International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

Vol. 32 No. 2 May 2022, pp. 359-379

Groundwater Sodium Levels Estimation of Proposed Irrigation Groundwater Source for the Kumahumato Settlement of the Dadaab Subcounty, North Eastern Kenya

Dr. Meshack Owira Amimo[1] and Jibril A. Shune[2]

[1]Water Resources Authority Headquarters
Nairobi, Kenya

[2]Assistant Hydrologist, Northern Water Works Development Agency
Garissa, Kenya

***Abstract*—The Project Area of Kumahumato is a locality located on the fringes of the Merti aquifer within a radii ranging from 5 to 10 kilometer metric units. The area is primarily inhabited Nomadic pastoralists who have limited experience with matters farming and allied agricultural techniques. Owing to the rapid change in climate patterns and with massive death toll of livestock resulting from prolonged droughts and unpredictable rains, the community leadership have deemed it fit to focus on irrigation-aided agriculture. One problem noted is that the sodium levels in the soils may be exacerbated by irrigation farming, if the groundwater sodic levels area already way above the thresholds deemed safe by the WHO, both for human usage and for soil chemistry. The sodium levels may increase with progressive usage of borehole water in the farming projects, up to a point deemed way beyond salvage-meaning the destroyed fertility may not be reclaimed or restored once the damage is done. To mitigate against the potential disastrous and irreversible consequence, the study team undertook a geophysical surveys as well as hydrochemical surveys and data analysis to understand the likely consequence of a prolonged usage of irrigation –based agriculture in the Kumahumato centre. To achieve this, eight algorithms were employed, namely, Neural Networks, Naïve Bayes, Support Vector Machines, Logistic Regression, Decision Trees, K-Nearest Neighbor and Random forests algorithm amongst others. The final three algorithms mentioned here emerged out as the best performers, registering between 95 to hundred percent precisions levels during detailed data analysis. A point picked at random in the Kumahumato area which showed promise of good groundwater potential was analysed and found to be at suitable aquifer sodium levels, which will not be a threat to small scale agriculture envisaged in the program. Machine Learning was thus employed and proved a useful decision making tool in the Project Planning and Design Phase for the proposed food security meant to be a practical resilience response to climate change hazards.**

Keywords—Decision Tree, Distal Merti Aquifer, Python, R, Precision, Accuracy.

I. INTRODUCTION

The Project targets a population of at least 1500 persons and 12000 animal shots, alongside 1000 or so camels heads, on the average, respectively for domestic and livestock use. The project area, **Kumahumato**, is located some 130.0 kilometers away from the Garissa regional/provincial headquarters, in the Republic of Kenya.

The present report presents the findings of a borehole site survey carried out within the **Kumahumato** settlement and its environs, located in the Dadaab district of NE Province. The local population has been steadily growing over the years, hence the need to site and drill a borehole for the purpose of getting sweeter, more potable water of superior quality.

Water tankering has been going on to get the locals water, particularly those settled at Rama Village. there is no sufficient water and **local-leadership** had to arrange to ferry water to the villagers to meet their basic demands. This has proved to be inadequate. The water quality is not that acceptable, going by the TDS levels reported, giving it a familiar hard taste.

Corresponding Author: Dr. Meshack Owira Amimo

The project area is covered by the 1:100,000 Survey of Kenya topographic map sheet of No. SK-81, whose extract is appended at the end of the report. The Hydrogeologist, NWSB, was commissioned by **LOCAL-LEADERSHIP** to carry out **the tasks in May 2022.** The general Project area and its environs consist of arid and semi-arid land characterized by thick to sparse vegetation and no surface water. The parcel of land is a government/community land which has been set aside for purposes of water development infrastructure and range management/dryland research.

Community's Water Demand: The water from the proposed borehole is mainly for domestic and minor irrigation use by the local community in order to ease the population pressure-aided strains on the existing water resources, mainly the borehole in **Kumahumato**. The water demand from the proposed borehole is estimated to be about 40-m^3 per day for domestic and minor livestock use.

Climate: The drought ravaged area has the arid climatic parameters. **The Kumahumato-area** shares the same climatic parameters as the North Eastern area of Kenya. It experiences an arid to semi-arid type of climate. In the Agro Climatic Zone Map of Kenya of 1980, the area falls within the very arid zone. Dominant features are the relatively high temperatures and the modest seasonal variation in the temperature regime, low humidity, low annual rainfall, and high potential evaporation. Evaporation far exceeds rainfall in this area. The area, as in other parts of the country, has two rainy seasons controlled by the movement of Inter-tropical Convergent Zone (ITCZ) which crosses the equator in March - May and again October-December. The two rainy seasons are called "the long rains and "the short rains" respectively. The other months have very low precipitation with the period June - September being considered as dry. Temperatures range between 24^{o}C to 42^{o}C.

Geology, hydrogeology and Groundwater Potential of the project area: The study site is underlain by Mesozoic, Tertiary and Quaternary sediments. **These are locally composed of sandstones grits, conglomerates, limestones, calcrete and superficial deposits of both Pleistocene and Holocene periods. The superficial deposits comprise alluvium** which contains sand and silt, found along the Lagha courses, as well as colluviums composed mainly of crudely stratified mixtures of clay, silt, rock fragments, sandstones, grits and conglomerates. Aquifers have been encountered in these rudaceous and arenaceous sediments.

Based on geophysical/hydrogeological results, the study area is situated in a zone with medium to high groundwater potential and sits **on the Distal Merti Aquifer** comprising **sandstones, grit, clays and allied calciferous sediments**. Thus, chances of striking productive and reliable aquifers are good.

Geophysical Method Employed: Geophysical resistivity measurements were used to determine the potential for groundwater development in the study area. Underground anomalies were probed by means of Vertical Electrical Soundings (VES). The geophysical measurements indicated presence of heterogeneous sands, gravels and grits of varying textures and structures capable of harbouring potential aquifers. The water bearing zones are expected at various levels **between 135, 225 and 285-m bgl.**

Boreholes in the vicinity of the site: The vast expanse of land surveyed is a bare sandy-terrain, yet to be developed, with neither shallow wells nor boreholes. In the vicinity of the area up to a radius of about 40km, there are no wells existing at all. The locals have to rely on the sporadic trips made by a water bowser per month on daily subsistence water volumes.

Groundwater Quality: From the geology, we may predict a limestone aquifer dominating both upper sediments while sands and clayey sediments dominate the deep-seated zones in the area. The bicarbonate component in the waters will give it a characteristic hard taste. The water quality of the proposed borehole is expected to be chemically satisfactory, meaning within tolerable salinity range. However, owing to the proximity to the buried Quaternary river systems in the area, the lateral dilution effect to the boreholes' waters shall be such that the water may not be saline at all. The bacteriological content of the water from the aquifer is expected to be negligible. However, proper construction, by sealing off the borehole from surface contaminants should be observed. Proper bacteriological and chemical analysis should be carried out after completion. Further, bacteriological analysis is advised at least bi-annually.

Project Ownership*:* The proposed site is a public facility owned by the local Kumahumato Small Scale Farmers Association.

II. Literature Review

According to a study in India (Kumar et al, 2016), a study was conducted on groundwater hydrochemistry on quality aspects of water in Kerala. In that study, the sampled sites had the water samples subjected to laboratory analysis, for various

physicochemical parameters as well as analysis of the major ion hydrochemistry. The study went ahead to generate both the Gibbs and Wilcox plot models. Whereas the Gibbs diagram suggested strongly that the sampled aquifers water quality are controlled to a large extent by the country rock mineralogy, which is the dominating factor in determining aquifer hydrochemistry in the study area. Further into the study, the suitability of this water for irrigation purposes was also determined by analyzing the sodium adsorption ratio, also abbreviated as SAR, the residual sodium carbonate, aquifer sodium percent, the Kelly's ratio, residual sodium carbonate, soluble sodium percentage, permeability index, and water quality index. All these parameters tallied with the known set standards for drinking, irrigation and livestock drinking water TDS and EC thresholds deemed acceptable. Moreover, the Wilcox plots suggested that the water quality was well within the limits deemed acceptable for drinking purposes as well as irrigation usages, with respect to the WHO standards of water quality, and this included the Na+ ionic levels in the aquifers.

Overall, the study concluded that the water from the aquifers analyzed bear water quality generally held as acceptable- safe for drinking and irrigation purposes, save for a few aquifers, whose water samples were found to be way exceeding the limits, on the account of due to human and industrial-related activities. Those samples were recommended or use in irrigation, as the sodium levels therein were observed not to be a threat in any way to soil fertility.

Islam et al (2017) studied Groundwater hydrochemistry in Bangladesh. The study involved sampling and analyzing 20 samples from different tube wells whose depths ranged between 21 and 54m below ground level. The water quality assessment was undertaken by studying the physicochemical parameters, namely, aquifer temperature, pH, EC, TDS and major anions and cations in the aquifers as inferred from the analytical results. These were mainly the sodic ions (Na^+ ions), K^+ ions, Ca^{2+} ions, Mg^{2+} ions, Cl^- ions, SO_4^{2-} ions, NO_3^- ions, and , finally, the HCO_3^- ions. The study found out that the aquifers are slightly alkaline and brackish. Further graphical analysis were carried out on the major ions revealed in the study to delineate the relationship between the chemistry and aquifer parent rock geology or the history of the origin of the water in these aquifers. The study entailed investigations on the geology, geochemistry, hydrology and soil science as they relate to water quality in the Bangladesh area of study. The study focused on the groundwater hydrochemistry and water quality within most of the wells which were basically located in shallow aquifer systems. The present study aspires to assess the quality of the groundwater, to determine its utility and find out the major geochemical processes active in the study area. The study also delineated the spatial distribution of hydrochemical parameters to aid proper understanding of the aquifers and thereby generate info that shall be in use , in the future for matters relating to water resources management. The geochemical plots were then generated. These plots tended to suggest that aquifers are sodic. The trends of cations and anions were found to suggest this order of anionic-cationic prevalence: $Na^+ > Ca^{2+} > Mg^{2+} > K^+$ and of this other order, chloride type: $Cl^- > HCO_3^- > SO_4^{2-} > NO_3^-$, respectively and Na–Cl–HCO_3 is the dominant groundwater type. This implied that the major aquifer was a sodic ground water species.

Summarily, the study how vital a role sodium ions may play in shaping up the groundwater aquifer water quality, and this is a function of the rock types along which the subsurface flow systems traverse before reaching the aquifer. Another study in Bangladesh by Shammi et al (2019) notes that although rainwater consumption has the positive impact of low or no sodium intake at all, it impacts negatively by virtue of the fact that it does not possess the other essential vital minerals needed by the body, that is primary micro elements. In the course of the study, it was noted that there was a progressive increase in sodium ion levels, in the course of the study duration, mainly the span of the dry season. It was therefore, vital that there be increased awareness on water quality issues, and also adopting correct technological interventions, as well as training communities on the workings of the proposed measures proposed for management of these aquifers.

Aquifer water chlorinity, salinity, permeability indices, sodium adsorption ratio (SAR), residual bicarbonate and magnesium hardness area factors that impact on the quality of water in an aquifer and need to be studied in details before putting in place aquifer management strategies.

Anim–Gyampo et al (2019) studied groundwater chemistry in Ghana, evaluating all these parameters. Heavy metal contents of the aquifers were assessed and analyzed, so that factors like the hazard quotient, hazard index and cancer risk of analyzed heavy metals were estimated, in order to quantitatively to assess their possible risks to human health when ingested in the drinking waters. As noted earlier, the levels of sodium ion the groundwater area an issue in the distal Merti zones, on matters to do with irrigation. The soil sodic levels are already on such a threshold that any use of the sodium-enriched groundwater abstracted there will further deteriorate the soils and destroy its for life.

The results showed that some samples had concentrations of As, Zn and Pb exceeding respective WHO recommended limits considered safe, set by the WHO at 0.001 milligrams per liter, 0.006 milligrams per liter and 0.01 milligrams per liter. Moreover, the Fluoride concentrations in some samples exceeded the maximum WHO limit of 1.5 milligrams per liter, whereas e samples were found to be below the minimum limit of 0.5 milligrams per liter. The lower limits indicate that if the fluoride levels are too low, health effects shall come up, and, conversely, if levels are too high, negative impact such as dental sclerosis and skeletal fluorosis may occur. The water quality indices generated in the study revealed that as much eighty percent of the waters from these aquifers were generally of potable quality. A study was undertaken on the effects of the water quality on both children's health and on adult human health. When compared to adults, the children were found to than twice as vulnerable, to potential health damages relating to heavy metal ingestion, and that the effects may last a lifetime. The effects mainly affected the skin and stomach tissues. Generally, the groundwater was found to be of acceptable WHO thresholds recommended for irrigation purposes, especially for moderate salinity-tolerant crops such as maize, millet, sorghum, pepper, tomatoes, cabbage.

The water may be used for sanitation purposes like cleaning and also for construction industry. Lead levels and Zinc levels have been on such a low scale as to be not of interest as threats of pollution to the Merti aquifers, but are still found in the distal Merti aquifer zones and fringes

Wodira (2020) studied the sodium and other ions behavior in the groundwater systems in the Mwingi subcounty, Kenya. The Parameters analyzed cations such as potassium, sodium, calcium, magnesium, manganese and iron. The research also analysed the anions such as carbonates, bicarbonates, nitrates, nitrites, sulfates, chlorides, and fluorides.

By evaluating the analysed values of the median ionic contents in the aquifers mapped, it was observed that the concentrations of cations in groundwater increased, progressively, in the order Na+> Mg+>Ca2+>K+>Fe2+> Mn2+ while concentration of anions increases in the order Cl-> HCO3-> SO42-> CO3->NO3->F-. The values of EC, Na, Mg, Ca, Cl, and F were found to be way above the World health Organizations set threshold limits for Consumption and / or utilization. Shallow wells were observed to be the least mineralized groundwaters, while boreholes have highly mineralized groundwaters. This suggested that the deeper aquifers are more mineralized than the shallow aquifers in the study area. Moreover, the cation-exchange was observed to be the dominant rock-water interaction process. This naturally is responsible for the highly mineralized waters in study area. The relationship between Na+ ions and the Cl- ions for the majority of the groundwaters analyzed, suggested similarity in their hydrochemical origins. This is to say that the processes that formed the waters or the paths along which the water flowed prior to reaching the aquifers being studied was more or less the same. The geology and mineralogy of the sediments hosting the aquifers appear to be the same, implying rocks and soils of the same age and origin. The study observed that Mivukoni and Ngomeni areas have inferior quality groundwaters while Mumoni and the southern part of Kyuso bear potable aquifers and should be developed for abstraction. A comprehensive program may be to sink shallow wells of large diameter types in the seasonal streams and have the waters piped to the zones with anomalous salinity levels. A case in point is the Thunguthu River which may have fresh waters and which can be pumped to Ngomeni-Kamusilio areas.

III. HYDROGEOLOGY

A. Tertiary sediments

These consist of sandstones grits and conglomerates and are best exposed on the flanks of the Merti Plateau and Barchuma where they have been preserved from erosion by the capping of olivine Basalts. Other exposures are to the South-East of the Yamicha Plateau and in the area north of the Eldera-Modogashe road. Their exposure is poor on the featureless plains that extend from Merti all the way to Habaswein and Dadaab, since they easily erode on account of their friable nature which leaves them covered by thick soils. The grits are however exposed along rivers, notably the Galana Gof. The grey sandy soil is an indication of the grits presence as the soil is identical to that formed from them.

B. Quaternary sediments

These are composed of lacustrine sediments with limestones, calcrete and superficial deposits belonging to both Pleistocene and Holocene periods. The texture of the lacustrine sediments is coarse to fine grained. The superficial deposits comprise alluvium which contains sand and silt along the river courses, and colluvium composed mainly of crudely stratified mixtures of clay, silt and rock fragments along the slopes of large inselbergs.

C. Geology of the Study Site

The geology of the **Kumahumato** area is predominantly Loams, fine-to- medium grained sands, holo-crystalline glass, red clays, alongside marls at shallow depths, but with some Jurassic clay alongside medium-grained sandstones at great depths. The study area and its immediate environs are underlain by sedimentary formations belonging to Plio-Pleistocene age. These formations increase in thickness towards east and south. Most of the sediments in the area belong to Tertiary age, being underlain at greater depths by Jurassic and Triassic formations. **The oldest sedimentary succession exposed in the area is probably Upper Jurassic Limestone underlying the sandstones and limestone which outcrop locally.**

The Triassic sediments include variable marls and sandy marls, friable sandstones, sometimes pebbly and ranging through grey, pale red or green, or mottled, in colour. Locally they include shales and thick brown clays. The area is entirely covered by young sediments, especially wind-blown sands and depositional alluvial sediments. The local geology has been delineated based on the few visible outcrops and geological logs of boreholes that have been drilled in the area.

As observed during the fieldwork, the site is underlain by a thick layer of sandy soils. These are directly underlain by a thin layer of sandy clays, limestones and assorted sands and sandstones. The thickness of each of these layers varies significantly from site to site. Further, each of these sediments are intercalate with each other up to depths in excess of 3000 metres bgl.

Geo-structural parameters such as faults in the rocks often optimize storage, transmissivity and recharge, particularly when they occur adjacent to, or within, surface drainage systems. Faulting will have the highest impact on hard and massive rock types. Elastic formations such as tuffs and weakly consolidated deposits will bend (fold) rather than break (fault). As a result, their porosity will not increase in the area affected by the fault. Fractures and joints will break massive layers such as gneisses. This gives rise to increased (secondary) porosity, and consequently enhanced recharge, groundwater storage and transmissivity.

Only lineaments are well illustrated through the series of tree-lines along the river courses and allied neighbourhoods and this is a certain proof that the areas bear some fractures that are transmitting juvenile and meteoric waters in the subsurface.

Map 1: Map Showing The Location Of Kumahumato Study Area In The Dadaab Subcounty

D. Hydrology, Hydrochemistry and Structural Geology

1. *Recharge Mechanisms:* within the peripheral Merti Beds: Evidences abound of jointing and fracturing of the carbonate sediments on the surface, alluding to intense forces of fracturing, carbonation and quaternary tectonic faulting. Much of the south westerly – north easterly directed stress fields helped sculpture the terrain into its present geological state. Owing to the relatively high fractions of clays in the beds, there is no sufficient time available for maximum catchment input infiltrations into the sub surface zones lying on the adjacent aquifer units in the proposed well sites. This explains the anomalous salinity levels of the boreholes done to great depths in the area. The area has good quality water, in TDS and EC terms.

2. *Temperature and Humidity:* The scanty data available for the Wajir/Garissa Weather stations give an insight into the rainfall pattern here. Many are the times when the two bordering communities exchange visits for pastures and water whenever extremities are encountered in terms of rain failure.

Temperature variations across the year follow a bimodal distribution similar to that of rainfall, except that mean monthly temperatures are lower when rainfalls are higher. The time series length (3 – 4yrs) for which this data was collected is short; however the trend these data display is considered to be valid. Temperatures are fairly constant over the year, with an average of 28.5°C, a mean monthly maximum of approximately 30°C, and a monthly minimum close to 26°C. The January – May period is relatively warm, with mean monthly temperatures of 28.6 to 30.4°C. From June to September, monthly average temperatures commonly range between 26.8 and 27.8°C, with August being the coolest month.

3. Regional Rainfall: The region, located at an average altitude of 160 m asl, is semi-arid and characterised by low rainfall distributed bimodally across the year. Precipitation in this area is markedly erratic. Much rain falls as intense local convectional storms which may yield 50 to 60 mm in a single event. The area experiences two wet seasons, march to May and from October to December. June to September forms the driest period although the named wet seasons are fairly dry.

IV. GEOPHYSICS:

In order to determine the **Projects Area's hydrostragraphy and aquifer suitability, a total of 3No. Vertical** Electrical Soundings were undertaken using the modern **ABEM SAS 4000B Terrameter**. Schlumberger arrays were used so that current electrode spreads of up **to 320m** against potential spreads of between 5m and 25m were employed to conduct the surveys. Copper electrodes were used for the potentials, while steel iron electrodes were used for the currents.

A. Actual Field Findings

The geoelectrical data generated for the whole study was some 1No Vertical Electrical Soundings. The best data analysed is described hereunder:

Table 1: The Geonalysed Data

Resistivity Sounding Curve No	**Formation Depth Interval**	**Resistivity OhmM**	**Expected Geological Formation**	**Remarks**
R-001/2022	0-1	30	Top Soils	Barren sediments
Located some fifteen kilometers away from the **Darusalaam town**. Shown to area chief.	1-20	6.5	Loose Subsoils	Barren Sediments
	20-80	14	Clays/sandstones	Aquifer material
Recommended to a maximum depth of 270m	80-160	13	Calciferous deposit	Aquifers/sandstones
	160-250	12.5	Clays/sandstones	Aquifer Material
	Over 250	infinity	Clays/Sandstones	Aquiclude

B. Procedure Used in Data Analysis

The levels of Merti aquifer Sodium was analysed using the seven algorithms in such a way that the starting point was acquisition of data of boreholes already drilled over the years from the Ministry of Water and Irrigation, the Water Resources Management Authority (presently known as WRA), the United Nations High Commission for Refugees (UNHCR), Islamic Relief (K), Kenya Red cross society and the NWWDA office hydrological databank. The main data analysis tool was the kNN but all the other algorithms were also used to find a comparison between them and the performance of kNN in terms of precision and accuracy.

C. K-Nearest Neighbors

The kNN is termed as a non-parametric and lazy learning algorithm. It is essential that meanings be put to the two terms. The term 'non-parametric' means there is no assumption, to be taken during analysis and predictions of class, for the distribution of the underlying data. In other words, the model structure is to be determined from the dataset used by the modeler. Datasets may take non-theoretical distributions such as Gaussian. However, in real life, this may never be the case. The kNN algorithm may thus prove useful in practical applications, where most of the real-world datasets do not follow mathematical theoretical assumptions.

The field of groundwater hydrology provides one such a case, as the datasets that may be used may be too complex to be easily inferred as to its exact distribution type. Statistical analysis methodology of data presupposes, theoretically speaking, that it be assigned a distribution class, which in most cases, is Gaussian or Normal. The linearity or nonlinearity of relationships between various variables in the dataset may be a challenge to the researcher and getting an algorithm that does not require a prior knowledge of the distribution is of immense help to a researcher whose requirement is purely decision making on a subject class under a situation of uncertainty. This may be TDS of variable levels in a yet-to-be-developed proposed aquifer site.

There are also what are known as lazy algorithms in data science. The term Lazy algorithm implies that the algorithm doesn't not need to train the data points to generate a prediction model for decision making.

It means that all the training data shall be utilized during the testing phase, which is the phase at which a prediction model is generated for the entire dataset. There is a catch to this paradigm. It implies that training phase shall be faster whilst the testing phase shall be much slower. This testing shall also be more costly. The term costly as used here means that the testing phase shall consume more time and memory. The modeling phase will involve more time and memory when using kNN and the algorithm is therefore not suited to a large dataset for this reason. The large dataset implies many hundreds or thousands of rows, which shall deplete memory reserved for storing the training data portion. The model generation will thus end up taking quite a long time.

1. *The working of the kNN Algorithm:* In kNN, the variable k the number of nearest neighbors. This is crucial since if we have 30 data rows, the value of k will be the whole number gained by the square-root of k.

n=30

$k=(30)^{0.5}$

k=5.47

The odd-numbered value applicable for kNN modeling will thus be 5.0. This means that in a prediction of class or category with a dataset for a new row, if there are three of class "No" and two of class "yes", then the class of "No" wins. The new row is thus classified as of the class 'No'. This may be explained further hereunder. Suppose several rows give a class of either red or blue in a dataset for class.

The value assigned to this k number of neighbors is the most critical value during modeling. This is because if, during modeling there are even numbers of classes in the dataset, such that the class of red is 2 and class of, as blue is also 2, the tie will make it difficult to perform the prediction of the new variable whose row of data is being predicted using, say, the Euclidean distances.

In other words, some sort of voting takes place in the various classes out of the five instances considered (when value of k=5 as postulated earlier, so that the majority class wins.in order to evaluate the closest similar points in a data frame, one searches the radius between these points using distance measures, which includes the following:

a) Euclidean distance,

b) Hamming distance,

c) Manhattan distance and, finally,

d) Minkowski distance.

This research primarily employed the Euclidean distance formula owing to its relative simplicity and the fact that all dataset tried on it performed as desired-the zones that had saline water for example were predicted as so. This also happened to the zones with fresh water.

The kNN algorithm employs a step-by-step approach, which includes the following:

1. Computing the radii or distances
2. Assess the proximity/closeness of the neighbors in terms of the said radii or distances
3. Taking a vote as to the majority class that ends up winning

Assigning class of the new row on the basis of the number 3 above

V. ALGORITHMS

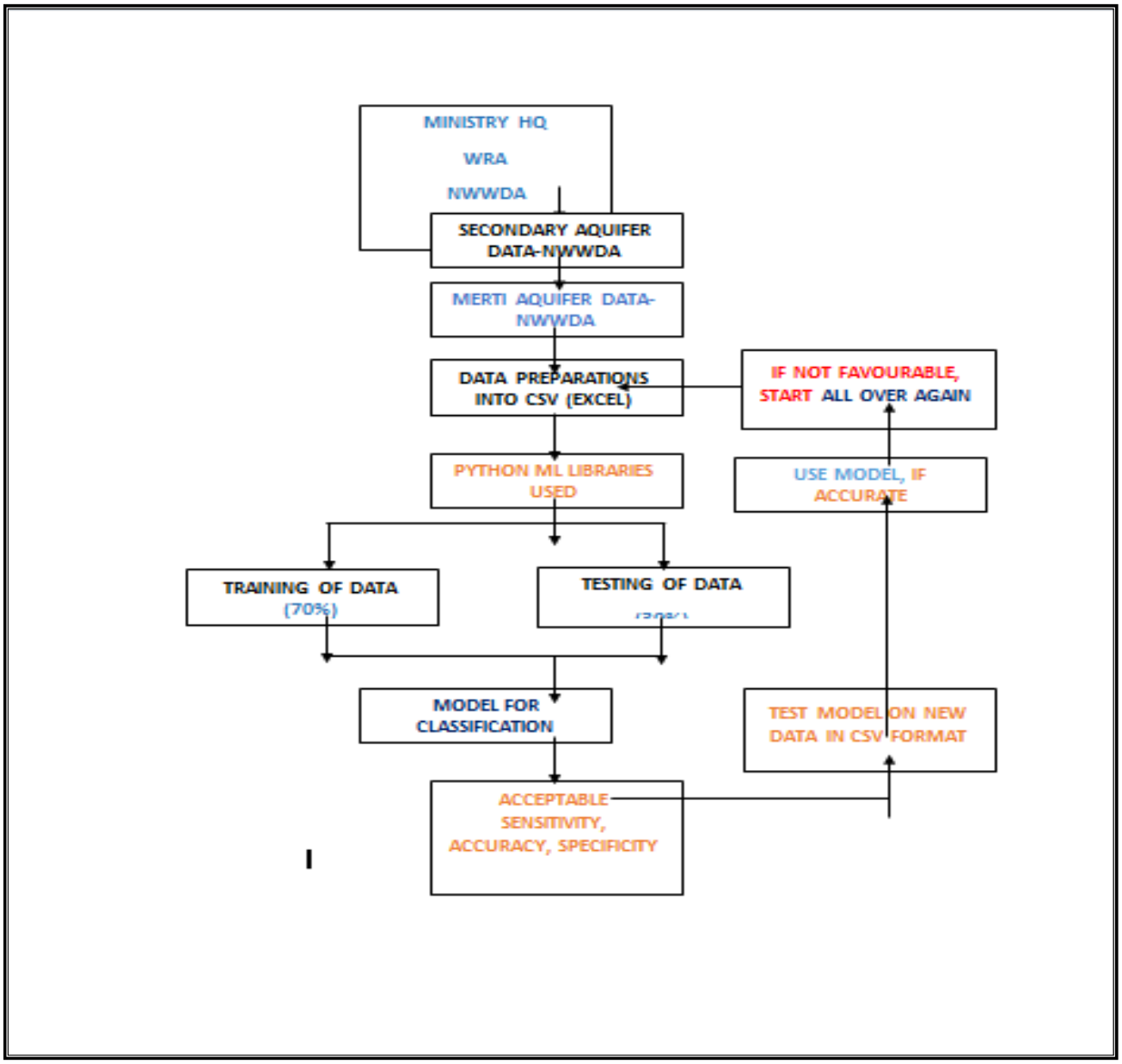


Fig 1: Flow Diagram of Analysis Using the Algorithms

Some insitu analysis were undertaken for some wells to aid data addition to the existing csv files prepared for detailed analysis using the ML algorithmic procedures employed in the study.

The study was essential as most waters drilled previously have been discovered to contain levels deemed toxic to human lives, so that the result has been to abandon them or leave them for use by livestock like camels, which are able to tolerate anomalous levels of sodium in real life, without undergoing any medical complications.

The sodic levels are important as the Dertu area of the Merti aquifer has once been a focus of intensive study to determine the relationship between sodic levels and the bone diseases and incidents of miscarriages seen in some pregnant women, said to have had a prolonged use of waters from a well in Dertu.

1. The procedure that was undertaken is illustrated overleaf, and involved generating the parent dataset with the predictors of sodium, as shall be highlighted later on in this study. The data was then used to form a predictive model in the python programming language.

2. To achieve this, python libraries were used when the data was split into 70 percent, and 30 percent ratios, to be used as training and testing data portions, respectively. This was necessary to aid understanding of the accuracy of the algorithm used in each instance, and takes the most promising algorithm for predictions of the new field datasets. It is this field dataset that helps determine whether a newly proposed area shall have appropriate/desired levels of sodium, or otherwise.

3. If the algorithm is not having desired accuracy levels, another one is tried on trial and error basis, until the algorithm or set of algorithms deemed appropriate is attained.

4. The flow chart sketched hereunder illustrates how the procedure works and how it was employed in the present study.

The results of the predictions have been summarized later in the data analysis portion of the study.

Several algorithms, eight in total were used in the study, and kNN and Decision Trees found to have hundred percent levels of precision.

A. Modeling Sodium ion levels using the K Nearest neighbor or kNN algorithm

It became necessary to model the Merti aquifer levels of sodium on the account of the suspected fluoride mineralization, as the statistical correlation between the two was found to be significant. The study looks into ways of analyzing the sodium levels using the values of sodium-levels and vice versa and also from the elements inferred by the correlation plot functions in R. The analysis of sodium levels provides useful information on chloride levels, as well, since the two bear a high value of correlation coefficient of over 90 percent. If one predicts the sodium levels, the values or outputs so derived may be used to predict chloride levels in the Merti aquifer.

The algorithm that was used is as follows:

```
109
110
111
112 # kNN Classifier in Python
113 y_pred = modelXX.predict(X_test)
114 y_pred
115 y_pred=pd.DataFrame(y_pred)
116 y_pred
117 from sklearn.metrics import confusion_matrix
118 from sklearn.metrics import accuracy_score
119 cm = confusion_matrix(y_test, y_pred)
120 cm
121 print(cm)
122 from sklearn.metrics import accuracy_score, confusion_matrix,precision_score, recall_score, r
123 accuracy_score(y_test, y_pred)
124 print(f"The accuracy of the model is {round(accuracy_score(y_test,y_pred),3)*100} %")
125
126
127
128
```

Fig 2- Output table using the kNN algorithm for Merti aquifer SODIUM LEVELS

Description

The kNN is an excellent performer with an accuracy of up to 100 percent and is therefore obviously a well-suited to the developing of prediction models for the Merti sodic levels. The generated output is tabulated hereunder, in the python spyder GUI. The output has both the confusion matrix indicating the three classes of sodium levels denoted as one (1), very good and also as zero (0) which represents unsuitability for both drinking and agriculture. The class of **1** is the sodic level ranging between zeros to 250 mg/L levels in the aquifer. The **unsuitable class** of sodium level is denoted by zero, indicating levels of over 250mg / L. category of **unsuitable in the dataframe** implies any levels of sodium exceeding the value of 250mg/L. The model was built

with the WHO thresholds in mind, which is 250mg/L, as the guiding factor, which stipulates that sodium ions concentration levels beyond 250 mg/L is deemed unsuitable for human health, as well as soil fertility and crop health, in the long run.

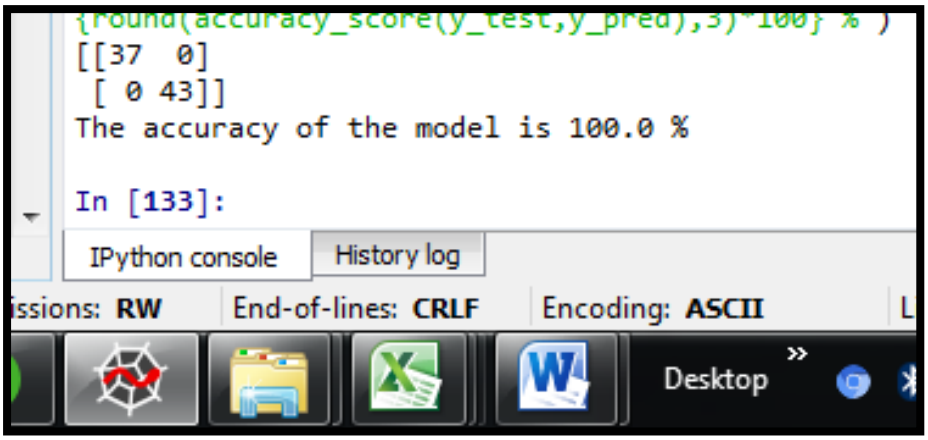


Fig 3: -The confusion matrix tables and accuracy levels with kNN

Summary

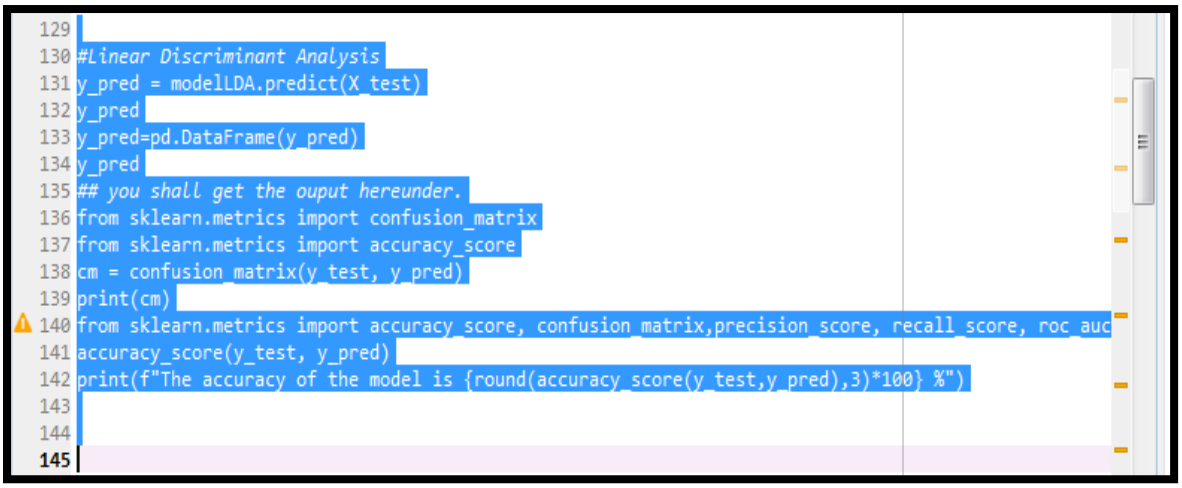
The accuracy of the sodium levels predictor model of kNN is a very good figure of 100% percent and is therefore among the best suited algorithms for mapping cations like sodium in the Merti. This will be useful for food security projects targeting well drilling for the purpose of irrigation in the study area, such as at Shantabak, Dertu and Damajalley. Previous attempts had given rise to over four wells that were later on abandoned due to anomalous levels of sodium way back in 2012-2013 due to intolerable levels of sodic ions that were projected to have detrimental effects on soil fertility in the long run.

There are two categories of sodium in the aquifer and are arranged diagonally, as captured in the model, in the matrix, and there are zero percent instances of misclassifications, mis-prediction using kNN. The kNN classifier algorithm is thus suitable for future work on sodium levels in the Merti aquifer.

B. Modeling SODIUM ion levels using the LDA algorithm

The next algorithm used was the Linear Discriminant Analysis or LDA. The study looks into sodium levels in the aquifer water via analyzing the sodium ion levels, using the values of groundwater aquifer Na+ level. This is also being done from the statistics of the elements inferred by the correlation plot functions in R.

The algorithm that was used is as follows:



```
129
130 #Linear Discriminant Analysis
131 y_pred = modelLDA.predict(X_test)
132 y_pred
133 y_pred=pd.DataFrame(y_pred)
134 y_pred
135 ## you shall get the ouput hereunder.
136 from sklearn.metrics import confusion_matrix
137 from sklearn.metrics import accuracy_score
138 cm = confusion_matrix(y_test, y_pred)
139 print(cm)
140 from sklearn.metrics import accuracy_score, confusion_matrix,precision_score, recall_score, roc_auc
141 accuracy_score(y_test, y_pred)
142 print(f"The accuracy of the model is {round(accuracy_score(y_test,y_pred),3)*100} %")
143
144
145
```

Fig 4 - Output table using the LDA algorithm for Merti aquifer SODIUM LEVELS

Description

The LDA is a found to be a moderate performer, with an accuracy of up to **only 80 percent** and is therefore obviously a fairly-suited to the developing of prediction models for the Merti sodic levels. The generated output is tabulated hereunder, in the python spyder GUI. The output has both the confusion matrix indicating the three classes of sodium levels denoted as one (1), very good and also as zero (0) which represents unsuitability for both drinking and agriculture. The class of one (**1)** is the sodic level ranging between zeros to 250 mg/L levels in the aquifer. The **unsuitable class** of sodium level is denoted by zero, indicating levels of over 250mg / L. The category of **unsuitable in the dataframe** implies any levels of sodium exceeding the value of 250mg/L. The model was built with the WHO thresholds in mind, which is 250mg/L, as the guiding factor, which stipulates that sodium ions concentration levels beyond 250 mg/L is deemed unsuitable for human health, as well as soil fertility and crop health, in the long run.

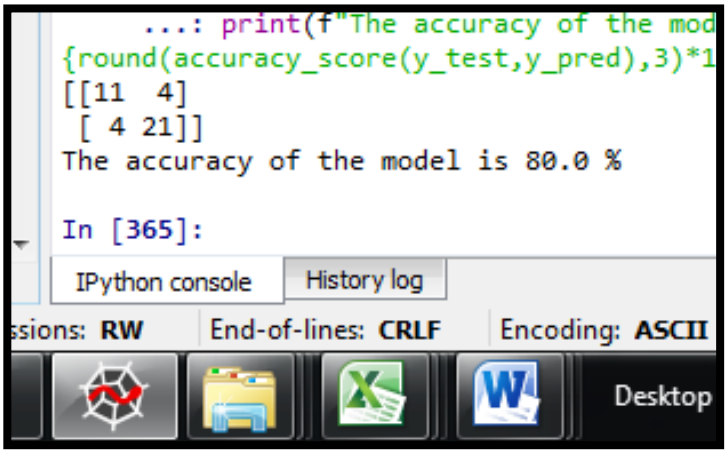


Fig 5: The confusion matrix tables and accuracy levels with LDA

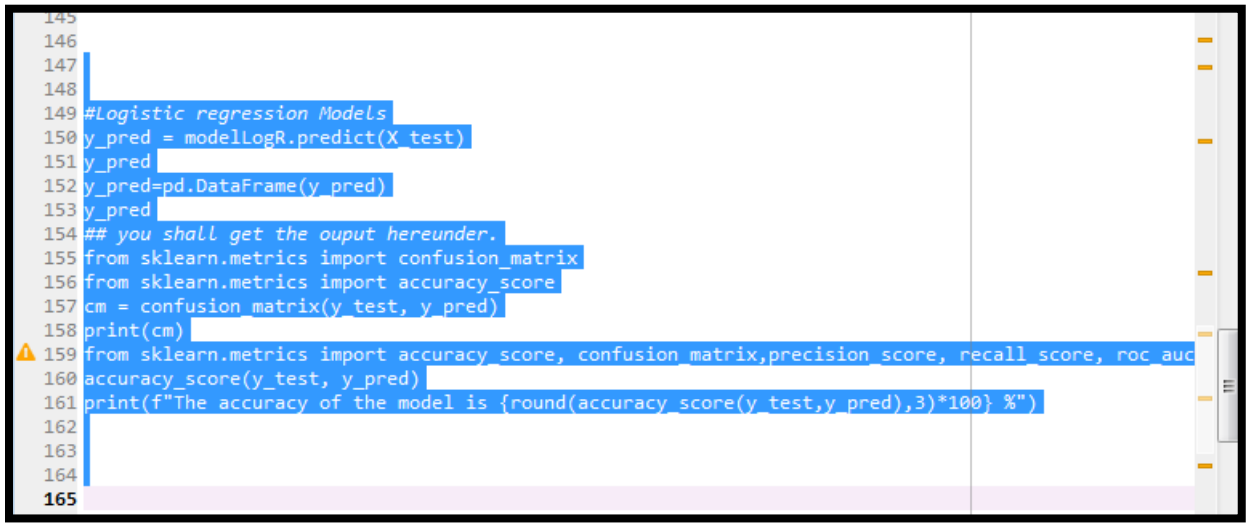
Summary

The accuracy of the sodium levels predictor model of LDA is a fairly promising figure of 80% percent, and is therefore a fairly well-suited algorithm, for mapping cations like sodium in the Merti. This will be useful for food security projects targeting well drilling for the purpose of irrigation in the study area, such as at Shantabak, Dertu and Damajalley. Previous attempts had given rise to over four wells that were later on abandoned due to anomalous levels of sodium way back in **2012-2013** due to intolerable levels of sodic ions that were projected to have detrimental effects on soil fertility in the long run. There are two categories of sodium in the aquifer and are arranged diagonally, as captured above, in the matrix, and there are 20 percent instances of misclassifications, mis-prediction using LDA. The LDA classifier algorithm is thus suitable for NOT future work on sodium in the Merti aquifer.

C. Modeling Sodium ion levels using the LogReg algorithm

The next algorithm used was the Logic Regression Model or LogReg. The study looks into sodium levels in the aquifer water via analyzing the sodium levels, using the values of groundwater aquifer Na+ level. This is also being done from the statistics of the elements inferred by the correlation plot functions in R.

The algorithm that was used is as follows:



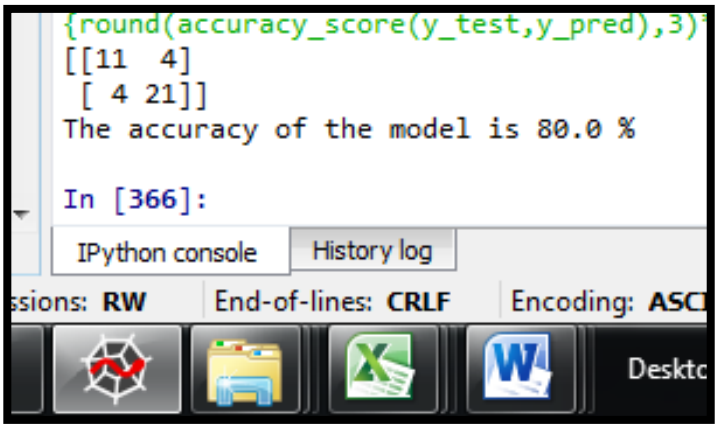
```
145
146
147
148
149 #Logistic regression Models
150 y_pred = modelLogR.predict(X_test)
151 y_pred
152 y_pred=pd.DataFrame(y_pred)
153 y_pred
154 ## you shall get the ouput hereunder.
155 from sklearn.metrics import confusion_matrix
156 from sklearn.metrics import accuracy_score
157 cm = confusion_matrix(y_test, y_pred)
158 print(cm)
159 from sklearn.metrics import accuracy_score, confusion_matrix,precision_score, recall_score, roc_auc
160 accuracy_score(y_test, y_pred)
161 print(f"The accuracy of the model is {round(accuracy_score(y_test,y_pred),3)*100} %")
162
163
164
165
```

Fig 6 - output table using the LOGREG algorithm for Merti aquifer SODIUM LEVELS

Description

The LogReg is found to be a moderate performer, with an accuracy of up to only 80 percent and is therefore obviously the fairly-suited to the developing of prediction models for the Merti sodic levels. The generated output is tabulated hereunder, in the python spyder GUI. The output has both the confusion matrix indicating the three classes of sodium levels denoted as one (1), very good and also as zero (0) which represents unsuitability for both drinking and agriculture. The class of one (**1)** is the sodic level ranging between zeros to 250 mg/L levels in the aquifer. The **unsuitable class** of sodium level is denoted by zero, indicating levels of over 250mg / L. The category of **unsuitable in the dataframe** implies any levels of sodium exceeding the value of 250mg/L. The model was built with the WHO thresholds in mind, which is 250mg/L, as the guiding factor, which stipulates that sodium ions concentration levels beyond 250 mg/L is deemed unsuitable for human health, as well as soil fertility and crop health, in the long run.

One encounters similarity between the LOGREG algorithm and the Logistic regression performances with the sodium data. They are both ill-suited for use as decision making tools for the research on anions and cation levels in the Merti Aquifer.



```
{round(accuracy_score(y_test,y_pred),3)
[[11  4]
 [ 4 21]]
The accuracy of the model is 80.0 %

In [366]:
```

IPython console | History log

ssions: RW | End-of-lines: CRLF | Encoding: ASCI

Desktc

Fig 7: The confusion matrix tables and accuracy levels with LogReg

Summary

The accuracy of the sodium levels predictor model of LogReg is a fairly promising figure of 80% percent, and is therefore a poorly-suited algorithm, for mapping cations like sodium in the Merti. This will be useful for food security projects targeting well drilling for the purpose of irrigation in the study area, such as at Shantabak, Dertu and Damajalley. Previous attempts had given rise to over four wells that were later on abandoned due to anomalous levels of sodium way back in **2012-2013** due to intolerable

levels of sodic ions that were projected to have detrimental effects on soil fertility in the long run. There are two categories of sodium in the aquifer and are arranged diagonally, as captured above, in the matrix, and there are 20 percent instances of misclassifications, mis-prediction using LogReg. The LogReg classifier algorithm is thus NOT suitable for future work on sodium in the Merti aquifer.

D. Modeling SODIUM ion levels using the MLP algorithm

The next algorithm used was the Multi-Layer Perceptron or MLP classifier. The study looks into sodium levels in the aquifer water via analyzing the sodium levels, using the values of groundwater aquifer Na+ level. This is also being done from the statistics of the elements inferred by the correlation plot functions in R.

The algorithm that was used is as follows:

```
168
169
170 y_pred = modelMLP.predict(X_test)
171 y_pred
172 y_pred=pd.DataFrame(y_pred)
173 y_pred
174 ## you shall get the ouput hereunder.
175 from sklearn.metrics import confusion_matrix
176 from sklearn.metrics import accuracy_score
177 cm = confusion_matrix(y_test, y_pred)
178 print(cm)
179 from sklearn.metrics import accuracy_score, confusion_matrix,precision_score, recall_score, roc_auc
180 accuracy_score(y_test, y_pred)
181 print(f"The accuracy of the model is {round(accuracy_score(y_test,y_pred),3)*100} %")
182
183
```

Fig 8: - Output table using the MLP algorithm for Merti aquifer SODIUM LEVELS

Description

The MLP is found to be a moderate performer, with an accuracy of up to only 23 percent and is therefore obviously the ill-suited to the developing of prediction models for the Merti sodic levels. The generated output is tabulated hereunder, in the python spyder GUI. The output has both the confusion matrix indicating the three classes of sodium levels denoted as one (1), very good and also as zero (0) which represents unsuitability for both drinking and agriculture. The class of one (**1**) is the sodic level ranging between zeros to 250 mg/L levels in the aquifer. The **unsuitable class** of sodium level is denoted by zero, indicating levels of over 250mg / L. The category of **unsuitable in the dataframe** implies any levels of sodium exceeding the value of 250mg/L. The model was built with the WHO thresholds in mind, which is 250mg/L, as the guiding factor, which stipulates that sodium ions concentration levels beyond 250 mg/L is deemed unsuitable for human health, as well as soil fertility and crop health, in the long run.

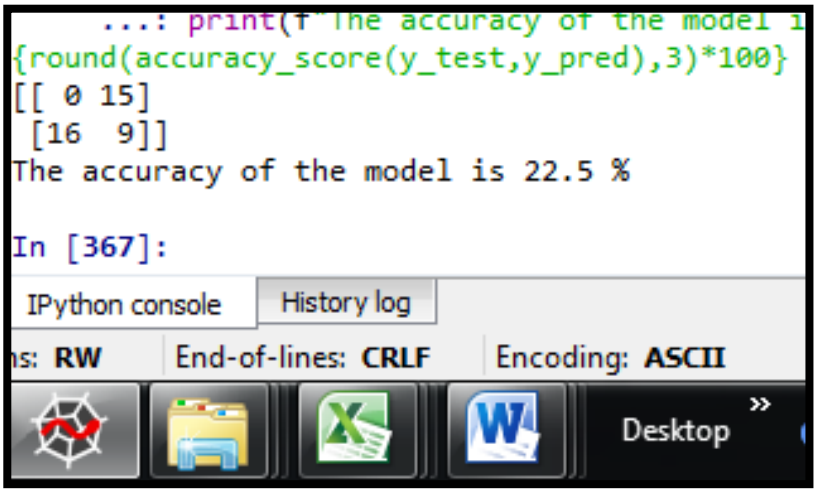


Fig 9: - The confusion matrix tables and accuracy levels with MLP

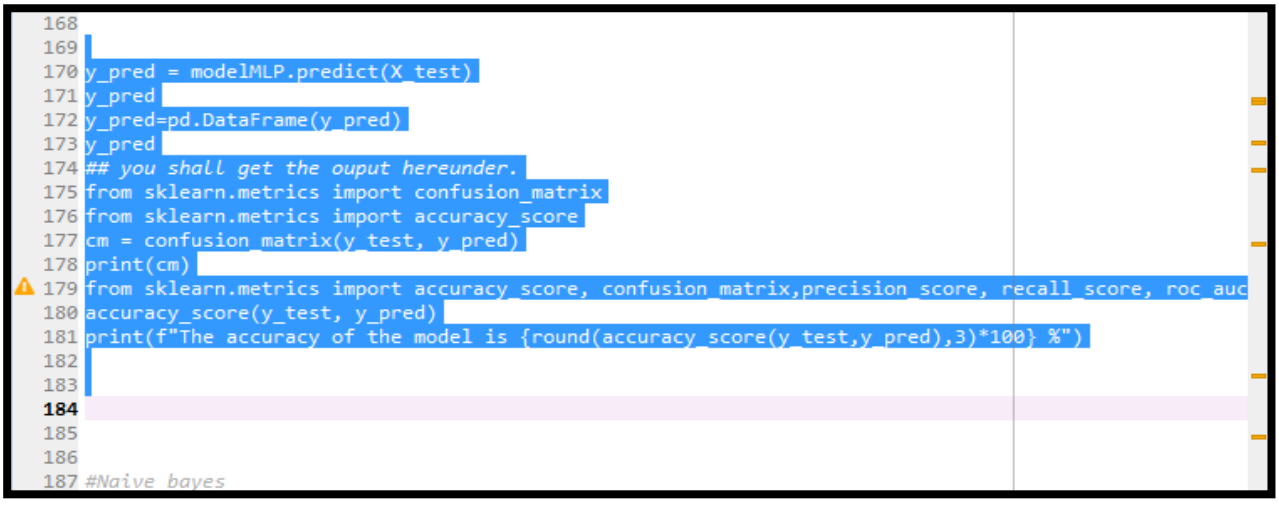
Summary

The accuracy of the sodium levels predictor model of MLP is a poor figure of 23% percent, and is therefore a poorly-suited algorithm, for mapping cations like sodium in the Merti. This will be useful for food security projects targeting well drilling for the purpose of irrigation in the study area, such as at Shantabak, Dertu and Damajalley. Previous attempts had given rise to over four wells that were later on abandoned due to anomalous levels of sodium way back in **2012-2013** due to intolerable levels of sodic ions that were projected to have detrimental effects on soil fertility in the long run. There are two categories of sodium in the aquifer and are arranged diagonally, as captured above, in the matrix, and there are 77 percent instances of misclassifications, mis-prediction using MLP. The MLP classifier algorithm is thus FAIRLY suitable for future work on sodium in the Merti aquifer.

E. Modeling SODIUM ion levels using the NAÏVE BAYES algorithm

The next algorithm used was the NAÏVE BAYES or the NB classifier. The study looks into sodium levels in the aquifer water via analyzing the sodium levels, using the values of groundwater aquifer Na+ level. This is also being done from the statistics of the elements inferred by the correlation plot functions in R.

The algorithm that was used is as follows:



```
168
169
170 y_pred = modelMLP.predict(X_test)
171 y_pred
172 y_pred=pd.DataFrame(y_pred)
173 y_pred
174 ## you shall get the ouput hereunder.
175 from sklearn.metrics import confusion_matrix
176 from sklearn.metrics import accuracy_score
177 cm = confusion_matrix(y_test, y_pred)
178 print(cm)
179 from sklearn.metrics import accuracy_score, confusion_matrix,precision_score, recall_score, roc_auc
180 accuracy_score(y_test, y_pred)
181 print(f"The accuracy of the model is {round(accuracy_score(y_test,y_pred),3)*100} %")
182
183
184
185
186
187 #Naive bayes
```

Fig 10 - output table using the NB algorithm for Merti aquifer SODIUM LEVELS

Description

The NB is found to be a moderate performer, with an accuracy of up to only 80 percent and is therefore obviously a fairly-suited to the developing of prediction models for the Merti sodic levels. The generated output is tabulated hereunder, in the python spyder GUI. The output has both the confusion matrix indicating the three classes of sodium levels denoted as one (1), very good and also as zero (0) which represents unsuitability for both drinking and agriculture. The class of one (**1)** is the sodic level ranging between zeros to 250 mg/L levels in the aquifer. The **unsuitable class** of sodium level is denoted by zero, indicating levels of over 250mg / L. The category of **unsuitable in the dataframe** implies any levels of sodium exceeding the value of 250mg/L. The model was built with the WHO thresholds in mind, which is 250mg/L, as the guiding factor, which stipulates that sodium ions concentration levels beyond 250 mg/L is deemed unsuitable for human health, as well as soil fertility and crop health, in the long run.

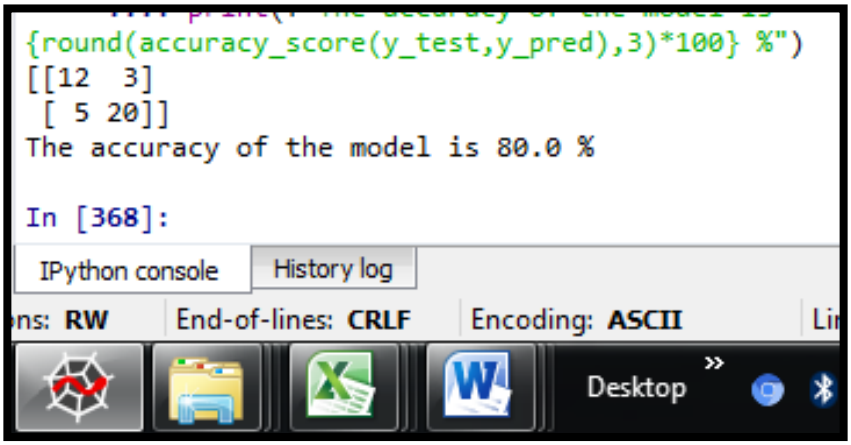


Fig 11-The confusion matrix tables and accuracy levels with NB

Summary

The accuracy of the sodium levels predictor model of NB is a fairly promising figure of 80% percent, and is therefore a well-suited algorithm, for mapping cations like sodium in the Merti. This will be useful for food security projects targeting well drilling for the purpose of irrigation in the study area, such as at Shantabak, Dertu and Damajalley. Previous attempts had given rise to over four wells that were later on abandoned due to anomalous levels of sodium way back in **2012-2013** due to intolerable levels of sodic ions that were projected to have detrimental effects on soil fertility in the long run. There are two categories of sodium in the aquifer and are arranged diagonally, as captured above, in the matrix, and there are 20 percent instances of misclassifications, mis-prediction using NB. The NB classifier algorithm is thus very poor for use, in any future work on sodium analysis in the Merti aquifer.

F. Modeling SODIUM ion levels using the SVM classifier (SVC) algorithm

The next algorithm used was the Support Vector Classifier or the SVC classifier. The study looks into sodium levels in the aquifer water via analyzing the sodium levels, using the values of groundwater aquifer Na+ level. This is also being done from the statistics of the elements inferred by the correlation plot functions in R.

The algorithm that was used is as follows:

```
204
205
206 y_pred = modelSVC.predict(X_test)
207 y_pred
208 y_pred=pd.DataFrame(y_pred)
209 y_pred
210 ## you shall get the ouput hereunder.
211 from sklearn.metrics import confusion_matrix
212 from sklearn.metrics import accuracy_score
213 cm = confusion_matrix(y_test, y_pred)
214 print(cm)
215 from sklearn.metrics import accuracy_score, confusion_matrix,precision_score, recall_score, roc_au
216 accuracy_score(y_test, y_pred)
217 print(f"The accuracy of the model is {round(accuracy_score(y_test,y_pred),3)*100} %")
218
219
220
221
222
```

Fig 12- Output table using the SVC algorithm for Merti aquifer SODIUM LEVELS

Description

The SVC is found to be a moderate performer, with an accuracy of up to only 82.5 percent and is therefore obviously a well-suited to the developing of prediction models for the Merti sodic levels. The generated output is tabulated hereunder, in the python spyder GUI. The output has both the confusion matrix indicating the three classes of sodium levels denoted as one (1), very good and also as zero (0) which represents unsuitability for both drinking and agriculture. The class of one (**1**) is the sodic level ranging between zeros to 250 mg/L levels in the aquifer. The **unsuitable class** of sodium level is denoted by zero, indicating levels of over 250mg / L. The category of **unsuitable in the dataframe** implies any levels of sodium exceeding the value of 250mg/L. The model was built with the WHO thresholds in mind, which is 250mg/L, as the guiding factor, which stipulates that sodium ions concentration levels beyond 250 mg/L is deemed unsuitable for human health, as well as soil fertility and crop health, in the long run.

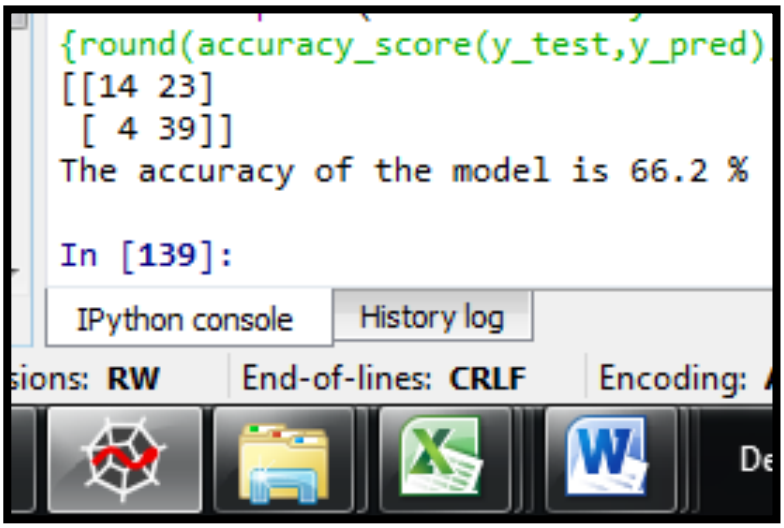


Fig 13-The confusion matrix tables and accuracy levels with SVC

Summary

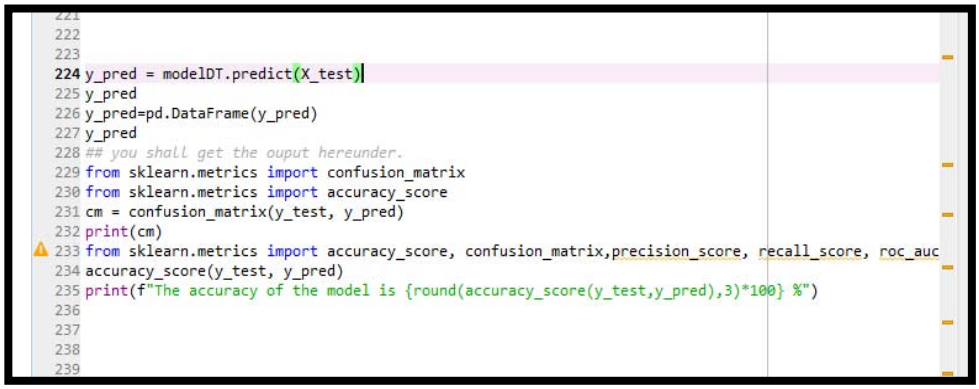
The accuracy of the sodium levels predictor model of SVC is a fairly promising figure of 82.5% percent, and is therefore a well-suited algorithm, for mapping cations like sodium in the Merti. This will be useful for food security projects targeting well drilling for the purpose of irrigation in the study area, such as at Shantabak, Dertu and Damajalley. Previous attempts had given rise to over four wells that were later on abandoned due to anomalous levels of sodium way back in **2012-2013** due to intolerable levels of sodic ions that were projected to have detrimental effects on soil fertility in the long run. There are two categories of sodium in the aquifer and are arranged diagonally, as captured above, in the matrix, and there are 17.5 percent instances of

misclassifications, mis-prediction using SVC. The SVC classifier algorithm is thus very poor for use, in any future work on sodium analysis in the Merti aquifer.

G. Modeling SODIUM ion levels using the Decision tree classifier (DTC) algorithm

The next algorithm used was the Decision Tree Classifier or the DTC classifier. The study looks into sodium levels in the aquifer water via analyzing the sodium levels, using the values of groundwater aquifer Na+ level. This is also being done from the statistics of the elements inferred by the correlation plot functions in R.

The algorithm that was used is as follows:



```
224 y_pred = modelDT.predict(X_test)
225 y_pred
226 y_pred=pd.DataFrame(y_pred)
227 y_pred
228 ## you shall get the ouput hereunder.
229 from sklearn.metrics import confusion_matrix
230 from sklearn.metrics import accuracy_score
231 cm = confusion_matrix(y_test, y_pred)
232 print(cm)
233 from sklearn.metrics import accuracy_score, confusion_matrix,precision_score, recall_score, roc_auc
234 accuracy_score(y_test, y_pred)
235 print(f"The accuracy of the model is {round(accuracy_score(y_test,y_pred),3)*100} %")
```

Fig 14 - Output table using the DTC algorithm for Merti aquifer SODIUM LEVELS

Description

The DTC is found to be a VERY EFFICIENT performer, with an accuracy of up to 100 percent and is therefore obviously the best-suited to the developing of prediction models for the Merti sodic levels. The generated output is tabulated hereunder, in the python spyder GUI. The output has both the confusion matrix indicating the three classes of sodium levels denoted as one (1), very good and also as zero (0) which represents unsuitability for both drinking and agriculture. The class of one (**1**) is the sodic level ranging between zeros to 250 mg/L levels in the aquifer. The **unsuitable class** of sodium level is denoted by zero, indicating levels of over 250mg / L. The category of **unsuitable in the dataframe** implies any levels of sodium exceeding the value of 250mg/L. The model was built with the WHO thresholds in mind, which is 250mg/L, as the guiding factor, which stipulates that sodium ions concentration levels beyond 250 mg/L is deemed unsuitable for human health, as well as soil fertility and crop health, in the long run.

VI. Summary Of Study

Sodium levels in aquifers have been an extensive subject of study by ground water research scientists and environmental engineers over the past decades. This is on the account of the high correlation values existing between sodium and fluorides, which is a harmful cation to both animals and plants if ingested in large quantities. Since fluorides and positively correlated with high values of TDS, it becomes necessary at times to use the known Fluoride levels of a study locality to estimate both the TDS and sodium levels. The values estimated in the present study for sodium level assessments further proves and attests to the usefulness of ML methods in aquifer assessments. Bhageri et al (2017) studied sodium levels, amongst other water quality paratemers using ML methods and mapped aquifers water quality to aid aquifer development planning and monitoring schemes.

Wagh et al (2016) also undertoook a similar task in Maharashtra in India on a local aquifer by employing the Artificial Neural Networks schemes. The present study used ANN but the ANN was not as accurate with the data employed as was the work in Maharashtra.

The present study favored the use of kNN, Random Forest and Decision tree ML techniques over the ANN used by Wagh et al in 2016 study. Jeihouni et al (Iran 2015) undertook a similar study in Iran, and went ahead to combine the artificial intelligence methods with the conventional known GIS kriging techniques, achieving excellent results.

As shall be seen later, the GIS schemes used to map the Merti aquifer sodium levels have been equally a great success. A study in China by Wu et al (2015) attest to the use of ML and GIS schemes to map aquifers for the purpose of generating info vital for decision making situations under uncertainty.

The ANN algorithm registered a paltry 23 percent accuracy levels for the Merti aquifer Sodium data, rendering it unsuitable for the task of mapping the aquifer.

A. Sample Application

The dataset used here projects the GPS coordinates of areas within the Merti aquifer, and these areas are stated relative to their positioning away from the Laghdera flow course, which is the key hydrological parameter determining drainage and recharge into the aquifer, hence the water quality expected. The code of 1 shows suitable (below 250mg/L) , while a code of 0 shows unsuitable, meaning way above 250mg/L levels in the groundwater. The column "wsodium" denotes sodium class.

	A	B	C	D	E	F	G
1	longtd	lattd	elev	dist	wSodium		
2	38.647	1.062	293	204	0		
3	39.6547	1.14458	256	133	0		
4	40.18014	0.34175	172	34	1		
5	38.69197	1.02384	291	172	0		
6	39.31385	0.90669	214	143	1		
7	40.20805	0.28844	135	30	0		
8	39.00675	2.07735	344	256	1		
9	40.32534	0.23798	159	28	0		
10	39.08145	1.96014	321	222	1		
11	39.7467	1.99968	312	224	1		
12	40.5794	0.175	124	45	1		
13	39.74757	0.458	164	74	1		
14	40.51227	0.18899	120	21	0		
15	40.018	1.618	260	180	0		
16	39.15833	1.81137	300	215	0		
17	40.08295	0.67915	160	67	1		

Fig 15: Sample Dataset

In the above table dataframe, sodium levels are predicted in terms of four parameters, three of which represent geology, while the final one (distance from Laghdera flow course dented as 'dist') represents the hydrology.

The details of such an excel table from the site in Kumahumato is shown hereunder:

	longtd	lattd	elev	dist
1				
2	40.079	0.3167	173	18
3				
4				

Fig 16: Kumahumato site data

The following excerpt is part of the R codes used to predict the values of sodium levels anticipated in the new site surveyed in

```
mydataX=read.csv("wsodium2021k.csv", header=T)

head(mydataX)

library(e1071)
library(class)

normalize <- function(x) {
 num <- x - min(x)
 denom <- max(x) - min(x)
 return (num/denom)
}
mydataX_norm <- as.data.frame(lapply(mydataX[1:4], normalize))

summary(mydataX_norm)
```

May 2022.

Fig 17: The new data set was then called into R and predictions made using the model generated in kNN algorithm in R.

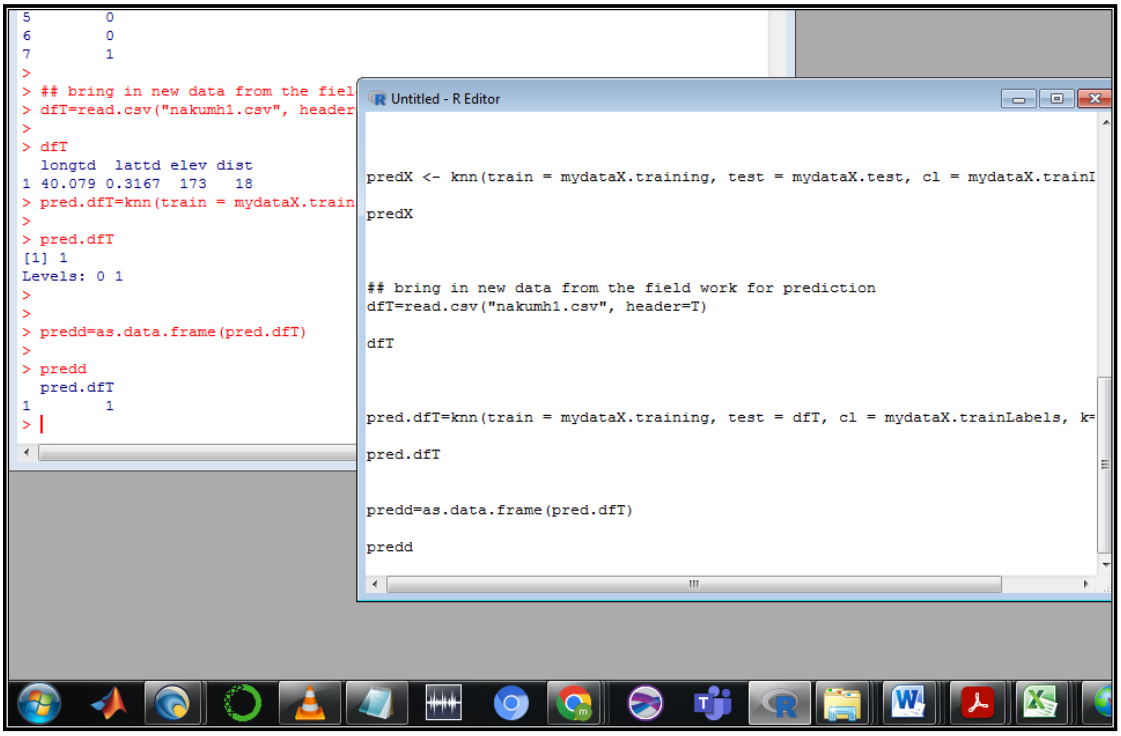


Fig 18: The predicted value of sodic levels lies in class 1, meaning suitable –less than 250mg/L.

VII. Recommendations And Conclusions

From the foregoing synthesis of the Project Areas hydrology, geophysics, hydrogeology and stratigraphy, the sites of the borehole in Kumahumato is found to contain levels of sodium suitable for agriculture and also for human consumption. The community is thus advised to proceed to drill the borehole to the depths recommended in the geophysical surveys, i.e. to the depth of 285m bgl.

VIII. Acknowledgements

I wish to thank friends and colleagues who helped me learn and be able to independently model data using Matlab, Python and R environments. Of special mention is Dr, Okoth Owuor, Ministry of Water, sanitation and Irrigation, presently the CEO of the Water Appeals Board.

References

[1] Anim-Gyampo, M., Anornu, G. K., Appiah-Adjei, E. K., & Agodzo, S. K. (2019). Quality and health risk assessment of shallow groundwater aquifers within the Atankwidi basin of Ghana. *Groundwater for Sustainable Development*, *9*, 100217.

[2] Bazvand, A., Bhagheri, M., & Ehteshami, M. (2017). Application of artificial intelligence for the management of landfill leachate penetration into groundwater, and assessment of its environmental impacts. *Journal of Cleaner Production, 149, 784-796.*

[3] He, S., & Wu, J. (2015). Hydrogeochemical characteristics, groundwater quality, and health risks from hexavalent chromium and nitrate in groundwater of Huanhe Formation in Wuqi county, northwest China. Exposure and Health, 11(2), 125-137.

[4] Islam, S. D. U., Bhuiyan, M. A. H., Rume, T., & Azam, G. (2017). Hydrogeochemical investigation of groundwater in shallow coastal aquifer of Khulna District, Bangladesh. *Applied Water Science*, *7*(8), 4219-4236.

[5] Jeihouni, M., Delirhasannia, R., Alavipanah, S. K., Shahabi, M., & Samadianfard, S., (2015). Spatial analysis of groundwater electrical conductivity using ordinary kriging and artificial intelligence methods (case study: Tabriz plain, Iran). *Geofizika,* 32(2), 192-208.

[6] Kumar, V. S., Amarender, B., Dhakate, R., Sankaran, S., & Kumar, K. R. (2016). Assessment of groundwater quality for drinking and irrigation use in shallow hard rock aquifer of Pudunagaram, Palakkad District Kerala. *Applied Water Science*, *6*(2), 149-167.

[7] Shammi, M., Rahman, M., Bondad, S. E., & Bodrud-Doza, M. (2019, March). Impacts of salinity intrusion in community health: a review of experiences on drinking water sodium from coastal areas of Bangladesh. In *Healthcare* (Vol. 7, No. 1, p. 50). Multidisciplinary Digital Publishing Institute.

[8] Kadam, A. K., Wagh, V. M., Muley, A. A., Umrikar, B. N., & Sankhua, R. N. (2019). Prediction of water quality index using artificial neural network and multiple linear regression modelling approach in Shivganga River basin, India. *Modeling Earth Systems and Environment*, *5*(3), 951-962.

[9] Rafique, T., Naseem, S., Ozsvath, D., Hussain, R., Bhanger, M. I., & Usmani, T. H. (2015). Geochemical controls of high fluoride groundwater in Umarkot sub-district, Thar Desert, Pakistan. *Science of the Total Environment*, *530*, 271-278.

[10] Wagh, V. M., Panaskar, D.B., Muley, A. A., Mukate, S. V., Lolage, Y. P., & Aamalawar, M. L. (2016). Prediction of groundwater suitability for irrigation using artificial neural network model: a case study of Nanded tehsil, Maharashtra, India." *Modeling Earth Systems and Environment,* 2(4), 1-10.