“Aplicando Machine Learning para determinar la calidad del agua subterránea en Lambayeque”

“Applying Machine Learning to determine the quality of groundwater in Lambayeque”

Elias Agapito Jose[1], Rodriguez Custodio Jhonatan[1], Sigueñas Ucañay Marcos[1]

[1]*Universidad Nacional Pedro Ruiz Gallo, Lambayeque, Perú*

Resumen

El departamento de Lambayeque enfrenta un desafío crítico en la gestión de sus aguas subterráneas, crucial para el suministro de agua potable debido al crecimiento poblacional y la expansión urbana. La sobreexplotación ha provocado la disminución del nivel freático, intrusión salina y degradación del agua. En este contexto, se aplicó un modelo LightGBM optimizado mediante Grid Search para predecir la calidad del agua subterránea.

Tras la optimización, el modelo mostró un rendimiento destacable, mejorando la métrica AUC-ROC al 85.33%. Esto indica una buena capacidad del modelo para distinguir entre agua de calidad aceptable e inaceptable. El LGBMClassifier se seleccionó como el mejor modelo, destacando su eficacia para manejar grandes conjuntos de datos y su eficiencia en el procesamiento.

El estudio proporciona un enfoque efectivo para la evaluación de la calidad del agua subterránea, con implicaciones significativas para la gestión y conservación de este recurso vital en Lambayeque.

Palabras clave: LightGBM, Grid Search, LGBMClassifier.

Abstract

The department of Lambayeque faces a critical challenge in the management of its groundwater, crucial for the supply of drinking water due to population growth and urban expansion. Overexploitation has caused a decrease in the water table, saline intrusion and water degradation. In this context, a Grid Search-optimized LightGBM model was applied to predict groundwater quality.

After optimization, the model showed remarkable performance, improving the AUC-ROC metric to 85.33%. This indicates a good ability of the model to distinguish between water of acceptable and unacceptable quality. The LGBMClassifier was selected as the best model, highlighting its effectiveness in handling large data sets and its processing efficiency.

The study provides an effective approach for the evaluation of groundwater quality, with significant implications for the management and conservation of this vital resource in Lambayeque.

Keywords: LightGBM, Grid Search, LGBMClassifier.

1. Introducción

En diversas regiones, incluyendo Lambayeque, la calidad del agua subterránea es una preocupación crítica debido a su importancia para el consumo humano y la agricultura. Esta calidad puede verse afectada por factores naturales y actividades humanas, subrayando la necesidad de un enfoque avanzado para su evaluación y gestión.

El uso de técnicas de Machine Learning (ML) se presenta como una herramienta innovadora para analizar y predecir la calidad del agua subterránea de manera eficiente, especialmente en regiones con complejos sistemas hidrogeológicos como Lambayeque. Este estudio se enfoca en evaluar modelos de ML para determinar la calidad del agua subterránea en la región, utilizando datos multidimensionales que abarcan una variedad de parámetros físicos, químicos y biológicos recopilados a lo largo del tiempo.

El objetivo principal es proporcionar a las autoridades locales, científicos y gestores del agua una herramienta avanzada para el monitoreo continuo y la predicción anticipada de la calidad del agua subterránea. Al comprender mejor los factores que afectan esta calidad, se facilita la implementación de estrategias de gestión hídrica sostenible y la toma de decisiones proactivas.

Al integrar tecnologías de vanguardia como el ML, se espera contribuir significativamente a la preservación de los recursos hídricos en Lambayeque, garantizando un suministro seguro y saludable de agua para las generaciones presentes y futuras.

2. Marco Teórico

2.1. Agua Subterránea

Según Ordoñez (2015), el agua subterránea se refiere a la porción de agua que se encuentra debajo de la superficie terrestre y que puede ser captada mediante perforaciones, túneles o galerías de drenaje. También incluye aquella que fluye de manera natural hacia la superficie a través de manantiales o se filtra hacia los cursos fluviales.

2.2. Autoridad Nacional del Agua (ANA)

Según el Ministerio de Agricultura y Riego (2017), la Autoridad Nacional del Agua es un organismo público encargado de la gestión integrada y sostenible de los recursos hídricos. Su misión incluye la planificación, regulación, supervisión y control de los recursos hídricos, así como la promoción de su uso eficiente y equitativo.

2.3. Índice de Calidad del Agua (ICA)

Según el ANA (2028), el Índice de Calidad del Agua (ICA) es una herramienta que se utiliza para evaluar y resumir la calidad general del agua de un cuerpo de agua determinado, como ríos, lagos o embalses. Este índice se calcula mediante la medición y evaluación de diversos parámetros físicos, químicos y biológicos del agua.
El ICA proporciona una puntuación o clasificación numérica que refleja la calidad general del agua y permite comparaciones a lo largo del tiempo y entre diferentes ubicaciones.

2.4. Parámetros de Calidad del Agua

Según el ANA (2028), la presencia de ciertas sustancias en el agua puede influir en su calidad, especialmente cuando alcanzan concentraciones que podrían resultar perjudiciales para los organismos (ya sean humanos, plantas o animales) o cuando superan los estándares establecidos para la calidad ambiental.

A continuación, se detallan los parámetros (elementos o compuestos) que figuran en las regulaciones ambientales y que se toman en cuenta al evaluar la calidad del agua.

N°	Parámetro	Unidades
01	Oxígeno disuelto (valor mínimo)	mg/L
02	Demanda Bioquímica de Oxígeno (DBO_5)	mg/L
03	Arsénico	mg/L
04	Cadmio	mg/L
05	Cobre	mg/L
06	Cromo Total	mg/L
07	Hierro	mg/L
08	Manganeso	mg/L
09	Plomo	mg/L
10	Mercurio	mg/L
11	Zinc	mg/L
12	Potencial de Hidrógeno (pH)	Unid. de pH
13	Coliformes Termotolerantes (44,5 °C)	NMP/100 ml

Cuadro de Parámetros considerados en la Categoría 1-A2 Poblacional y Recreacional

2.5. Machine Learning

El machine learning o aprendizaje automático se define como un sistema de análisis computacional que utiliza algoritmos capaces de aprender y mejorar a partir de los resultados obtenidos. Los procesos de ML tienen la capacidad de crear modelos matemáticos analíticos altamente precisos mediante dos fases esenciales: la fase de entrenamiento (training), en la que los algoritmos procesan un conjunto inicial de datos que incluye una o más variables conocidas como inputs, para identificar posibles interacciones y hacer predicciones sobre los valores de una variable de interés llamada output; y la fase de comprobación (testing), donde se utiliza un segundo conjunto de datos para validar el modelo generado, basándose en las interacciones establecidas durante la fase de entrenamiento (Mitchell, 1997).

2.6. Árboles de Decisión (Decision Tress)

Los Árboles de Decisión, según Breiman (1984), son modelos de Machine Learning utilizados para clasificación y regresión, perteneciendo al grupo de algoritmos supervisados. Se destacan por su eficacia en la toma de decisiones basada en reglas.

En un Árbol de Decisión, cada nodo interno representa una decisión basada en una característica específica, y las ramas salientes conducen a otros nodos o a nodos hoja que contienen la salida deseada. Estos nodos hoja representan las clases o valores de salida finales.

La construcción de un Árbol de Decisión implica dividir recursivamente el conjunto de datos en subconjuntos más pequeños, optimizando la separación de clases o la predicción de valores de salida en cada paso mediante la evaluación de diferentes características y la determinación de la mejor forma de dividir los datos.

2.7. Bosque Aleatorio (Random Forest)

El Bosque Aleatorio (Random Forest), según Louppe (2014), es un modelo de Machine Learning que pertenece a la categoría de ensambles, una técnica que combina múltiples modelos para mejorar la precisión y generalización del sistema. Este enfoque, aplicado principalmente a problemas de clasificación y regresión, se destaca por su efectividad.

La idea central del Bosque Aleatorio es construir varios Árboles de Decisión durante el entrenamiento y combinar sus resultados para obtener una predicción más robusta y generalizada. Para ello, se crea una colección de árboles, cada uno entrenado con un subconjunto aleatorio y diferente de los datos de entrenamiento. Durante la construcción de cada árbol, se introduce aleatoriedad en la selección de características a considerar en cada división de nodo, lo que ayuda a diversificar los árboles y reducir la correlación entre ello.

2.8. Gradient Boosting

Según Friedman (2002), Gradient Boosting es una técnica de aprendizaje automático que pertenece a los algoritmos de ensamble, específicamente a la familia de métodos de boosting. Destaca como un enfoque poderoso para la construcción de modelos predictivos, tanto para problemas de regresión como para clasificación.

El principio fundamental del Gradient Boosting es la construcción iterativa de modelos débiles y la corrección de los errores de los modelos anteriores. A diferencia de otros métodos de ensamble, como el Bosque Aleatorio, donde se construyen múltiples árboles de decisión de manera independiente, en Gradient Boosting cada nuevo modelo se enfoca en corregir las deficiencias del conjunto actual.

El proceso comienza con la construcción de un modelo inicial, generalmente un modelo simple como un árbol de decisión poco profundo. Luego, se evalúan los errores del modelo inicial y se construye un segundo modelo que se centra en corregir esos errores. Este proceso se repite iterativamente, con cada nuevo modelo enfocándose en las instancias que fueron mal clasificadas o predichas incorrectamente.

Una de las implementaciones más conocidas de Gradient Boosting es el algoritmo Gradient Boosting Machine (GBM). Además, existen extensiones y mejoras como XGBoost (Extreme Gradient Boosting), LightGBM y CatBoost, que optimizan y aceleran el proceso de Gradient Boosting.

2.9. Naive Bayes

Naive Bayes es un algoritmo de clasificación basado en el teorema de Bayes, utilizado en aprendizaje supervisado para la clasificación y, en ocasiones, para clasificación probabilística. Su base teórica se basa en el teorema de Bayes, que calcula la probabilidad condicional de una instancia perteneciente a una clase dada su conjunto de características. El término "naive" (ingenuo) se refiere a la suposición de independencia condicional entre las características, aunque esta suposición puede no ser realista. Naive Bayes es eficiente y fácil de implementar, siendo especialmente útil en conjuntos de datos de alta dimensionalidad y efectivo en problemas de clasificación de texto como spam o categorización de documentos.

2.10. Regresión Logística

Según Müller y Guido (2016), la Regresión Logística es un modelo de aprendizaje supervisado utilizado para la clasificación binaria y, mediante extensiones, para problemas de clasificación multiclase. Destaca como un enfoque para predecir la probabilidad de que una instancia pertenezca a una clase particular.

A pesar de su nombre, la Regresión Logística se emplea comúnmente para problemas de clasificación. Su relación con la regresión se debe a que utiliza una función logística para modelar la

probabilidad condicional de que una instancia pertenezca a la clase positiva.
Para la clasificación multiclase, se pueden utilizar extensiones de la Regresión Logística, como la Regresión Logística Multinomial (también conocida como Softmax Regression), que maneja más de dos clases.

3. Resultados

3.1. Problema de Negocio

"Gestión Sostenible de Cuencas y Pozos Acuíferos en Lambayeque"

Para el departamento de Lambayeque en gran crecimiento que depende en gran medida de las aguas subterráneas para satisfacer las necesidades de agua potable. Sin embargo, debido al aumento de la población y la expansión urbana, la sobreexplotación de los recursos hídricos subterráneos se ha convertido en un problema crítico. Esto ha llevado a la disminución del nivel freático, la intrusión salina en acuíferos y la degradación de la calidad del agua.
El desafío radica en encontrar soluciones para garantizar una gestión sostenible de las aguas subterráneas, equilibrando la demanda creciente con la necesidad de conservar y proteger este recurso vital.

3.2. Datos

Los datos están en formato xlsx en donde cada línea representa un atributo:

- **CLAVE:** Identificador único para una ubicación o muestra específica.
- **SITIO:** Indica el lugar donde se recogió la muestra.
- **ESTADO:** Indica el estado en el que se encuentra el sitio.
- **MUNICIPIO:** Especifica el municipio de la ubicación del sitio.
- **ACUÍFERO:** Se refiere al acuífero asociado o cercano al sitio.
- **ORGANISMO_DE_CUENCA:** Se refiere a la organización de la cuenca asociada al sitio.
- **SUBTIPO:** Proporciona información adicional de clasificación o subtipo sobre el tipo de agua subterránea donde se recogió la muestra.
- **LONGITUD/LATITUD:** Coordenadas geográficas que especifican la ubicación exacta del sitio/muestra.
- **PH:** Mide la acidez/alcalinidad del agua en una escala de 0-14, donde 7 es neutro, <7 ácido, >7 alcalino.
- **Hardness:** Se mide en miligramos por litro (mg/L), que indica la concentración de iones de calcio y magnesio en el agua.
- **Solids:** Se mide en miligramos por litro (mg/L) y representa la cantidad total de partículas disueltas en el agua, incluyendo minerales, sales, metales, etc.
- **Sulfate:** Se mide en miligramos por litro (mg/L) es un anión común en el agua que puede influir en su sabor y puede tener efectos adversos en la salud humana en concentraciones elevadas.
- **Conductivity:** Se mide en microsiemens por centímetro (µS/cm) y representa la capacidad del agua para conducir la corriente eléctrica.
- **Organic_carbon:** Se mide en miligramos por litro (mg/L) y refleja la cantidad de carbono orgánico presente en el agua.
- **Trihalomethanes:** Se mide en microgramos por litro (µg/L) y

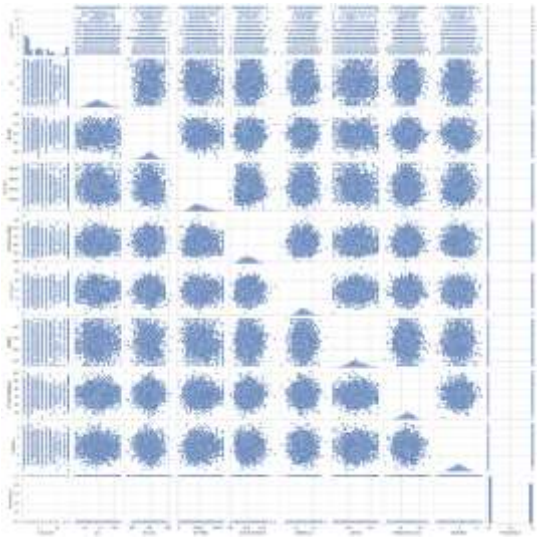
representa la concentración de trihalometanos en el agua.

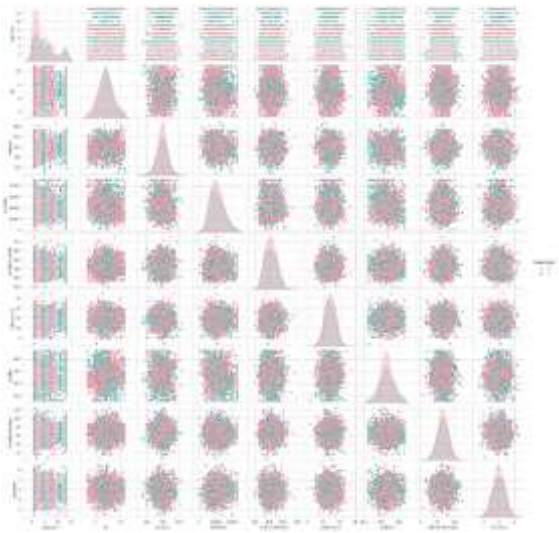
- **Turbidity:** Mide la turbidez o claridad del agua en unidades de nefelometría turbia (NTU).
- **Potability:** Indica si el agua es potable o no (1: potable, 0: no potable).

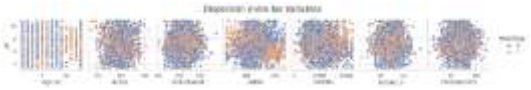
Los datos que te ha enviado la compañía están en el archivo Aguas-Subterraneas.xlsx.

3.3. Análisis Multivariado - Correlación entre Variables

3.3.1. Gráfico de Dispersión







- variables son dispersas y no muestran una clara asociación entre ellas, lo que sugiere una correlación casi nula.
- Además, se nota que hay más datos que indican que el agua no es potable en comparación con los que indican que sí lo es.

3.3.2. Matriz de Correlación

	Potabilidad
Potabilidad	1.000000
orgCarb	0.144455
sulfToDis	0.042080
ph	0.009819
turbidez	0.009758
trihalometanos	0.000519
conductividad	-0.011981
dureza	-0.018750
sulfato	-0.021077
carbono_O	-0.023892

3.3.3.VIF para detección de multicolinealidad

	Variable	VIF
0	orgCum	2.540572
1	ph	22.812613
2	dureza	31.300665
3	soliToDis	7.335332
4	conductividad	25.771869
5	carbono_O	18.610829
6	sulfato	45.678097
7	trihalometanos	16.440169
8	turbidez	22.847729

Los sulfatos son elementos naturales presentes en minerales, suelos y rocas. Son importantes para la calidad del agua subterránea, ya que su presencia puede indicar la presencia de contaminantes o minerales que afectan su utilidad para el consumo humano y otros usos. Por lo tanto, monitorear los niveles de sulfatos es esencial para garantizar la seguridad del agua subterránea.

Conclusión

Entendiendo la idea del negocio con mis variables predictoras y teniendo en cuenta que se realizó anteriormente un Feature Importances y eliminado previamente la variable subtipo entendemos que las demás variables se consideran importantes para nuestro modelo.

3.4. Construcción de los Modelos

Estandarización de datos

```
scaler = RobustScaler()
scaler.fit(X_train)
```

```
RobustScaler()
```

```
X_train_scaled = scaler.transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

```
pd.DataFrame(X_train_scaled).head()
```

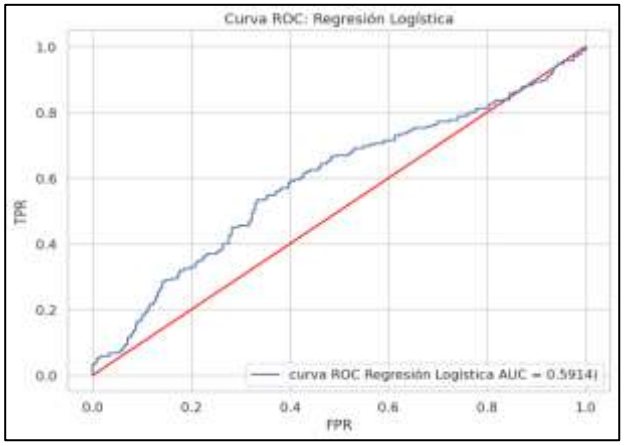
	0	1	2	3	4	5	6	7	8
0	0.6	0.912899	0.736651	0.240923	0.219868	0.003171	1.323536	-1.114366	-0.161689
1	1.4	0.453482	0.200947	-0.697900	0.238752	0.300349	-0.016346	-0.518702	0.562052
2	1.8	-0.476167	-0.095019	0.184818	-1.075261	-0.190427	0.166179	0.564982	-0.300428
3	-0.6	-0.476238	-1.474411	0.055753	0.222560	-0.018936	0.135308	-1.639122	-1.301440
4	-0.6	0.098072	-0.194956	0.046045	-0.272825	0.014520	0.492132	0.695341	0.419643

```
pd.DataFrame(X_test_scaled).head()
```

	0	1	2	3	4	5	6	7	8
0	0.2	-0.149868	0.344764	0.826292	0.613336	1.265573	-0.109982	-0.248404	-0.971943
1	-0.6	0.597559	0.199468	-0.295640	-0.649568	-0.407261	-1.097960	0.829343	1.125682
2	1.8	1.302034	0.095513	-0.069589	-0.744183	0.516510	-0.557338	0.233264	-0.867321
3	0.8	-0.698600	1.134091	0.946529	0.637158	1.886886	-2.024664	0.238936	1.874328
4	0.0	0.000000	0.835560	0.136556	-0.283628	0.262758	-0.217409	-1.519021	-0.084692

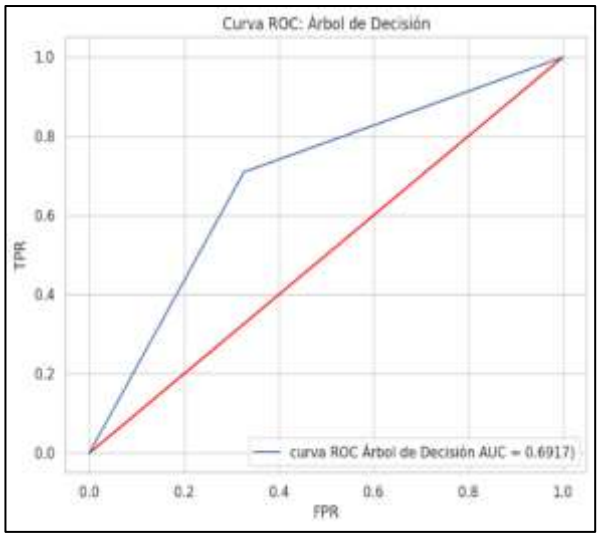
3.4.1. Regresión Logística

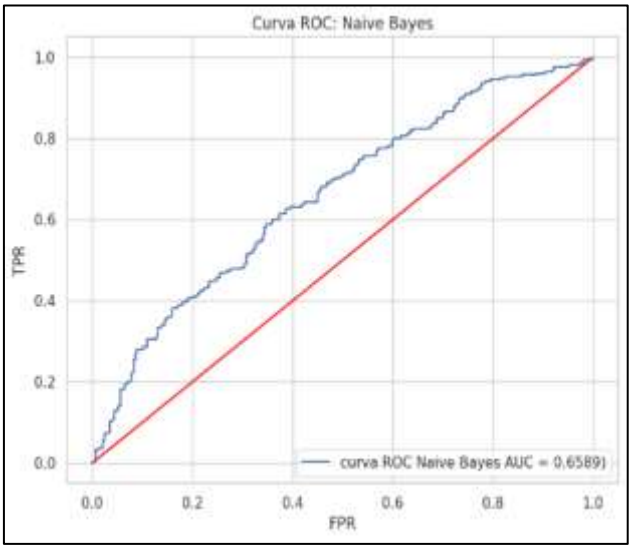
	No_Potable	Potable	accuracy	macro avg	weighted avg
precision	0.613909	0.537143	0.591216	0.575526	0.580842
recall	0.759644	0.368627	0.591216	0.564136	0.591216
f1-score	0.679045	0.437209	0.591216	0.558127	0.574876
support	337.000000	255.000000	0.591216	592.000000	592.000000



3.4.2. Árbol de Decisión

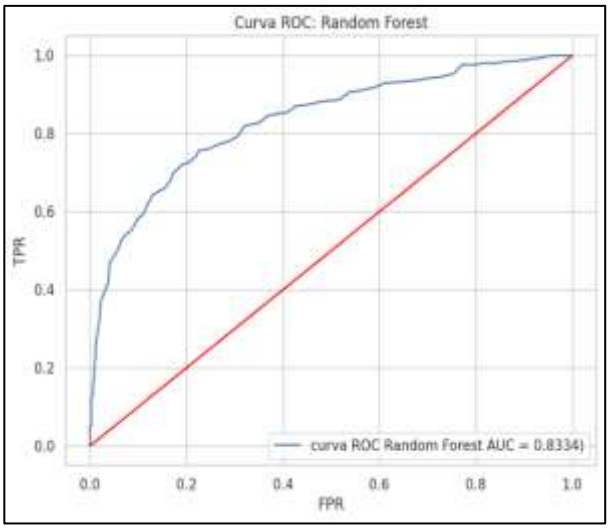
	No_Potable	Potable	accuracy	macro avg	weighted avg
precision	0.754153	0.621993	0.689189	0.688073	0.697226
recall	0.673591	0.709804	0.689189	0.691697	0.689189
f1-score	0.711599	0.663004	0.689189	0.687301	0.690667
support	337.000000	255.000000	0.689189	592.000000	592.000000





3.4.3. Random Forest

	No_Potable	Potable	accuracy	macro avg	weighted avg
precision	0.798220	0.733333	0.770270	0.765776	0.770270
recall	0.798220	0.733333	0.770270	0.765776	0.770270
f1-score	0.798220	0.733333	0.770270	0.765776	0.770270
support	337.000000	255.000000	0.770270	592.000000	592.000000

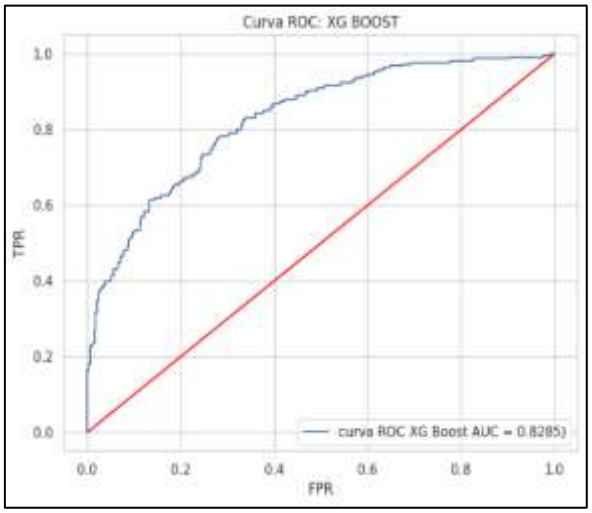


3.4.4. Naive Bayes

	No_Potable	Potable	accuracy	macro avg	weighted avg
precision	0.646617	0.590674	0.628378	0.618645	0.622519
recall	0.765579	0.447059	0.628378	0.606319	0.628378
f1-score	0.701087	0.508929	0.628378	0.605008	0.618316
support	337.000000	255.000000	0.628378	592.000000	592.000000

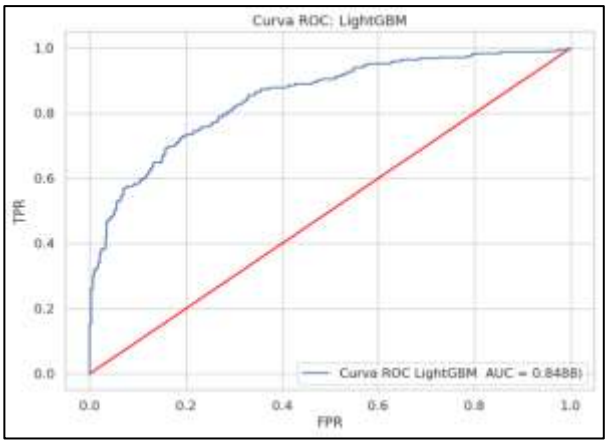
3.4.5. XG Boost

	No_Potable	Potable	accuracy	macro avg	weighted avg
precision	0.768072	0.684615	0.731419	0.726344	0.732124
recall	0.756677	0.698039	0.731419	0.727358	0.731419
f1-score	0.762332	0.691262	0.731419	0.726797	0.731719
support	337.000000	255.000000	0.731419	592.000000	592.000000

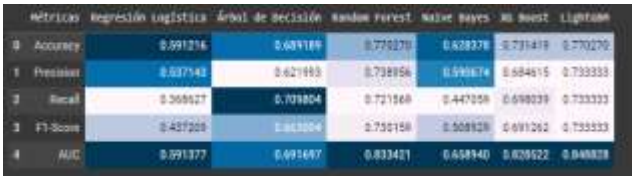


3.4.6. LightGBM

	No_Potable	Potable	accuracy	macro avg	weighted avg
precision	0.793003	0.738956	0.770270	0.765979	0.769722
recall	0.807122	0.721569	0.770270	0.764345	0.770270
f1-score	0.800000	0.730159	0.770270	0.765079	0.769916
support	337.000000	255.000000	0.770270	592.000000	592.000000

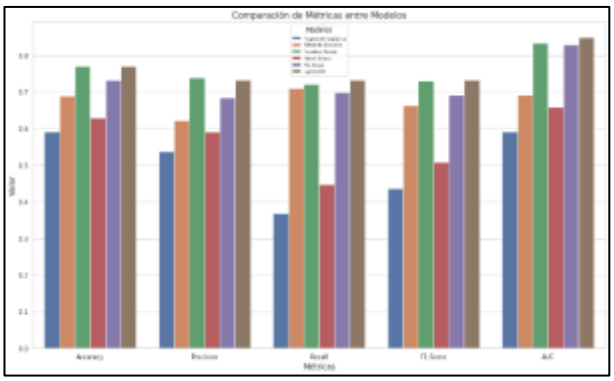


3.5. Agrupando Métricas - Comparando los modelos



	Métricas	variable	value
0	Accuracy	Regresión Logística	0.591216
1	Precision	Regresión Logística	0.537143
2	Recall	Regresión Logística	0.368627
3	F1-Score	Regresión Logística	0.437209
4	AUC	Regresión Logística	0.591377
5	Accuracy	Árbol de Decisión	0.689189
6	Precision	Árbol de Decisión	0.621993
7	Recall	Árbol de Decisión	0.709804
8	F1-Score	Árbol de Decisión	0.663004
9	AUC	Árbol de Decisión	0.691697
10	Accuracy	Random Forest	0.770270
11	Precision	Random Forest	0.738956
12	Recall	Random Forest	0.721569
13	F1-Score	Random Forest	0.730159
14	AUC	Random Forest	0.833421
15	Accuracy	Naive Bayes	0.628378
16	Precision	Naive Bayes	0.590674
17	Recall	Naive Bayes	0.447059
18	F1-Score	Naive Bayes	0.508929
19	AUC	Naive Bayes	0.658940
20	Accuracy	XG Boost	0.731419

3.6. Conclusiones de la Comparación



- Realizamos la comparación de nuestros 6 modelos, La regresión Logistica, El árbol de Decisión, Random Forest , Naive Bayes, XG Boost, LightGBM. Aplicamos varias métricas para cada uno de nuestros modelos, entre las cuales tenemos: Accuracy, Precision, Recall, F1Score y AUC la cual nos vamos a centrar en nuestra metrica AUC porque mis datos se encuentran desbalanceados.
- Es importante tener en cuenta que estamos trabajando con 2 clases: 0-agua no potable y 1-agua potable. Debido a esto separamos nuestro cuadro de clasificación de métricas para estas dos clases, a fin de visualizar de mejor manera los valores para las clases.
- Todas nuestras métricas mostradas en la gráfica son evaluadas para nuestra clase 1-agua potable, Lo que se busca es entender cómo el modelo está realizando predicciones específicamente para esa clase.

$$\text{Precisión} = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Positivos}}$$

- La Precisión, nos evalúa sobre nuestra predicción, esto nos dice que de todos los Verdaderos Positivos predichos que tantos son Verdaderos Positivos y que tantos nos salieron como Falsos positivos. Esto quiere decir que mientras más se acerque al valor de 1 nuestra métrica, nuestra predicción será más buena. Pero esto no exime de que algunos Positivos se vayan a nuestros Falsos negativos. Debido a esto no podemos quedarnos solo con esta métrica, si es importante a tener en cuenta, pero no tiene que ser la única.

$$\text{Recall} = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Negativos}}$$

- El Recall, nos evalúa sobre nuestras muestras reales, esto nos dice que de todos nuestros datos reales Positivos que tantos son realmente Positivos y que tantos son Falsos Negativos. Esto quiere decir que mientras más se acerque el valor a 1, tendremos menos falsos negativos y mejor predicción de los Reales Positivos, esto es beneficioso porque si tenemos menos falsos Negativos, esto quiere decir que pocos verdaderos positivos se escapan de ser predichos como Positivos. Esta métrica nos ayuda mucho a saber que realmente nuestros datos positivos tengan predicción positiva.

$$\text{F1-Score} = 2 \times \frac{\text{Precisión} \times \text{Recall}}{\text{Precisión} + \text{Recall}}$$

- El F1-Score, esta métrica nos evalúa el promedio ponderado de la Precisión y del Recall. Esto es importante debido a que como observamos muchas veces ciertos modelos tienen más alta precisión que otros, pero en el Recall quedan relegados. Con esta métrica nos apoyamos de un aspecto estadístico para poder ponderar estos valores y deducir qué modelo podría comportarse de mejor manera. Normalmente se considera mucho este valor en data desbalanceada como en nuestro caso.

La formula del área bajo la Curva ROC (AUC) es:

$$AUC = \int_0^1 \text{TPR} \, d\text{FPR}$$

TPR : Tasa de Verdaderos Positivos (Sensibilidad o Recall),
FPR : Tasa de Falsos Positivos.

- Como última Métrica a Analizar, y también una de las más importantes es el AUC, el cual nos mide la comparación entre la Tasa de Verdaderos Positivos y los Falsos Positivos. Es importante tener en cuenta que lo que queremos es conseguir mayor número de verdaderos Positivos ya que queremos conseguir predecir a nuestra clase 0-Good customer. Por esta razón, mientras más alto sea el valor de AUC, esto nos indica que tendremos mayor cantidad de verdaderos Positivos con menos cantidad de Falsos Positivos.

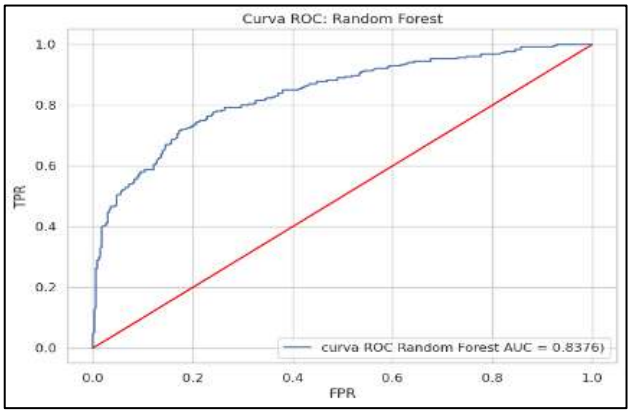
- En nuestro análisis, encontramos que tanto LightGBM como Random Forest tienen un buen rendimiento en términos de AUC, siendo LightGBM el ganador con un 84.8%. A pesar de esta ligera ventaja, vamos a automatizar ambos modelos usando GridSearchCV para ver cómo se comportan, ya que la diferencia en AUC entre los dos

modelos es mínima, de alrededor del 1%.

Finalmente, luego de analizar todas nuestras métricas que consideramos importantes para nuestra predicción, y sabiendo que lo que queremos conseguir es predecir de mejor manera nuestra clase 1, teniendo la menor cantidad de Falsos positivos y Falsos Negativos, concluimos que: El modelo a elegir es el Light GBM y Random Forest, pero después de la automatización me voy a quedar con uno.

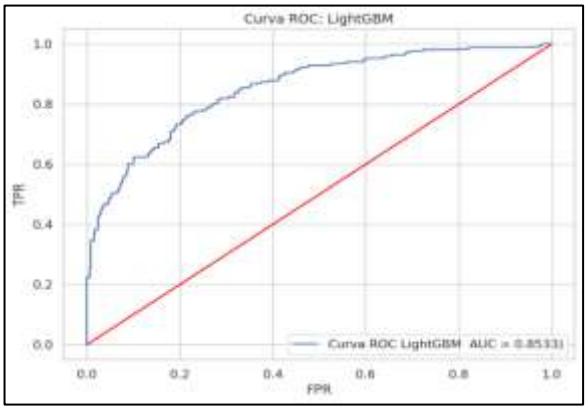
3.7. Optimizando nuestro modelo con Hiperparámetros (Random Forest)

	No_Potable	Potable	accuracy	macro avg	weighted avg
precision	0.804217	0.730769	0.771959	0.767493	0.772580
recall	0.792285	0.745098	0.771959	0.768691	0.771959
f1-score	0.798206	0.737864	0.771959	0.768035	0.772214
support	337.000000	255.000000	0.771959	592.000000	592.000000



3.8. Optimizando nuestro modelo con Hiperparámetros (LightGBM)

	No_Potable	Potable	accuracy	macro avg	weighted avg
precision	0.804217	0.730769	0.771959	0.767493	0.772580
recall	0.792285	0.745098	0.771959	0.768691	0.771959
f1-score	0.798206	0.737864	0.771959	0.768035	0.772214
support	337.000000	255.000000	0.771959	592.000000	592.000000



- **Conclusión:**

Después de optimizar los modelos usando Grid Search nos quedamos con el modelo LightGBM

Se logró mejorar la métrica AUC-ROC a 85.33%, lo que indica un buen rendimiento del modelo en términos de discriminación entre clases. Esta métrica es especialmente útil para evaluar modelos en problemas de clasificación binaria como la calidad del agua subterránea, donde es crucial identificar correctamente las muestras positivas (agua de calidad inaceptable).

La métrica AUC-ROC se basa en la tasa de verdaderos positivos (recall) y la tasa de falsos positivos. En este caso, un valor de 0.851 sugiere que el modelo es capaz de clasificar correctamente el 85.33% de los casos positivos (agua de buena calidad). Esto indica una buena capacidad del modelo para distinguir entre las dos clases.

Además, se seleccionó el modelo LGBMClassifier como el mejor para predecir la calidad del agua subterránea, lo que sugiere que este algoritmo es efectivo en este contexto. Esto podría atribuirse a su capacidad para manejar grandes

conjuntos de datos y su eficiencia en términos de tiempo de procesamiento.

En conclusión, con un modelo optimizado y una métrica AUC-ROC mejorada, se puede confiar en las predicciones del modelo LGBMClassifier para identificar la calidad del agua subterránea de manera efectiva, lo que puede ser fundamental para la toma de decisiones en la gestión y conservación de los recursos hídricos.

3.9. Prueba con instancia

```
nuevaCaracteristica_aguaSub = [{
    'orgCuen': 1,
    'ph': 7,
    'dureza': 221,
    'soliToDis': 17240,
    'conductividad': 442,
    'carbono_O': 16,
    'sulfato': 322,
    'trihalometanos': 78,
    'turbidez': 4.7,
}]
```

```
nuevaCaracteristica_aguaSub = pd.DataFrame.from_dict(nuevaCaracteristica_aguaSub)
instance_x = scaler.transform(nuevaCaracteristica_aguaSub)
```

```
# Predecir con el modelo
variant = model.predict(instance_x)
variant_proba = model.predict_proba(instance_x)
```

```
# Obtener el porcentaje de efectividad y el valor predicho
porcentaje_efectividad = round(variant_proba[0][variant[0]] * 100, 2)
valor_predicho = "Buena" if variant[0] == 1 else "Mala"

# Imprimir el resultado
print(f"La calidad del agua subterránea es {valor_predicho} con un {porcentaje_efectividad}% de confianza.")
```

4. Conclusión Final

Basándose en el análisis realizado con el modelo LightGBM, se concluye que la calidad del agua subterránea es buena, con un nivel de confianza del 96.9%. Este resultado respalda la efectividad del modelo para predecir la calidad del agua subterránea con precisión. Este hallazgo refuerza la confianza en la calidad y fiabilidad del modelo en la evaluación de la calidad del agua subterránea.

Referencias Bibliográficas

ANA (1 de junio de 2018). Metodología para la determinación del Índice de Calidad de Agua de los Recursos Hídricos Superficiales en el Perú. Obtenido de https://repositorio.ana.gob.pe/handle/20.500.12543/2440

ANA (13 de mayo de 2020). Lineamientos para la elaboración de los Diagnósticos de la calidad de los Recursos Hídricos Superficiales. Obtenido de https://www.ana.gob.pe/sites/default/files/normatividad/files/lineamientos%20diagnosticos%20de%20calidad_0.pdf

Bellido Davila, R. (2021). Evaluación del agua subterránea del sector de remanso de Characato Arequipa, factibilidad del uso del agua para consumo humano.Trabajo de fin de maestría.Universidad Nacional de San Agustín

Breiman, L. (1984). Classification and regression trees. Biometrics, 40(3), 874. https://doi.org/10.2307/2530946

Castellanos Díez Hector (2021). “Aplicación de Machine Learning al análisis de contaminantes en aguas subterráneas”. Trabajo de fin de maestría. Industriales UPN.

Chapoñan Cayao, J. C. (2023). Caracterización hidrogeoquímica y calidad del agua subterránea del distrito de Jayanca, departamento de Lambayeque.Tesis para obtener título.Universidad Católica Santo Toribio de Mogrovejo

Figueroa Condori E., Chavez Quispe B. et al. (2022). “Calidad de aguas superficiales y subterráneas en la zona de influencia de una cantera de yeso en el Perú”. Universidad Nacional de Moquegua. DOI: 10.37761/rsqp.v88i2.383

Friedman, J. H. (2012). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4), 367-378. https://doi.org/10.1016/s0167-9473(01)00065-2

Gil Marín J., Celeidys Vizcaino y Veliz Eduardo (2019). “Evaluación de la calidad del agua subterránea utilizando el índice de calidad del agua (ICA). Caso de Estudio: Acuíferos de Maturín, Estado Monagas, Venezuela”. Revista Científica UNTRM. DOI:10.25127/aps.20193.488

Guevara Manrique Jeimy (2022). “Monitoreo de la calidad del agua subterránea del Centro Poblado ‘El Carmen’ - Tambogrande utilizando el ICA como instrumentos de evaluación en el periodo mayo - agosto del 2022”. Tesis para obtener título. Universidad Nacional de Piura.

Guevara Vigo, C. J. (2021). Desarrollo de un sistema de remoción de arsénico de las aguas subterráneas del pozo tubular Comunidad: Punto 1 Sector 2 distrito Mochumí, Lambayeque.Tesis para obtener título. Universidad Nacional Pedro Ruiz Gallo.

López Velandia Cristian (2023). “Evaluación de la calidad del agua subterránea utilizando métodos de índice y análisis estadístico multivariado: cuenca del río Pavas (Colombia)”. South Sustainability, 4(1), e072. DOI: 10.21142/SS-0401-2023-e072

Louppe, G. (2014). Understanding random forests: From Theory to practice. arXiv (Cornell University). http://export.arxiv.org/pdf/1407.7502

Mitchell, T. M. (1997). Machine learning. McGraw-Hill Science/Engineering/Math.

Müller, A., & Guido, S. (2016b). Introduction to Machine Learning with Python: a guide for data scientists. http://cds.cern.ch/record/2229831

Novoa Julca Javier (2019). “Eficiencia y calidad del agua en 14 fuentes subterráneas, Baños del Inca”. Tesis para la obtención de título. Universidad Privada del Norte.

Ordoñez Gálvez Juan Julio (2015). Aguas subterráneas - Acuíferos. Sociedad Geográfica de Lima. Obtenido de https://www.gwp.org/globalassets/global/gwp-sam_files/publicaciones/varios/aguas_subterraneas.pdf

Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. IBM journal of research and development, 44(1.2), 206-226. https://doi.org/10.1147/rd.441.0206