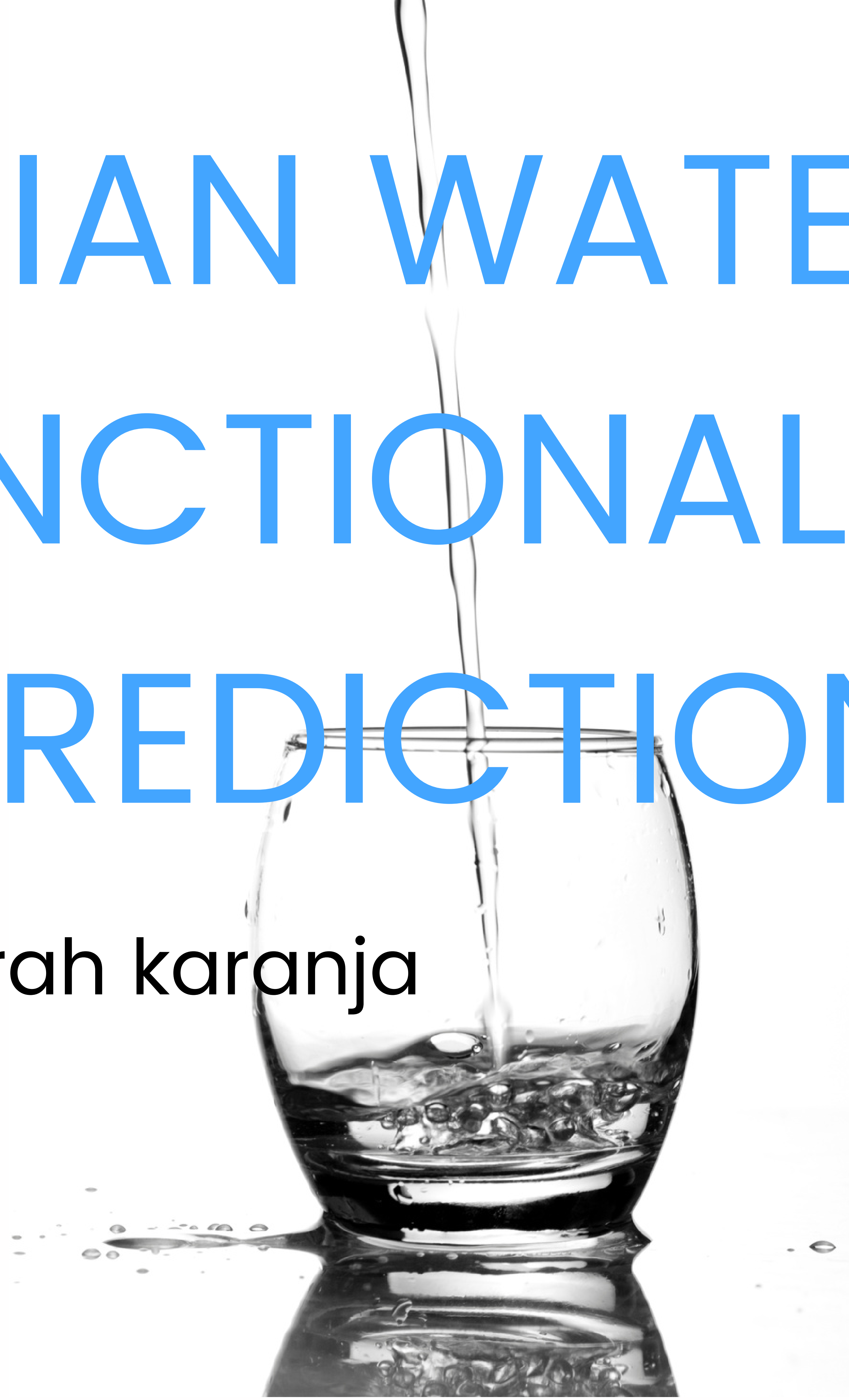
TANZANIAN WATER PUMP FUNCTIONALITY PREDICTION



Presented by Sarah karanja

Overview

- Problem: Faulty water pumps leading to water shortages and health risks.
- Objective: Develop a predictive model to identify faulty water pumps.
- Features: Pump age, location, type, condition.
- Impact: Prioritize maintenance to ensure clean and safe water for all Tanzanians.

Business Understanding

- Stakeholder: Ministry of Water in Tanzania.
- Responsibility: Ensure water supply systems are well-maintained.
- Opportunity: Use predictive modeling to allocate maintenance resources efficiently.
- Outcome: Improved reliability and accessibility of clean water for Tanzanian communities.

Modeling Approach

Objective: Predict the functionality of water pumps based on historical data.

Modeling Strategy:

- Baseline Model: Decision Trees
- Alternative Model: Random Forests

Benefits:

- Accurate predictions for timely maintenance.
- Generalization to unseen data.

Baseline Model - Decision Trees

Training and Prediction:

- Features extracted from training data.
- Test data aligned with training data columns.

Model Fitting:

- Decision Tree classifier trained on training data.

Cross-Validation:

- Model evaluated using 5-fold cross-validation.
- Accuracy used as evaluation metric.

Baseline Model Evaluation

Performance Metrics:

- Mean Cross-Validation Accuracy: ~0.775
- Standard Deviation: ~0.002

Interpretation:

- Decent accuracy in predicting water pump functionality.

Alternative Model - Random Forests

Model Overview:
Ensemble of decision trees.
Aggregation reduces overfitting.

Benefits:
Robust predictions.
Effective with high-dimensional data.

Random Forest Classifier

Model Overview:

Extension of decision trees.

Ensemble learning reduces overfitting.

Performance Metrics:

Mean CV Accuracy: ~81.76%

Standard Deviation: ~0.0028

Test Set Predictions:

Predicted labels for test set: functional/non-functional.

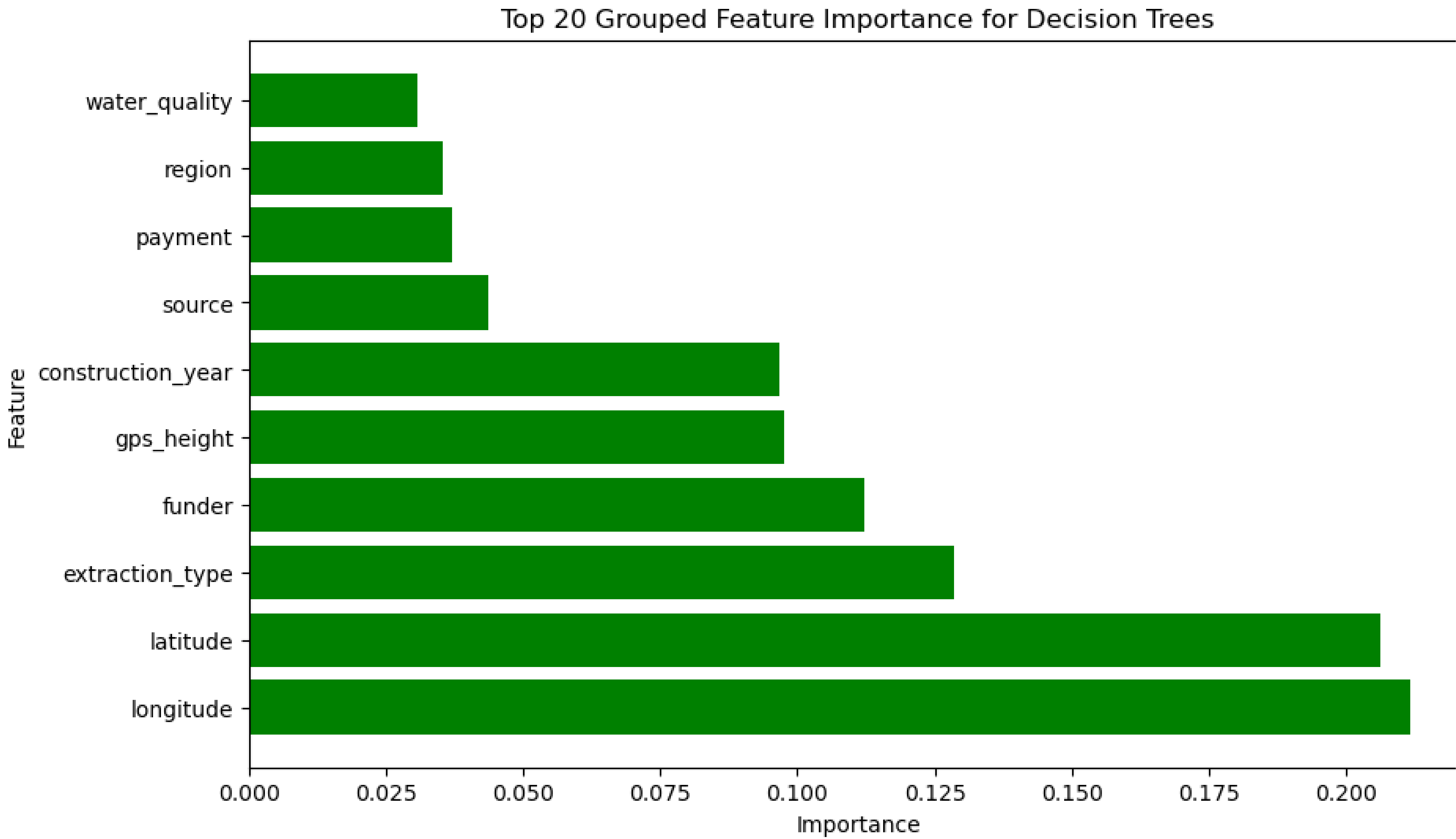
Feature Importance Analysis

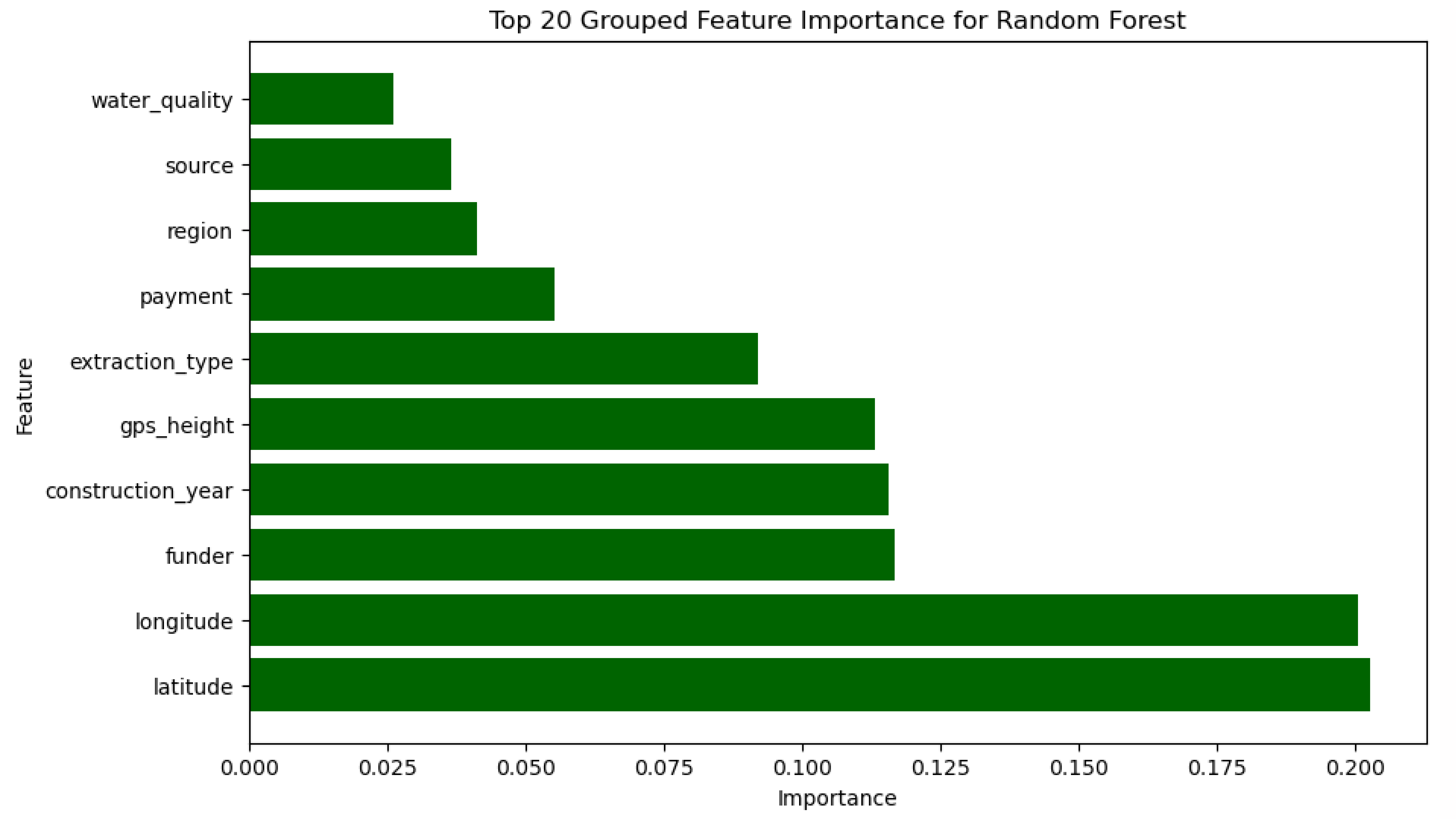
Decision Trees:

Extraction Type (0.45): Highest importance.
Payment (0.12): Second most important.
Construction Year (0.12): Age impacts functionality.

Random Forest:

Latitude (0.2) and Longitude (0.2): Strong spatial patterns.
Funder (0.125): Funding organization's impact.
Construction Year (0.125): Year of construction significant.


Top 20 Grouped Feature Importance for Decision Trees
water_quality
region
payment
source
construction_year
gps_height
funder
extraction_type
latitude
longitude
Feature
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Importance



Top 20 Grouped Feature Importance for Random Forest
water_quality
source
region
payment
extraction_type
gps_height
construction_year
funder
longitude
latitude
Feature
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Importance

Model Visualization results

Decision Trees:

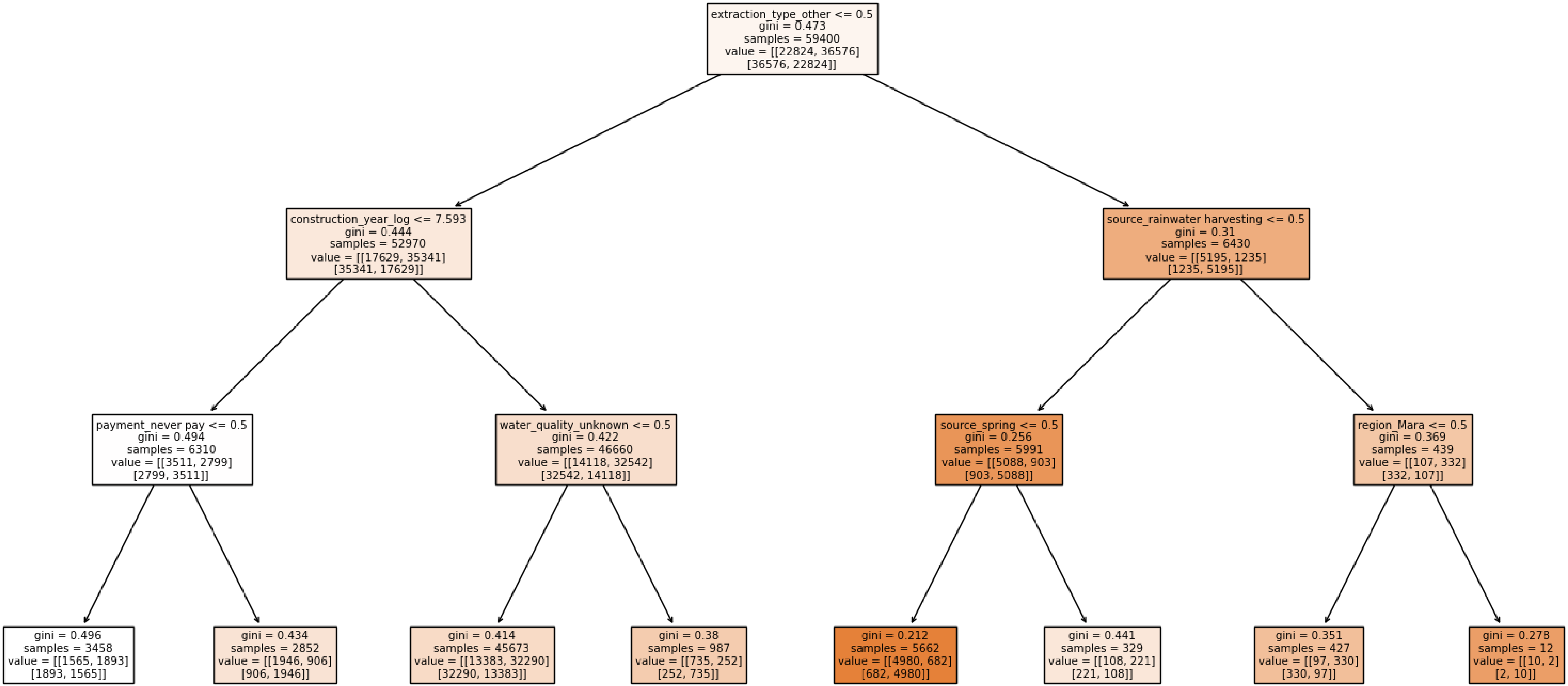
Extraction Type, Payment, Construction Year highly influential.

Random Forest:

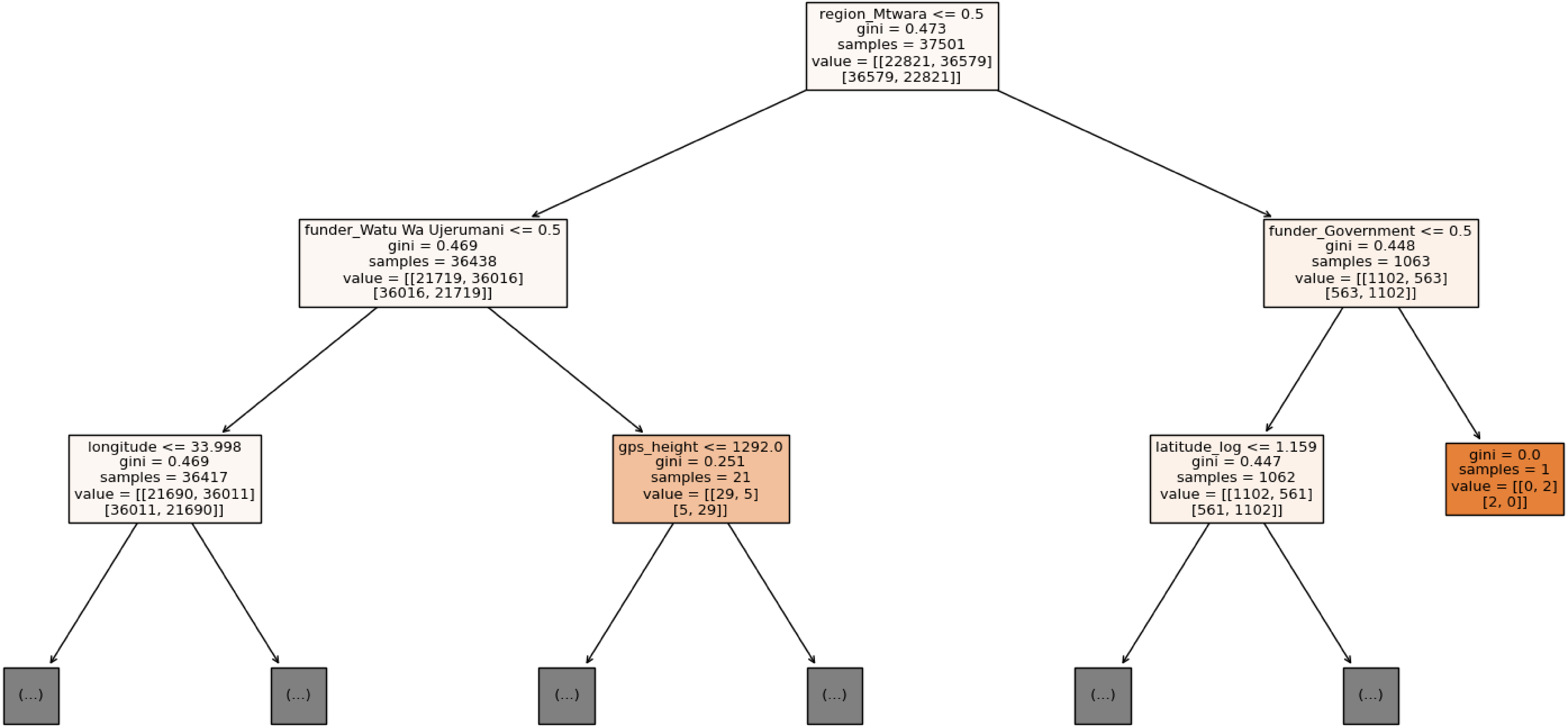
Latitude, Longitude, Construction Year prominent.

Geographic features contribute to spatial understanding.

DECISION TREES MODEL VISUALIZATION



RANDOM FOREST MODEL VISUALIZATION



Implications and Conclusion

Implications:

Model performs well on unseen data.

Consistent performance across different folds.

Conclusion:

Complex factors influence water pump functionality.

Insights can inform interventions for improved water access and infrastructure maintenance, especially in challenging regions.

Decision Tree Classifier Model Tuning

Tuned Decision Tree Model:

Best Parameters: max_depth of 20.

Best Cross-validation Score: ~0.783.

Mean CV Accuracy: ~0.783.

Standard Deviation: ~0.0014.

Summary:

Tuned model outperforms untuned model.

Higher accuracy and consistency in cross-validation.

Random Forest Classifier Model Tuning

Best Parameters:

max_depth: 30

Best Cross-validation Score: ~0.8114

Cross-validation Scores:

[0.8159, 0.8087, 0.8091]

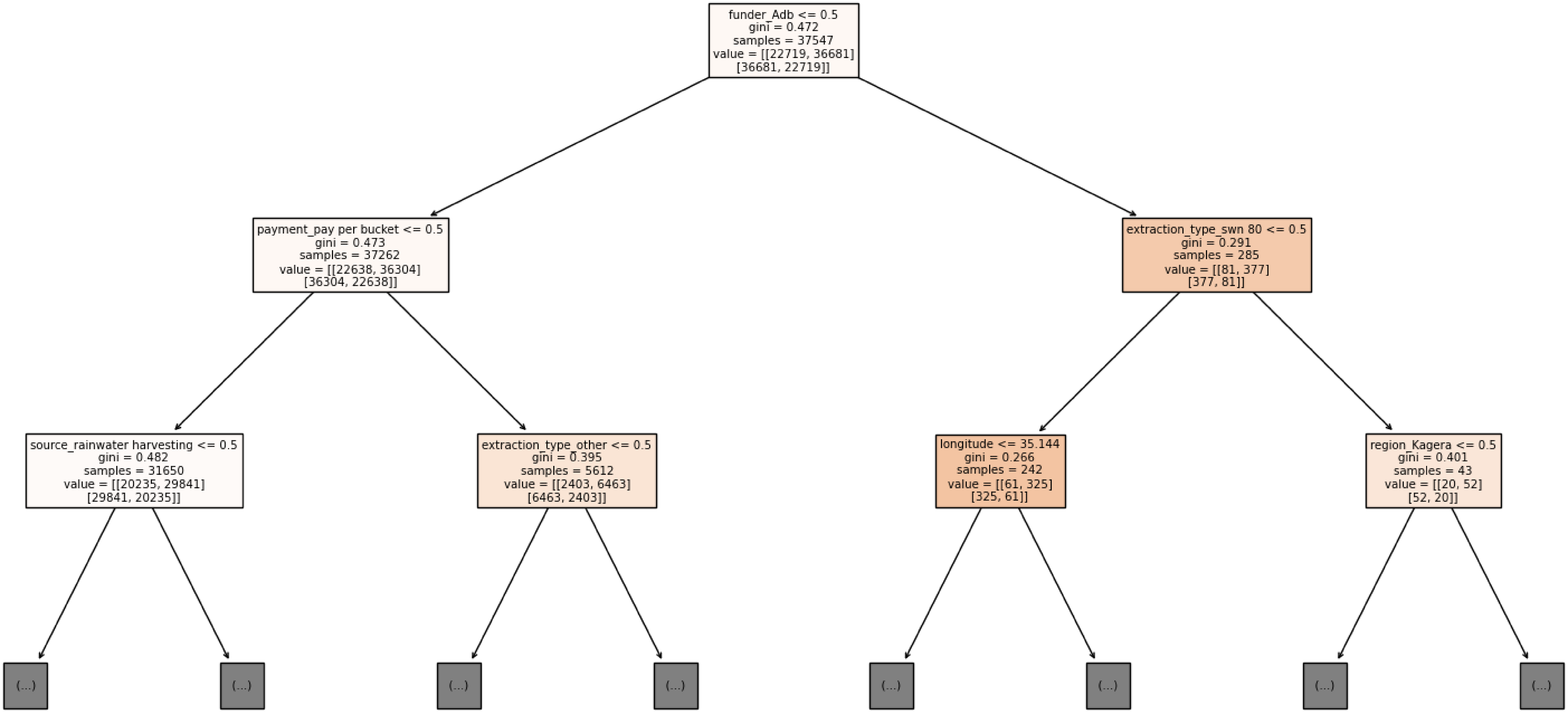
Mean CV Accuracy: ~0.8112

Standard deviation: ~0.0033

Test Set Predictions:

Predicted labels: functional/non-functional

TUNED RANDOM FOREST MODEL VISUALIZATIONS



Conclusion

Tuned Random Forest Classifier:

Achieved mean CV accuracy of ~81.12%.
Despite slightly lower accuracy than default model, outperforms other models.
Recommended for practical applications due to robustness and accuracy.

Recommendations

Implement Routine Maintenance Programs

- Action: Establish regular inspection and maintenance schedules for water pumps, including checks on mechanical components and water quality.
- Benefits: Timely detection and repair of faults can prevent breakdowns, ensuring continuous access to clean water for communities.

Target High-Risk Regions

Action: Utilize geographic data to identify regions with a high prevalence of non-functional water pumps.

Benefits: Target interventions, such as repair and rehabilitation efforts, to areas with the greatest need, optimizing resource allocation and impact.

Introduce Flexible Payment Plans

Action: Introduce flexible payment options for water services, including subsidized or tiered pricing models based on income levels.

Benefits: Improve affordability and accessibility of water services, reducing the financial burden on low-income communities and increasing revenue for maintenance and infrastructure upgrades.

Foster Collaboration with Funders

Action: Engage with funders and donor organizations to align priorities and strategies for water infrastructure projects.

Benefits: Secure long-term support and investment in water projects, leveraging partnerships to access funding for maintenance, upgrades, and capacity-building initiatives.

Embrace Data-Driven Decision Making

Action: Invest in robust data collection systems and analytics capabilities to track water pump functionality and performance metrics.

Benefits: Enable evidence-based decision making, including trend analysis, predictive maintenance, and resource allocation based on real-time insights, leading to more effective and efficient management of water infrastructure.

Conclusion

In conclusion, the project has provided a comprehensive exploration of predicting water pump functionality and optimizing maintenance strategies. By leveraging machine learning techniques and data analysis, we have gained insights into the factors influencing water pump functionality, identified high-risk regions, and proposed actionable recommendations for improving access to clean water. Through collaborative efforts and data-driven decision-making, we aim to contribute to the sustainable management of water infrastructure and ensure the well-being of communities relying on these vital resources.

THANK YOU!!

https://www.linkedin.com/in.sarah-nyambura-a22276267/