Documentación
Modelos multidatos

https://github.com/fabianhuertas1992/DatosML.git

Octubre de 2024

Introducción

Este proyecto captura y utiliza diversas fuentes de datos para desarrollar modelos de aprendizaje automático enfocados en la predicción de biomasa aérea (AGB) en áreas geográficas.
En este proyecto se utiliza machine learning para predecir la biomasa aérea (AGB) en áreas forestales, un indicador clave para la gestión de recursos naturales. A través de modelos como Regresión Lineal, Random Forest y XGBoost, se exploran técnicas que combinan datos geoespaciales, como el NDVI y el diámetro de los árboles (DAP), con la altura y el año de la muestra, para generar predicciones en polígonos de estudio.

Lo que Hacemos

1. Datos Reales: Usamos datos forestales y satelitales (NDVI) para construir modelos que predicen biomasa.
2. Modelos Comparados: Probamos tres modelos de machine learning, visualizando sus predicciones en mapas interactivos.
3. Limitaciones: Aclaramos que, debido a la escasez de datos, los modelos no alcanzan su máximo potencial.

Recomendaciones

Para mejorar los resultados, se recomienda ampliar la cantidad de datos y mejorar el preprocesamiento, permitiendo a los modelos generar predicciones más robustas y precisas.

Temas Claves y Objetivos

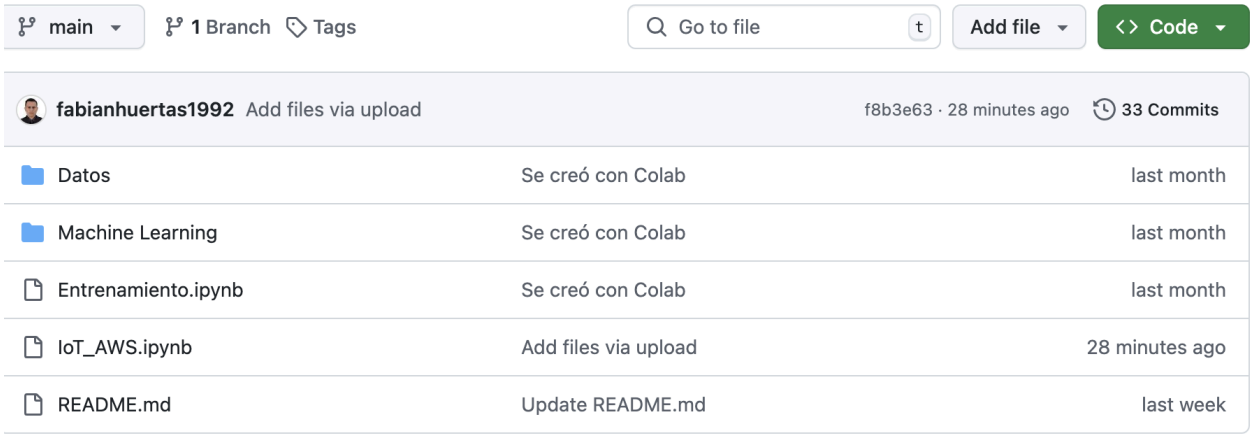
Los objetivos de este manual son asegurar que los usuarios puedan:

1. **Desarrollar modelos predictivos para la biomasa aérea** usando datos forestales, satelitales y de imágenes RGB capturadas por dron, explorando diferentes enfoques de machine learning.
2. **Implementar DeepForest para la detección de copas de árboles** en imágenes RGB, mejorando la comprensión de la estructura forestal y su relación con la biomasa.
3. **Comparar el rendimiento de tres modelos de machine learning** para determinar cuál ofrece mejores resultados en la predicción de biomasa.
4. **Visualizar las predicciones de biomasa en mapas interactivos**, facilitando la interpretación de resultados y la toma de decisiones.
5. **Identificar las limitaciones actuales de los modelos**, específicamente la falta de datos suficientes, y recomendar pasos para mejorar la precisión de las predicciones.
6. **Proporcionar una base para futuras investigaciones** en la predicción de biomasa con un enfoque en la sostenibilidad ambiental.

Índice de Contenidos

Estructura del proyecto

El repositorio contiene notebooks Jupyter desarrollados en Google Colab para el análisis y modelado de datos.



Carpeta 1: Datos

Notebooks para preprocesamiento de datos y generación de características.

Carpeta 2: Machine Learning

Notebooks para la implementación de modelos de machine learning (Regresión Lineal, Random Forest, XGBoost) y la predicción de biomasa.

Cada notebook puede ser ejecutado directamente en Google Colab, asegurando que las dependencias necesarias estén instaladas previamente.

Predicción de Biomasa Aérea con Modelos de Machine Learning

Entrenamiento.ipynb

Es un notebook que realiza la predicción de biomasa aérea (AGB) utilizando tres modelos de machine learning: Regresión Lineal, Random Forest y XGBoost. El enfoque de este análisis es generar un polígono de estudio basado en puntos de muestreo y realizar predicciones de AGB dentro de ese polígono.

Características Clave

- **Modelos Utilizados:** Se entrenan tres modelos de machine learning para la predicción de la biomasa:
    - Regresión Lineal
    - Random Forest
    - XGBoost
- **Datos de Entrada:**
    - Coordenadas GPS de los árboles.
    - Diámetro del árbol a la altura del pecho (DAP).
    - Altura de los árboles.
    - Año de muestreo.
- **Generación de Polígono:**
    - Se utiliza una envolvente convexa para generar un polígono que cubre el área de interés basada en los puntos de muestreo.
    - Dentro de este polígono, se genera una cuadrícula de puntos, a los que se les asignan valores promedio de diámetro, altura y año.

- **Predicción de Biomasa Aérea (AGB):**
    - Los modelos previamente entrenados son utilizados para predecir la AGB en cada punto de la cuadrícula generada dentro del polígono.

- **Visualización de Resultados:**
    - Los resultados se agrupan en intervalos de AGB y se visualizan en tres mapas, uno por cada modelo. Cada mapa colorea los puntos del polígono según el intervalo de AGB predicho, facilitando la comparación visual entre los tres modelos.

Instalación de Librerías.

Para ejecutar este notebook, asegúrate de instalar las siguientes librerías en tu entorno de trabajo:

```
!pip install rasterio
!pip install SQLAlchemy aiomysql
!pip install geopandas
!pip install folium
!pip install boto3
```

Ejecución del Notebook

Puedes ejecutar este notebook directamente en Google Colab utilizando el siguiente enlace: https://colab.research.google.com/github/fabianhuertas1992/DatosML/blob/main/Entrenamiento.ipynb

Resultados y Visualización

Los resultados de las predicciones se visualizan en forma de mapas, donde cada punto dentro del polígono está coloreado según el rango de AGB predicho:

- Mapa 1: Predicción de biomasa utilizando el modelo de Regresión Lineal.
- Mapa 2: Predicción de biomasa utilizando el modelo de Random Forest.
- Mapa 3: Predicción de biomasa utilizando el modelo de XGBoost.

Análisis de Datos de IoT

IoT_AWS.ipynb

En este proyecto, trabajamos con datos provenientes de sensores IoT que miden variables como la **temperatura** y la **humedad** en un entorno determinado. Estos datos fueron analizados con el objetivo de predecir las condiciones ambientales en el futuro, utilizando dos modelos de machine learning: **Random Forest y ARIMA.**

El uso de dispositivos IoT permite la recolección continua y en tiempo real de datos ambientales, lo que proporciona una base sólida para realizar análisis predictivos y entender cómo estas variables se comportarán en los próximos días. Los sensores de IoT ofrecen un flujo constante de información que se puede utilizar para anticipar cambios en el clima o en las condiciones locales de manera más eficiente.

Procesos Realizados:

1. **Cargar y Preprocesar los Datos:**
    - Los datos de los sensores IoT fueron organizados y convertidos a un formato numérico adecuado para su análisis.
    - Se eliminaron valores no válidos y se reorganizaron los datos de temperatura y humedad para facilitar su uso en los modelos de predicción.
2. **Ingeniería de Características:**
    - Se crearon características adicionales a partir de los datos originales, como diferencias entre mediciones consecutivas, promedios móviles y características cíclicas (hora del día y mes del año). Estas características son esenciales para mejorar la capacidad predictiva de los modelos.
3. **Modelos de Predicción:**
    - Utilizamos dos tipos de modelos de machine learning para realizar las predicciones:
        - **Random Forest:** Un modelo basado en la construcción de múltiples árboles de decisión para hacer predicciones robustas.

    - **ARIMA:** Un modelo especializado en series temporales, que analiza los patrones en los datos a lo largo del tiempo para predecir valores futuros.
  - Estos modelos fueron entrenados para realizar predicciones de temperatura y humedad en dos horizontes temporales: 7 días y 30 días.
4. **Visualización de Resultados:**
  - Se generaron gráficas que muestran las predicciones de ambos modelos para los próximos 7 días y 30 días. Estas visualizaciones permiten comparar cómo cada modelo anticipa los cambios en las variables ambientales.

Ejecución del Notebook

Puedes ejecutar este notebook directamente en Google Colab utilizando el siguiente enlace: https://colab.research.google.com/drive/1_uYBmQjpQJzwA--yjPdgqisv1zUz8VJq?usp=sharing

Detección de Copas y Cálculo de Biomasa con DeepForest

DeepForest.ipynb

Este script utiliza DeepForest para detectar copas de árboles en imágenes RGB capturadas por un dron y estima la biomasa aérea (AGB) utilizando ecuaciones específicas para cada especie de árbol.

Descripción General

El análisis se enfoca en:

- Dividir una imagen RGB en mosaicos para procesarla más eficientemente.
- Detectar copas de árboles en cada mosaico utilizando un modelo DeepForest preentrenado.
- Calcular el diámetro a la altura del pecho (DAP) estimado y la biomasa aérea (AGB) por especie para cada copa detectada.
- Visualizar los resultados en una imagen con copas detectadas y mostrar el conteo total de copas y la biomasa estimada por especie.

Flujo del Script

1. **Carga de la Imagen RGB**
    - La imagen se carga usando rasterio, dividiendo las bandas para obtener una imagen RGB normalizada.
    - La imagen utilizada debe estar en formato .tif, obtenida a partir de un dron.
2. **Inicialización de DeepForest**
    - Se utiliza un modelo DeepForest preentrenado, con un umbral de confianza (score_threshold) ajustado a 0.5 para la detección de copas.
    - El modelo puede ser finetuneado previamente con imágenes específicas, cargando pesos personalizados.
3. **División de la Imagen en Mosaicos**
    - La imagen completa se divide en mosaicos de 4000x4000 píxeles para permitir un procesamiento más eficiente.
    - Esto ayuda a DeepForest a manejar grandes imágenes de manera óptima.
4. **Calcular la Biomasa por Especie**
    - Se implementa una función que utiliza ecuaciones alométricas para calcular la biomasa aérea (AGB) en función del DAP y la especie del árbol.
    - Las ecuaciones utilizadas son específicas para cada especie:
    - Banana: 0.030 * (DAP ** 2.13)
    - Cacao: 0.1208 * (DAP ** 1.98)
    - Fruit: 0.1466 * (DAP ** 2.223)
    - Timber: 21.3 - 6.95 * DAP + 0.74 * (DAP ** 2)
5. **Encontrar el DAP más Cercano y la Especie**
    - Se compara la posición de la copa detectada con datos de campo, buscando el árbol más cercano para obtener su DAP y especie correspondiente.
6. **Visualización de Resultados**
    - Se genera una imagen con rectángulos dibujados alrededor de las copas detectadas.
    - Se imprime el número total de copas detectadas y la biomasa total estimada, así como la biomasa por especie.

Resultados

El script proporciona:

- Número total de copas detectadas.
- Biomasa total estimada (AGB).
- Biomasa estimada por especie, con los siguientes valores por defecto:
- Banana
- Cacao
- Fruit
- Timber
- Other (si no se reconoce la especie).

Ejecución en Google Colab

Este script está diseñado para ejecutarse en Google Colab. Puedes cargar la imagen .tif y el archivo de datos de campo (field_data_Nestor.csv) desde Google Drive, asegurándote de tener las rutas correctas al inicio del script.

Ejecuta este notebook en Google Colab: https://colab.research.google.com/github/fabianhuertas1992/DatosML/blob/main/DeepForest.ipynb

Carpeta 1: Datos

Multidatos.ipynb

Este notebook está diseñado para la captura y análisis de datos de dispositivos, con especial enfoque en parámetros como temperatura y humedad. Se manipulan datos de diferentes dispositivos, almacenados en un DataFrame, y se realiza un análisis exploratorio.

El archivo Multidatos.ipynb realiza las siguientes acciones:

1. **Conexión a varias fuentes de datos:**

- Oracle: Para obtener imágenes almacenadas en una base de datos.
- KoboToolbox: Para consultar datos de ubicación (GPS) relacionados con encuestas o análisis de campo.

- IoT | AWS: Para recuperar datos de dispositivos IoT almacenados en un bucket S3 en AWS, principalmente relacionados con la temperatura y la humedad.

2. **Instalación de dependencias:** Asegura que se instalen las librerías necesarias para el procesamiento y la visualización de los datos, como rasterio, SQLAlchemy, geopandas, folium, y boto3.

3. **Consulta y visualización de datos:**

- Base de datos Oracle: Consulta imágenes anteriores y posteriores almacenadas en la base de datos.
- KoboToolbox: Recupera datos de GPS y los filtra, mostrando ubicaciones precisas y realizando una consulta a una API para obtener datos catastrales y polígonos asociados a cada coordenada.
- IoT | AWS: Consulta dispositivos en AWS para recuperar valores de temperatura y humedad, transformando los datos y formateando fechas y horas para su análisis.

4. **Visualización:**

- Muestra los tres DataFrames resultantes: Oracle, KoboToolbox, y AWS IoT.

Ejecuta este notebook en Google Colab: https://colab.research.google.com/github/fabianhuertas1992/DatosML/blob/main/Datos/Multidatos.ipynb

Query_DB.ipynb

Este notebook realiza consultas a múltiples bases de datos (Oracle, KoboToolbox, IoT a través de AWS) y procesa las respuestas para generar datasets y visualizaciones. También se aplican modelos de aprendizaje para realizar análisis y simulaciones de los datos obtenidos.

Funcionalidades principales:

- Oracle: Consulta datos de imágenes y NDVI a través de un sistema de base de datos, procesando los resultados para crear visualizaciones y estadísticas sobre vegetación y análisis geoespacial.
- KoboToolbox: Recupera datos de geolocalización para su análisis y visualización en un mapa con datos catastrales de parcelas.

- IoT | AWS: Consulta dispositivos IoT que registran valores de temperatura y humedad, almacena los datos en S3 y los procesa para obtener estadísticas y gráficos de estos valores por dispositivo..

Visualizaciones

- Mapa interactivo con datos catastrales y ubicaciones GPS.
- Gráficos de barras y de calor para mostrar la distribución de temperatura y humedad de los dispositivos.
- Gráfico 3D interactivo que visualiza los datos de temperatura y humedad por dispositivo.
- Simulación de temperaturas a lo largo del tiempo con gráficos de tendencias.

Ejecuta este notebook en Google Colab: https://colab.research.google.com/github/fabianhuertas1992/DatosML/blob/main/Datos/Query_DB.ipynb

Carpeta 2: Machine Learning

Esta carpeta contiene notebooks que implementan y comparan diferentes modelos de machine learning para predecir la biomasa aérea (AGB) en Mg/ha. Los modelos evaluados incluyen Regresión Lineal, Random Forest y XGBoost. A continuación, se detallan los contenidos de cada archivo en esta carpeta:

Modelo_ML.ipynb

Este notebook se enfoca en la extracción, transformación y carga de datos (ETL) desde diversas fuentes, incluyendo KoboToolbox. Después de procesar los datos, se entrenan diferentes modelos de machine learning para predecir la biomasa en función de variables como el NDVI, el diámetro a la altura del pecho (DAP) y la altura de los árboles.

Principales características:

- Procesamiento de datos desde KoboToolbox.
- Comparación inicial de modelos de aprendizaje automático.

Ejecuta este notebook en Google Colab:
https://colab.research.google.com/github/fabianhuertas1992/DatosML/blob/main/Machine%20Learning/Modelo_ML.ipynb

ML_comparativos.ipynb

Este notebook implementa una comparación detallada de tres modelos de machine learning (Regresión Lineal, Random Forest y XGBoost) para predecir la biomasa en Mg/ha.

Características del notebook:

- **Carga y preprocesamiento de datos:** Se eliminan valores nulos y se generan nuevas características como NDVI_DAP y HT_NDVI.
- **Selección de características clave:** NDVI, DAP, altura, NDVI_DAP y HT_NDVI.
- **Comparación de modelos:** Se evalúan los modelos utilizando MSE y R^2, con visualizaciones gráficas que muestran las diferencias entre las predicciones y los valores reales de biomasa.
- **Conclusiones:** Se visualiza cómo cada modelo captura la relación entre las características y la biomasa.

Ejecuta este notebook en Google Colab:
https://colab.research.google.com/github/fabianhuertas1992/DatosML/blob/main/Machine%20Learning/ML_comparativos.ipynb

ML_Parcelas.ipynb

Este notebook realiza el entrenamiento de modelos agrupando los datos por parcelas, lo que permite observar la variabilidad espacial en las predicciones de biomasa.

Aspectos clave:

- **Agrupación de datos por parcelas:** Permite que los modelos enfoquen el análisis en áreas específicas.
- **Modelos utilizados:** Se implementan Regresión Lineal, Random Forest y XGBoost.
- **Evaluación de rendimiento:** Se evalúan los modelos con métricas como R^2 y MSE.

- **Visualización de resultados:** Los valores predichos por cada modelo se comparan gráficamente con los valores reales de biomasa.

Ejecuta este notebook en Google Colab: https://colab.research.google.com/github/fabianhuertas1992/DatosML/blob/main/Machine%20Learning/ML_Parcelas.ipynb

ML_Parcelas(entrenamiento).ipynb

Similar al notebook ML_Parcelas.ipynb, este archivo entrena modelos de machine learning para predecir la biomasa en función de características como NDVI, DAP y altura de los árboles, pero se enfoca más en el proceso de entrenamiento.

Aspectos clave:

- **Problemas en los datos:** Se concluye que los datos actuales no son suficientes para obtener predicciones precisas, lo que sugiere la necesidad de más datos o mejores técnicas de preprocesamiento.
- **Comparación entre modelos**: Al igual que en otros notebooks, se utilizan XGBoost, Random Forest y Regresión Lineal.

Ejecuta este notebook en Google Colab: https://colab.research.google.com/github/fabianhuertas1992/DatosML/blob/main/Machine%20Learning/ML_Parcelas(entrenamiento).ipynb

Conclusión General

El conjunto de notebooks proporcionados implementa un análisis profundo para la predicción de biomasa aérea (AGB) en un área forestal utilizando técnicas de machine learning, basadas en modelos como Regresión Lineal, Random Forest, y XGBoost. A lo largo de los notebooks, se abordaron diferentes enfoques para predecir la biomasa, incluyendo la agrupación por parcelas, el uso de cuadrículas geoespaciales y la visualización de los resultados en mapas interactivos.

Logros

1. **Extracción y Preprocesamiento de Datos:**
    - Los datos fueron obtenidos de fuentes diversas como KoboToolbox y se realizaron transformaciones importantes como la generación de nuevas características basadas en el NDVI, el diámetro a la altura del pecho (DAP), y la altura de los árboles.
2. **Implementación y Comparación de Modelos:**
    - Se implementaron y compararon modelos de machine learning para evaluar su rendimiento en la predicción de biomasa. Aunque los modelos presentaron algunas diferencias en su capacidad de predicción, los resultados mostraron que las variaciones entre los modelos pueden depender de los datos disponibles y las características utilizadas.
3. **Visualización de Resultados:**
    - Los resultados se presentaron en forma de mapas interactivos que permiten visualizar la distribución espacial de la biomasa aérea, proporcionando una herramienta útil para la comparación entre los modelos y para la toma de decisiones en la gestión forestal.

Limitaciones y Recomendaciones

Sin embargo, una conclusión clave derivada de este trabajo es que la cantidad de datos utilizados es insuficiente para obtener predicciones precisas y confiables. La variabilidad y cantidad limitada de los datos actuales limitan la capacidad de los modelos para generalizar y predecir adecuadamente la biomasa en un área más amplia o en condiciones más complejas.

Recomendaciones:

- **Ampliar el conjunto de datos:** Es crucial aumentar la cantidad y la variedad de los datos disponibles, especialmente para cubrir una mayor cantidad de muestras y condiciones ambientales. Esto proporcionará a los modelos un conjunto de entrenamiento más robusto que les permita hacer predicciones más precisas.
- **Mejorar el preprocesamiento de datos:** Un enfoque más detallado en la limpieza y generación de características puede ayudar a mejorar la precisión de los modelos.
- **Evaluación con más métricas:** Incluir más métricas de evaluación para obtener una visión más clara del rendimiento de los modelos bajo diferentes escenarios.

Aunque los modelos implementados han mostrado potencial para predecir la biomasa aérea, el éxito de estos algoritmos depende en gran medida de la cantidad y calidad de los datos de entrada. Para avanzar en la precisión de las predicciones y la utilidad práctica de estos modelos, se requiere la recolección de más datos y la exploración de técnicas adicionales de machine learning y preprocesamiento.