Moringa School

PHASE 3 PROJECT.

JEANMARIE WACHIRA


Tanzania Water Wells

Overview

Presentation

- Business and Data Understanding
- Modeling
- Evaluation
- Recommendations

Project Introduction

"In Tanzania, over 57 million people face a major challenge of accessing clean water. A classifier will be developed to predict the condition of water wells based on factors like pump type and installation date. This will benefit NGOs focused on repairing wells or the government looking to improve new well construction."

Business Problem

Our goal is to create a classifier that predicts the condition of water wells based on factors such as the type of pump and installation date

Data understanding

The data used in this analysis was obtained from DrivenData and was sourced from Taarifa and the Tanzanian Ministry of Water.

Water Pump Labels

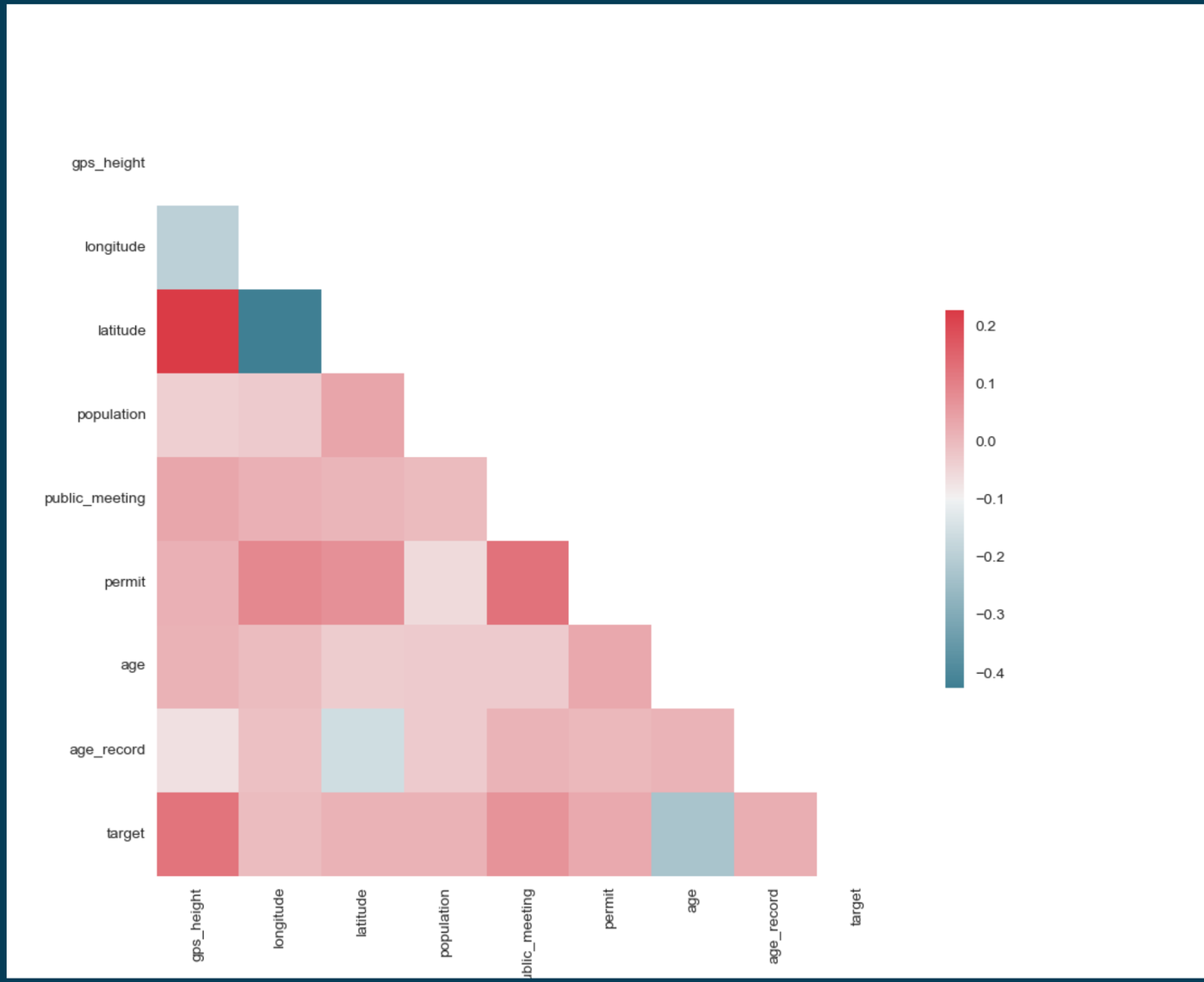
- Functional
- non Functional
- Functional needs repairs

MODELING

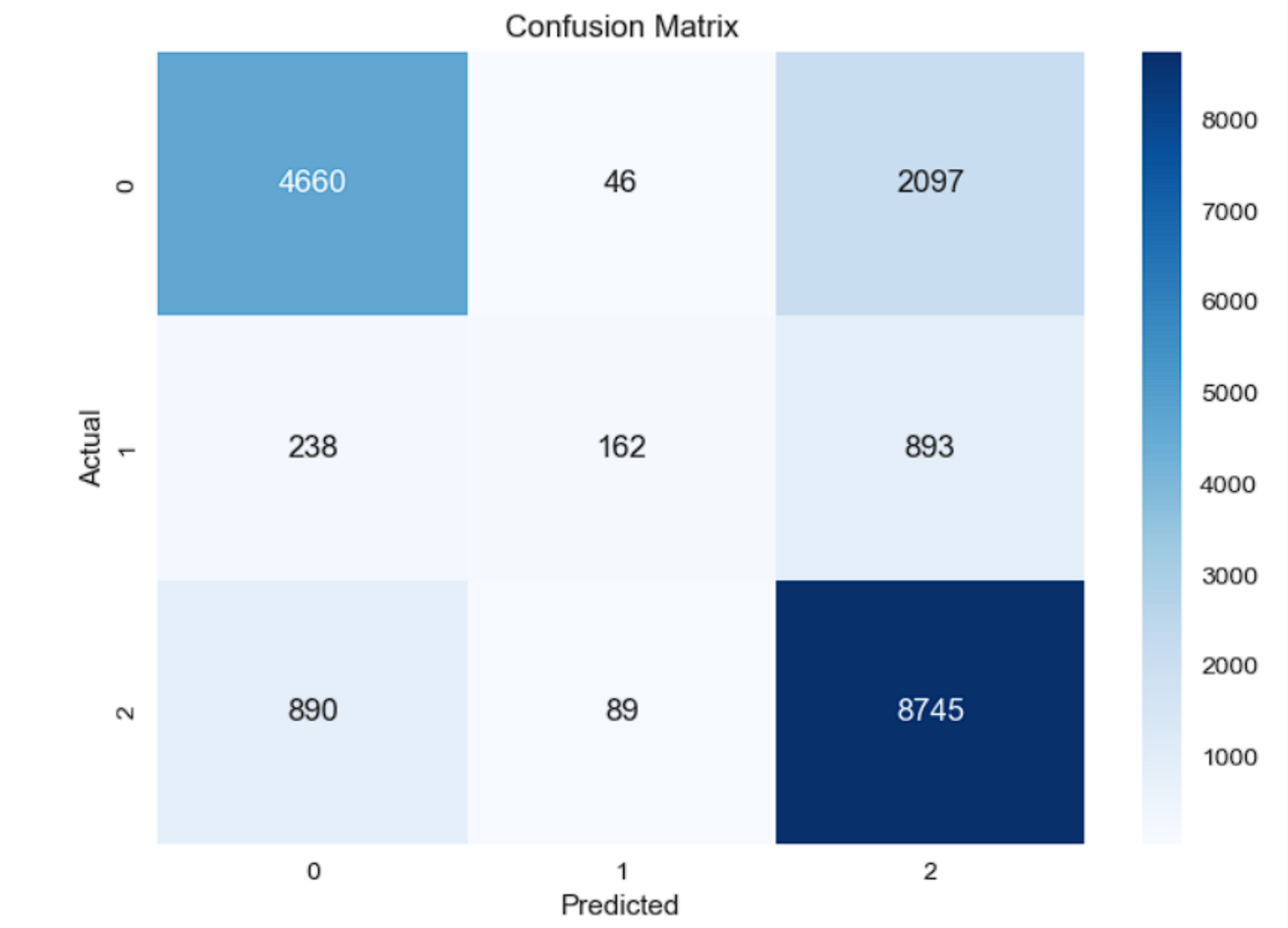
Correlation

Ocean pH levels since 2013

- The correlation matrix shows the relationship between features and the target value.
- Helps understand the degree of association between features and target.
- Enables making informed decisions regarding which features to include in the models.



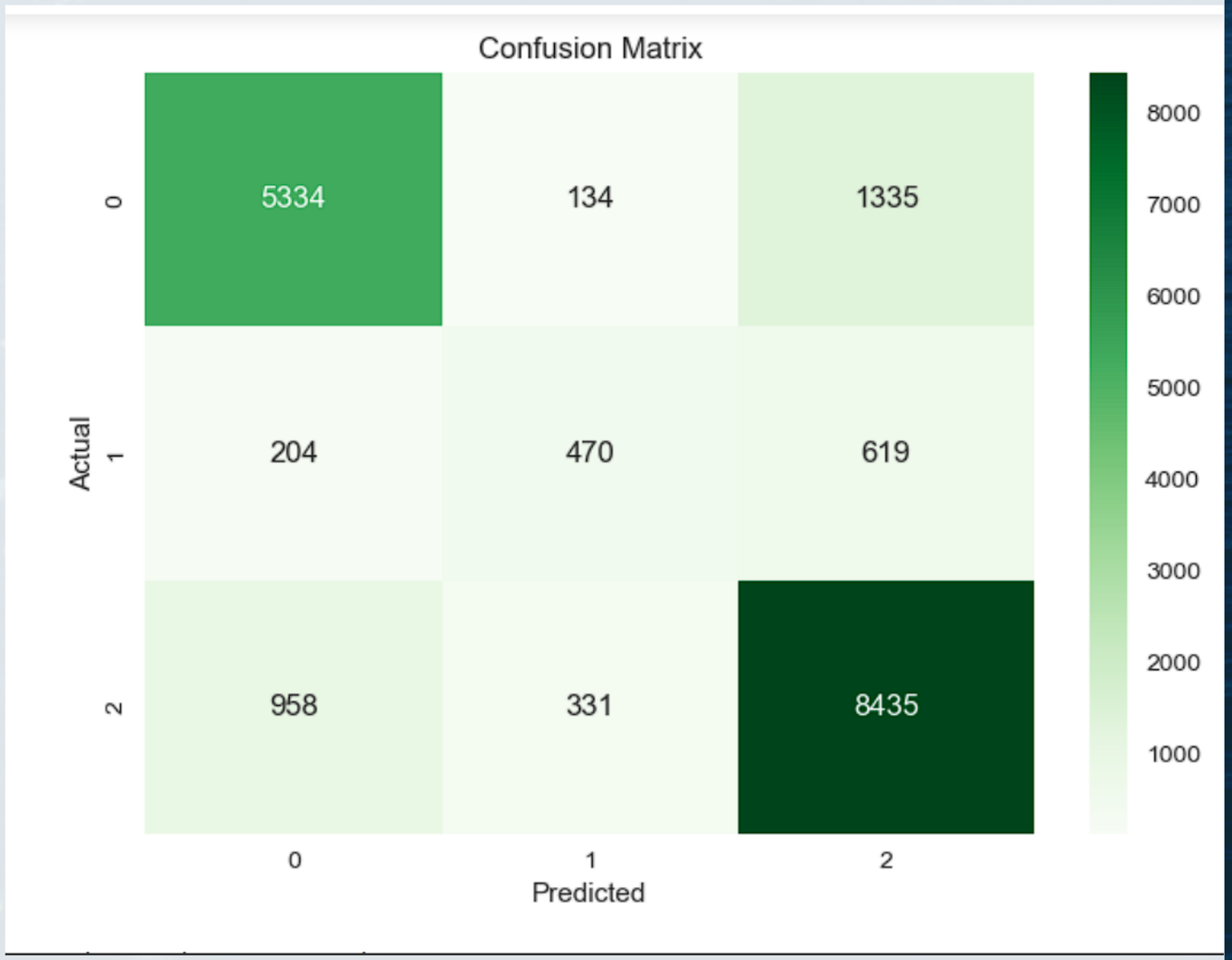
Logistic Regression



- The model performed well in predicting the "functional" class but was average in predicting the "repair" class and poorly in predicting the "non-functional" class.
- False positives are important to avoid, so the model needs improvement.

Random Forest

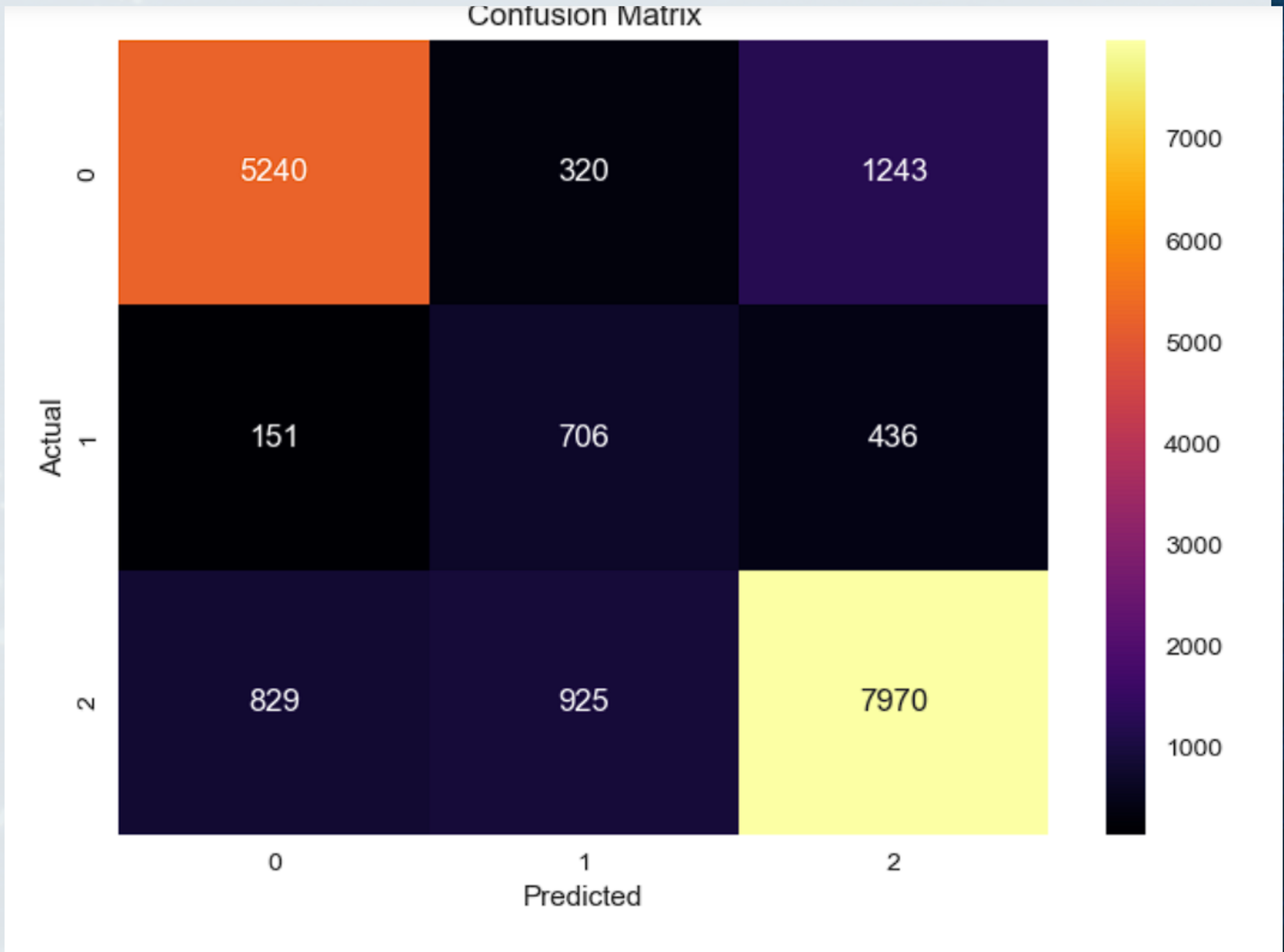
- Fine-tune the hyperparameters of the decision tree model for better performance.
- Use Grid Search to determine the best set of hyperparameters.
- Evaluate the performance of different hyperparameter combinations using cross-validation.



SMOTE

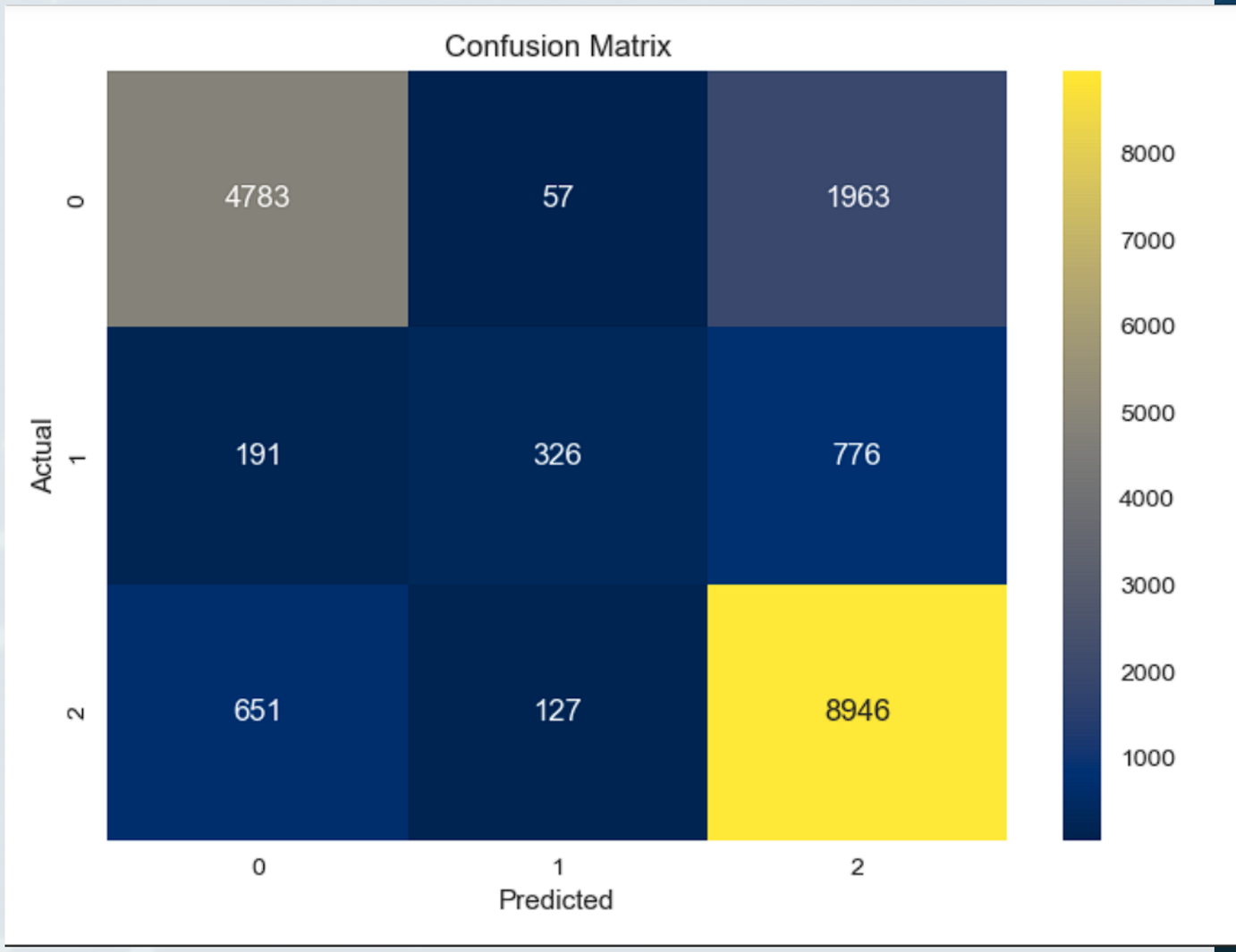
- The score of the model has decreased but the accuracy of Non-Functional and Repair pumps predictions has improved.
- The model is more accurate in its predictions for specific classes.
- The model is an improvement over previous models

.



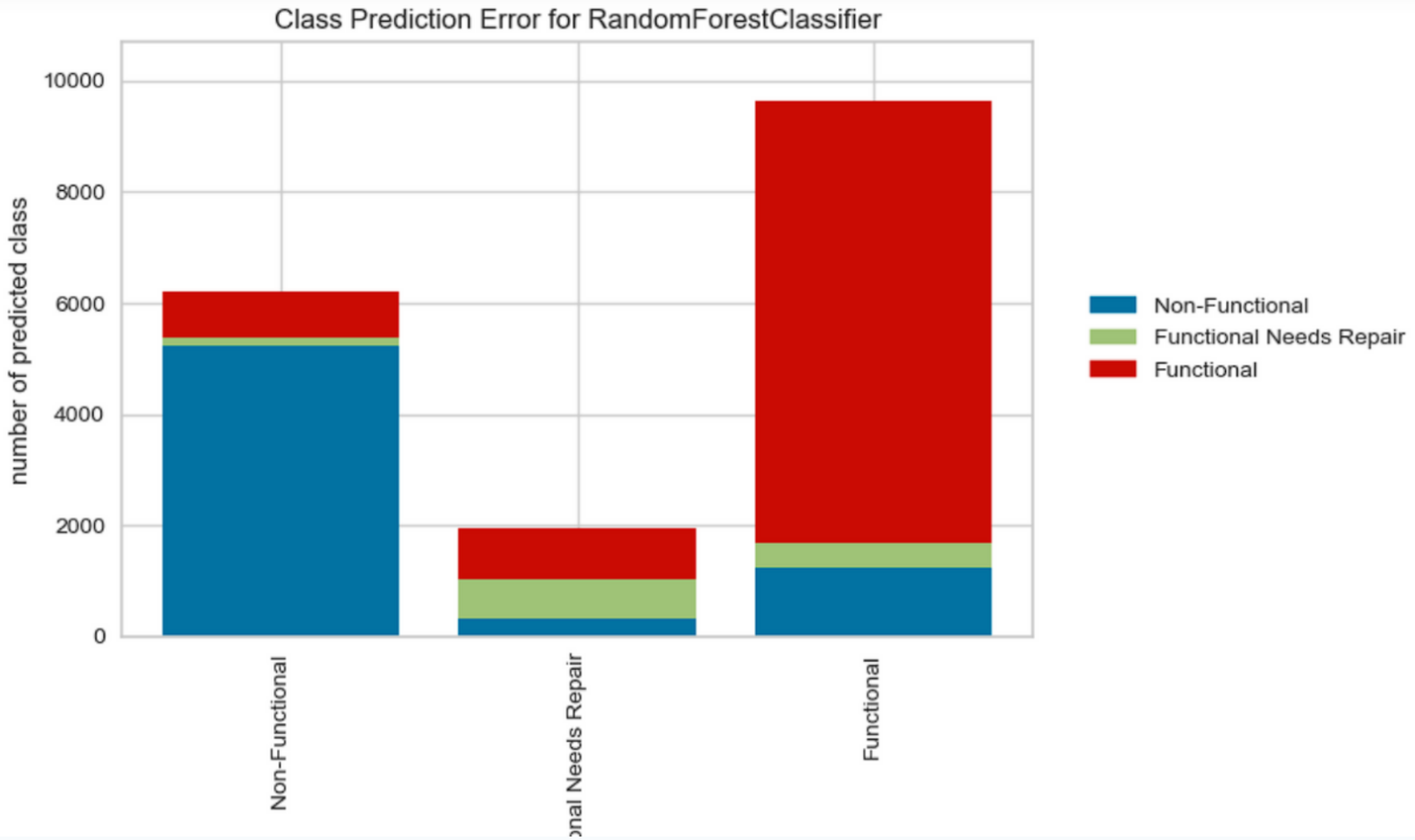
XG Boost

- The decision tree model is not performing as well as the Random Forest model.
- The performance of the decision tree model can be improved through hyperparameter tuning and balancing the data using SMOTE.
- Explore these avenues to determine if the decision tree model can surpass the Random Forest model.

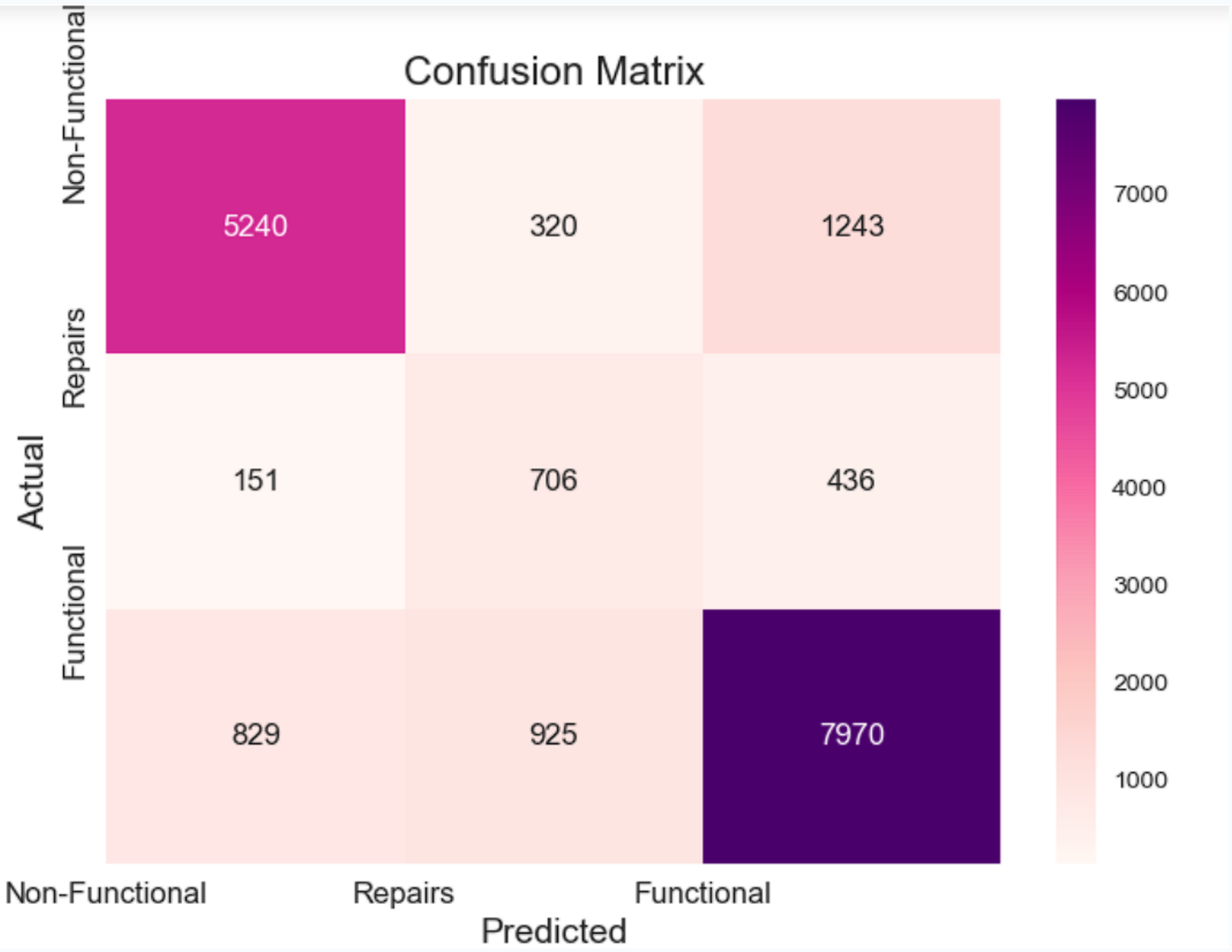


EVALUATION

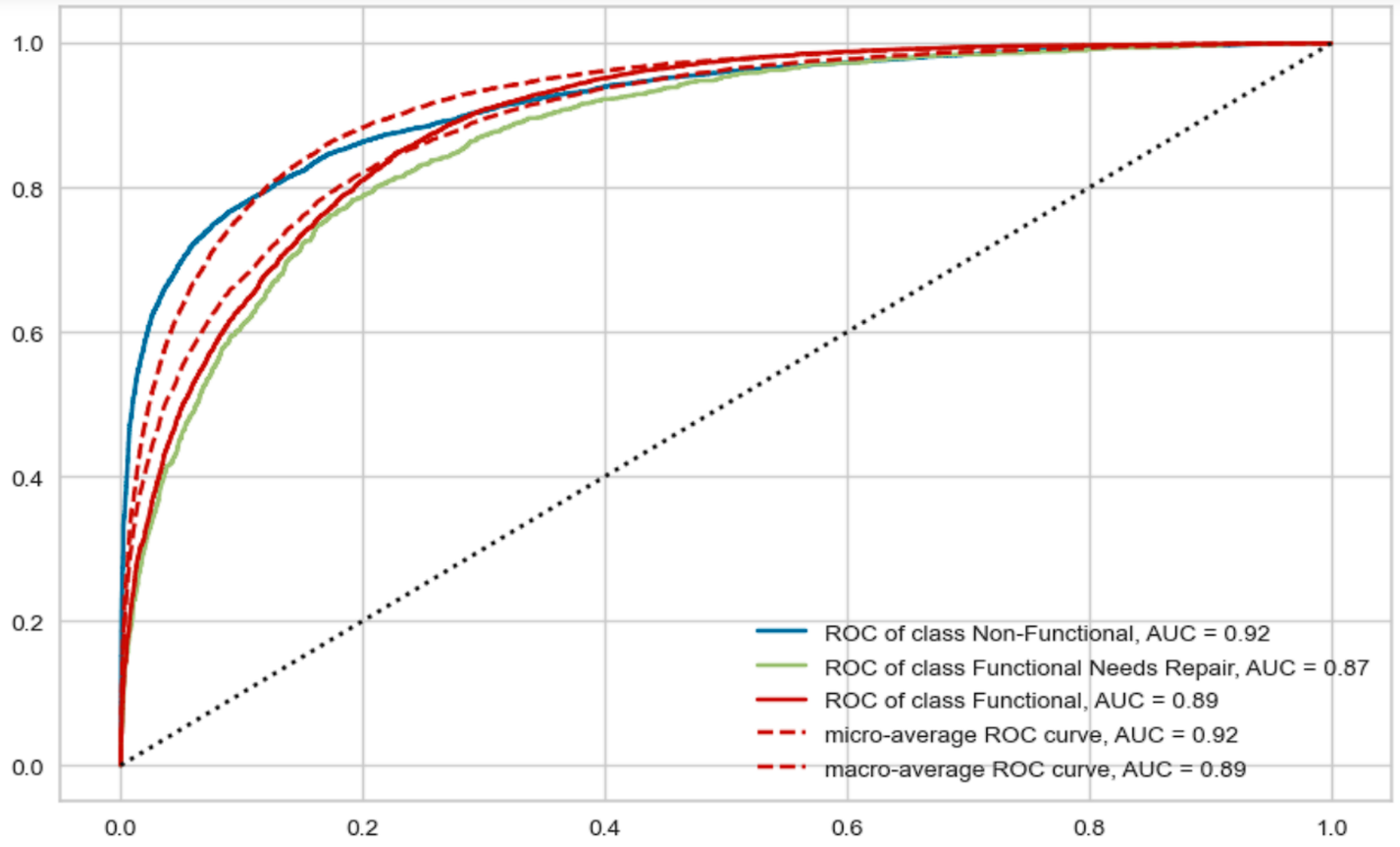
- XGBoost and Random Forest have comparable results.
- Random Forest performs slightly better in predicting the "Repair" class.
- XGBoost has a slightly higher overall accuracy.
- Further investigation and fine-tuning is necessary to determine the best fit.
- Consider other algorithms or models to see if they perform better.
- Random Forest Classifier is the best fit for the dataset.
- Random Forest performed faster than XGBoost.



- Try different classifiers that handle multi-class problems better.
- Add additional features to the data to improve distinction between "Functional Needs Repair" and "Functional".
- Experiment with different ensemble models and resampling techniques.
- Test the model on unseen data for a more generalizable evaluation



- False predictions of functional need Repairs
- Performed well at minimizing false positives



- ROC curve displays the model's performance in making predictions for each class.
- Closer to the top left corner and the higher the value approaches 1, the better the performance.
- Model performs well in classifying "Non-Functional" and "Functional".
- Room for improvement in classifying "Functional Needs Repair".

Recommendations

- Focus was primarily on human wellbeing, and cost analysis was limited due to lack of cost data
- Future studies should prioritize collection of cost data for comprehensive cost-benefit analysis and consider available resources for improved efficiency
- Further investigation of alternative models/algorithms may be beneficial for the unique characteristics and business requirements of the dataset.