
Machine Learning

Data Mining the Water Table

TAs
ntu.mlta@gmail.com

Task description

- Using data from Taarifa and the Tanzanian Ministry of Water, can you predict which pumps are functional, which need some repairs, and which don't work at all?
- A smart understanding of which waterpoints will fail can improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania.



Goal

- Predict the operating condition of a waterpoint.

Data

- Training data: 59400
- Testing data: 14851
- Feature: date_recorded, funder, installer...
- Multi-class classification: functional, non functional, functional needs repair.

Format

Feature data example

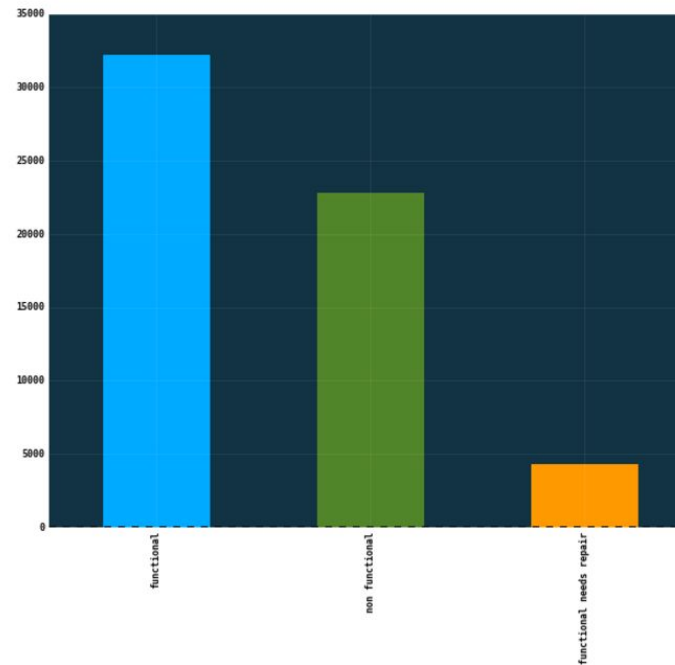
id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name
69572	6000	2011-03-14	Roman	1390	Roman	34.93809275	-9.85632177	none
8776	0	2013-03-06	Grumeti	1399	GRUMETI	34.6987661	-2.14746569	Zahanati
34310	25	2013-02-25	Lottery	Club	686	World	vision	37.46066446
67743	0	2013-01-28	Unicef	263	UNICEF	38.48616088	-11.15529772	Zahanati
19728	0	2011-07-13	Action	In	A	0	Artisan	31.13084671
9944	20	2011-03-13	Mkinga	Distric	Coun	0	DWE	39.1727956
19816	0	2012-10-01	Dwsp	0	DWSP	33.36240982	-3.76636472	Kwa
54551	0	2012-10-09	Rwssp	0	DWE	32.62061707	-4.22619802	Tushirikiane
53934	0	2012-11-03	Wateraid	0	Water	Aid	32.71110001	-5.14671181
46144	0	2011-08-03	Isingiro	Ho	0	Artisan	30.62699053	-1.25705061
49056	0	2011-02-20	Private	62	Private	39.20951812	-7.03413939	Mzee
50409	200	2013-02-18	Danida	1062	DANIDA	35.77025785	-10.57417468	Kwa
36957	0	2012-10-14	World	Vision	0	World	vision	33.79810612
50495	0	2013-03-15	Lawatefuka	Water	Supply	1368	Lawatefuka	water
53752	0	2012-10-20	Biore	0	WEDECO	34.36407268	-3.62933335	Mwabasabi
61848	0	2011-08-04	Rudep	1645	DWE	31.44412134	-8.27496163	Kwa
48451	500	2011-07-04	Unicef	1703	DWE	34.64243884	-9.10618458	Kwa
58155	0	2011-09-04	Unicef	1656	DWE	34.56926611	-9.08551497	Kwa

Sample submission

```
id,status_group
69572,functional
8776,functional
34310,functional
67743,non functional
19728,functional
9944,functional
19816,non functional
54551,non functional
53934,non functional
46144,functional
49056,functional
50409,functional
36957,functional
50495,functional
53752,functional
61848,functional
48451,non functional
58155,non functional
34169,functional needs repair
18274,functional
48375,functional
6091,functional
58500,functional needs repair
37862,functional
51058,functional
22308,functional needs repair
55012,functional
20145,functional
19685,non functional
```

Characteristics

- Mostly categorical data: one-of-n encoding may lead to data sparsity.
- imbalanced label space.
- Missing value.



Evaluation

Accuracy.