Cost-Effective and Accurate Cropland Mapping in Arid Regions using AI/ML and Remote Sensing

---

1. Executive Summary

This project aims to develop a cost-effective and accurate method for cropland mapping in arid regions using machine learning and remote sensing data. Using time-series satellite imagery (Sentinel-1 and Sentinel-2), the project leverages geospatial features and a Random Forest model to classify cropland vs non-cropland areas in **Fergana (Uzbekistan)** and **Orenburg (Russia)**. The final model achieves high accuracy while being computationally efficient and reproducible using open-source tools.

---

2. Problem Statement

Mapping cropland accurately in arid and semi-arid regions is challenging due to **spectral similarities with pastures and steppe land**. This project addresses the challenge using **open-source tools** and **public satellite imagery** to distinguish between cropland and other land types for sustainable land use monitoring.

---

3. Data Description

The dataset includes **Sentinel-1** and **Sentinel-2** time-series satellite imagery.

- Each entry includes a spatial location and corresponding multi-band reflectance values.
- The **target variable** is a binary label:
    - **1** = cropland
    - **0** = non-cropland

---

4. Methodology

- Data was preprocessed to remove irrelevant columns and merged across sensors.
- Time-series features such as **NDVI**, **SAVI**, and raw bands were used.
- Data imbalance was addressed using **SMOTE** oversampling.
- Features were scaled using **StandardScaler** or **RobustScaler**.

- A **Random Forest Classifier** was trained with hyperparameter tuning using **GridSearchCV**.
- Accuracy and F1 scores were calculated on validation data.

---

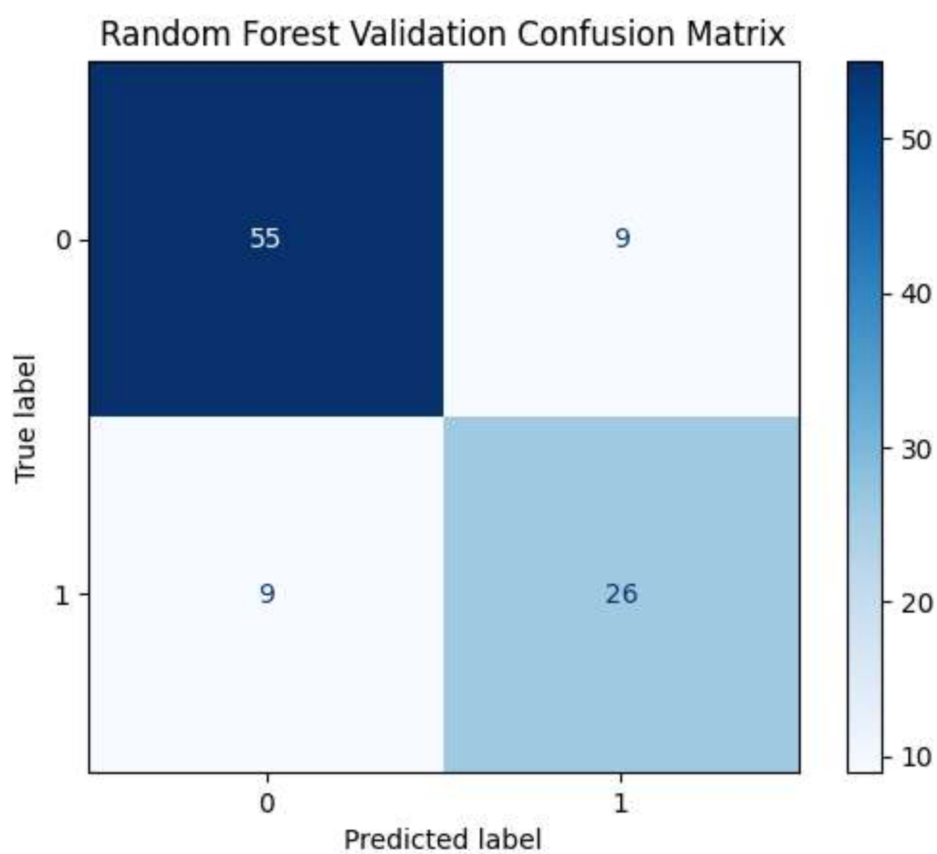
5. Results

The final Random Forest model achieved:

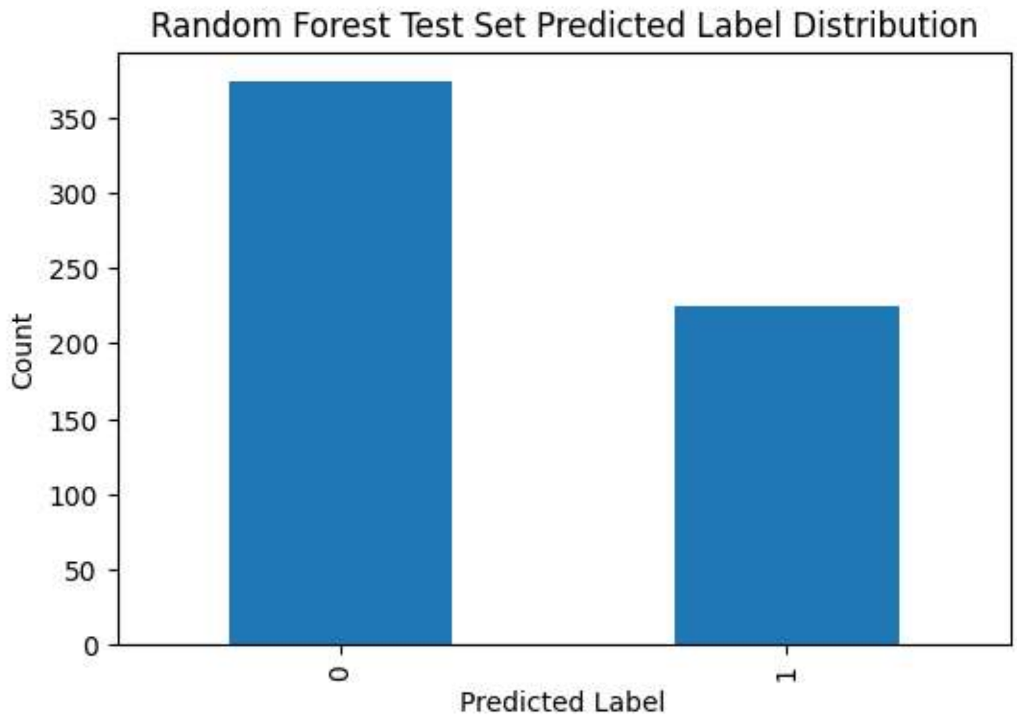
- **High classification accuracy**
- **Balanced F1 Score**

```
Random Forest Validation Accuracy: 0.8181818181818182
Random Forest F1 Score: 0.8181818181818182
```

- A well-distributed **confusion matrix** with minimal false positives and false negatives.



These results confirm the effectiveness of using ensemble models with engineered geospatial and temporal features for cropland mapping in dry regions.



6. Recommendations and Future Work

- Explore additional vegetation indices such as **EVI**, **NBR**, and **NDWI** for better representation.
- Include **weather**, **soil**, and **elevation data** to improve the temporal and spatial context.
- Experiment with **deep learning models** (e.g., LSTM, 1D CNN) to learn from longer time series.
- Improve spatial generalization by training on **diverse agro-ecological zones**.

---

7. Appendix

- **Libraries used:** scikit-learn, imbalanced-learn, geopandas, matplotlib, seaborn, pyproj
- **Environment:** Python 3.10+, runs on Google Colab or local machines (8GB RAM minimum)
- **Expected Runtime:** 5–15 minutes per model training iteration

- **Submission Format:**
  A CSV file with two columns:

ID, Target

ID_C7AV4GEJP9, 1

ID_AFVZYGLXXY, 0