Pump it Up:
Data Mining the Water Table
Focus on Tanzania

PHASE 3 PROJECT PRESENTATION BY KIMANI J. IRUNGU.

PROBLEM STATEMENT.

The goal of this analysis and model determine with water pumps are faulty and hence enable access to clean water accross Tanzania reliably.

OBJECTIVE

The purpose of this analysis is to use classification modeling tecniques to accurately predict which water pumps are faulty and allow for stakeholders to repair/replace such and ensure consistent supply of clean water accross Tanzania.

DATA UNDERSTANDING.

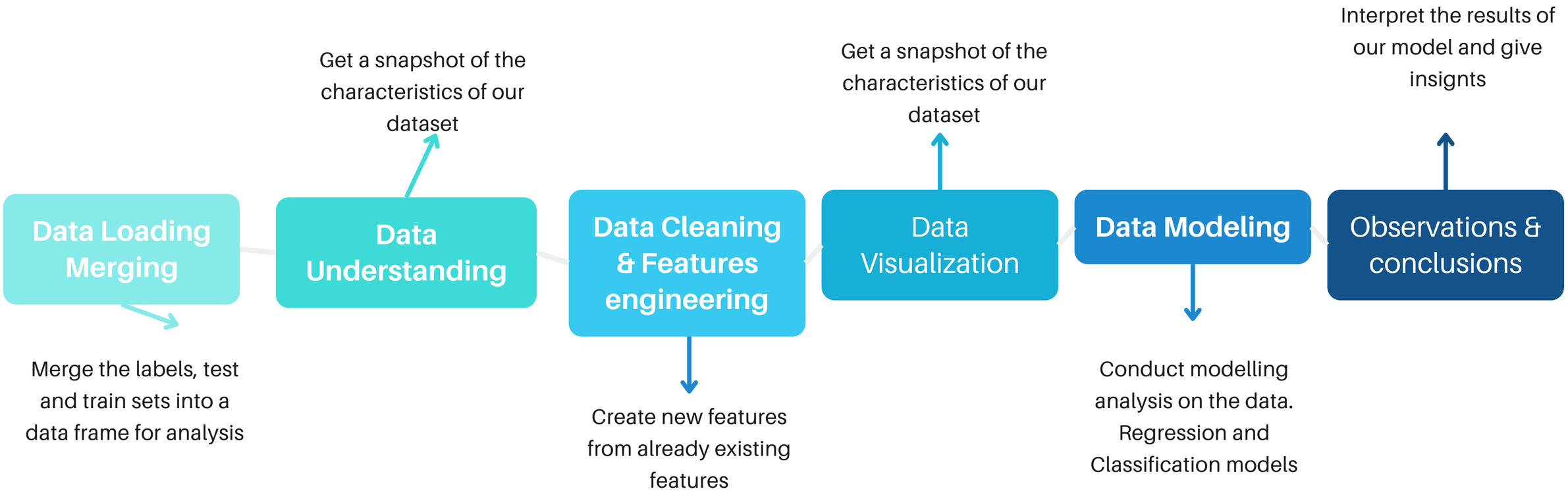
The dataset used in this modelling is derived from **https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/25/.**

The data contain information on water pumps recorded across the country of Tanzania. The Target variable is the status group of a pump given various characteristics of the pump including but not limited to Geographic location, Water quantity, GPS height location and Construction year among other features

SUCCESS CRITERIA

The objective is to obtain a prediction model that can correctly predict the condition of a water pump by 80% accuracy.

METHODOLOGY



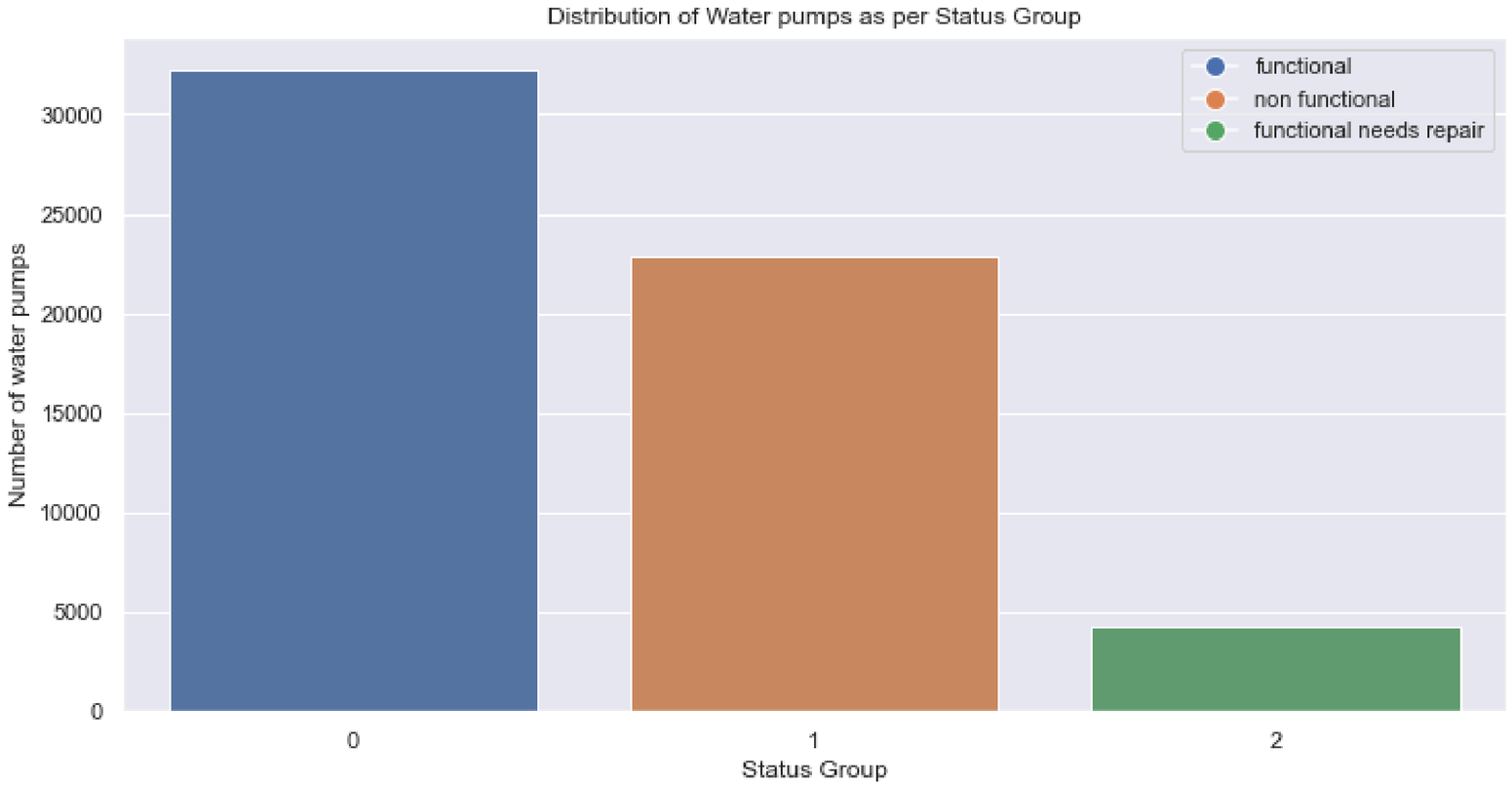
DATA UNDERSTANDING

Total Number of Pumps recorded 59,400

Functional Pumps = 54.3%

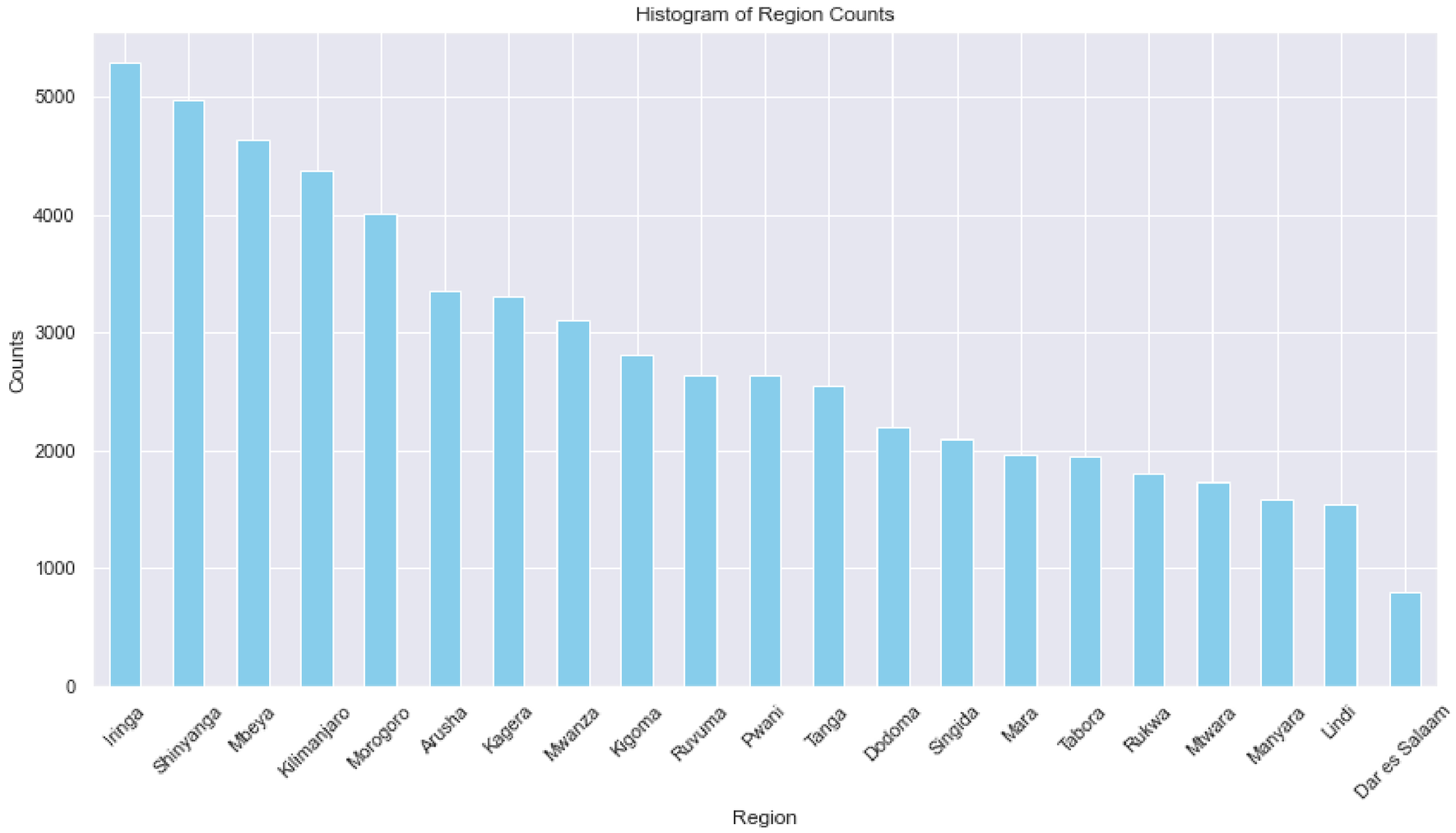
Non- functional = 38.4%

Functional need repair = 7.3%



DATA UNDERSTANDING

Distribution of Water Pimps per Region



DATA MODELING.

Multiple linear Regression

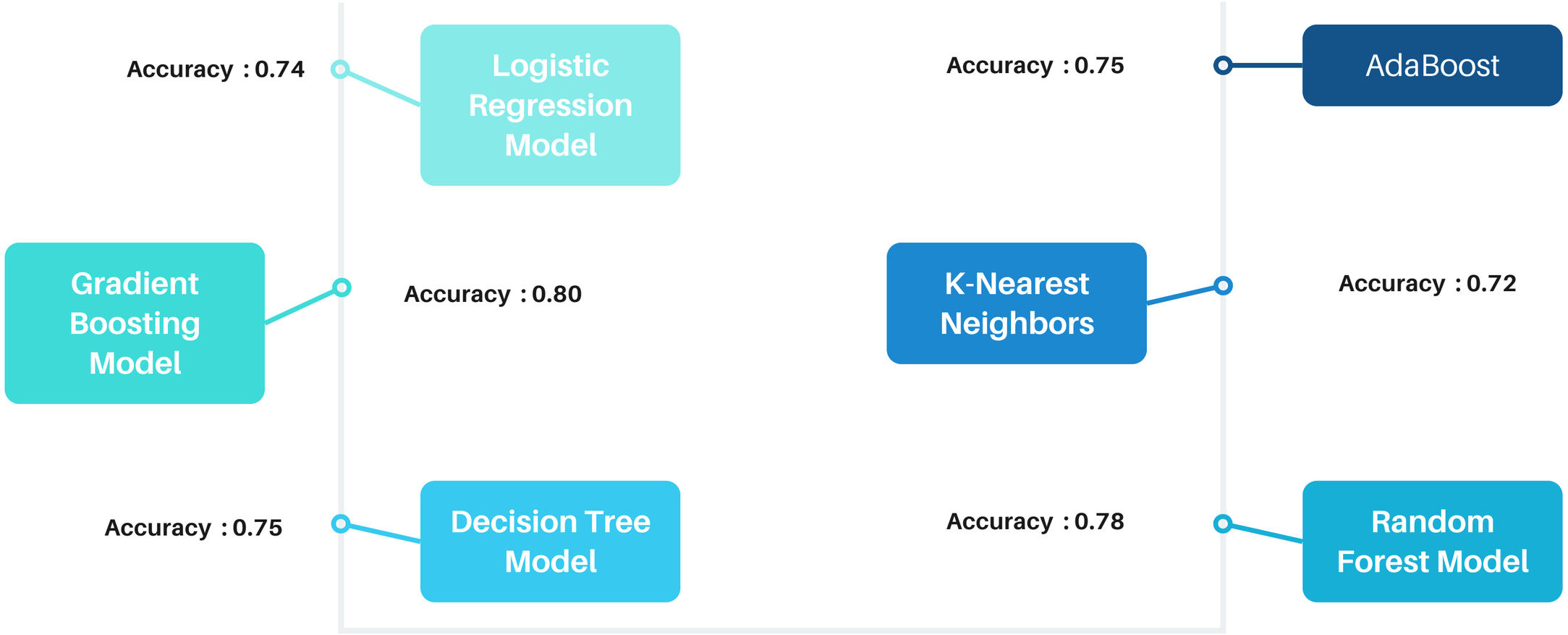
OLS Regression Results

Dep. Variable:	status_group	R-squared:	0.205
Model:	OLS	Adj. R-squared:	0.204
Method:	Least Squares	F-statistic:	139.2
Date:	Wed, 22 May 2024	Prob (F-statistic):	0.00
Time:	21:44:33	Log-Likelihood:	-49837.
No. Observations:	59400	AIC:	9.990e+04
Df Residuals:	59289	BIC:	1.009e+05
Df Model:	110		
Covariance Type:	nonrobust		

The Feature variables only explain 20% in the variability of our target variable
Further analysis is required to enable us increase the predictive power of our model, using classification models

CLASSIFICATION MODELING RESULTS

A brief history of the accuracy of each model.



CONCLUSION

Overall, The best perfoming classsication model is the Gradient Boosting Model with an accurancy of 80%.
Further, the random forest model performed well with an accuracy of 78%

The End