not.

The proposed system began with a strong desire to predict water fraud customer using an effective model, and to achieve a high level of accuracy when compared to previous systems by employing an ensemble learning method. Machine learning algorithms outperform other models by a significant margin. So, in further studies the researchers can work on a model that can predict the deceitful behaviour in water consumption using fewer attributes, as well as providing the required precautions that infected patients must take.

The main purpose of this proposal is to detect the abnormal behaviour in water consumption and predict the customer who has done the water fraud. Therefore, we present a predictive system that will determine whether a customer is a water fraudster or not based on the previous water consumption data. Therefore, the proposals use machine learning metrics using ensemble method to predict abnormal behaviour in water consumption and the water fraudster.

One of the most-scarce renewable resources is Water. There are 17 countries in the world that are extremely suffering from water scarcity and 1.1 billion people lack access to water and a total of 2.7 billion find water scarce for at least one month of the year. The water supplying companies are helpful in those areas of water scarcity. These companies are severely facing the loss of income due to technical loss (TL) which is related to problems in the production system, the transmission of water through the network (i.e., leakage), and the network washout and the non-technical loss (NTL) which is the amount of delivered water to customers but not billed, resulting in loss of revenue and reduce the monthly bill of water consumption. By using the previous data over present data of water consumption of a customer, using a predictive model of two machine learning classification techniques (SVM and KNN), it is possible to predict the water fraudster. But no technique is perfect by itself as the real time situations are generally continuously changing and the system has to adapt itself to change in the continuously changing circumstances.

2. RELATED WORK

[1] M. Seshubhavani et al. tells that various data mining techniques that can be used to detect and identify different types of frauds, there is little research that synthesizes various facets of This research explores the use of two classification techniques (SVM and KNN) to detect suspicious fraud water customers. The SVM based approach uses customer load profile attributes to expose abnormal behaviour that is known to be correlated with non-technical loss activities. The data has been collected from the historical data of the company billing system. To deploy the model, a decision tool has been built using the generated model. The system will help the company to predict suspicious water customers to be inspected on site. The conducted experiments showed that a good performance of Support Vector Machines (SVM) and KNearest Neighbours (KNN) had been achieved with overall accuracy around 70% for both. [2] Jos´e Carlos Carrasco-Jim´enez et al. stated that Water distribution system constantly aims at improving and efficiently distributing water to the city. Thus, understanding the nature of irregularities that may interrupt or exacerbate the service is at the core of their business model. The detection of technical and non-technical losses allows water companies to improve the sustainability and affordability of the service. Anomaly detection in water consumption is at present a challenging task. Manual inspection of data is tedious and requires a large workforce. Fortunately, the sector may benefit from automatized and intelligent workflows to reduce the amount of time required to identify abnormal water consumption. The aim of this research work is to develop a methodology to detect anomalies and irregular patterns of water consumption. We propose the use of algorithms of different nature such as Mean STL Decomposition, Median STL Decomposition and Seasonal Hybrid ESD Test that approach the problem of anomaly detection from different perspectives that go from searching deviations from typical behaviour to identification of anomalous pattern changes in prolonged periods of time. The experiments reveal that

different approaches to the problem of anomaly detection provide complementary clues to contextualize household water consumption. In addition, all the information extracted from each approach can be used in conjunction to provide insights for decision-making. [3] Angelica Goglio et al. tells that this paper mainly focuses on Machine learning algorithms that are applied towards the detection of non-technical losses are increasingly becoming a go-to solution. This project, which is part of a larger project called BD4OPEM, aims to propose a preliminary solution to detect Smart Meters tempering through machine learning algorithms. The first part of the thesis gives a theoretical overview of losses in the distribution grid, particularly it provides an extensive review of the non-technical ones, from the causes to the reasons why it is essential to reduce them. Then an explanation about the current use of Big Data and Machine Learning in the field of fraud detection is given. In particular, the different methodologies for non-technical losses detection are reported. The experimental section of the thesis is divided into three parts. In the first one, an algorithm that an algorithm that creates synthetics frauds from real load profiles is implemented. In the second section, three threat models are developed and effectively employed according to a methodology presented in the literature. The clustering algorithm k-means and fuzzy c-means, and the classification algorithm SVM are implemented and tested on different fraudulent data sets. The results are reported, and their performance is evaluated through the use of a confusion matrix. Lastly, a synthetic grid with nine regular Smart Meters and a fraudulent one is created, and the algorithms are tested on it. This example is used to see if the implemented methodology would work in a realistic scenario. the algorithms fuzzy c-means and k-means are proved to give a better solution than classification techniques; in fact, their flexibility works better if the type of fraud changes. Moreover, to check their reliability, they have been tested on different manipulated data sets, one for each possible kind of fraud, giving a positive result in most cases. They have also been tested on a simulated grid to see how the algorithms would work in a possible real case scenario, and, from this test, the two algorithms can detect the tempered smart meters at least in between ten of them. [4] SamerNofal et al. presents a use case of anomaly detection for identifying the unusual water consumption of consumers. Unusual water consumption may be due to a faulty water meter, fraudulent tampering with a water meter, or a leak in the water pipes within the consumer's property. We apply several anomaly detection methods to a real dataset of 22,877 mechanical water meters located in Amman, the capital city of Jordan. The dataset is unlabelled such that no discrimination is given for any meter whether it records a normal water consumption or not. The objective of this study is to test the hypothesis that abnormal water consumption (registered by a given water meter) can be identified based on previous records of water consumption measured by the same meter. We tested our hypothesis using well-known anomaly detection methods, namely: z-score (ZS), local outlier factor (LOF), density-based spatial clustering of applications with noise (DBSCAN), minimum covariance determinant (MCD), one-class support vector machine (OCSVM), and isolation forest (iForest). In the settings of our experiments, we observed that ZS, LOF, OCSVM and iForest support our hypothesis, contrasting with DBSCAN and MCD. we note that LOF, OCSVM, iForest, and ZS supported our hypothesis because their F1 scores were greater than 0.80 with respect to most of the reference points (except for DBSCAN and MCD). In contrast, DBSCAN and MCD seem to undermine our hypothesis as their F1 scores were less than 0.60 with respect to all reference points.

3. PROPOSED METHOD

System Analysis is the process of analysing a system with the potential goal of improving or modifying the system. Analysis is breaking down the problem into smaller elements of study and ultimately providing a better solution. During the process of system development, Analysis is an important aspect. This involves gathering and interpreting facts, diagnosing the problem and using the information to recommend improvements to the system ultimately, the goal is to give a computerized solution

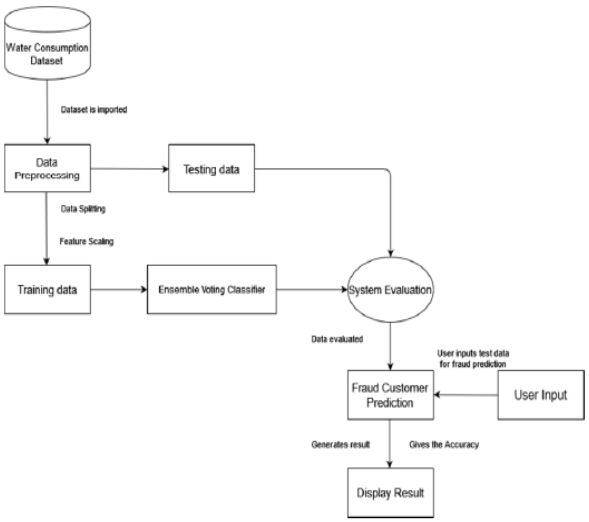

Figure 1: System architecture for detecting deceitful behaviour in Water consumption

System will take input from datasets. The datasets undergo pre-processing and the unnecessary information is removed from it and the data types of the columns are changed if required. Jupyter notebook/ Google Colaboratory and python libraries are used in the above step. Label Encoder technique is used in the initial step. For fraud water consumer, we have to train the system using dataset. Before entering to the detection of deceitful behaviour in water consumption, entire dataset is divided into two datasets. 80% is used for training and 20% is used for testing. During training, Ensemble Voting Classifier where the SVM and KNN are used to train the model using the train dataset. In testing, from test dataset the user gives a customer data as input and the output is predicted. After the testing time, the predicted output and the actual output are compared using confusion matrix obtained. The confusion matrix gives the information regarding the number of correct and wrong predictions in the case of fraud customer or not. The accuracy is calculated and given from Ensemble Voting Classifier.

ALGORITHM FOR THE PROPOSED SYSTEM

Step 1: Start

Step 2: Input data is taken from a dataset .csv file

Step 3: Pre-processing of data is done and dataset is divided into 2 parts: training data and testing data.

1. Feature Scaling is used in pre-processing helps in scaling the data which are havinghuge difference in range of data (high and low ranges) and makes the machine learning model a better one.

2. Label Encoder converts labels into a numerical form so as to convert them into the machine-readable format i.e., 0's and 1's.

Step 4: Ensemble Voting Classifier is used to build a predictive model using the trained data.

Step 5: Confusion matrix is obtained

Step 6: Accuracy is calculated and displays the result.

4. RESULTS AND DISCUSSIONS

To execute the code in Google Colaboratory, simply press the Ctrl+F9 on your keyboard or go to Tool bar as shown in the figure 2 and click on Runtime Tab→ Run all tab to run all cells with code

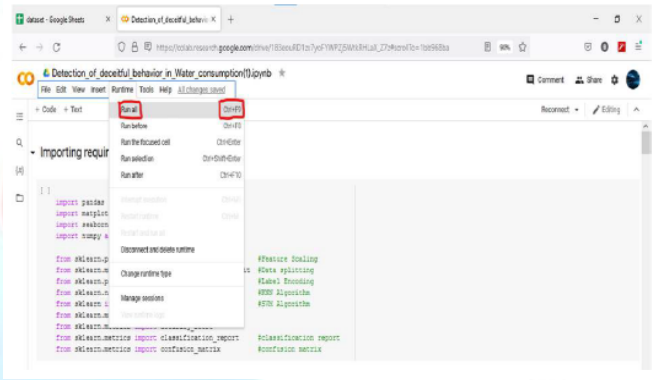

Figure 2 Running the code

After the code has run, the water consumption dataset will be imported from given path as shown in figure 3

```
df = pd.read_csv(r'/drive/My Drive/dataset.csv')
df.head()
```

	Condition	Case	Diameter	Year	Month	Reference	Consumption	Label
0	1	1	0.5	9	1	764.0	13.0	NO
1	2	1	0.5	9	1	576.0	15.0	NO
2	131	2	0.5	9	1	0.0	32.0	YES
3	131	3	0.5	9	1	23.0	11.0	YES
4	131	4	0.5	9	1	543.0	0.0	YES

Figure 3 Dataset is imported

After importing the dataset, Data Pre-processing is done where the data other than numerical data is converted to numerical one i.e., Label column is encoded into numerical form as shown in the figure 4

```
en = LabelEncoder()
df['Label'] = en.fit_transform(df['Label'])
df.head()
```

	Condition	Case	Diameter	Year	Month	Reference	Consumption	Label
0	1	1	0.5	9	1	764.0	13.0	0
1	2	1	0.5	9	1	576.0	15.0	0
2	131	2	0.5	9	1	0.0	32.0	1
3	131	3	0.5	9	1	23.0	11.0	1
4	131	4	0.5	9	1	543.0	0.0	1

Figure 4 Data Pre-processing on the imported data

The figure 5 gives the Overall Accuracy and the Confusion Matrix of KNN Algorithm which is shown in below page.

```
[96] print("Accuracy of KNN Algorithm: ")
     print(knn.score(x_test,y_test)*100)

Accuracy of KNN Algorithm:
97.25171767645222

print("Confusion Matrix of KNN Algorithm:\n")
cm1=confusion_matrix(y_test,y_pred2)
sns.heatmap(cm1, annot = True)

Confusion Matrix of KNN Algorithm:

<matplotlib.axes._subplots.AxesSubplot at 0x7f6129634450>
```

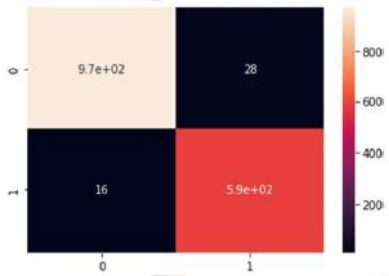

Figure 5 Accuracy and Confusion Matrix of KNN Algorithm

The figure 6 gives the Overall Accuracy and the Confusion Matrix of SVM Algorithm which is shown below.

```
[103] print("Accuracy of SVM Algorithm:")
      accuracy_score(y_test,y_pred5)*100

Accuracy of SVM Algorithm:
89.75640224859463

print("Confusion Matrix of SVM Algorithm\n")
cm4 = confusion_matrix(y_test,y_pred5)
sns.heatmap(cm4, annot = True)

Confusion Matrix of SVM Algorithm

<matplotlib.axes._subplots.AxesSubplot at 0x7f612954ddd0>
```

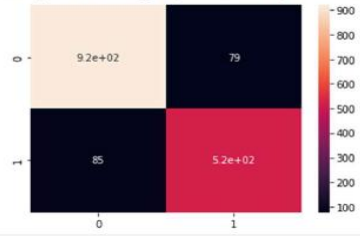

Figure 6: Accuracy and Confusion Matrix of SVM Algorithm

Here, the system prompts the user to enter the Water Consumption details, then the User must enter the details for detecting the fraud has done or not which is shown in below figure 7

```
Enter User water Consumption Detials:
Enter condition
133
Enter case
5
Enter diameter
.5
Enter year
9
Enter month
1
Enter reference
554
Enter consumption
22
```

Figure 7: User Input for detecting the Water Consumption Fraud

After giving user input, the Ensemble Voting Classifier uses the SVM and KNN results on the data and evaluates those and generate an output of whether the customer did a fraud or not and gives an accuracy of how well the model is detecting fraud by taking testing data as shown in the below figure 8

```
Yes [1]
Water Consumption Fraud detected

Accuracy: 99.87507807620237

classification report

              precision    recall  f1-score   support

           0       1.00      1.00      1.00       997
           1       1.00      1.00      1.00       604

    accuracy                           1.00      1601
   macro avg       1.00      1.00      1.00      1601
weighted avg       1.00      1.00      1.00      1601
```

Figure-8 Water Fraud Detection, Gives Accuracy and generates a Classification report for Ensemble Voting Classifier

The figure 9 shows the confusion matrix which displays the 4 types of outcomes i.e., True Negative, False Positive, False Negative, True Positive.

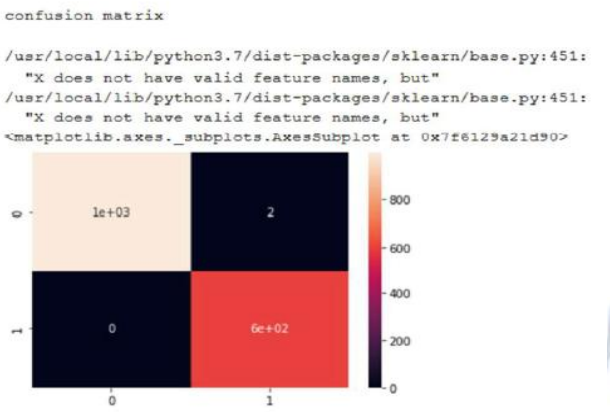

Figure 9 Confusion Matrix of the Ensemble Voting Classifier

The figure 10 displays the accuracy comparison of KNN Algorithm, SVM Algorithm and Ensemble Voting Classifier where we got high accuracy for Ensemble Voting Classifier.

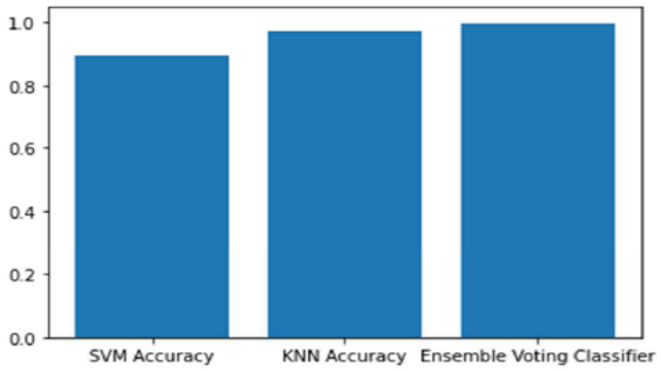

Figure 10 Accuracy Comparisons of SVM, KNN and Ensemble Voting Classifier

Test Case Input: Condition, Case, Diameter, Year, Month, Reference, Consumption

Result: PASS or FAIL

Table 1: Test Cases

Test Case ID	Test Case	Expected Output	Predicted Output	Result
1.	133, 5, 0.5, 9, 1, 554, 22	1(Water Fraud Detected)	1(Water Fraud Detected)	PASS
2.	1, 1, 0.5, 9, 1, 764, 13	0(No Water Fraud Detected)	1(Water Fraud Detected)	FAIL
3.	2, 7, 0.5, 9, 2, 3095, 20	0(No Water Fraud Detected)	0(No Water Fraud Detected)	PASS
4.	82, 5, 0.5, 9, 1, 0, 20	1(Water Fraud Detected)	0(No Water Fraud Detected)	FAIL

5. CONCLUSIONS

Detection of Fraud Water Consumption is a system which helps to predict whether the customer is doing water fraud or not using machine learning techniques. The data set consists of 7 independent variables and 1 dependent variable to predict the wate fraud. Many of the previous research papers yield different accuracy on different datasets using machine learning algorithms such as Support Vector Machine, K-Nearest Neighbour Algorithms. This proposed model used Ensemble voting classifier which combines the above machine learning algorithms to predict the water fraud based on majority probability of the models. The proposed method achieved an accuracy of 99% compared to above ML models in which KNN algorithm got 97% whereas SVM algorithm got 89% accuracy.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] M.Seshubhavani, K.Yashoda, T.Manjusha, M.Pushkala. "A Dynamic Detection Of Deceitful Behaviour In Water Depletion", In Crossref, Volume 04, Issue 05, May 2020, ISSN 2581 – 4575

[2] Jos´e Carlos Carrasco-Jim´enez, Filippo Baldaro and Fernando Cucchietti "Detection of Anomalous Patterns in Water Consumption: An Overview of Approaches", IntelliSys 2020, AISC 1250, pp. 19–33, 2021.

[3] Angelica Goglio "Meter tampering detection through short-lived patterns clustering", Escola Tècnica Superior d'Enginyeria Industrial de Barcelona(ETSEIB), June 19, 2021

[4] SamerNofal , Abdullah Alfarrarjeh and Amani Abu Jabal "A use case of anomaly detection for identifying unusual water consumption in Jordan", Water Supply Vol 22 No 1, 1131 doi: 10.2166/ws.2021.210

[5] Monedero I., Biscarri F., Guerrero J., Roldán M., and León C. "An Approach to Detection of Tampering in Water Meters", In Procedia Computer Science, 2015, 60: pp 413- 421.

[6] Vercruyssen, V., Meert, W., Verbruggen, G., Maes, K., Baumer, R., Davis, J.: Semisupervised anomaly detection with an application to water analytics. In: IEEE International Conference on Data Mining (ICDM), pp. 527–536, November 2018

[7] Dogo, E.M., Nwulu, N.I., Twala, B., Aigbavboa, C.: A survey of machine learning methods applied to anomaly detection on drinking-water quality data. Urban Water J. 16(3), 235–248 (2019)

[8] Christodoulou, S.E., Kourti, E., Agathokleous, A.: Waterloss detection in water distribution networks using wavelet change-point detection. Water Resources Manage. 31, 979–994 (2017)

[9] Romano, M., Kapelan, Z., Savi´c, D.A.: Automated detection of pipe bursts and other events in water distribution systems. J. Water Resour. Plann. Manage. 140(4), 457–467 (2014)

[10] J´unior, L.A.P., Ramos, C.C.O., Rodrigues, D., Pereira, D.R., de Souza, A.N., da Costa, K.A.P., Papa, J.P.: Unsupervised non-technical losses identification through optimum-path forest. Electr. Power Syst. Res. 140, 413–423 (2016)

[11] Roccetti, M., Casini, L., Delnevo, G. &Bonfante, S. 2021 Dimensionality reduction and the strange case of categorical data for predicting defective water meter devices. In: Human Interaction, Emerging Technologies and Future Applications III (Ahram, T., Taiar, R., Langlois, K. &Choplin, A., eds). pp. 155–159, Paris, Springer.

[12] Muharemi, F., Logof˘atu, D., Leon, F.: Machine learning approaches for anomaly detection of water quality on a real-world data set. J. Inf. Telecommun. 3(3), 294– 307 (2019)