# Predicting Water Pump Failures in Tanzania

## A Machine Learning Project by

## Austine Denis Otieno

## Phase3

# Overview

- Rural Tanzanians rely heavily on water pumps.
- Many pumps fail without early warnings, disrupting access to clean water.
- My goal: Use data and machine learning to predict pump failures and help decision-makers prioritize maintenance.

# Business and Data Understanding

- Dataset from Taarifa and DrivenData: 59,000plus water points.
- Problem: Predict the pump's status:  Functional, Needs Repair, or Non-Functional.
- Business value: Enables proactive maintenance, reducing outages and improving public health.

# The Data

- Features include:
  - GPS coordinates, region, installer
  - Water quality, pump type, extraction method
  - Construction year, quantity of water, etc.
- Challenges:
  - Many categorical variables
  - Imbalanced class distribution (most pumps functional)

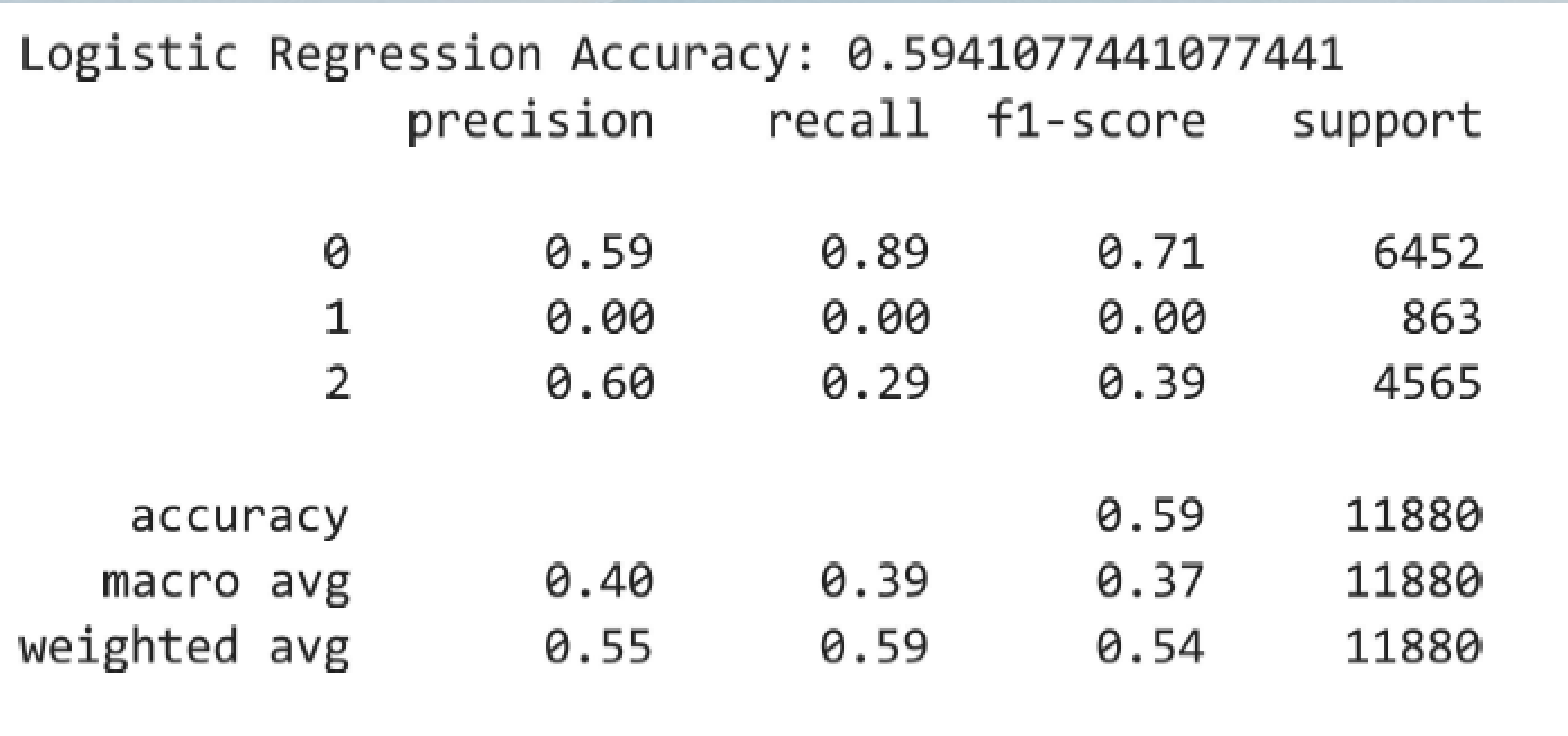
# Data Preparation

- Cleaned missing data and standardized inputs
- Combined rare categories into 'Other' to simplify patterns
- Applied SMOTE to address class imbalance
- One-hot encoding used for categorical features

# Modeling

- Tried several models: Logistic Regression, Decision Trees, Random Forest
- Random Forest performed best:
  - Handles categorical and numerical data well
  - Robust to overfitting

# Logistic regression

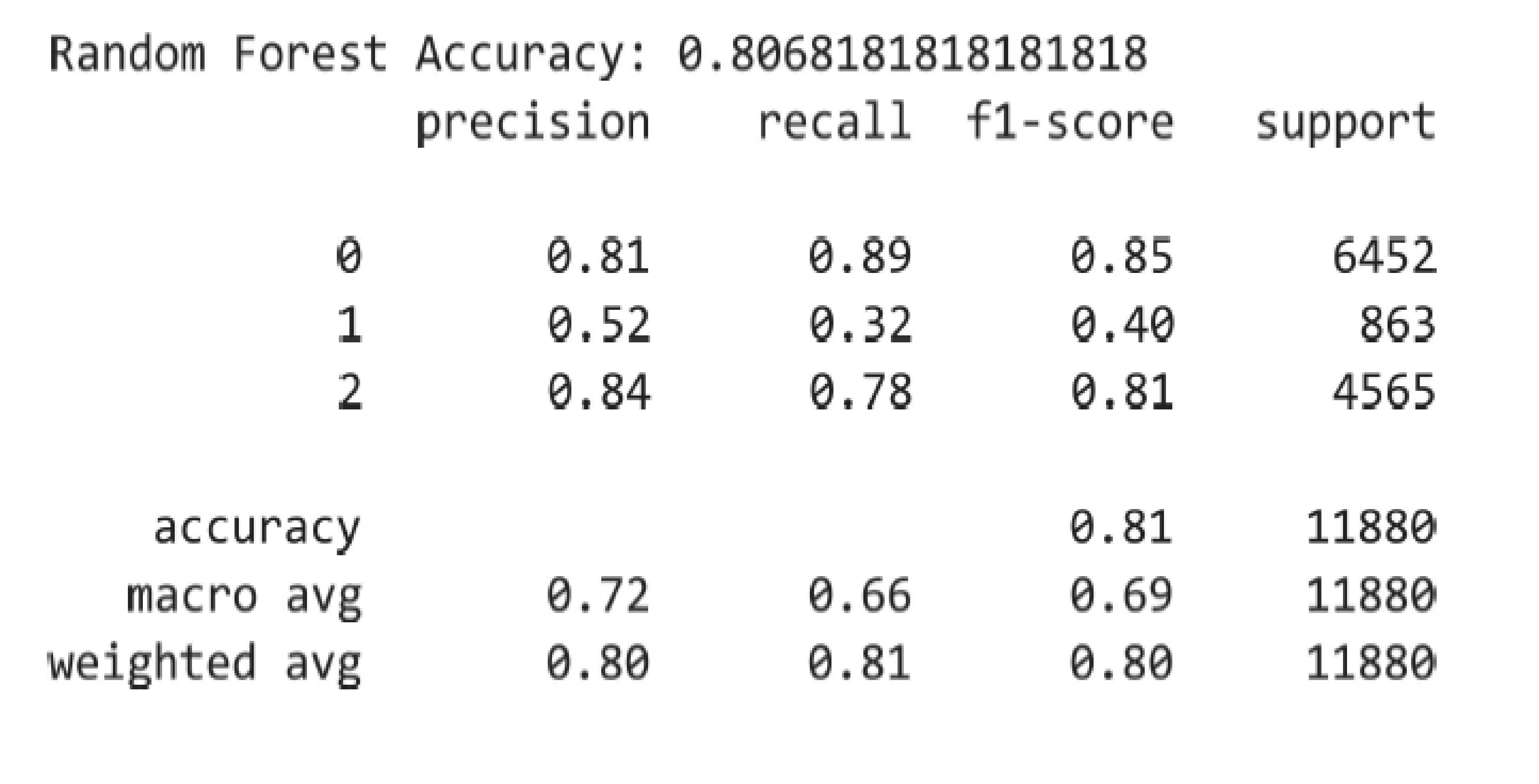


```
Logistic Regression Accuracy: 0.5941077441077441
              precision    recall  f1-score   support

           0       0.59      0.89      0.71      6452
           1       0.00      0.00      0.00       863
           2       0.60      0.29      0.39      4565

    accuracy                           0.59     11880
   macro avg       0.40      0.39      0.37     11880
weighted avg       0.55      0.59      0.54     11880
```

Class 0 has a lot more samples than class 1 → the model learns to predict class 0 often.

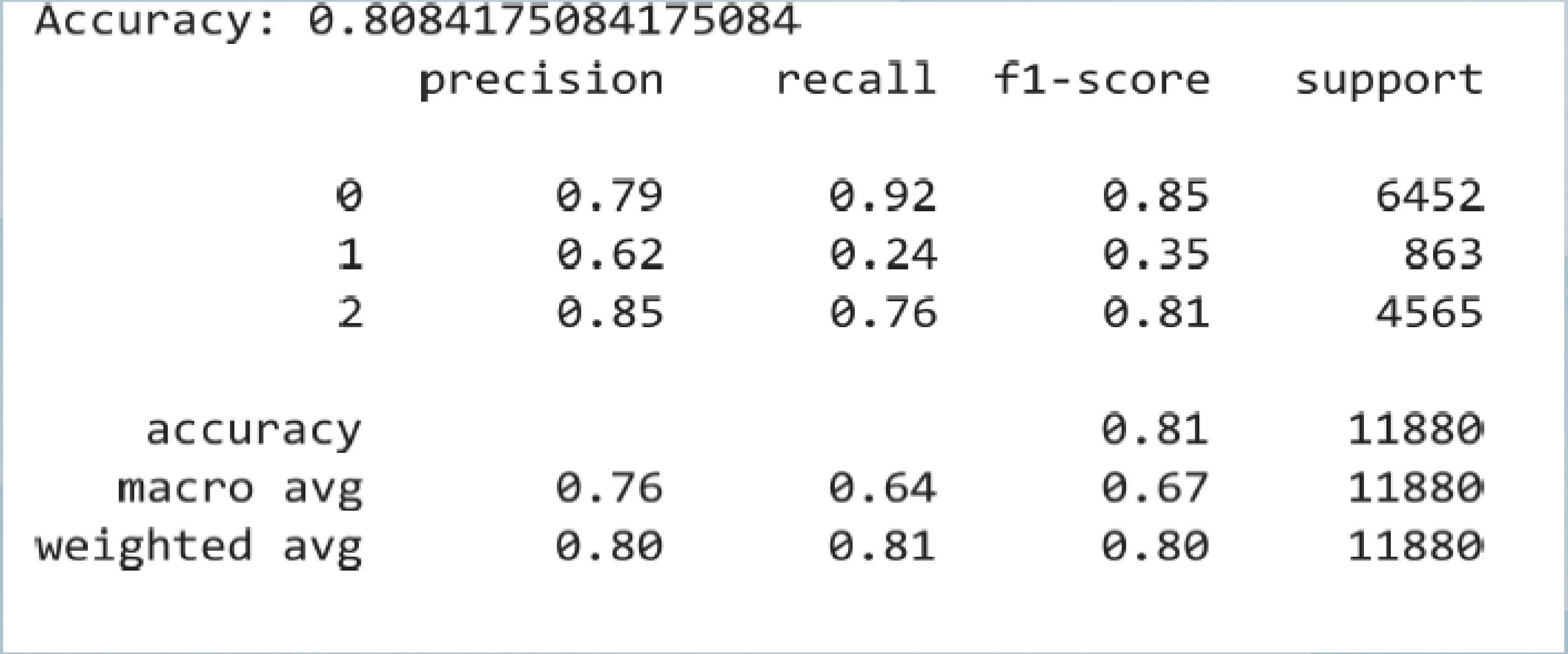


```
Random Forest Accuracy: 0.8068181818181818
              precision    recall  f1-score   support

           0       0.81      0.89      0.85      6452
           1       0.52      0.32      0.40       863
           2       0.84      0.78      0.81      4565

    accuracy                           0.81     11880
   macro avg       0.72      0.66      0.69     11880
weighted avg       0.80      0.81      0.80     11880
```

Overall accuracy is up to 80.8%—a big jump from 59%.



```
Accuracy: 0.8084175084175084
              precision    recall  f1-score   support

           0       0.79      0.92      0.85      6452
           1       0.62      0.24      0.35       863
           2       0.85      0.76      0.81      4565

    accuracy                           0.81     11880
   macro avg       0.76      0.64      0.67     11880
weighted avg       0.80      0.81      0.80     11880
```

The model strongly predicts class 0 and class 2, which together make up the majority of the dataset.

# Evaluation

- **Used Accuracy, Precision, Recall, F1-Score**
- **Final Accuracy: 81%**
- **High performance on Functional and Non-Functional classes**
- **Moderate performance on Needs Repair (due to few examples)**

# Interpretation of Results

- **Class 0 (Functional):** Most predictions accurate
- **Class 1 (Needs Repair):** Often confused with other classes
- **Class 2 (Non-functional):** High recall and precision
- Indicates value for prioritizing emergency interventions

# Recommendations:

- Use model to highlight pumps at high risk of failure
- Improve field data quality and consistency
- Update model regularly with new data
- Train local staff to interpret and act on predictions

# Next Steps

- Integrate model with mobile or web-based reporting tools
- Collaborate with government agencies for deployment
- Explore use of satellite or weather data to enhance predictions

# THANK YOU