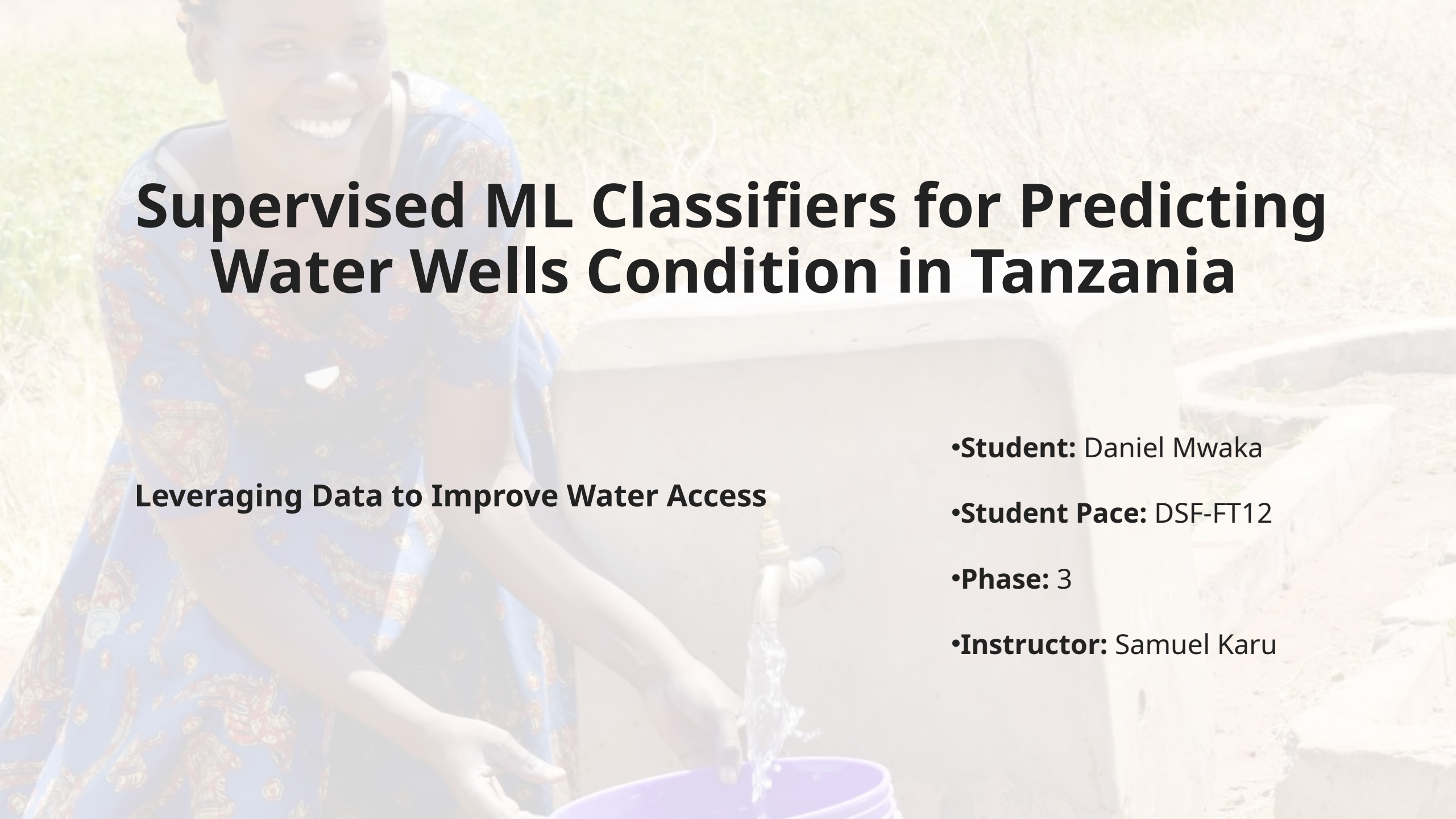


Supervised ML Classifiers for Predicting Water Wells Condition in Tanzania

Leveraging Data to Improve Water Access

- **Student:** Daniel Mwaka
- **Student Pace:** DSF-FT12
- **Phase:** 3
- **Instructor:** Samuel Karu

Problem Statement

- Access to clean water is a critical challenge in Tanzania.
- Many established water points are in disrepair or non-functional, impacting millions leading to water scarcity.
- The lack of an effective, data-driven framework for predicting a well's functional status is a leverageable improvement area.
- **The Goal:** To recommend an evidence-based supervised ML classification model for predicting the functional condition of water wells in Tanzania.

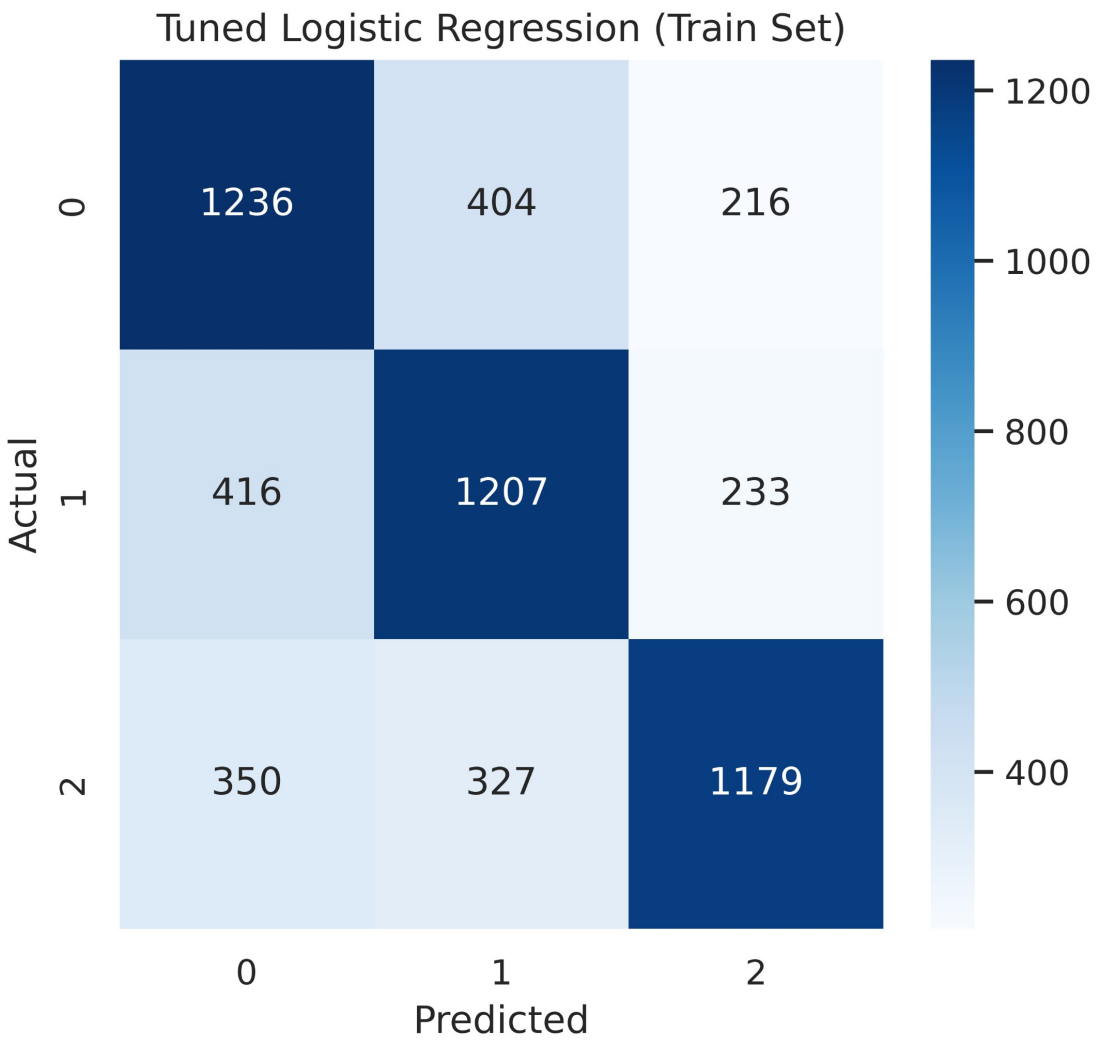
Business Understanding

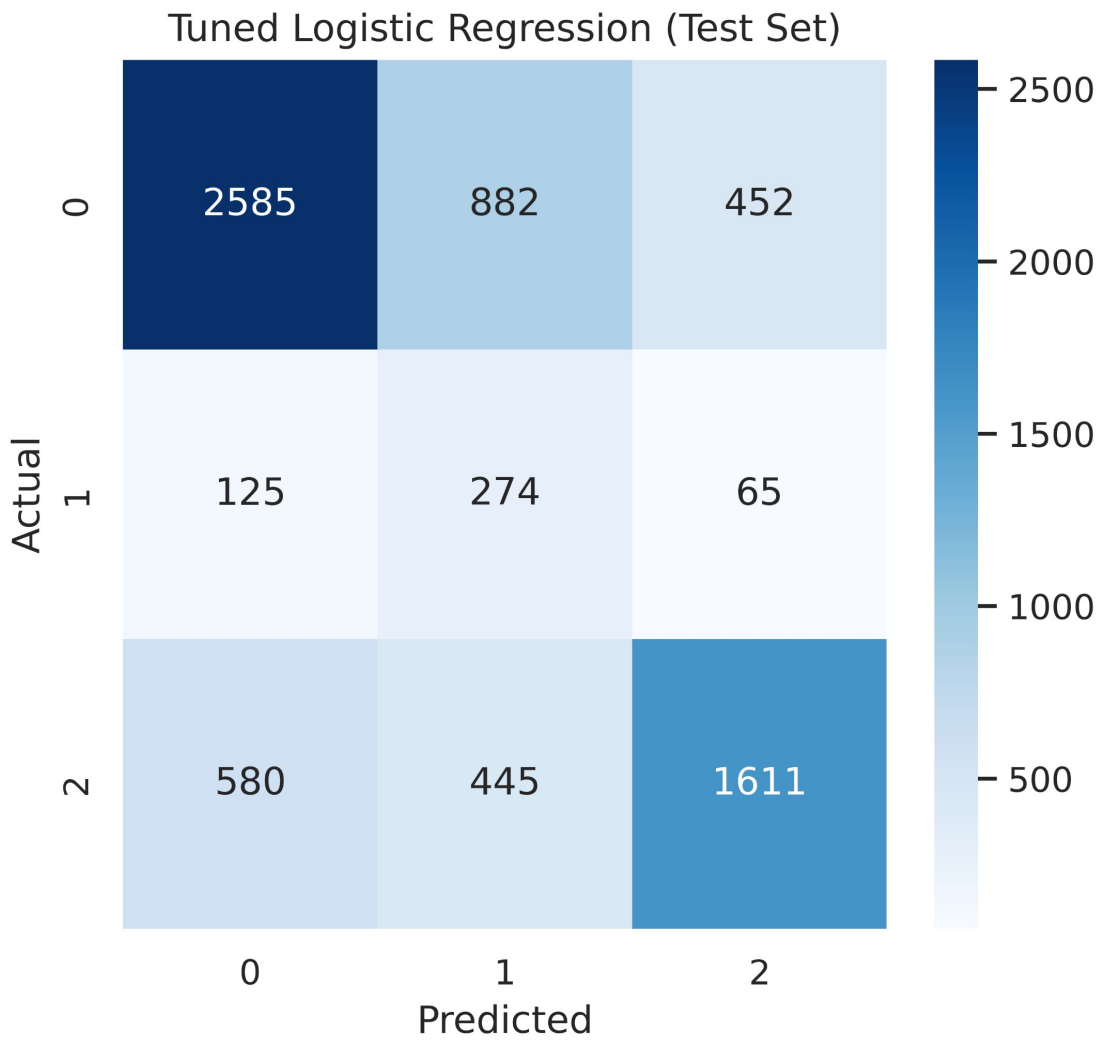
- **Business Problem:** Predicting the condition of water wells can enable stakeholders to prioritize maintenance, and allocate resources efficiently to support water security.
- **Core Question:** Can the operational status of water-wells be accurately predicted using available data and machine learning models?
- **Project Scope:**
  - Examine features related to a Water-well's function to identify potential predictor variables.
  - Build multiple classifiers, tune them, evaluate, and compare their respective prediction performance to determine the best fit model.
  - Verify the reliability of the best-fit model by using it to predict the target variable for a previously unseen dataset (testdata.csv).
  - Deduce feasible business recommendations and viable next steps.

Data Understanding

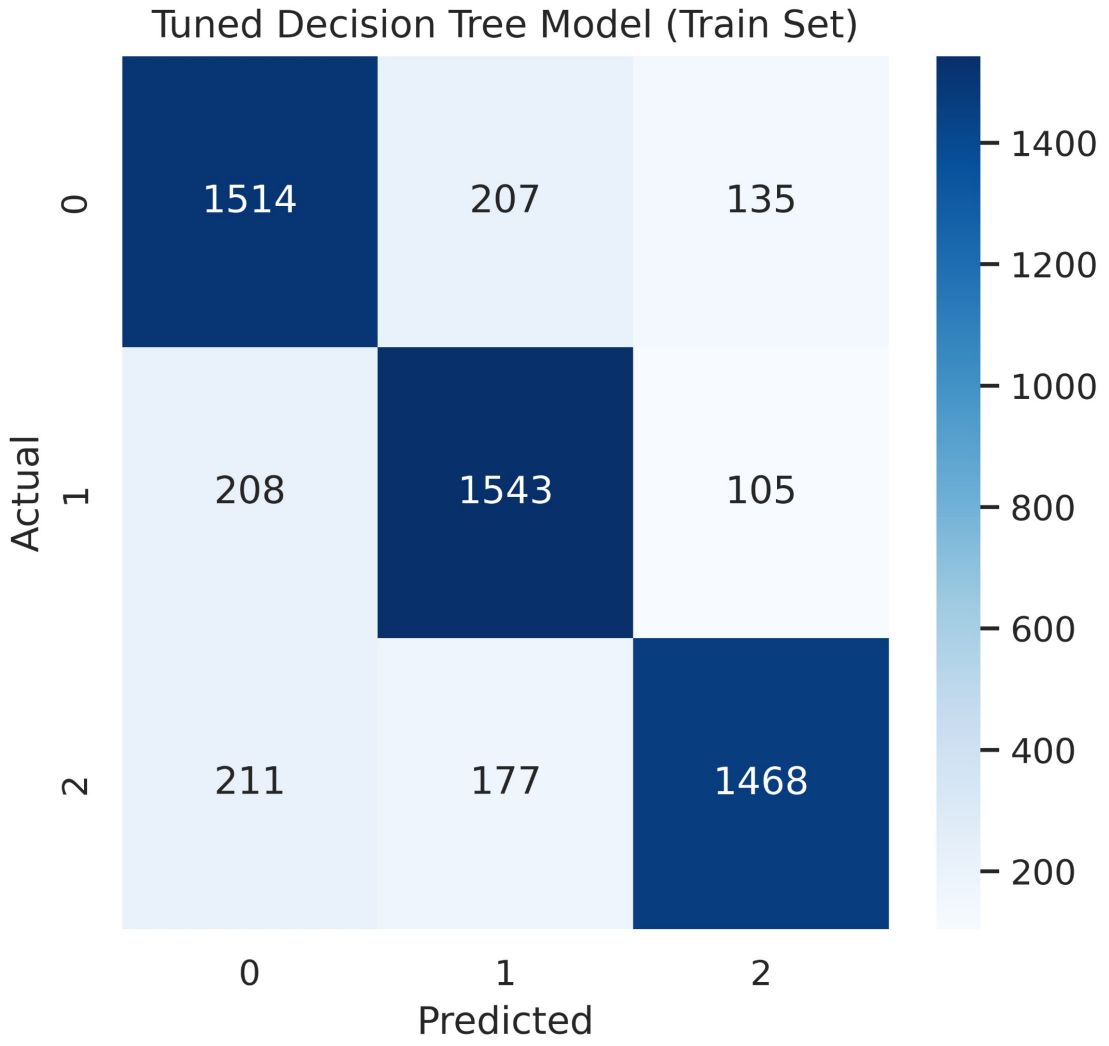
- **Datasets Source:** Kaggle competition dataset on Tanzanian Water Wells from https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/data/
- **Datasets:**
  - trainingset.csv (40 columns and 59,400 entries).
  - trainingsetlabels.csv (2 columns and 59,400 entries).
  - Testdata.csv (40 columns and 14,850 entries).
- **Target Variable:** Water-well Status (Functional, Functional needs repair, or Non-functional).
- **Selected Predictor Features:** gps height, population, well age, basin, region, permit, extraction type, management group, payment type, water quality, water quantity, source type, and water-point type.

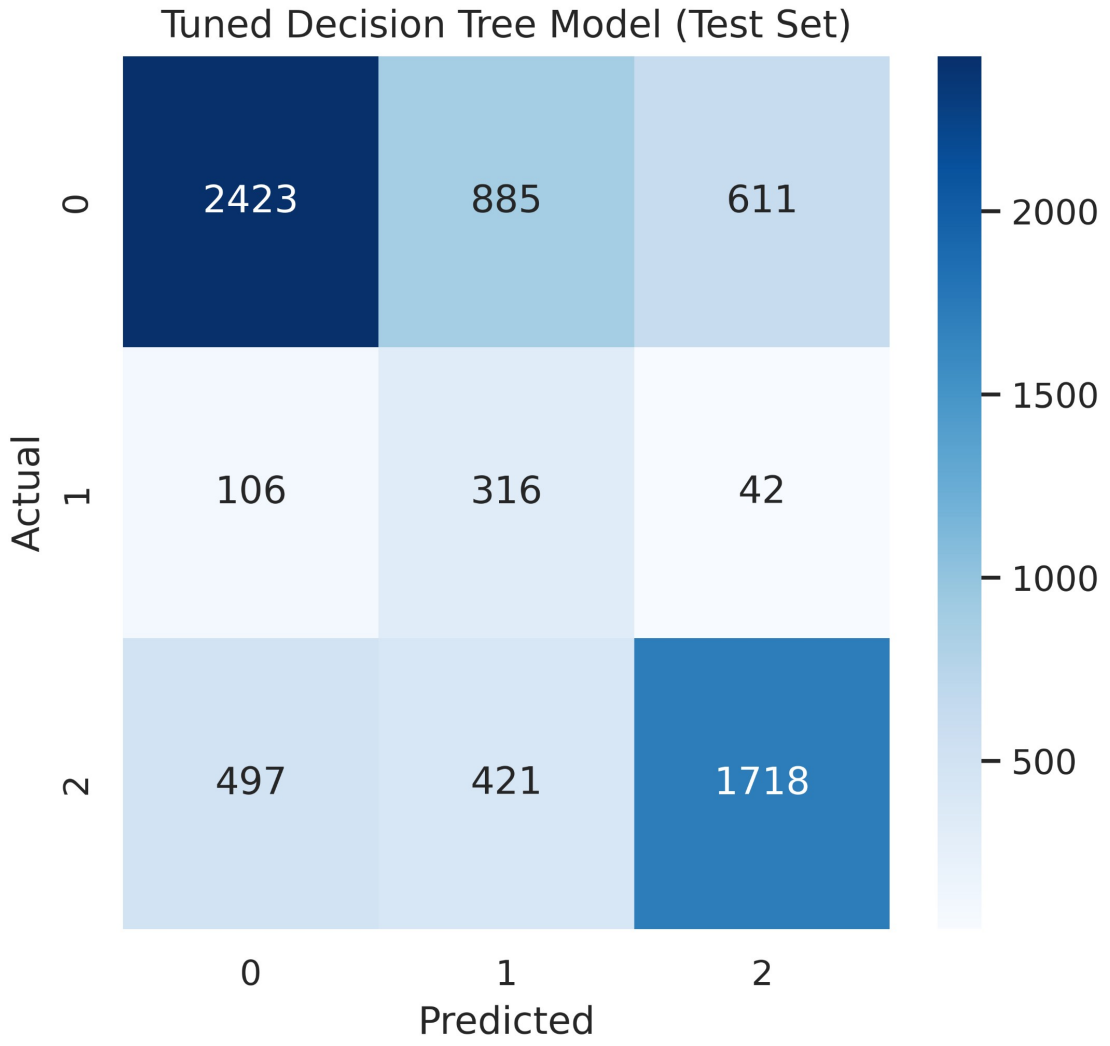
Model Evaluation: Logistic Regression Model



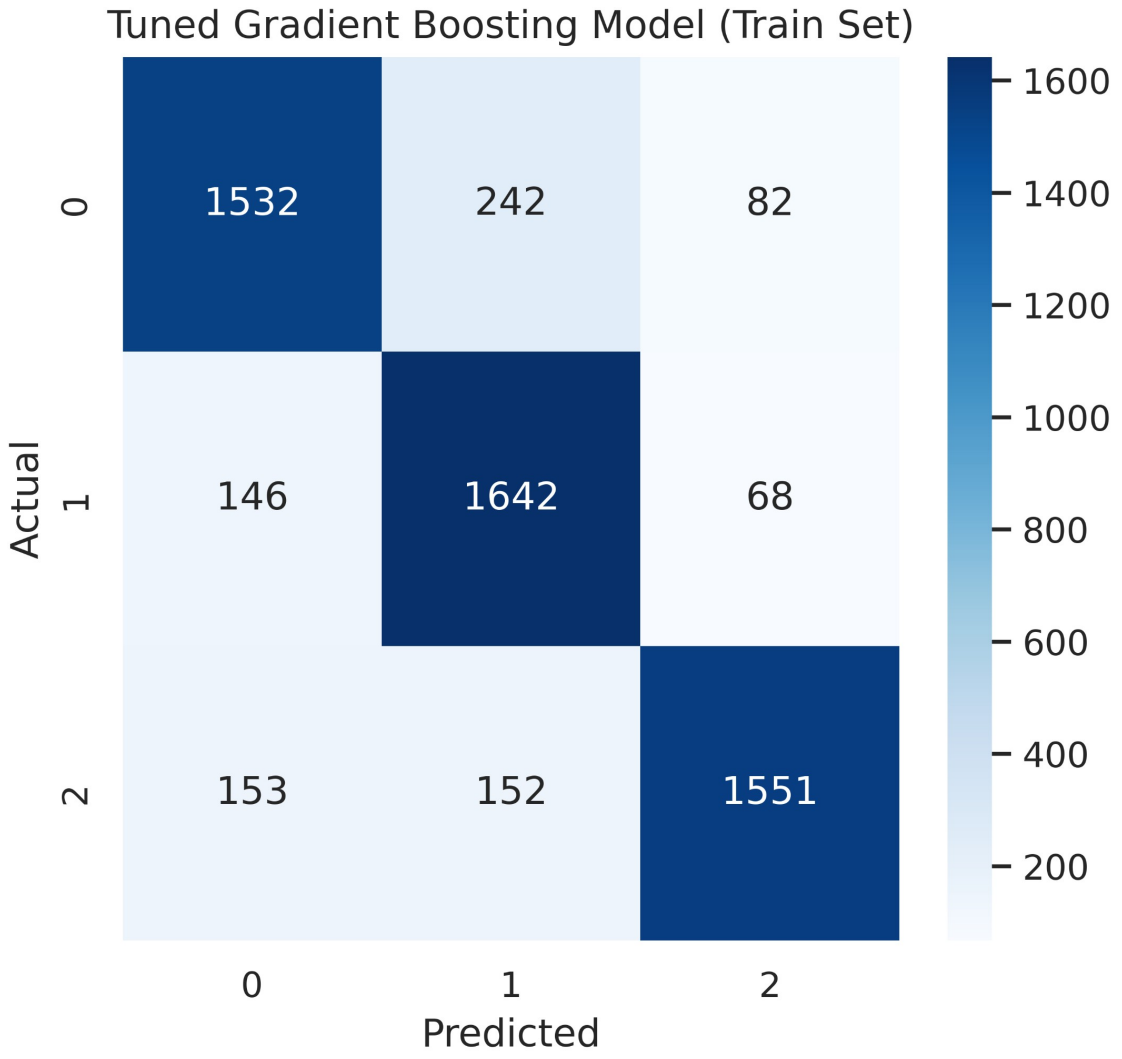


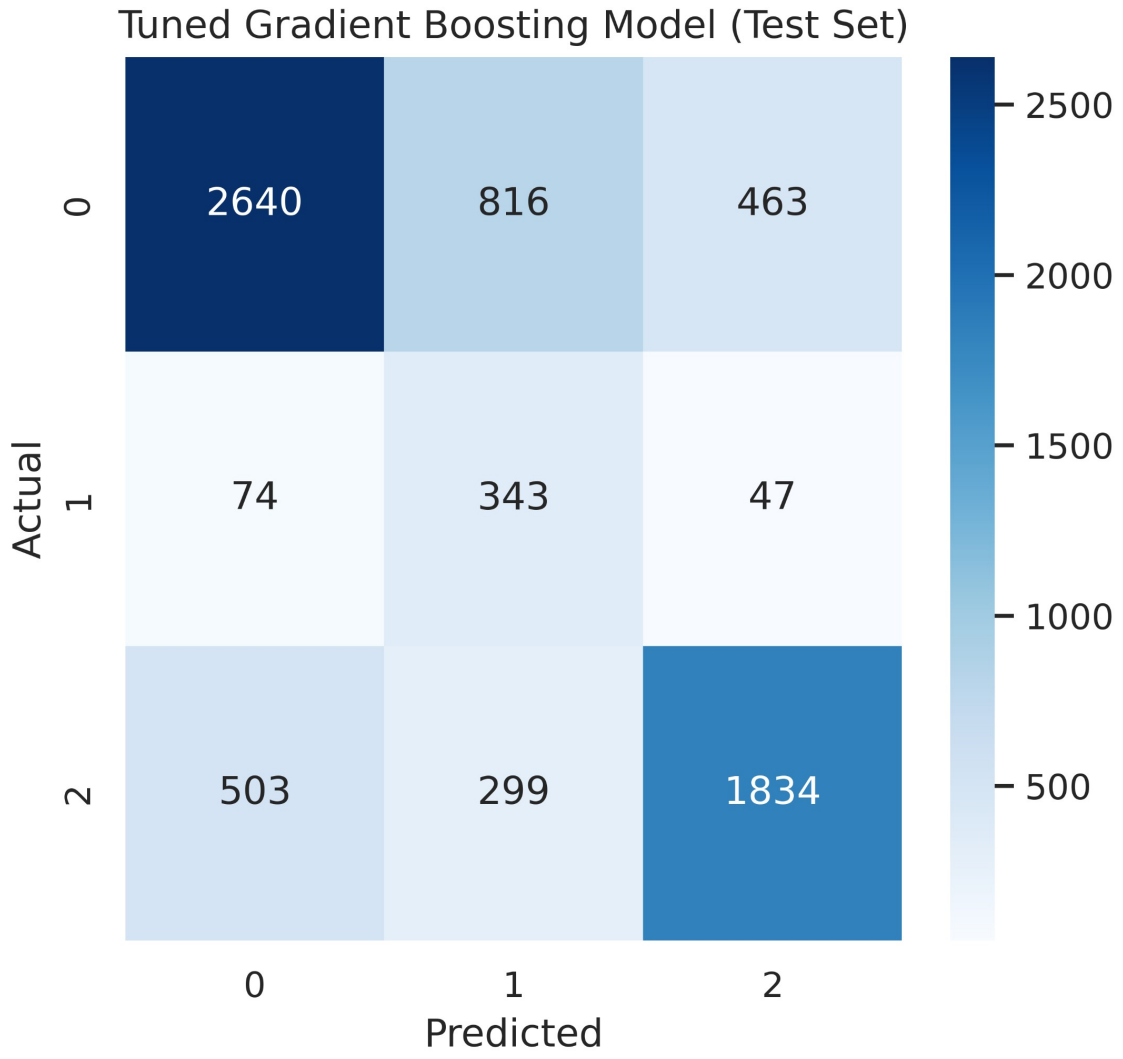
Model Evaluation: Decision Tree Classifier





Model Evaluation: Gradient Boosting Classifier





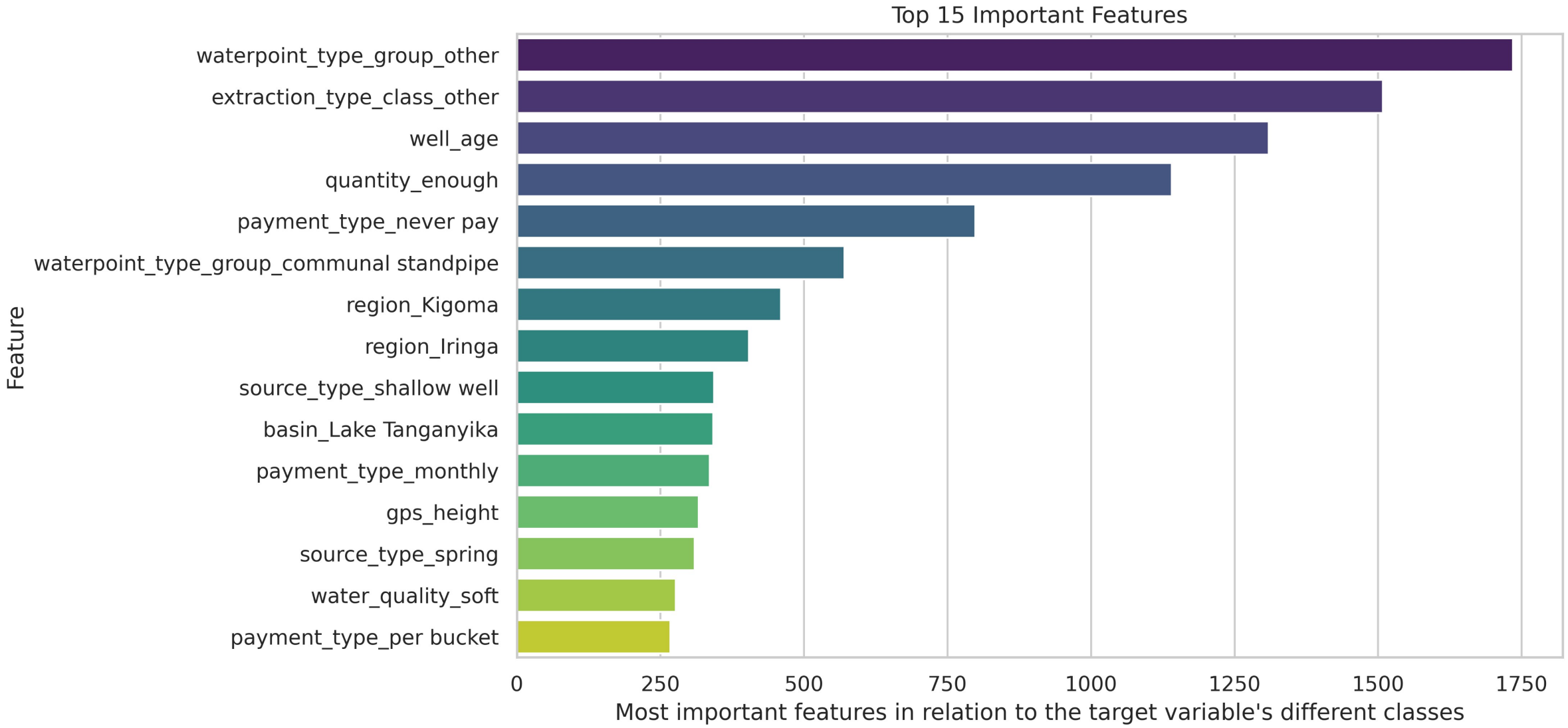
Model Performance Comparison

	Model	Train Accuracy	Test Accuracy	Train Precision	Test Precision	Train Recall	Test Recall	Train F1-score	Test F1-score	Train ROC-AUC	Test ROC-AUC
0	Logistic Regression (Untuned)	0.649425	0.640120	0.653538	0.735050	0.649425	0.640120	0.650206	0.674341	0.831456	0.816311
1	Logistic Regression (Tuned)	0.650503	0.636843	0.654797	0.734322	0.650503	0.636843	0.651274	0.671957	0.832015	0.816505
2	Decision Tree (Untuned)	0.999461	0.625445	0.999462	0.727775	0.999461	0.625445	0.999461	0.657666	1.000000	0.731613
3	Decision Tree (Tuned)	0.812680	0.634991	0.814483	0.732079	0.812680	0.634991	0.812901	0.667340	0.953505	0.801969
4	Gradient Boosting (Untuned)	0.720366	0.668329	0.724715	0.754088	0.720366	0.668329	0.721155	0.698047	0.879053	0.843328
5	Gradient Boosting (Tuned)	0.848599	0.686280	0.851667	0.767590	0.848599	0.686280	0.848964	0.713329	0.958991	0.866277

Selected Model for Deployment

- The ***tuned Gradient Boosting Classifier*** is selected for deployment.
- Reasons for Selection:
  - Consistently delivered superior predictive performance.
  - Balanced precision and recall across all classes.
  - Demonstrated strong generalization to unseen data.
  - Small gap between train and test performance metrics indicates robustness and minimal overfitting.
- This model is the most reliable, effective, and best-choice for predicting the status of water wells in Tanzania.

Feature Importance Insights



Business Recommendations

- **Prioritize Maintenance:** Use model predictions to identify high-risk wells and allocate maintenance resources efficiently.
- **Improve Data Collection:** Enhance data quality, particularly for key features influencing well condition, such as management, payment types, and environmental factors.
- **Stakeholder Engagement:** Share insights with local authorities and NGOs to inform decision-making, schedule maintenance routines, and find patterns on factors impacting long-term functionality.
- **Inform New Infrastructure:** Use insights on factors contributing to failure to inform the design and construction frameworks for new groundwater projects.

Next Steps

- **Model Deployment:** Integrate the Tuned Gradient Boosting Classifier into a user-friendly dashboard for real-time predictions.
- **Integration into Planning:** Use model predictions to optimize maintenance schedules and target interventions.
- **Pilot Targeted Interventions:** Use the model to pilot interventions in high-risk areas/well types and measure the impact.
- **Feature Expansion:** Incorporate additional data sources (e.g., weather, usage patterns) to enhance model accuracy and update the mode regularly with new data to maintain performance and relevance.