Water Well Functionality Prediction Report

Introduction

The purpose of this data report is to present a detailed overview of the findings and insights generated from the analysis of water pump functionality data in Tanzania. The data contains information on the type, amount of water, and the quality of water from various sources, as well as the infrastructure and geolocation information of the pumps.

Exploratory Data Analysis (EDA)

The results of the analysis revealed that there were a total of 59400 data points in the data set, with 41 variables representing different aspects of the water pumps. Some of the variables included:

- amount_tsh: Total static head (amount water available to water point)

- gps_height: Altitude of the well

- population: Population around the well

- status_group: the functioning status of the pump (functional, functional but needs repair, or non-functional)

The data was analyzed using various data exploration techniques such as univariate, bivariate and multivariate analysis.

Modelling

In order to build a predictive model for the water pump functional status, I used the Gradient Boosting Classifier and Random Forest Classifier algorithms. I first split the available data into training and validation sets to assess the performance of our models. Then, I performed a grid search over different hyperparameters for the Random Forest Classifier to find the best set of hyperparameters that would give us the highest validation accuracy.

Results

The best model I found using the grid search had a validation accuracy of 0. 796. The best hyperparameters found for the Random Forest Classifier were: 'min_samples_split': 8, 'n_estimators': 1000.

Conclusion

The model is able to predict the functionality status of a water pump with an accuracy of 0.796, which is a relatively good performance. This can be used to aid in decision making and maintenance planning for the water pumps in Tanzania.

However, it is important to note that this is a limited sample and there is room for improvement with more data or additional feature engineering.

Recommendations

- Further feature engineering could be done to improve the model performance.

- The model could be tested on additional data to ensure its generalizability.

- The model could be improved with the use of other algorithms or a combination of multiple algorithms.