Improving Access to Clean Water In Tanzania

Group Members:

1) Ahmed Haji
2) Caroline Njoroge
3) Amadi Growman
4) James Nyamu
5) Robert Gesembe
6) Lynns Waswa
7) Peter Muhia

Business Understanding

In Tanzania, access to clean and reliable water sources is a pressing issue, affecting millions.

Our project aims to use data science to predict water pump faults, helping ensure continuous access to clean water.

Objectives

- To predict the functionality of water pumps based on installation features. (Main Objective)
- To evaluate factors influencing pump functionality.
- To identify and model feature combinations for accurate prediction.
- To test and validate the accuracy of the predictive model.
- To draw conclusions and offer recommendations based on the findings.

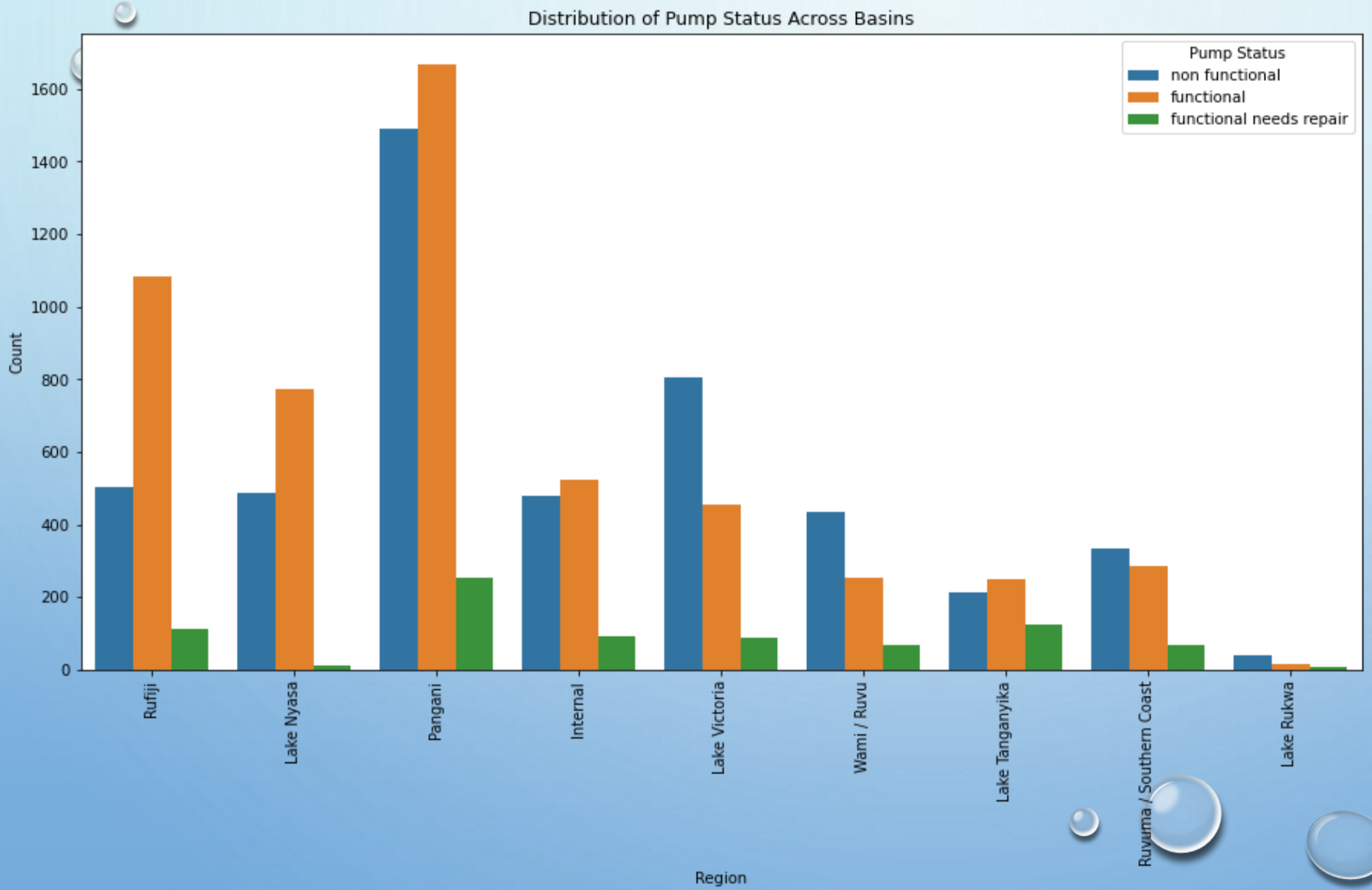
Data Understanding

The dataset originates from the Taarifa waterpoints dashboard, aggregating reports from the Tanzania Ministry of Water to predict the operational status of water pumps.

Data Preprocessing

- **Data Cleaning:** We made sure our water pump data was complete by filling in any missing information, especially for who funded the pump and who installed it.
- **Feature Engineering**: We figured out each pump's age from when it was built to help us understand which pumps might need more attention.
- **Encoding**: We organized the pumps' details, like water quality and location, into clear categories so our analysis can be more accurate.
- **Class balancing:** We adjusted the numbers to make sure no single detail about the pumps would be more important than another in our study

Exploratory Data Analysis



Pangani has the highest number of water pumps, both functional and non functional.

Methodology

We evaluated multiple machine learning models to predict water pump status.

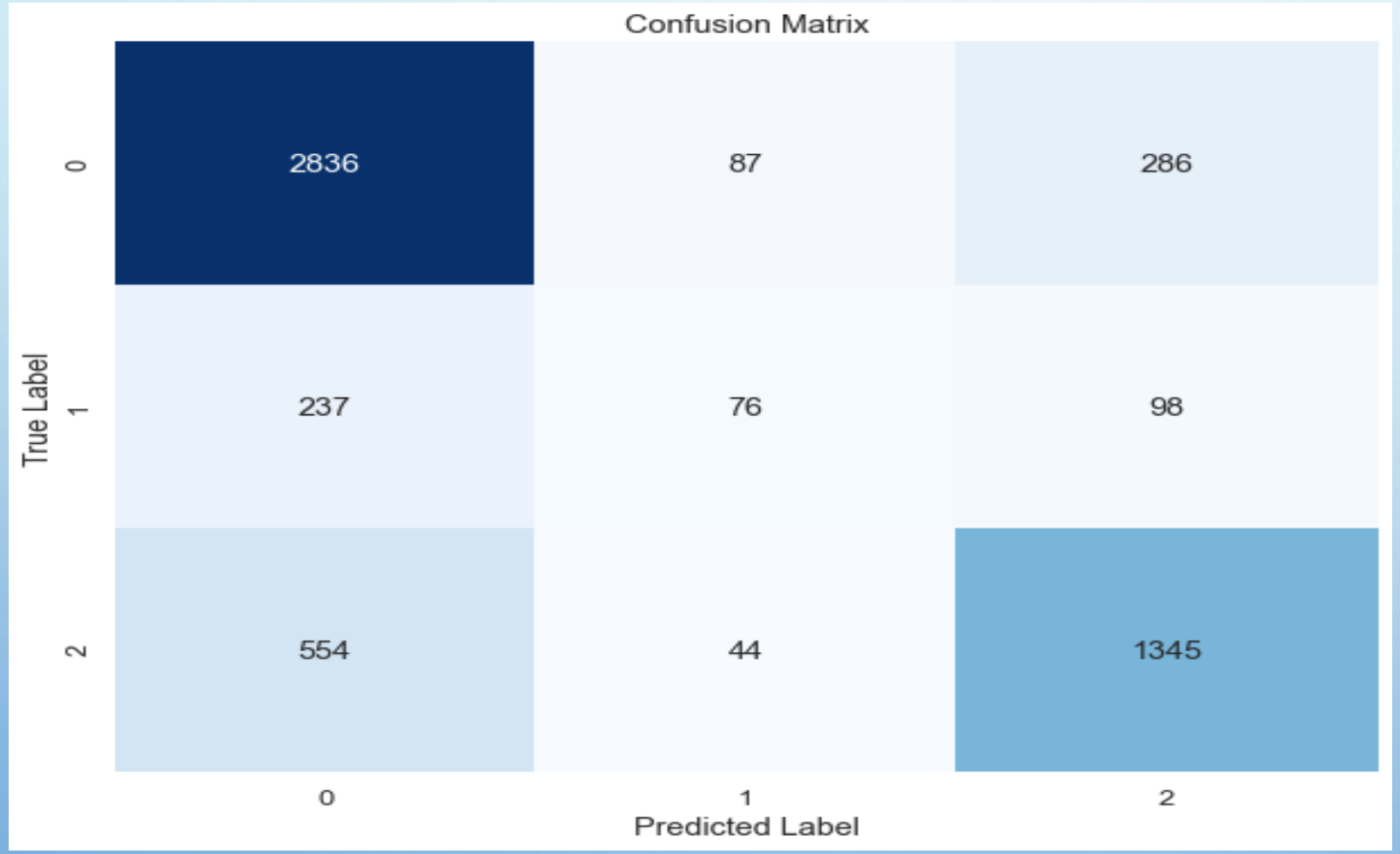
Models include:

- Logistic Regression
- Decision Tree
- KNN
- Random Forest
- XGBoost.

Model Evaluation Metrics

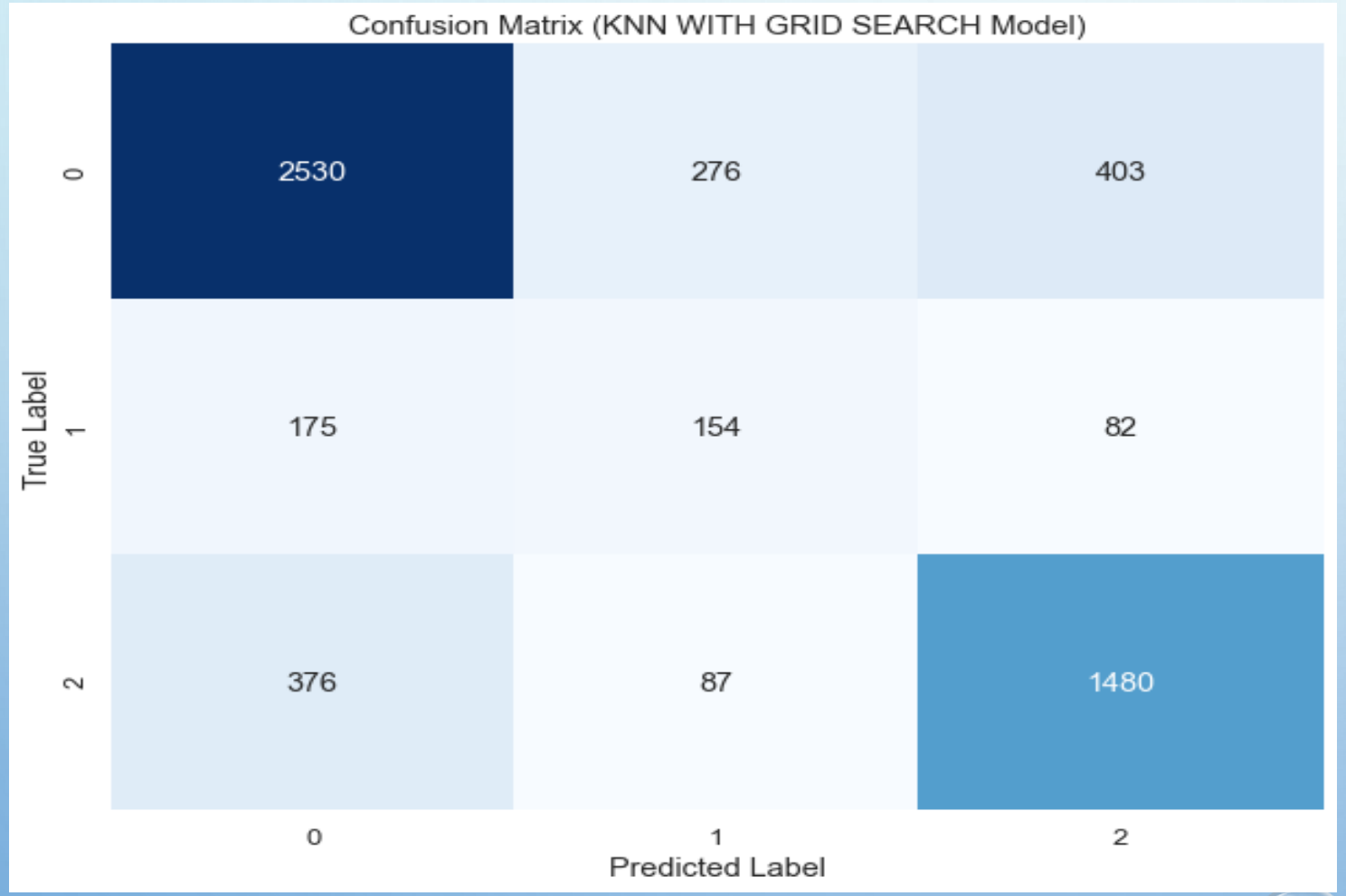
- **Accuracy** tells us the overall success rate of our predictions.
- **Precision** highlights the exactness of our positive identifications.
- **Recall** shows our ability to catch all necessary repairs.
- **F1-score** combines precision and recall for a quick model quality check.
- **SMOTE** helps us avoid bias by making sure all pump conditions are equally represented in the data.

Logistic Regression(Base Model)



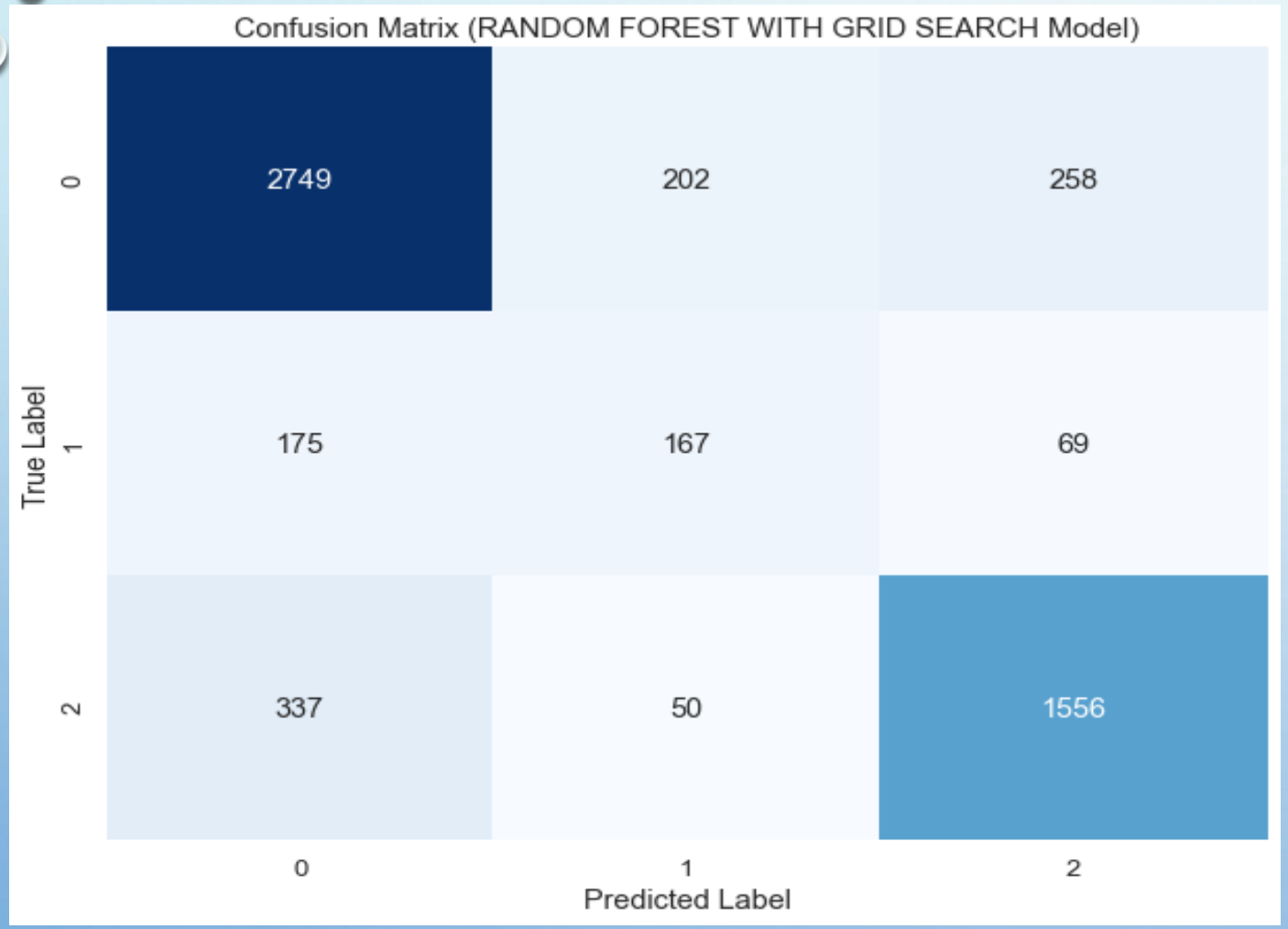
Fairly accurate prediction of labels, but not accurate enough.

XGBoost and KNN with GridSearchCV



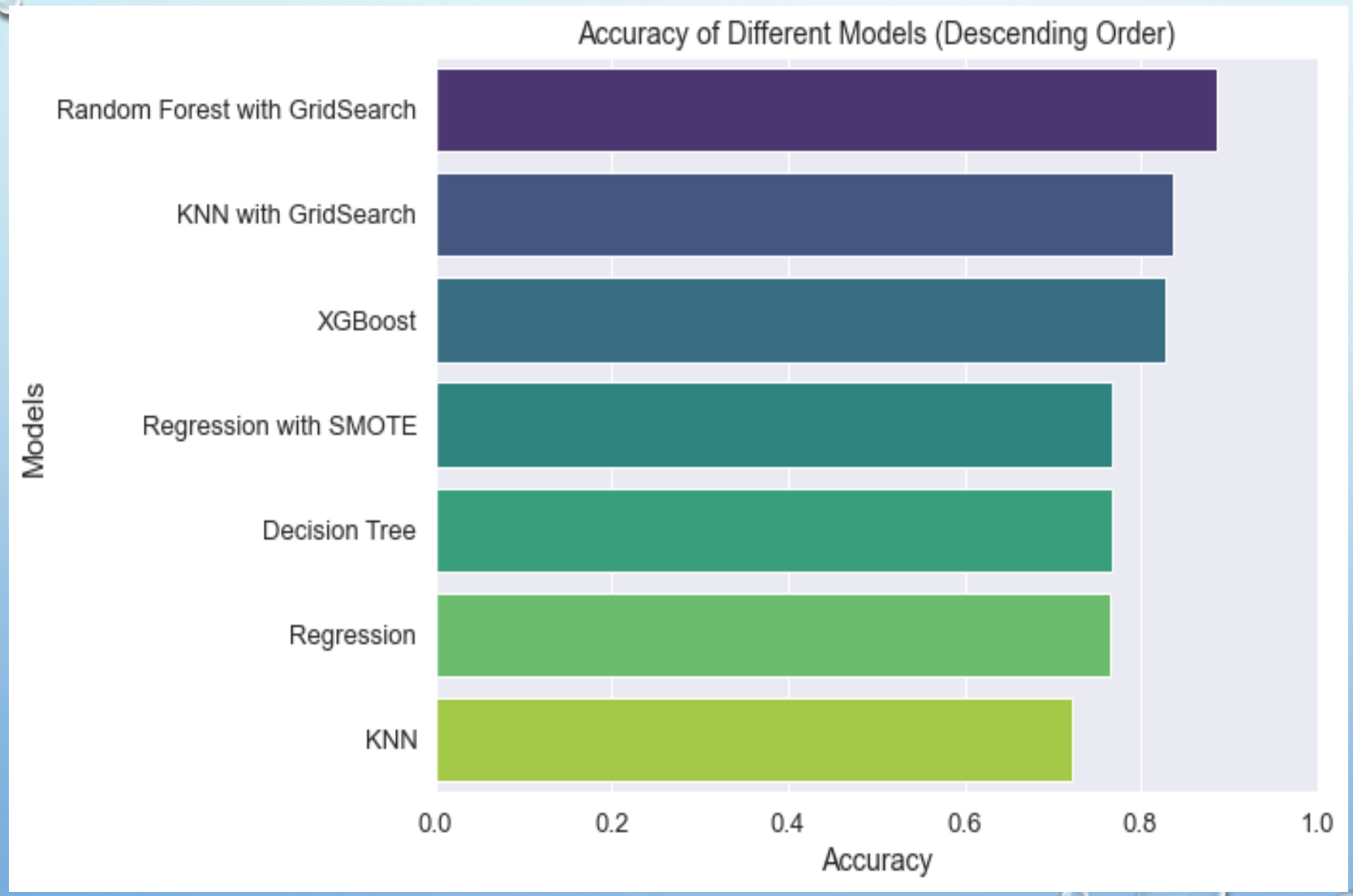
Comparatively better classification of labels than logistic regression

Random Forest with GridSearchCV



The model correctly classified a majority of instances across all classes

Model Comparison



Random Forest with grid search is the most accurate model

Conclusions

- Machine learning algorithms prove to be effective in predicting the status of waterpoints in Tanzania with notable accuracy.
- Regression models show significant promise, highlighting the predictive potential for waterpoint conditions with accuracies reaching over 76%.
- Decision Tree and KNN models offer reasonable predictions with accuracies around 78% and 72%.
- Random Forest and XGBoost models excel in performance, with Random Forest achieving the highest accuracy at 88.62%.

Recommendations

1. Pump Improvement & Data Collection
Research and deploy efficient pumps.
Implement structured data collection on pump performance.

2. Government Oversight & Collaboration
Strengthen oversight and quality assurance.
Encourage partnerships among government, NGOs, and communities.

3.Financial Strategies
Explore innovative financing for pump maintenance.
Utilize grants and community contributions

4. Continuous Data Monitoring
Set up real-time monitoring systems.
Engage community in data collection and reporting.

Questions