Pump it Up: Aplicación de *Machine Learning* frente a crisis hídrica en Tanzania

Sergio A. Silva-Rubio

30 de Octubre de 2020

1. Introducción

La disponibilidad de agua potable y limpia es un problema de salud pública importante en las partes menos desarrolladas del mundo. Además, la infraestructura de suministro de agua todavía está compuesta en gran parte por bombas manuales, que requieren un mantenimiento periódico.

Un país que depende en gran medida de estas bombas es el país de Tanzania, en el este de África. Según su Ministerio de Agua, se pueden encontrar más de 74.000 bombas y su mantenimiento es responsabilidad de la comunidad local. Desafortunadamente, el costo de mantenimiento de las bombas son altos y las comunidades no son conscientes de la necesidad de realizar.

El año 2016, el Ministerio del Agua se asoció con Taarifa (www.taarifa.org) para patrocinar una competencia de desafío de datos para evaluar si es posible predecir el estado funcional de una bomba de agua en función de una variedad de indicadores cuantitativos y cualitativos.

El propósito de este estadio fue implementar un modelo de predictivo de clasificación basado en arboles de decisión y evaluar su rendimiento. Si los modelos son precisos, esto podría ayudar al gobierno de Tanzania a ahorrar mucho tiempo y dinero.

2. Metodología

Los datos para este proyecto se obtuvieron del sitio web DrivenData.org que alberga desafíos mediante el uso de datos. Específicamente, la información se puede encontrar a través del siguiente enlace web:

https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table

2.1. Análisis exploratorio de datos

El conjunto de datos se compone de atributos que describen un total de 59.400 bombas de agua de Tanzania. Cada bomba de agua está representada por un total de 39 atributos cualitativos y cuantitativos que se presentan en el Cuadro 1. A partir de esta información, se realizó una "limpieza" de los datos, debido a que muchas columnas contienen la misma información.

atributo	**descripción**	**atributo**	**descripción**
cantidad_tsh	altura estática total	scheme_name	quién opera el punto de agua
date_recorded	la fecha en que se ingresó	permit	si el punto de agua está permitido
funder	quién financió el pozo	construction_year	año en que se construyó
gps_height	altitud del pozo	extraction_type	que utiliza el punto de agua
installer	la organización que instaló el pozo	extraction_type_group	tipo de extracción que utiliza
longitude	coordenadas GPS	extraction_type_class	tipo de extracción que utiliza
latitude	coordenadas GPS	management	cómo se gestiona el punto de agua
wpt_name	nombre del punto de agua	management_group	cómo se gestiona el punto de agua
num_private	-	payment	lo que cuesta el agua
basin	cuenca de agua geográfica	payment_type	lo que cuesta el agua
subvillage	ubicación geográfica	water_quality	calidad del agua
region	ubicación geográfica	quality_group	calidad del agua
region_code	ubicación geográfica (codificada)	quantity	cantidad de agua
district_code	ubicación geográfica (codificada)	quantity_group	cantidad de agua
lga	ubicación geográfica	source	fuente del agua
ward	ubicación geográfica	source_type	fuente del agua
population	Población alrededor del pozo	source_class	fuente del agua
public_meeting	True/False	waterpoint_type	tipo de punto de agua
recorded_by	Grupo que ingresó datos	waterpoint_type_group	tipo de punto de agua
Scheme_management	Quién opera el punto de agua		

Cuadro 1: Atributos registrados de las bombas.

2.2. Distribución de los estados de las bombas de agua

Se complementa la información anterior con los datos del estado de las bombas a través del atributo *status_group*. Hay tres estados diferentes en la columna *status_group*: *functional*, *functional needs repair* y *non functional*.

```
> table(data$status_group)
```

functional	**functional needs repair**	**non functional**
32259	4317	22824

```
> prop.table(table(data$status_group))
```

functional	**functional needs repair**	**non functional**
0.54308081	0.07267677	0.38424242

2.3. Modelos predictivo de clasificación: Árbol de decisión

Un árbol de decisión es un modelo de predicción que dado un conjunto de datos se fabrican diagramas de construcciones lógicas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.

2.4. Arquitectura

1. La primera etapa consiste en realizar la limpieza de datos, debido a que en este caso, existían muchas columnas que tienen la misma información. Además, hay muchos valores nulos. Hay errores ortográficos en algunas columnas que crean muchas categorías.

2. La segunda etapa corresponde al entrenamiento de los datos, que nos permite encontrar relaciones, desarrollar comprensión, tomar decisiones y evaluar su confianza a partir de los datos de entrenamiento. Además, para este caso contamos con información que nos nos permite clasificar los datos mediante una etiqueta.

3. La tercera etapa consiste en seleccionar un algoritmo que reciba y analice los datos de entrada para predecir los valores de salida dentro de un rango aceptable.

4. A partir de la tercera etapa podemos crear nuestro modelo predictivo que a con nueva información podrá clasificar en alguna de las etiquetas previamente definidas.

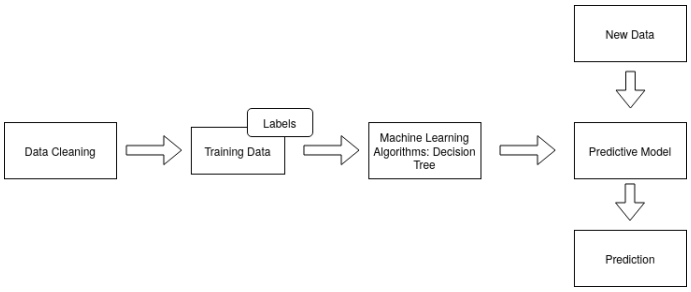

Figura 1: Arquitectura de experimentos.

3. Resultados

Los atrinbutos utilizados para prededecir el estado de la bomba fueron: **quantity, construction_year, region_code, region, gps_height, payment, extraction_type_class, source, waterpoint_type**. Con ello se encontro el arbol de decision que se muestra en la Figura 3. Donde cada nodo nos muestra la siguiente información:

- El primer valor es la clasificación que predomina en esa opción de los datos.
- El segundo valor son los porcentajes para las distintas clasificaciones definidas.
- El tercer valor es el porcentaje de los datos correspondientes a dicha opción.
- Finalmente, bajo cada nodo se indican los atributos considerados para cada opción.

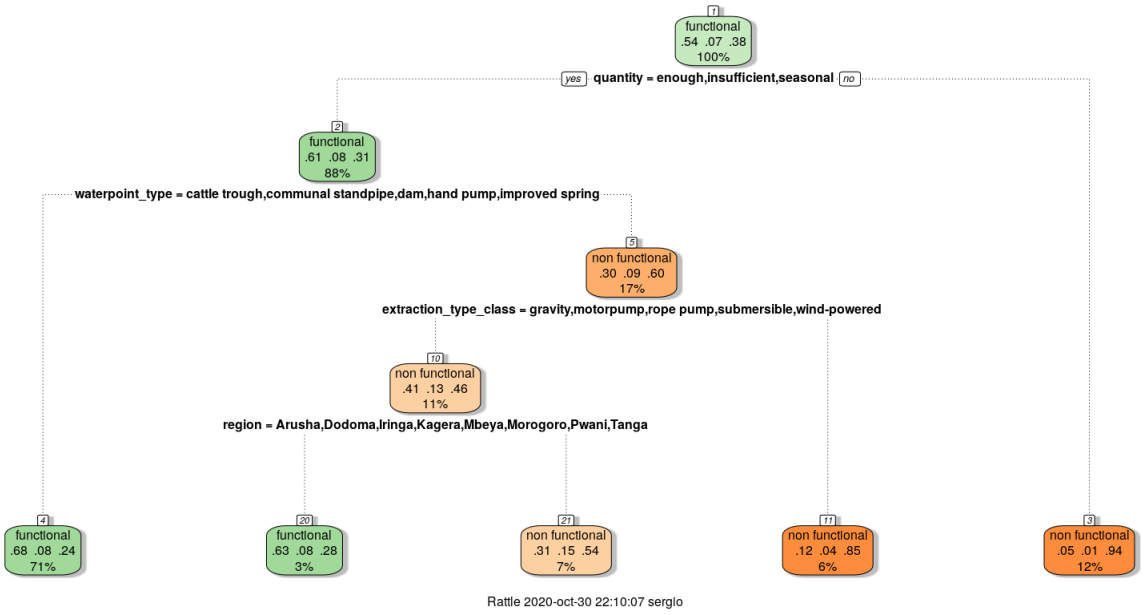

Figura 2: Estructura del Árbol de decisión.

A continuación se presentan los resultados obtenidos utilizando el conjunto de nuevos datos.

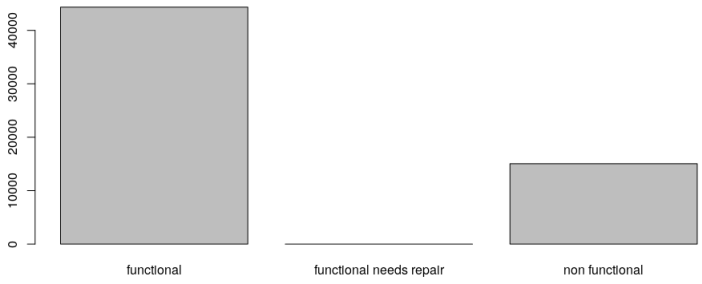

Figura 3: Resultados nuevos datos.

4. Conclusión

Como primer análisis se puede decir que el conjunto de datos era difícil de clasificar. Una de las razones fue la redundancia de información de los atributos. En cuanto a los resultados obtenidos solo se logro clasificar en dos categorías los nuevos datos, si bien no era lo esperado, el porcentaje de exito es consiste a nuestro conjunto de entrenamiento. Como trabajo futuro se sugiere la seleccion de los atributos de mejor forma y una categorización de ellos.