Cropland Classification in Arid & Semi-Arid Regions

- **Methodology:** CRISP-DM (Cross-Industry Standard Process for Data Mining)
  Authors:

  - James Wachira
  - Tim Musungu
  - Vivian Kwamboka
  - Hashim Ibrahim
  - Calvin Mutua

  Date: 22 Aug 2025
  Repository/Notebook: *https://github.com/James-Wachira/phase5_project.git*
  Stakeholders: Agricultural policymakers, NGOs, satellite analytics providers, food security initiatives

---

Executive Summary

Goal: Develop a machine learning pipeline using Sentinel-1 (radar) and Sentinel-2 (optical) satellite data to identify cropland in arid/semi-arid regions.
Data: Training dataset with cropland labels (`Train.zip`), unlabeled test dataset (`Test.csv`), Sentinel-1 (`Sentinel1.csv`), Sentinel-2 (`sentiment2` supplemental dataset).
Approach: Applied CRISP-DM framework with exploratory analysis, feature engineering, classical ML baselines, and deep learning models (including transformer architectures).
Headline Findings (replace with actual results from notebook): - Cropland can be reliably distinguished with combined Sentinel-1 + Sentinel-2 data. - Transformer models outperformed baselines by **+x% F1 score**. - Seasonal signals (NDVI trends, backscatter variability) are strong discriminators in arid landscapes. - Identified risk factors: cloud cover in Sentinel-2 imagery; spatial imbalance across training samples.

Recommendations: Deploy ensemble of CNN/transformer models for operational cropland mapping; prioritize gap-filling strategies for missing Sentinel-2 data.

1. Business Understanding

1.1 Background & Problem Statement

Food security monitoring in arid/semi-arid regions requires accurate cropland maps. Existing global datasets lack resolution and regional specificity. Sentinel satellites provide free, high-resolution radar/optical imagery, but require advanced ML to classify cropland under challenging conditions (cloud cover, sparse vegetation).

1.2 Objectives & Success Metrics

- **Primary KPI:** F1-score for cropland vs non-cropland classification.
- **Secondary KPIs:**
    - Precision/Recall balance.
    - Robustness across geographic regions.
    - Scalability for operational deployment.

1.3 Constraints & Assumptions

- Cloud cover reduces Sentinel-2 optical availability.
- Sentinel-1 radar provides all-weather data but lower interpretability.
- Data labels (Train.zip) may have geographic bias.
- GPU resources constrain deep learning training scale.

1.4 Risks

- **Bias:** Class imbalance (cropland << non-cropland).
- **Data quality:** Missing values in Sentinel-2 bands.
- **Transferability:** Model trained in one region may underperform in another.

2. Data Understanding

2.1 Sources

- **Train.zip**: Contains labeled cropland/non-cropland samples.
- **Test.csv**: Unlabeled samples for submission.
- **Sentinel1.csv**: Radar features (VV, VH polarization backscatter; temporal composites).
- **Sentinel2 (sentiment2 dataset)**: Optical features (bands B2–B12, NDVI, EVI, temporal composites).

2.2 Data Dictionary (Excerpt)

Field	Type	Description
`id`	string	Unique identifier
`label`	int	1 = cropland, 0 = non-cropland
`s1_vv_mean`	float	Mean VV backscatter
`s1_vh_std`	float	Std. dev. VH backscatter
`s2_B4_mean`	float	Mean Red band
`s2_B8_ndvi`	float	NDVI index
`lon`, `lat`	float	Coordinates

2.3 Initial Quality Profile

- Missing values in Sentinel-2 bands (due to clouds).
- Skewed class distribution: cropland samples ≈ *x%* of training set.
- Sentinel-1 radar features relatively complete (all-weather).

2.4 Exploratory Analysis

- Histograms: NDVI distribution → cropland clusters at higher values.
- Radar backscatter patterns: Cropland shows distinctive seasonal variance.
- Geographic plots: Training labels concentrated in a few regions (possible bias).

3. Data Preparation

3.1 Cleaning

- Handle missing Sentinel-2 bands (interpolation, median imputation, temporal composites).
- Remove duplicates and invalid coordinates.
- Normalize features (per band standardization).

3.2 Feature Engineering

- **Indices:** NDVI, EVI, NDWI, radar ratios (VV/VH).
- **Temporal metrics:** Seasonal amplitude, variance, harmonic features.
- **Spatial context:** Buffer statistics around points (if available).
- **Interaction features:** Radar × optical combined indices.

3.3 Train/Validation/Test Strategy

- Split by geography (not random) to test generalization.
- Stratify by class to address imbalance.
- Use k-fold CV with region-aware folds.

---

4. Modeling

4.1 Problem Framing

Binary classification: cropland vs non-cropland.
Inputs = Sentinel-1 + Sentinel-2 features; Outputs = binary label.

4.2 Algorithms Evaluated

- **Classical baselines:** Random Forest, XGBoost.
- **Deep learning:** CNNs (for spectral bands).
- **Ensembles:** Voting/stacking to combine strengths.

4.3 Model Performance

Model	F1	Precision	Accuracy	Notes
Random Forest	59.18	63.01	64.24	Baseline
XGBoost	65.02	67.94	67.81	Strong tabular baseline
Neural Networks	92.0	91.0	92.0	Best performing

4.4 Feature Importance / Interpretation

- Sentinel-2 NDVI & red-edge bands strongly predictive.
- Sentinel-1 VH variance adds value where Sentinel-2 missing.
- Temporal features critical for arid regions (seasonality of planting cycles).

---

5. Evaluation (Business)

5.1 Business Interpretation

- Cropland detection feasible with high F1 (≥x%).
- Combining radar + optical improves resilience against cloud cover.
- Model outputs can directly support agricultural monitoring and policy.

5.2 Fairness, Bias & Ethics

- Avoid over-reliance on biased training samples.
- Validate across multiple geographies.
- Use outputs to support—not penalize—farmers.

5.3 Limitations

- Cloud cover gaps in Sentinel-2.
- Geographic transferability.
- High compute cost for deep learning models.

---

6. Deployment

We developed a 'Cropland Prediction Dashboard' using Streamlit as a platform. It had the following features:

- Data upload via a CSV file
- Model Selection via a Dropdown Menu
- DataFrame generated that summarized Predictions and Confidence Level of Predictions
- A Mapbox visualizing the corresponding location coordinates of our Predictions
- A Visualization of important Features corresponding to each model

This tool was geared towards helping agricultural researchers and policy planners identify, map and carry out necessary interventions on arid and semi-arid areas.

7. Recommendations

7.1 Recommendations

- Use transformer ensemble as primary model.
- Deploy pipeline with both Sentinel-1 + Sentinel-2 features.
- Augment with cloud gap-filling techniques (e.g., temporal smoothing, radar substitution).
- Continuously retrain with new labeled samples from target regions.

7.2 Next Steps

- Acquire more labeled data for underrepresented geographies.
- Test pipeline in operational monitoring system.
- Publish open dataset + model weights for transparency.
- Extend classification to **crop type** (beyond binary cropland/non-cropland).

---

8. Conclusion

The CRISP-DM framework guided the cropland classification project from business understanding to deployment. Sentinel-1 and Sentinel-2 data, when combined with modern ML techniques (transformers), enable accurate mapping of cropland in challenging arid regions. This supports food security initiatives by providing scalable, cost-effective monitoring of agricultural land.