The k-Nearest Neighbor Evaluation of Groundwater Quality in the Distal Merti Aquifer, Modogashe Area

Meshack OwiraAmimo[1], Dr K.S.S. Rakesh[2], and, Jibril Shune[3]

[1]*Research Scholar, IIC University of Technology, Cambodia*
[2]*CEO,Gradxs,India*
[3]*Intern Hydrologist, Northern Water Works Development Agency, Garissa Town*

[1]bmoamimo@gmail.com
[2]kssrakesh@gmail.com
[3]jibreelshune@gmail.com

***Abstract*—The trans-boundary Merti aquifer possesses various water quality categories, namely, saline, fresh and hard species of water. Some boreholes in the Modogashe area are found to be of fresh water quality, yet the vast majority of the wells are either saline or outright brackish. The Northern Water Works Development Agency wished to fund the sinking of a well ,and the company CEO wished to make an informed decision on whether or not to proceed with this project as any water of unacceptable quality would means being rejected by the community, way after millions have been sunk into the project. The objective of the present study is to help develop a groundwater exploration decision-making under uncertainty, so that the Board Technical Services Manager (TSM) and the CEO get the technical info backing the decision to proceed with the program. Consequently, much of the data used to develop the models were sourced from the Kenyan rather than the Somalia-side of the Merti Aquifer.**

To achieve this, a list of existing data of groundwater sources of the Merti aquifer were assembled and processed so that we now had longitudes, latitudes, depth, elevation, mean resistivity of the main aquifer, in the first five dataframe columns. The Total dissolved Solids (TDS) or Electrical conductivity (EC) or the respective rows of data were then analysed against the Kenya Bureau of Standards (KEBS) for water quality, so that a sixth column emerged, code-named as water category. This column expressed the status of the water: whether it was saline, fresh or hard. If fresh or hard, it was categorized as good. If saline, it was categorized as bad. To infer the water quality of the new field data points of a proposed drilling spot whose depths have been determined using Vertical Electrical soundings , a geoelectrical mapping tool, new rows of longitudes, latitudes, elevations, depth, and resistivity (abbreviated as rho) were brought in and predicted. This field dataset lacked the final column, as it is the one to be predicted. The predicted findings were hundred percent correct, for the expected water quality of the new spots.The kNN algorithm was used to generate a prediction algorithm with an accuracy of way over 90.9 percent. With this high level of accuracy, the model was deemed fit for use in predictions of new dataset class of water category, whether the new dataset would give rise to good or badwater species.

***Keywords*—Vertical Electrical Soundings, kNN (k-nearest neighbors), Merti aquifer, Euclidean Distance.**

I. INTRODUCTION

The k-Nearest Neighbor algorithm is an easy-to-use machine learning language for researchers and predictors handling noncomplex data on environmental and other scientific nature like floods mapping (Shahabi et al, 2020).

The machine learning algorithm that is fairly easy to understand and apply in solving real life problems happens to be the k-nearest neighbor. Its scientific principles dwells on neighborhoods and analogically or figuratively and as its very name suggest, it typically implies that you are as green as your neighbor, as cleans as your neighbor-incase that neighbor of yours happens to be somebody clean and maybe as rich as your neighbor-if the neighborhood within which you two live is reserved for the rich. It simply clusters you in the same neighborhood as the person you are closest to in the context of some mathematically defined radii- like the Euclidean distance or the Manhattan distances, for example. Researchers have used this KNN algorithm to map nitrate contamination (Motevalli et al, 2019).

Both continuous and categorical variables may be the following diagram gives an idea on what KNN clustering does to data in terms of attributes or nearness to each other -it clusters them together in shades of pink, blue, green purple, et cetera.

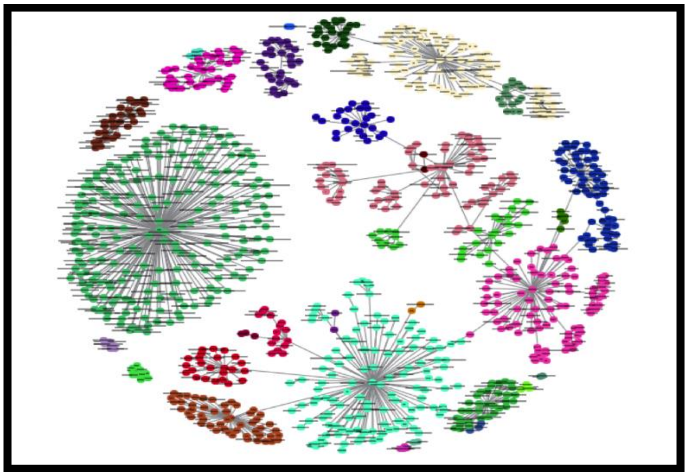


Fig 1: Model of KNN clustering discriminating objects of similar radii together.

In the Modogashe water quality assessment study variables like longitudes, latitudes, elevations, aquifer depths, and mean aquifer resistivity were used to help predict the anticipated quality of water in the groundwater aquifer systems after drilling. This study shall highlight the general principles behind KNN, explore ways used to calculate the distances between points and then apply the same principles using the python software to predict water quality of the study area.The prediction result determines whether or not the drilling shall proceed or otherwise.

II. LITERATURE REVIEW

Raseman et al (2020) undertook a study into the use of kNN algorithms alongside optimization methods to enhance the understanding of influent water quality variability. The study was done around the Poudre River area in Colorado. This was deemed essential to the long-term planning and water resources management templates developed for potable water systems in the area of study. To aid the kNN approach in this regard, time-series methods involving Holt-Winters and Box-Jenkins approaches were also utilized. Using the kNN approach, feature vectors or the class variable being assessed/ predicted were resampled at a given time, thereby generating values at subsequent intervals predefined. Random perturbations enhanced the overall efficacy of performance of the algorithm. The study noted that the k-NN algorithm has been employed on a successful note, to analyses datasets in studies involving stochastic models in hydrology and hydro-climatology. The algorithm was verified to have captured the requisite statistical distribution of the historical records employed in the Colorado study.

Nabahan et al (2018) also undertook a study using kNN to analyses water quality parameters. In the 2018 study, a kNN-based algorithm was proposed for use. The Attribute Weighting Based K-Nearest Neighbor Using Gain Ratio was proposed and used as a study tool: It was a parameter used to evaluate the existing relationships between each variable in the study data-frame. Another attribute, the Gain Ratio was equally utilized in the study as the basis for weighting each attribute of the data-set used in the study. The use of these kNN-based algorithms boosted the levels of accuracy. The accuracy of results so obtained via the n algorithms was then compared to the conventional kNN method, which used using ten-fold cross validation s, using several dataframes. On the basis of the results thus obtained, the newly-proposed kNN-based algorithm managed to raise the classification accuracy levels.

Moderasi et al (2018, Iran) undertook a study on Monthly stream flow forecasting using kNN. The Monthly forecasting of stream flow was of singular importance in water resources management, given that design of dams, it is used to generate the rule curves, used to determine dam viability. In the study, the performance of four data-based models generated using different architecture, including Artificial Neural Network (ANN), Generalized Regression Neural Network (GRNN), Least Square Support Vector Regression (LS-SVR), and k-Nearest Neighbor Regression (KNN) were evaluated in great details, in order to help predict monthly inflow output, to the Karkheh dam. This makes it possible for design engineers to design a volume which is commensurate with the simulated inflow expected.

III. K-NEARSET NEIGHBOR: THEORY BEHIND THE METHOD

A. How to grasp the concept of kNN in a simple manner

The Euclidean distance metrics is the principal idea behind this efficient algorithm for classification. The study begins by illustrating what the kNN is all about, with a sample dataset which is not real, but is used deliberately to make the point clear in terms of what the kNN algorithm does with data points, during the modeling of actual data. The illustration takes aquifer depth to be the x-axis and y-axis to represent the Electrical Conductivity of the aquifer, or the EC in short.

TABLE I
SAMPLE DEMO DATA

SNo	Borehole depth	Borehole EC	Borehole groundwater level
1	12	200	7
2	15	313	8
3	15	345	8
4	20	451	7.5
5	23	655	8
6	25	564	9
7	34	700	8
8	31	725	8
9	45	800	8.5
10	60	991	9
11	72	1000	9.4
12	85	?	10

For a clearer understanding of this, below is the display of a plot of borehole EC versus Depth, and take this just a general example to illustrate what kNN is all about.

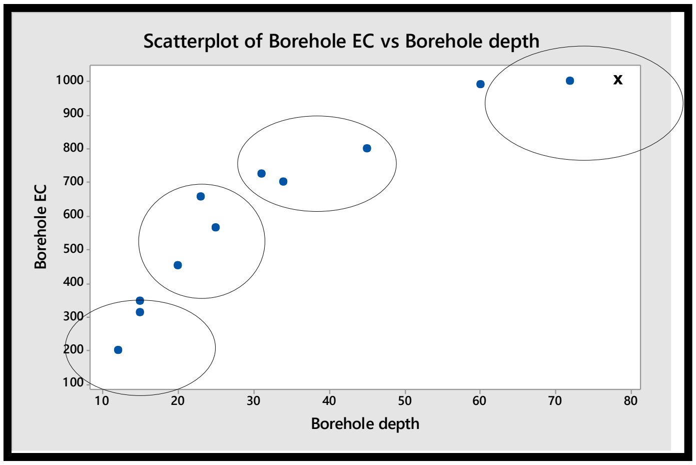


Fig 2: Graph of scatter plots of EC against depths to aquifers

In the above graph, the y-axis represents the EC of water which is a variable determining salinity of the water source, whereas the x –axis rep the aquifer depths. If we were asked to predict the EC of water at depth 85m bgl, the graph shows that it shall be something in the range of 1000 units and above.That is fairly straight forward, as the data is designed to show that every three first data sets belong to the same cluster, and are neighbors, in that respect.

B. *The workings of the kNN algorithm*

As already illustrated above, KNN algorithm can be used for both classification and regression problems. The KNN algorithm uses '**feature similarity**' to predict the values of any new data points. This implies that the new point is assigned a value, based on how closely it resembles the points in the training set. From our example, we know that borehole no 12 with depth of 85m bgl has depths and EC more or less similar to boreholes number 10 and number 11.

If the classes of the boreholes are ranked into 1, 2, 3 and 4 in order to respectively define very fresh water, fresh water, fresh to hard water and hard water, and then the following classification scheme works just fine.

TABLE 2
BOREHOLE EC CLASSIFICATION

SNo	Borehole depth	Borehole EC	EC class
1	12	200	1
2	15	313	1
3	15	345	1
4	20	451	2
5	23	655	2
6	25	564	2
7	34	700	3
8	31	725	3
9	45	800	3
10	60	991	4
11	72	1000	4
12	85	?	4

Below is a stepwise explanation of the algorithm on how it would cluster out and classify the dataset shown here for demo:

First, evaluate the distance between the new point and each training point. See figure 3 below.

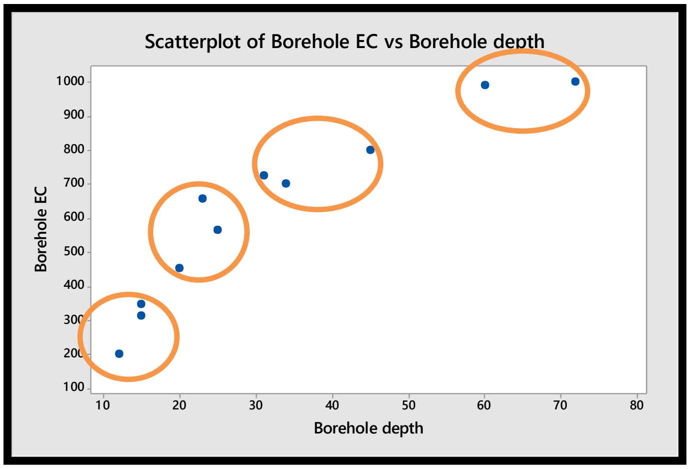


Fig 3: Graph showing the scatterplot of Borehole EC against Borehole depth

The closest 4 data points are selected (based on the distance). In this example, points 1, 2, 3 & 4, 5, 6 &7, 8, 9 and 10, 11, 12 will be selected if the value of k is 4. See the graph shown herewith.

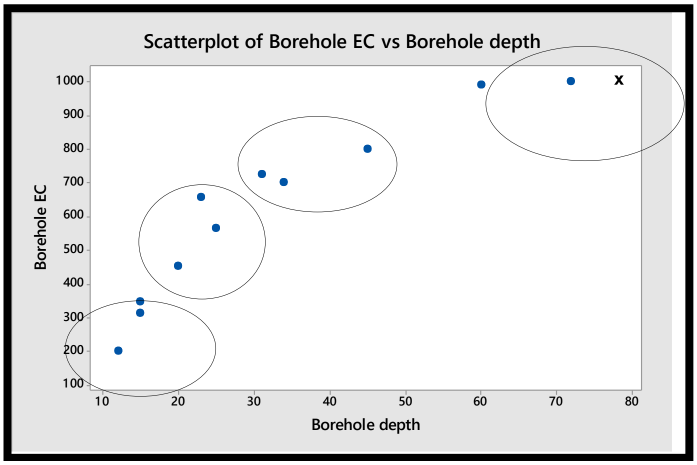


Fig 4: Scatter plot showing the clustered points and the missing value X to be predicted

The average of the last three data points is the approximate final prediction for the new point with a depth of 85m bgl as aquifer depth drilled. Here, we have meant of EC as follows:

= (800+991+1000)/3 = 931 mg/L

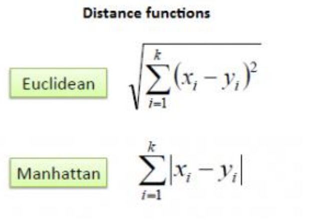
C. *Calculating distances between points plotted*

The first step is to calculate the distance between the new point and each training point. There are various methods for calculating this distance, of which the most commonly known methods are – Euclidian, Manhattan (for continuous) and Hamming distance (for categorical).

1. *1. Euclidean Distance:* The most famous algorithm for KNN classification happens to be the Euclidean distance. This Euclidean distance is calculated as the square root of the sum of the squared differences between a new point (x) and an existing point (y). In our case, the X represents the depths to aquifer

whereas t-values shall represent the aquifer chemistry defining salinity.

2. *Manhattan Distance:* This is the distance between real vectors using the sum of their difference.



$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

$$\sum_{i=1}^{k}|x_i - y_i|$$

.

3. *Hamming Distance*: the algorithm is used in non-linear regression analysis of a dataset. It is mainly used for categorical variables defining class. If the value (x) and the value (y) are the same, the distance D will be equal to 0. Otherwise D=1. This is ideal for classification problems involving attributes like fresh water, hard water and saline water, if we were to use a 1, 2, and 3 coding schemes for the respective water quality.

$$D_H = \sum_{i=1}^{k}|x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

Once the distance of a new observation from the points in our training set has been measured, the next step is to pick the closest points. The number of points to be considered is defined by the value of k. As mentioned earlier, k is the subjective class or different categories in the case of classifications.

D. *Choosing the value of k for computation*

The **second step** is to select the k value. This determines the number of neighbors we look at when we assign a value to any new observation. In our example, for a value k = 4, the closest points are cluster 1,2,3 and 4 as shown in the graph, with the numerical values assigned moving in the ascending order.

E. *Analyzing real groundwater data using kNN algorithm in python for the Modogashe well*

The algorithm works in such a simple scheme and will be used to make predictions using the python codes(highlighted later) for the newly sited borehole site in Modogashe Township.

The python software was used –the anaconda GUI was used and the codes implemented which gave a 90.9 percent accuracy levels, when the machine learning kNN models were used.Then dataset that were used are actually two-the parent data set for the whole Garissa county and also the field data set for the areas that were visited around **Modogashe area** for geoelectrical mapping, and whose generated depths and average aquifer resistivity were then used for data modeling as thus:

i) The parent data was used to generate the model calculator of water quality category in python codes

ii) The parent data was also used to generate the accuracy score percentage and this was around 91 %.

iii) Test data from the field was then brought bin for predictions and indeed the site with average resistivity of 28ohmM was predicted ton bear water with expected water quality.

iv) Recommendations were then made appropriately to the office regarding the fate of drilling 1No well at modogashe, to be laterally recharged via lamina flow from the sand dam being designed.

F. *Summarizing*

It is noted that the KNN algorithm falls in the **supervised learning algorithms**. The implication here that there is a dataset with labels or class training measurements (x, y) and the model would be used to find the link between x and y. the research goal is therefore to discover a function h:X→Y, so that having an unknown observation x, h(x) can positively predict the identical output ,y.

The KNN algorithm is handy in modeling data of environmetrics nature like precipitation and climate change dynamics altogether (Mehdizadeh, 2020).

G. *Working*

First, we will talk about the working of the KNN classification algorithm. In the classification problem, the K-nearest neighbor algorithm essentially said that for a given value of K algorithm will find the K nearest neighbor of unseen data point and then it will assign the class to unseen data point by having the class which has the highest number of data points out of all classes of K neighbors.

For distance metrics, we will use the Euclidean metric.

$$d(x, x') = \sqrt{\left(x_1 - x'_1\right)^2 + \ldots + \left(x_n - x'_n\right)^2}$$

Finally, the input x gets assigned to the class with the largest probability.

$$P(y = j | X = x) = \frac{1}{K}\sum_{i \in \mathcal{A}} I(y^{(i)} = j)$$

While performing regression predictions, the technique will be the same, so that instead of the classes of the neighbors, the model shall:

i) Pick the value of the target class being inferred, and also

ii) To find the target value for the unseen data point by taking an average, mean or any suitable function that may be necessary.

H. The Ideal Value for k for use in kNN algorithmic predictions

Now most probably, one may wonder how to decide the value for variable K and how it will affect your classifier. Well, like most machine learning algorithms, the K in KNN is a hyper-parameter that you, as a data scientist, must decide in place to get the most suitable fit for the data set.

When K is small, we are holding the region of a given prediction and pushing our classifier to be “more blind” to the overall distribution. A small value for K provides the most adjustable fit, which will have low bias but high variance. Graphically, our decision boundary will be more irregular. On the other hand, a higher K averages more voters in each prediction and hence is more flexible to outliers.

Larger values of K will have a smoother decision boundary which means lower variance but increased bias.

IV. RESEARCH METHODOLOGY

The prediction of the water quality entailed combining the field generated data; mainly water struck level depths imputed rom the Vertical electrical soundings using band ABEM SAS Terrameter, and the resistivity of the main aquifer, which correlates well with aquifer EC-electrical conductivity. This parameter determines the aquifer TDS and water quality.

A. Getting the data for statistical analysis

There is a databank in the office for all boreholes in the NWWDA Offices and the WRA offices, of Garissa. This is the county headquarters. The data was used to generate the prediction model which was subsequently used to predict water quality, an aspect of environmental concern for both humans and livestock.

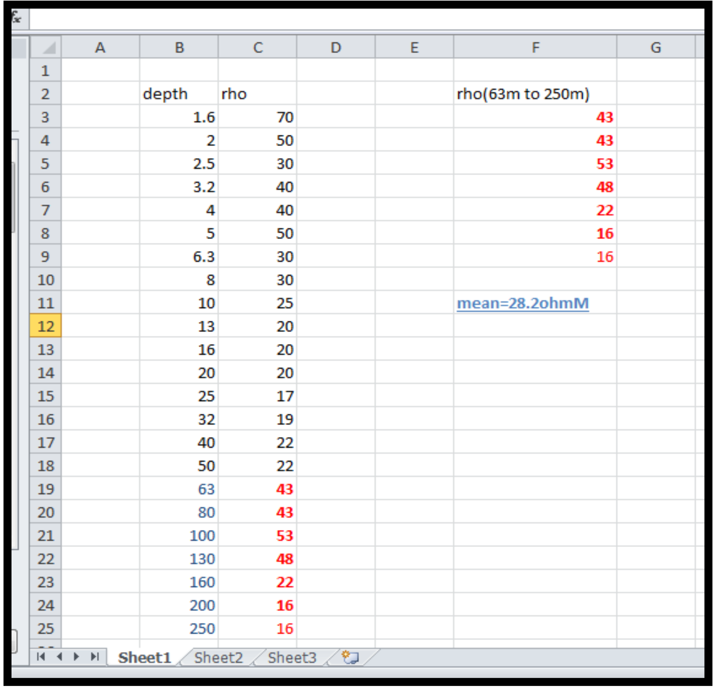
B. Getting the data for testing/predictions of aquifer to be drilled.

Here, the geophysical investigations-aided hydrogeological surveys provided the data to be used as it required the elevations, longitudes, latitudes and depths as well as resistivity values to be modeled for predictions.

From the known hydrology of study area, the depths 63m to 250m determine the flow that recharges the boreholes. It is with this in mind the average of the resistivity values as from 63m to 250m was used in the dataframe for predictions.

The results predicted were indicative of good quality water.

Model 1-The screenshot for the best site with an average aquifer resistivity of 28.2 OhmM



	A	B	C	D	E	F	G
1							
2		depth	rho			rho(63m to 250m)	
3		1.6	70			43	
4		2	50			43	
5		2.5	30			53	
6		3.2	40			48	
7		4	40			22	
8		5	50			16	
9		6.3	30			16	
10		8	30				
11		10	25			mean=28.2ohmM	
12		13	20				
13		16	20				
14		20	20				
15		25	17				
16		32	19				
17		40	22				
18		50	22				
19		63	43				
20		80	43				
21		100	53				
22		130	48				
23		160	22				
24		200	16				
25		250	16				

VES 001/2020 the first site surveyed near the bridge, very promising site in terms of aquifer fractures and water quality anticipated from the values of resistivity.

TABLE 3
DATA ANALYSIS AND GEO-INFERENCES OF SEDIMENTS MINERALOGY

Resistivity Curve No-R	**Schlumberger Probe Depth Interval(m)**	**Resistivity In OhmM**	**Expected Geological sediment/Formation**	**Comments**
001/2020 The first site near the main road, 80m away from the Modogashe bridge	0-1	70	Top Soils	Wetness with water as from 5m to 13m, this is aquifer One. The other aquifers are as from 63M onwards. The water shall be hard, and not saline. The worst case scenario for the aquifer is hardness and not salinity. Results would be better if drilling is abandoned at around 130m bgl
	1-2.5	35	Subsoils	
	2.5-5	50	Wet Clays	
	5-6	30	Calcrete and clays	
	6-8	30	Clays and fine sands	
	8-20	20	Sandstones	
	20-25	17	Clays	
	25-40	22	Fine sands	
	40-50	22	Grits and calcrete	
	50-63	43	Barren Coarse sandstones	
	63-80	43	Sandstones	
	80-100	53	Grits/gravels	
	100-200	16	Clays	
	200-250	16	Fine sandstones	
	Over 250	Infinity	Clays and shalestones	

The VES 001/2020 model plotted showing aquifers in the proposed site

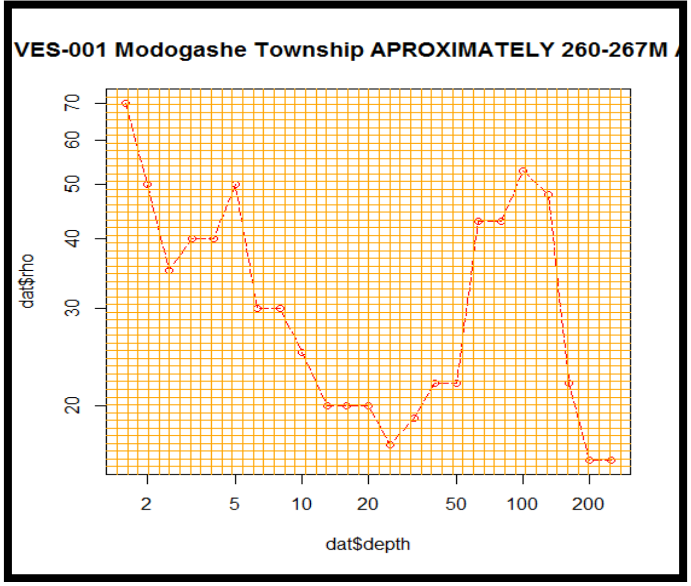
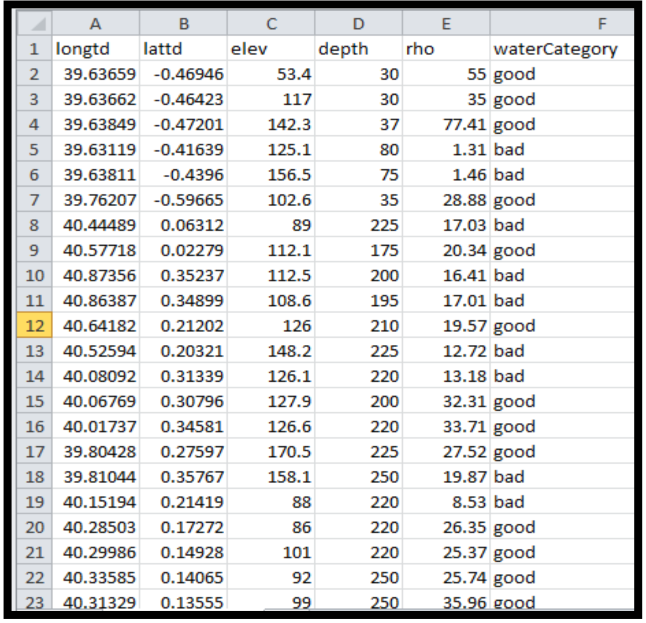


Fig 5: A graph of resistivity against depth showing the aquifers in the proposed site

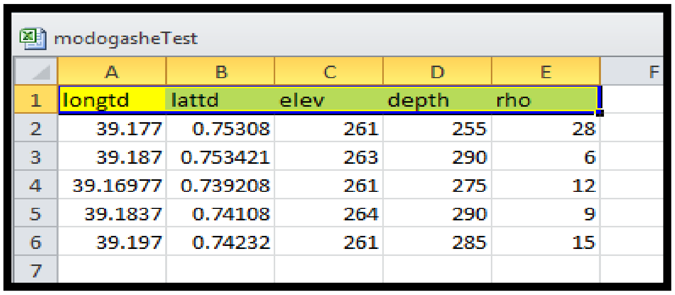
C. *Python Coding*

Having discussed the theory of how the kNN algorithm predicts, the data was analysed using python software. Python coding in kNN analytics has been used in research, with obvious advantages being that python a free language programming software. The present study wrote short script deemed appropriate for the study. The implementation proceeds as scripted hereunder for both the parent data set and the field data from modogashe. (Makkah, et al, 2017). In the year 2018, a survey was done for groundwater abstraction and this method was used to predict the quality of water before drilling (**Amimo, 2018).** The parent data which was analysed to generate predictive model calculator is 'modogasheWQ'. This is the historical data for groundwater for the Garissa County, factoring the predominantly merti aquifer wells.



	A	B	C	D	E	F
1	longtd	lattd	elev	depth	rho	waterCategory
2	39.63659	-0.46946	53.4	30	55	good
3	39.63662	-0.46423	117	30	35	good
4	39.63849	-0.47201	142.3	37	77.41	good
5	39.63119	-0.41639	125.1	80	1.31	bad
6	39.63811	-0.4396	156.5	75	1.46	bad
7	39.76207	-0.59665	102.6	35	28.88	good
8	40.44489	0.06312	89	225	17.03	bad
9	40.57718	0.02279	112.1	175	20.34	good
10	40.87356	0.35237	112.5	200	16.41	bad
11	40.86387	0.34899	108.6	195	17.01	bad
12	40.64182	0.21202	126	210	19.57	good
13	40.52594	0.20321	148.2	225	12.72	bad
14	40.08092	0.31339	126.1	220	13.18	bad
15	40.06769	0.30796	127.9	200	32.31	good
16	40.01737	0.34581	126.6	220	33.71	good
17	39.80428	0.27597	170.5	225	27.52	good
18	39.81044	0.35767	158.1	250	19.87	bad
19	40.15194	0.21419	88	220	8.53	bad
20	40.28503	0.17272	86	220	26.35	good
21	40.29986	0.14928	101	220	25.37	good
22	40.33585	0.14065	92	250	25.74	good
23	40.31329	0.13555	99	250	35.96	good

Fig 6: screenshot of the data that was used to generate the Model Predictor in Python GUI



	A	B	C	D	E	F
1	longtd	lattd	elev	depth	rho	
2	39.177	0.75308	261	255	28	
3	39.187	0.753421	263	290	6	
4	39.16977	0.739208	261	275	12	
5	39.1837	0.74108	264	290	9	
6	39.197	0.74232	261	285	15	
7						

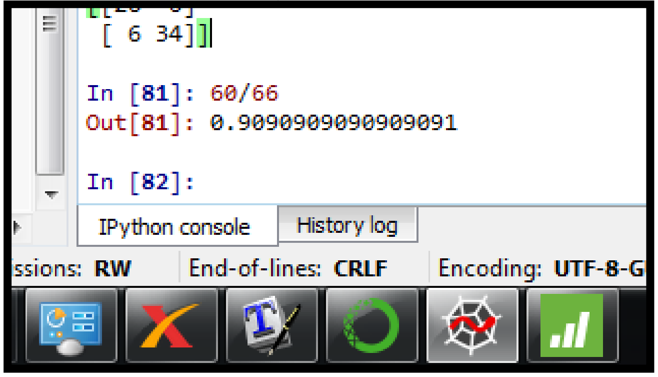
Fig 7: screenshot of the data that was generated in the field and whose water category class is being predicted in python

D. Python Scripts Used In the Model Generation and Predictions

```
## invite data into python

importnumpy as np

fromsklearn.model_selection import train_test_split
fromsklearn.linear_model import LogisticRegression

## call in the metrics that will be used to evaluate accuracy
fromsklearn import metrics
importseaborn as sn
import pandas as pd
fromsklearn.model_selection import train_test_split

## import pandas-to aid calling csv data into python
import pandas as pd
path =
"C:/Users/Amimo/Desktop/python37/modogasheWQ.csv"
data= pd.read_csv(path)
print (data)
data.head()
df=data
df
df.tail()

## from the dependent variable, and the predictors
X = df.drop("waterCategory",axis=1) # Features
y = df.waterCategory # Target variable
importnumpy as np
fromsklearn.model_selection import train_test_split
# split X and y into training and testing sets
#from sklearn.cross_validation import train_test_split
X_train,X_test,y_train,y_test =
train_test_split(X,y,test_size=0.25,random_state=0) #in this
case, you may choose to set the test_size=0. You should get
the same prediction here
fromsklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier (n_neighbors=15)
modelKNN=knn.fit(X, df.waterCategory)
modelKNN
##preview the predictions of the water class using test data
y_pred=modelKNN.predict(X_test)
y_pred
```

```
## preview the power of the model in predicting.it is 90.0
%
fromsklearn.metrics import confusion_matrix
fromsklearn.metrics import accuracy_score
cm = confusion_matrix(y_test, y_pred)
print(cm)

##see the score of performance of the model
score = modelKNN.score(X_test, y_test)
print(score)

###call in ur new field data  into python
path =
"C:/Users/Amimo/Desktop/python37/modogasheTest.csv"
data2= pd.read_csv(path)
print (data2)
data2.head()
df2
df2
## make predictions on this dataset using KNN algorithm
y_pred1=modelKNN.predict(df2)
y_pred1
ypredtable1=pd.DataFrame(y_pred1)
ypredtable1
```

1. *Model accuracy* –this was determined to be around 90.9 % and this suggested that it is a good model worthy of use for predictions of water quality



```
 [ 6 34]]

In [81]: 60/66
Out[81]: 0.9090909090909091

In [82]:
```

Fig 8 Accuracy at 90.9 %screenshot of model

2. *Confusion matrix*–here the model generated the matrix to indicate that 60 out of 66 predictions were correct and this is what gave us the 90.9 per cent accuracy, determine as thus:
=60/66
=90.9 percent
=approximately 91 percent

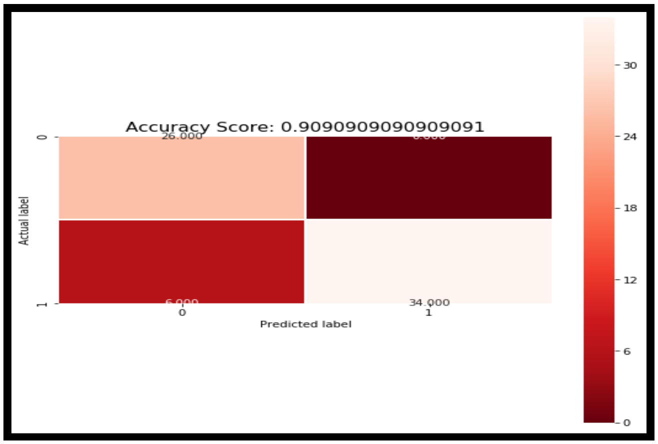
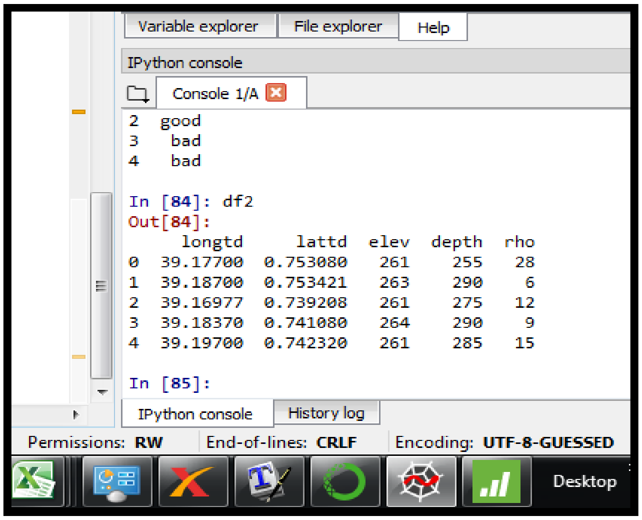


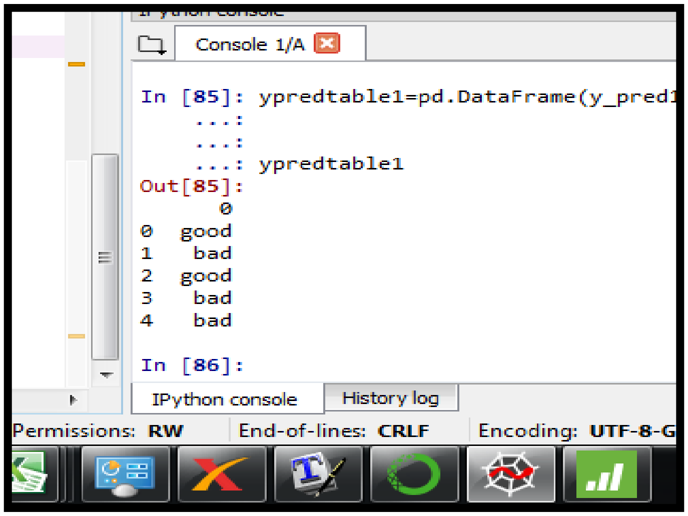
Figure 9: See the confusion matrix in the sea-born python graphics showing hundred percent predictions accuracy by the algorithm



```
2   good
3    bad
4    bad

In [84]: df2
Out[84]:
     longtd     lattd  elev  depth  rho
0  39.17700  0.753080   261    255   28
1  39.18700  0.753421   263    290    6
2  39.16977  0.739208   261    275   12
3  39.18370  0.741080   264    290    9
4  39.19700  0.742320   261    285   15

In [85]:
```

Permissions: **RW** End-of-lines: **CRLF** Encoding: **UTF-8-GUESSED**

Fig 10: The field data on which the model calcualtor genarated by the parent data set was tested. See data in row 1 which has resistivity value at 28.0ohmM.



```
In [85]: ypredtable1=pd.DataFrame(y_pred1
    ...:
    ...:
    ...: ypredtable1
Out[85]:
      0
0  good
1   bad
2  good
3   bad
4   bad

In [86]:
```

Permissions: **RW** End-of-lines: **CRLF** Encoding: **UTF-8-G**

Figure 11: The field data on which the model calcualtor generated by the parent data set was tested. See data in row 1 which has now been predicted as haivng a high probability of striking good quality water.

V. RECOMMENDATIONS& CONCLUSIONS

The Merti aquifer concludes that the water quality is desirabel with an accuracy level of 90.9 % It is thus recommended that the CEO of Northern water works development Agency, earlier known as the Northern Water Services Board, may proceed and drill the borehole to a maximum depth of 255m bgl. The study employed the knowledge of GIS details of proposed study sites and the geoelectrical data analysis of the resistivity generated at variuos depths during groundwater exploration in the area. The depths of boreholes penetrates different types of aquifer starum sets or strata with varying resistivity, which is a factor in the predicted water quality.

The borehole may be utilised for :

a) Sanitation purposes
b) Domestic/drinking purposes

The study recommends that such a study should thus inform future drilling expeditions in study area to ensure a minimal number of saline wells sank in the study area.

REFERENCES

[1] AmimoMeshack, Hydrogeological surveys of Modogashe-Skanska Area (2018), Northern Water Services Board, Kenya

[2] Makkar, T., Kumar, Y., Dubey, A. K., Rocha, Á, &Goyal, A. (2017, December). Analogizing time complexity of KNN and CNN in recognizing handwritten digits. In 2017 Fourth International Conference on Image Information Processing (ICIIP) (pp. 1-6). IEEE.

[3] Mehdizadeh, S. (2020). Using AR, MA, and ARMA time series models to improve the performance of MARS and KNN approaches in monthly precipitation modeling under limited climatic data. Water Resources Management, 34(1), 263-282.

[4] Modaresi, F., Araghinejad, S., &Ebrahimi, K. (2018). A comparative assessment of artificial neural network, generalized regression neural network, least-square support vector regression, and K-nearest neighbor regression for monthly streamflow forecasting in linear and nonlinear conditions. Water Resources Management, 32(1), 243-258.

[5] Motevalli, A., Naghibi, S. A., Hashemi, H., Berndtsson, R., Pradhan, B., &Gholami, V. (2019). Inverse method using boosted regression tree and k-nearest neighbor to quantify effects of point and non-point source nitrate pollution in groundwater. Journal of cleaner production, 228, 1248-1263.

[6] Nababan, A. A., &Sitompul, O. S. (2018, April). Attribute weighting based K-nearest neighbor using gain ratio. In Journal of Physics: Conference Series (Vol. 1007, No. 1, p. 012007). IOP Publishing.

[7] Raseman, W. J., Rajagopalan, B., Kasprzyk, J. R., &Kleiber, W. (2020). Nearest neighbor time series bootstrap for generating influent water quality scenarios. Stochastic Environmental Research and Risk Assessment, 34(1), 23-31.

[8] Shahabi, H., Shirzadi, A., Ghaderi, K., Omidvar, E., Al-Ansari, N., Clague, J. J., ...& Ahmad, A. (2020). Flood detection and susceptibility mapping using sentinel-1 remote sensing data and a machine learning approach: Hybrid intelligence of bagging ensemble based on k-nearest neighbor classifier. Remote Sensing, 12(2), 266.