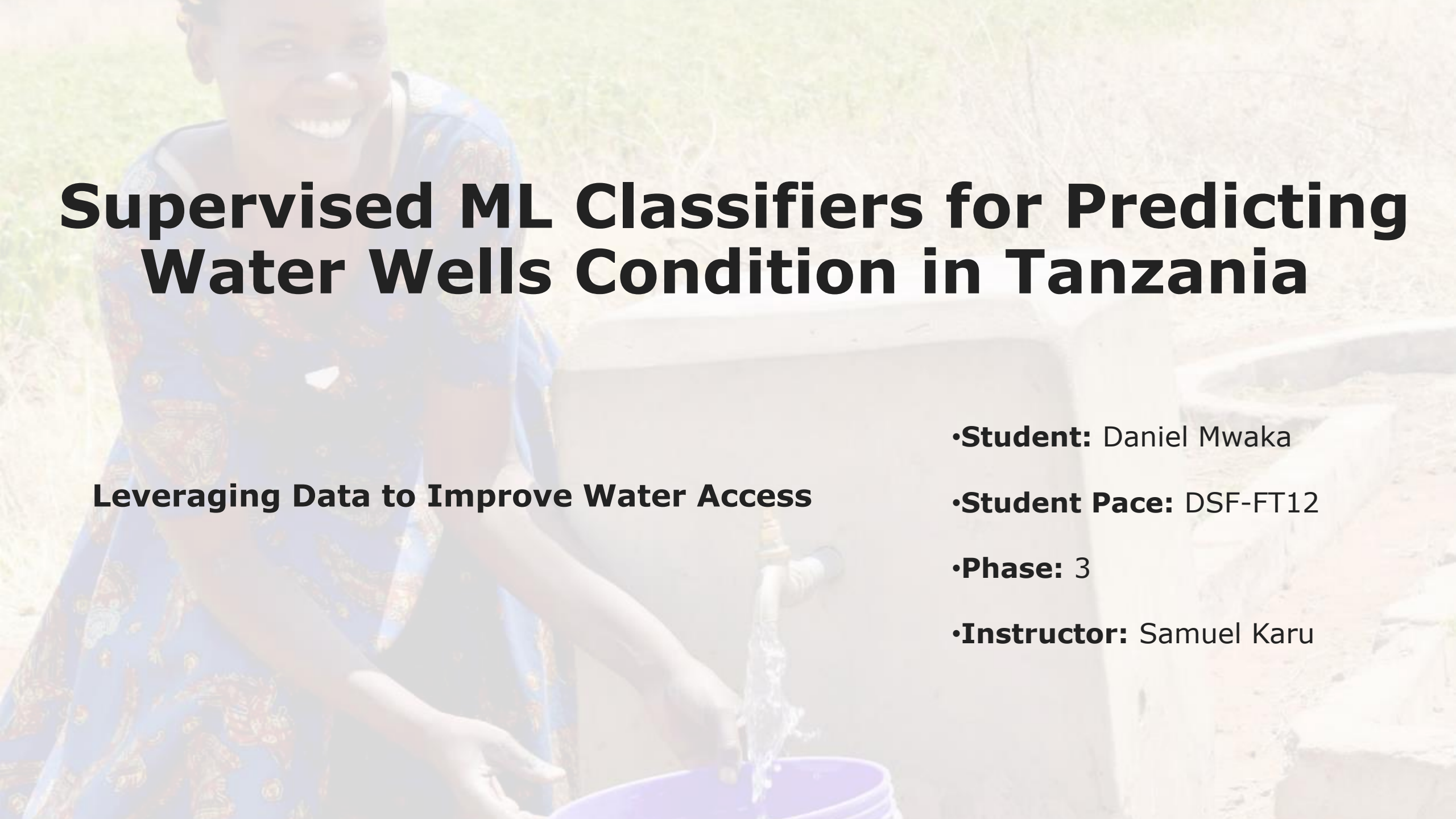


Supervised ML Classifiers for Predicting Water Wells Condition in Tanzania

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.

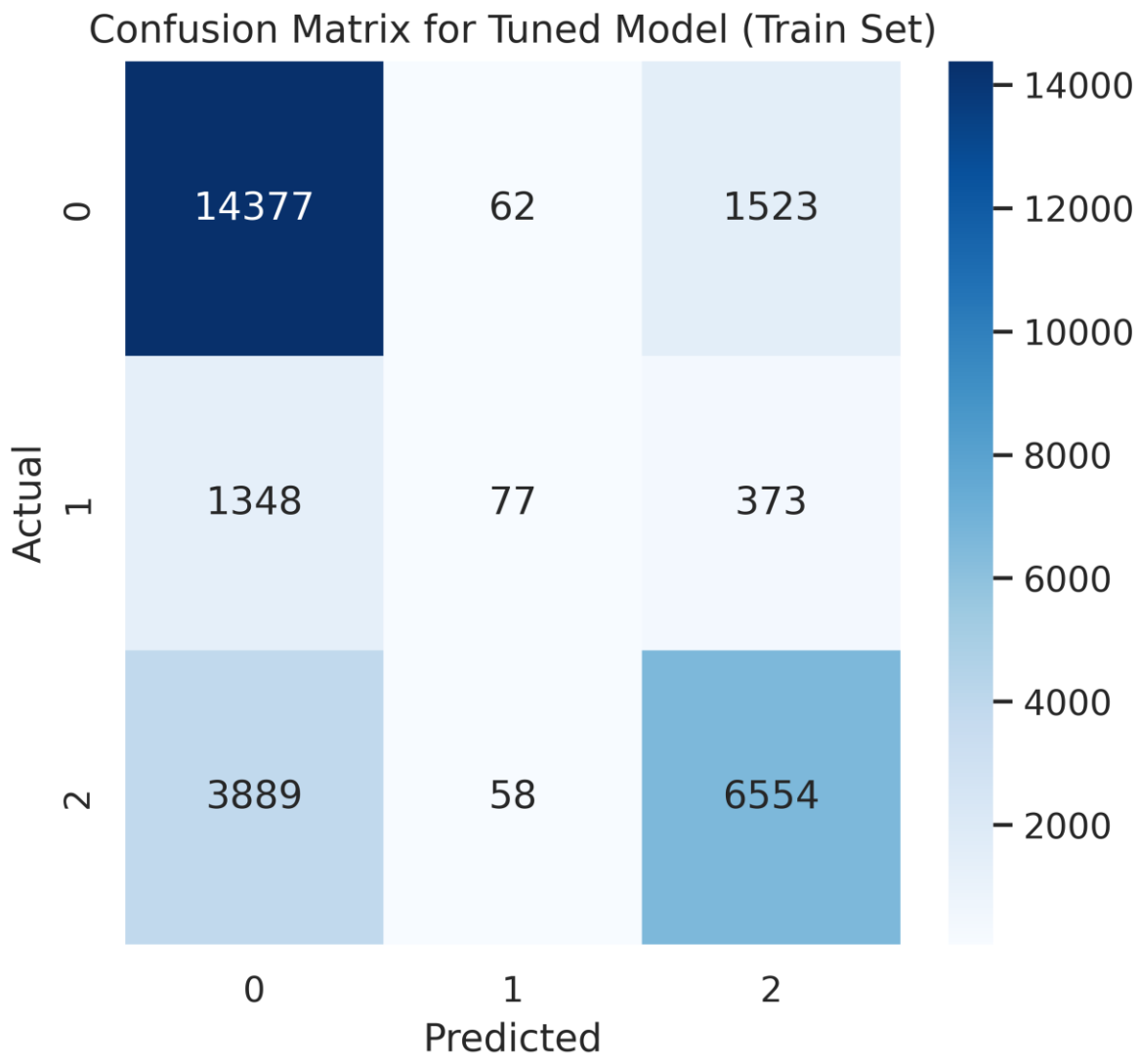
Business Understanding

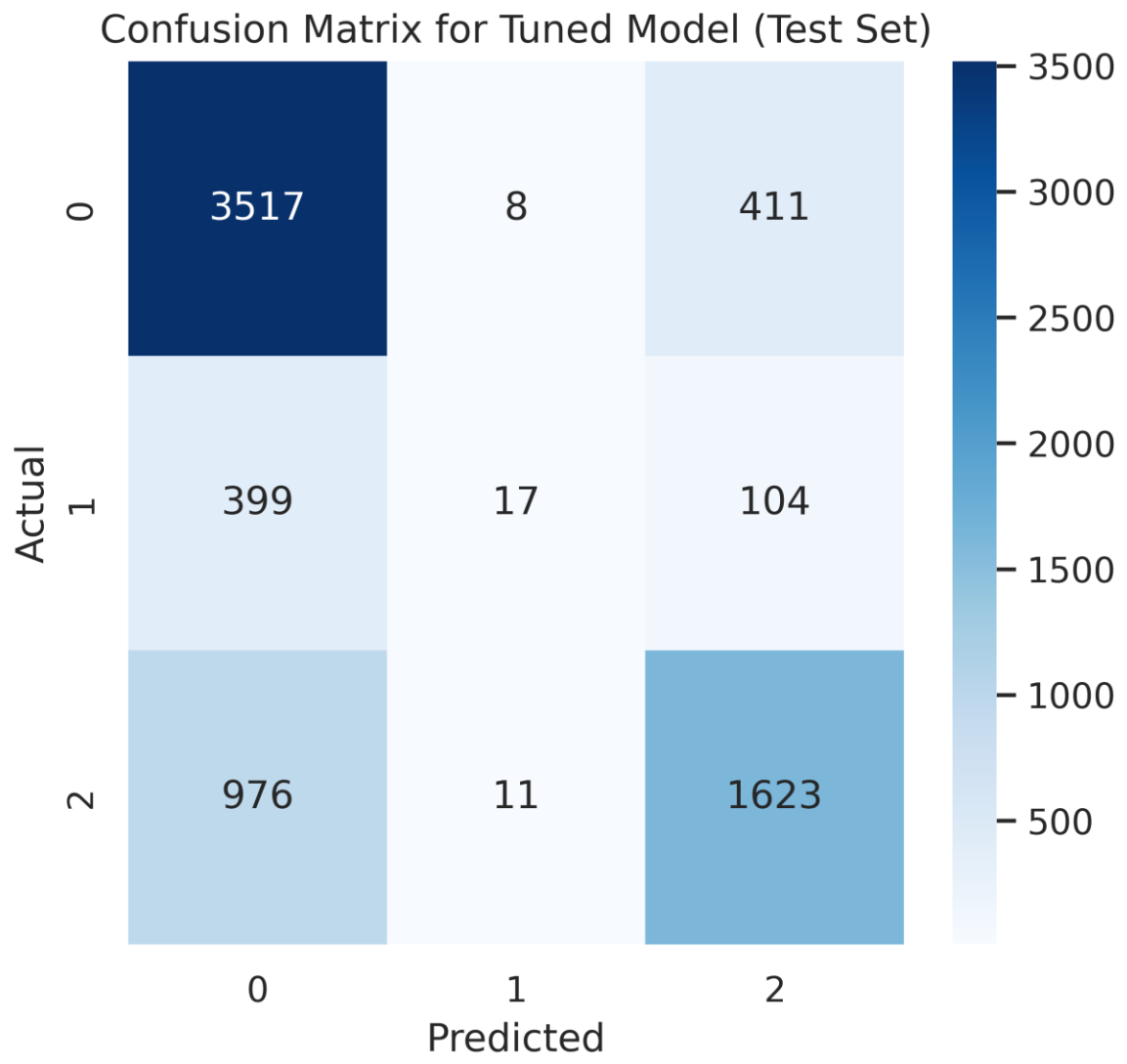
- **Business Problem:** Predicting the condition of water wells can enable stakeholders to prioritize maintenance, and allocate resources efficiently to support water security.
- **Core Question:** Can the operational status of water-wells be accurately predicted using available data and machine learning 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 deploying it to predict the target variable for a previously unseen dataset.
  - Deduce feasible business recommendations and viable next steps.

Data Understanding

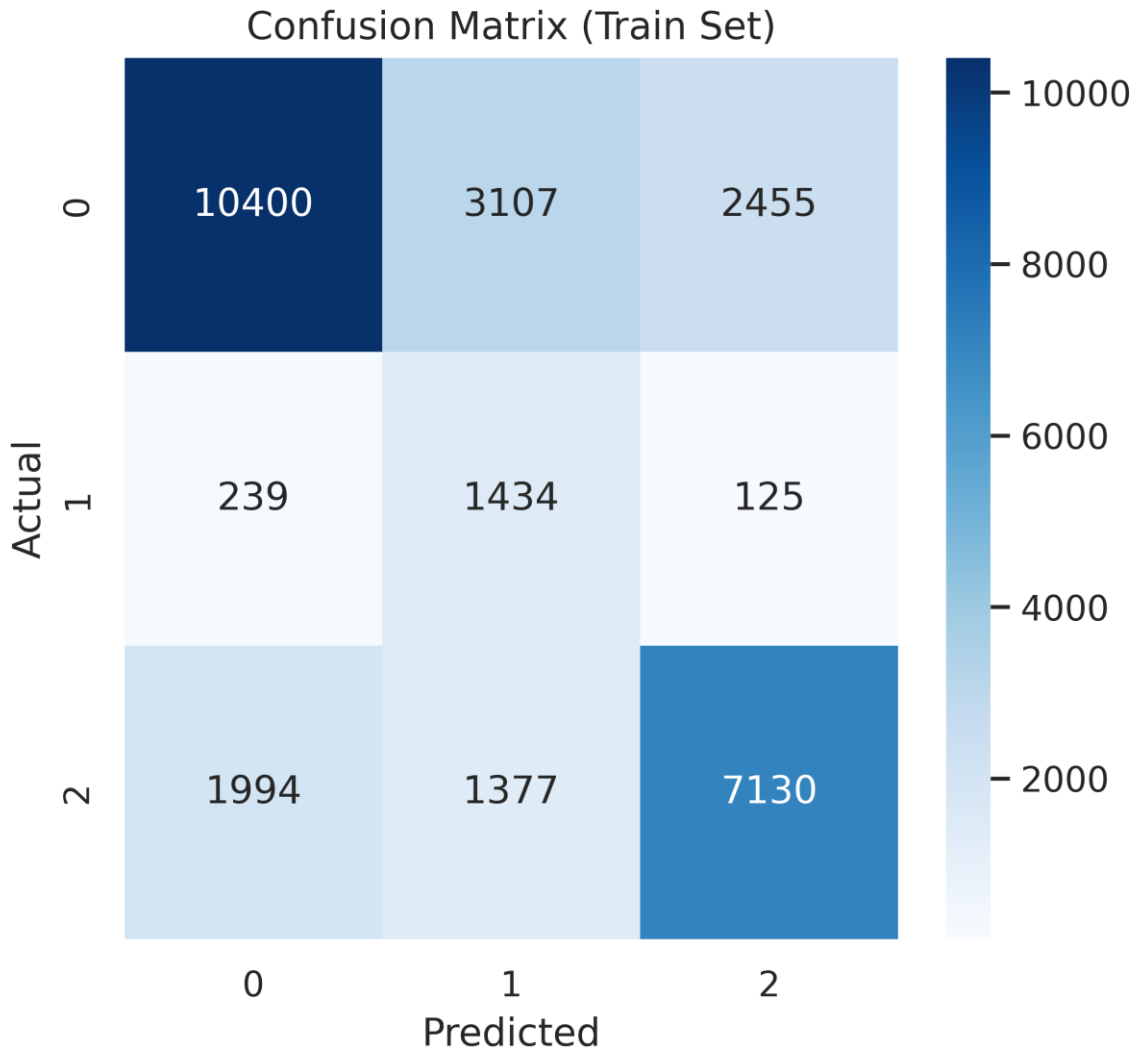
- **Dataset Source:** Kaggle competition dataset on Tanzanian Water Wells from https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/data/
- **Dataset Features:** training-set (41 columns and 59400 entries), test-data (40 columns and 4874 entries).
- **Target Variable:** Water-well Status (Functional, Function but in need of 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.

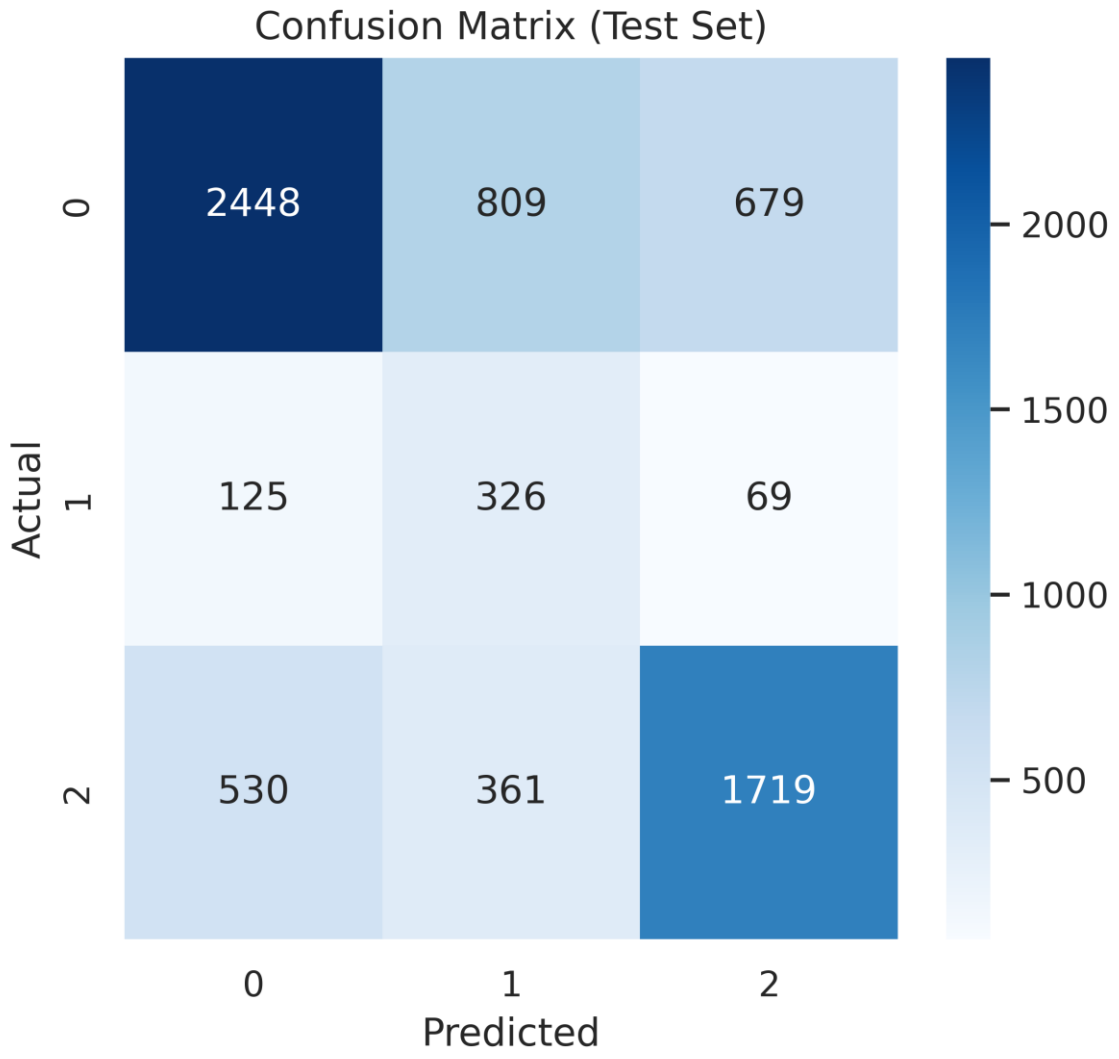
Model Evaluation: Logistic Regression Model



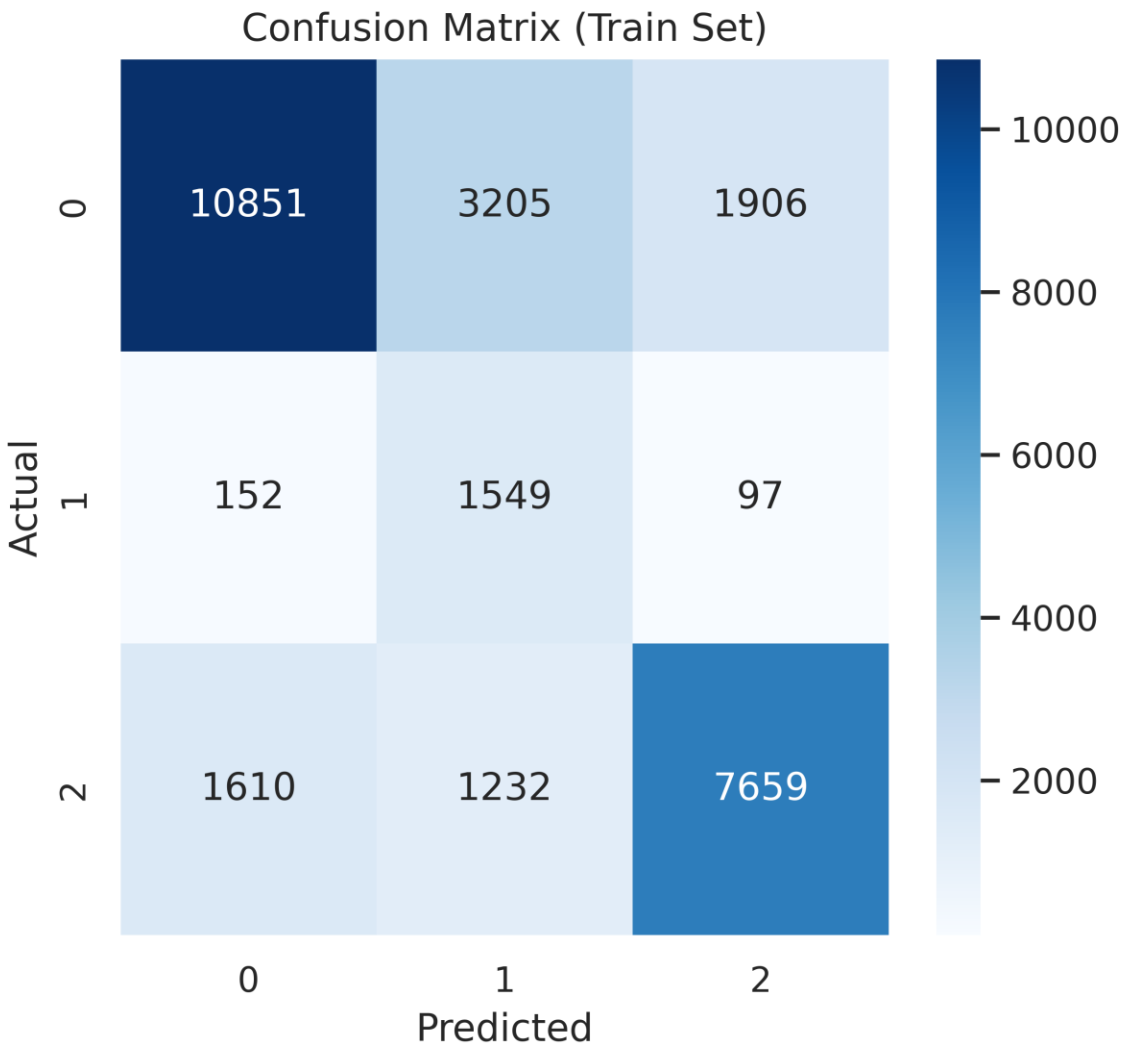


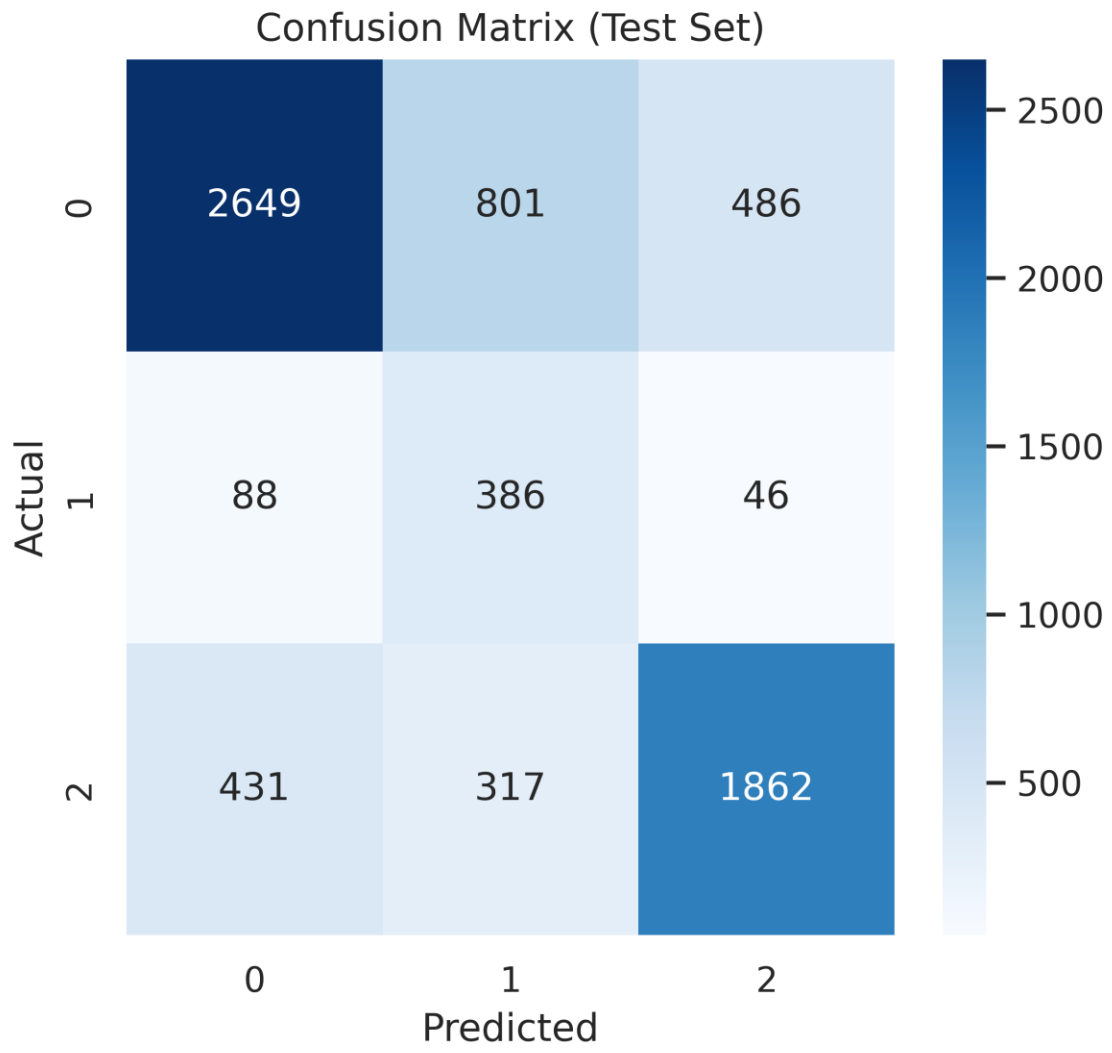
Model Evaluation: Decision Tree Classifier





Model Evaluation: Gradient Boosting Classifier





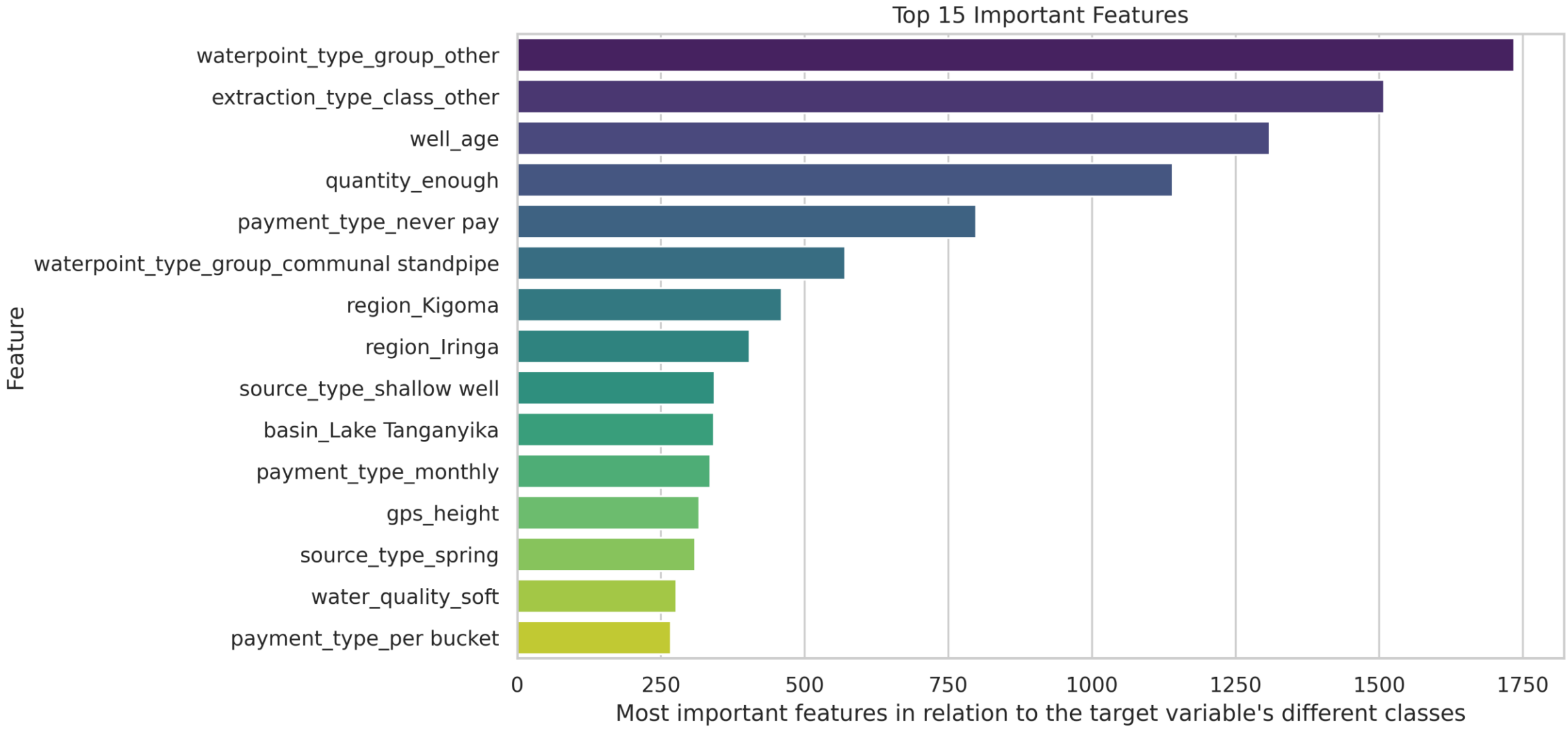
Model Performance Comparison

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
Gradient Boosting (Tuned)	0.709777	0.693037	0.796911	0.771956	0.709777	0.693037	0.736564	0.718383	0.893645	0.862429
Logistic Regression (Tuned)	0.743357	0.729833	0.727068	0.715619	0.743357	0.729833	0.718420	0.700860	0.832416	0.828979
Gradient Boosting (Untuned)	0.674463	0.672658	0.771189	0.755260	0.674463	0.672658	0.706673	0.700170	0.856597	0.844373
Decision Tree (Tuned)	0.671031	0.635862	0.753231	0.712868	0.671031	0.635862	0.696653	0.661377	0.844391	0.786505
Logistic Regression (Untuned)	0.645377	0.648316	0.738565	0.727730	0.645377	0.648316	0.678582	0.676042	0.822257	0.820428
Decision Tree (Untuned)	0.687945	0.620011	0.777609	0.706562	0.687945	0.620011	0.710038	0.646578	0.816522	0.719962

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

- **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.
- **Continuous Improvement:** Regularly update the model with new data to maintain performance and relevance.