

TANZANIA WATER WELLS



PREDICTING WATER PUMP STATUS
USING MACHINE LEARNING
CLASSIFICATION MODELS.



INTRODUCTION

- PROBLEM STATEMENT

- Water access is critical in rural Tanzania.
- Many pumps fail or require maintenance, impacting communities.
- **Goal:** Predict whether a water pump is:
 1. Functional
 2. Functional needs repair
 - 3 . Non-functional



DATASET OVERVIEW

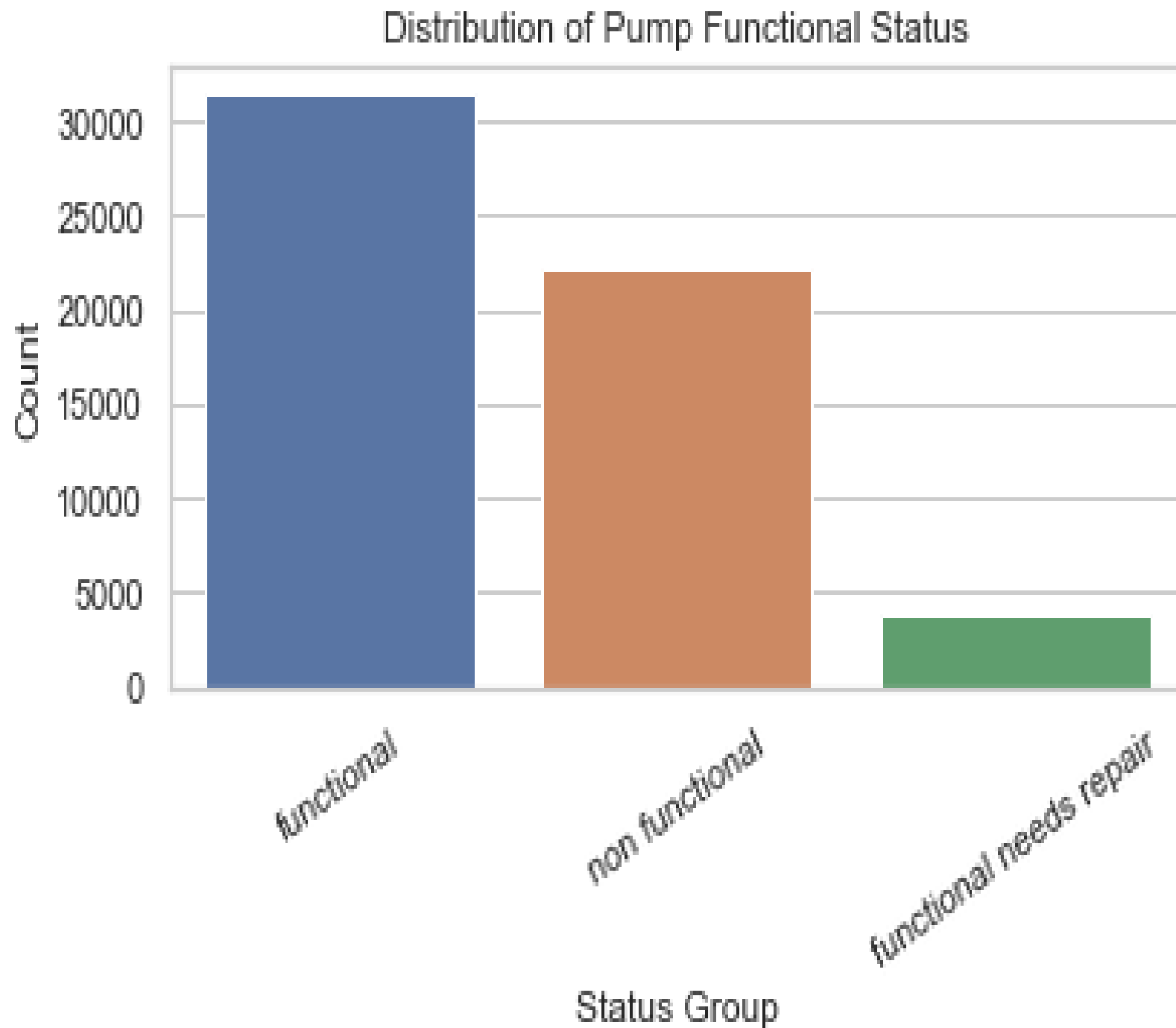
- Source: DrivenData "Pump it Up" competition.
- **Total records:** ~59,400 pumps.
- Key features:
 - Categorical: basin, region, payment_type, etc.
 - Numerical: gps_height, population, amount_tsh, etc.
 - Binary: permit, public_meeting.



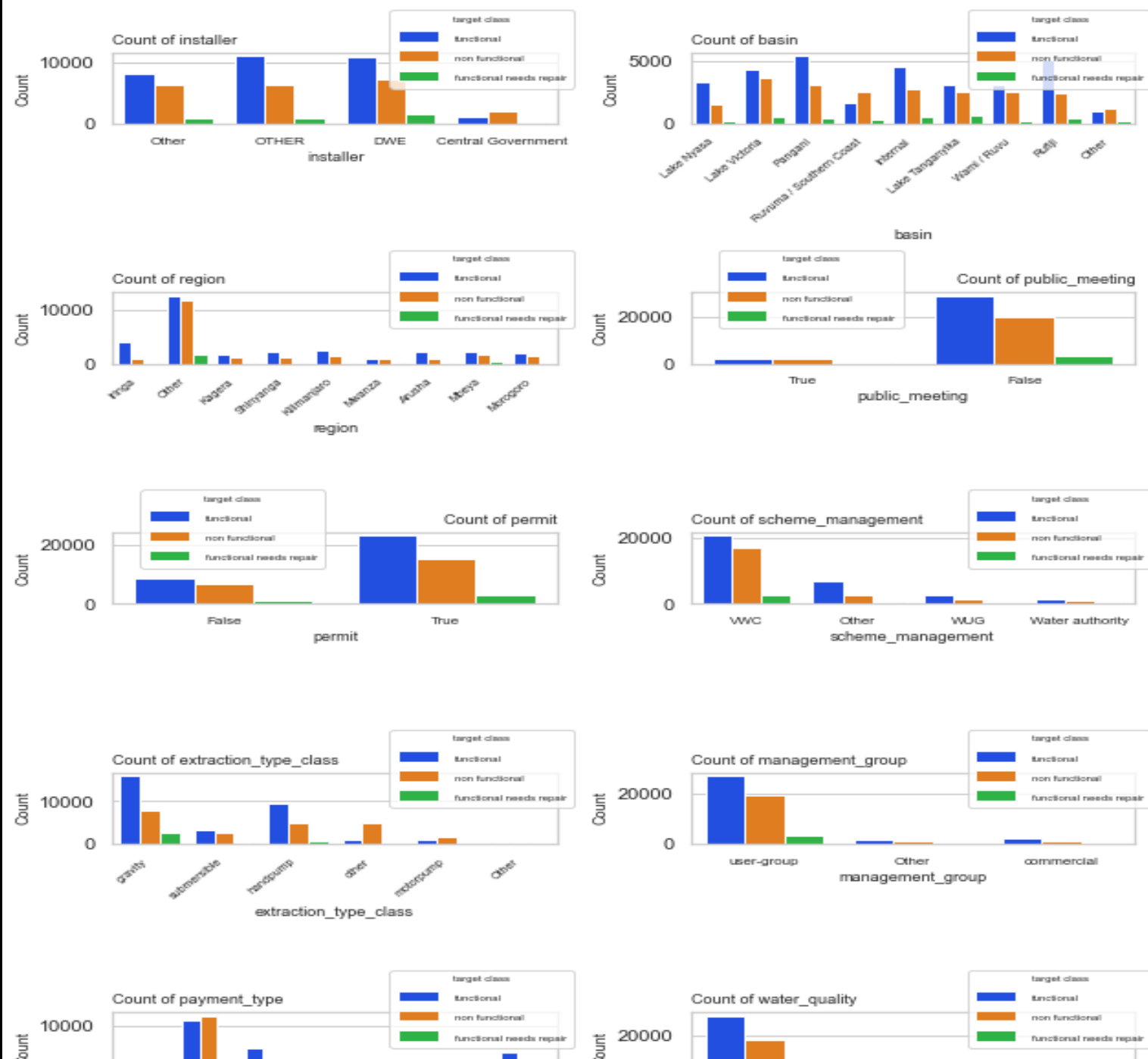
DATA PREPROCESSING

- Missing value handling
- One-hot encoding for categorical variables
- Train-test split: 80/20
- Final feature count after encoding: 21

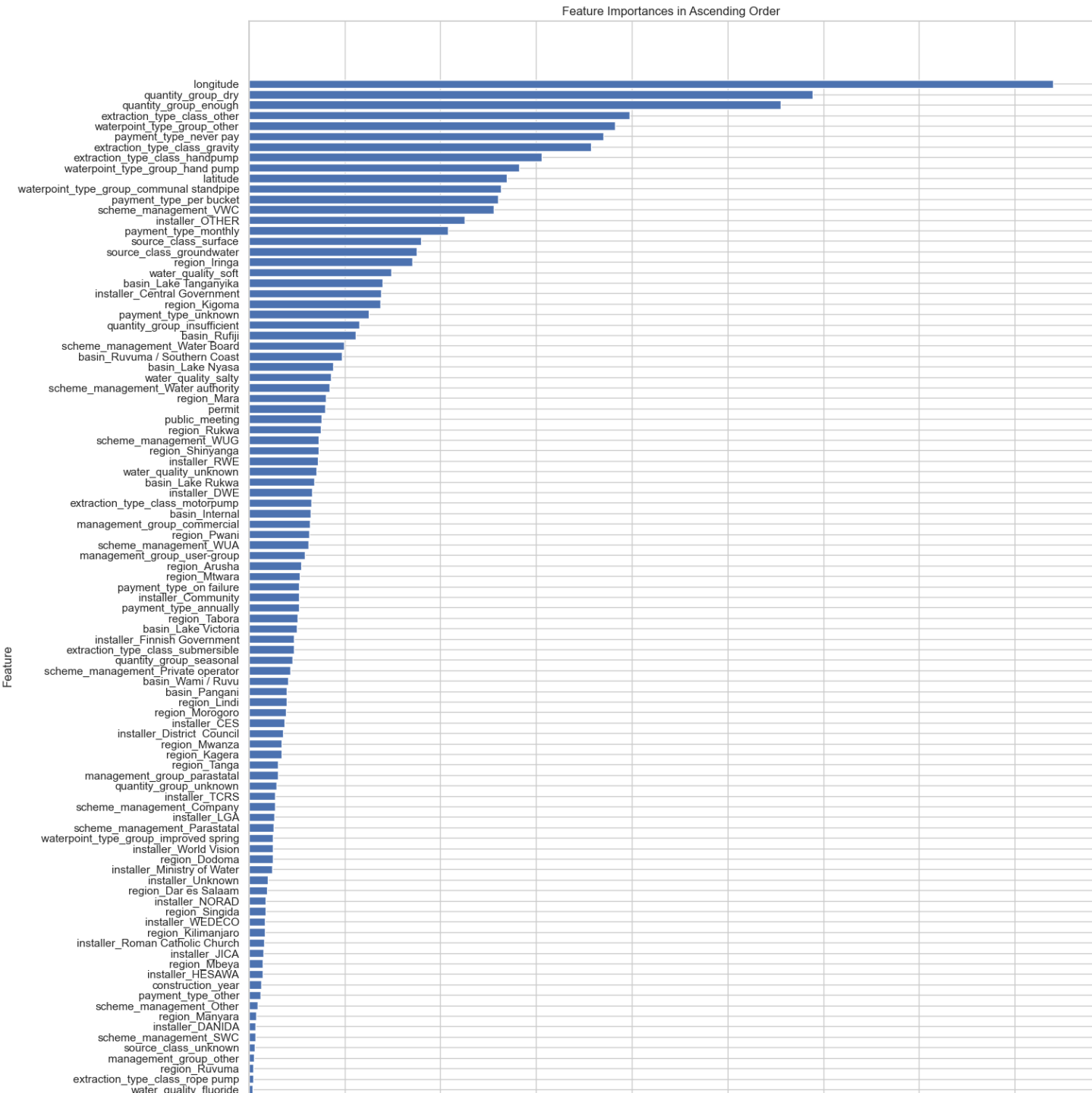
DISTRIBUTION OF TARGET CLASS (STATUS)



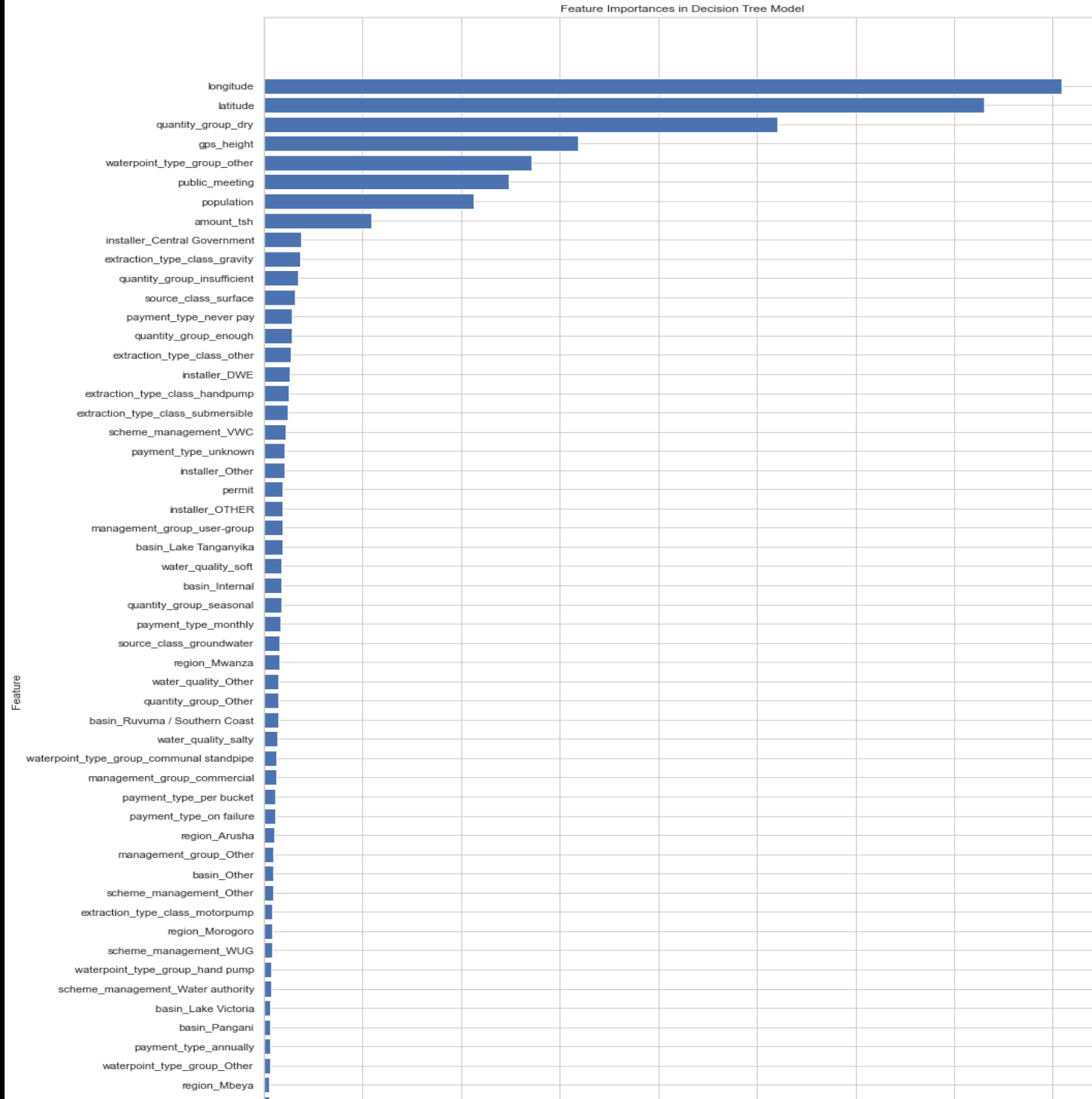
DISTRIBUTION OF TARGET CLASS BY CATEGORICAL FEATURES



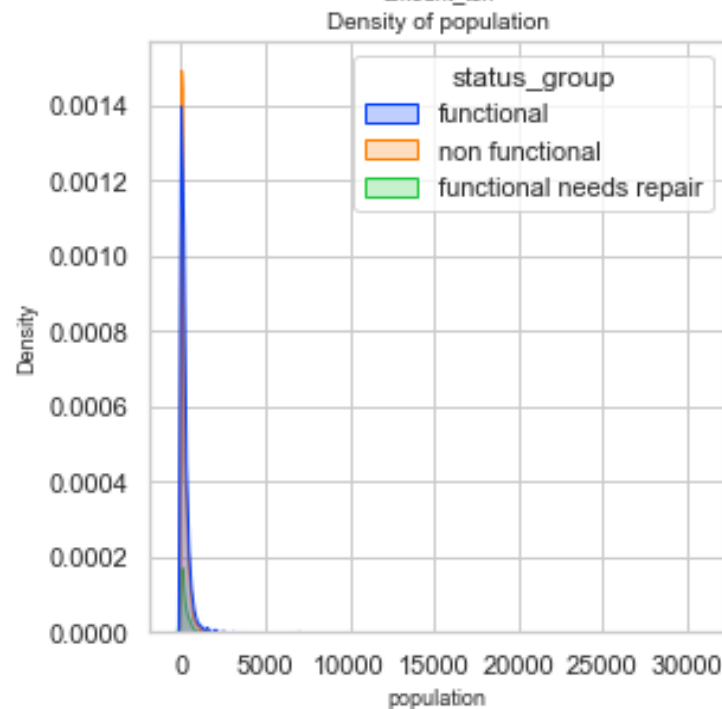
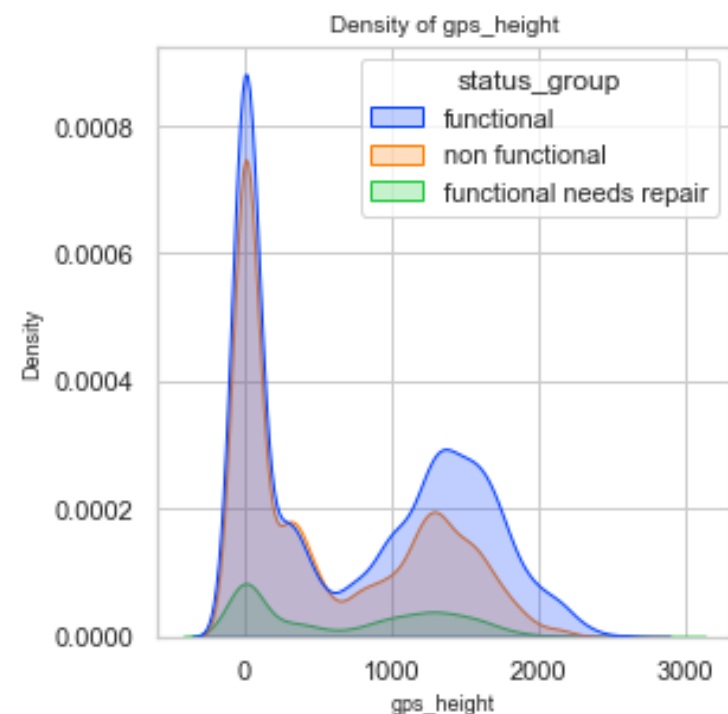
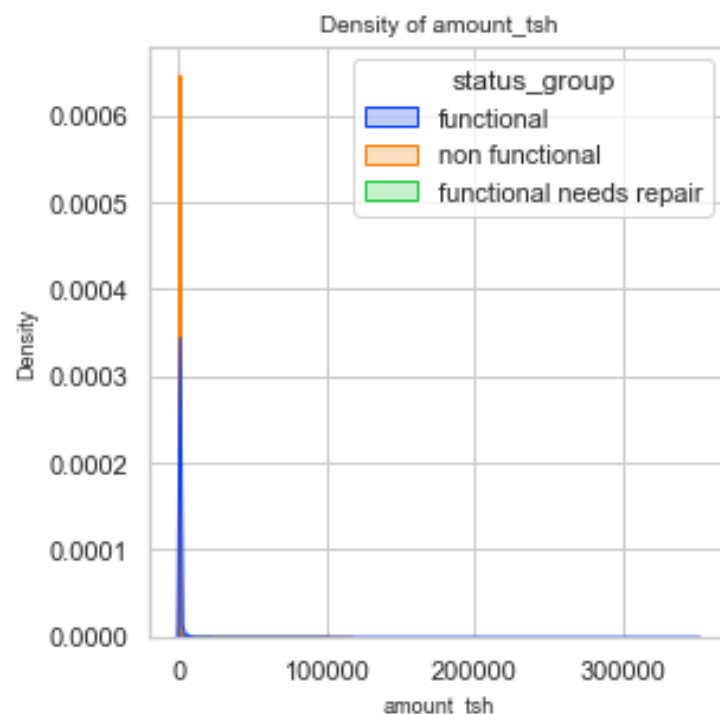
FEATURE IMPORTANCE IN LOGISTIC MODEL



FEATURE IMPORTANCE IN DECISION TREE MODEL



DISTRIBUTION OF TARGET CLASS BY NUMERICAL FEATURES



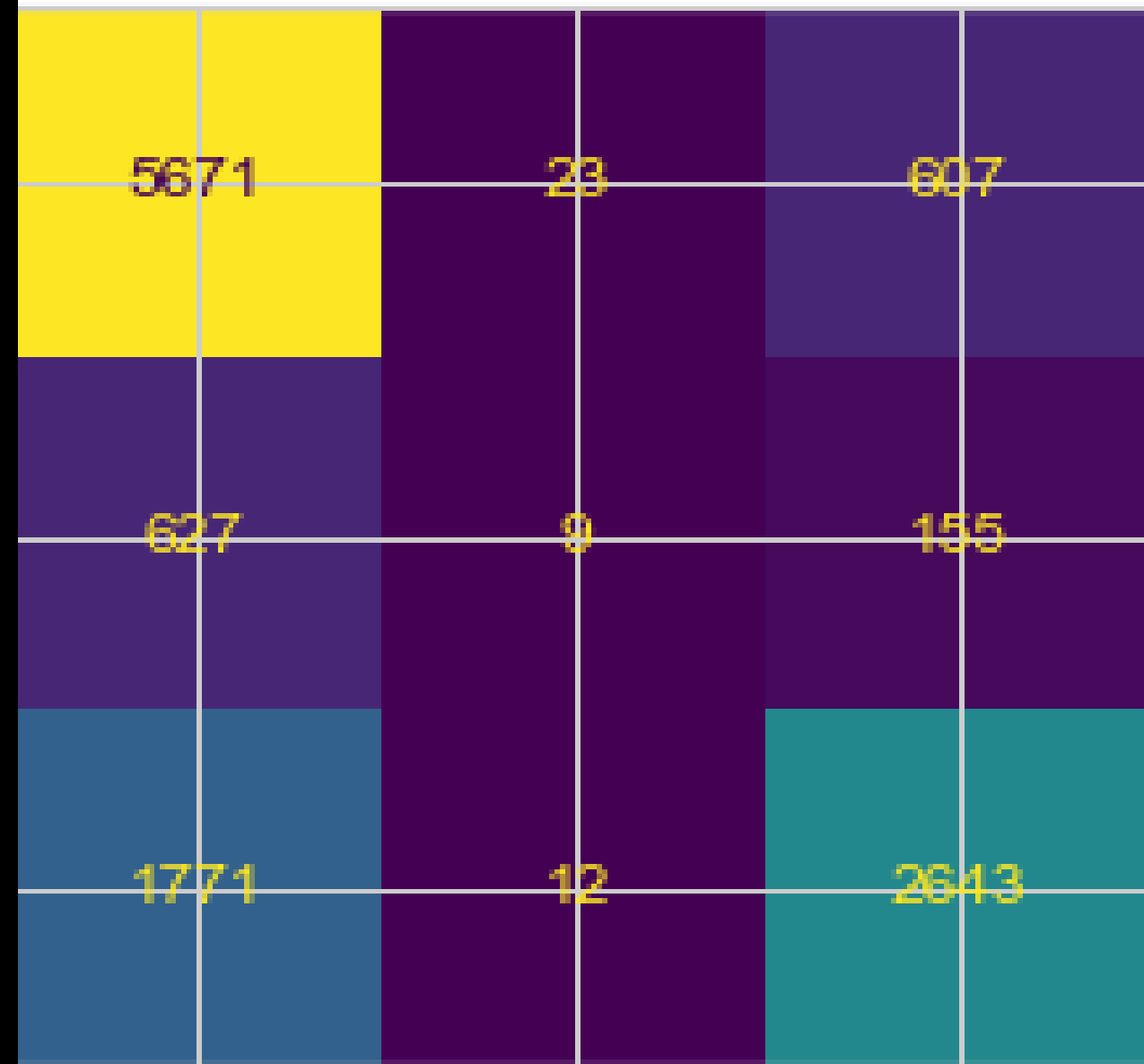


MODELLING


1. LOGISTIC REGRESSION MODEL

- Simple classifier with no tuning
- Results on test set:
 - Accuracy: 72.3%
 - Weighted F1-score: 69.3%
- Struggled with “functional needs repair” class

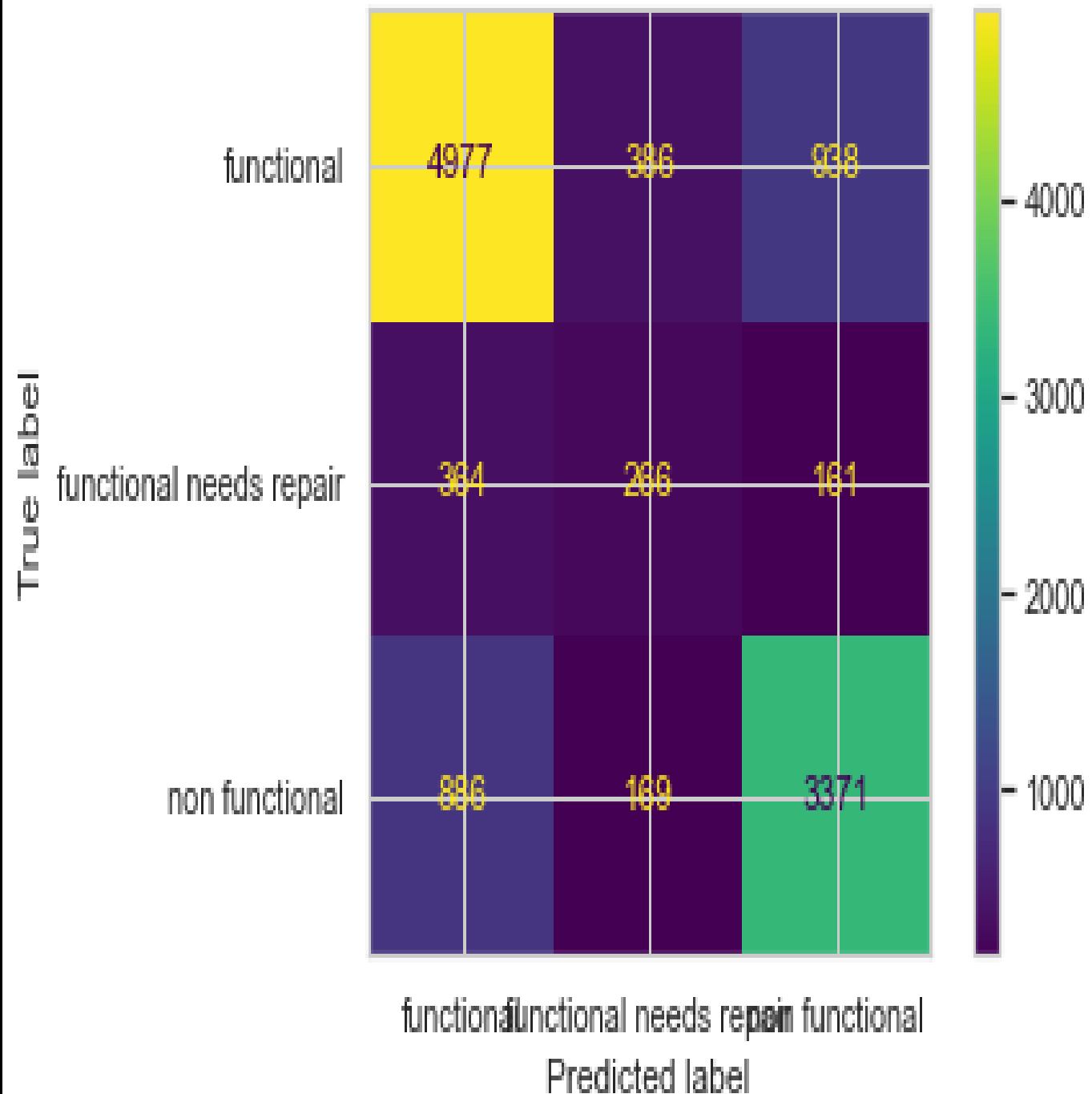
CONFUSION MATRIX FOR LINEAR REGRESSION MODEL



functional functional needs repair functional
Predicted label

- 
- 2. DECISION TREE
 - Unpruned Decision Tree
 - Results on test set:
 - Accuracy: 74.8%
 - Weighted F1-score: 74.9%
 - Better recall and balance across all classes

CONFUSION MATRIX FOR DECISION TREE





MODEL COMPARISON

Metric	Logistic Regression	Decision Tree
Accuracy	72.3%	74.8%
Precision	69.7%	74.9%
Recall	72.3%	74.8%
F1 Score	69.3%	74.9%

Decision tree outperforms logistic regression in all metrics.

CONCLUSION

- Decision trees provide more accurate and balanced predictions.
- “Functional needs repair” remains the hardest to predict.
- **Next steps:**
 - Try ensemble models (Random Forest, XGBoost)
 - Apply SMOTE or class-weighting to improve minority class

THANK YOU



ANY QUESTIONS?

LINKEDIN: JUDAH SAMUEL