



TANZANIA WATER WELLS

CONTENT

- 01 INTRODUCTION
- 02 MODELS EVALUATED
- 03 BASELINE MODEL PERFORMANCE
- 04 MODEL PERFORMANCE
- 05 CONCLUSION



MODEL EVALUATION FOR CLASSIFICATION TASK

The objective of this analysis is to evaluate different machine learning models particularly focusing on handling class imbalance and selecting the best-performing model to aid the Tanzanian Government in deciding which wells need to be visited first in order to repair and aid in the crisis the country is facing.



METHODOLOGY

Due to the many categorical features that can influence functionality, this project investigates features and their effects through two types of models in an attempt to best predict the well condition. This is done through the use of logistic regression and decision trees.

MODELS EVALUATED

The two main models evaluated are the Logistic Regression Model and the Decision Tree Model. These two models were then analyzed with SMOTE (Synthetic Minority Over-sampling Technique) to handle the imbalance.

Logistic Regression is a simple yet effective linear model often used for binary classification tasks. It's widely used due to its simplicity and effectiveness, especially when the relationship between the features and the target variable is linear.

**LOGISTIC
REGRESSION**

Decision Trees are non-linear models that can capture complex patterns in the data but are prone to overfitting.

DECISION TREE

SMOTE is a technique used to balance the dataset by generating synthetic instances of the minority class. This helps the model learn better from an imbalanced dataset by providing more examples of the underrepresented class..

**HANDLING
IMBALANCE
(USING SMOTE)**





Baseline Model Results

Logistic Regression:


- Accuracy: 93.64% – Indicates the proportion of correctly classified instances out of the total instances.
Accuracy shows how many predictions the model got right overall. With 93.64% accuracy, the Logistic Regression shows that it correctly classifies most instances.
- F1-Score: 92.18% – Balances precision and recall, providing insight into how well the model handles both classes. The considers both precision (the accuracy of positive predictions) and recall (how well the model captures positive instances). A score of 92.18% suggests that Logistic Regression is handling both the majority and minority classes well.



Baseline Model Results

Decision Tree:

- Accuracy: 56.06% – This is much lower than Logistic Regression, suggesting issues with generalization. This lower accuracy indicates that it's struggling to generalize beyond the training data, possibly due to overfitting.
- F1-Score: 60.03% – Lower score indicating poor performance, likely due to overfitting. The F1-score further confirms that the Decision Tree is not handling the minority class well, and overall, it's not performing reliably.
- Overall, Logistic Regression significantly outperforms Decision Tree in both accuracy and F1-score, making it the better baseline model.




SMOTE Model Performance

Logistic Regression:

- Accuracy: 93.65% – Slight improvement after applying SMOTE, showing better handling of the minority class.

After applying SMOTE, the accuracy of Logistic Regression improved slightly to 93.65%, indicating that balancing the classes helped the model perform even better.

- F1-Score: 92.45% – Improved F1-score indicates better balance between precision and recall after addressing the class imbalance. The F1-score improving suggests that the model now has a better balance between precision and recall, due to the more balanced dataset.



SMOTE Model Performance

Decision Tree:

- Accuracy: 91.69% – There is significant improvement compared to the baseline, but still not as strong as the Logistic Regression.
This significant boost in accuracy shows that SMOTE helped it generalize better. However, it still falls short of Logistic Regression model.
- F1-Score: 91.87% – Better F1-score, but slightly lower than that of Logistic Regression.
Similarly, the F1-score improves which is good, however, it doesn't quite match the performance of Logistic Regression with SMOTE.
- Applying SMOTE improves the performance of both models, but Logistic Regression still provides the best results overall.

CONCLUSION

It is clear that the best model is the Logistic Regression with SMOTE.

This is because the Logistic Regression model with SMOTE has the highest accuracy (93.65%) and a strong F1-score (92.45%), demonstrating that it effectively handles both majority and minority classes after applying SMOTE. It provides a good balance between simplicity, interpretability, and performance, making it a robust choice for this classification task.

Therefore, based on the analysis, Logistic Regression with SMOTE is recommended as the model for deployment in this classification task.

