Eurasian Journal of Analytical Chemistry
ISSN: 1306-3057 OPEN ACCESS 2018 13 (SP): 369-379

A Survey of Existing Approaches for Water Pump Status Prediction

K. Karthik Prasad, Jayakrishna Bachu, E. Poovammal

Received 08 November 2018 ▪ Revised: 30 November 2018 ▪ Accepted: 06 December 2018

Abstract: Water Distribution Systems play an important role in shaping the quality of life of the public. However, we must realize that these systems are aging and have been deteriorated to the point that their service capabilities are now a matter of concern. Due to this fact there has been a growing awareness to improve the conditions of the water distribution system. The objective of the proposed system is to integrate the approaches followed by other systems to predict the functioning of a water pump by considering the various diverse factors that influence the water pump functioning such as water quality, source of the water, pressure on the system, age of the system, population density, various environmental factors, etc., and come up with a system that can predict the same with maximum efficiency.

INTRODUCTION

Water distribution system plays a vital role in the urban development and these pipeline systems have been laid and developed long time back. These systems have reached a point of deterioration often causing pipe burst leading to lot of social and economic stress. Water pipeline burst can be categorized as (1) structural deterioration and (2) functional deterioration.

Social and economic cost of pipe line failure is very expensive and very difficult to detect. Thus, raising need for strategy to solve the multi-objective problem. Most of the existing systems consider the factors that influence the distribution systems greatly, such as soil, water source, location, pressure on the system, etc and predict whether the system is in good condition, needs to be repaired or whether it is beyond repair and must be replaced completely.

For solving these problems, various evolutionary algorithms have been developed. Different mathematical simulations have been run on different load. Selection of exact algorithm for solving problem is difficult. Some algorithms may be better than other algorithms for a particular problem while at the same time could lead to a poor solution in other situations.

The system approach planned extends the present working systems with certain modifications which may lead to greater efficiency.

CURRENT TECHNIQUES

Deciding the Optimal Location to Set up the Representative Station

This system specifies methods to set up a representative station for the distribution system in an area depending upon certain factors. Setting up of a representative station is done as per the regulations specified by the U.S. Environmental Protection Agency that requires the water reaching the public to be safe for drinking which mandates all the drinking-water authorities to monitor the water quality in their distribution systems. The regulations (”Safe” 1974) prescribe the sampling frequency and the water-quality parameters to be monitored [1]. One catch is that the regulations do not specify how to set up the representative station.

K. Karthik Prasad, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India. E-mail: karthikprasad887@gmail.com

Jayakrishna Bachu, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India. E-mail: bachujaya1997@gmail.com

E. Poovammal, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India. E-mail: poovammal.e@ktr.srmuniv.ac.in

The system assumes that if a station is selected as the representative station and the quality of water is calculated at the station, then the quality of water at stations that are nearer to the representative station is known automatically. It also assumes that if the water quality at the representative station be determined, then the quality of water at stations that are upstream with respect to the representative station are similar to that at the representative station as the same water flows to the representative station.

The method to select to a representative station is given by: Each node is associated with a known demand. Thus, if the quality of water at a node i is known along with demand of the node i, i.e., di and the total demand D, then the fraction di/D of the total demand is known [1]. Then a cut-off percentage is determined above which a station can be considered as a representative station. Mathematically, the objective is to maximize the function given by equation 1:

$$\sum_{i=1}^{n} d_i y_i \qquad (1)[1]$$

Figure 1 illustrates how the selection of the representative station is made. In the figure, node 5 is assumed to be the representative station and the fractional flow to every node is cal- culated where d indicates the demand and fi indicates the flow of water through that station.

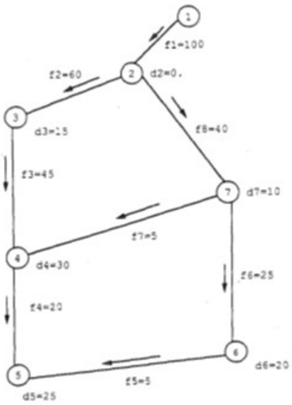


Figure 1: Water Distribution System

The mentioned system takes help from existing models such as the Kentucky Pipe Model and the WADSIO in order to determine the best location for a representative station. An addition to the system could have been to mention a solution to compare multiple sampling sites as there exists no quantitative measure to do the same.

Urban Water Quality Prediction

This system highlights the fact that predicting water quality in urban areas is more challenging than it sounds since the quality of water in an urban environment varies non-linearly and depends on various environmental factors such as meteorology, water-usage patterns, and land uses [2]. The entire system is built upon datasets of different domains such as meteorology, pipe networks, structure of road networks, etc., which are then combined using multi-task multi- view learning method.

The system consists of two components: multi-task learning and multi-view learning respectively and each component has a unique role. The multi-view learning tries to acquire two different views within a station while the multi-task learning acquires the relationships of the pipe networks among different stations. This method can help improve the predictive capability of all the stations [2].

The formula for the spatio-temporal multi-task multi-view learning is given by equation 2:

$$\min_{\mathbf{W}} \quad \frac{1}{2}\sum_{l=1}^{M} \|\mathbf{y}_l - \frac{1}{2}\mathbf{X}_l\mathbf{w}_l\|_2^2 + \lambda \sum_{l=1}^{M} \|\mathbf{X}_l^s\mathbf{w}_l^s - \mathbf{X}_l^t\mathbf{w}_l^t\|_2^2$$
$$+\gamma \sum_{l,m=1}^{M} C_{l,m}\|\mathbf{w}_l - \mathbf{w}_m\|_2^2 + \theta\|\mathbf{W}\|_{2,1}, \qquad (2)[2]$$

where the first term is the loss function, the second term is the multi-view learning component, the third term is multi-task learning component and λ, γ and θ are regularization parameters.

Next, the system applies the multi-view learning to gather data from various domains and combines them into one single component which will be able to describe the properties of the object of study furthermore and thus enhance the results. The proposed system achieves a good success percentage by combing the data from various domains, but the system could have used a much simpler approach as the data needed should be in real-time and there is an increased cost factor due to installations of real-time monitoring systems.

Real-Time Water Quality Prediction

This system makes use of a hypothetical city A in which the model operates to determine the energy utilized by an old plant and a new plant and also the amount of pressure the entire distribution system experiences due to addition of new water sources. The system mainly focuses on the pipe network of the city A.

The first thing the system considers is the current age of the pipe network and the area they are serving. To proceed, the model utilizes two formulae given by the equations 3 and 4:

$$\sum Q_{in} - \sum Q_{out} = Q_{demand} \quad (3)[3]$$

$$H_i - H_j = H_l = \frac{4.73LQ^{4.73}}{C^{1.852}D^{4.87}} \quad (4)[3]$$

where equation 3 represents the Hazen-William equation used to derive a relationship between energy loss and pipe flow. Using the above equation, a hydraulic model was constructed and an analysis regarding the energy loss was carried out and certain benchmarks were set which the model had to satisfy, which by the way it did. A conclusion was drawn stating the constraints that the system had to manage to achieve the objective.

Prediction of Water Distribution System Based on Bayesian Distribution System

This study predicts any break in the pipeline network in water distribution system. It uses the Bayesian neural network for diagnostic and predictive management of water distribution system.

The system uses data such as average household age at census tract and population density of the area for determining the average age of the pipeline network. They took different samples of soil to determine its quality as well. They used United States Department of Agriculture National Resource Conservation Services (NRCS) soil data mart to capture the soil characteristics. [4]

Bayesian Neural Network (BNN) was learned using following learning algorithm: (i) Grow-shrink Algorithm

1. RsMax2 Algorithm
2. Gradient Decent Algorithm

Statistical learning may be hypothesis generating and might be less resourceful. Flexibility is required in both continuous data and discreet data. There is a need for improved discretization. Discrete observation may be a natural knowledge base for the BNN otherwise discretization of each variable has to justify and the interpretation of resultant BNN model.

Pipe Failure Methods in Water Distribution System

This system states that the water distribution system can break due to many reasons. Some of the reasons are: i) There is a single source of element and there is failure at any point in the pipeline. ii) Water demand is not met, this can be because of failure of booster pump.

The first type of failure is further classified into: a) failure in branch network and b) failure in a loop network. When a failure occurs in branch network, i.e., a network without any loops, then the water supply to the entire network is cut-off as the network is linear and the water has to pass through the entire network sequentially. However, when a failure occurs in a loop network, only a portion of the water supply is cut-off which increases the demand pressure on the remaining nodes due reduced water supply. In both the cases, if a short circuit occurs, i.e., if the water from the source finds a path to reach the last node without flowing through the network then the water supply to the network ceases.

The second type of failure is caused due excess demand in the network, restricting certain paths in the network due to maintenance and repair, or the failure maybe progressive. All in all, the pumps are made to work more than what they can handle due to which they fail which leads to the entire water supply being cut-off [5].

The formula for calculating the Micro Flow Distribution in performance assessment is given by the equation 5:

$$C_i = C_i^* \left[1 - a_i \exp\left(-b_i \frac{SP_i}{P_i^*}\right)\right]$$ (5)[5]

where C = nominal demand at a node, SP = service pressure at a node and a, b, P are constants.

The reliability estimation using Micro Flow Distribution is given by equation 6:

$$R = 1 - \sum_{ij} p_{ij}$$ (6)[5]

Uncertainty of Parameter Selection and Its Effects on Determining Water Quality

Monte Carlo simulations (MCS) was used to determine the effect on water quality due to uncertainty in selection of parameters. Advective transport and mass conservation at a node determines the water quality. Formula for conservation of mass for steady system at junction of two or more pipes is given by equation 7:

where Ql refers to pipe flow and qi refers to external demand or supply. Mass conservation during advective flow is given by equation 8:

$$\sum_{l \in J_{\text{in},i}} Q_l - \sum_{l \in J_{\text{out},i}} Q_l = q_i$$ (7)[6]

$$\partial C/\partial t + V\partial C/\partial x = r(C)$$ (8)[6]

where dfC/dx refers to rate of change of concentration in inflow to outflow, dfC/dt refers to rate of change of constituent concentration of different elements and r(C) refers to reaction relationship.

The reaction is due to decay that is mainly caused by a reaction between water and minerals of pipe and is given by equation 9:

$$r(C) = k(C - C^*)C^{nc-1}$$ (9)[6]

where k is the reaction constant and k = kb+kwall.

It was found that output uncertainty increased with input uncertainty by about 1.75 times and standard deviation increased by doubling the coefficient value of kw. Due to bulk decay coefficient and global coefficient the output uncertainty increased faster than linear as coefficient values were increased [6].

Estimation of Leakage for Water Distribution System

The system proposes two methods for quantifying the water leaks in the water distribution system. First is the minimum night flow and second is the methodology of estimation of water leakage using advanced computational calculations.

Amount lost using minimum night flow method is calculated as mentioned in equation 10:

$$DRLV = F_{nd} \times Q_{mn}$$ (10)[7]

where DRLV refers to Daily Real Volume Loss, Fnd refers to night-day factor and Qmn

refers to avg minimum night leak. Fnd is calculated by the sum of pressure acquired for 24 hours as shown in equation 11:

$$F_{nd} = \sum_{i=0}^{24} \left(\frac{P_i}{P_{3-4}}\right)^{N1}$$ (11)[7]

where P is the average pressure and P3-4 is the average pressure during 3-4pm.

In the second method epanet calibrator software is used for calibrating the water losses. Pressure data and observed flow are inserted into the software for calibration. The formula used for calibration is given by equation 12:

$$Q = C_e H^{N1}$$ (12)[7]

where N1 is 0.5 for metallic pipes, Ce is discharge coefficient and H is the nodal head.

This study suggested to use computational method for quantifying the water leakage in the distribution system [7].

Relationship between Pressure-Leakage of Some Failed Water Pipes

This study presented the data according to which the leakage exponent for corrosion holes was between 0.67 and 2.3, for round holes it was 0.5, for circumferential holes it was between 0.41 to 0.52 and for longitudinal cracks it was between 0.79 and 1.85.

In this study two ends of a broken water pipeline system were fitted with water distribution system and calibrated pressure system. Pressure transducer and flow meter reading was collected on a logger. Readings were taken after every 30 seconds. It was understood that flow and pressure included fluctuations in short term. The amplitude of fluctuations is low with low pressure and flow. This is caused due to effect of dampening on throttle valve.

It was observed that by using the asbestos cement pipes the distribution system failed. There were longitudinal cracks and the leakage component was much above 0.5. It was also observed that pipes which had parallel longitudinal crack had maximum leakage and diagonal cracks had minimum leakage.

It was also observed that with steel pipes higher the corrosion higher the exponent. Due to reduced supporting area because of corrosion, stress will increase thus causing increased pressure on the pipe. But for round hole pipe exponent is approximately equal to the theoretical value.

In case of longitudinal cracks exponent value is much higher than exponential value. With increase of length there was increase in exponent. For circumferential cracks the exponential value was much higher than the longitudinal stress due to the reason that circumferential stresses cracks elongates [8].

Water Distribution System with Both Distribution Leaks and Pressure Dependent Demands

This system proposes an efficient model to predict better results when new scenarios were run with the values of time period and coefficient related to the node. It is proposed that modulation of flow rate at nodes should be preserved.

Equation 13 is used for determining the total consumption at a node:

$$Q_{c,i} = Q_{d,i} + q_i \quad (13)[9]$$

where Qc,i refers to consumption at node i, Qd,i refers to demand at node i and qi refers to quantity loss.

Leakage could be characterized using equation 14:

$$q_i = K_i \left(p_i - p_o\right)^{\beta} \quad (14)[9]$$

where pi stands for pressure upstream, p0 stands for pressure downstream and Ki stands for orifice constant.

Finally, the demand can be determined using equation 15:

$$Q_{d,i}(k) = Q_{c,i}(k) - q_{1,i}(k) - q_{2,i}(k) \quad (15)[9]$$

Using the above different calculations an integrated water distribution system can be developed. In this paper the author has differentiated the leakage from modelled network and from non-modelled network. Parameters such as age of the pipe, total demand on them and average pressure at each node were considered [9].

Population Growth on Water Quality

In today's world, water pollution due to human activities is a major cause of concern and beyond a certain threshold limit, it poses a threat to the quality of water in the water body [10, 11]. A study shows that urban activities is one of the major causes for water pollution in the African and Asian countries [11]. Many studies that have been conducted to determine the amount of pollution mainly focus on the following parameters: i) biochemical oxygen demand (BOD), ii) dissolved nitrates-nitrogen (NO3-N), iii) dissolved oxygen (DO), iv) soluble phosphorous (PO4-P), v) total coliform (TC), vi) chlorophyll, and other substrates to base their conclusions [10, 11]. All the above parameters use the unit of parts per million (ppm) or milligram per liter (mg/L).

A study was conducted in Lake Victoria, Kenya to determine the amount of pollution due to human activities. A methodological way was followed wherein first, the physical examination of the area was carried out and the physical characteristics of the lake was determined such as the area of the lake, the depth, its longitudinal and latitudinal positions, percentage of lake shared between the countries it falls in and also the population density range in those regions from a given year to the present. Next, the area around the lake was examined to find out the amount of wildlife the lake supported, the forest coverage, and the major industrial projects of which the lake was a part, such as, the hydroelectric project. In the final step, information about the environmental factors was collected, like rainfall, temperature ranges in a year and the terrain characteristics. Then, a sample of water from the lake was collected and sent to the Lake Victoria Environmental Management Project (LVEMP) laboratories for analysis. The data was then compared with the previous data acquired in a similar manner to estimate the amount of pollution caused. The results showed an increase in the level of NO3-N and PO4-P contents which increased the nutrient content of the lake due to increased surface runoffs which intern increased the amount of water hyacinth in the region which depleted the dissolved oxygen in the lake which posed a problem [10].

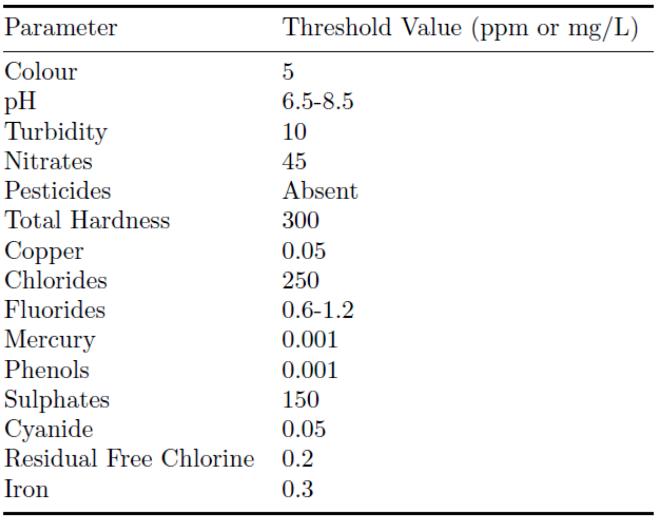
A similar study was conducted to determine the influence of population growth on the water quality of the Kelani River, Sri Lanka. The parameters considered were BOD, DO and TC. The study used a Bayesian Network classification approach to determine the appropriate range of population in a region around the waterbody for the water to be suitable for drinking and other domestic purposes. Next, actual data was collected at three points in the course of the river, i.e., at the two ends and in the middle section. The methods used for the analysis process was as specified by the Central Environmental Agency (CEA) of Sri Lanka and the quality control tests were as per the regulations of the American Water Works Association (AWWA), American Public Health Association (APHA) and the 2005 standards of the Water Environment Federation (WEF). The study then collected monthly data from the start of 2003 to the end of

2013 with some missing values due to technical and human errors [11]. The conclusion the study arrived at was like that of the study of Lake Victoria, Kenya, that the population in the regions surrounding the river kept rising due to which the nutrient content in the river also increased which made the water not fit for drinking purposes.

Water Quality Parameters for Domestic and Drinking Purposes

It is known that water covers around 70 percent of the Earth's surface and only a fraction of it can be used for human purposes. So, it is very essential to maintain the quality of the fresh water bodies which are very vulnerable, finite and play an important role in sustaining life. Scientists estimate that the total water on the Earth's surface contain physicochemical parameters such as, color, pH, BOD, chemical oxygen demand (COD), DO, total dissolved solids (TDS), turbidity, total suspended solids (TSS), and others, which make-up about 1.36 billion cubic kilometers of the total water content [12]. These parameters can be used to estimate the water quality of a water body. Table 1 illustrates some of the water quality parameters and their threshold limit as specified by the Indian Standard Specifications for Drinking Water, IS:10500.

Table 1: Water Quality Parameters



Parameter	Threshold Value (ppm or mg/L)
Colour	5
pH	6.5-8.5
Turbidity	10
Nitrates	45
Pesticides	Absent
Total Hardness	300
Copper	0.05
Chlorides	250
Fluorides	0.6-1.2
Mercury	0.001
Phenols	0.001
Sulphates	150
Cyanide	0.05
Residual Free Chlorine	0.2
Iron	0.3

Some of the other parameters include temperature which directly impacts the aquatic communities, conductivity which indicates the presence of free ions mainly due to increase in salinity, TSS which indicates the amount of erosion upstream from a particular place, TDS which indicates the amount of inorganic salts and dissolved materials in the water which mainly includes chemical compounds such as nitrates, calcium, magnesium, sodium, potassium, carbonates, chlorides, sulphates, etc., BOD which indicates the amount of organic pollution to both surface water and waste, COD which is used along with BOD indicates the amount of organics in water, Ammonia Nitrogen which in high levels may harm the aquatic life by increasing the body pH or by altering the metabolism, Potassium which indicates the amount of fertilizers being used in an area as it directly relates to plant growth being one of the macro elements required by the plants, DO which indicates the amount of oxygen dissolved in water which is essential for sustaining life [12].

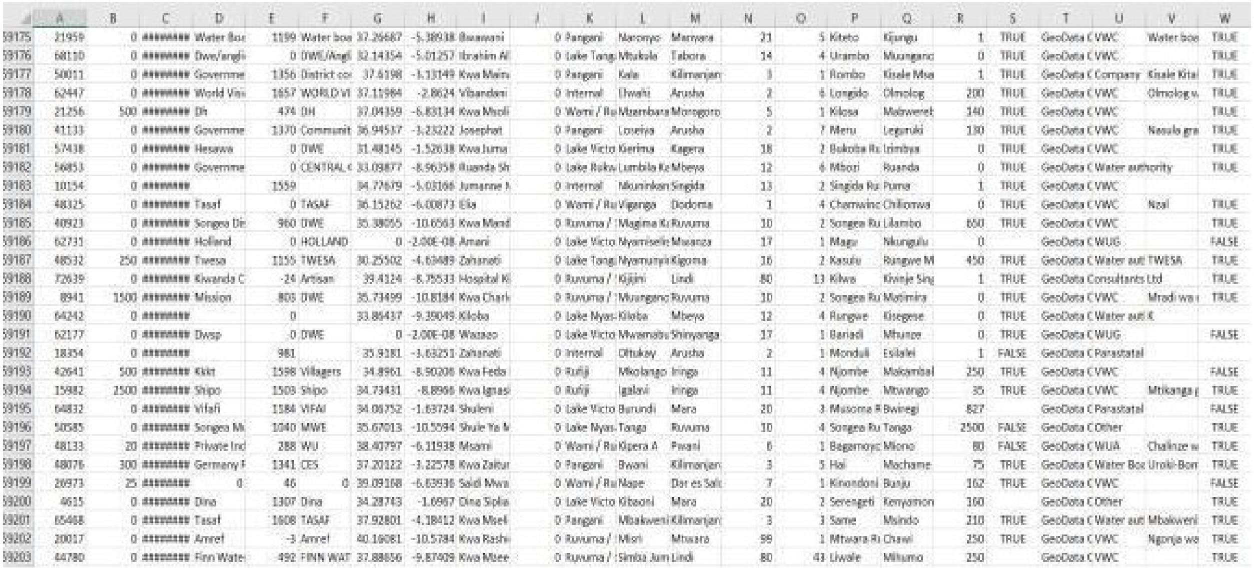
A study was conducted to estimate the water quality parameters of the Cauvery River in the Erode region, Tamil Nadu, India. Cauvery river originates in the foothills of the Western Ghats in Karnataka, India and flows for about 800 km before emptying into the Bay of Bengal. During its course, it flows through the Erode District in Tamil Nadu where the study was conducted. The study revealed the pH as 7.86, TSS as 690mg/L, TDS as 1004mg/L, total solids as 1580mg/L, total hardness as 340mg/L, chloride as 380mg/L, DO as 5.59mg/L, BOD as 38mg/L, COD as 304mg/L, phosphate as 6.0mg/L, electrical conductivity as 920S/cm, and sulphates as 60mg/L [13]. When the results of the study were compared with the IS:10500 standard, it can be observed that most of the water quality parameters were well beyond their prescribed limit which indicates the water was not fit for drinking purposes and had to be maintained better to make it fit for drinking and other domestic purposes.

PROPOSED SYSTEM AND TRAINING DATA

The proposed system is that of a simple linear Support Vector Machine (SVM) to model the given data and achieve the objective of predicting the status of the water pumps. The SVM was selected particularly because though it is a very simple model, it can easily model real world problems, performs well on dataset with many attributes and its functional form is very similar to that of the neural networks and most importantly it is cost effective. One speed breaker is that the SVM works only on numerical data, but the presented dataset has only character data due to which the dataset has to be first converted into a numerical dataset and all the empty values must be managed properly which otherwise will lead to faulty prediction. Since the dataset is quite large, the method of cross validation is applied wherein the training dataset is divided into two parts, one for training and the other for testing. Once the training and cross validation is done, a rough estimate can be drawn regarding the attributes to be used based on the percentage efficiency of the model. It also paves the way to many techniques that can be applied on the model to further improve the prediction efficiency.

If the above system fails to achieve the desired result then Artificial Neural Networks (ANN) will be employed to achieve the desired solution.

Figure 2 shows a part of the dataset the model must work on. The dataset nearly contains 41 attributes covering a wide variety of domains and nearly 59401 training values and 14851 test values.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
59175	21959	0	########	Water Boa	1199	Water boa	37.26687	-5.38938	Bwawani		0	Pangani	Naronyo	Manyara	21	5	Kiteto	Kijungu	1	TRUE	GeoData C	VWC	Water boa	TRUE
59176	68110	0	########	Dwe/angli	0	DWE/Angl	32.14354	-5.01257	Ibrahim Al		0	Lake Tang	Mtukula	Tabora	14	4	Urambo	Muungano	0	TRUE	GeoData C	VWC		TRUE
59177	50011	0	########	Governme	1356	District co	37.6198	-3.13149	Kwa Mains		0	Pangani	Kala	Kilimanjaro	3	1	Rombo	Kisale Msa	1	TRUE	GeoData C	Company	Kisale Kitai	TRUE
59178	62447	0	########	World Visi	1657	WORLD VI	37.11984	-2.8624	Vibandani		0	Internal	Elwahi	Arusha	2	6	Longido	Olmolog	200	TRUE	GeoData C	VWC	Olmolog w	TRUE
59179	21256	500	########	Dh	474	DH	37.04359	-6.83134	Kwa Msoli		0	Wami / Ru	Mzambara	Morogoro	5	1	Kilosa	Mabwerebe	140	TRUE	GeoData C	VWC		TRUE
59180	41133	0	########	Governme	1370	Communit	36.94537	-3.23222	Josephat		0	Pangani	Loseiya	Arusha	2	7	Meru	Leguruki	130	TRUE	GeoData C	VWC	Nasula gra	TRUE
59181	57438	0	########	Hesawa	0	DWE	31.48145	-1.52638	Kwa Juma		0	Lake Victo	Kierima	Kagera	18	2	Bukoba Ru	Irimbya	0	TRUE	GeoData C	VWC		TRUE
59182	56853	0	########	Governme	0	CENTRAL G	33.09877	-8.96358	Ruanda Sh		0	Lake Rukw	Lumbila Ka	Mbeya	12	6	Mbozi	Ruanda	0	TRUE	GeoData C	Water authority		TRUE
59183	10154	0	########		1559		34.77679	-5.03166	Jumanne M		0	Internal	Nkuninkan	Singida	13	2	Singida Ru	Puma	1	TRUE	GeoData C	VWC		
59184	48325	0	########	Tasaf	0	TASAF	36.15262	-6.00873	Ella		0	Wami / Ru	Vigingo	Dodoma	1	4	Chamwino	Chilonwa	0	TRUE	GeoData C	VWC	Nzal	TRUE
59185	40923	0	########	Songea Di	960	DWE	35.38055	-10.6563	Kwa Mand		0	Ruvuma /	Magima Ki	Ruvuma	10	2	Songea Ru	Lilambo	650	TRUE	GeoData C	VWC		TRUE
59186	62731	0	########	Holland	0	HOLLAND	0	-2.00E-08	Amani		0	Lake Victo	Nyamiselin	Mwanza	17	1	Magu	Nkungulu	0		GeoData C	WUG		FALSE
59187	48532	250	########	Twesa	1155	TWESA	30.25502	-4.63489	Zahanati		0	Lake Tang	Nyamunyis	Kigoma	16	2	Kasulu	Rungwe M	450	TRUE	GeoData C	Water aut	TWESA	TRUE
59188	72639	0	########	Kiwanda C	-24	Artisan	39.4124	-8.75533	Hospital Ki		0	Ruvuma /	Kijijini	Lindi	80	13	Kilwa	Kivinje Sing	1	TRUE	GeoData Consultants Ltd			TRUE
59189	8941	1500	########	Mission	803	DWE	35.73499	-10.8184	Kwa Chark		0	Ruvuma /	Muungano	Ruvuma	10	2	Songea Ru	Matimira	0	TRUE	GeoData C	VWC	Mradi wa	TRUE
59190	64242	0	########		0		33.86437	-9.39049	Kiloba		0	Lake Nyas	Kiloba	Mbeya	12	4	Rungwe	Kisegese	0	TRUE	GeoData C	Water aut	K	
59191	62177	0	########	Dwsp	0	DWE	0	-2.00E-08	Wazazo		0	Lake Victo	Mwamabu	Shinyanga	17	1	Bariadi	Mhunze	0	TRUE	GeoData C	WUG		FALSE
59192	18354	0	########		981		35.9181	-3.63251	Zahanati		0	Internal	Oltukay	Arusha	2	1	Monduli	Esilalei	1	FALSE	GeoData C	Parastatal		
59193	42641	500	########	Kkkt	1598	Villagers	34.8961	-8.90206	Kwa Feda		0	Rufiji	Mkolango	Iringa	11	4	Njombe	Makambal	250	TRUE	GeoData C	VWC		FALSE
59194	15982	2500	########	Shipo	1503	Shipo	34.73431	-8.8966	Kwa Ignasi		0	Rufiji	Igalavi	Iringa	11	4	Njombe	Mtwango	35	TRUE	GeoData C	VWC	Mtikanga	TRUE
59195	64832	0	########	Vifafi	1184	VIFAFI	34.06752	-1.63724	Shuleni		0	Lake Victo	Burundi	Mara	20	3	Musoma R	Bwiregi	827		GeoData C	Parastatal		FALSE
59196	50585	0	########	Songea Mu	1040	MWE	35.67013	-10.5594	Shule Ya M		0	Lake Nyas	Tanga	Ruvuma	10	4	Songea Ru	Tanga	2500	FALSE	GeoData C	Other		TRUE
59197	48133	20	########	Private Ind	288	WU	38.40797	-6.11938	Msami		0	Wami / Ru	Kipera A	Pwani	6	1	Bagamoyo	Miono	80	FALSE	GeoData C	WUA	Chalinze w	TRUE
59198	48076	300	########	Germany F	1341	CES	37.20122	-3.22578	Kwa Zaitun		0	Pangani	Bwani	Kilimanjaro	3	5	Hai	Machame	75	TRUE	GeoData C	Water Boa	Uroki-Bom	TRUE
59199	26973	25	########	0	46	0	39.09108	-6.63936	Saidi Mwa		0	Wami / Ru	Nape	Dar es Sala	7	1	Kinondoni	Bunju	162	TRUE	GeoData C	VWC		FALSE
59200	4615	0	########	Dina	1307	Dina	34.28743	-1.6967	Dina Siplia		0	Lake Victo	Kibaoni	Mara	20	2	Serengeti	Kenyamon	160		GeoData C	Other		TRUE
59201	65468	0	########	Tasaf	1608	TASAF	37.92801	-4.18412	Kwa Mseli		0	Pangani	Mbakweni	Kilimanjaro	3	3	Same	Msindo	210	TRUE	GeoData C	Water aut	Mbakweni	TRUE
59202	20017	0	########	Amref	-3	Amref	40.16081	-10.5784	Kwa Rashi		0	Ruvuma /	Misri	Mtwara	99	1	Mtwara Ru	Chawi	250	TRUE	GeoData C	VWC	Ngonja wa	TRUE
59203	44780	0	########	Finn Wate	492	FINN WAT	37.88656	-9.87409	Kwa Mzee		0	Ruvuma /	Simba Jum	Lindi	80	43	Liwale	Mihumo	250		GeoData C	VWC		TRUE

Figure 2: Part of the Training Data Set

RESULTS AND DISCUSSION

After applying the simple linear SVM on the dataset, a percentage success of 58.41 was achieved. Next K- Nearest Neighbor algorithm was applied on the same dataset and a percent-age success of 70 was achieved. Finally, K-means Clustering technique was used and it resulted in a percentage success of mere 41.81. Figure 3 and 4 shows the comparison between the three different approaches on the same dataset.

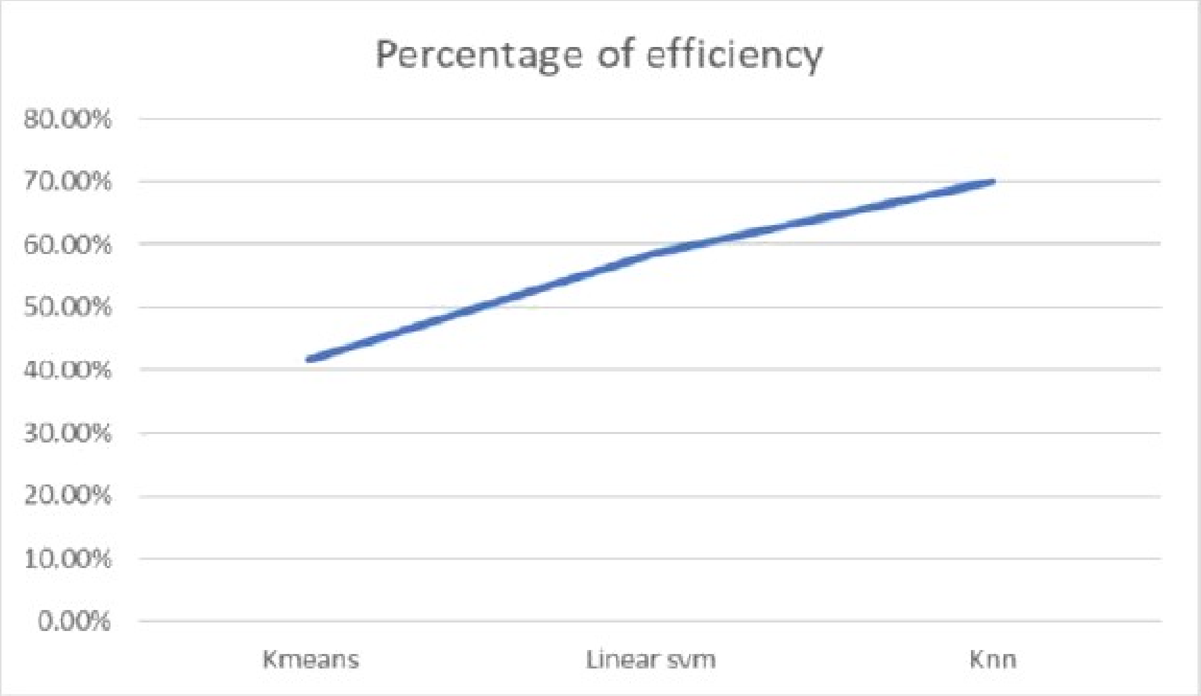


Figure 3: Line Graph Determining the Percentage Efficiency

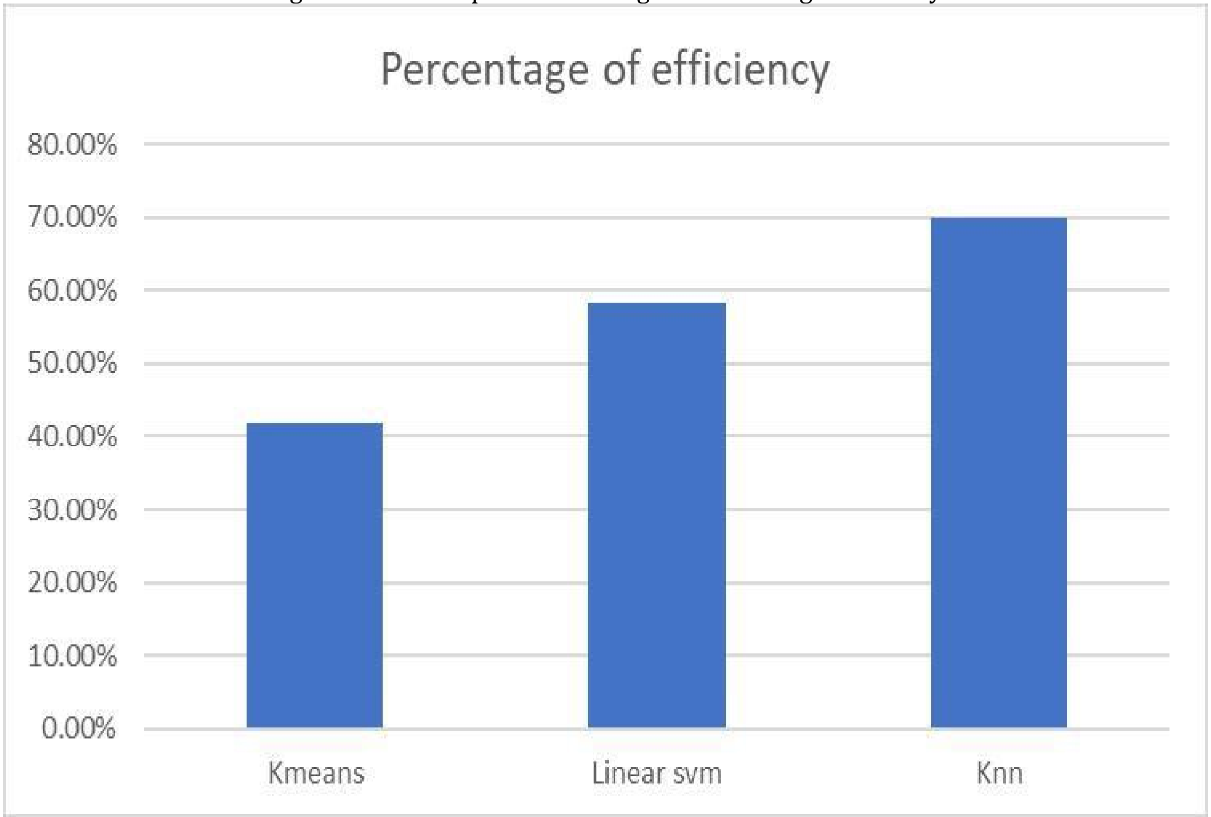
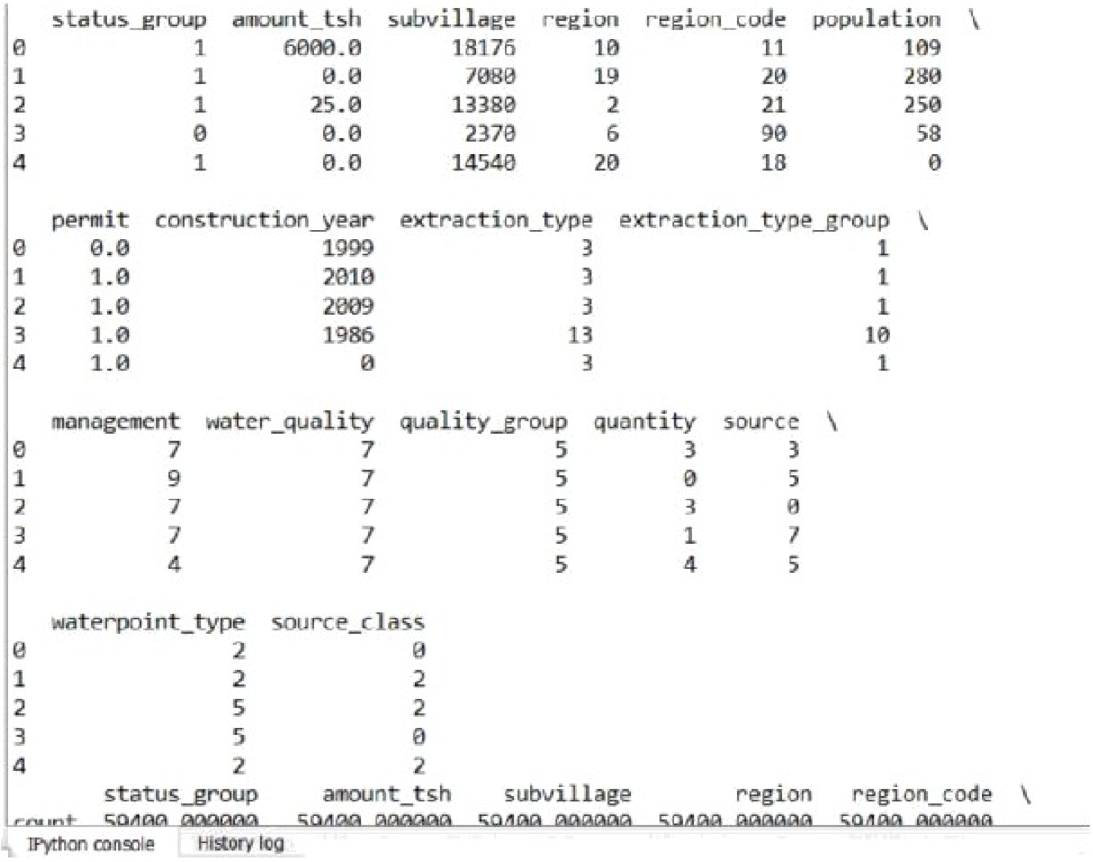


Figure 4: Line Graph Determining the Percentage Efficiency

Figures 5, 6 and 7 show the dataset after converting the character data to numerical data and cleaning, i.e., only keeping the attributes necessary for the prediction process.



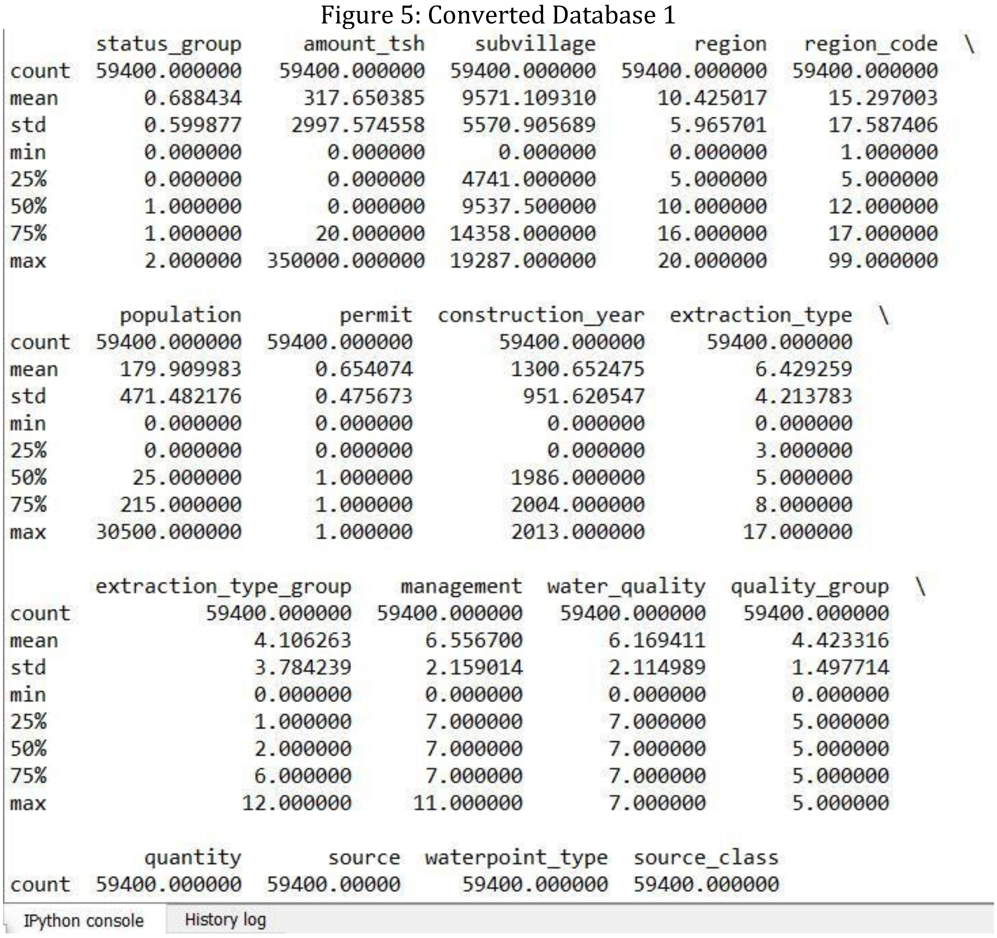
```
   status_group  amount_tsh  subvillage  region  region_code  population  \
0             1      6000.0       18176      10           11         109
1             1         0.0        7080      19           20         280
2             1        25.0       13380       2           21         250
3             0         0.0        2370       6           90          58
4             1         0.0       14540      20           18           0

   permit  construction_year  extraction_type  extraction_type_group  \
0     0.0               1999                3                      1
1     1.0               2010                3                      1
2     1.0               2009                3                      1
3     1.0               1986               13                     10
4     1.0                  0                3                      1

   management  water_quality  quality_group  quantity  source  \
0           7              7              5         3       3
1           9              7              5         0       5
2           7              7              5         3       0
3           7              7              5         1       7
4           4              7              5         4       5

   waterpoint_type  source_class
0                2             0
1                2             2
2                5             2
3                5             0
4                2             2
       status_group     amount_tsh     subvillage         region    region_code  \
count  59400.000000   59400.000000   59400.000000   59400.000000   59400.000000
```

Figure 5: Converted Database 1



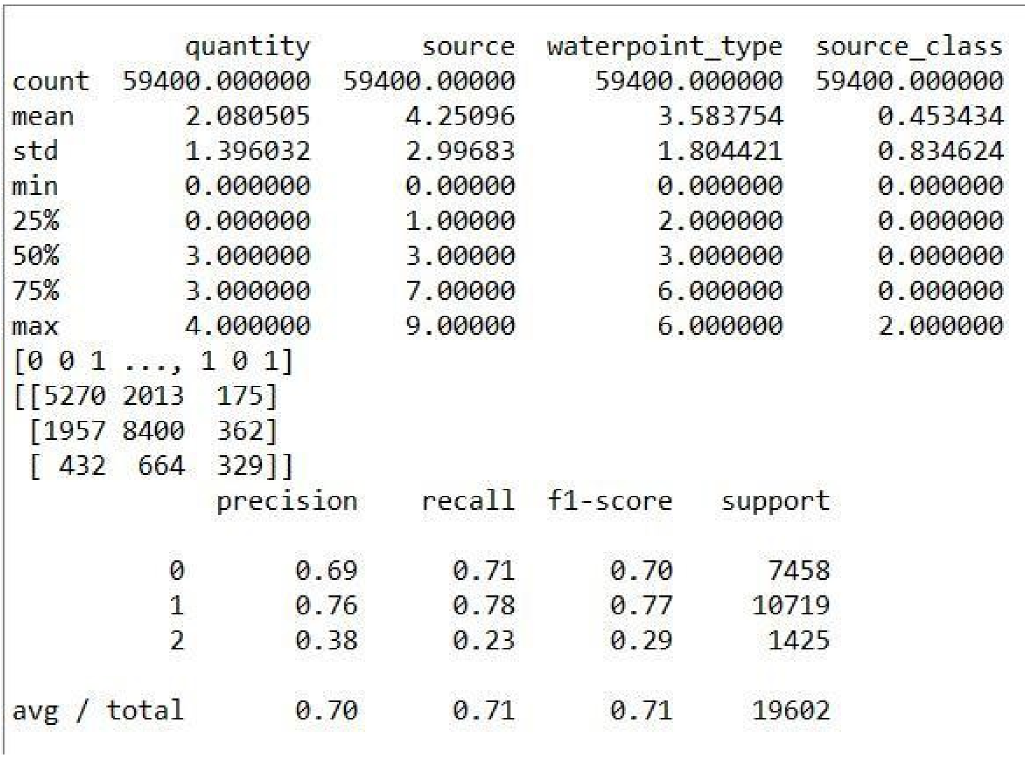
```
       status_group     amount_tsh    subvillage        region   region_code  \
count  59400.000000   59400.000000  59400.000000  59400.000000  59400.000000
mean       0.688434     317.650385   9571.109310     10.425017     15.297003
std        0.599877    2997.574558   5570.905689      5.965701     17.587406
min        0.000000       0.000000      0.000000      0.000000      1.000000
25%        0.000000       0.000000   4741.000000      5.000000      5.000000
50%        1.000000       0.000000   9537.500000     10.000000     12.000000
75%        1.000000      20.000000  14358.000000     16.000000     17.000000
max        2.000000  350000.000000  19287.000000     20.000000     99.000000

         population        permit  construction_year  extraction_type  \
count  59400.000000  59400.000000       59400.000000     59400.000000
mean     179.909983      0.654074        1300.652475         6.429259
std      471.482176      0.475673         951.620547         4.213783
min        0.000000      0.000000           0.000000         0.000000
25%        0.000000      0.000000           0.000000         3.000000
50%       25.000000      1.000000        1986.000000         5.000000
75%      215.000000      1.000000        2004.000000         8.000000
max    30500.000000      1.000000        2013.000000        17.000000

       extraction_type_group    management  water_quality  quality_group  \
count           59400.000000  59400.000000   59400.000000   59400.000000
mean                4.106263      6.556700       6.169411       4.423316
std                 3.784239      2.159014       2.114989       1.497714
min                 0.000000      0.000000       0.000000       0.000000
25%                 1.000000      7.000000       7.000000       5.000000
50%                 2.000000      7.000000       7.000000       5.000000
75%                 6.000000      7.000000       7.000000       5.000000
max                12.000000     11.000000       7.000000       5.000000

           quantity       source  waterpoint_type  source_class
count  59400.000000  59400.00000     59400.000000  59400.000000
```

Figure 6: Converted Database 2



```
            quantity       source  waterpoint_type  source_class
count  59400.000000  59400.00000     59400.000000  59400.000000
mean       2.080505      4.25096         3.583754      0.453434
std        1.396032      2.99683         1.804421      0.834624
min        0.000000      0.00000         0.000000      0.000000
25%        0.000000      1.00000         2.000000      0.000000
50%        3.000000      3.00000         3.000000      0.000000
75%        3.000000      7.00000         6.000000      0.000000
max        4.000000      9.00000         6.000000      2.000000
[0 0 1 ..., 1 0 1]
[[5270 2013  175]
 [1957 8400  362]
 [ 432  664  329]]
             precision    recall  f1-score   support

          0       0.69      0.71      0.70      7458
          1       0.76      0.78      0.77     10719
          2       0.38      0.23      0.29      1425

avg / total       0.70      0.71      0.71     19602
```

Figure 7: Converted Database 3

To improve the efficiency further, the method of information gain is suggested and is being applied on the model currently.

REFERENCES

[1] Lee B.H and Deininger R.A 1992 Jan Optimal locations of monitoring stations in water distribution system *Journal of Environmental Engineering* **118(1)** 4-16

[2] Liu Y, Liang Y, Liu S, Rosenblum D.S and Zheng Y 2016 Oct 29 Predicting urban water quality with ubiquitous data *arXiv preprint arXiv:1610.09462*

[3] Shihu S, Dong Z, Suiqing L, Mingqun M, Ming Z, Yixing Y, Jinliang G and Hongbin Z 2010 May 7 Decision support system of water distribution network expansion *2010 International Conference on E- Business and E-Government. IEEE* (pp. 1520-23)

[4] Francis R.A, Guikema S.D and Henneman L 2014 Oct 1 Bayesian belief networks for predicting drinking water distribution system pipe breaks *Reliability Engineering & System Safety* **130,** 1-11.

[5] Jowitt P.W and Xu C 1993 Jan Predicting pipe failure effects in water distribution networks *Journal of Water Resources Planning and Management* **119(1)** 18-31

[6] Pasha M.F and Lansey K 2010 Jan 1 Effect of parameter uncertainty on water quality predictions in distribution systems-case study *Journal of hydro informatics* **12(1)** 1-21

[7] Cheung P.B, Girol G.V, Abe N and Propato M 2010 Night flow analysis and modeling for leakage estimation in a water distribution system *Integrating water systems* 509-13

[8] Greyvenstein B and Van Zyl J.E 2007 Mar 1 An experimental investigation into the pressure-leakage relationship of some failed water pipes *Journal of Water Supply: Research and Technology-AQUA* **56(2)** 117-24

[9] Martinez F, Conejos P and Vercher J 1999 Developing an integrated model for water distribution systems considering both distributed leakage and pressure-dependent demands *In WRPMD'99: Preparing for the 21st Century* (pp. 1-14)

[10] Juma D.W, Wang H and Li F 2014 Apr 1 Impacts of population growth and economic development on water quality of a lake: case study of Lake Victoria Kenya water *Environmental Science and Pollution Research* **21(8)** 5737-46

[11] Liyanage C and Yamada K 2017 Aug Impact of population growth on the water quality of natural water bodies *Sustainability* **9(8)** 1405

[12] Tiwari S 2015 Water quality parameters–A review *International Journal of Engineering Science*

Invention Research & Development **1(9)** 319-24

[13] Appavu A, Thangavelu S, Muthukannan S, Jesudoss JS and Pandi B 2016 Study of water quality parameters of cauvery river water in erode region *Journal of Global Biosciences* **5(9)** 4556-67

[14] Vishwakarma S, Varma A and Saxena G 2013 Apr 30 Assessment of water quality of Betwa River, Madhya Pradesh, India *International Journal of Water Resources and Environmental Engineering* **5(4)** 117-222