

HYDROFORECAST: TANZANIA WATER WELLS PREDICTION

Josephine Maro

TABLE OF CONTENTS



OVERVIEW

BUSINESS UNDERSTANDING

DATA UNDERSTANDING

MODELLING

EVALUATION

CONCLUSION

RECOMMENDATIONS

OVERVIEW

Introduction

- This project aims to analyze Tanzanian water wells to identify patterns and factors that contribute to well functionality status.
- Predictive models and segmenting wells based on their characteristics and performance provide actionable insights to improve well maintenance, optimize resource allocation, and enhance water access.

Rationale

- Various models chosen for their diverse capabilities.
- Logistic Regression provides baseline performance and interpretable results.
- KNN captured local patterns, ideal for heterogeneous data.
- Decision Tree offers insights into feature importance and decision-making.
- Random Forest combines decision trees to enhance accuracy and robustness.





BUSINESS UNDERSTANDING

Problem Statement

- In Tanzania, maintaining the functionality of water wells is critical for ensuring sustainable access to clean water.
- Through this prediction, stakeholders can prioritize maintenance efforts, allocate resources efficiently, and safeguard access to clean water for local communities.

Stakeholders

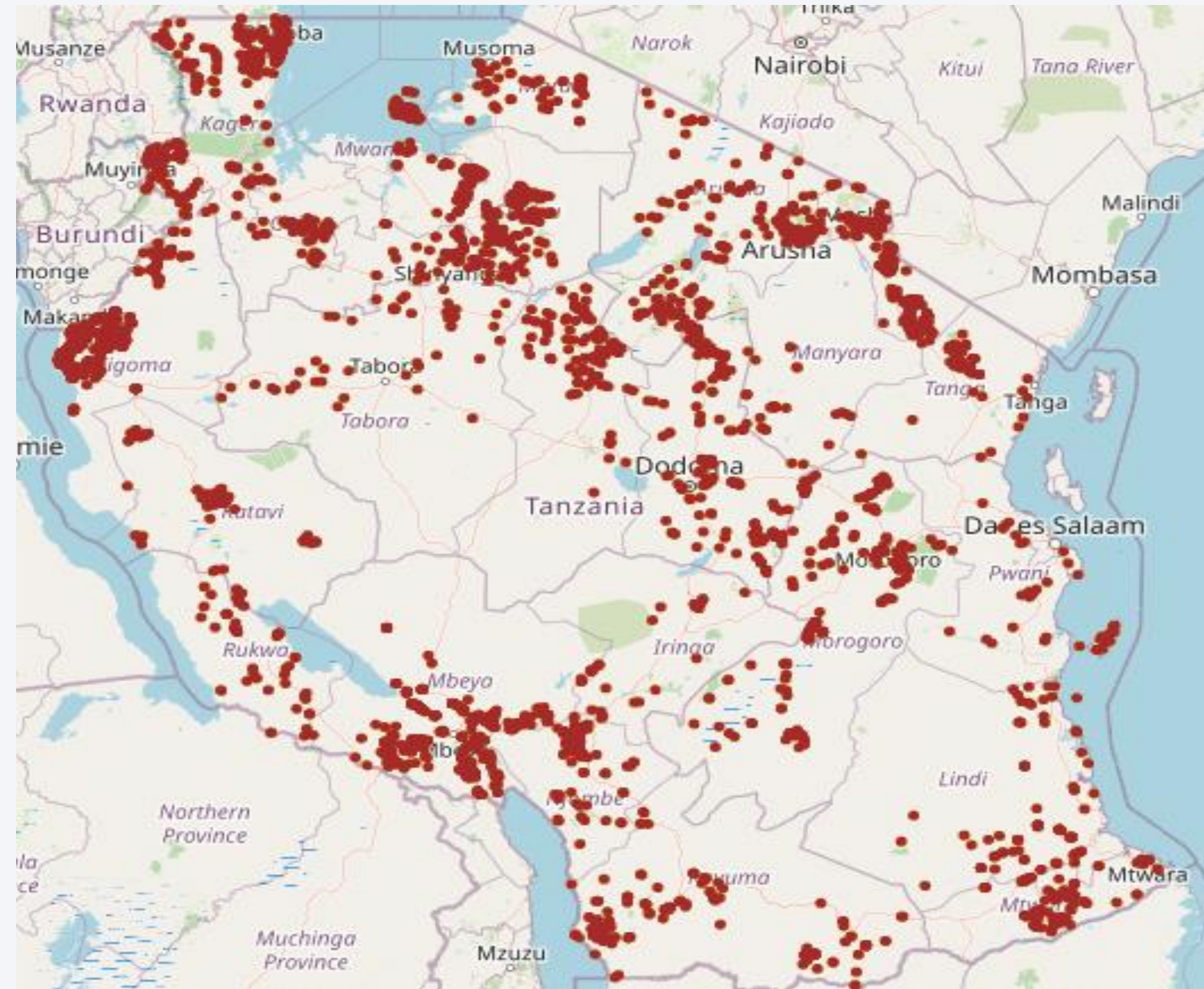
- Tanzanian government agencies for water resource management
- NGOs
- Local communities

Business Impact

- Resource allocation.
- Mitigate the risk of waterborne diseases
- Economic development
- Ensuring uninterrupted water supply for agriculture and other activities.

DATA UNDERSTANDING

- Dataset used contains information about existing water wells in Tanzania.
- Four CSV files
- "Training set values" containing independent features data for the training set
 - "Training set labels" containing dependent variable data
 - "Test set values" for prediction
 - "Submission format" for result formatting



DATASET

01

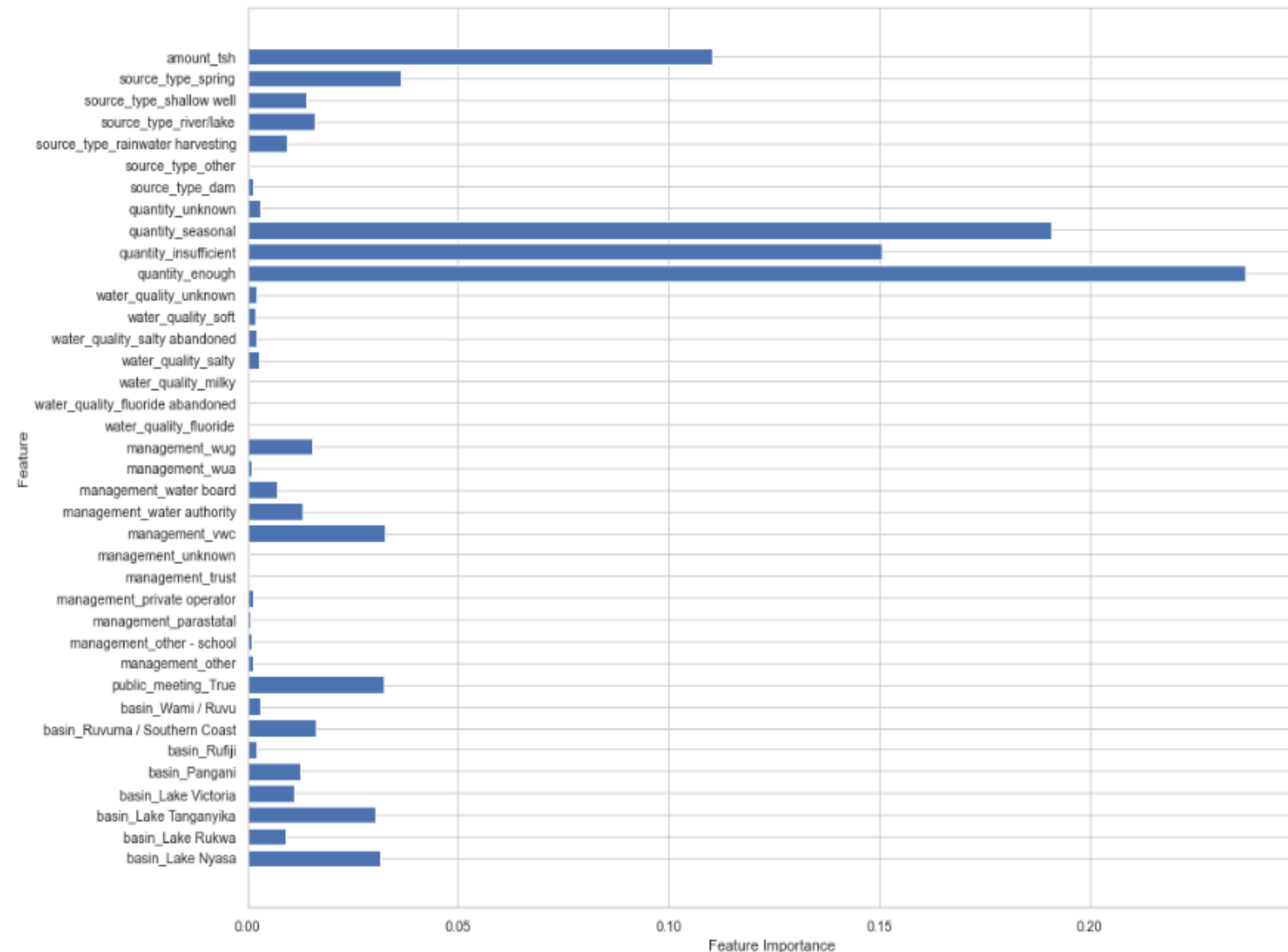
- The dataset comprises 59,400 records and 40 columns, with 31 identified as categorical and 9 as numerical.

02

- Further classification grouped the columns into general features captured by the dataset.

DATA ANALYSIS

- We explored the functionality status of water wells using various ML models.
- We began with exploratory data analysis to understand feature distributions and relationships.
- Logistic regression, K-Nearest Neighbors, Decision Tree, and Random Forest models were trained and evaluated.
- The Decision Tree classifier performed the best, with an accuracy of 71.0%.
- Feature importance analysis revealed key predictors.

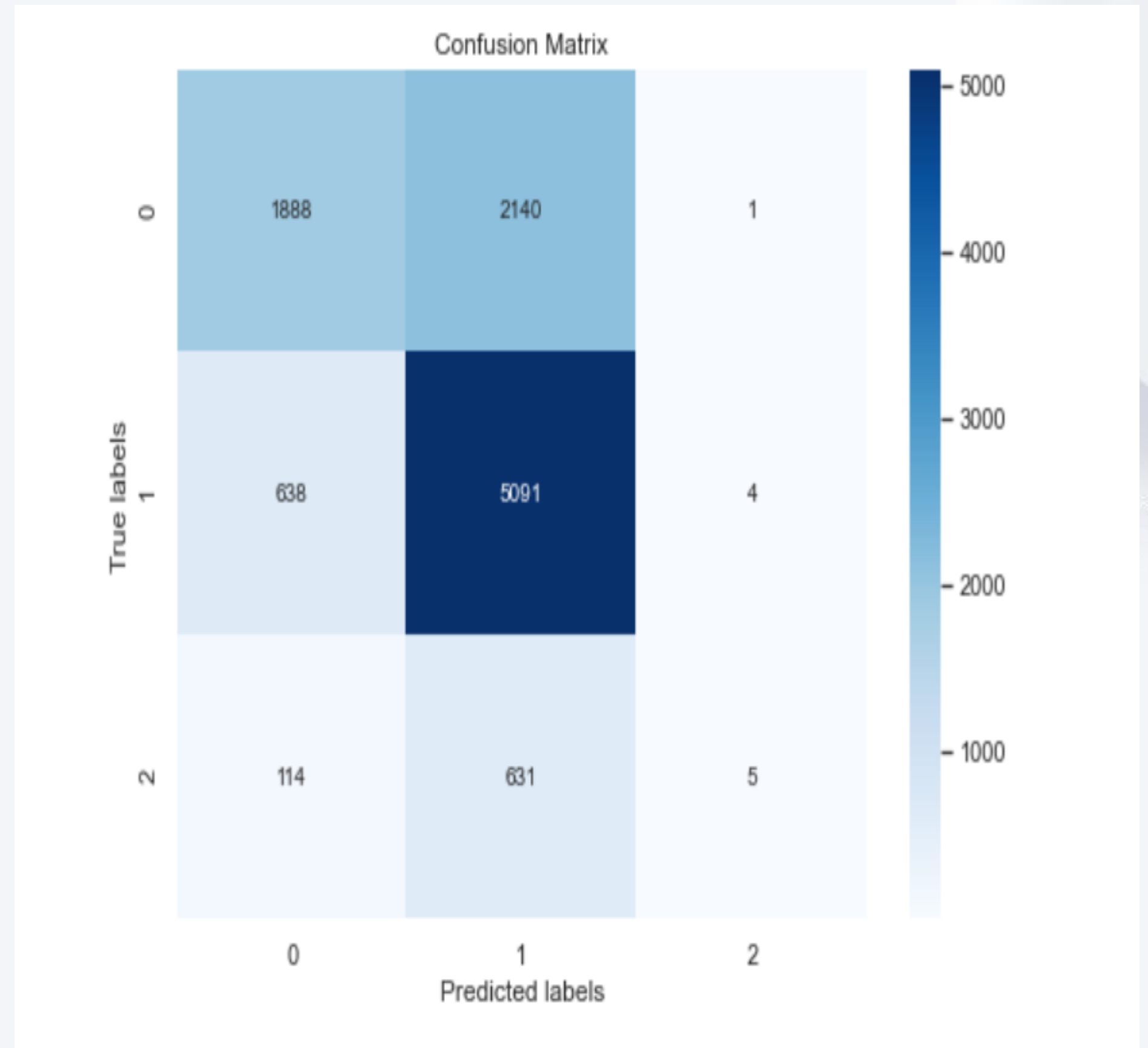


MODELLING

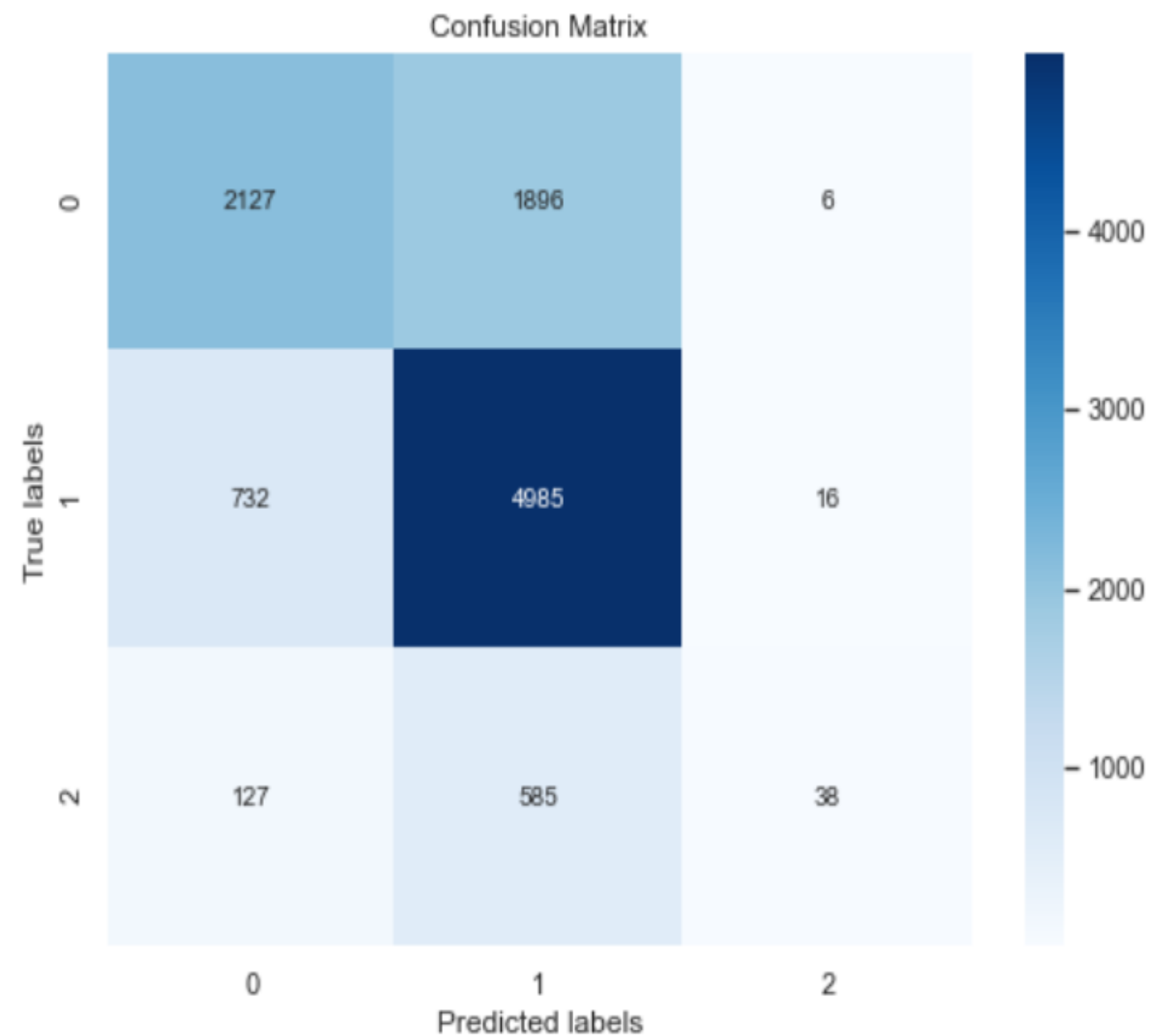
- We employed various machine learning techniques to predict water well functionality status based on provided data.
- Each model underwent training on the training set and subsequent evaluation on the testing set to gauge its predictive accuracy and generalization ability..

MODELS USED

- Logistic regression served as our initial baseline model, providing a probabilistic interpretation of predictions with an accuracy of approximately 66.4%.
- Following this, we explored the K-Nearest Neighbors (KNN) algorithm, achieving an accuracy of approximately 67.2%.
- Decision Tree and Random Forest classifiers were also employed, with the Decision Tree model yielding the highest accuracy of approximately 71.0%.
- Random Forest model achieved an accuracy of approximately 68.0%.




HYPERPARAMETER TUNING



- We employed hyperparameter tuning techniques, particularly using GridSearchCV, to systematically explore a range of hyperparameters and identify the optimal configuration for our models.
- In the random forest classifier, we defined a grid of hyperparameters including criteria for splitting, maximum depth, maximum number of features, and the number of estimators.
- GridSearchCV then exhaustively searched through this parameter grid, evaluating each combination through cross-validation to determine the set of hyperparameters that yielded the best performance.

EVALUATION

- The evaluation compared the performance of Logistic Regression, KNN, Decision Tree, and Random Forest models in classifying water well statuses.
 - Decision Tree outperformed others with an accuracy of 71.0%, while Logistic Regression and KNN struggled with accurately predicting wells needing repair.
- 

METRICS



- We relied on metrics including Accuracy, Precision, Recall, and F1-Score.
- Logistic Regression achieved an accuracy of 66.4%, with notable precision and recall for functional wells but struggled with wells needing repair.
- KNN achieved an accuracy of 67.2% and facing difficulties in accurately predicting wells needing repair.
- Decision Tree achieved an accuracy of 71.0% and showcasing balanced precision and recall for functional and non-functional wells.
- Random Forest, with an accuracy of 68.0%, encountered challenges in accurately predicting wells needing repair.



CONCLUSION

- The decision tree model showed the best performance among the models evaluated, with the highest accuracy and better precision and recall for functional and non-functional wells.
- All models struggled to accurately predict wells needing repair, indicating a need for further improvement in identifying and addressing maintenance issues.

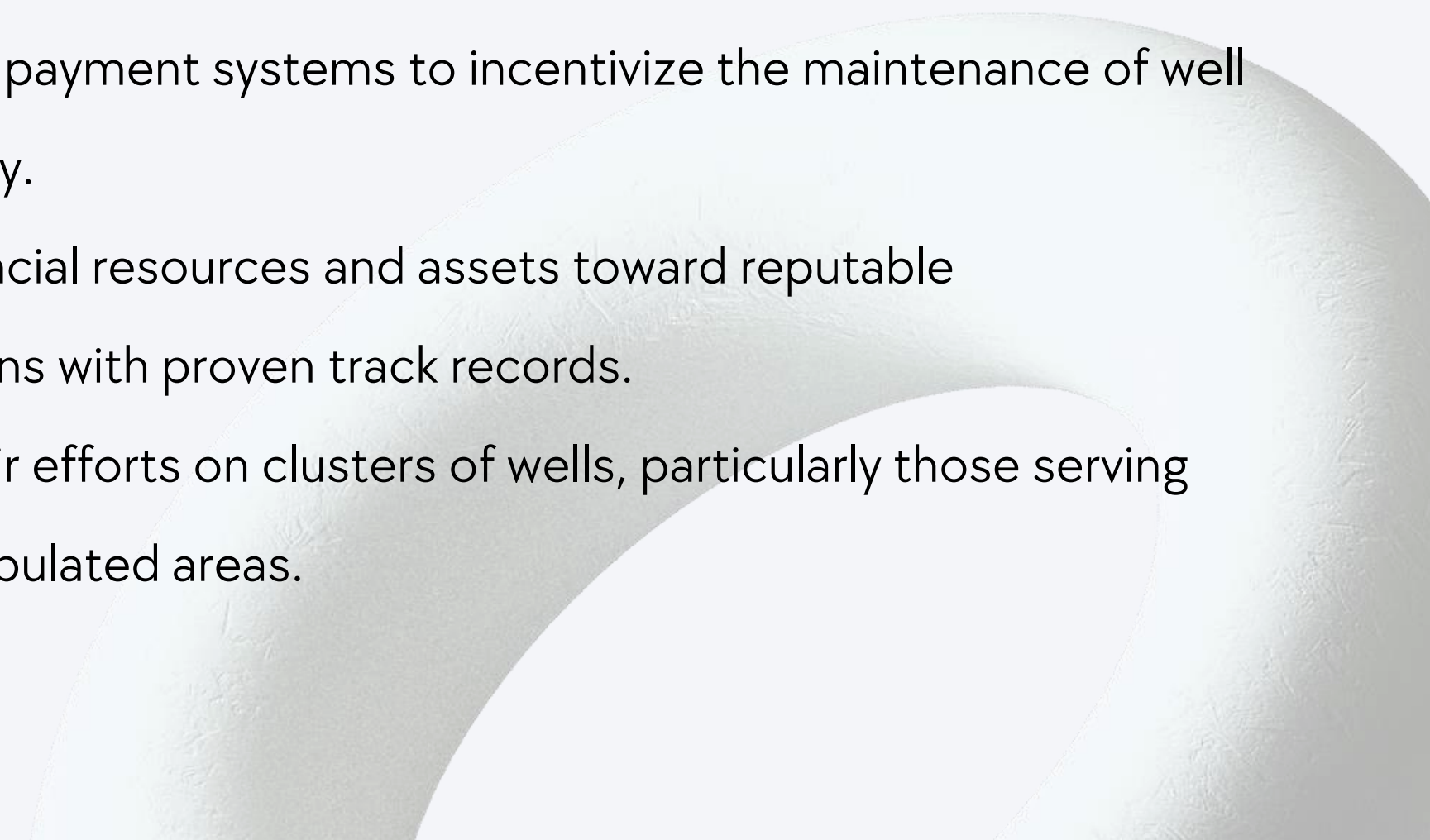
Limitations

- Time constraints, among other factors, constrained my efforts.
- Certain computationally intensive techniques, like hyperparameter tuning, were constrained due to their prolonged execution times, necessitating their selective application.
- Improving the model in the future could involve engineering new features that may be more pertinent to the current problem.





RECOMMENDATIONS

- Give precedence to repairing operational wells that dispense clean water.
 - Implement payment systems to incentivize the maintenance of well functionality.
 - Direct financial resources and assets toward reputable organizations with proven track records.
 - Focus repair efforts on clusters of wells, particularly those serving densely populated areas.
- 

CONTACT INFO

Name: Josephine Maro

Mail to:

josephinemaro@student.moringa
school.com



