Tanzanian Water Wells

PHASE 3 PROJECT:

BETTY KOILA

DSF-PT08

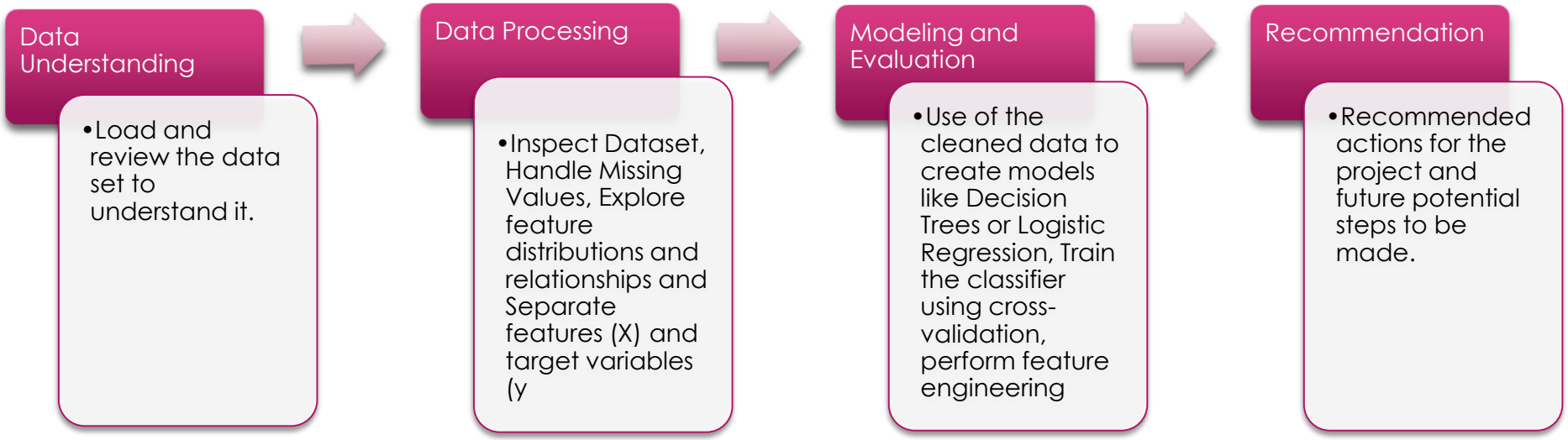
Project Outline

- Business understanding
- Data Understanding and Processing
- Modelling
- Evaluation
- Recommendations and Conclusions

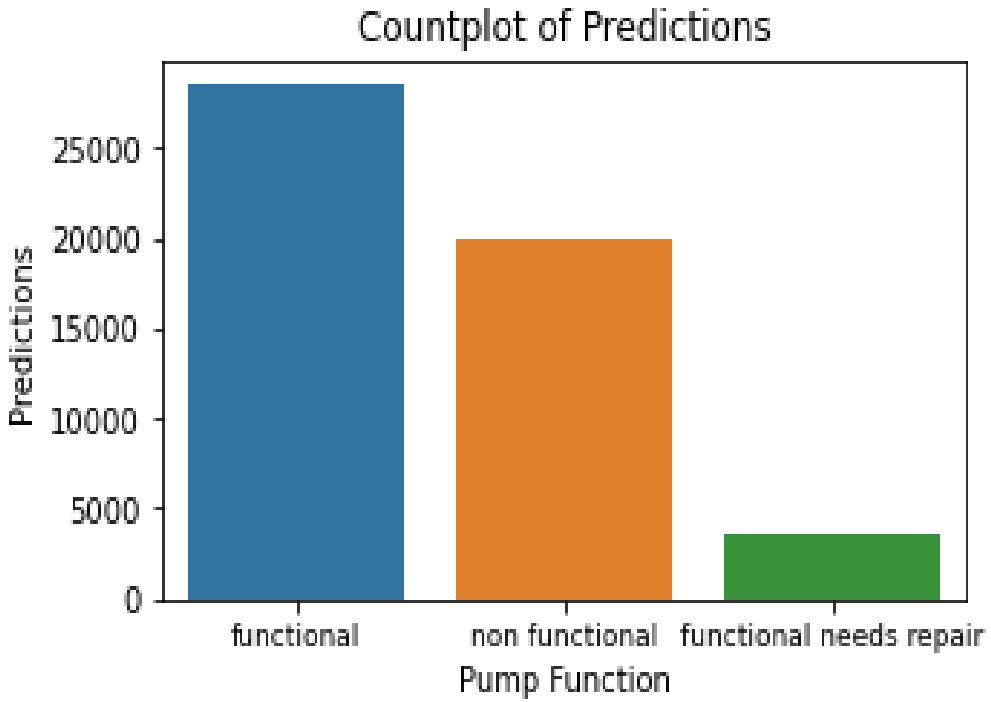
Business understanding

- This project aims to predict the operational status of water pumps across Tanzania, specifically identifying which pumps are functional, which require repairs, and which are non-functional.
- The prediction is based on various factors such as pump type, installation date, and management practices.
- The project uses data from Taarifa and the Tanzanian Ministry of Water, with the challenge provided by Driven Data in 2015.
- By accurately forecasting pump failures, maintenance efforts can be optimized, ensuring that communities in Tanzania maintain reliable access to clean and potable water.

Data Process

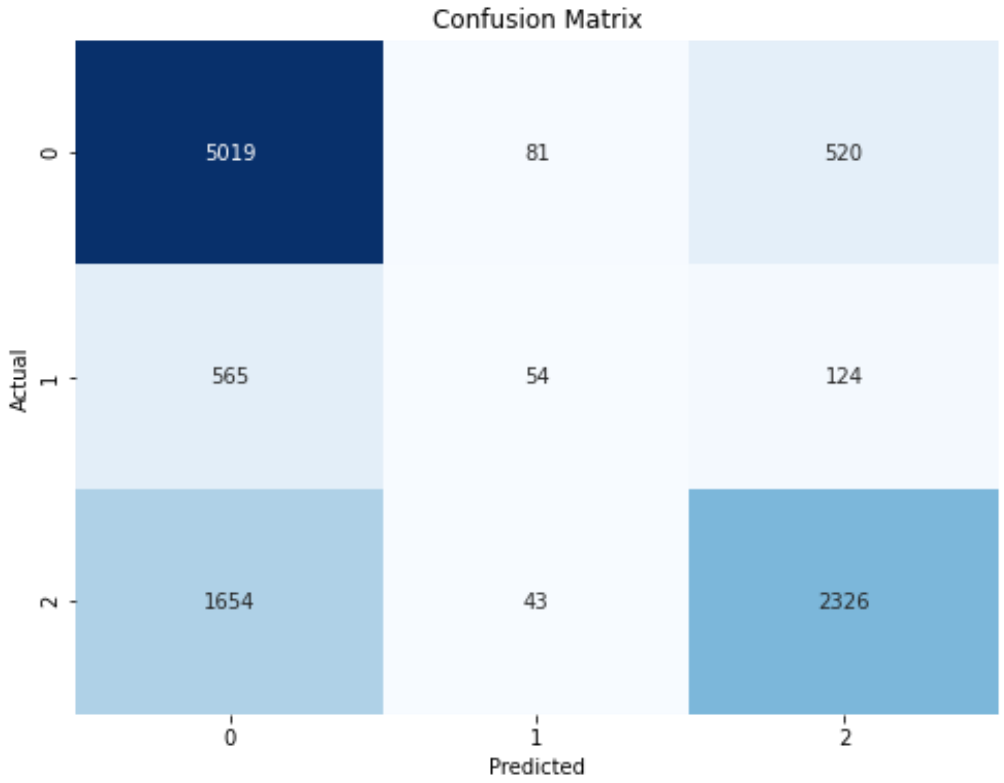


Data Understanding and Processing



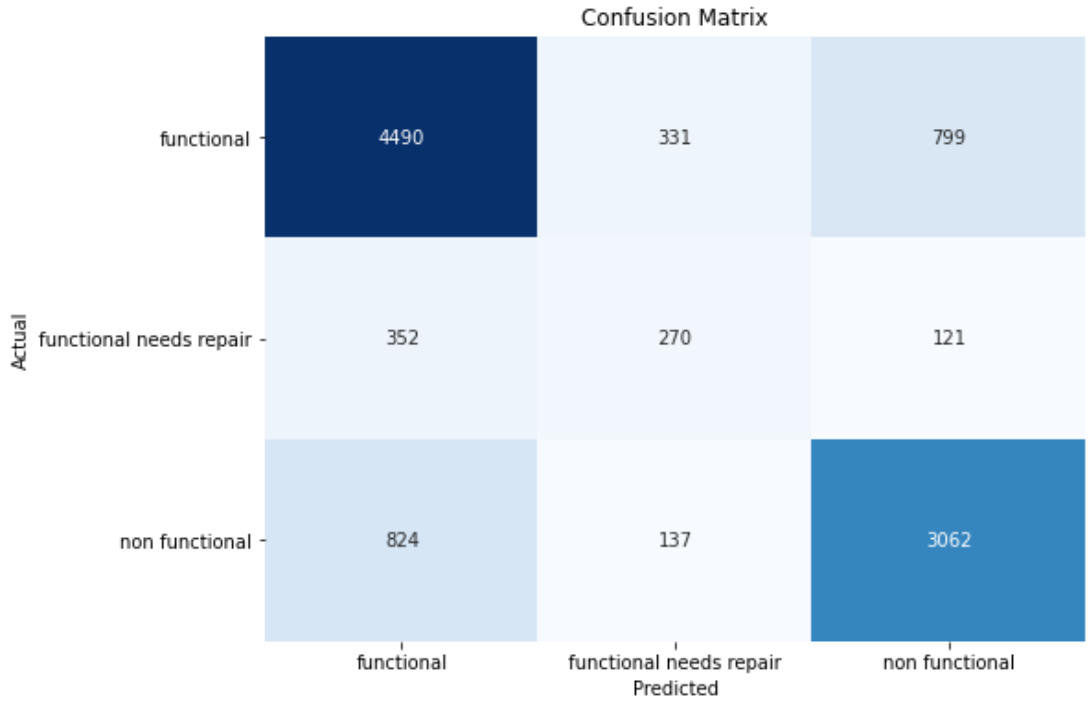
- The dataset was examined to understand its structure.
- Missing values were identified and appropriately handled.
- Data distributions and correlations were explored to uncover insights.
- Feature engineering was applied where necessary to enhance the data.
- The data was then split into features (X) and the target variable(s) (y) for model training.

Modelling and Evaluation: Logistic Regression



- Accuracy: Logistic regression achieves 70.8% accuracy but has room for improvement.
- Class Performance: High precision and recall for functional pumps; poor performance for pumps needing repair with very low recall (0.01%).Non-functional Pumps: Reasonable performance with 74% precision, but recall is moderate (61%).
- Key Issue: The model struggles to identify "functional needs repair" pumps due to class imbalance.

Modelling and Evaluation : Decision Tree



- Overall Accuracy: 75.31% accuracy is satisfactory but room for improvement, especially in distinguishing between classes.
- Functional Pumps: Strong performance with 79% precision, 80% recall, and 80% F1-score.Non-Functional Pumps: Solid results with 77% precision, 76% recall, and 77% F1-score.Functional Pumps Needing Repair: Weak performance with 37% precision, 36% recall, and 36% F1-score, indicating difficulty in identifying this class.
- Key Insight: The model performs well with majority classes but struggles significantly with the "functional needs repair" category, likely due to class imbalance and model limitations in distinguishing these pumps.

Model Comparison

	Metric	Decision Tree	Logistic Regression
0	Accuracy	0.757	0.708
1	Functional Precision	0.790	0.700
2	Functional Recall	0.800	0.870
3	Functional F1-Score	0.800	0.770
4	Functional Needs Repair Precision	0.370	0.420
5	Functional Needs Repair Recall	0.350	0.010
6	Functional Needs Repair F1-Score	0.360	0.010
7	Non-functional Precision	0.770	0.740
8	Non-functional Recall	0.770	0.610
9	Non-functional F1-Score	0.770	0.670

- Accuracy: Decision Tree (75.31%) and Logistic Regression (70.8%) performed well but leave room for improvement.
- Precision/Recall/F1-Score: Decision Tree had strong performance for functional and non-functional pumps but struggled with pumps needing repair, while logistic regression faced challenges across all classes, especially with the repair category.
- Feature Impact: Pump type and installation age had significant influence on the model's predictions.
- Model Performance: Majority classes (functional and non-functional) were predicted well, but "functional needs repair" pumps were consistently misclassified

Recommendations and Conclusions

- Recommended Model for the Problem
- Decision Tree

Why it's recommended:

- Achieves solid performance with 75.7% accuracy.
- Highly interpretable, offering clear decision-making criteria.
- Useful for real-world applications where transparency is important.

- Limitation:

- Struggles with accurately classifying the "functional needs repair" category.
- Low precision and recall for this class.
- Can be improved through optimization or model combination.