Informe de laboratorio: Modelado y Predicción de Floraciones de Cianobacterias en los Lagos Atitlán y Amatitlán

Fecha: 17 de agosto de 2025

1. Introducción y Objetivos

Las floraciones de cianobacterias representan un grave riesgo ecológico y para la salud pública en los cuerpos de agua dulce de Guatemala, particularmente en los lagos de Atitlán y Amatitlán. La capacidad de monitorear y predecir estos eventos es fundamental para la gestión ambiental y la toma de decisiones informada.

El presente Laboratorio tuvo como objetivo principal desarrollar y evaluar una serie de modelos de aprendizaje automático para detectar y predecir la presencia de cianobacterias en ambos lagos, utilizando datos de teledetección satelital y variables climáticas.

Los objetivos específicos del laboratorio fueron:

1. **Modelo de Serie Temporal:** Desarrollar un modelo para predecir el avance de la cianobacteria a lo largo del tiempo.
2. **Modelo de Clasificación:** Elaborar un modelo para determinar la probabilidad de presencia de cianobacteria en un punto geográfico específico.
3. **Modelo Híbrido:** Combinar las técnicas anteriores para crear un modelo predictivo más robusto que incorpore tendencias temporales y factores externos.
4. **Análisis de Modelos:** Discutir la utilidad, el rendimiento y las limitaciones de los modelos desarrollados.
5. **Visualización:** Presentar los resultados de los modelos en mapas geoespaciales interactivos.

---

2. Metodología y Desarrollo

Para abordar los objetivos, se siguió una metodología incremental, comenzando con modelos simples y avanzando hacia enfoques más complejos. Los datos de entrada principales consistieron en índices espectrales derivados de imágenes satelitales (NDCI, FAI, NDWI) y coordenadas geográficas.

2.1. Intento de Modelo de Serie Temporal (ARIMA)

El primer enfoque fue utilizar un modelo ARIMA (Promedio Móvil Integrado Autoregresivo) para predecir la concentración media de Clorofila-a (un indicador de biomasa de algas) en el futuro. Este tipo de modelo se basa exclusivamente en los valores históricos de la serie para identificar tendencias y estacionalidades.

El código implementado para esta tarea fue:

```
def predict_cyanobacteria_index(df, lake_name):
    """
    Utiliza un modelo ARIMA para predecir el índice de cianobacterias.
    """
    if df is None or len(df) < 10: # Se necesita suficientes datos para el modelo
        print(f"No hay suficientes datos para el lago {lake_name} para hacer una predicción.")
        return

    # Preparar y dividir los datos
    df_series = df.set_index('date')['mean_chl_a']
    train, test = df_series[0:int(len(df_series)*0.8)], df_series[int(len(df_series)*0.8):]

    # Modelo ARIMA
    history = [x for x in train]
    predictions = list()
    for t in range(len(test)):
        model = ARIMA(history, order=(5,1,0))
        model_fit = model.fit()
        output = model_fit.forecast()
        yhat = output[0]
        predictions.append(yhat)
        history.append(test[t])
    # ... (código de evaluación y visualización)
```

2.2. Modelo de Clasificación Espacial (Random Forest)

El segundo modelo se centró en un problema de clasificación espacial: para un píxel específico en una imagen satelital, ¿existe una alta concentración de cianobacterias? Para esto, utilizamos un clasificador de `Random Forest`, un modelo robusto y eficaz que, además, nos permite evaluar la importancia de cada variable predictora.

Las características utilizadas fueron los índices NDCI, FAI, NDWI y las coordenadas normalizadas (X, Y) de cada píxel.

```
def train_classification_model(features, labels, lake_name):
    # División de datos
    X_train, X_test, y_train, y_test = train_test_split(
        features, labels, test_size=0.3, random_state=42, stratify=labels
    )
    # Escalamiento y entrenamiento
    scaler = StandardScaler()
    X_train_scaled = scaler.fit_transform(X_train)
    X_test_scaled = scaler.transform(X_test)
    rf_model = RandomForestClassifier(
        n_estimators=100,
        random_state=42,
        class_weight='balanced'
    )
    rf_model.fit(X_train_scaled, y_train)
    # ... (código de evaluación)
    return rf_model, scaler, X_test_scaled, y_test, y_pred, y_pred_proba
```

2.3. Modelo Híbrido

Finalmente, se desarrolló un modelo híbrido para combinar la predicción temporal con variables exógenas (climáticas). El enfoque fue:

1. Crear un modelo simple de regresión lineal para capturar la tendencia temporal de la Clorofila-a.
2. Utilizar la predicción de este modelo como una característica de entrada para un clasificador más potente (`Gradient Boosting`).
3. Añadir a este segundo modelo datos climáticos (temperatura, precipitación, etc.) y de presión urbana para enriquecer el contexto de la predicción.

El objetivo era predecir si la condición general del lago en una fecha futura sería de "floración" o "no floración".

---

3. Resultados Obtenidos

3.1. Modelo de Serie Temporal

La ejecución de este modelo arrojó el siguiente resultado de inmediato:

```
No hay suficientes datos para el lago Atitlán para hacer una predicción.
No hay suficientes datos para el lago Amatitlán para hacer una predicción.
```

Conclusión: Los modelos de series de tiempo como ARIMA requieren un historial de datos largo y consistente para ser efectivos. Nuestro conjunto de datos, con solo 5 observaciones temporales, fue completamente insuficiente, por lo que este enfoque tuvo que ser descartado.

3.2. Modelo de Clasificación

Este modelo fue el más exitoso del laboratorio. Los resultados fueron excelentes para ambos lagos.

Lago Atitlán:

- **Precisión (Accuracy): 97.8%**
- **Reporte de Clasificación:**

	precision	recall	f1-score	support
Sin cianobacteria	0.99	0.98	0.99	27480
Con cianobacteria	0.93	0.98	0.96	9152

- **Análisis Visual:** Las gráficas (Figura 1) muestran una matriz de confusión casi perfecta y una clara separación en la distribución de probabilidades.
- **Importancia de Características:** El **NDCI fue, por mucho, el predictor más importante, con un 83.5%** de la influencia en la decisión del modelo.

Figura 1: Análisis del Modelo de Clasificación - Atitlán

Lago Amatitlán:

- **Precisión (Accuracy): 95.3%**
- **Reporte de Clasificación:**

	precision	recall	f1-score	support
Sin cianobacteria	0.97	0.97	0.97	3441
Con cianobacteria	0.90	0.91	0.91	1146

- **Análisis Visual:** Similar a Atitlán, el modelo muestra un rendimiento sólido (Figura 2).
- **Importancia de Características:** Aunque el NDCI sigue siendo el más importante (49.3%), otras variables como el FAI (19.0%), NDWI (11.7%) y las coordenadas tienen un peso considerablemente mayor que en Atitlán.

Figura 2: Análisis del Modelo de Clasificación - Amatitlán

3.3. Modelo Híbrido

El entrenamiento del modelo híbrido arrojó resultados aparentemente perfectos:

- **Precisión del modelo híbrido (Atitlán): 1.000**
- **Precisión del modelo híbrido (Amatitlán): 1.000**

Sin embargo, estos resultados se lograron utilizando únicamente **5 puntos de datos** para el entrenamiento. Una precisión del 100% en estas condiciones es una señal inequívoca de **sobreajuste (overfitting)**. El modelo no aprendió a generalizar, sino que simplemente memorizó las 5 muestras que le proporcionamos. La inestabilidad del modelo se hizo evidente en la importancia de características, donde para Atitlán atribuyó el 100% de la importancia a la velocidad del viento, un resultado poco plausible.

3.4. Visualización de Resultados

Utilizando el **modelo de clasificación (el único modelo fiable)**, se generaron mapas de riesgo para cada lago.

- **Mapa de Atitlán:** Se identificaron **1,676 puntos de calor**, con **1,503 clasificados como de alto riesgo**. Esto indica una presencia significativa y extendida de cianobacterias en la fecha del último dato satelital.
- **Mapa de Amatitlán:** Se identificaron **125 puntos de calor**, con **93 de alto riesgo**. Aunque el problema es presente, parece estar más localizado en comparación con Atitlán.

Se generaron y guardaron tres archivos HTML interactivos: `mapa_prediccion_atitlan.html`, `mapa_prediccion_amatitlan.html` y `mapa_comparativo_lagos.html`, que permiten una exploración detallada de estas zonas de riesgo.

---

4. Análisis y Discusión de Resultados

El laboratorio revela una dicotomía clara entre la capacidad de **diagnosticar (clasificar)** y **pronosticar (predecir a futuro)**.

El Éxito del Diagnóstico Espacial: El modelo de Random Forest demostró ser una herramienta extremadamente poderosa y precisa para el **monitoreo** de la cianobacteria. Con una precisión superior al 95%, podemos tomar una imagen satelital reciente y generar un mapa de riesgo detallado casi en tiempo real. La diferencia en la importancia de las características entre los dos lagos sugiere que las condiciones en Amatitlán son más complejas (posiblemente por mayor turbidez o coexistencia con otras especies de algas), lo que obliga al modelo a depender de una combinación de índices más rica.

El Desafío del Pronóstico y la Limitación de Datos: Nuestros intentos de predecir la evolución de la cianobacteria a futuro (tanto con ARIMA como con el modelo híbrido) fracasaron. La razón no fue una falla en la metodología conceptual, sino una **limitación fundamental: la escasez de datos temporales**. Con solo 5 puntos en el tiempo, es estadísticamente imposible para cualquier modelo aprender patrones de tendencia o la influencia de variables climáticas. El sobreajuste del modelo híbrido es una lección clave: un modelo complejo con pocos datos produce resultados engañosos y no confiables.

Validez de los Mapas: Es crucial entender que los mapas de riesgo generados son una **"fotografía" del estado actual** basada en nuestro robusto modelo de clasificación. No representan una predicción a futuro. Son una herramienta de diagnóstico, no de pronóstico.

---

5. Conclusiones y Recomendaciones

1. **Conclusiones Clave:**
    - **Es posible clasificar la presencia de cianobacterias con muy alta precisión (>95%)** utilizando datos de índices espectrales satelitales a través de un modelo de Random Forest.
    - El **NDCI** es el predictor más significativo, confirmando su validez científica para la detección de cianobacterias.
    - La **predicción a futuro** (pronóstico) es actualmente **inviable** debido a la falta de una serie de datos históricos suficientemente larga.
    - La **principal limitación** de este laboratorio no fue la técnica de modelado, sino la **disponibilidad de datos**.
2. **Recomendaciones para el Uso de los Modelos:**
    - **Implementar el Modelo de Clasificación:** Se recomienda utilizar este modelo de forma operativa para generar mapas de riesgo periódicos (ej. mensuales). Esto puede guiar a las autoridades en la toma de muestras de agua, la gestión de recursos y la emisión de alertas a la población.
    - **No Utilizar el Modelo Híbrido:** Las predicciones de este modelo deben ser descartadas. Su uso para la toma de decisiones sería irresponsable dado el severo sobreajuste detectado.
3. **Recomendaciones para Futuros Trabajos:**
    - **Adquisición de Más Datos:** La prioridad absoluta para mejorar la capacidad predictiva es construir un conjunto de datos histórico. Se necesita recopilar y procesar imágenes satelitales (ej. Sentinel-2, Landsat) de forma mensual o bimensual, abarcando un período de **al menos 5 años**.
    - **Integrar Datos Climáticos Históricos:** Paralelamente, se debe recopilar una serie de tiempo de datos climáticos (temperatura, precipitación, radiación solar) de estaciones cercanas para el mismo período histórico.
    - **Re-entrenar el Modelo Híbrido:** Solo con un conjunto de datos robusto se podrá entrenar un modelo híbrido o de serie temporal que pueda generar predicciones a futuro fiables y útiles para la gestión proactiva de las floraciones de cianobacterias.

Link al repositorio

https://github.com/DiegoValdez10/Lab-5-data-sciece.git