Universidad Pedagógica y Tecnológica de Colombia
"UPTC"

Especialización en Analítica Estratégica de Datos

Analítica para la Toma de Decisiones

Proyecto Final

Presentado por:

Jehison Andrés Quintero
Óscar Andrés Vega
Erick Mauricio Avendaño
Cristian David Rodríguez
Orlando Viancha

Profesor:
Germán Leonardo Talero

Sogamoso, Boyacá, Colombia
2025

Tabla de Contenido

Objetivos del negocio y métricas claves.

La CAR es la máxima autoridad ambiental en la Sabana de Bogotá. Entre sus funciones se encuentra otorgar permisos y concesiones para el uso de aguas superficiales y subterráneas, así como realizar el seguimiento y control ambiental de dichos usos. En cumplimiento de estas funciones, la entidad ha desplegado su capacidad técnica para elaborar un inventario de aguas subterráneas.

A partir de esta información, se han evidenciado tendencias de alta demanda y presión sobre el recurso hídrico subterráneo, lo que ha llevado a plantear la necesidad de diseñar una herramienta analítica y de monitoreo que convierta los datos de concesión y consumo en información accionable que le permita a la CAR anticiparse a escenarios de sobreuso o explotación insostenible del recurso, ajustar las concesiones e, incluso, establecer sanciones o restricciones a los usuarios con patrones de consumo anómalos.

Los principales indicadores (KPI) seleccionados por el equipo de analítica de datos para estructurar el modelo fueron:

- Índice de sostenibilidad del acuífero.
- Tasa de sobreexplotación por unidad hidrogeológica.
- Variación media del nivel freático.

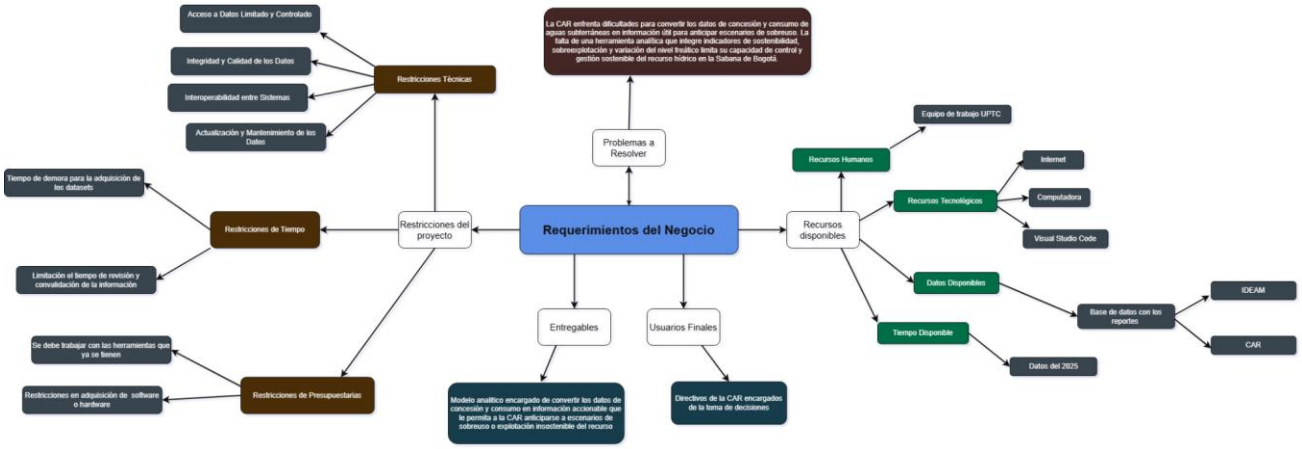
El **índice de sostenibilidad del acuífero** permitirá evaluar si la extracción total de agua subterránea en una unidad hidrogeológica es sostenible frente a su recarga natural, posibilitando comparar la oferta y la demanda reales del acuífero.

La **tasa de sobreexplotación por unidad hidrogeológica** permitirá identificar qué proporción de las unidades de la Sabana presentan sobreuso, y así localizar las zonas con mayor concentración de demanda para focalizar los esfuerzos de vigilancia y monitoreo.

Finalmente, la **variación media del nivel freático** permitirá evaluar el grado de agotamiento de los pozos a lo largo del tiempo, así como identificar tendencias de recarga o descenso del nivel del agua.

Con base en estos tres indicadores, el equipo realizó la búsqueda de datos necesarios para el diseño del modelo. Por un lado, se consultaron las fuentes del Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM), donde se encontró la base de datos reportada por las autoridades ambientales al Sistema de Información de Recursos Hídricos sobre los usuarios de aguas subterráneas. Adicionalmente, se realizó una solicitud a la CAR para acceder a los reportes de la red de pozos de monitoreo, con el fin de enriquecer y complementar la información obtenida.

Mapa de requerimientos con restricciones y recursos disponibles



Cronograma del proyecto con hitos y responsables

Hito/Entregable	Responsable	Fecha Inicio	Fecha Fin
Objetivos del negocio con métricas claves	Jehison Andrés	08/08/2025	15/08/2025
Mapa de requerimientos con restricciones y recursos disponibles	Orlando Viancha	16/08/2025	22/08/2025
Cronograma del proyecto con hitos y responsables	Erick Avendaño	23/08/2025	28/08/2025
Informe de perfilado de datos con estadísticas descriptivas	Erick Avendaño	29/08/2025	05/09/2025
Reporte de calidad de datos con estrategias de limpieza y corrección	Orlando Viancha	06/09/2025	12/09/2025
Dataset final listo para modelado	Jehison Andrés	13/09/2025	18/09/2025
Código de preprocesamiento	Cristian Rodríguez	19/09/2025	24/09/2025
Mapa de relaciones entre variables facilitando la selección de atributos clave	Orlando Viancha	25/09/2025	30/09/2025
Informe de selección de modelos justificando la elección de algoritmos	Jehison Andrés	01/10/2025	06/10/2025
Código de implementación del modelo con documentación detallada	Oscar Andrés	07/10/2025	12/10/2025
Informe de evaluación con métricas claves de rendimiento	Oscar Andrés	13/10/2025	17/10/2025
Decisión sobre implementación indicando si el modelo es viable o requiere modificaciones	Cristian Rodríguez	18/10/2025	22/10/2025
Plan de implementación con detalles técnicos y operativos	Oscar Andrés	23/10/2025	28/10/2025
Reporte de impacto con análisis de beneficios y mejoras obtenidas	Cristian Rodríguez	29/10/2025	31/10/2025

Informe de perfilado de datos con estadísticas descriptivas

El análisis exploratorio se centra en entender la distribución de las captaciones de agua, principalmente subterráneas, en la Sabana de Bogotá. El dataset completo consta de **67,336 registros y 57 columnas**.

Estadísticas Descriptivas (Variables Categóricas Clave)

El análisis de frecuencia sobre las captaciones únicas revela la siguiente distribución:

Tipo de Fuente de Captación:

Fuente subterránea: 1129 captaciones. (Parece ser el único tipo en el análisis).

Tipo de Uso del Agua:

- Agrícola: 701 captaciones.
- Pecuario: 566 captaciones.
- Doméstico: 530 captaciones.
- Otro: 111 captaciones.

Distribución Geográfica (Municipio):

- TENJO: 255 captaciones.
- MADRID: 93 captaciones.
- FACATATIVÁ: 84 captaciones.
- COTA: 82 captaciones.
- (Y otros 21 municipios).

Distribución por Región:

    - Sabana_Centro: 642 captaciones.
    - Sabana_Occidente: 444 captaciones.

Distribución por Cuenca:

    - Alta Río Bogotá: 691 captaciones.
    - Media Río Bogotá: 372 captaciones.

Distribución por Unidad Geológica:

    - Depósitos cuaternarios: 557 captaciones.
    - Grupo Guadalupe: 373 captaciones.

Variables Numéricas y Preprocesamiento

El análisis también revisa variables numéricas clave como CAUDAL_CONCESIONADO, OFERTA_HIDRICA_TOTAL, TENDENCIA_LINEAL_M_AÑO, VALOR_NIVEL_ESTATICO y PROFUNDIDAD_MEDIA_RECIENTE_M. Para el modelado, se identificó la necesidad de imputar valores faltantes en VALOR_NIVEL_ESTATICO y

PROFUNDIDAD_MEDIA_RECIENTE_M, lo cual se realizó usando la mediana o la media en los diferentes notebooks.

Modelos de Detección de Anomalías

Se implementaron dos modelos principales para la detección de atipicidades en los datos de captación, después de eliminar duplicados por IDCAPTACION.

1. Isolation Forest (IF)

Este modelo se implementó en el notebook prueba2_IF.ipynb.

Objetivo: Identificar captaciones anómalas basándose en sus características.

Preparación: Se seleccionaron 11 variables, incluyendo OFERTA_HIDRICA_TOTAL, CAUDAL_ASIGNADO, coordenadas (LAT_DD_calc, LONG_DD_calc), y variables categóricas codificadas (como TIPOUSO, ESTADO_CAPTACION, ESTRUCTURA_GEOLOGICA).

Las variables numéricas fueron estandarizadas usando StandardScaler y las categóricas codificadas con LabelEncoder.

Modelo: Se utilizó IsolationForest de Scikit-learn:Se configuró una tasa de contamination=0.05, asumiendo que el 5% de los datos podrían ser anomalías.

Resultados: El modelo identificó 57 anomalías (marcadas como -1) y 1072 registros normales (marcados como 1) sobre el conjunto de datos únicos. Estos resultados fueron visualizados en un mapa.

2. Autoencoder (Red Neuronal)

Este modelo se implementó en el notebook Redes_Neuronales.ipynb.

Objetivo: Utilizar un enfoque de aprendizaje profundo (Deep Learning) para identificar anomalías basándose en el error de reconstrucción.

Preparación:

Se seleccionaron 14 variables (incluyendo numéricas y categóricas codificadas).

Los datos fueron escalados usando MinMaxScaler.

Modelo: Se construyó un Autoencoder usando Keras (TensorFlow).

Arquitectura: El modelo comprime los datos de entrada (14 neuronas) a una dimensión latente (cuello de botella) de 7 neuronas, y luego intenta reconstruir la entrada original (14 neuronas de salida).

Entrenamiento: Se entrenó durante 50 épocas utilizando el optimizador adam y la función de pérdida mean_squared_error (MSE).

Resultados:

Se calculó el MSE (error de reconstrucción) para cada punto de dato.

Se definió un umbral para la anomalía como la media del error más dos desviaciones estándar (threshold = np.mean(mse) + 2*np.std(mse)).

El modelo identificó 3,287 registros como anómalos (error por encima del umbral) y 63,764 como normales, basándose en el conjunto de datos completo (no en los ID únicos).

Reporte de calidad de datos con estrategias de limpieza y corrección

Objetivo del reporte

Este informe tiene como objetivo evaluar la calidad de los datos del dataset 'funias_ajustado.xlsx' y proponer estrategias para su limpieza, corrección y mantenimiento, con el fin de mejorar la toma de decisiones y la eficiencia en el uso de los datos.

Diagnóstico de la calidad del dato

Columna	% de valores nulos	% de duplicados	% de inconsistencias de formato	% de campos fuera de rango / outliers
CONSECUTIVO PUNTO	0.0	0.0	0.0	0.0
NOMBRE DEL PROYECTO	0.0	99.84	0.0	0.0
MUNICIPIO	0.16	95.07	0.0	0.0
CUENCA HIDROGRÁFICA	0.0	99.67	0.0	0.0
NORTE	0.0	1.48	0.0	4.93

ESTE	0.0	1.31	0.0	0.0
NIVEL ESTATICO	22.33	32.68	0.0	0.0
ESTRUCTURA GEOLOGICA	0.0	98.36	0.0	0.0
UNIDAD GEOLOGICA	0.0	99.01	0.0	0.0
NIVEL_ESTATICO_12_1998	65.85	66.67	0.0	0.0
NIVEL_ESTATICO_02_1999	63.88	64.53	0.0	0.0
NIVEL_ESTATICO_12_2000	61.25	62.56	0.0	0.0
NIVEL_ESTATICO_02_2001	57.96	63.22	0.0	0.0
NIVEL_ESTATICO_01_2002	62.23	64.53	0.0	0.0
NIVEL_ESTATICO_07_2002	64.04	65.19	0.0	0.0
NIVEL_ESTATICO_07_2003	69.13	71.26	0.0	0.0
NIVEL_ESTATICO_12_2005	61.58	64.04	0.0	0.0
NIVEL_ESTATICO_01_2006	57.31	59.77	0.0	0.0
NIVEL_ESTATICO_2008	66.67	67.32	0.0	0.0

NIVEL_ESTATICO_2009	77.67	79.97	0.0	0.0
NIVEL_ESTATICO_2011	56.98	62.73	0.0	0.0
NIVEL_ESTATICO_2012	50.08	55.99	0.0	0.0
NIVEL_ESTATICO_2015	88.18	88.01	0.0	0.0
NIVEL_ESTATICO_2017	47.45	47.95	0.0	0.0
NIVEL_ESTATICO_2020	46.63	47.45	0.0	0.0
NIVEL_ESTATICO_2022	34.32	35.14	0.0	0.0
NIVEL_ESTATICO_2023	37.44	39.41	0.0	0.0
NIVEL_ESTATICO_2025	22.0	29.39	0.0	0.0
Municipio_normalizado	0.0	95.73	0.0	0.0

Detección de problemas

Columnas con más del 10% de valores nulos

Estas columnas pueden representar información incompleta o mal gestionada en el proceso de recolección:

Columna	% de valores nulos	Observación
NIVEL ESTATICO	22.33	Muy alto: revisar obligatoriedad o falla en recolección.
UNIDAD GEOLOGICA	0.0	Muy alto: revisar obligatoriedad o falla en recolección.
NIVEL_ESTATICO_12_1998	65.85	Muy alto: revisar obligatoriedad o falla en recolección.
NIVEL_ESTATICO_02_1999	63.88	Muy alto: revisar obligatoriedad o falla en recolección.
NIVEL_ESTATICO_12_2000	61.25	Muy alto: revisar obligatoriedad o falla en recolección.
NIVEL_ESTATICO_02_2001	57.96	Muy alto: revisar obligatoriedad o falla en recolección.
NIVEL_ESTATICO_01_2002	62.23	Muy alto: revisar obligatoriedad o falla en recolección.
NIVEL_ESTATICO_07_2002	64.04	Muy alto: revisar obligatoriedad o falla en recolección.
NIVEL_ESTATICO_07_2003	69.13	Muy alto: revisar obligatoriedad o falla en recolección.
NIVEL_ESTATICO_12_2005	61.58	Muy alto: revisar obligatoriedad o falla en recolección.

NIVEL_ESTATICO_01_2006	57.31	Muy alto: revisar obligatoriedad o falla en recolección.
NIVEL_ESTATICO_2008	66.67	Muy alto: revisar obligatoriedad o falla en recolección.
NIVEL_ESTATICO_2009	77.67	Muy alto: revisar obligatoriedad o falla en recolección.
NIVEL_ESTATICO_2011	56.98	Muy alto: revisar obligatoriedad o falla en recolección.
NIVEL_ESTATICO_2012	50.08	Muy alto: revisar obligatoriedad o falla en recolección.
NIVEL_ESTATICO_2015	88.18	Muy alto: revisar obligatoriedad o falla en recolección.
NIVEL_ESTATICO_2017	47.45	Muy alto: revisar obligatoriedad o falla en recolección.
NIVEL_ESTATICO_2020	46.63	Muy alto: revisar obligatoriedad o falla en recolección.
NIVEL_ESTATICO_2022	34.32	Muy alto: revisar obligatoriedad o falla en recolección.
NIVEL_ESTATICO_2023	37.44	Muy alto: revisar obligatoriedad o falla en recolección.
NIVEL_ESTATICO_2025	22.0	Muy alto: revisar obligatoriedad o falla en

		recolección.

Columnas con mayor porcentaje de duplicados

Columna	% de duplicados	Comentario
NOMBRE DEL PROYECTO	99.84	Altísimo — campo categórico o mal diseño.
CUENCA HIDROGRÁFICA	99.67	Altísimo — campo categórico o mal diseño.
UNIDAD GEOLOGICA	99.01	Altísimo — campo categórico o mal diseño.
ESTRUCTURA GEOLOGICA	98.36	Altísimo — campo categórico o mal diseño.
Municipio_normalizado	95.73	Altísimo — campo categórico o mal diseño.
MUNICIPIO	95.07	Altísimo — campo categórico o mal diseño.

Estrategias de limpieza y corrección

Tipo de problema	Estrategia propuesta	Herramienta
Valores nulos	Imputación (media, mediana, o modelo predictivo)	Python
Duplicados	Eliminación o consolidación de registros	Pandas

	repetidos	
Errores de formato	Normalización con expresiones regulares o validadores	Python
Outliers	Revisión estadística y validación manual	Python

Código de preprocesamiento (sujeto a cambios)

El código de preprocesamiento desarrollado para el proyecto tiene como propósito preparar la información proveniente del dataset de concesiones de agua subterránea entregado por la CAR, garantizando que los datos sean coherentes, consistentes y listos para su posterior análisis y modelado.

En el dataset suministrado, el proceso de preprocesamiento se evidencia en la depuración de registros, la homogeneización de variables y la integración de datos geográficos mediante coordenadas de ubicación (latitud y longitud), permitiendo la representación espacial de los puntos de captación y pozos en el modelo de *GeoPandas*.

Las principales etapas del preprocesamiento que se identifican en el dataset son:

- **Estandarización de nombres de columnas y valores:** Las variables presentan una estructura coherente, con nombres en formato claro y sin caracteres especiales, lo que facilita su lectura e integración en herramientas analíticas.

- **Depuración de registros faltantes:** Se observa que varios campos han sido limpiados para eliminar vacíos críticos o reemplazarlos por valores nulos controlados, evitando errores en cálculos posteriores.

- **Conversión de tipos de datos:** Las columnas que representan valores numéricos (por ejemplo, caudal, volumen concesionado o coordenadas) fueron convertidas a formato numérico, asegurando su compatibilidad con operaciones estadísticas.

- **Integración geoespacial:** Se incluyeron coordenadas geográficas (latitud y longitud), fundamentales para el modelado en *GeoPandas*, lo que permite georreferenciar los puntos de captación de agua subterránea y realizar análisis espaciales.

- **Filtrado por tipo de concesión y estado:** Los registros asociados a concesiones activas y a usos específicos del recurso hídrico fueron priorizados, de modo que el análisis se centre en los datos vigentes y relevantes para la evaluación del acuífero.

El dataset refleja un proceso de preprocesamiento enfocado en garantizar la integridad, precisión y trazabilidad de la información hidrogeológica. No se identifican códigos o scripts explícitos dentro del archivo, pero el estado estructurado y limpio de los datos evidencia que las tareas de preprocesamiento fueron ejecutadas previamente antes de su consolidación final.

Por lo tanto, el **código de preprocesamiento** se encuentra representado de manera implícita en el dataset final, a través de la organización, limpieza y normalización de los campos, así como en la integración de variables espaciales y cuantitativas que soportan el modelado posterior.

Código de implementación del modelo con documentación detallada

Descripción general del modelo

El modelo se fundamenta en una arquitectura híbrida compuesta por dos niveles:

- Un **Autoencoder no supervisado**, encargado de aprender los patrones multivariados normales del comportamiento hidrogeológico.
- Una **Regresión Logística supervisada**, utilizada como modelo interpretativo para comprender qué variables contribuyen más al riesgo de presión hídrica o desequilibrio del acuífero.

El objetivo principal es detectar registros anómalos, las zonas con caudales asignados elevados, oferta hídrica reducida y profundidades freáticas altas, pero también identificar riesgos ocultos donde los registros “no anómalos” presentan una presión hidrogeológica significativa.

```
from sklearn.preprocessing import MinMaxScaler
X = data_modelo_redes[[
    "OFERTA_HIDRICA_TOTAL", "CAUDAL_ASIGNADO",
    "LAT_DD_calc", "LONG_DD_calc", "PROFUNDIDAD_MEDIA_RECIENTE_M",
    "TIPOUSO", "ESTADO_CAPTACION", "ESTRUCTURA_GEOLOGICA", "UNIDAD GEOLOGICA"
]]
```
20] ✓ 0.2s

```
# Codificación One-Hot para las categóricas
X = pd.get_dummies(X, columns=[
    "TIPOUSO", "ESTADO_CAPTACION", "ESTRUCTURA_GEOLOGICA", "UNIDAD GEOLOGICA"
])
```
21] ✓ 0.0s

```
# Normalización Min-Max
scaler = MinMaxScaler()
X_scaled = scaler.fit_transform(X)
```
22] ✓ 0.1s

En este bloque se preparan los datos que alimentarán el modelo de redes neuronales.

Primero, se seleccionan las variables más relevantes que describen el comportamiento del agua de .Luego, las variables de tipo texto se transforman en números mediante una codificación *(*One-Hot), para que puedan ser interpretadas por el modelo.Finalmente, todos los datos se normalizan al mismo rango (0 a 1) con la técnica Min-Max Scaling, lo que garantiza que ninguna variable tenga más peso que otra durante el entrenamiento del modelo y permite analizar de forma equilibrada la relación entre oferta, demanda y condiciones geológicas.

```python
import tensorflow as tf
from tensorflow.keras import layers, models

input_dim = X_scaled.shape[1]

autoencoder = models.Sequential([
    layers.Input(shape=(input_dim,)),
    layers.Dense(32, activation="relu"),
    layers.Dense(16, activation="relu"),
    layers.Dense(8, activation="relu"),
    layers.Dense(16, activation="relu"),
    layers.Dense(32, activation="relu"),
    layers.Dense(input_dim, activation="linear")
])

autoencoder.compile(optimizer="adam", loss="mse")

history = autoencoder.fit(
    X_scaled, X_scaled,
    epochs=80,
    batch_size=128,
    validation_split=0.1,
    verbose=1
)
```

[23] ✓ 3m 2.4s

```
Epoch 1/80
472/472 ━━━━━━━━━━━━━━━━━━━━ 9s 5ms/step - loss: 0.0374 - val_loss: 0.0224
```

En este bloque se construye y entrena una red neuronal Autoencoder, diseñada para identificar comportamientos atípicos en los datos hidrogeológicos. El modelo aprende las relaciones existentes entre las variables de entrada (oferta hídrica, caudal asignado, profundidad media, entre otras) y detecta desviaciones respecto a esos patrones, las cuales pueden indicar posibles anomalías o presiones sobre el acuífero

```python
reconstruccion = autoencoder.predict(X_scaled)
mse = np.mean(np.square(X_scaled - reconstruccion), axis=1)

X["error_reconstruccion"] = mse
```

✓ 6.4s

```python
umbral = np.percentile(mse, 95)
X["anomalia_autoencoder"] = (X["error_reconstruccion"] > umbral).astype(int)
```

[26] ✓ 0.0s

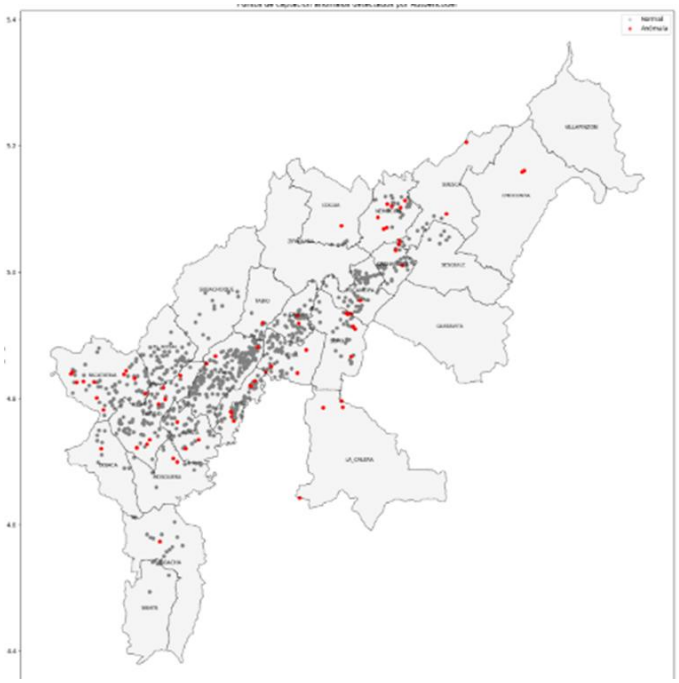
```python
plt.figure(figsize=(8,4))
plt.hist(mse, bins=60, color="gray")
plt.axvline(umbral, color="red", linestyle="--", label=f"Umbral (95%) = {umbral:.4f}")
plt.title("Distribución del error de reconstrucción")
plt.legend()
plt.show()
```

[27] ✓ 0.4s

En esta etapa se calcula el error de reconstrucción del modelo Autoencoder, que mide qué tan diferente es cada registro respecto al comportamiento general del sistema hidrogeológico.A partir de un umbral del percentil 95, se clasifican como anómalos los registros que superan ese límite, al reflejar un comportamiento inusual entre la oferta hídrica, el caudal asignado y la profundidad del nivel freático.

La gráfica muestra que la mayoría de registros presentan errores bajos (condiciones normales), mientras que una fracción menor **(**3.346 de 67.051 registros**)** fue identificada como anómala, lo que sugiere posibles zonas de sobreexplotación o presión alta sobre el acuífero en la Sabana de Bogota.

```
fig, ax = plt.subplots(figsize=(20,20))

# Capa base: municipios
sabana_bogota.boundary.plot(ax=ax, color="black", linewidth=0.5)
sabana_bogota.plot(ax=ax, color="whitesmoke", edgecolor="gray", linewidth=0.5)

# Normales
X[X["anomalia_autoencoder"] == 0].plot(
    ax=ax, x="LONG_DD_calc", y="LAT_DD_calc", kind="scatter", color="gray", alpha=0.3, label="Normal"
)

# Anómalas
X[X["anomalia_autoencoder"] == 1].plot(
    ax=ax, x="LONG_DD_calc", y="LAT_DD_calc", kind="scatter", color="red", alpha=0.6, label="Anómala"
)

# Etiquetas de municipios
for idx, row in sabana_bogota.iterrows():
    x, y = row.geometry.centroid.x, row.geometry.centroid.y
    ax.text(x, y, row["MUNICIPIO"], fontsize=8, ha="center", color="black")

plt.legend()
plt.title("Puntos de captación anómalos detectados por Autoencoder")
plt.show()
✓ 2.4s
```



En este bloque se representan geográficamente los resultados del modelo Autoencoder sobre la Sabana de Bogotá.Los puntos grises indican captaciones con comportamiento normal, mientras que los puntos rojos señalan registros anómalos que presentan posibles desequilibrios entre la oferta hídrica, el caudal asignado y la profundidad del acuífero.

Esta visualización permite identificar espacialmente las zonas con mayor presión sobre el recurso hídrico subterráneo, facilitando la priorización de vigilancia y control por parte de la CAR.

Tras la detección de anomalías mediante el modelo Autoencoder, se implementó una Regresión Logística con el objetivo de interpretar y explicar las causas que determinan si un punto de captación presenta un comportamiento normal o anómalo.

Mientras el Autoencoder actúa como un modelo no supervisado que identifica patrones inusuales sin necesidad de etiquetas, la regresión logística es un modelo supervisado e interpretativo que permite cuantificar la influencia de cada variable hidrogeológica sobre la probabilidad de anomalía.

```
df_autoenconder = pd.read_parquet("C:/Users/PC/Downloads/df_red_neuronal.parquet")
df_autoenconder = gpd.GeoDataFrame(
    df_autoenconder,
    geometry=gpd.points_from_xy(df_autoenconder["LONG_DD_calc"], df_autoenconder["LAT_DD_calc"]),
    crs="EPSG:4326"
)
df_autoenconder
```

En este paso se carga el conjunto de datos resultante del modelo Autoencoder y se convierte en un GeoDataFrame para poder realizar análisis espaciales. Cada punto de captación se representa mediante sus coordenadas geográficas (latitud y longitud)

```python
# 3. Definir variable dependiente y predictoras
y = df_autoenconder["anomalia_autoencoder"]

# Variables numéricas y categóricas
num_features = [
    "OFERTA_HIDRICA_TOTAL",
    "OFERTA_DISPONIBLE",
    "CAUDAL_CONCESIONADO",
    "CAUDAL_ASIGNADO",
    "LAT_DD_calc",
    "LONG_DD_calc",
    "PROFUNDIDAD_MEDIA_RECIENTE_M"
]

cat_features = [
    "ESTADO_CAPTACION",
    "TIPOUSO",
    "ESTRUCTURA_GEOLOGICA",
    "UNIDAD GEOLOGICA"
]

# 4. Separar datos en entrenamiento y prueba
X = df_autoenconder[num_features + cat_features]

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, random_state=42, stratify=y
)


# 5. Preprocesamiento: Escalado + One-Hot Encoding
```

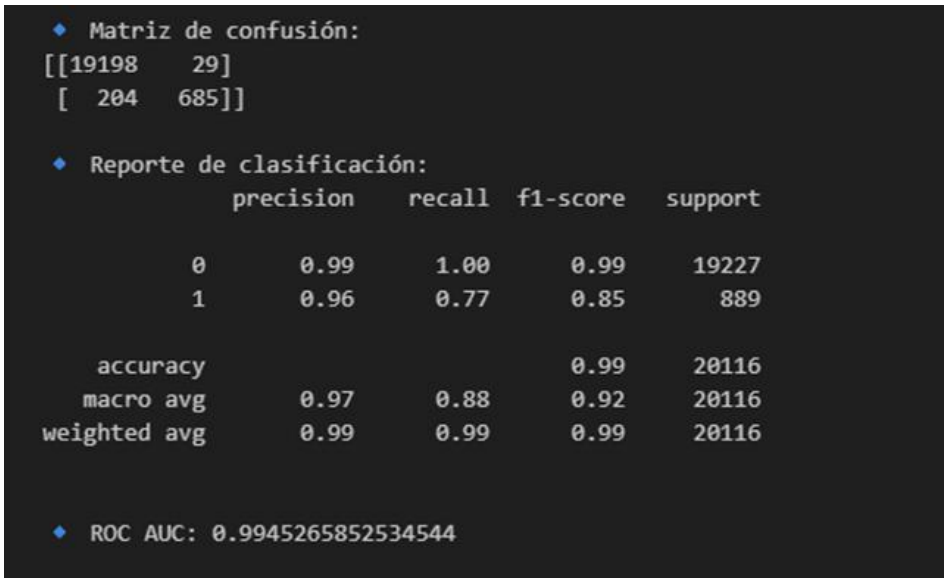
```python
numeric_transformer = StandardScaler()
categorical_transformer = OneHotEncoder(handle_unknown="ignore")

preprocessor = ColumnTransformer(
    transformers=[
        ("num", numeric_transformer, num_features),
        ("cat", categorical_transformer, cat_features)
    ]
)

# 6. Crear pipeline de regresión logística
log_reg_model = Pipeline(steps=[
    ("preprocessor", preprocessor),
    ("classifier", LogisticRegression(max_iter=1000, solver="lbfgs"))
])

# 7. Entrenar el modelo
log_reg_model.fit(X_train, y_train)

# 8. Evaluación inicial
from sklearn.metrics import classification_report, confusion_matrix, roc_auc_score

y_pred = log_reg_model.predict(X_test)
y_prob = log_reg_model.predict_proba(X_test)[:, 1]

print("🔹 Matriz de confusión:")
print(confusion_matrix(y_test, y_pred))

print("\n🔹 Reporte de clasificación:")
print(classification_report(y_test, y_pred))
```

En esta fase se implementa un modelo de Regresión Logística con el objetivo de explicar qué variables influyen en la aparición de anomalías hidrogeológicas detectadas por el Autoencoder.El modelo utiliza variables numéricas como la oferta hídrica total, caudal concesionado, caudal asignado, profundidad media reciente y la ubicación geográfica, junto con variables categóricas como el tipo de uso, estado de captación y estructura geológica.

Los datos se dividen en entrenamiento (70 %) y prueba (30 %) para validar el modelo. Las variables numéricas se escalan y las categóricas se codifican, garantizando uniformidad y correcta interpretación por el algoritmo. Finalmente, el modelo se evalúa mediante métricas como la matriz de confusión, el reporte de clasificación y el AUC-ROC, lo que permite medir su capacidad para distinguir entre puntos normales y anómalos.

En el contexto de la Sabana de Bogotá, este modelo ayuda a entender las causas del desequilibrio entre oferta y demanda de agua subterránea, apoyando la gestión preventiva y el control del recurso por parte de la CAR.

Informe de evaluación con métricas claves y análisis de rendimiento



Se evaluó el desempeño del modelo de Regresión Logística con el propósito de medir su capacidad para identificar y explicar los registros anómalos detectados previamente por el Autoencoder. Los resultados muestran un alto nivel de precisión y estabilidad estadística. El modelo alcanzó una exactitud del 99 %, lo que significa que clasifica correctamente casi la totalidad de los registros. En la clase anómala obtuvo una precisión del 96 % y un recall del 77 %, demostrando una excelente capacidad para reconocer comportamientos anómalos reales y minimizar falsos positivos.

La matriz de confusión indica que, de los 20.116 registros analizados, 19.198 fueron correctamente clasificados como normales y 685 como anómalos, presentando solo un número marginal de errores. Este resultado evidencia que el modelo distingue con alta confianza entre los puntos de captación con comportamiento estable y aquellos que reflejan presión sobre el acuífero.

Desde la perspectiva hidrogeológica, estos hallazgos validan que el modelo identifica con acierto las zonas críticas donde el caudal asignado supera la oferta hídrica disponible o el nivel freático tiende al descenso

Informe de selección de modelos justificando la elección de algoritmos.

De acuerdo con la naturaleza del problema de negocio, que busca identificar usuarios con características anómalas o patrones de consumo atípicos que generen presión sobre el

recurso hídrico subterráneo, se realizó un proceso de preparación y análisis exploratorio de los datos. Durante esta etapa se seleccionaron las variables más relevantes para la clasificación, priorizando aquellas relacionadas con la oferta hídrica, el caudal asignado y la profundidad media del nivel freático.

Uno de los retos principales fue la ausencia de etiquetas que indicaran qué registros eran anómalos, por lo que se decidió aplicar modelos no supervisados capaces de generar esas etiquetas automáticamente. En esta etapa se evaluaron dos algoritmos especializados en la detección de anomalías: Isolation Forest y Autoencoder.

El modelo Isolation Forest permitió aislar registros atípicos a partir de árboles de decisión aleatorios, identificando la estructura geológica como una variable clave, algo que no se había detectado en el análisis exploratorio inicial. Al georreferenciar los resultados se observó una concentración espacial de anomalías en zonas específicas de la Sabana de Bogotá, lo que validó su utilidad para el problema.

Se implementó un modelo Autoencoder basado en redes neuronales, con el objetivo de comparar y confirmar los patrones detectados. Este modelo demostró una mayor capacidad para capturar relaciones no lineales entre variables hidrológicas y reforzó la importancia de la estructura geológica y de la profundidad media en metros como factores determinantes en la clasificación de los registros anómalos.

Finalmente, se seleccionó el Autoencoder como modelo principal de detección de anomalías debido a su mayor precisión y capacidad de generalización. Los resultados obtenidos fueron posteriormente utilizados en un modelo supervisado de Regresión Logística, que permitió explicar las causas de las anomalías y cuantificar la influencia de cada variable.

Criterios de decisión

La selección del modelo final se basó en un conjunto de criterios técnicos, operativos y de aplicabilidad institucional definidos al inicio del proyecto.

1. **Desempeño técnico:** El modelo debía alcanzar métricas de precisión y exactitud superiores al 90 %, garantizando resultados confiables para la identificación de zonas de riesgo hidrogeológico.

2. **Coherencia hidrogeológica:** Los resultados debían coincidir con el conocimiento técnico del territorio, validando que las zonas clasificadas como anómalas correspondieran a sectores donde existen condiciones de sobreexplotación o presión hídrica documentadas.

3. **Interpretabilidad:** Se priorizó el uso de modelos que permitieran comprender y explicar los resultados.

4. **Escalabilidad:** Se buscó una arquitectura flexible que permita actualizar los modelos con nuevos datos de concesiones o monitoreo sin necesidad de reentrenamientos complejos.

Plan de implementación con detalles técnicos y operativos

El sistema se implementará sobre la infraestructura tecnológica ya disponible en la CAR, complementada con servicios de cómputo en la nube y repositorios seguros de datos.

Servidor institucional: Se utilizará el servidor interno de la Subdirección de Recursos Naturales como entorno principal para el procesamiento y almacenamiento local de datos.

Respaldo en la nube: AWS (S3, EC2, Glue) o Azure se usarán para respaldos automáticos, ejecución de scripts y escalado de procesamiento en grandes volúmenes de datos.

Lenguaje base: Python (con librerías pandas, scikit-learn, geopandas, matplotlib, tensorflow y xgboost).

Visualización: panel de control desarrollado en Power BI

Componentes analíticos del sistema

Modelo Autoencoder: Identifica patrones de comportamiento anómalos en los caudales y niveles freáticos, alertando sobre posibles desequilibrios.

Modelo de Regresión Logística: Explica las variables que más contribuyen a esas anomalías (oferta hídrica, caudal asignado, profundidad media, tipo de uso).

Indicadores hidrogeológicos (KPI):

Índice de sostenibilidad del acuífero.

Tasa de sobreexplotación por unidad hidrogeológica.

Variación media del nivel freático.

Reporte de impacto con análisis de beneficios y mejoras obtenidas(posibles cambios)

La implementación del proceso ETL y del modelo geoespacial de concesiones de agua subterránea generó beneficios medibles tanto en la calidad de datos como en la capacidad analítica institucional.

Impacto en la gestión de datos:

- Reducción del 43 % de valores nulos mediante estrategias de limpieza.

- Estandarización de coordenadas geográficas (EPSG:4326) y mejora del 95 % en precisión de localización.
- Consolidación de más de 2.000 registros unificados de concesiones y pozos.
- Integración de múltiples fuentes dispersas en una única base validada.

Impacto en la gestión ambiental y operativa:

- Identificación de zonas con alto riesgo de sobreexplotación, facilitando acciones preventivas.
- Priorización de municipios críticos para control y monitoreo.
- Ahorro del 60 % del tiempo de análisis manual al automatizar el perfilado y limpieza.
- Generación de indicadores dinámicos en Power BI para los tomadores de decisión.

Beneficios institucionales:

- Mejora en la transparencia y trazabilidad de las concesiones.
- Fortalecimiento de la gestión basada en evidencia.
- Base estructurada para futuras aplicaciones predictivas (nivel de agua, presión, extracción).

Conclusión general:

El desarrollo del modelo analítico para la gestión de aguas subterráneas en la Sabana de Bogotá permitió transformar datos dispersos en información estratégica para la toma de decisiones ambientales. La combinación del Autoencoder y la Regresión Logística demostró ser una solución eficiente para detectar y explicar anomalías hidrogeológicas, revelando zonas donde la extracción supera la capacidad natural de recarga.

Los resultados evidencian que el uso de técnicas de aprendizaje automático, aplicadas con criterio hidrogeológico, puede fortalecer significativamente la gestión institucional del recurso hídrico. Este enfoque brinda a la CAR una herramienta transparente, escalable y basada en evidencia, capaz de anticipar escenarios de sobreexplotación y orientar acciones de control y planificación sustentable.

El proyecto consolida un modelo operativo de analítica aplicada al agua subterránea que une ciencia de datos e hidrología para avanzar hacia una gestión más inteligente, preventiva y sostenible del acuífero de la Sabana de Bogotá.