ISY5002 Continuous Assessment 1

Predicting water point functionality in Sierra Leone

Report

Introduction

Water pumps are widely used in rural areas in sub-Saharan Africa, but their maintenance remains a challenging issue even as new water supply infrastructure are being built. In Sierra Leone, about 25% of water point systems are estimated to be non-functional, depriving people of access to safe drinking water [1]. A number of factors has been found to be associated with non-functionality of water points including age of the system, absence of fee payment and lack of technical support [2].

The objective of the study is to predict the functionality of water points in Sierre Leone. Having knowledge of the working status of water pumps would enable officials to identify areas with most need for structural improvements and allow for precise budgeting for repair of damaged water systems.

Dataset

The Sierra Leone water point dataset was obtained from Sierra Leone WASH data portal [1]. Additional 2015 census data on the population density of various administrative regions was manually scrapped from the website City Population [3] and added to the first dataset to explore whether population density could be a determinant of water point functionality.

Ensemble model

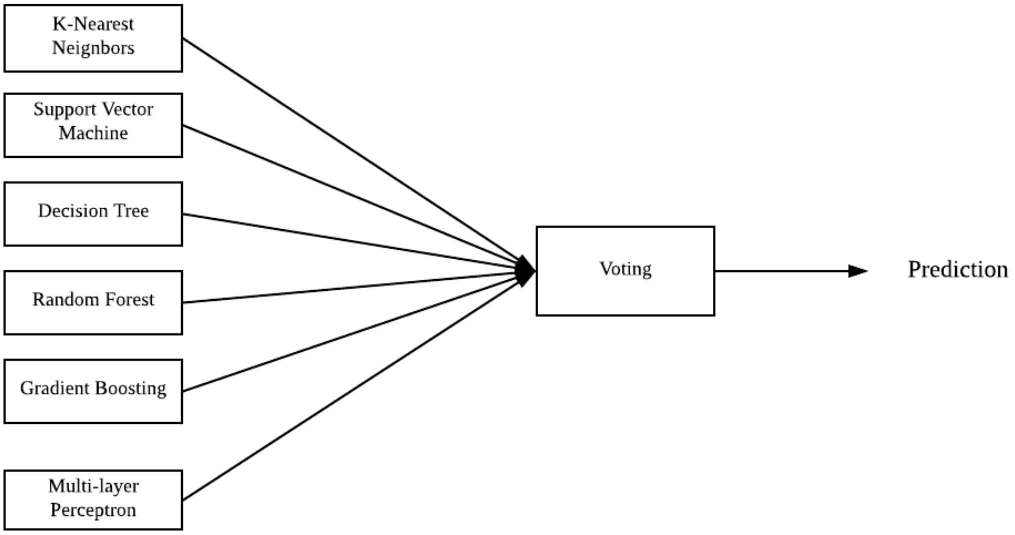


Figure 1 Ensemble with voting classifier and six sub-models

The dataset was split into training and testing in the ratio 80:20, and training set was further split into training and validation in the ratio 80:20 as well, such that training, validation and testing datasets were in the ratio 64:16:20.

In the ensemble model, processed data was passed into all or some of the following six sub-models: K Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting and Multi-layer Perceptron (). The output from each sub-model was then fed into a voting classifier to obtain predictions.

Exploratory data analysis

Exploratory data analysis was first performed to understand the characteristics of the dataset.

The data set comprised of 49 attributes, of which 41 were categorical and the remaining numeric. These attributes described 31540 water points in Sierra Leone, including type of pump, location, elevation, longitude, latitude, installation funding source, management method, year constructed, water source, and the quality of water delivered by the system.

The objective of this assessment was to predict the functionality of the water point. The functionality of water points is classified into six groups: Yes-Functional (and in use) ; Yes-Functional (but not in use); Yes - But damaged; No - Broken down; No - Still under construction; and No - under rehabilitation. The water points in this dataset were predominantly in the functional (and in use) category.

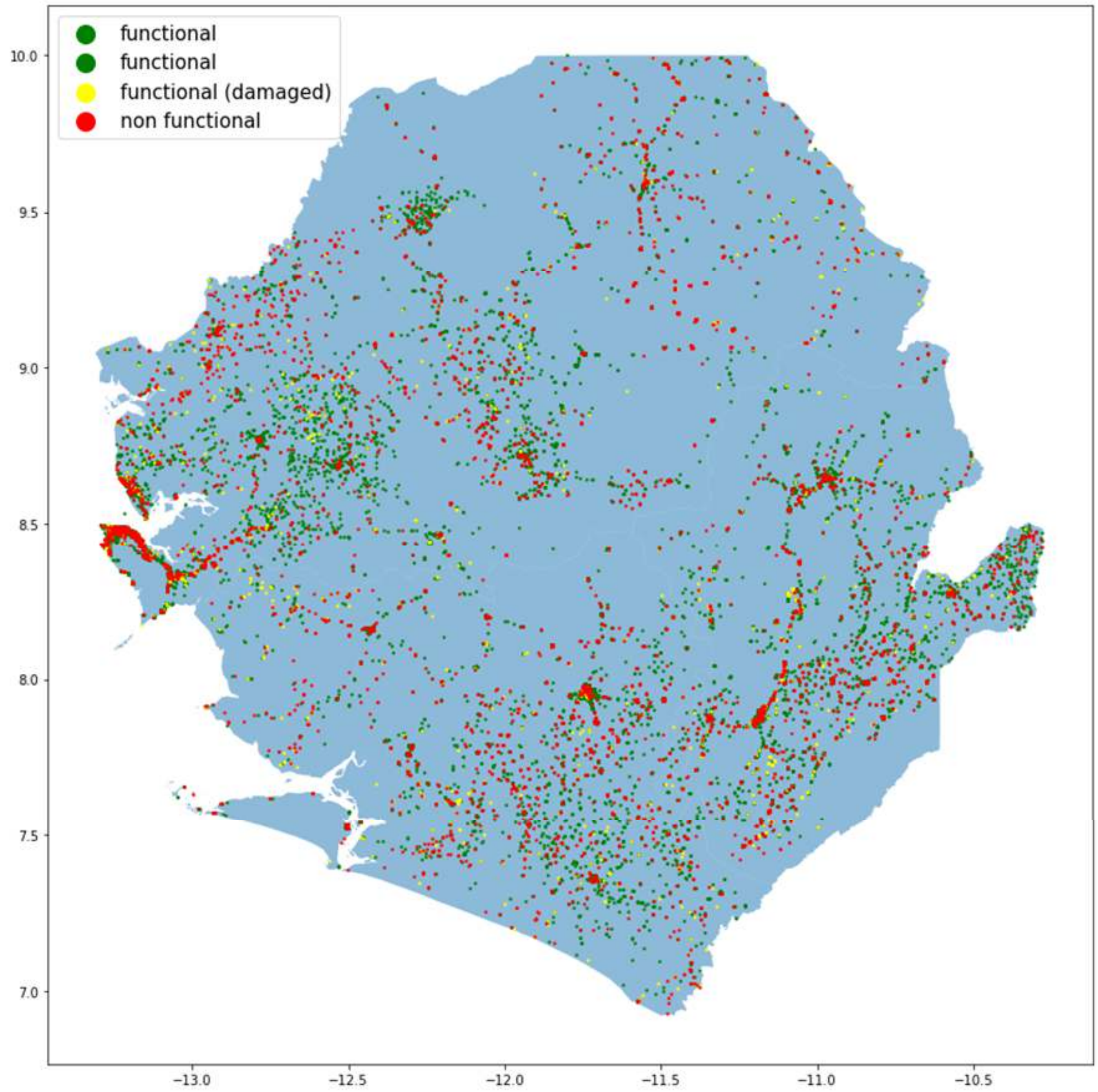


Figure 2 Functionality of pumps in Sierra Leone

Based on the visualisation in *Figure 2*, certain regions in the country seemed to have a higher concentration of non-functional water pumps.

Missing data was frequent in the dataset, with slightly less than half of the attributes (23) having more than 20% missing values (***Error! Reference source not found.***). Erroneous values were also occasionally seen in the data. For example, certain values for latitude and longitude were invalid.

Preprocessing and results

The data was separately preprocessed by each team member in order for everyone to have a chance to work on the data and each person took a different approach in processing the data. As the objective was to predict the working status of water points, those under construction or rehabilitation were excluded from analysis.

Approach

In this approach, it is decided to use PCA and LDA to select the features that were most significant.

Preprocessing

Most variables were dropped (*Table 1*), either because there was too much missing data or the information was deemed to be unrelated to functionality of the water point.

Selected variables were processed. For instance, values in type of water point and extraction system type that represented "other" were grouped into a single category.

Table 2 shows the variables that were retained as predictors for classification.

Variable	Reason for elimination
Submission Date	Information used to calculate age of water pump
16230052|EA number	Unsure what information represents
7420032|Community Name	Difficult to classify unstructured data
7430032|Water point Name	Irrelevant
5420051|Location	Location information captured in latitude and longitude
--GEOELE--|Elevation	Irrelevant
--GEOCODE--|Geo Code	Irrelevant
4430050|Photo	Irrelevant

440041|Pump type	Irrelevant
1410054|Number of taps at this point	Frequent missing data
5440041|Are you able to measure the depth of well? 0.308232	Frequent missing data
9410050|Measure the depth of the well (in metres) 0.894172	Frequent missing data
470044|When did the water point break down? 0.778853	Frequent missing data
6430039|Is/was this point monthly or regularly chlorinated? 0.283289	Frequent missing data
5420052|Does this Water point have any damage? 0.023560	Irrelevant
4390041|Is water available throughout the year? 0.000000	Irrelevant
2480001|During the seasonal drought of the well, how long is it not available? (months) 0.543972	Frequent missing data
4390042|Is/was this point used for drinking water 0.000000	Irrelevant
1450005|Why is this point not used for drinking water? 0.912533	Frequent missing data

7420038|Is the water paid for at this point?	Irrelevant
9410052|How reliable is the water point?	Frequent missing data
3480045|Is the water clean or is there a quality problem? 0.303081	Frequent missing data
4430055|Year of construction	Information used to calculate age of water pump
6400047|Installer / implementing agency	Frequency missing data
6430041|Others Installer / implementing agency 0.118848	Frequent missing data
4380054|Who is maintaining the water point (routine repairs)? 0.000000	Irrelevant
510001|Is the WASH management committee functioning? 0.580981	Frequent missing data
7430040|Is there a trained mechanic available at this point? 0.118752	Frequent missing data
6430044|Were trained mechanics provided with toolkits? 0.678081	Frequent missing data
460037|How many minutes does it take to reach the nearest spare part supplier? 0.000000	Irrelevant
1500002|Has the community been declared ODF? 0.000048	Sanitation facilities information irrelevant (ODF refers to open defeacation free)

2470038|Do you think the community is still ODF? 0.733308	Sanitation facilities information irrelevant
6540001|Are there functioning latrines in this village? 0.733308	Sanitation facilities information irrelevant
2540001|Do the latrines have handwashing facilities? 0.753577	Sanitation facilities information irrelevant
8480002|Are there trained natural ODF leaders in this community? 0.733308	Sanitation facilities information irrelevant
1530002|Are the trained natural ODF leaders performing their role effectively? 0.810998	Sanitation facilities information irrelevant
4690001|Observations about toilet	Sanitation facilities information irrelevant
6740003|Observe presence of water at the specific place for hand washing 0.020269	Sanitation facilities information irrelevant
3830002|Observe what device is present for hand washing 0.020269	Sanitation facilities information irrelevant
3810002|Record if soap or detergent is present at the specific place for hand washing 0.020269	Sanitation facilities information irrelevant
Unnamed: 48	Unsure what information represents

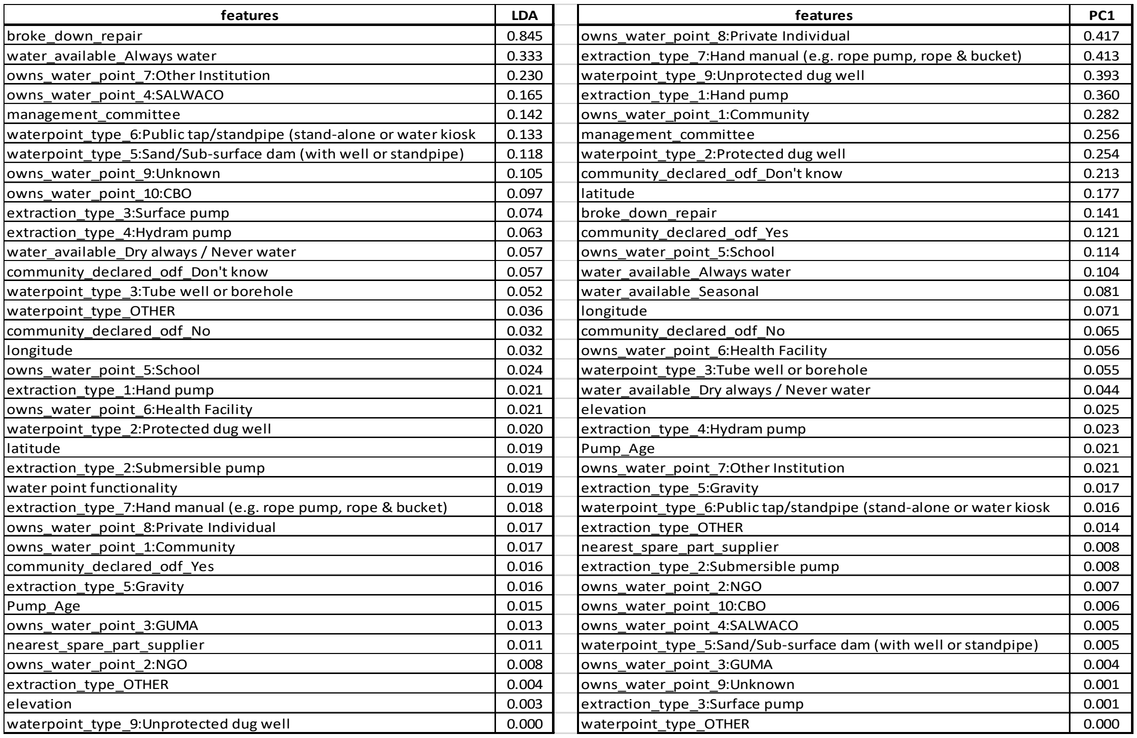
Table 1 Excluded variables in approach 1

Variable	Preprocessing
2420047|Latitude	- Observations with invalid values filtered - Standardization done
--GEOLON--|Longitude	- Observations with invalid values values filtered - Standardization done
5450040|Type of water point	- Values that represent "other" grouped into one category - One-hot encoding done
4420041|Extraction system type	- Values that represent "other" grouped into one category - One-hot encoding done
7430035|Water point Functionality	- Label encoding done
4380053|Last time the water point broke down, how long did it take to repair?	- Label encoding done
4390044|Who owns the water point?	- Label encoding done
7380052|Is there a WASH management committee?	- Label encoding done
Age of pump	- Standardization done

Table 2 Preprocessing in Approach 1

Feature importance selection

PCA and LDA were used to evaluate the features contribution weightage and assisting in finding which were the features to be removed and which were the features account for most variance in the data with the aim of retaining as much information as possible (*Figure 3*).



features	LDA	features	PC1
broke_down_repair	0.845	owns_water_point_8:Private Individual	0.417
water_available_Always water	0.333	extraction_type_7:Hand manual (e.g. rope pump, rope & bucket)	0.413
owns_water_point_7:Other Institution	0.230	waterpoint_type_9:Unprotected dug well	0.393
owns_water_point_4:SALWACO	0.165	extraction_type_1:Hand pump	0.360
management_committee	0.142	owns_water_point_1:Community	0.282
waterpoint_type_6:Public tap/standpipe (stand-alone or water kiosk	0.133	management_committee	0.256
waterpoint_type_5:Sand/Sub-surface dam (with well or standpipe)	0.118	waterpoint_type_2:Protected dug well	0.254
owns_water_point_9:Unknown	0.105	community_declared_odf_Don't know	0.213
owns_water_point_10:CBO	0.097	latitude	0.177
extraction_type_3:Surface pump	0.074	broke_down_repair	0.141
extraction_type_4:Hydram pump	0.063	community_declared_odf_Yes	0.121
water_available_Dry always / Never water	0.057	owns_water_point_5:School	0.114
community_declared_odf_Don't know	0.057	water_available_Always water	0.104
waterpoint_type_3:Tube well or borehole	0.052	water_available_Seasonal	0.081
waterpoint_type_OTHER	0.036	longitude	0.071
community_declared_odf_No	0.032	community_declared_odf_No	0.065
longitude	0.032	owns_water_point_6:Health Facility	0.056
owns_water_point_5:School	0.024	waterpoint_type_3:Tube well or borehole	0.055
extraction_type_1:Hand pump	0.021	water_available_Dry always / Never water	0.044
owns_water_point_6:Health Facility	0.021	elevation	0.025
waterpoint_type_2:Protected dug well	0.020	extraction_type_4:Hydram pump	0.023
latitude	0.019	Pump_Age	0.021
extraction_type_2:Submersible pump	0.019	owns_water_point_7:Other Institution	0.021
water point functionality	0.019	extraction_type_5:Gravity	0.017
extraction_type_7:Hand manual (e.g. rope pump, rope & bucket)	0.018	waterpoint_type_6:Public tap/standpipe (stand-alone or water kiosk	0.016
owns_water_point_8:Private Individual	0.017	extraction_type_OTHER	0.014
owns_water_point_1:Community	0.017	nearest_spare_part_supplier	0.008
community_declared_odf_Yes	0.016	extraction_type_2:Submersible pump	0.008
extraction_type_5:Gravity	0.016	owns_water_point_2:NGO	0.007
Pump_Age	0.015	owns_water_point_10:CBO	0.006
owns_water_point_3:GUMA	0.013	owns_water_point_4:SALWACO	0.005
nearest_spare_part_supplier	0.011	waterpoint_type_5:Sand/Sub-surface dam (with well or standpipe)	0.005
owns_water_point_2:NGO	0.008	owns_water_point_3:GUMA	0.004
extraction_type_OTHER	0.004	owns_water_point_9:Unknown	0.001
elevation	0.003	extraction_type_3:Surface pump	0.001
waterpoint_type_9:Unprotected dug well	0.000	waterpoint_type_OTHER	0.000

Figure 3 Comparison of features extracted with LDA and PCA

Results

After preprocessing, the data was passed into different classifiers: K-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, and Gradient Boosting. The accuracies obtained are shown in *Table 3*. The output from three of the models were then fed into the hard voting classifier and an accuracy of 75.90% was obtained.

Classifier	Accuracy (%)			
	Training	Validation	Testing	Ensemble
Random Forest	90.30	77.00	78.80	75.90
Gradient Boosting	100	76.30	77.90	
K-Nearest Neighbors	90.30	76.20	77.60	
Support Vector Machine	100	75.60	76.80	-

Table 3 Accuracies of different classifiers

6. Discussion

The dataset was processed in three different ways in this study. Eight features were selected in the first approach and the number of dimensions further reduced using LDA. But lower accuracies were obtained with this approach.

An unequal distribution of in the functionality of water points was observed. Despite the use of undersampling and over-sampling techniques, recall did not improve.

Based on the random forest classifier, water point damage and reliability of the water point were important features in predicting functionality of the water point. Population density also had a rather high feature importance, with a ranking of 22 out of 245 features.

The accuracy rate was the only measure used to determine performance of the models, but receiver operating curves could have been plotted for each class to visualize the performance and area under the curve (AUC) calculated to compare the classifiers.

Hard voting was used as it gives more weight to highly confidence votes and tend to give better performance than soft voting. However, accuracies did not improve and the wrong classifications were mainly for the minority groups.

Conclusion

Feature selection is crucial during preprocessing and eliminating too many features can lead to loss of information and a decline in prediction accuracy. Techniques for dimension reduction such as LDA must also be used with caution and with understanding of the data, otherwise the wrong features might be extracted.

The ensemble model using either the hard or soft voting classifier failed to improve on the accuracy of the sub-models. Other ensemble models such as an ensemble of neural networks can be attempted in future work to try and improve prediction accuracy further.

References

[1] “Sierra Leone WASH data portal,” [Online]. Available: https://washdata-sl.org/water-point-data/water-point-functionality/. [Accessed 03 September 2019].

[2] K. T, C. R, S. KF and B. J, “A categorization of water system breakdowns: Evidence from Liberia, Nigeria, Tanzania, and Uganda,” *Sci Total Environ,* vol. 1, pp. 619-620, 2018.

[3] “Sierre Leone Adeministrative Division,” City Popultation, [Online]. Available: https://www.citypopulation.de/php/sierraleone-admin.php. [Accessed 03 September 2019].

Appendix

Variable	No of missing values	Missing rate (%)
Submission Date	5210	16.57
16230052|EA number	8467	26.92
7420032|Community Name	5213	16.57
7430032|Water point Name	5212	16.57
5420051|Location	5213	16.57
2420047|Latitude	5213	16.57
--GEOLON--|Longitude	5213	16.57
--GEOELE--|Elevation	5263	16.73
--GEOCODE--|Geo Code	5228	16.22
4430050|Photo	5212	16.57
5450040|Type of water point	5210	16.57
4420041|Extraction system type	8191	26.04
440041|Pump type	19213	61.09
1410054|Number of taps at this point	24754	78.71
5440041|Are you able to measure the depth of well?	12775	40.62
9410050|Measure the depth of the well (in metres)	28784	91.52
7430035|Water point Functionality	5212	16.57
470044|When did the water point break down?	26226	83.39
4380053|Last time the water point broke down, how long did it take to repair?	5660	18.00
6430039|Is/was this point monthly or regularly chlorinated?	12186	38.75
5420052|Does this Water point have any damage?	5750	18.28

4390041|Is water available throughout the year?	5212	16.57
2480001|During the seasonal drought of the well, how long is it not available? (months)	19445	61.83
4390042|Is/was this point used for drinking water 0.000000	5211	16.57
1450005|Why is this point not used for drinking water?	29084	92.47
7420038|Is the water paid for at this point?	11120	7420038
9410052|How reliable is the water point?	11120	35.36
3480045|Is the water clean or is there a quality problem? 0.303081	12397	39.42
4430055|Year of construction	5212	16.57
6400047|Installer / implementing agency	9828	31.25
6430041|Others Installer / implementing agency	8303	26.40
4390044|Who owns the water point?	5212	16.57
4380054|Who is maintaining the water point (routine repairs)?	5212	16.57
7380052|Is there a WASH management committee?	5212	16.57
510001|Is the WASH management committee functioning?	21663	68.88
7430040|Is there a trained mechanic available at this point?	9833	31.26
6430044|Were trained mechanics provided with toolkits?	23868	75.89
460037|How many minutes does it take to reach the nearest spare part supplier?	5212	16.57
1500002|Has the community been declared ODF?	5213	16.57
2470038|Do you think the community is still ODF?	25211	80.16
6540001|Are there functioning latrines in this village?	25211	80.16
2540001|Do the latrines have handwashing facilities?	25681	81.65

8480002|Are there trained natural ODF leaders in this community?	25211	80.16
1530002|Are the trained natural ODF leaders performing their role effectively?	27210	86.52
4690001|Observations about toilet	436	1.39
6740003|Observe presence of water at the specific place for hand washing	436	1.39
3830002|Observe what device is present for hand washing	436	1.39
3810002|Record if soap or detergent is present at the specific place for hand washing	437	1.39
Unnamed: 48	5210	16.57

Table 4 Percentage of missing values in WASH dataset

Results: Approach 2

Functionality of water point	Precision	Recall	F1-score	Support
Yes – Functional (and in use)	0.89	0.97	0.93	3151
Yes - Functional (but not in use)	0.49	0.27	0.35	225
Yes - But damaged	0.58	0.36	0.44	332
No - Broken down	0.87	0.82	0.84	967

Table 5 Classification report for K-Nearest Neighbors classifier

3046	*8*	*77*	*20*
74	*61*	*1*	*89*
194	*4*	*119*	*15*
112	*52*	*7*	*796*

Table 6 Confusion matrix for K-Nearest Neighbors classifier

Functionality of water point	Precision	Recall	F1-score	Support
Yes – Functional (and in use)	0.91	1.00	0.95	3151
Functional (but not in use)	0.70	0.13	0.22	225
Yes - But damaged	0.80	0.06	0.11	332
No - Broken down	0.83	0.99	0.90	967

Table 7 Classification report for Support Vector Machine classifier

3146	*0*	*5*	*0*
0	*30*	*0*	*195*
312	*0*	*20*	*0*
0	*13*	*0*	*954*

Table 8 Classification report for Support Vector Machine classifier

Functionality of water point	Precision	Recall	F1-score	Support
Yes – Functional (and in use)	0.90	1.00	0.95	3151
Functional (but not in use)	0.81	0.15	0.25	225
Yes - But damaged	1.00	0.00	0.01	332
No - Broken down	0.83	0.99	0.91	967

Table 9 Classification report for Random Forest classifier

3151	*0*	*0*	*0*
0	*34*	*0*	*191*
331	*0*	*1*	*0*
0	*8*	*0*	*959*

Table 10 Confusion matrix for Random Forest Classification classifier

Functionality of water point	Precision	Recall	F1-score	Support
Yes – Functional (and in use)	0.93	0.97	0.95	3151
Functional (but not in use)	0.59	0.28	0.38	225
Yes - But damaged	0.53	0.33	0.41	332
No - Broken down	0.85	0.96	0.90	967

Table 11 Classification report for Decision Tree classifier

3054	0	97	0
0	63	0	162
222	0	110	0
0	43	0	924

Table 12 Confusion matrix for Decision Tree classifier

Functionality of water point	Precision	Recall	F1-score	Support
Yes – Functional (and in use)	0.94	0.97	0.96	3151
Functional (but not in use)	0.74	0.46	0.57	225
Yes - But damaged	0.65	0.44	0.53	332
No - Broken down	0.88	0.96	0.92	967

Table 13 Classification report for Gradient Boosting classifier

3071	0	80	0
0	104	0	121
185	0	147	0
0	37	0	930

Table 14 Confusion matrix for Gradient Boosting classifier

Functionality of water point	Precision	Recall	F1-score	Support
Yes – Functional (and in use)	0.93	0.98	0.95	3151
Functional (but not in use)	0.68	0.40	0.51	225
Yes - But damaged	0.62	0.30	0.41	332
No - Broken down	0.87	0.96	0.91	967

Table 15 Classification report for Multi-layer Perceptron classifier

3090	0	61	0
0	91	0	134
231	0	101	0
0	43	0	924

Table 16 Confusion matrix for Multi-layer Perceptron classifier

Results: Approach 3

Functionality of water point	Precision	Recall	F1-score	Support
Yes – Functional (and in use)	0.88	0.98	0.93	2624
Functional (but not in use)	0.82	0.45	0.56	221
Yes - But damaged	0.66	0.35	0.45	297
No - Broken down	0.98	0.96	0.97	901

Table 17 Classification report for K-Nearest Neighbors

2560	*8*	*57*	*1*
116	*97*	*6*	*3*
181	*2*	*111*	*3*
17	*2*	*0*	*882*

Table 18 Confusion matrix for K-Nearest Neighbors classifier

Functionality of water point	Precision	Recall	F1-score	Support
Yes – Functional (and in use)	0.94	0.98	0.96	2624
Yes – Functional (but not in use)	1.00	1.00	1.00	221
Yes - But damaged	0.74	0.41	0.53	297
No - Broken down	1.00	1.00	1.00	901

Table 19 Classification report for Support Vector Machine classifier

2581	0	43	0
2	221	0	0
174	0	123	0
1	0	0	900

Table 20 Confusion matrix for Support Vector Machine classifier

	Precision	Recall	F1-score	Support
Functional (and in use)	0.92	0.98	0.95	2624
Yes – Functional (but not in use)	1.00	1.00	1.00	221
Yes - But damaged	0.64	0.29	0.40	297
No - Broken down	1.00	1.00	1.00	901

Table 21 Classification report for Decision Tree with D3

2576	0	48	0
0	221	0	0
211	0	86	0
0	0	0	901

Table 22 Confusion matrix for Decision Tree classifier

	Precision	Recall	F1-score	Support
Functional (and in use)	0.88	1.00	0.94	2624
Yes – Functional (but not in use)	1.00	0.72	0.84	221
Yes - But damaged	1.00	0.01	0.02	297
No - Broken down	1.00	1.00	1.00	901

Table 23 Classification report for Random Forest classifier

2624	0	0	0
62	159	0	0
294	0	3	0
0	0	0	901

Table 24 Confusion matrix for Random Forest classifier

	Precision	Recall	F1-score	Support
Functional (and in use)	0.95	0.98	0.96	2624
Yes – Functional (but not in use)	1.00	1.00	1.00	221
Yes - But damaged	0.73	0.51	0.60	297
No - Broken down	1.00	1.00	1.00	901

Table 25 Classification report for Gradient Boosting classifier

2568	0	56	56
0	221	0	0
147	0	150	150
0	0	0	901

Table 26 Confusion matrix for Gradient Boosting classifier

	Precision	Recall	F1-score	Support
Functional (and in use)	0.94	0.97	0.95	2624
Yes – Functional (but not in use)	1.00	1.00	1.00	221
Yes - But damaged	0.64	0.47	0.55	297
No - Broken down	1.00	1.00	1.00	901

Table 27 Classification report for Multi layer Peceptron (MLP) with D3

2545	0	79	0
0	221	0	0
156	0	141	0
0	0	0	900

Table 28 Confusion matrix for Multi-layer Perceptron classifier