Flooded Area Coverage Prediction

Team Name: Group 8
Members: Elena Liao, Xinyi (Jessica) Wang

Problem/Motivation Statement

The Nature Conservancy (TNC) and partners have created BirdReturns to pay farmers to flood their fields to better support migratory wetland birds. As continue to scale, technology is needed to cost-effectively monitor enrolled fields to ensure they are flooded. Our ML objective is to use machine learning algorithms to predict flood coverage. We plan to investigate the following and more approaches for water coverage estimation.

Dataset and Analytic Goals

Dataset:

1. We used Google Earth Engine and Sentinel 2 satellites to collect data on reflected light bands. Specifically, we collected 10 bands (Band2, Band3, Band4, Band5, Band6, Band7, Band8, Band8A, Band11, and Band12) for our analysis.

2. TNC survey data and threshold(to calculate the Normalize the Difference Weather Index - NDWI) saved in Google drive.

Analytic goals:

Our goal is to use machine learning algorithms to predict flood coverage.

Data Engineering Pipeline

Here is the automated pipeline using Airflow for extraction data from Google Earth Engine API and Google Drive, uploading them to Google Cloud (GCS) and transferring to MongoDB, and finally using Databricks to implement machine learning:

1. Use the Google Earth Engine Python API to retrieve image data from Sentinel 2 satellites.
2. Use Airflow to download the survey data from Google Drive.
3. Save the image and survey data to GCS.
4. Use Airflow to insert the data from GCS to MongoDB.
5. Use Databricks to read the data from MongoDB and implement machine learning algorithms(K-means) to predict flood coverage.

Data Preprocessing

Cluster Specification:

- Databricks Runtime Version: 7.3 LTS (includes Apache Spark 3.0.1, Scala 2.12)
- Worker type: i3.xlarge
- Driver types: i3.xlarge

- number of workers: 5(max workers), 2(min workers), 2(current)

Preprocessing goals:

Before feeding it to K-means, we performed scaling on the data before feeding it to K-means model. Since K-means is the distance-based algorithm and bands data range widely for different bands. Without scaling, features with larger values will have a greater impact on the clustering than features with smaller values. By scaling the data, we can ensure that all features have equal weighting in the clustering process.

We also used the VectorAssembler to convert all the features (normalized band values) into one single vector to feed to the K-means model.

Algorithms: Standard Scaler

Time efficiency:

- Preprocessing Pipeline: 1.43 sec

Machine Learning Modeling

ML Goals:

Our objective is to use machine learning algorithms, specifically K-means clustering, to estimate the flood coverage and compare it to the traditional estimate approach using NDWI, which is the normalized difference between Band3 and Band8. We plan to investigate opportunities to use more bands and improve accuracy.

Modeling Outcomes:

- Clustering Model Evaluation: for our model, the Silhouette score is 0.44, which is not perfect but within reasonable range. The Silhouette score measures how well each data point fits into its assigned cluster compared to other clusters. It ranges from -1 to 1, and a higher Silhouette score indicates better clustering performance, indicating the data point is very similar to its assigned cluster and dissimilar to others.

- Result Comparision between Two approaches

We compared the model clustering results and the NDWI calculation results for all the dates. The result for the date of March 7, 2023 and summarized in the table below:

Table: NDWI and K-means Results Comparision

NDWI Calculation (Label)	K-means Clustering	Count
1	0	32
1	1	516
0	0	98
0	1	2756

Based on the results above, about 82% of the clustering outcomes from two approaches matches. Here the K-Means clustering label is different from the NDWI label.

- Final Results Visualization

We visualized the results for three selected days, comparing satellite images based on color (RGB), K-means clusters, and labeled water coverages. The results for Feb 27 and Mar 3 are quite similar, therefore only the results for Feb 27 and Mar 7 are shown below.

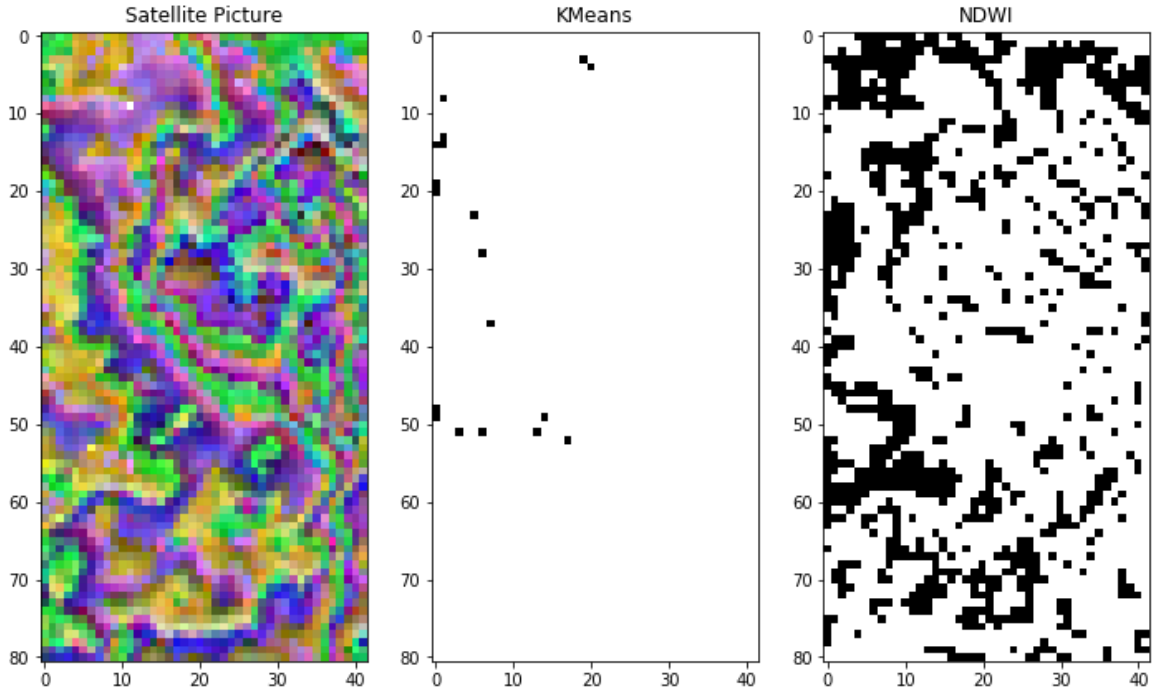

Figure 1. Results for Feb 27, 2023

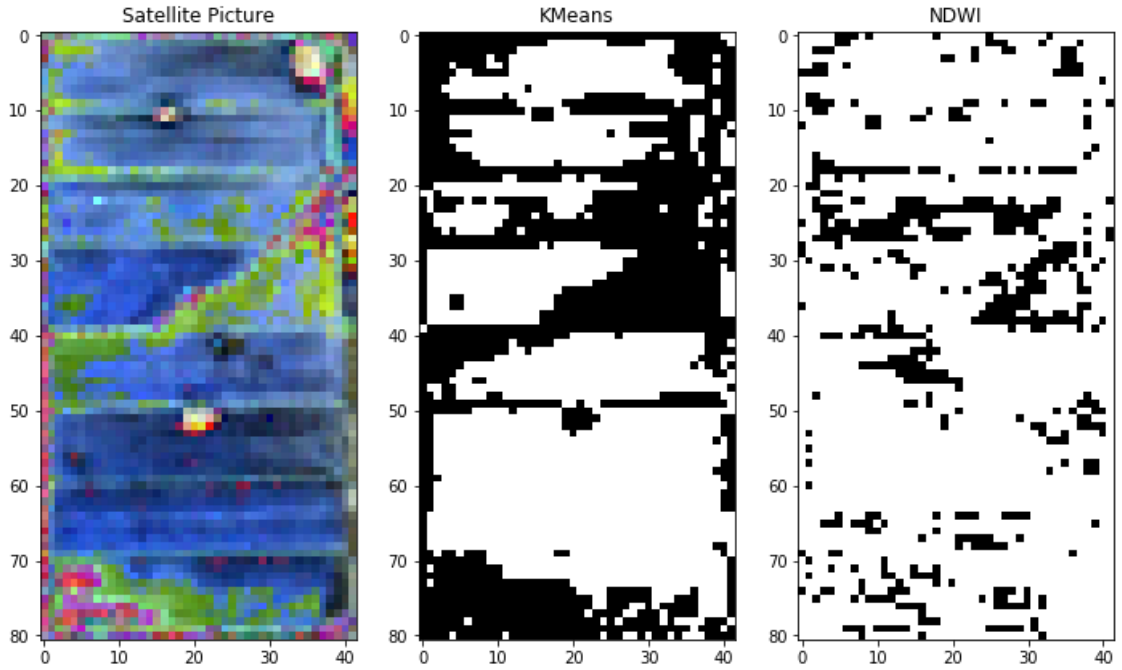

Figure 2. Results for Mar 7, 2023

For Feb 27, 2023, the satellite image doesn't reflect the selected farmland. The possible reason is that the field was covered by cloud on that day. The cloud coverage impacts the model clustering significantly.

For the date of Mar 7, it is clear that there was no cloud coverage and the plot shows promising clustering outcomes from KMeans. For the K-means plot above, the white areas are identified as a flooded and the black area are identified as land. The clustering results match the satellite picture presentation and more sensitive the the shallow flooded areas.

Execution Time Efficiency:

- Implement Model: 5.45 sec
- Evaluate Model: 3.54 sec

Lesson Learned

Through this project, we gained valuable insights into the use Airflow for data pipeline management. We are originally using GitHub Action workflow for the automated pipeline. We found out that GitHub Actions is a more flexible tool than Airflow as it does not require a local environment and can be used for various tasks, such as continuous integration and deployment.

Additionally, this project highlighted the importance of collaboration between different tools and platforms to achieve a common goal. By using various tools like GitHub, Databricks, and AWS, we were able to stream our workflow and improve the overall efficiency of our project.

In addition, using Spark ML with distributed computing shows time efficiency benefits. It could be a good option for scaling up and processing more and larger satellite images.

Conclusion

For this project, we successfully implement the data pipeline with Airflow, GCS, Spark, and MongoDB. We also investigated the clustering approach to estimate farmland water coverage using the Spark ML K-means algorithm.

We discovered that clustering algorithms can be an alternative option for water coverage estimation. We plan to investigate the opportunities to improve accuracy by exploring other clustering and feature selection approaches.