Creating and Comparing Classification Models for Predicting Water Point Status Using DrivenData's Pump it Up Dataset: A Data Science Approach

Aparna Krishnakumar
School of Computer Science
University of Nottingham
Nottingham, United Kingdom
psxak18@nottingham.ac.uk

Neha Unni
School of Computer Science
Universirty of Nottingham
Nottingham, United Kingdom
psxnu1@nottingham.ac.uk

***Abstract*—The focus of this paper is to analyze and predict the functionality of water pumps in Tanzania. We employed different data mining techniques on the data set provided, which contains several features of the water pump, including the quality, quantity and location. Multiple machine learning algorithms were then used to train and evaluate the models. By analyzing the various features of the water points, the models were able to pinpoint the factors that contribute towards the operational status of the water pumps with a reasonable level of accuracy. We implemented Random Forest, XGBoost, CatBoost, GradientBoost and TabNet algorithms along with a stack classifier to predict the accuracy of the model. We were able to achieve 80.68 as the highest accuracy after comparing the models.**

***Index Terms*—Classification, Random Forest, Stack Classifier, Deep Learning, Hyper parameter tuning**

I. Introduction

The data collection was obtained via a drivendata.org online data mining competition. The information is derived from the Taarifa waterpoints dashboard, a Tanzanian innovation initiative that compiles data from the ministry of water there. The collection comprises 40 variables, such as waterpoint type, water quality, and water amount, and information on more than 59,400 water points. The goal of this dataset is to predict the functional status of a water point, which can be classified as "functional", "functional but needs repair", or "non-functional". Based on the data analysis the research question on focus is, how does geographic factors such as geographic location, extraction type, source of water, management group and population predict the overall performance (functionality, water quality, cost) of waterpoints in Tanzania's different water basins?

The dataset contains both numerical and categorical variables. The summary statistics reveal that the range and mean values differ significantly among the numerical variables. Similarly, the categorical variables have a varying number of unique values, with some columns having only one unique value. Overall, the dataset contains a wide range of information that can be explored further to gain insights into the water wells in Tanzania. A statistical visualization with the numerical and categorical variables is provided by the below charts (Fig. 1., Fig. 2., Fig. 3.)

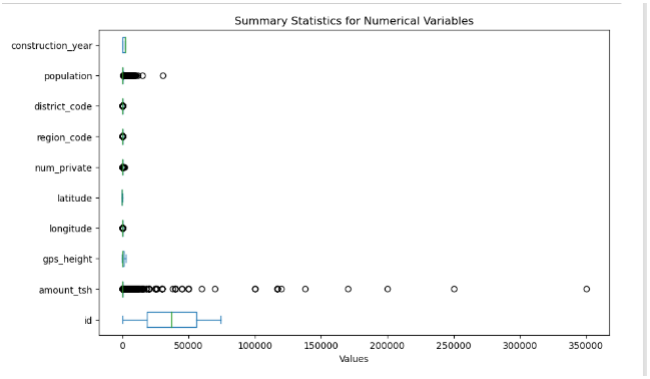

Fig. 1. Summary statistics for numeric variable

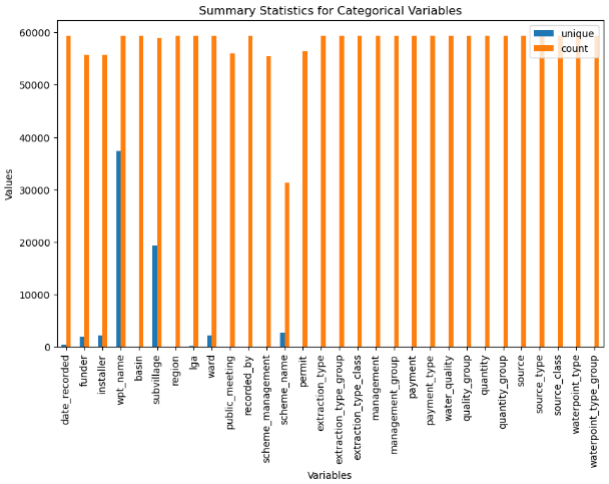

Fig. 2. Summary statistics for categorical variable

II. Literature Review

During the literature review, several approaches and methods were discovered for addressing the problem of predicting

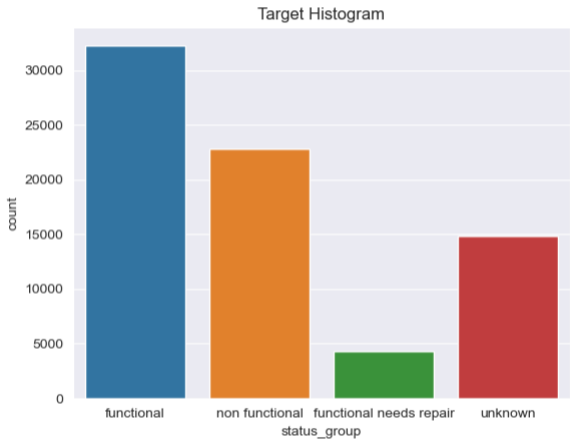

Fig. 3. Summary statistics for target variable

the functionality of water pumps. Different machine learning algorithms were applied in various studies, showcasing their effectiveness in this domain.

One notable approach is TabNet, which combines tree-based algorithms with neural networks to enable attentive tabular learning [2]. TabNet was able to identify complex relationships between water points and outperformed traditional algorithms such as XGBoost and LightGBM. Even in scenarios where the data was imbalanced or scarce, TabNet achieved an accuracy(in percentage) of 83.6, surpassing other models that scored around 79 and 78.

In another comparative study [3], various machine learning models were evaluated, including Logistic regression, Decision trees, Random forests, Support vector machines, XGBoost, LightGBM, and TabNet. Again, TabNet emerged as the best-performing model with an accuracy score of 93.2, highlighting its superior predictive capabilities.

Furthermore, deep learning methods were explored in the prediction of water pump functionality [1]. A fully connected neural network, specifically a 9-layer model, yielded the highest accuracy among the tested architectures. The model employed the Adam optimizer, which incorporates RMSprop and momentum for efficient learning.

Additionally, geospatial features were leveraged in a related work [4], emphasizing their importance in understanding and predicting water pump functionality. Decision trees, Neural Networks, Multinomial Logistic regression, and Random Forest were employed to classify water pumps based on geospatial factors. Random Forest demonstrated the highest accuracy in classifying the pumps, as determined by the validation misclassification rate.

Among these approaches, TabNet was consistently favored as the most efficient and accurate model in multiple studies. Several factors contribute to its superiority, including its ability to automatically learn the importance of complex features through novel feature selection. TabNet also exhibits robustness when dealing with noisy data, further enhancing its performance.

III. Methadology

The dataset required some pre-processing before it could be used effectively for further analysis. Preparing the data included identifying and handling the missing values and outliers using different techniques. The data was cleaned by fixing any errors in the data such as misspellings or incorrect values, converting variables to much appropriate data types and removing any irrelevant (like variables with large amount of zeros) or redundant variables.By applying appropriate techniques such as imputation or removal, we ensure that the dataset is complete and free from extreme values that may skew the analysis or model training. New features were created from the existing variables that can provide additional information or insights. For example, A new feature age was created that represents the approximate age of the water point from construction year and the recorded year.Machine learning algorithms typically require numerical inputs, so categorical variables need to be transformed into a numerical representation [5]. Categorical variables were converted into numerical variables using appropriate encoding techniques such as one-hot encoding and factorisation. Cleaning of the data effectively increased the consistency of the data and improved its accuracy.

We analysed the dataset in-depth in this study using a wide range of machine learning models. We first chose well-known models as our baseline models, such as Random Forest, XGBoost, CatBoost, GradientBoost, and TabNet[6]. These foundational models were selected because to their shown proficiency in handling a range of datasets. We used them in our analysis with the intention of establishing a performance benchmark against which to evaluate alternative models and methodologies. Using default parameter settings, these models were trained and assessed.

We understood the significance of hyperparameter adjustment and how to further improve the performance of our models [7]. Therefore, we conducted hyperparameter tuning for Random Forest, XGBoost, and CatBoost. Hyperparameter tuning involves systematically searching through different combinations of hyperparameters to identify the optimal settings that maximize the models' predictive capabilities. This iterative process allowed us to fine-tune the models and achieve better results.In order to navigate the vast parameter space and identify the optimal model configuration, the hyperparameter tuning step is crucial. Through hyperparameter adjustment, we aimed to improve prediction performance by enhancing the models' capacity to perceive the underlying relationships and patterns in the dataset. Using a validation dataset, the models were evaluated throughout the tuning phase, and the hyperparameter values that gave the greatest performance were selected.

In this investigation, we included a Stack Classifier in addition to the baseline models and hyperparameter adjustment. Random Forest and CatBoost, two potent models with contrasting traits, were merged to create the Stack Classifier. The Stack Classifier's underlying premise is that stacking the

predictions of many models can produce a classification that is more precise and reliable [8].With the Stack Classifier, many models are trained, and their combined predictions serve as the basis for the final prediction. We wanted to take use of the unique benefits of both models and harness their range of capabilities by merging Random Forest and CatBoost. By identifying a broader range of patterns and connections in the data, this ensemble technique has the potential to enhance classification results.

By employing this comprehensive approach, we aimed to thoroughly explore the capabilities of various models and provide a comprehensive analysis of their effectiveness for the specific dataset. This methodology allowed us to assess the strengths and weaknesses of different models and gain insights into their performance characteristics.

IV. Results

A. Data Analysis

During Exploratory data analysis, the relationships between various factors and the status of water pumps were explored using visualizations and correlation analysis. The analysis begins by examining the impact of different variables on the functionality of water pumps. Plots are generated to investigate how factors such as region, basin, scheme management, extraction type class, management group, payment type, water quality, source, waterpoint type, and age are associated with the status groups of the pumps. The plots provide a visual representation of the distribution of status groups within each variable, allowing for a better understanding of their influence on pump functionality. The data analysis revealed several important findings regarding the relationship between different variables especially geographic factors and the status of water pumps.

- On observing the plot for regions, Region 11 had the highest number of functional pumps, while Region 8 had far fewer pumps and a higher number of them were non-functional pumps. This indicates a significant variation in pump functionality across different regions (Fig 4.(2)).
- Similarly for basins, Lake Victoria had the highest number of pumps, with almost an equal distribution between functional and non-functional pumps. However, the Rufiji basin had the highest number of functional pumps compared to non-functional pumps (Fig 4.(1)).
- Examining payment types, it was found that pumps for which payment was never made had the highest number of non-functional pumps (Fig 4.(4)).
- The analysis of extraction types revealed that most pumps had gravity based extraction type. More than half of the pumps using gravity were functional, but still there is evidence that shows considerable amount of pumps are non functional under gravity type. On the other hand, wind-powered pumps were rarely used (Fig 4.(3)).
- When considering scheme management, the VWC had the highest number of pumps in the continent. Among these, an equal number were functional and non-functional, while there were fewer pumps with an unknown scheme management (Fig 5.(1)).
- Considering water quality, while the plot shows that more than half of the pumps using soft water are functional, the evidence also suggests that high number of non functional pumps also use soft water (Fig 5.(2)).
- Shallow wells were associated with the highest number of non-functional pumps among different water sources (Fig 6.).Furthermore, communal standpipes were predominantly non-functional among different waterpoint types (Fig 5.(4)).
- The analysis of age indicated that as the age of the water pumps increased, the number of non-functional pumps also increased. This suggests that older pumps are more likely to require maintenance regularly.

These findings provide valuable insights into the relationships between various factors and the functionality of water pumps.

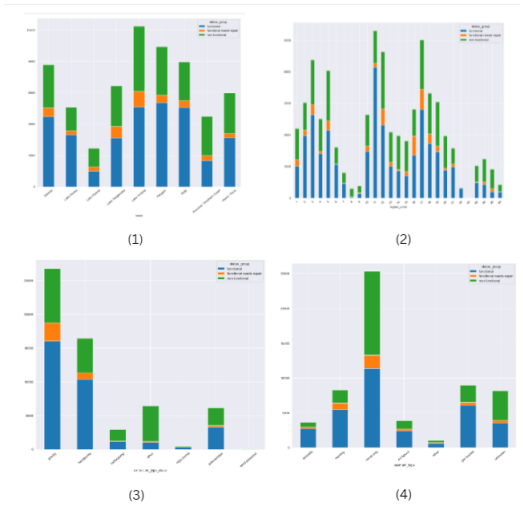

Fig. 4. Plot of different variables with status groups(a)

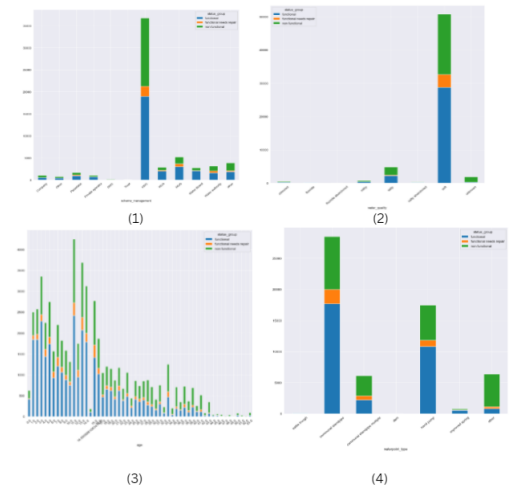

Fig. 5. Plot of different variables with status groups(b)

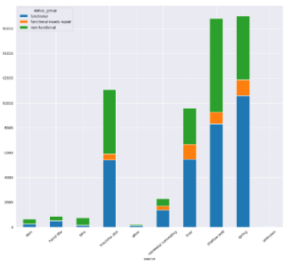

Fig. 6. Plot of different variables with status groups(c)

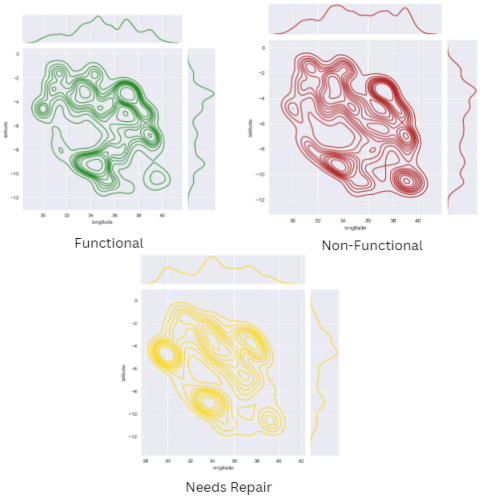

Fig. 7. Geographical distribution of pumps according to the status

Incorporating these findings into the data analysis section enhances the understanding of the factors affecting pump functionality and can inform strategies for improving maintenance and repair efforts.

The geographical distribution of the pumps is also analyzed as shown in Fig.7. Joint plots are created to visualize the density of functional, non-functional, and functional-needs-repair pumps based on their latitude and longitude coordinates. This analysis helps identify any spatial patterns or clusters associated with specific status groups. By analysing the geographical distribution of the pumps we can deduce that each type of pump is distributed in a varying manner across the region. Furthermore, a correlation matrix is computed to assess the relationships between selected variables (Fig. 8). The matrix evaluates both the magnitude and the direction of the linear connection between two pairs of variables. Insights into the relationships between data like GPS height, population log, longitude, latitude, and age may be obtained by looking at the correlation coefficients. This knowledge aids in spotting probable connections and links between the variables. Overall, the data analysis section provides a detailed exploration of the relationships, patterns, and dependencies within the dataset.These revelations help to clarify the variables influencing the operation of water pumps and can direct

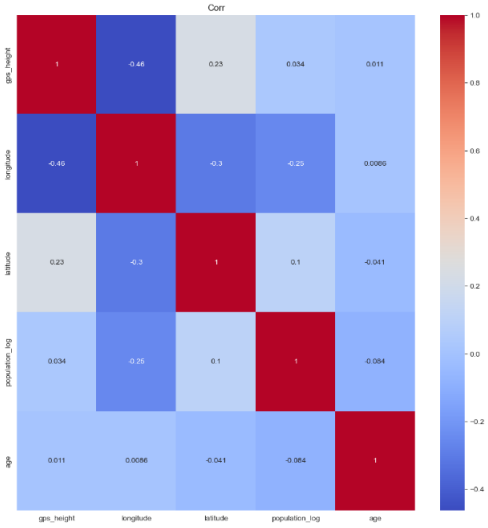

Fig. 8. Heat map of the correlation matrix

future modelling and analytical work to enhance water pump performance and maintenance.

B. Pre-processing

The preprocessing stage involved applying various techniques to handle missing values, outliers, and errors in the data. Initially, the data was merged into a single dataframe, combining the information from three different files. The following steps were performed on each feature:

- Subvillage: Existing null values were imputed with ”other,” and remaining error values (e.g., data with single letters) were also replaced with ”other.”
- Amount tsh: Since the data contained numerous zeros, the values were imputed with the log value. However, considering the large number of unusable values, the column was ultimately dropped.
- Date recorded: The column was split into year and month, which were later used for calculating the age of the water points.
- Funder/Installer: Similar to the approach taken for subvillage, null values were replaced with ”other.” Clear errors were changed to ”other” as well (e.g., 1, A, M). The top 20 funder and installer categories were considered for modeling, and category entries were replaced as required.
- Longitude/Latitude: There were clear errors in the entries, such as entries with values of 0, were replaced with the mean value.
- GPS height: The same process as for longitude/latitude was applied.
- Region/Region code: These columns were redundant, so the ”region” column was dropped. The district code column was also dropped for the same reason.
- Population: Entries with population as 0 were imputed with the mean population. However, the data still exhibited skewness after analysis. To address this, the loga-

rithm of the population was taken as the final population feature.

- Values like ”pub meeting” and ”permit” were factorized.
- Scheme management/Scheme name: The ”scheme name” column was dropped due to redundancy.
- Construction year: Zeros in the construction year were imputed with the median, as the year cannot be null.
- Extraction type: Null values and errors were replaced with the value ”other.”
- Payment type and payment: These columns were redundant, so the ”payment” column was dropped.
- Water quality/Quantity group/Quantity/Quality group: Quantity group and Quality group, these groups were dropped due to redundancy.

Feature Engineering: Two new features were introduced during the preprocessing stage:

- Age: The approximate age of the water point was calculated using the recorded year and the construction year.
- Season: It was possible to determine the season during which the water points were recorded based on the pertinent months. These brand-new fields were created to supplement the dataset’s metadata and to capture any emerging patterns or insights.

C. Classification

In this paper, we employed eight different machine learning algorithms to classify and predict the accuracy of our model. The algorithms used were as follows: Baseline XGBoost, Baseline CatBoost, Baseline GradientBoost, Baseline TabNet, Random Forest with hyperparameter tuning, XGBoost with hyperparameter tuning, CatBoost with hyperparameter tuning, and Stack Classifier (using Random Forest and CatBoost as base classifiers).

First, we split the data into training and testing sets using a 70:30 ratio. Then, we trained each algorithm on the training set and evaluated their performance on the testing set. Here are the accuracy score for each algorithm:

- Random Forest: Achieved an accuracy of 79.67.
- XGBoost: Achieved an accuracy of 79.57.
- CatBoost: Achieved an accuracy of 79.63.
- GradientBoost: Achieved an accuracy of 74.69.
- TabNet: Achieved an accuracy of 75.12.

Next, we performed hyperparameter tuning for the algorithms to further improve their performance. Using grid search and random search methods, we explored different combinations of hyperparameters for each algorithm. Here are the results after hyperparameter tuning:

- Random Forest with hyperparameter tuning: Achieved an accuracy of 79.73.
- XGBoost with hyperparameter tuning: Achieved an accuracy of 79.40.
- CatBoost with hyperparameter tuning: Achieved an accuracy of 79.21.

Among the tuned models, Random Forest and CatBoost yielded the highest accuracies. Therefore, we proceeded to create a Stack Classifier using Random Forest and CatBoost as base classifiers. The Stack Classifier combines the predictions of multiple models to make the final prediction. With the Stack Classifier, we achieved an accuracy of 80.68, which is considered the best accuracy among all the models evaluated.

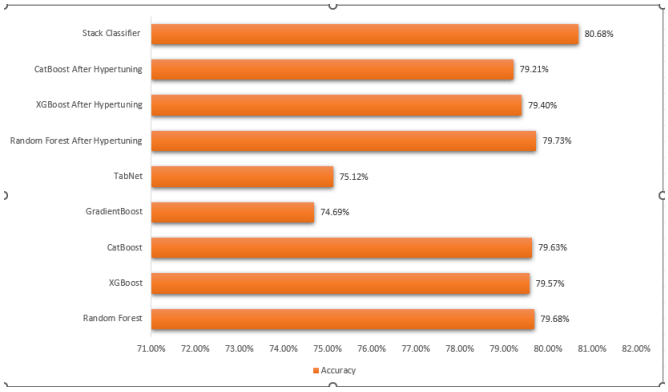

Fig. 9. Accuracy Comparison of the classifiers

Overall, the Stack Classifier using Random Forest and CatBoost as base classifiers provided the highest accuracy in predicting the water point status. This demonstrates the effectiveness of combining different models to improve classification performance.

Discussion

In this paper, we analyzed the work of two authors (Author 1(Aparna krishnakumar) and Author 2(Neha Unni)) who approached the classification problem using different methods and strategies. We will compare and critique their work based on the provided information, considering the earlier classification results as well.

Author 1 took a comprehensive approach to data analysis and cleaning. They joined the train and test data before cleaning, but the label data was not merged with this combined data. On the other hand, Author 2 joined the train and test data along with the labeled data before the cleaning process, which was was the chosen method for final model.

Regarding variable selection, Author 1 analyzed and cleaned most of the variables individually. This approach can provide a thorough understanding of the data. In contrast, Author 2 focused on variables with null data, which is a more targeted and efficient strategy. When it comes to handling null values, both authors used different approaches. Author 1 imputed null values with the value ”other,” and for numerical values imputing was done using mean values. In contrast, Author 2 used mean imputation or a general variable to fill null values, and additionally, addressed spelling mistakes and general errors appropriately.In The final model the both the techniques were equally adopted.

Author 1 employed multilevel imputation for variables like longitude and latitude based on grouping with other variables. This approach can capture more hidden patterns in the data. This approaches seemed appropriate and can improve the accuracy of the models.

Regarding feature engineering, both authors took different steps. Author 1 dropped the "amt" feature after taking the log due to a high number of zeros. This decision can be justified as the log transformation can mitigate the impact of extreme values and improve model performance. Additionally, both Author 1 and Author2 imputed population using its log, which can handle skewed distributions and provide better representation of the data.

Both authors conducted basic exploratory data analysis (EDA) for variables with "statusgroup." EDA is essential for understanding the relationships between variables and identifying patterns that can aid in model building. Therefore, this step was crucial for both authors.In terms of variable selection, Author 1 dropped irrelevant variables after thorough processing. This is a reasonable approach to reduce model complexity and focus on relevant features. Author 2 also dropped irrelevant columns, but this was done at early data processing stage.

In terms of categorical variables, both authors used different techniques. Author 1 employed factorization and one-hot encoding, which are common approaches for handling categorical data. Author 2 used mapping and one-hot encoding, which is also a valid approach. The choice of encoding method was based on personal preferences and the requirements of the models.

For classification, both authors implemented different classifiers. Author 1 implemented XGBoost (and XGBoost with hyper parameter tuning), CatBoost (and CatBoost with hyper parameter tuning),TabNet and stack classifier. Similarly, Author 2 employed Random Forest, Random Forest with hyperparameter tuning, and Gradient Boost. All the classifiers were used in the final model. Accuracy scores of each of the model was compared with one another which resulted in choosing the stack classifier as the best model with an accuracy score of 80.68.

On comparison of the work done by the authors of the paper listed in the literature review and this work,the main difference is the approach adopted in each of the stages of the analysis.All the works had one thing in common, that was to find the most appropriate model with the highest predicted accuracy. More than one of the literatures specified TabNet as the best model with the highest accuracy, but we can see that the TabNet model did not perform well for the data we prepared. This shows that there is a significant variation in the techniques we used before the modelling stage.

V. Conclusion

Based on the analysis of different machine learning models we decided to go with the stack classifier, having the highest accuracy of 80.68 to predict the water pump repair status in the region of Tanzania. From the selected model prediction we can also observe that the geographical feature along with the age of the pump play a bigger role in determining the status of the water pumps. This work could be used to emphasize the point that data mining and machine learning methods could be used to effectively help in water resource management. This could help the people in prioritizing the maintenance and repair on water points in several affected areas.

References

[1] A. Pham, B. Backus, and L. Zhu, "Pump it up: Mining the water table," 2018.

[2] K. Pathak, "Pump It Up: Predict Water Pump Status using Attentive Tabular Learning." Available: https://arxiv.org/ftp/arxiv/papers/2304/2304.03969.pdf

[3] D. Johnson and S. Smith, "A Comparative Analysis of Predictive Models for Water Pump Functionality," 2023.

[4] "Data Mining the Water Pumps: Determining the functionality of Water Pumps in Tanzania using SAS® Enterprise Miner." Available: http://www.scsug.org/wp-content/uploads/2016/11/SS-Determining-the-functionality-of-Water-Pumps-in-Tanzania.pdf

[5] A. Engel, "Categorical Variables for Machine Learning Algorithms," Medium, Mar. 16, 2022. https://towardsdatascience.com/categorical-variables-for-machine-learning-algorithms-d2768d587ab6

[6] V. Borisov, T. Leemann, K. Sessler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep Neural Networks and Tabular Data: A Survey," IEEE Transactions on Neural Networks and Learning Systems, pp. 1–21, 2022, doi: https://doi.org/10.1109/tnnls.2022.3229161.

[7] P. Probst, A.-L. Boulesteix, and B. Bischl, "Tunability: Importance of Hyperparameters of Machine Learning Algorithms," Journal of Machine Learning Research, vol. 20, pp. 1–32, 2019, Available: https://jmlr.org/papers/volume20/18-444/18-444.pdf

[8] S. Džeroski and B. Ženko, "Is Combining Classifiers with Stacking Better than Selecting the Best One?," Machine Learning, vol. 54, no. 3, pp. 255–273, Mar. 2004, doi: https://doi.org/10.1023/b:mach.0000015881.36452.6e.