

Ingeniería de Sistemas y Computación
Escuela de Posgrado
MINE-4101: Ciencia de Datos Aplicada

Santiago Najar Gomez 202021647
Nestor Ivan Ramirez 201315220

Juan Diego Velásquez 201413344
Carlos Alberto Niño Ramirez 201516916

Proyecto Final - Primera Entrega

1.Problemática

La organización escogida es el ministerio de agua de Tanzania, este tiene como misión mejorar el acceso de agua potable segura y de garantizar servicios sanitarios para todos los ciudadanos. Esta entidad está encargada de administrar los recursos hídricos para asegurar la seguridad alimentaria nacional y un desarrollo económico sostenible.

Tanzania está ubicado en el suroriente africano, posee una población de 66.6 millones de personas, con una densidad de 67 habitantes por km² (). Cuenta con una superficie de 947.000 km² y 6.2% de ellos corresponden a cuerpos hídricos. Para ponerlo en perspectiva Colombia cuenta con 52.2 millones, una densidad de 46 personas por km², 1'140.000 km² de superficie de los cuales 8.8 son cuerpos hídricos, lo que significa que Tanzania tiene menos recursos hídricos y debe suplir de agua a una mayor cantidad de personas.

Para el 2015 en África subsahariana sólo el 61% de personas tenían acceso a una fuente de agua potable mejorada, 40% no tenían acceso a agua canalizada en el hogar y el 25% de las personas debían gastar 30 minutos para poder obtener agua de la fuente más cercana.

Para Tanzania específicamente el panorama es el mismo, el porcentaje de personas que tenían acceso a una fuente de agua potable mejorada es únicamente del 61%, este sin embargo varía de acuerdo a la zona geográfica, por ejemplo en las islas de Zanzíbar el porcentaje es mucho mayor, el 98%. Por otro lado, en el casco urbano el porcentaje llega a 86% y finalmente las zonas rurales son las más afectadas teniendo únicamente un 49% de cobertura.

La problemática a trabajar consiste en diseñar un modelo que identifique con alta precisión el estado de los pozos de agua (funcionales, funcionales pero requieren reparaciones o no funcionales) para tomar medidas proactivas de mantenimiento y reparación de estos y con ello garantizar el servicio de agua potable a todos los ciudadanos del país.

1.1.Objetivo y KPIs

El principal objetivo del proyecto es la creación de un producto de data que apoye al ministerio del Agua y a las entidades gubernamentales de Tanzania para mejorar el procedimiento de mantenimiento y reparación de los pozos de agua del país.

Los principales KPIs del ministerio son:

- Porcentaje de cobertura a nivel nacional del servicio de agua potable.
- Número de pozos reparados al mes (que previamente no funcionaran).
- Porcentaje de pozos en funcionamiento.

Adicionalmente ya que se cuenta con recursos limitados la principal métrica a tener en cuenta es la precisión del modelo para identificar los pozos que requieren atención, de esta forma optimizar al máximo los recursos disponibles.

2.Ideación

2.1.Identificación de Potenciales Usuarios

- Personal del Ministerio de Agua: Responsables de la gestión y mantenimiento de pozos.
- Ingenieros y Técnicos: Encargados de realizar inspecciones y reparaciones en los pozos.
- Comunidad Local: Beneficiarios del acceso a agua potable y afectados por la funcionalidad de los pozos.

2.2.Procesos Actuales y Dolores Relacionados

Procesos Actuales

- Monitoreo de pozos: Inspecciones manuales y reportes de estado.
- Mantenimiento y Reparaciones: Programación y ejecución de reparaciones basadas en informes.

Dolores Relacionados

- Fallos en el monitoreo: Dificultades para identificar pozos que necesitan atención de manera proactiva. Esto genera retraso en la toma de decisiones estratégicas al interior de la compañía que opera el pozo.
- Uso Ineficiente de Recursos: Asignación de recursos sin datos precisos que guíen las decisiones. Claramente se traduce en generación de sobre costos al no saber con exactitud el estado del pozo y no poder priorizar su mantenimiento en una temprana etapa.

2.3.Requerimientos del Producto de Datos

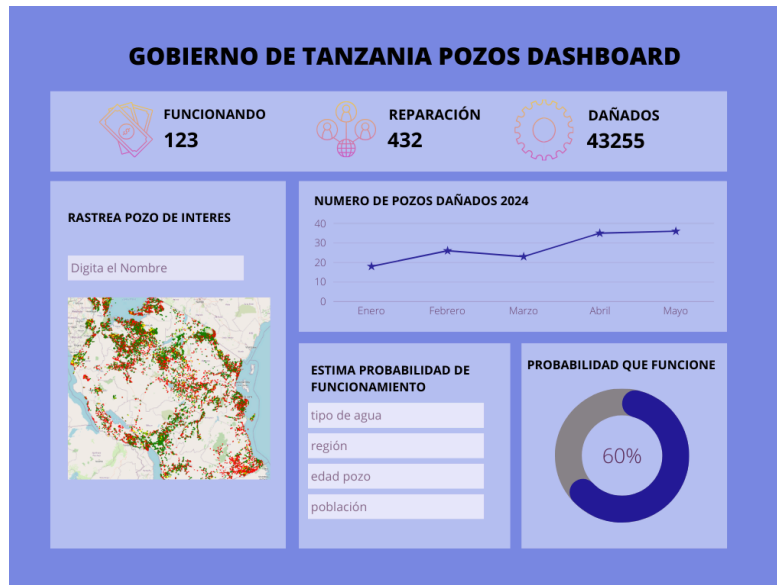
Funcionales

- **Carga de Datos:** Permitir la carga de un archivo Excel y actualizar automáticamente el dashboard con los datos procesados en tiempo real.
- **Visualización Interactiva:** Incluir un mapa interactivo y gráficos dinámicos que respondan rápidamente a la interacción del usuario.
- **Filtros Clave:** Incorporar al menos cinco filtros (Año, Estado del pozo, Calidad del agua, Fuente de agua y Región) para personalizar las visualizaciones.
- **Modelo Predictivo:** Integrar un modelo que prediga si un pozo es funcional, necesita reparación o no funcional.
- **Accesibilidad Web:** Desplegar la herramienta como una página web accesible de forma remota con tiempos de respuesta bajos.

No Funcionales

- Escalabilidad: Capacidad de manejar un aumento en el número de pozos y usuarios.
- Accesibilidad: Accesible en página web para que los técnicos puedan acceder a la información en el campo.

2.4.Mockup



Modelo de Machine Learning

Optaremos por un modelo de aprendizaje automático basado en árboles de decisión y sus variantes, ideales para calcular probabilidades del estado del pozo (funcional, necesita reparación, no funcional). Los árboles de decisión son una excelente opción debido a su capacidad para manejar datos categóricos y numéricos, además de su fácil interpretabilidad. Al segmentar las variables clave, estos modelos pueden identificar patrones que predicen fallos y ayudan a planificar el mantenimiento preventivo, optimizando recursos y tiempo.

API

Desarrollaremos una API REST que proporcionará los datos de predicción en formato JSON, permitiendo que el gobierno de Tanzania acceda a la información en tiempo real desde cualquier ubicación. Esta integración permitirá que los datos de mantenimiento de pozos se puedan consumir fácilmente y se integren con los sistemas existentes del gobierno, facilitando así su implementación y uso en otras plataformas que ya empleen. La API tomará como entrada los parámetros usados para entrenar el modelo y devolverá la predicción hecha por el mismo.

Dashboard

El dashboard ofrece estadísticas clave y permite a los ingenieros en Tanzania introducir variables específicas para predecir el estado de los pozos. Esto ayuda a reducir la necesidad de desplazamientos y a optimizar recursos económicos, permitiendo al gobierno priorizar pozos que presenten alta probabilidad de fallo. Además, el dashboard proporcionará una visión general para facilitar la toma de decisiones y estará integrado con la API para una comunicación fluida y actualizada con otros sistemas del gobierno.

3.Responsables

Para abordar los aspectos éticos, de privacidad, confidencialidad, transparencia y regulatorios en el contexto del reto de "Pump it Up: Data Mining the Water Table", es crucial tener en cuenta las siguientes consideraciones:

3.1.Implicaciones éticas:

- **Equidad en el acceso a los recursos de agua:** El modelo que prediga el estado de las bombas de agua debe garantizar que las poblaciones más vulnerables no sean desatendidas. La discriminación en la asignación de recursos, basada en factores socioeconómicos o geográficos, podría agravar las desigualdades en el acceso al agua potable. Es esencial garantizar que los resultados del modelo no perpetúen estas desigualdades, sino que promuevan una distribución justa de los recursos.
- **Impacto en las comunidades locales:** Las decisiones basadas en el modelo afectarán directamente a las comunidades que dependen de estas fuentes de agua. Por lo tanto, el modelo debe considerar no solo la funcionalidad técnica de las bombas, sino también el impacto social y económico de las reparaciones o sustituciones.

3.2.Transparencia:

- **Explicabilidad del modelo:** Las predicciones del modelo sobre el estado de las bombas deben ser comprensibles para los usuarios finales, como los encargados de mantenimiento o los funcionarios gubernamentales. Esto significa que el modelo debe ser interpretable, proporcionando explicaciones claras sobre las decisiones tomadas.
- **Documentación y acceso a las decisiones:** El proceso de desarrollo y las decisiones automatizadas deben estar debidamente documentadas, permitiendo auditorías independientes y garantizando la confianza de la comunidad en el uso del modelo.

3.3.Aspectos Regulatorios:

- **Cumplimiento con regulaciones locales:** Dado que el modelo trabaja con datos de infraestructuras de agua en Tanzania, debe cumplir con las leyes y regulaciones locales relacionadas con la protección de datos y la gestión de infraestructuras críticas. Es posible que se requiera la consulta con entidades regulatorias locales para garantizar que el modelo cumpla con las normativas vigentes en cuanto al uso de datos y la toma de decisiones automatizada.

3.4.Responsabilidad Social:

- **Impacto a largo plazo:** El uso de IA en la predicción del estado de las bombas de agua debe estar alineado con un enfoque de desarrollo sostenible, ayudando a las autoridades a tomar decisiones más informadas que benefician a las comunidades locales en el largo plazo.

4.Enfoque analítico

Debido a que el objetivo general del problema es predecir el estado de los pozos y que contamos con data tabular etiquetada, la exploración y limpieza de los datos se basará primero en la clase objetivo (faltantes, outliers, etc) y posteriormente en identificar variables relevantes para la clasificación. Para lograr esto utilizaremos análisis univariado buscando filtrar constantes, variables con muy poca varianza y variables únicas como nombres o identificadores. Estas variables se descartarán del análisis y el modelo pues no tienen incidencia en la variable objetivo.

Una vez realizada esta primera limpieza de los datos, iniciaremos la exploración de los datos para identificar correlaciones, tendencias y posibles variables derivadas de las existentes que puedan enriquecer el análisis, sobre estas variables seleccionadas tomaremos estrategias para reemplazar datos atípicos y vacíos para garantizar la calidad de los datos seleccionados. Las herramientas que usaremos para esto son indicadores como: % nulos o vacíos, # distintos, valores mínimo y máximo y técnicas de visualización de la distribución como Boxplots, violinplots, barcharts y scatterplots, y mapas geográficos.

Una vez realizado el análisis preliminar de los datos y la limpieza de los mismos desarrollaremos el modelo de clasificación, para esto iniciaremos entrenando modelos sencillos como k-nn, árboles y svm antes de métodos de ensamble como random forest o gradient boosting. Todos estos se entrenarán y evaluarán utilizando validación cruzada, el dataset contiene cerca de 59.000 registros por lo que habrá que encontrar un número de folds sensato para la complejidad de los algoritmos.

5.Recolección de datos

Los datos por trabajar en el proyecto provienen de los dashboards de puntos de agua de Taarifa, empresa tercerizada encargada de unificar los datos del Ministerio del Agua de Tanzania. El diccionario de este y sus comentarios se encuentra en el documento *Pumps_diccionary.xlsx*.

Adicional a esto se plantea para próximos avances el uso de información como el clima por región o el ingreso per cápita por región como variables que pudieran aportar al modelo final.

6.Entendimiento de los datos

Cómo se mencionó anteriormente la principal fuente de información proviene de datos registrados por el gobierno de Tanzania. Este dataset consta de 59.400 registros y de 40 campos, incluida la variable objetivo, de los cuales 32 son variables categóricas y 8 cuantitativas.

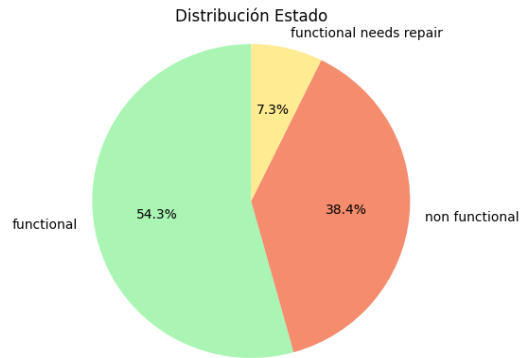
Se realizó el análisis de cada una de las variables y de las 40 iniciales, 20 fueron escogidas cómo útiles. En el diccionario de datos *Pumps_diccionary.xlsx* se pueden ver el detalle de las variables y la razón de por qué fueron descartadas.

6.1.Análisis Univariado

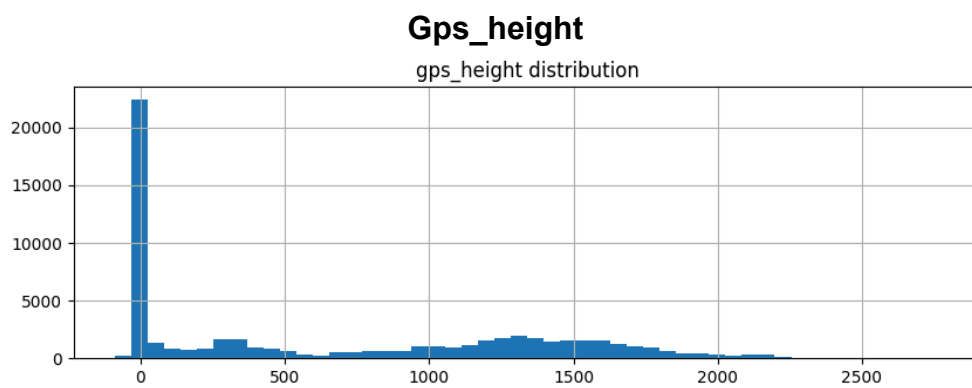
A continuación, explicamos los análisis univariados y multivariados más relevantes, si desea revisarlos, estos están disponibles en los notebooks del repositorio de github.

Status group (variable objetivo)

Es la variable objetivo del proyecto, cómo se menciona al inicio del reporte está distribuida en 3 categorías: funcional, funcional pero necesita reparaciones y no funcional.



Esta se encuentra desbalanceada ya que la categoría **funcional pero necesita reparaciones** solo representa el 7.3% del dataset.



Variable cuantitativa continua que representa la altura a la cual se ubica el pozo. Tiene como valor mínimo -90 y máximo 2770 metros. El 34% de los datos se encuentra con valor de 0.

6.2. Análisis Multivariado

Altura gps vs Estado

Se puede observar una clara tendencia que los pozos que no son funcionales tienen en promedio una menor altura respecto a los pozos funcionales. Esto debe deberse a una tercera variable no conocida relacionada con las anteriores como por ejemplo el clima de las zonas más bajas del país.

Población vs Estado

Al comparar la variable objetivo con la distribución de la población podemos observar que las distribuciones tienen colas largas, pero las poblaciones más altas, se encuentran en el grupo funcional, y las más pequeñas en el grupo no funcional, lo que indica que la cantidad de población está relacionada con el grupo funcional.

Esto también se evidencia en la agrupación de funcionales y no funcionales en el mapa donde existe un patrón de mayores pozos funcionales cerca a grandes centros poblacionales y por el contrario muchos no funcionales en zonas rurales o apartadas del país.

Año de construcción vs Estado

Al comparar el estado con el año de construcción de las bombas, hay una tendencia incremental del número de bombas en estado funcional con el paso del tiempo. Las bombas nuevas cumplen con normas de calidad más estrictas, requieren menos mantenimiento y están mejor protegidas contra la corrosión y la obstrucción, lo que las hace más confiables y menos propensas a fallos en comparación con las bombas más antiguas.

Cantidad de Agua vs Estado

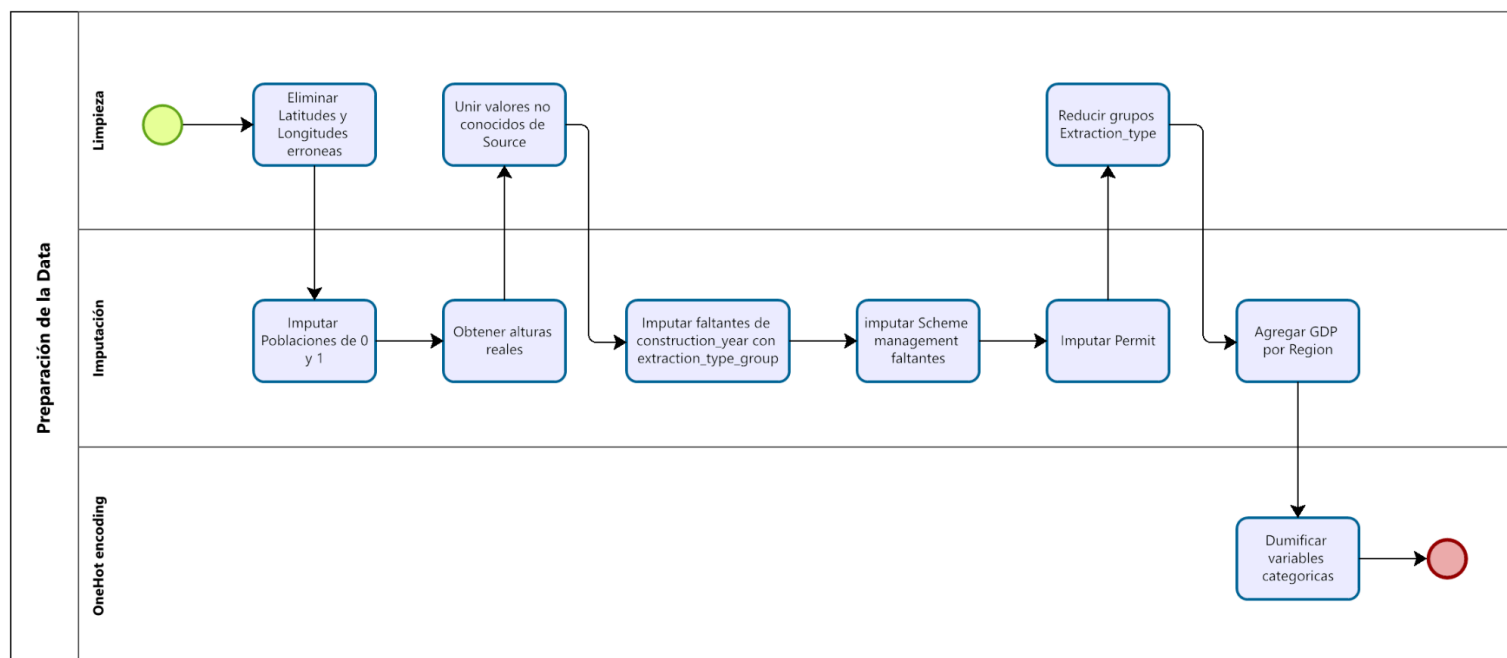
La tendencia muestra que las bombas tienden a sufrir daños cuando la fuente de agua se queda sin suministro. Esto ocurre porque las bombas requieren un flujo constante de agua para funcionar correctamente; al quedarse sin agua, experimentan fricción y sobrecalentamiento, lo que acelera su desgaste y puede llevar a fallos o daños graves.

7. Conclusiones iniciales

- Se va a desarrollar un modelo de clasificación robusto, como Random Forest o Gradient Boosting, para priorizar el mantenimiento preventivo de los pozos.
- Es fundamental priorizar la precisión del modelo para optimizar recursos, asegurando que los pozos identificados para reparación realmente lo necesiten, dada la importancia del transporte y la asignación de recursos en Tanzania.
- El análisis de cobertura de agua muestra gran variabilidad geográfica: 49% en zonas rurales, 86% en áreas urbanas y 98% en Zanzíbar, destacando la necesidad de un modelo que adapte patrones según la región para priorizar intervenciones.
- Los pozos funcionales tienden a estar ubicados a mayor altura, es necesario encontrar la razón por la cual esto sucede para tenerlo en cuenta en el modelo final.
- Los pozos en áreas de mayor densidad poblacional suelen estar en mejor estado, probablemente por una mayor asignación de recursos. Esto indica la necesidad de ajustar el modelo para considerar la densidad poblacional y priorizar áreas de baja densidad, donde hay mayor riesgo de fallos.
- Las bombas más antiguas tienen mayor probabilidad de ser no funcionales, lo que subraya la necesidad de actualizar infraestructuras.

Entrega 2

8.Preparación de datos:



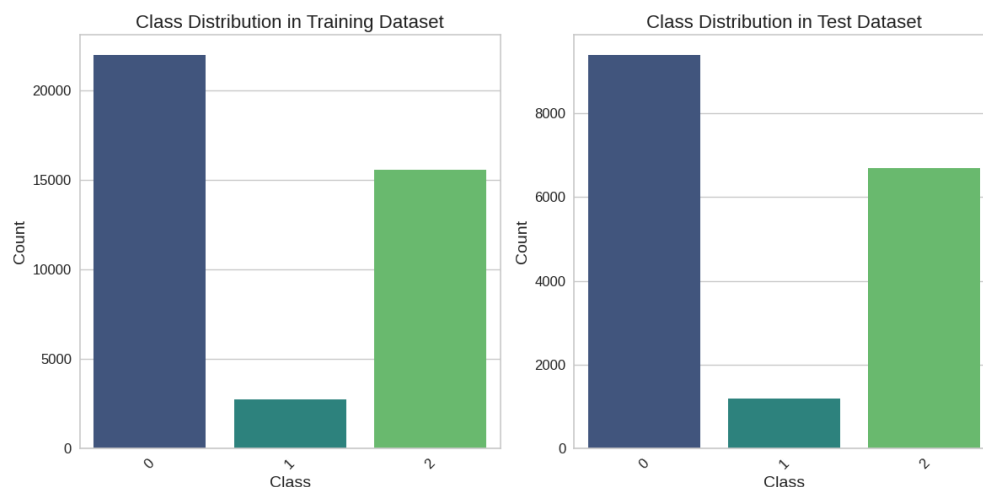
- **Latitud y Longitud:** se evaluó la veracidad de la latitud y longitud de todos los registros y se encontró que 1812 registros tenían longitudes fuera de Tanzania por lo que estos registros fueron eliminados ya que no hay forma de imputar estas variables.
- **Población:** el 46% de registros poseían valor de 0 o 1, cosa que evidentemente es un error, para imputar estos valores inicialmente intentamos usar el promedio de la población de los 3 pozos más cercanos (KNN). Sin embargo, nos dimos cuenta que la distribución geográfica de los pozos sin poblaciones se concentraba en ciertas regiones por lo que KNN no sería funcional.
Con esto en cuenta aplicamos dos métodos diferentes para imputar los datos separados de acuerdo al porcentaje de pozos con información de cada autoridad del gobierno local (lga). Si el porcentaje de faltantes era mayor al 60% se imputa la población del pozo como la población total del lga dividida por la cantidad de pozos en la misma, tomando como supuesto que la cantidad de personas se distribuye equitativamente en todos los pozos. Por otro lado cuando el lga tiene menos del 60% faltante es posible usar el método de KNN.
- **Altura:** La altura tiene un problema parecido al de población, el 34% de los datos tiene valor de cero, lo que evidentemente era un error. Para corregir esto se usó un api open-source con el cual basado en una latitud y longitud se obtiene la altura del punto.
- **Source:** El elemento source está granulado con variables no dicientes como “unknown” o “other” y al no tener un actor con quien podernos comunicar decidimos unir esta granularidad en una sola.
- **Construction_year:** al igual que la altura, el 33% de los valores estaban en cero. Para imputar este campo optamos por usar el extraction_type para asignarle el promedio de año

de construcción a los faltantes de cada clase, se toma cómo supuesto que los pozos con determinado tipo de extracción tienen años de construcción similares.

- **Scheme_management:** se utilizaron campos similares como management_group y management para con la moda de las combinaciones de estos campos se poble los valores faltantes.
- **Permit:** se optó por utilizar la moda de esta variable considerando los campos public_meeting y management_group para poblar los faltantes.
- **Extraction_type:** en este caso no hay valores faltantes pero se tiene un gran nivel de granularidad para disminuir esto se agrupan ciertas categorías en una sola.
- **GDP por Region:** obtuvimos esta información por fuentes externas.
- **OneHot encoding:** Para poder procesar correctamente los valores categóricos dentro de los modelos a usar realizamos la separación de cada variable categórica en sus respectivas columnas por cada posible valor.

9. Estrategia de validación y selección de modelo

Para entrenar el modelo utilizamos un split aleatorio del conjunto de datos, asignando el 30% de los datos para testeo. La distribución de la clase objetivo en los dos grupos es la siguiente:



Debido a que se muestrea de forma aleatoria se puede observar que se mantiene la distribución de las clases en los dos conjuntos de datos.

El conjunto de datos de entrenamiento fue agrupado en validación y entrenamiento utilizando cross validation con 5 folds, por lo que cada grupo posee 8000 registros. Adicionalmente, ejecutamos un Grid Search para encontrar la mejor combinación de hiper parámetros por modelo utilizando la métrica de precisión como indicador de desempeño.

La estrategia mencionada fue aplicada a cada uno de los modelos de ensamble seleccionados los cuales fueron: RandomForest, Light Gradient Boost y Extreme Gradient Boost.

10. Construcción y evaluación del modelo

Para entrenar el modelo utilizamos tres mallas de búsqueda de hiperparámetros con 5 folds de validación cruzada, las mallas utilizadas con cada uno de los algoritmos fueron las siguientes:

XGBoost

```
"n_estimators": [50, 100, 150],      "max_depth": [3, 5, 7],
"learning_rate": [0.01, 0.1, 0.2],    "subsample": [0.6, 0.8, 1.0],
"colsample_bytree": [0.6, 0.8, 1.0]
```

Random forest

```
"n_estimators": [50, 100, 150, 200],  "max_depth" : [3, 4, 5, 6],
'max_features': ['sqrt', 'log2'],      "class_weight": ["balanced", None]
```

LGBBoost

```
'num_leaves': [31, 50, 70],      'learning_rate': [0.01, 0.1, 0.2],
'n_estimators': [100, 200, 300]
```

Utilizando la precisión como métrica de desempeño obtuvimos los siguientes hiper parámetros:

XGBoost:

```
{'colsample_bytree':0.6,      'learning_rate':0.2,    'max_depth':7,
'n_estimators':150, 'subsample':1.0}
```

Random Forest:

```
{'class_weight':'balanced',    'max_depth':6,    'max_features':'sqrt',
'n_estimators':100}
```

LGBBoost:

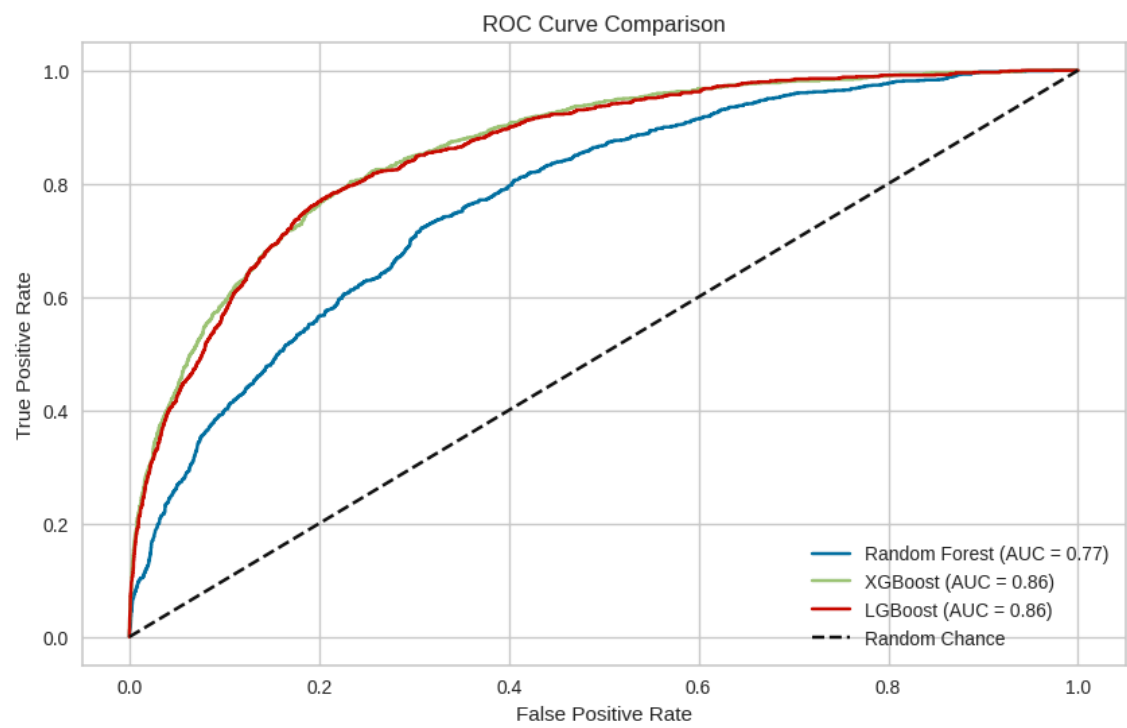
```
{'learning_rate': 0.1, 'n_estimators': 300, 'num_leaves': 70}
```

Evaluación cuantitativa

Las métricas de desempeño para los distintos modelos son las siguientes:

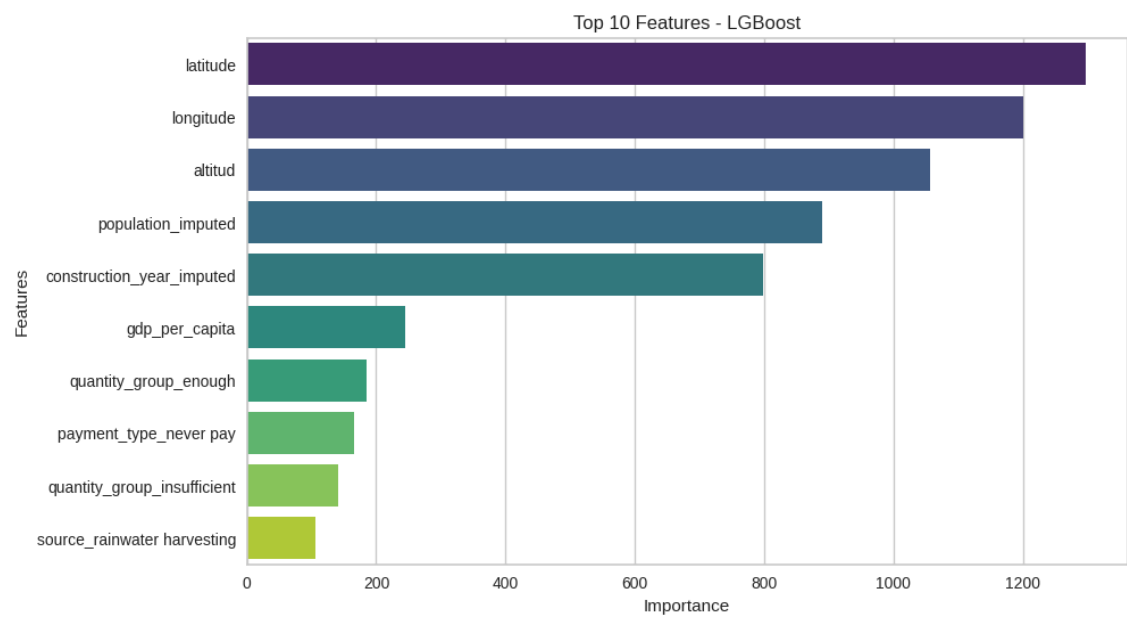
Metrica - Clase \ Modelo		RandomForest	LGBBoost	XGBoost
Precisión	Funcional	77	76	79
	Needs Repair	16	65	61
	No Funcional	73	84	83
f1-score	Funcional	69	83	83
	Needs Repair	25	27	31
	No Funcional	65	77	79

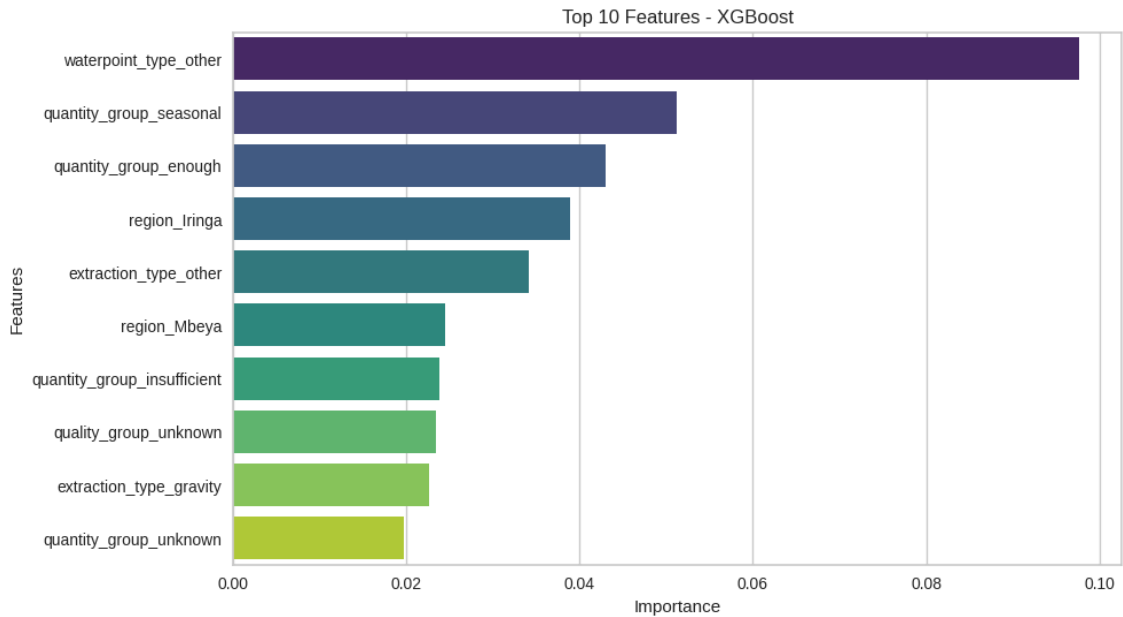
Las curvas ROC:



Evaluación cualitativa

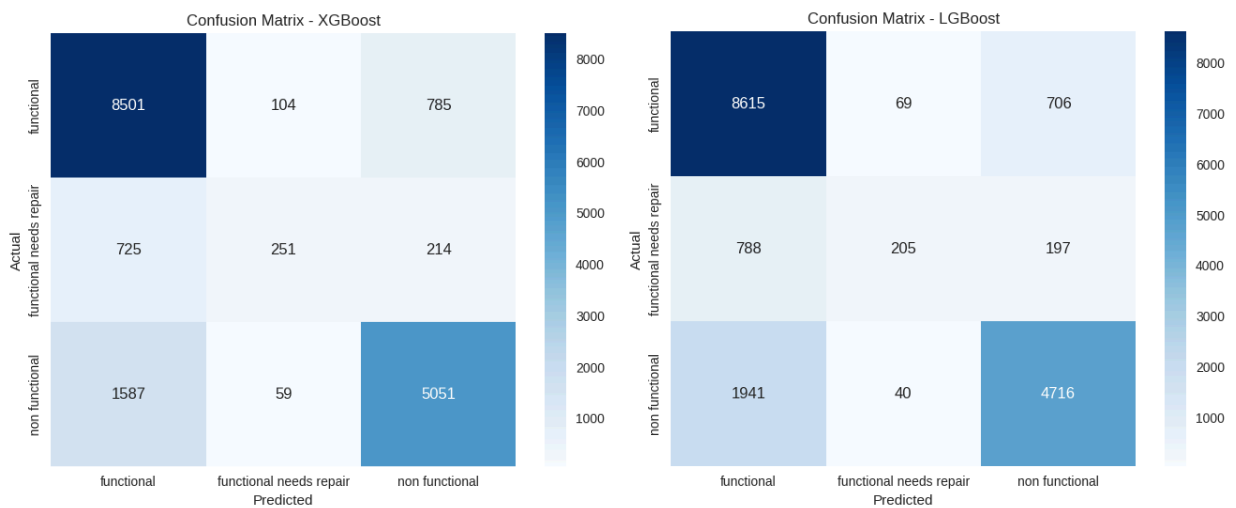
Feature importance de los dos mejores modelos:





Como se puede observar, el XGBoost utiliza en mayor medida las variables categóricas mientras que el LGBost utiliza las numéricas.

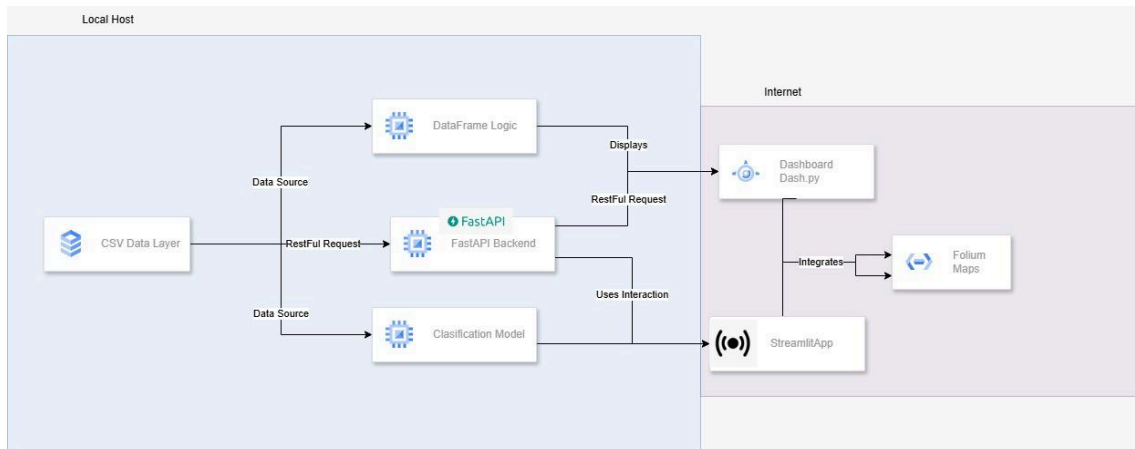
Matriz de confusión



Para identificar diferencias entre los modelos se realizó una prueba T, y el P-value fue de 0.88 por lo que no hay diferencias significativas entre las predicciones de los modelos, debido a esto realizaremos la elección del modelo por características cualitativas.

El XGBoost pese a tener una precisión ligeramente inferior en las clases distinta a la funcional, que es la métrica de análisis que nos interesa, confunde menos las no funcionales con las que necesitan reparación debido a que tiene mejor recall que se ve reflejado en el F-score. Por esta razón elegimos el XGBoost para el producto de datos.

11. Construcción del producto de datos



Este diagrama representa una arquitectura para desplegar una solución que combina un sistema local con interfaces accesibles desde Internet. En el entorno local, los datos en formato CSV son procesados por una lógica de datos (DataFrame Logic) y utilizados por un modelo de clasificación para realizar predicciones. Un backend basado en FastAPI centraliza el acceso a estos datos procesados y resultados de predicción, exponiéndolos mediante endpoints RESTful para que puedan ser utilizados por interfaces externas.

Desde Internet, los usuarios pueden interactuar con dos interfaces principales: un dashboard construido con Dash, que incluye visualizaciones avanzadas como mapas interactivos generados con Folium, y una aplicación web desarrollada en Streamlit para consultas dinámicas o en tiempo real. Esta arquitectura permite que toda la solución, desarrollada y probada localmente, sea posteriormente desplegada de forma pública para que usuarios remotos puedan consumir los datos y resultados, integrando análisis, predicciones y mapas de forma interactiva.

DashBoard



Este dashboard ofrece al gobierno de Tanzania una herramienta visual interactiva para analizar el estado histórico de los pozos de agua en el país. Muestra indicadores clave como el número de pozos funcionales, en reparación y no funcionales, junto con filtros dinámicos

para refinar los datos por año, estado del pozo, calidad y fuente del agua, región y rango de edad de los pozos. Incluye un mapa interactivo que localiza geográficamente los pozos y una gráfica de distribución que resume el estado de los mismos en porcentajes, ayudando a priorizar recursos y estrategias para la gestión hídrica en el país.

App Streamlit

Para la implementación del prototipo funcional, se desarrolló una API REST utilizando FastAPI. Esta API permite realizar predicciones sobre el estado de los pozos de agua al recibir datos como la altura, población, año de construcción, entre otros, y devolver una clasificación en tres categorías: funcional, funcional con reparaciones necesarias o no funcional. El diseño de la API se enfocó en la eficiencia y la simplicidad, proporcionando un esquema claro de endpoints para integrarse fácilmente con otras plataformas. Su despliegue local se realiza mediante el comando: `uvicorn api:app --reload --host 127.0.0.1 --port 8000`.

Por otro lado, se desarrolló una aplicación en Streamlit como interfaz predictiva. Esta herramienta ofrece una experiencia interactiva, donde los técnicos pueden introducir datos específicos del pozo directamente desde el campo y obtener predicciones inmediatas. La aplicación está diseñada para ser fácil de usar, con un enfoque en accesibilidad y portabilidad. El despliegue local de la aplicación se realiza mediante el comando: `streamlit run app.py`.

Predicción del estado de las bombas de agua

¿Cómo desea ingresar los datos?

☐ Un solo registro

☒ Múltiples registros (archivo CSV)

Subir archivo CSV

Cargue su archivo CSV

Drag and drop file here

Limit 200MB per file • CSV

Browse files

Bomba ID: 34169

	Estado	Probabilidad
0	funcional	13.37%
1	funcional needs repair	62.21%
2	non functional	24.42%

Estado Predicho: **funcional needs repair**

La arquitectura modular permite su despliegue en servidores locales o en la nube en etapas futuras, lo que garantiza la escalabilidad y la adaptabilidad a diferentes entornos operativos. Este enfoque técnico asegura que las soluciones sean eficientes, accesibles y listas para integrarse con los sistemas existentes del Ministerio del Agua de Tanzania.

12.Conclusiones

- Se obtuvo un producto de datos cuyo modelo, con una precisión del 84%, representa una herramienta importante para la optimización de recursos por parte de las entidades gubernamentales de Tanzania.
- El proceso de limpieza e imputación de los datos faltantes fue el proceso más arduo, ya que el Gobierno de Tanzania tiene poca información disponible y, además, se presentaban muchos valores erróneos.

- Se espera que con el uso de la herramienta el número de pozos reparados al mes aumente ya que se dispondrá de mejor forma los recursos disponibles para mantenimiento.
- Para obtener un modelo con mejores resultados sería ideal contar con datos actualizados, valores categóricos bien definidos y valores balanceados geográficamente. Igualmente pueden existir variables adicionales que aporten al modelo como comportamientos sociales o climáticos.

Referencias

- Programu. (n.d.). The United Republic of Tanzania Ministry of Water. Retrieved October 2, 2024, from <https://www.maji.go.tz/pages/programme>. Accessed 22 October 2024.
- Population. (2024, July 11). Our World in Data. <https://ourworldindata.org/grapher/population-unwpp?facet=none&country=TZA~COL>
- Tanzania Demographic and Health Survey and Malaria Indicator Survey (TDHSMIS), 2015–2016; SNV, WaterAid and UNICEF, School WASH Mapping in 16 districts, 2010; National Bureau of Statistics (NBS) et al., Tanzania Service Provision Assessment Survey 2014–2015, 2016; Benova et al., *Where there is no toilet*, 2014
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*.
- Floridi, L., & Taddeo, M. (2016). What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*.
- Open-Meteo. (n.d.). Elevation API. Recuperado el 13 de noviembre de 2024, de <https://open-meteo.com/en/docs/elevation-api>
- Statista. (s. f.). *Tanzania: Producto interno bruto (PIB) per cápita desde 1985 hasta 2024*. Statista. Recuperado el 15 de noviembre de 2024, de <https://www.statista.com/statistics/447608/gross-domestic-product-gdp-per-capita-in-tanzania/>