



PUMP IT UP

DATA MINING THE WATER TABLE

PROGETTO DI STATISTICA

Giorgio De Caro - Pelucchi Mauro

Obiettivo

Tanzania

Individuare i pozzi di acqua non funzionanti per minimizzare i costi di manutenzione e garantire alla maggior parte della popolazione acqua potabile pulita.

Il Ministero dell'acqua del paese africano, dal 2006, ha avviato progetti, politiche e fatto investimenti per ottimizzare la gestione e la mappatura degli impianti per la distribuzione dell'acqua.



Solo il 55% della popolazione ha accesso all'acqua potabile (dati 2006)

Popolazione
51.820.000

Acqua
6,4%

Matrice dei costi

Predizione

Osservazioni

		To Repare	No Repare
To Repare	0 TP	15 FN	
No Repare	5 FP	0 TN	

5

Il costo di un'uscita a vuoto di una squadra di manutenzione sul territorio.

15

Il costo sanitario della non riparazione di un pozzo non funzionante.

Classificare un evento nel quadrante dei falsi negativi è tre volte più costoso rispetto all'uscita di una squadra su un impianto non funzionante.

Decidiamo di assegnare un costo 5 agli eventi "falso positivi" ed un costo 15 agli eventi "falsi negativi".

La scelta di questi costi è dettata dal fatto che è molto oneroso inviare una squadra di manutentori su un impianto da non riparare ("uscita a vuoto") ma è molto più costoso (in termini sanitari e di salute pubblica) non riparare un impianto guasto ("falsi positivi").

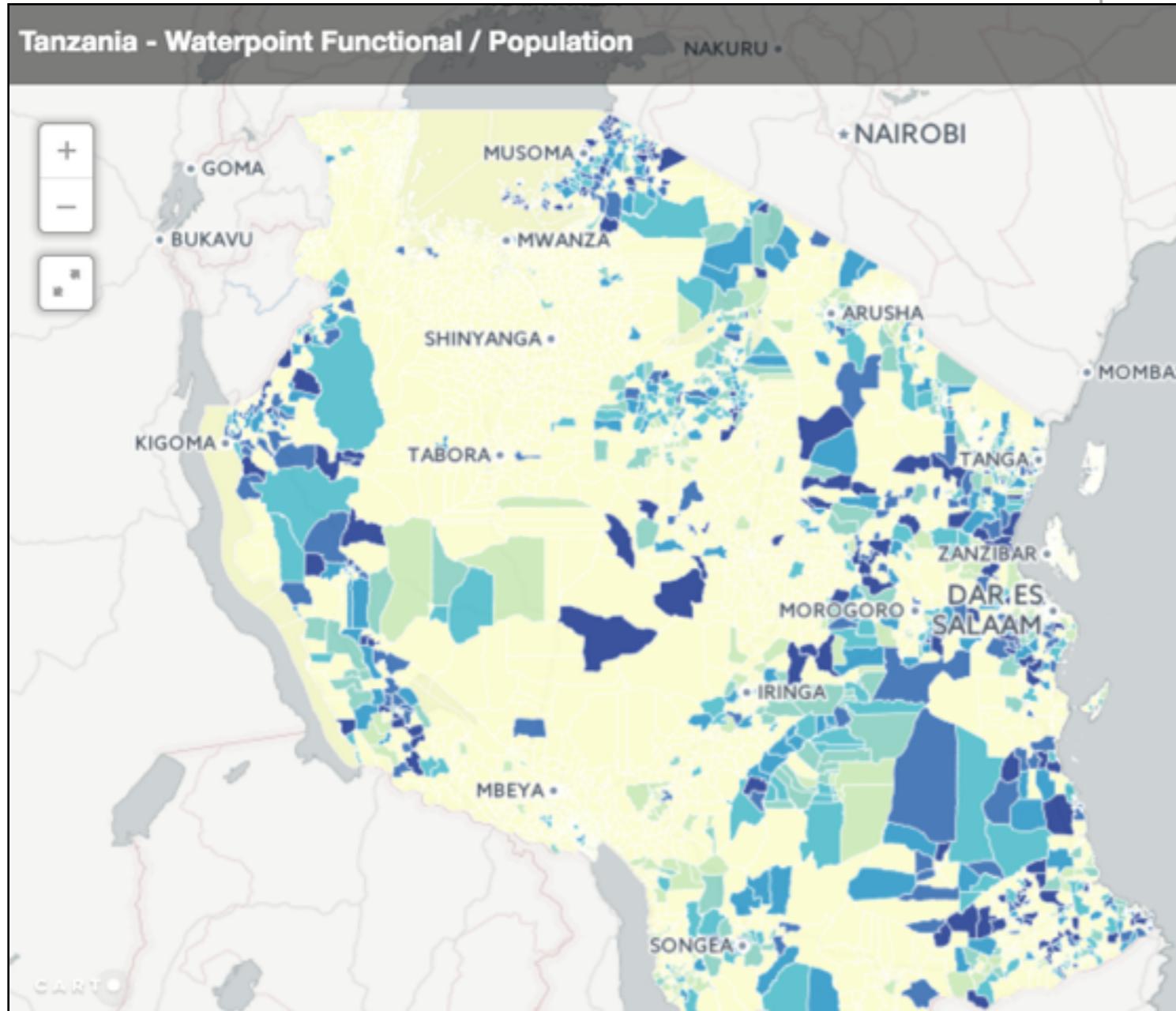
$$p^*_{1i} > \frac{1}{1 + \frac{\delta_{TP} - \delta_{FN}}{\delta_{TN} - \delta_{FP}}}$$

0.25

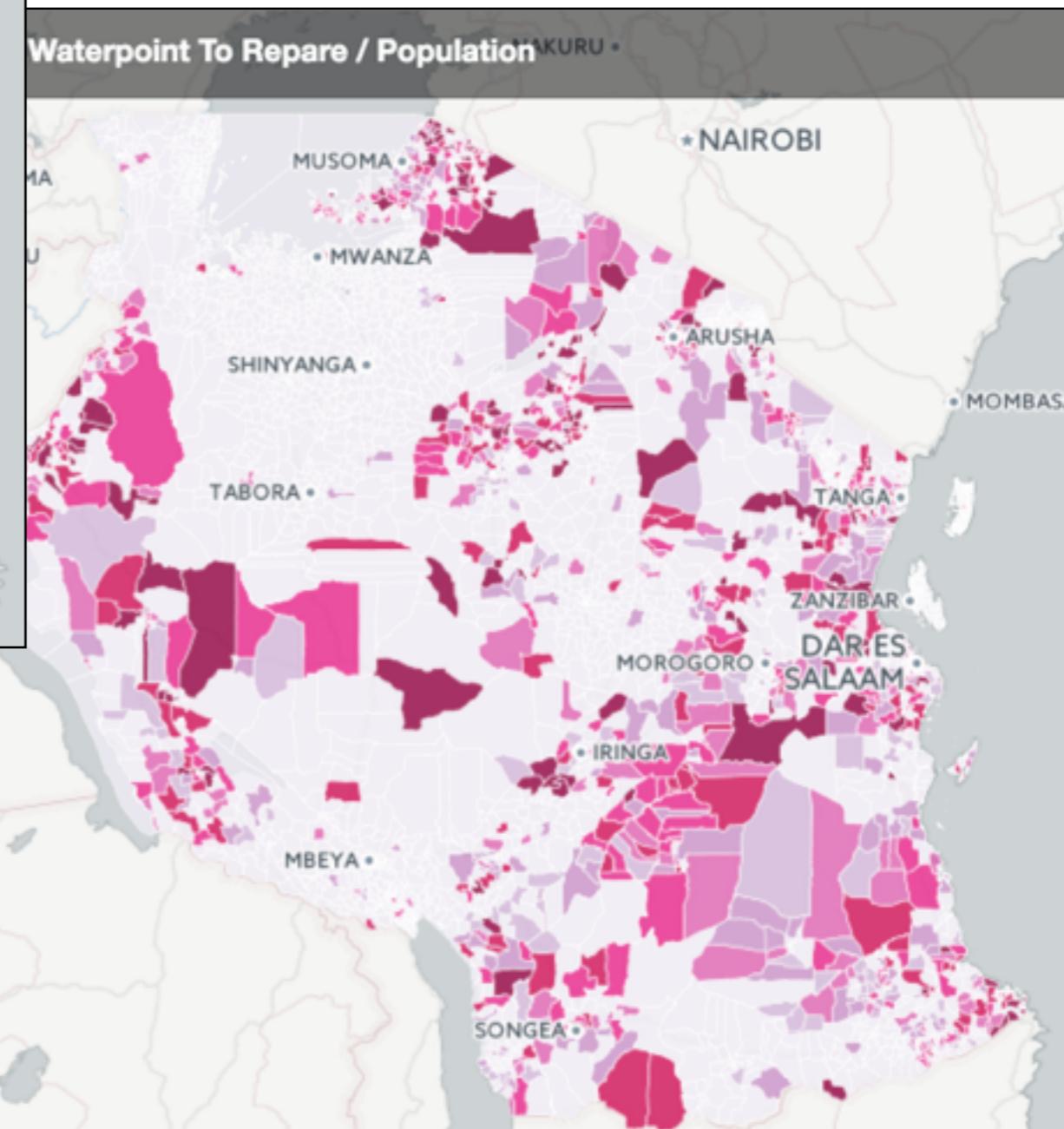
**Obiettivo: avere un
FNR < 6%**

< 6%

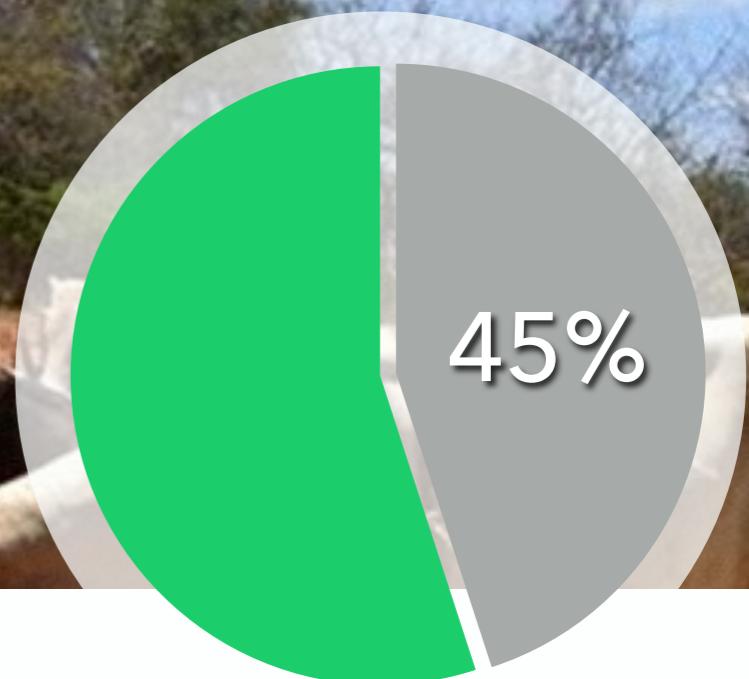
La situazione dei pozzi in Tanzania



I dati raccontano una storia: il 45% dei pozzi necessita riparazione. Non è possibile tenere il 100% di acqua potabile.



Output del progetto potrebbe essere un sistema intelligente che aiuti il Ministero dell'acqua e i vari user-group a indirizzare al meglio lo sforzo di manutenzione e di gestione favorendo la cooperazione fra user-group.



Il **45%** dei pozzi risultano non funzionanti nonostante gli interventi e i progetti del Ministero dell'acqua.

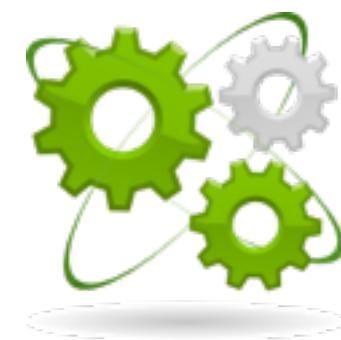
I malfunzionamenti si concentrano in certi distretti amministrativi o zone geografiche del paese?

In altre parole esiste un legame causa-effetto: vogliamo verificare se gli impianti sono da riparare per **cause naturali, problemi amministrativi** o di **mal gestione degli impianti**.

Metodologia e fasi del lavoro



Comprensione del campo applicativo:
ricerche sul progetto del Ministero dell'acqua
della Tanzania



Creazione di un insieme di dati per l'analisi: caricamento dei dataset in SAS, integrazione e pulizia



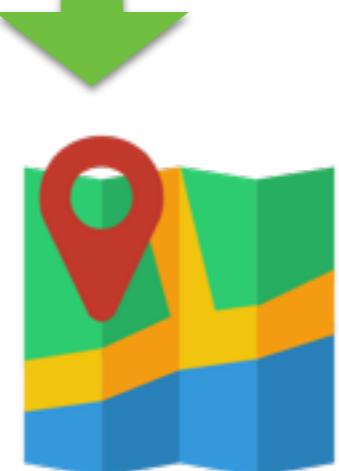
Esplorazione dei dati:
valutazione delle variabili più importanti



Analisi dei dati e ricerca di pattern: verifica dei risultati ottenuti e applicazione dei dati al set di score



Scelta degli algoritmi di data mining: selezione degli algoritmi e progetto in Miner



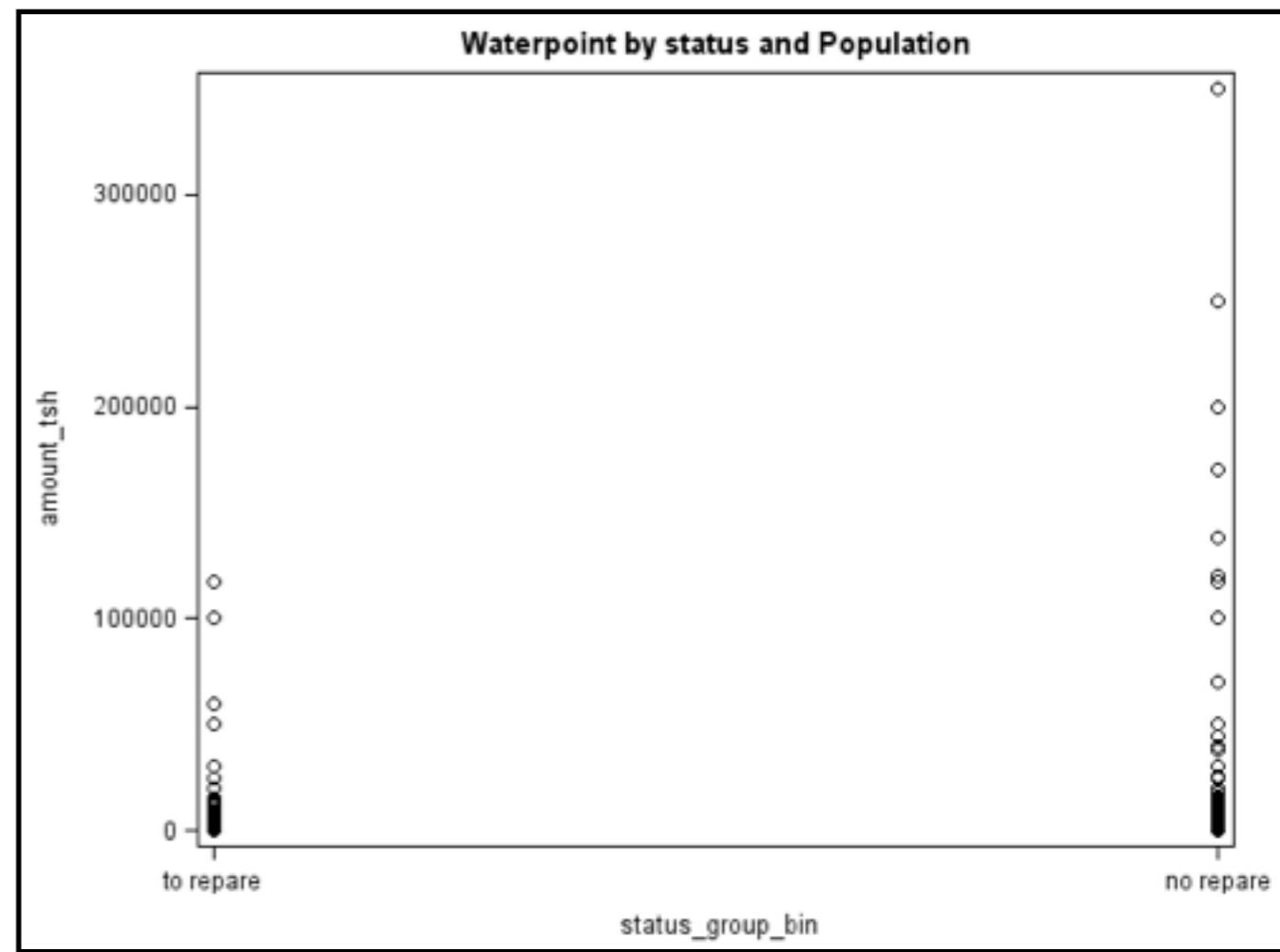
Presentazione dei risultati: commento ai risultati e creazione della mappa; individuazione di eventuali legami causa-effetto.

Data Quality ed esplorazione



Amount_tsh

La covariata **“Amount_tsh”** rappresenta il volume di acqua disponibile nel pozzo. Come si vede dallo strip plot, i pozzi più grandi non hanno problemi di manutenzione. La variabile presenta molti dati missing (andiamo a imputarli attraverso la mediana).



Basin

La variabile **Basin** indica il bacino di riferimento del waterpoint. Vediamo dei casi anomali:

- le pompe relative al lago Rukwa e alla costa sud, non funzionanti, sono superiori del 50% rispetto a quelle funzionanti (un rapporto di 1,5);
- gli impianti situati sul fiume Ruvuma, e sulla costa sud hanno un rischio di rotture 3 volte superiore a quelli del lago Nyasa.

	basin	count	c1	c0	odds
1	lake nyasa	5085	3324	1761	0.5297833935
2	rufiji	7976	5068	2908	0.5737963694
3	pangani	8940	5372	3568	0.6641846612
4	internal	7785	4482	3303	0.7369477912
5	wami / ruvu	5987	3136	2851	0.909119898
6	lake victoria	10248	5100	5148	1.0094117647
7	lake tanganyika	6432	3107	3325	1.0701641455
8	lake rukwa	2454	1000	1454	1.454
9	ruvuma / southern coast	4493	1670	2823	1.6904191617



Payment Type

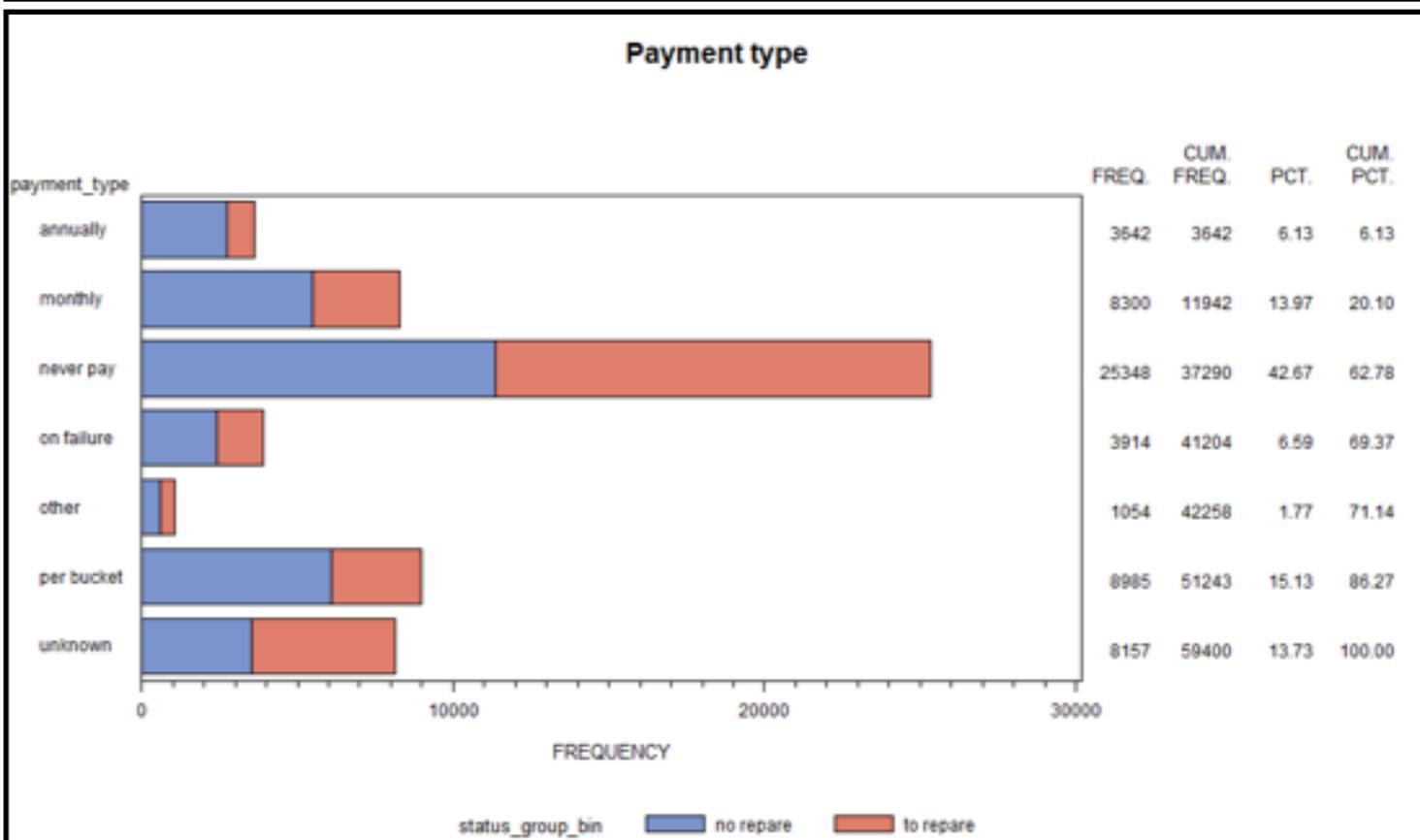
Il tipo pagamento indica la modalità di pagamento che gli utenti devono affrontare. Sicuramente vediamo come pozzi totalmente gratuiti abbiano un'attitudine maggiore ad essere malfunzionanti. Il valore "unknow" è da imputare (attraverso un albero decisionale possiamo derivare la variabile).

Il dataset presenta molte covariate simili ad un grado diverso di granularità. In questo caso (payment, payment type sono identiche) andiamo ad eliminare il grado che spiega di meno per rendere più robusti i modelli (ed evitare collinearità).

Management

La tipologia di gestione è un dato importante: in base a questo indicatore possiamo capire se è migliore la gestione degli user-group (WUA, WUA, VWC, ...), del Ministero dell'Acqua o dei privati. In generale possiamo dire che gli waterpoint gestiti direttamente dal VWC hanno un'attitudine a malfunzionamenti 3 volte superiori a quelli gestiti da operatori privati. I dati missing o other possono essere imputati attraverso un albero.

	payment_type	count	c1	c0	odds
1	annually	3642	2740	902	0.3291970803
2	per bucket	8985	6090	2895	0.4753694581
3	monthly	8300	5482	2818	0.5140459686
4	on failure	3914	2429	1485	0.6113627007
5	other	1054	611	443	0.7250409165
6	never pay	25348	11379	13969	1.2276122682
7	unknown	8157	3528	4629	1.3120748299



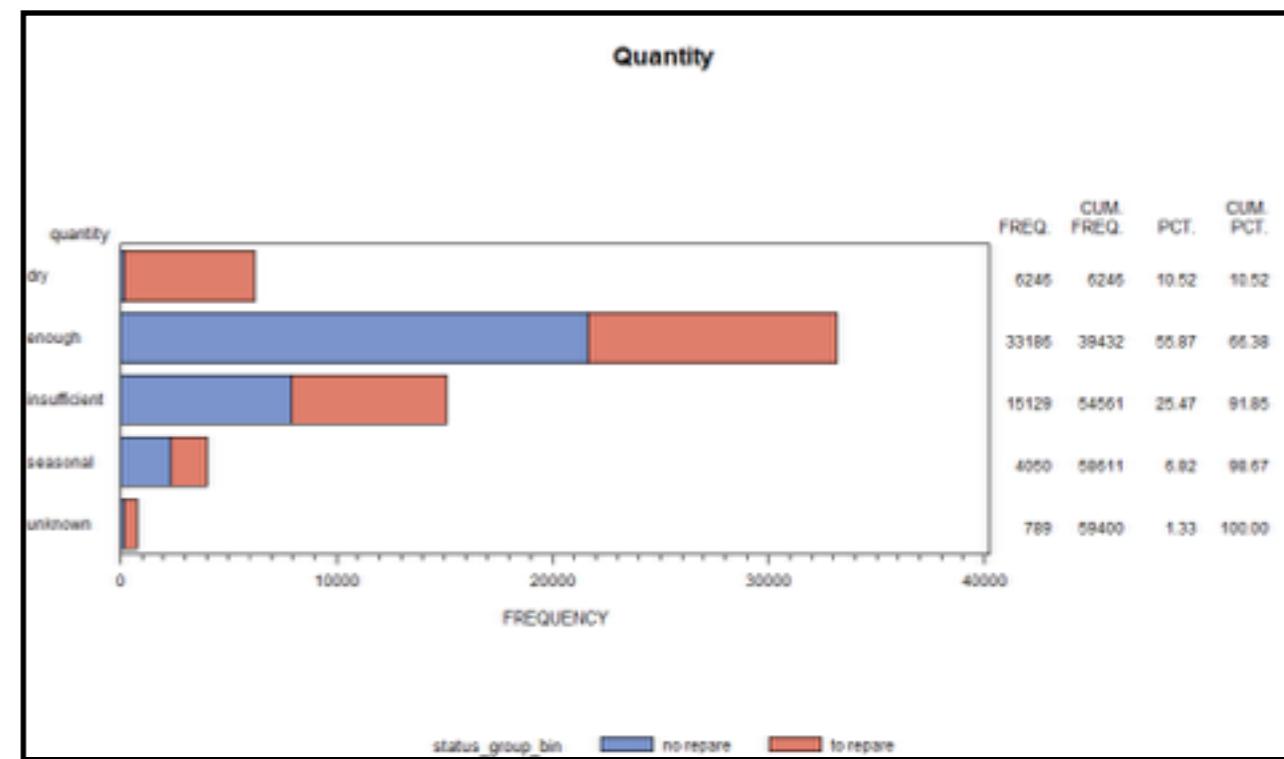
	management	count	c1	c0	odds
1	private operator	1971	1476	495	0.3353658537
2	water board	2933	2170	763	0.3516129032
3	wua	2535	1751	784	0.4477441462
4	wug	6515	3906	2609	0.6679467486
5	other	844	505	339	0.6712871287
6	trust	78	46	32	0.6956521739
7	parastatal	1768	1020	748	0.7333333333
8	vwc	40507	20425	20082	0.9832068543
9	water authority	904	446	458	1.0269058296
10	unknown	561	224	337	1.5044642857
11	company	685	267	418	1.5655430712
12	other - school	99	23	76	3.3043478261



Quantity

La variabile **quantity** indica la quantità di acqua osservata in uscita dal pozzo. Spiega molto bene il fenomeno; sicuramente abbiamo una correlazione fra questa variabile e lo stato della pompa di acqua.

Abbiamo circa 790 casi di dati missing (valore unknown); osserviamo inoltre che quando il pozzo è segnalato secco il rischio di trovarlo malfunzionante è di circa 38 volte rispetto alla possibilità di trovarlo in buono stato.



	quantity	count	c1	c0	odds
1	enough	33186	21648	11538	0.5329822616
2	seasonal	4050	2325	1725	0.7419354839
3	insufficient	15129	7916	7213	0.9111925215
4	unknown	789	213	576	2.7042253521
5	dry	6246	157	6089	38.78343949

Source

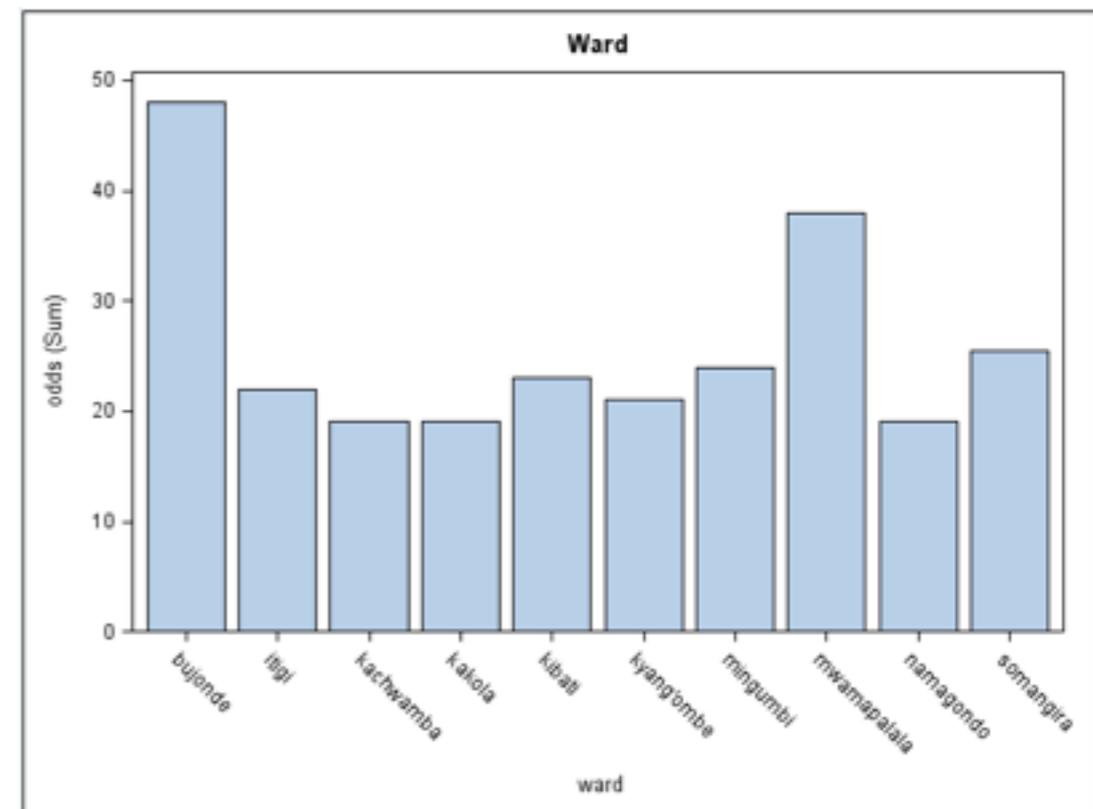
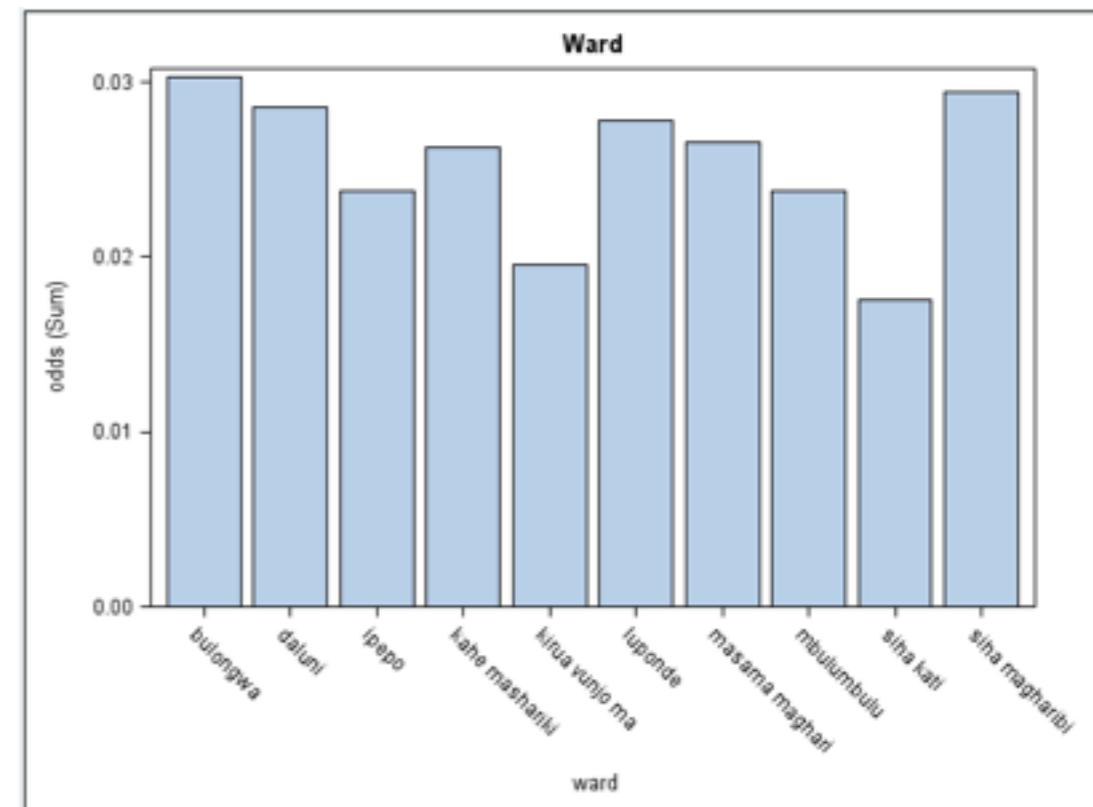
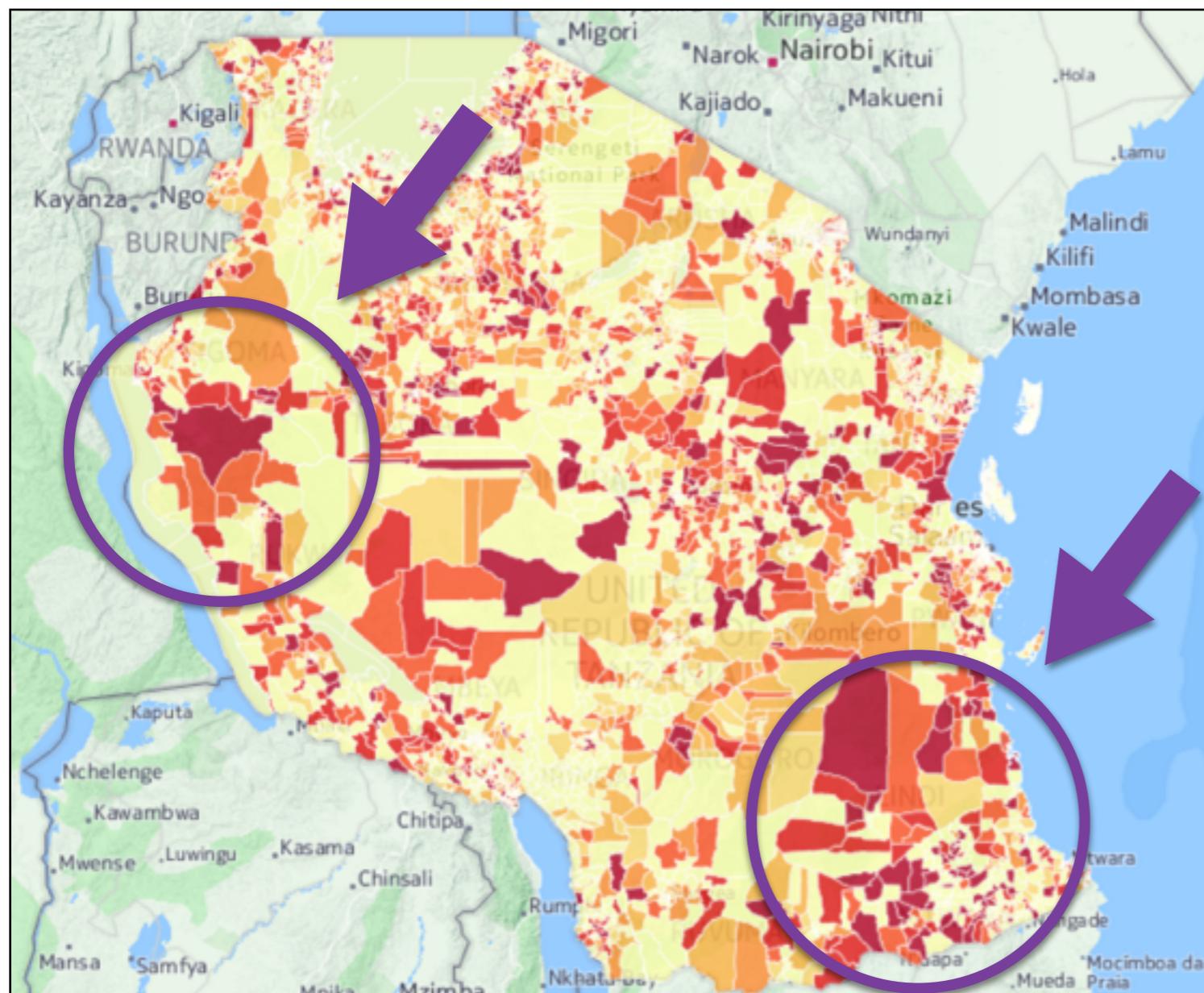
Mentre la covariata **source_class** è troppo generica, la variabile **source** mostra che quando la sorgente è un lago (o un bacino idrico) il rischio di trovare un impianto malfunzionante è di 3,7 e 1,5 volte quello di trovarlo funzionante. Attenzione alla collinearità di questo dato con il bacino di riferimento e la posizione geografica.

	source	count	c1	c0	odds
1	spring	17021	10592	6429	0.6069675227
2	rainwater harvesting	2295	1386	909	0.6558441558
3	.	212	126	86	0.6825396825
4	hand dtw	874	497	377	0.7585513078
5	river	9612	5465	4147	0.7588289113
6	shallow well	16824	8324	8500	1.0211436809
7	machine dbh	11075	5422	5653	1.0426042051
8	unknown	66	32	34	1.0625
9	dam	656	253	403	1.5928853755
10	lake	765	162	603	3.7222222222



Ward

La Tanzania è divisa in distretti amministrativi (ward). Ogni distretto attua le proprie politiche di gestione e manutenzione dei pozzi. Abbiamo distretti dove è altissimo il rischio di trovare un pozzo malfunzionante, contro distretti dove gli waterpoint sono sempre funzionanti.



[Link to CardoDB Map](#)

Regressione logistica



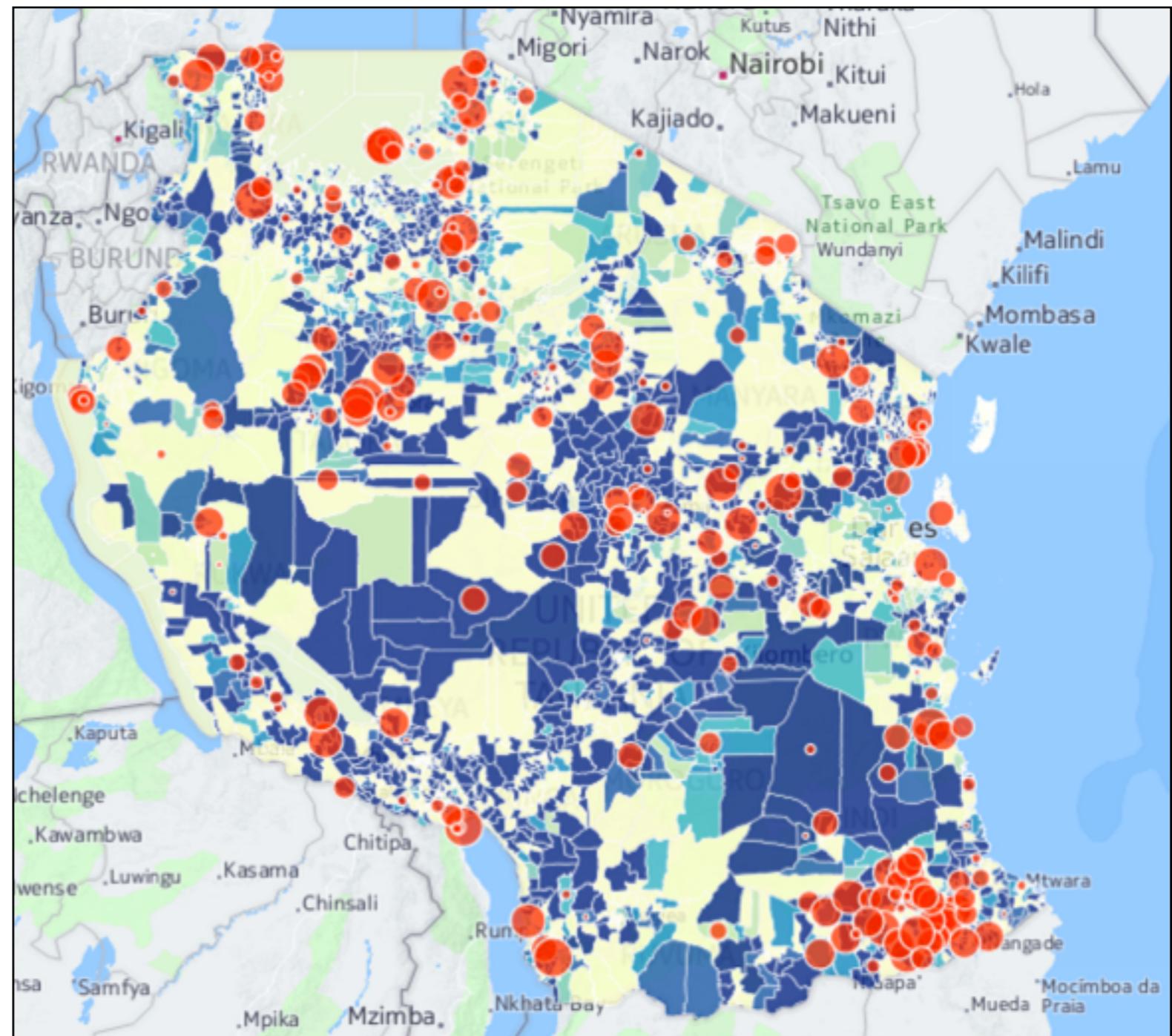
Verifica dei dati e pulizia: il dataset risulta bilanciato e le prior sono reali. Il 54% dei pozzi risultano agibili. Attraverso una regressione logistica in SAS troviamo le variabili importanti ed eventuali outlier (usando l'indice Pearson Chi square FIT Statistic).

```
ods graphics on;
proc logistic data=lib.training_set_bin;
class
status_group_bin (ref="to repare") waterpoint_type
source quantity payment_type
extraction_type basin region ward management
/param=ref
;
model status_group_bin=
waterpoint_type source quantity payment_type
extraction_type basin region ward management
/selection=stepwise pevent=0.46 details lackfit;
output out=lib.training_set_bin_c DIFCHISQ=difchi ;
run; quit;
ods graphics off;
```

The FREQ Procedure				
status_group_bin	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no repare	32259	54.31	32259	54.31
to repare	27141	45.69	59400	100.00
Number of Observations Read			59400	
Number of Observations Used			59400	
Response Profile				
Ordered Value	status_group_bin	Total Frequency		
1	no repare	32259		
2	to repare	27141		

Step	Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	quantity		4	1	8625.3830		<.0001
2	ward		2091	2	13499.8247		<.0001
3	waterpoint_type		6	3	3727.8558		<.0001
4	payment_type		6	4	983.6659		<.0001
5	extraction_type		16	5	616.8636		<.0001
6	source		9	6	498.5827		<.0001
7	management		11	7	346.9256		<.0001
8	region		20	8	160.2345		<.0001
9	basin		8	9	31.1761		0.0001

- Il **distretto amministrativo** è significativamente correlato al funzionamento o meno di un pozzo.
- Il **tipo di gestione** (riassunto in **pubblico**, **privato**, **utenti** e **altro**) influisce, e questo fa pensare a cause amministrative o di scarsa organizzazione: nella mappa sono evidenziati in blu i distretti con una percentuale di pozzi "auto gestita" più elevata. I cerchi rossi identificano il rischio di trovare un pozzo non funzionante: la maggior parte di distretti con maggioranza di waterpoint non funzionanti sono collocati dove la gestione è in mano agli utilizzatori.



[Link to CardoDB Map](#)

- Le cause naturali, definite dalle variabili come il **bacino**, la **tipologia di pozzo** e la **sorgente**, influiscono sul funzionamento: una sorgente stagionale oppure l'acqua piovana sono indicatori di possibili malfunzionamenti.



OUTLIER

- Troviamo nel dataset **2501** outlier (waterpoint chiaramente distinti dagli altri e quindi che necessitano di un'analisi particolare).
- Analizzando i dati troviamo qualcosa di interessante: nel distretto di **Sengerema**, in particolare nel ward amministrativo "**Katunguru**" ci sono moltissimi outlier. L'area ha elevato rischio di pozzi non funzionanti. Il fenomeno potrebbe essere dovuto alla diffusione nell'area di pompe manuali (che sono più soggette a malfunzionamenti).

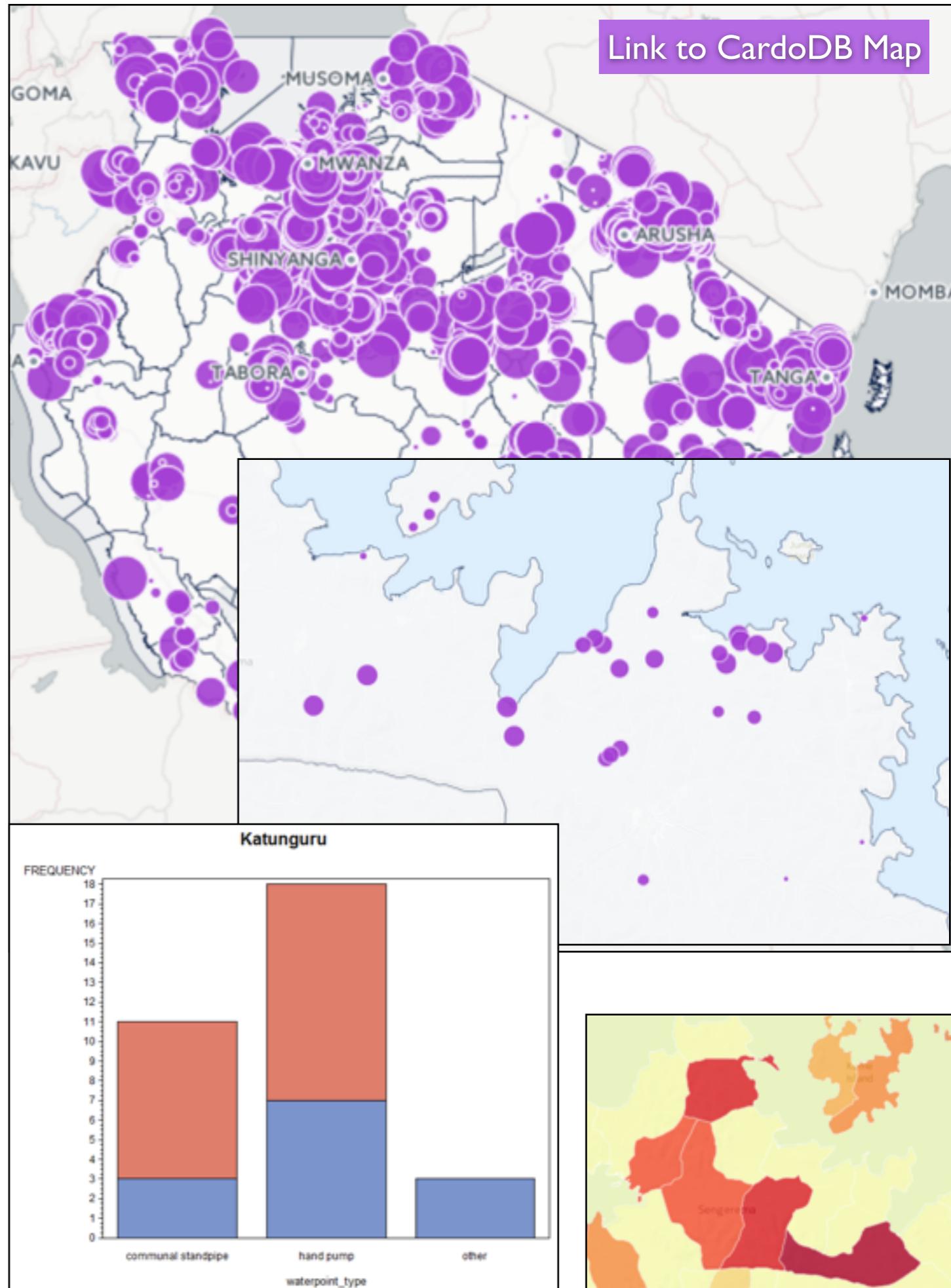
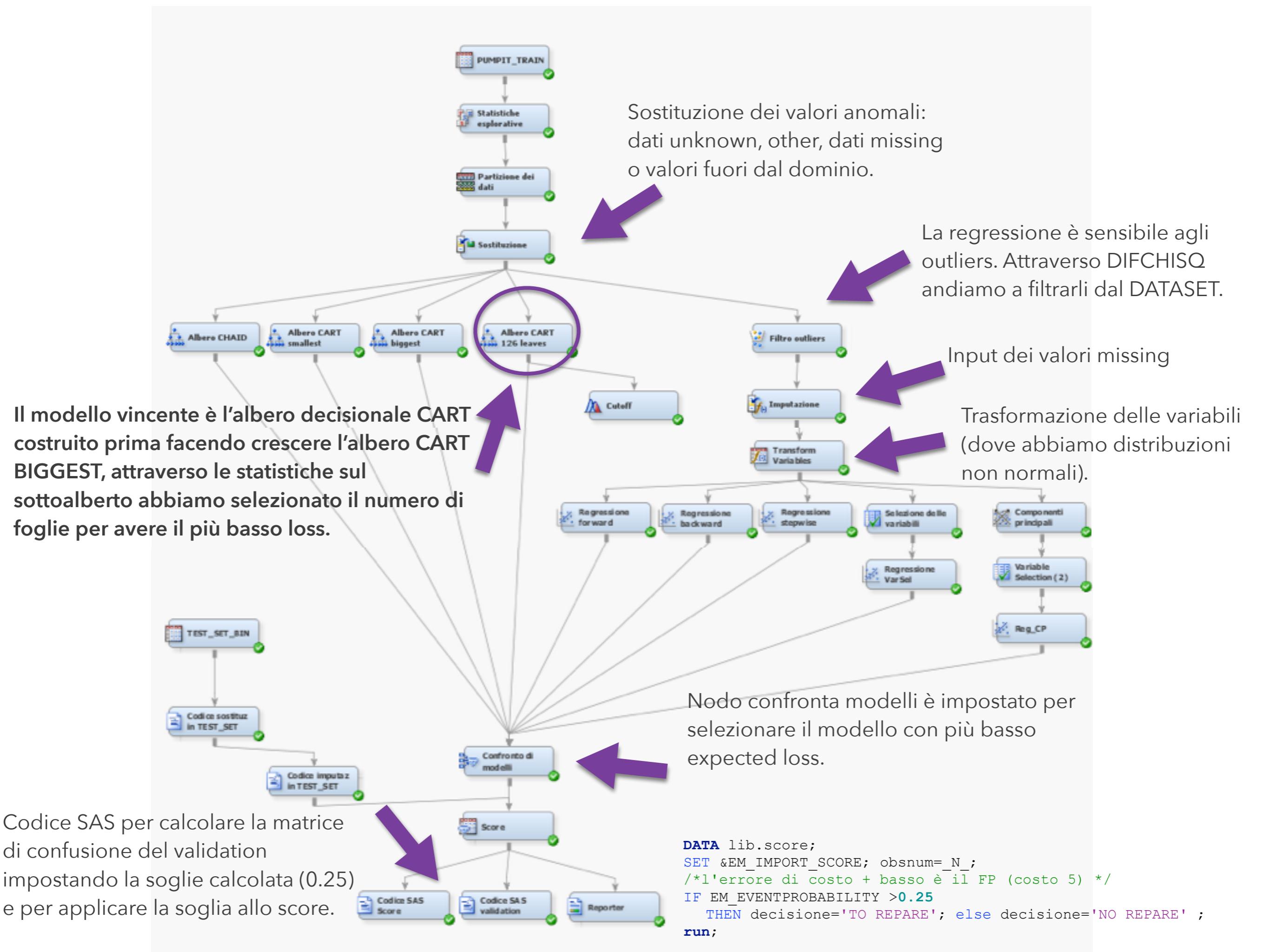


Diagramma SAS Miner

Covariata	Descrizione	Ruolo	Livello	Distribuzione	Missing values	Note
Amount_tsh	Amount water available	Input	Continuo	Non è una normale. Skewness = 32,48	Y	Nelle regressione andiamo ad usare la variabile trasformata con la funzione Log ed a imputare i dati mancanti o a 0
Gps_height	Altitude of the well	Input	Continuo	Normale	Y	
Longitude e latitude	GPS coordinate	Input	Continuo		N	Non contiene valori missing ma molti dati sono a 0. Andiamo a sostituire i dati imputando il valore attraverso la mediana
Basin	Geographic water basin	Input	Nominale		N	molto significativa
Region	Geographic location	Input	Nominale		N	
Population	Population around the well	Input	Continuo	Non è una normale. Skewness = 10,98	Y	Nelle regressione andiamo ad usare la variabile trasformata con la funzione Log ed a imputare i dati mancanti o a 0
Permit	If the waterpoint is permitted	Input	Binario		Y	
Construction_year	Year the waterpoint was constructed	Input	Continuo		Y	Circa 2000 osservazioni sono indicate con 0
Extraction_type	The kind of extraction the waterpoint uses	Input	Nominale		N	
Management	How the waterpoint is managed	Input	Nominale		N	
Payment_type	What the water costs	Input	Nominale		N	Meglio di payment, in quanto ha una dicitura più compatta
Water_quality	The quality of the water	Input	Nominale		N	
Quantity	The quantity of water	Input	Nominale		N	
Source	The source of the water	Input	Nominale		N	
waterpoint_type	The kind of waterpoint	Input	Nominale		N	



Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	41578	17822
MISC	Misclassification Rate	0.150176	0.159466
MAX	Maximum Absolute Error	0.996269	1
SSE	Sum of Squared Errors	8532.496	3772.075
ASE	Average Squared Error	0.102608	0.105826
RASE	Root Average Squared Error	0.320325	0.32531
DIV	Divisor for ASE	83156	35644
DFT	Total Degrees of Freedom	41578	
ALOSS	Average Loss for status_group...	1.022055	1.063012
LOSS	Total Loss for status_group_bin	42495	18945



PERFORMANCE VALIDATION

Frequency				
Percent				
Row Pct				
Col Pct	NO REPAR TO REPAR Total			
	E	E		
no repare	6811	2868	9679	
	38.22	16.09	54.31	
	70.37	29.63		
	95.69	26.79		
bad				
to repare	307	7836	8143	
	3.77	43.97	45.69	
	4.51	96.23		
		73.21		
Total	7118	10704	17822	
	39.94	60.06	100.00	

- Buone performance classificative. La sensitivity (TPR) è al **96.23 %** e abbiamo un FNR di **3.77%** (solo 307 pozzi vengono individuati come da non riparare quando in realtà lo sono).

Osservazioni

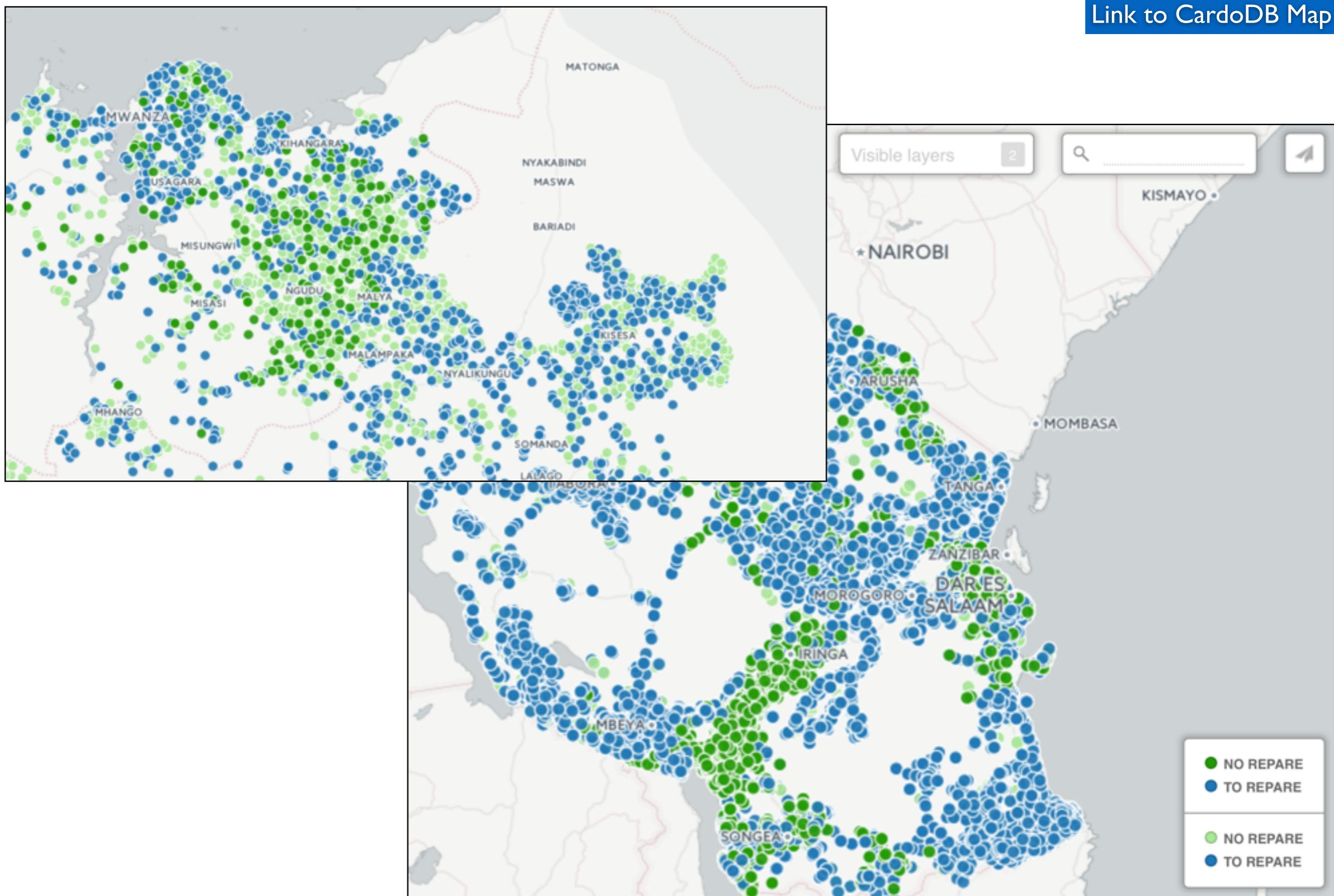
Predizione

	To Repare	No Repare
To Repare	0 TP	15 FN
No Repare	5 FP	0 TN

Average Expected Loss = 1,06
Total Expected Loss = 18.945

Presentazione della mappa finale

[Link to CardoDB Map](#)



EXTREME GRADIENT BOOSTING



H2O RANDOM FOREST



xgboost (eXtreme Gradient Boosting) è un'implementazione opensource in R/Python del modello gradient boosting di Friedman. L'idea principale è quella di migliorare il modello molte volte (nrounds è un parametro del modello) andando a "pesare" di più i soggetti misclassificati.

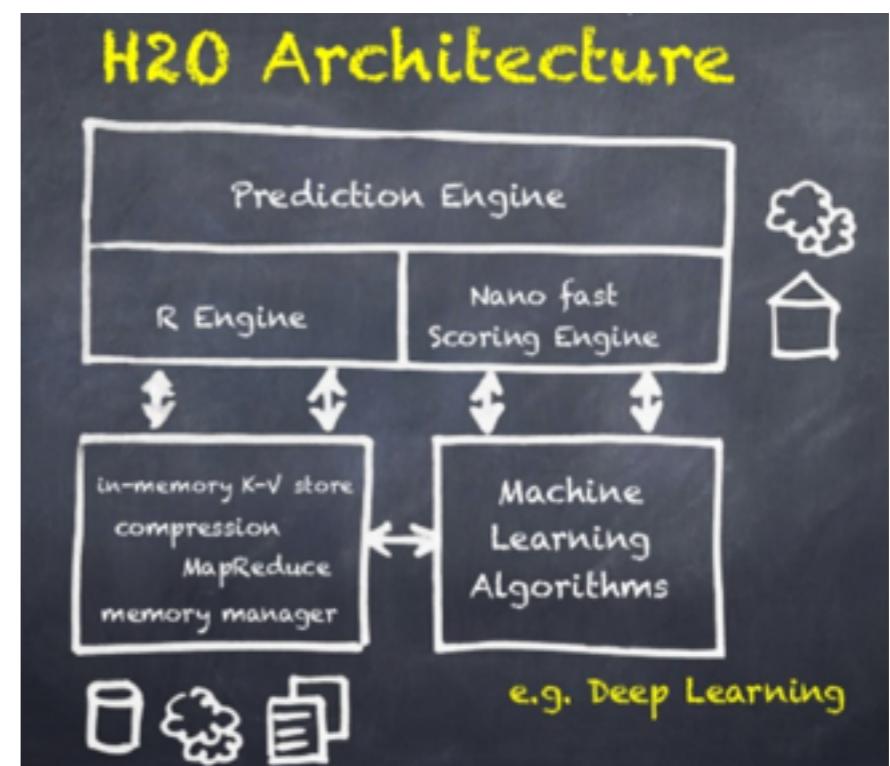
Parametri:

- eta = 0.05 **Step size shrinkage per prevenire overfitting**
- subsample = 0.7 **subsample ratio of the training instance**
- colsample_bytree = 0.7 **subsample ratio of columns**
- max_depth = 12 **maximum depth of a tree**

Reference			
Prediction	no repare	to repare	
no repare	7960	3990	
to repare	121	5695	
Accuracy : 0.7686			
95% CI : (0.7623, 0.7748)			
No Information Rate : 0.5451			
P-Value [Acc > NIR] : < 2.2e-16			

- Valutiamo solo le performance classificative: la **sensitivity** (TPR) è al **97,9 %** e abbiamo un **FNR** di **2,08%**

H2O è una libreria di ML open source, per calcolo distribuito implementata in JAVA.



> h2o.confusionMatrix(rfHex)

```
Confusion Matrix for max f1 @ threshold = 0.514849112774218:  
no repare to repare Error Rate  
no repare 21514 5537 0.204687 =5537/27051  
to repare 1729 30439 0.053749 =1729/32168  
Totals 23243 35976 0.122697 =7266/59219
```

- Valutiamo solo le performance classificative: la sensitivity (TPR) è al **94 %** e abbiamo un FNR di **5,3%**