JES
Journal of Engineering Sciences

Vol 13, Issue 02, FEB /2022
ISSN NO:0377-9254

A DATA MINING BASED MODEL FOR DETECTION OF FRAUDULENT BEHAVIOUR IN WATER CONSUMPTION

MANINDRANATH KOMMINENI[1], POOJITHA YATA[2], GURRAM NIKHIL[3], KALYANAPU RAJA[4], YAJURVED JAYAVARAPU[5]

#1#2#3#4#5 Student,SRM Institute of Science and Technology.

Mail id : manichowdary710@gmail.com

Mail id : poojithareddy558@gmail.com

Mail id : gurramnikhil20@gmail.com

Mail id : Rajakalyanapu1999@gmail.com

Mail id : yajurved9898@gmail.com

Abstract:

It's a major concern for water supply companies and authorities to deal with the fraudulent use of drinking water. Non-technical losses are the most common as a result of this type of behaviour, which causes significant financial harm. Recent years have seen a lot of work in the area of developing metrics that are effective at spotting fraud. The use of advanced data mining tools can assist water supply businesses detect and prevent these fraudulent acts, hence reducing their losses. SVM and KNN, two classification approaches being investigated, are being used to identify water customers who may be engaging in fraudulent behaviour. The primary goal of this study is to help the Yarmouk Water Company (YWC) in Irbid, Jordan, overcome its earnings loss by developing a new water treatment system. Non-technical loss activities have been shown to be associated with aberrant customer load profile features in SVM-based approaches. The information was gleaned from the company's billing system's archived records. The produced model's accuracy was higher than the YWC's present manual prediction techniques, at around seventy-four percent. A decision-making tool based on the derived model was created for the purpose of deploying the model. The device will assist the corporation in anticipating suspect water users who will be checked out on-site by the appropriate authorities.

I. Introduction:

Household, industrial, and agricultural purposes all require water. Jordan, like many other countries, suffers from water scarcity, which poses a danger to the sustainability of all sectors that rely on water for their development and prosperity [1]. Water and irrigation issues have long been a major impediment to Jordan's economic development, according to Jordan's water ministry. An increase in population over the past two decades has exacerbated this crisis situation. It is difficult to provide better water and sanitation services because of a limited quantity of renewable freshwater resources [2]. As a long-term solution to Jordan's water woes, the country's Ministry of Water and Irrigation has developed a strategy to reorganise and repair water networks, lower non-revenue water rates, provide new water sources, and maximise the efficiency of existing resources. Ministers' attempts to limit water consumption and discover water losses are also underway [2.] [3] Water delivery firms lose a lot of money due to fraudulent water use operations. Customers that commit water metre fraud do so to reduce or avoid their monthly water bill. Technical loss (TL) and network washout are two types of water loss in the real world. In the context of TL,

problems in the production system and water transport across the network are discussed (i.e. leakage). This non-technical loss is referred to as "water given but no money is collected" (NTL). Management of Jordan's Yarmouk Water Company is particularly worried about lowering its losses due to NTLs, which are estimated to be above 35% in 2012 for the whole service area. As a result of the lack of sophisticated computerised technology, the commercial department is responsible for overseeing the current detection methods. NLT suffers from a considerable amount of customer fraud. NTL is a big problem for Yarmouk Water Company (YWC). In 2012, the NTL was more than 35%, with district-specific percentages ranging from 31% to 61%, resulting in an annual loss of $13 million.. However, the model described in this study offers a helpful tool to help YWC teams recognise consumers who have stolen, lowering the NTL and improving profit. It is currently done at random. Non-technical losses (NTL) in the detection of energy theft have a plethora of study, however studies on water usage are scarce. For the purposes of this investigation, historical customer billing records from YWC have been analysed. Support Vector Machines (SVM) and K-Nearest Neighbor are two well-known data mining techniques that can be used to build a model that can detect questionable customers based on their water metered use history (KNN).

More than 90 percent of Jordan's terrain receives less than 100 millimetres (mm) of annual precipitation, while less than 3 percent receives more than 300 millimetres (mm). Because of Jordan's long history of severe weather, all three of these elements of precipitation — rain, runoff, and evaporation — are vulnerable to considerable variations in time and space. Since 1998–2005, population growth in Jordan has averaged roughly 3%, making it the ninth fastest-growing country in the world, according to figures from the Department of Statistics. In addition to groundwater, Jordan's water supply comes from baseflows and reservoirs, treated wastewater that doesn't flow into reservoirs, and other water sources. Surface water from the Yarmouk River, water from the peace accord, and non-traditional sources including desalination and non-renewable aquifer groundwater [1] have all been added to our current water supply. Percentage of each water resource in million cubic metres (Mm 3), as determined by Jordan's Ministry of Water and Irrigation, for the year 2005 Due to excessive over-pumping of the few remaining non-renewable resources in the country, Jordan's water sector is in jeopardy. [2] Since 1946, the amount of renewable water resources available to the population has dropped from 3600 m3 to 160 m3 per capita, a substantial decline.. There has been a significant reduction in the amount of nonrenewable resources due to population growth, agricultural and industrial development, as well as a rapid influx of refugees. Drought-induced water limitations reduced the total amount of water required in 2001 to just 774 Mm 3. Research on Jordan's water sector in the past has usually shown that it is necessary to anticipate the future impact of water shortages in resource planning and development [2–6]. Institutional reforms, new pricing approaches, water imports and desalination are all necessary components of Jordan's regional water management strategy, which also includes desalination. When formulating a new water resource management plan, decision-makers should take into account the influence of regional political instability.

The privatisation of electric power suppliers has led to increased competition on the national market in some countries' electrical networks. The fundamental objectives of business expenditures aimed at boosting productivity, efficiency, and profitability are financial and technological advancements. [1] Preventing theft or fraud is one way to reduce energy losses in electric power providers. It's important to note that both technical and nontechnical losses are based on the gap

between what you generate or buy and what you bill. A financial burden is placed on electricity providers as a result of these technical losses because of the high costs associated with repairing and replacing equipment that causes system problems [2]. As a result of charging customers for energy rather than giving it away for free, commercial losses, also known as non-technical losses, occur. Total losses and technical losses resulting from unauthorised distribution system connections have been demonstrated to be significantly impacted by illegal connections[3]. Electric power companies across the country and around the world have suffered significant financial losses as a result of power metre thefts and failures [4]. Losses in the commercial sector are notoriously difficult to quantify since it is practically impossible to pinpoint their exact location [5]. There is no room for complacency when it comes to tackling the problem of illegal power hookups. Due to a number of scams, authorities may not be able to collect taxes and tributes in this situation [1]. Because of these losses there will be more investments in product quality, protection of the property of the concession and a cheaper cost for public services [7]. It is becoming increasingly necessary to conduct research in this area despite recent breakthroughs, such as the use of various techniques to measure electric energy, to produce more flexible and adaptive methods, such as intelligent algorithm-based models of computational procedures. Nagi et al. [8] used SVMs for non-technical loss analysis, while Nizar et al. [7] used data mining approaches. Monedero et al. [10] advocated using ANNs [11] and statistical analysis to detect fraud in electrical usage. Genetic Algorithms - GA [12] and SVM [13] were also utilised for the detection of non-technical losses.

It appears that "book-cooking" accounting practises are becoming more commonplace. Accounting had a "bad" year, according to Koskivaara, who claims that fraud is still going on (Koskivaara, 2004). According to some estimates, US corporations lose $400 billion annually as a result of fraud (Wells, 1997). According to Spathis, Doumpos, and Zopounidis, fraudulent financial statements have become more widespread in recent years (2002). Managerial fraud in financial reporting can be defined as the deliberate deception of management that hurts investors and creditors by providing substantially inaccurate financial information. During the audit, the auditors must make an educated prediction about the likelihood of managerial fraud. The American Institute of Certified Public Accountants (AICPA) explicitly acknowledges this obligation for fraud detection (Cullinan & Sutton, 2002). Analytical review techniques are used by auditors during the audit process to estimate account balances without having to look at specific transactions. Fraser, Hatherly, and Lin have categorised simple and advanced quantitative analytical review techniques (1997). Statistics and artificial intelligence are two fields that have contributed to the development of complicated quantitative approaches. Fraud detection and financial statement verification have recently come to the forefront in Greece as the number of Greek companies listed on the Athens Stock Exchange has increased and cash has been generated through public offerings. Efforts have been made to lower profit taxes as well. In Greece, there has been a constant need for phoney financial accounts and qualified judgments as evidence of corporate failure. There is rising demand on financial statements to include more information, to be more open, and to be more consistent (Spathis, Doumpos, & Zopounidis, 2003).

Several countries lose tens of billions of dollars annually as a result of health care fraud and misuse. [4] Medical fraud can take a variety of forms. Health care fraud and corruption can be committed by anyone, including doctors, nurses, managers, and vendors, according to [6]. Fraud can be a severe problem for a company even if there are no immediate legal repercussions since prevention methods are not perfect. [7]. Fraud detection and fraud prevention are not the same thing [4]. Actions made to prevent fraud are known as "fraud prevention." When it comes to detecting fraud, on the other hand, the goal is to do it as quickly as possible after the fact. A person (or entity) willfully defrauds another person (or entity) in order to gain an advantage for themselves or others, according to the National Health Care Anti-Fraud Association (NHCAFA). Abuse of the healthcare system is defined as services that are not medically necessary or do not meet industry standards for health care [8]. The provider's actions are in contrast with generally recognised financial, commercial, and medical norms when this occurs. To catch health insurance fraud and misuse, you'll need a lot of specialised expertise, but that's about it. The policies must be flexible enough to keep up with the changing trends and at the same time serve as preventative measures. A data mining-based iterative technique to knowledge discovery is known as KDD [9.]. It's possible that [10] can help with the automatic extraction of this data. The ability to identify fraudulent claims, providers, and beneficiaries has improved the efficiency of health care fraud detection and investigative resources. [11] For commercial and government data mining applications, automated fraud detection has become one of the most commonly employed. In many countries, including the United States, healthcare fraud is a problem. With regard to both the financial impact and the level of expertise involved [12], medical insurance fraud has a broad scope. The cost of health care in many nations is split between the public and private sectors through health insurance programmes. [13] [14]. In the 1980s, Chile pioneered the idea of decentralising primary health care and fostering the development of a private health insurance market. [16]. Insured Chilean workers and their dependents have the option of channelling their statutory 7 percent payroll contributions toward health insurance through the FONASA (National Health Fund) and the privately run pre-paid healthcare plans known as ISAPREs (Institutos de Salud Previsional). The ISAPRES system, which entered into operation in 1981, was inspired by health maintenance organisations (HMOs) in the United States, which aimed to expand the provision of private health care and provide consumers with more options through a more competitive health insurance market. In some cases, private health insurance companies' affiliates may pay an additional fee for a particular health plan. There is a limit on how much can be spent on health care, and the price determines how much coverage is provided.

II. Literature survey:

Pedro A. Ortega, Cristian J. Figueroa, Gonzalo A. Ruz" A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile"

This study outlines a data mining-based method employed by a Chilean private health insurance company to detect medical claim fraud and abuse. Health insurance firms in Chile have become increasingly concerned about fraud and abuse in medical claims in recent years due to a rise in revenue losses. There are only a few number of medical professionals tasked with approving, revising, or rejecting the subsidy requests they receive within a short period of time after they are received. An MLP-based detection system is proposed, with one committee for each of the four parties involved in medical claims, affiliates, medical professionals, and employers committing fraud or abuse. With a

detection rate of around 75 fraudulent and abusive instances each month, the fraud detection system has saved the company 6.6 months over the previous method. Implementing an automated fraud detection system has transformed a non-standard medical claims screening process into a standardised procedure that helps to combat new, atypical, and known fraudulent/abusive activities.

Efstathios Kirkos a,1, Charalambos Spathis b,*, Yannis Manolopoulos c,2" Data Mining techniques for the detection of fraudulent financial statements"

Fake financial statements (FFS) and characteristics associated with them can be identified using DM classification algorithms, according to this article. Detecting management fraud can be made easier with the help of Data Mining techniques. All of these techniques are tested in this study to see if they can be used to identify fraudulent financial statements. The input vector is a collection of financial ratios. The three models are compared head-to-head. It's a no-fly zone. It was published by Elsevier Ltd. in 2006

SVMs and Decision Trees for Detecting Credit Card Fraud by Y. Sahin and E. Duman

Advances in information technology and improved communication methods are fueling an increase in global fraud. The most typical forms of fraud, such as fraudulent credit card usages over virtual POS terminals or postal orders, are not prevented by known fraud prevention systems like CHIP and PIN. Fraud detection has become a necessary tool in the fight against this type of fraud. For the purpose of identifying credit card fraud, decision trees and support vector machines (SVM) classification models are being developed and tested. For the first time, real credit card fraud data was used to compare SVM and decision tree algorithms for credit card fraud detection.

Caio C. O. Ramos, Andre N. Souza, Joao P. Papa, Alexandre X. Falcao" Fast Non-Technical Losses Identification Through Optimum-Path Forest"

Detecting fraudulent conduct in energy systems by unregistered consumers is the greatest way to study the non-technical losses suffered by electric power companies. Artificial Neural Networks and Support Vector Machines have been used to construct automated commercial fraud detection, however they suffer from slow convergence and high computing loads. Neuronal networks have been demonstrated to be inferior and Support Vector Machines to be comparable, while Optimal-Path Forest classifier has been proved to be substantially faster than both of these methods. Comparisons are also done between different classifiers and displayed.

Z. S. Tarawneh, N. Ahadin, and Ahmad N. Bduh: Policies to Strengthen Jordan's Water Sector

For Jordan's water sector, we proposed a set of broad policies that may be applied to improve its long-term viability and usability. Jordan's water laws should be amended to allow the private sector to have a greater role in the maintenance of the country's water infrastructure, according to a new report. Water resources should also be redistributed across competing sectors so that those who will reap the most benefits in terms of economics and society are given priority. Activating public awareness programmes may promote public engagement in the development and acceptance of new policies linked to water management, according to the authors of this report.

III. Methodology:

CRISP-DM (Cross Industry Standard Process for Data Mining) was used in this study [26]. Standard data mining approach developed by NCR systems engineering, DaimlerChrysler AG, SPSS Inc., and OHRA as part of a consortium of four firms. Model building, evaluation, and deployment are all part of

CRISP-business DM's understanding, data comprehension, and data preparation.

An Introduction to the Business World:

Yarmouk Water Company (YWC), which serves the northern governorates of Irbid, Mafraq, Jerash, and Ajloun, is owned by the Water Authority of Jordan (WAJ). At YWC's corporate headquarters, the departments of Commercial, Operations Support, Technical Affairs, Finance, Human Resources, and Information Technology all work together to support the company's overall objectives. Ten branches distribute water and provide customer service in the concession area. These branches are referred to as Regional Organizational Units (ROUs). For a better understanding of the billing process for our clients, we conducted interviews with staff, reviewed documentation, and examined the billing system. Three sets of consumers had their metre readings taken within one month of each other, and each group receives a bill for three months at a time. The commercial section connects to the company's Geographic Information Systems GIS to design reading routes for group reading. Customers' metres are read by meter-reading teams who visit their homes. Handheld units are used to enter data from metres, issue bills and deliver them to customers once a metre reader takes a reading. The system does not print a bill if there is an issue (such as a high or low reading). An Oracle application and FTP tools transport billing data from the HHUs to the COBOL billing system, which is directly connected to the server. Afterward, the commercial department's billing auditor reviews the invoices and fixes any issues they may have (i.e., high reading). When all issues have been resolved, bills are signed into law. In the mid-1980s, the COBOL programming language was used to create the venerable billing system. A mainframe running OPEN-VMS hosts it, and it's still in use. In October 2011, Yarmouk Water Company began using an HHU billing system. The bills will be printed on-site using this technique. The COBOL-based billing system is connected with the HHU billing system, and it is this system that computes and issues all bills. The ROU's business sections are working to stop water theft. A random inspection of the properties and water connections of customers is undertaken when water is supplied to a zone. If a theft case is discovered, the customer is penalised and the information is entered into a special form document that must be returned to the department.

Ease of Data Interpretation:

The data itself is a critical component of the data mining process. The acquired data is described in detail in the following sections. This article focuses on data from the billing system, which is primarily used to bill customers for their water consumption. Customer billing data can be extracted from COBOL programmes into a text-formatted database. Similar to COBOL data files, Oracle tables are prepared in a similar format.

TABLE I. CUSTOMERS' CONSUMPTIONS TABLE

Column	Description
DIST_NO (PK)	The district number
TOWN_NO (PK)	The village/town number
CONS_NO (PK)	Customer number
BILL_NO	The number of the bill
BILL_STAT	The status of the bill
ISSUE_DATE	Date of bill issue
PRINT_FLAG	A flag mentions if the bill is printed
OLD_MET_PREV_RDNG	The old meter previous reading
FORWARD_BAL	The forward balance
ADVANCE_PMNT	Advanced payment
OUTSTANDING_AMOUNT	Outstanding balance
CALC_CONS_FLAG	A flag mentions the consumption was calculated
METER_NO	Water meter metal number
METER_STATUS	Water meter valid or corrupted
ISSUE_DATE	Bill issue date
CONS_TYPE	Customer service type (industrial, housing,..)
FROM_DATE	Start of consumption period for the bill
TO_DATE	end of consumption period for the bill
UNITS_CONS	units consumed in cubic meter
UNITS_CONS_VAL	The price of the consumed water
SEWER_VAL	The price of Sewerage service
METER_RENT	The meter fees
CYCLE_NO	Consumption Cycle number
CYCLE_YY	Consumption Cycle year

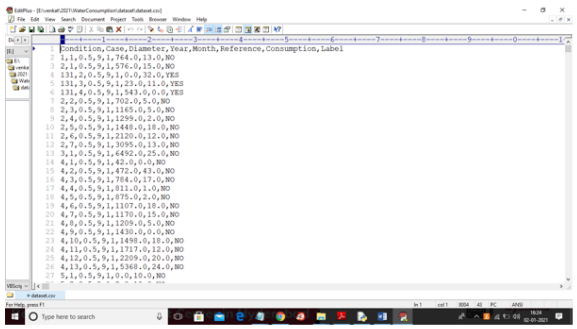
IV.Results:

Model for Detection of Fraudulent Water Consumption Based on Data Mining

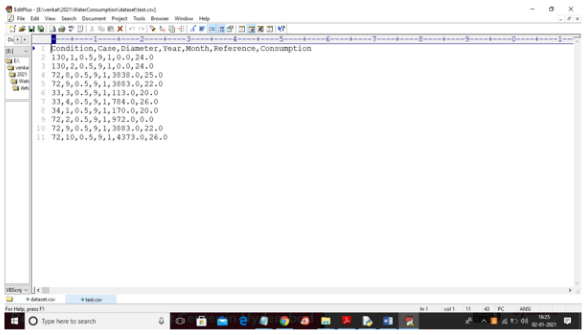
Water fraud can be detected by looking at past usage patterns, which is what the author of this research refers to as "behavioural" analysis. Government water supply providers will lose money if clients modify their water metres so that they don't operate according to

consumption. To avoid the time and errors involved in personally examining customers' homes, the author has turned to machine learning methods like SVM and KNN. Customer history data will be used to train SVM and KNN models that will subsequently be used to classify if a customer metre is normal or a fraud. SVM outperforms the other algorithms in terms of accuracy.

To implement this project, we used water usage data and we changed this dataset to meet the application's needs. The dataset used to train algorithms is shown in the images below.



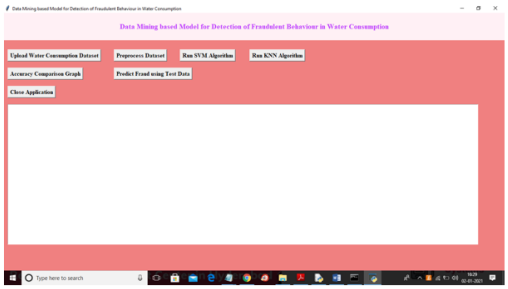
In above training dataset last column contains values as YES which means customer meter is fraud and NO means normal customer water meter. After training with above dataset we will used below test data to predict whether customer meter is normal or fraud



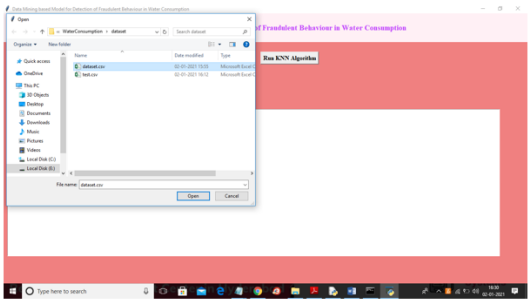
In above test dataset we don't have label as YES or NO and when we apply above dataset on trained model then application will predict YES or NO label for above dataset.

SCREEN SHOTS

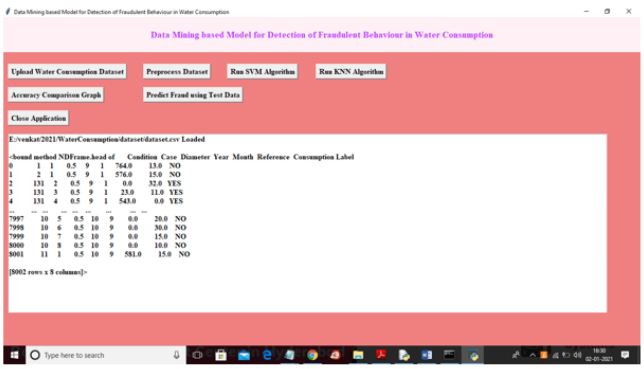
To run project double click on 'run.bat' file to get below screen



In above screen click on 'Upload Water Consumption Dataset' button and upload dataset

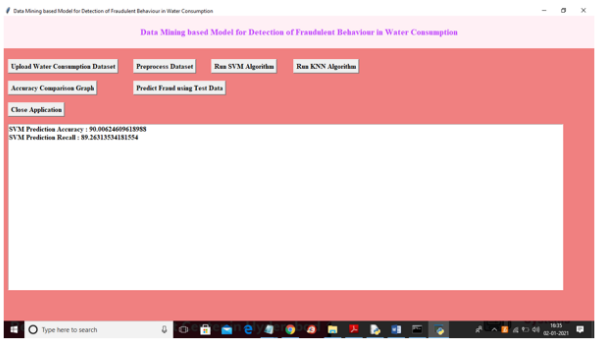


In above screen selecting and uploading 'dataset.csv' file and then click on 'Open' button to load dataset and to get below screen

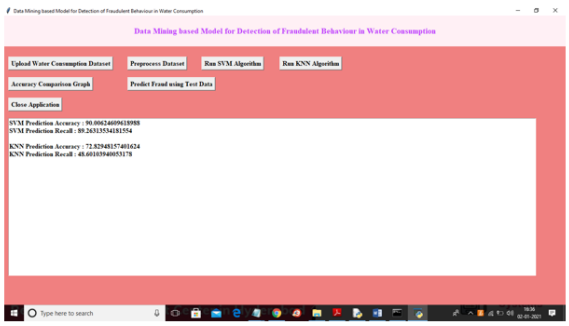


In above screen we can see some records are displaying from dataset and in last column we have string values as YES or NO but SVM or KNN only accept numeric values so we need to apply preprocessing to convert string to numeric data and then apply MINMAX scaler to convert all values between 0 and 1 and then remove all missing and imbalance values and to pre-process click on 'Preprocess Dataset' button and to get below screen

In above screen all values converted to numeric values and then dataset contains total 8002 records and then application using 6401 records to train algorithms and 1601 records to test algorithms prediction accuracy. Now both train and test data is ready and now click on 'Run SVM Algorithm' button to apply SVM on train and test data to get accuracy

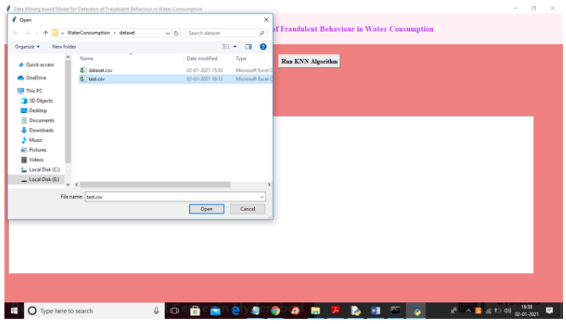


In above screen we got SVM accuracy as 90% and recall as 89% and now click on 'Run KNN Algorithm' button to apply train and test data on KNN to get its accuracy



In above screen KNN got 72% accuracy and SVM is better than KNN and now both algorithms trained model is ready and now click on 'Accuracy Comparison Graph' button to get below graph

In above screen x-axis represents algorithm name and y-axis represents accuracy of those algorithms and from above graph we can conclude that SVM is better than KNN and now click on 'Predict Fraud using Test Data' button to upload test data and predict fraud from that test data



In above screen selecting and uploading 'test.csv' file and then click on 'Open' button to get below prediction result

In above screen in brackets we can see uploaded test data and then after bracket we can see prediction result as 'No Fraud detected' or 'Fraud detected'. Similarly u can upload any test data and predict as fraud or no fraud

Conclusion:

Classification methods were employed in this study to identify customers that engage in fraudulent water usage practises. Models for detecting customers who appeared to be fraudsters were developed using classifiers based on SVM and KNN. The Cross Industry Standard Process for Data Mining was utilised to develop the models based on customer historical metered consumption data (CRISP-DM). Yarmouk Water Company (YWC) polled Qasabat Irbid ROU customers to gather water consumption data over a five-year period from 1.5 million unique consumers. Data mining classifiers like SVM and KNN required extensive pre-processing and formatting at this stage. Study participants were able to achieve an overall accuracy of 70% for both SVM and K-Neighbors (KNN) models. Nearly two-thirds of the YWC teams have a success rate greater than 60%, well exceeding the 1% of random manual checks. This model presents an intelligent tool that may be used by YWC in order to identify fraudulent clients and reduce losses. The proposed methodology can help Yarmouk water employees save time and effort by quickly identifying billing issues and corrupted metres in water metres. Utilities can improve their cost recovery by, for example, conducting onsite inspections of suspicious fraud customers in order to reduce administrative Non-Technical Losses (NTLs) and improve the productivity of their inspection employees.

References:

[1] N/A, "Jordan Water Sector Facts & Figures, Ministry of Water and irrigation of Jordan". Technical Report. 2015.

[2] N/A, "Water Reallocation Policy, Ministry of Water and irrigation of Jordan". Technical Report. 2016.

[3] C. Ramos , A. Souza , J. Papa and A. Falcao, "Fast non-technical losses identification through optimum-path forest". In Proc. of the 15th Int. Conf. Intelligent System Applications to Power Systems, 2009, pp.1-5.

[4] E. Kirkos, C. Spathis and Y. Manolopoulos, "Data mining techniques for the detection of fraudulent financial statements", Expert Systems with Applications, 32(2007): 995–1003.

[5] Y. Sahin and E. Duman, "Detecting credit card fraud by decision trees and support vector machines", IMECS, 2011, Vol I, pp. 16 – 18.

[6] S. Panigrahi, A. Kundu, S. Sural and A. Majumdar, "Credit card fraud detection: a fusion approach using dempster–shafer theory and bayesian learning, information fusion", 2009, 10(4): 354–363.

[7] N. Carneiro, G. Figueira and Costa M., "A data mining based system for credit-card fraud detection in e-tail decision support systems", Decision Support Systems, 2017, 95(C): 91-101.

[8] Ortega P., Figueroa C., and Ruz G. "A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile", In proc of DMIN, 2006.

[9] B. Kusaksizoglu, "Fraud detection in mobile communication networks using data mining", Bahcesehir University, The Department of computer engineering, Master Thesis. 2006.

[10] C. Liang-Chun, H. Chien-Lung, L.Nai-Wei, Y. Kuo-Hui and L. PingHsien, "Fraud analysis and detection for real-time messaging communications on social networks", IEICE Trans. Inf. & Syst., 2017, Vol. E100–D, No.10, pp: 2267-2274.