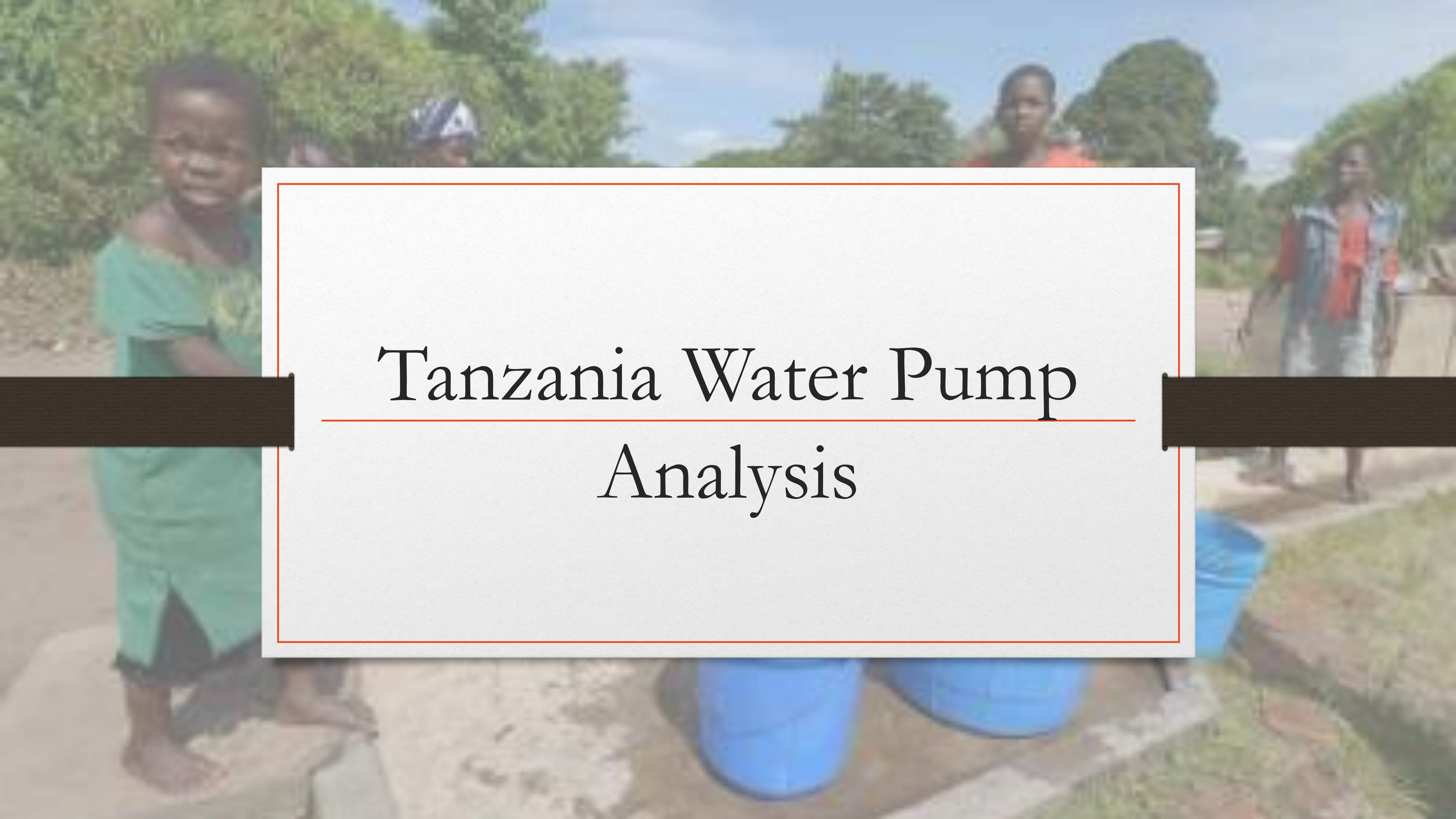


Tanzania Water Pump Analysis

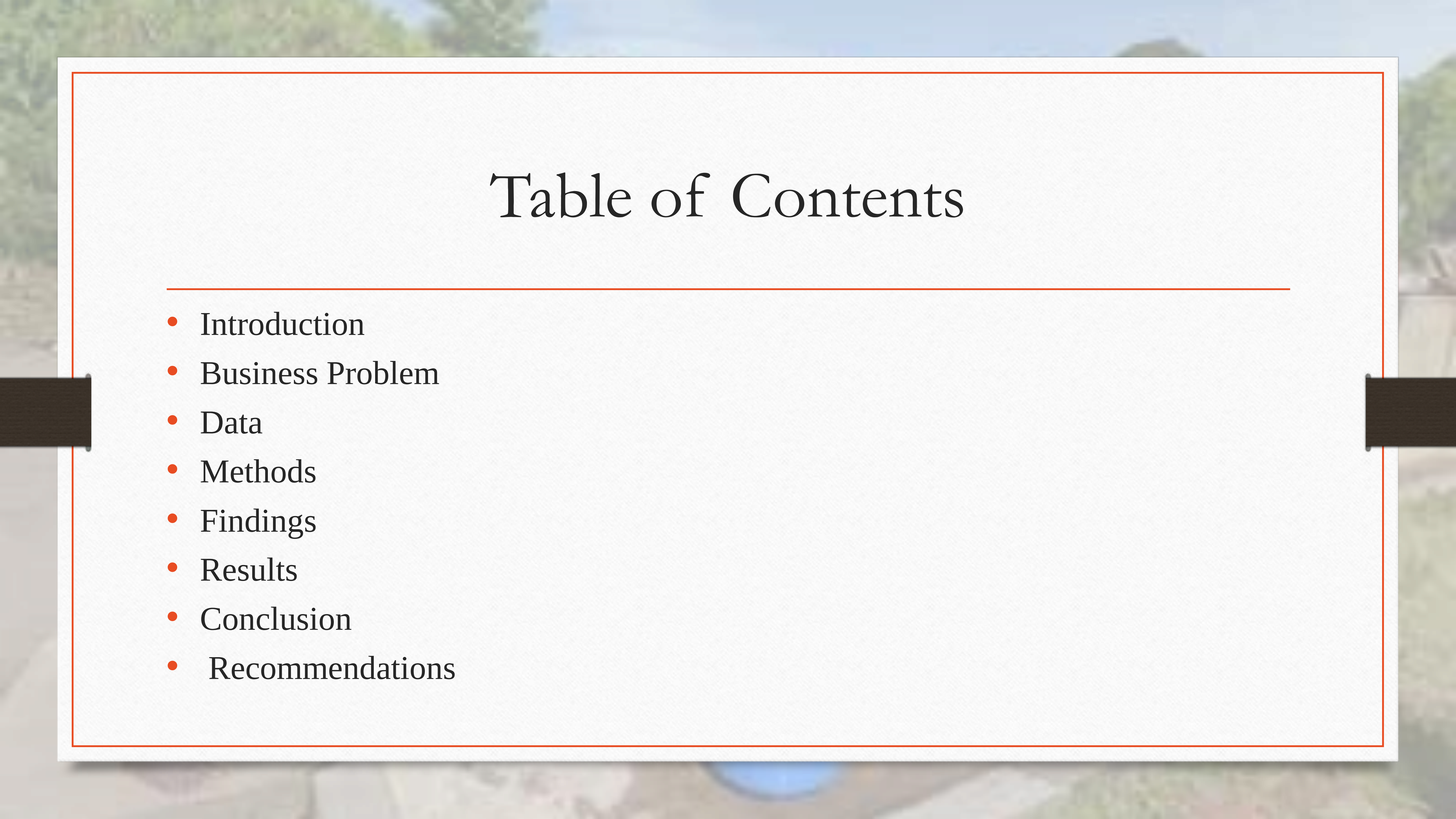


Table of Contents

- Introduction
- Business Problem
- Data
- Methods
- Findings
- Results
- Conclusion
- Recommendations

Introduction

- This is an analysis of Tanzanian water pump data in order to provide the Tanzania Government a tool with which to determine water pump functionality and information on how to improve pump maintenance efficiency. This analysis will focus to determining which factors needs attention to increase efficiency of the water pumps. The Tanzanian Government can use this analysis to improve prediction and identification of which pumps are non functional or may need repair, therefore increasing access to potable water across Tanzania.

Business problem

- Tanzanian Government wants to improve water pump maintenance operations
- They also need a way to better predict functionality status of water pumps
- It is also important for them to know characteristics that might lead to a non functional pump in the future

Data

- The data set contains various variables describing pump functionality status (the target variable), pump geographic location, what kind of pump is operating, when it was installed, how it is managed, etc. It includes data on 59,400 individual pumps recorded from 2011-2013.

Methods

- Created Decision tree classifier,Random Forest classifier model and Knearest Neighbour classifier model
- Resulting Decision tree model had an overall accuracy of 100%, meaning it could accurately predict the status of a given pump 100% of the time hence a sign that the model is overfit.
- Resulting Random Forest model had an overall accuracy of 78%, meaning it could accurately predict the status of a given pump 78% of the time
- Resulting K nearest neighbor model had an overall accuracy of 72%, meaning it could accurately predict the status of a given pump 72% of the time

Methods

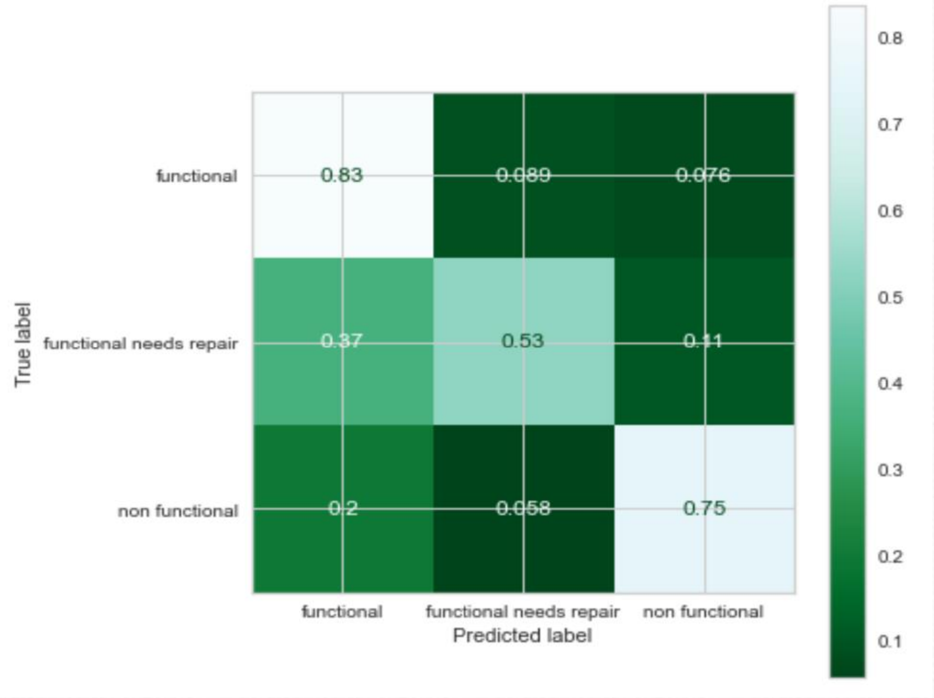
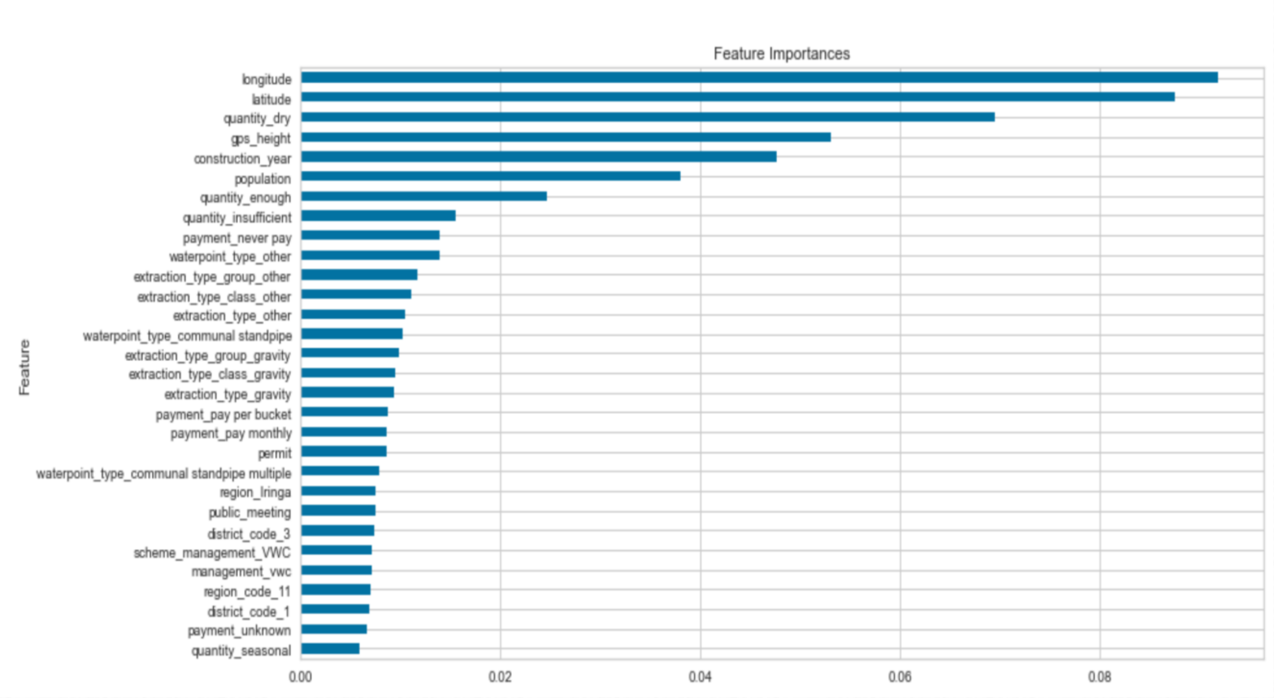


Figure shows the accuracy of the RandomForest model for predicting each status group

- Correctly predicts functional pumps 83% of the time
- Correctly predicts needs repair pumps 53% of the time
- Correctly predicts non functional pumps 75% of the time

Findings

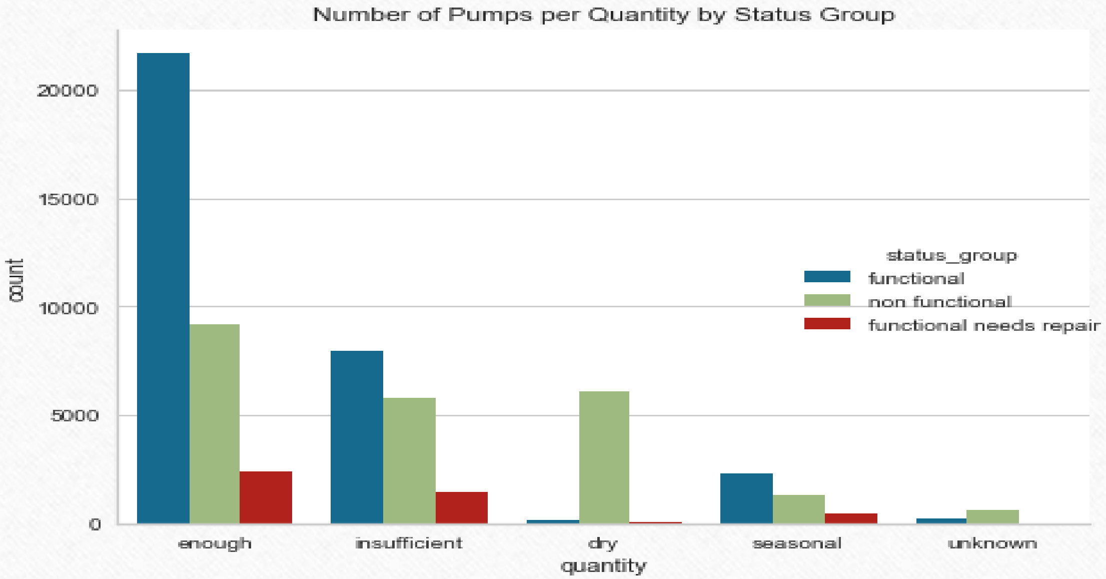


Random forest classifier model identifies several characteristics that are most important in identifying pump status:

- Pump location
- Pump water quantity
- Population surrounding pump
- Age of pump

Results

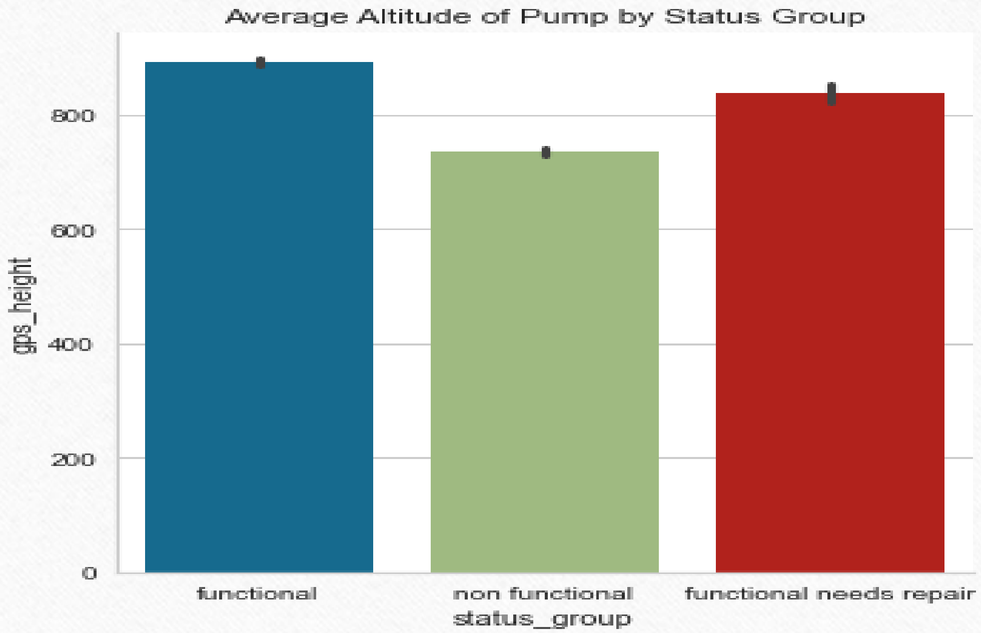
Water quantity vs the status of the pump



Pumps with lower water quantities may be more likely to be non functional or needing repair.

Results

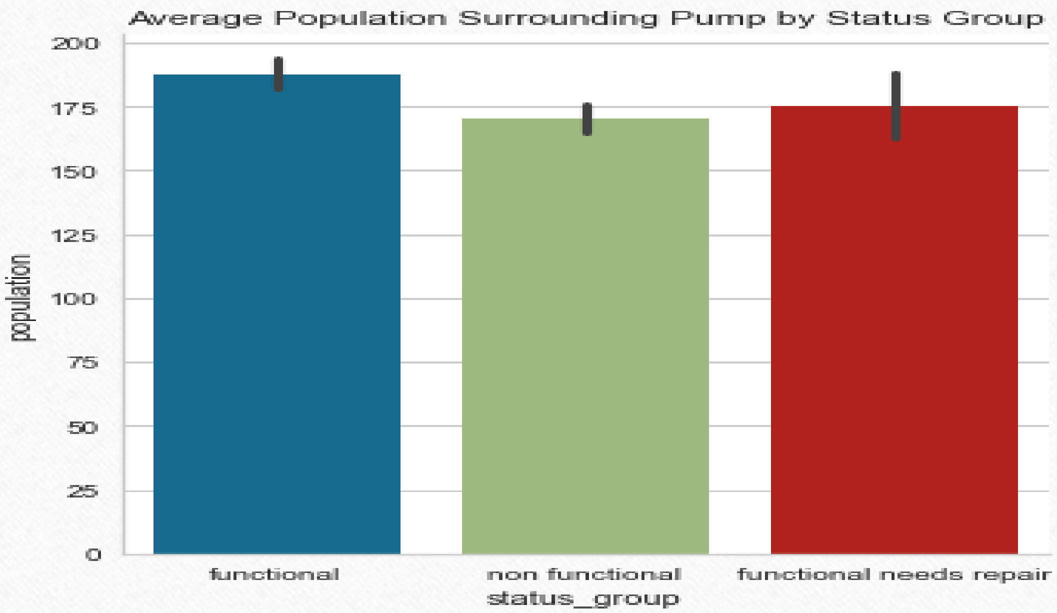
Location vs the status of the pump



The figure shows that on average, pumps at lower altitudes are more likely non functional or needing repair

Results

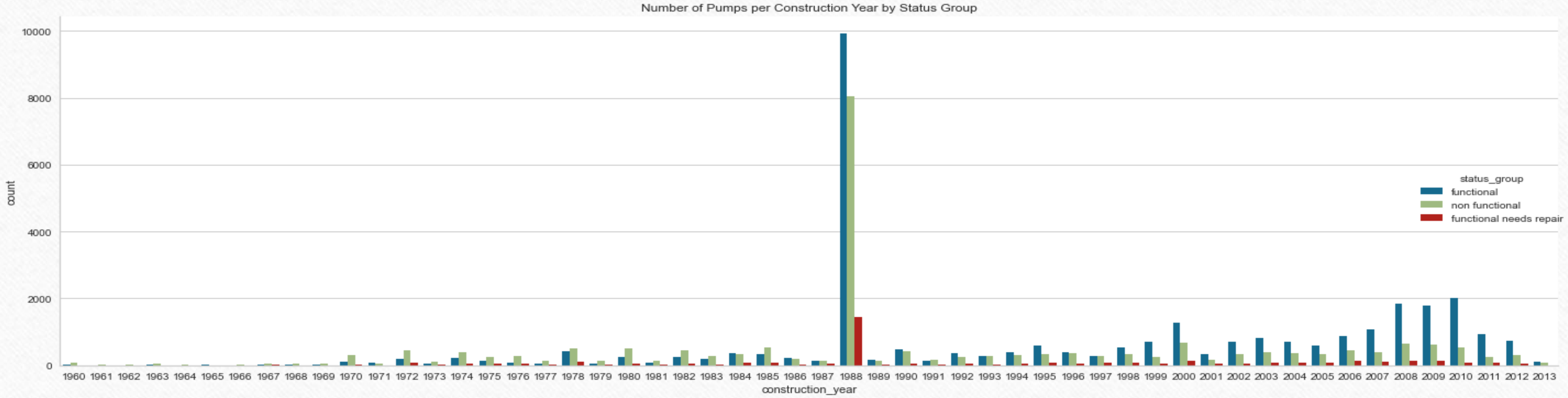
Population vs the status of the pump



Pumps in lower population areas may be more likely to be non functional or needing repair.

Results

Construction year vs the status of the pump



From the analysis above Older pumps may be more likely to be non functional or needing repair

Conclusion

- Location: The government should frequently monitor lower altitude pumps as they are likely to break down
- Quantity: The government should focus resources on pumps with low quantities of water.
- Population: The government should focus resources on low population areas, as they may not be receiving enough.
- Construction Year: The government should focus resources on modernizing older pumps

Recommendation

- The model and analysis are not complete solutions
- Model is overfit and still struggles withidentifying ‘functional needs repair’ pumps
- More data cleaning and use of various other models such as logistic regression and Xgboost may improve accuracy and reduce overfitting.