COMP4030: Predicting Water Pump Functionality in Tanzania using Machine Learning Techniques

Nicholas Fruin
Data Science
University of Nottingham
Nottingham, United Kingdom
pmynjf@nottingham.ac.uk

Timothy Murphy
Computer Science
University of Nottingham
Nottingham, United Kingdom
psxtm6@nottingham.ac.uk

Digvijay Kasana
Data Science
University of Nottingham
Nottingham, United Kingdom
psxdk9@nottingham.ac.uk

***Abstract*—Access to clean water is a significant issue in Tanzania, with many water pumps across the country being non-functional or in need of repair. This study aims to predict the functionality of water pumps in Tanzania using machine learning algorithms, based on a comprehensive dataset that contains many different features. We used a range of machine learning techniques, including K-Nearest Neighbours (KNN), Random Forest, XGBoost, CatBoost, and TabNet, to develop predictive models and compare their performance using precision, recall, F1 score, and accuracy metrics. The results showed that gradient boosting algorithms, XGBoost and CatBoost in particular, consistently outperformed other models, with Random Forest also showing strong performance. The highest accuracy achieved was 0.79, suggesting that there is still room for improvement. This study demonstrates the potential of machine learning in predicting water pump functionality, which could help inform maintenance and repair efforts in Tanzania, and would ultimately improve access to clean water. Future research should focus on improving and refining the models, as well as incorporating domain knowledge from experts in the field, and further exploring the use of ensemble learning methods to improve predictive performance.**

***Index Terms*—water pump functionality, machine learning, Tanzania, XGBoost, Random Forest**

I. Introduction

Tanzania is the largest country in East Africa and it faces significant challenges in providing access to clean water for its population of over 69 million people [1]. Many areas in Tanzania lack access to functional water pumps and, of these pumps, a large amount are non-functional or in need of repair. The Tanzanian Ministry of Water, in collaboration with Taarifa, has collected and created a comprehensive dataset containing information about water pumps across the country, and have made it publicly available [2]. This dataset provides us with the opportunity to assess various machine learning techniques and approaches to predict the functionality status of water pumps and to identify the factors that influence this the most.

The dataset contains information on 59,400 water pumps and includes 41 features. These features can be roughly categorised into geographical information, water pump characteristics, and management details. The geographical information includes features such as latitude, longitude, elevation, region, and district. This information helps to understand the geographical distribution of water pumps across the country and to identify any regional patterns.

Water pump characteristics include features like the amount of available water, extraction type, water quantity, and water quality. These features show the technical aspects of the water pumps and how they might impact their functionality. For example, pumps with good, soft water might be more functional than pumps with salty water, or pumps with hand-pump mechanisms might be more functional than pumps with motorised mechanisms.

Management details include information about the installer and funder of the pump, as well as whether they have a permit and what their management scheme is, for example. Information like this can also reveal interesting patterns, such as the relationship between certain installers and the functionality of their pumps.

The dataset also contains other relevant features such as the date of recording and population served by the pump. Features like this can be used to provide additional context to the functionality of the water pumps.

By analysing this dataset, we aim to discover any patterns and relationships that relate to water pump functionality, and to ultimately produce a solution that can accurately predict water pump functionality.

The main research questions addressed in this project are:

1) Can we accurately predict the functionality status of water pumps in Tanzania based on the available features?
2) How do different machine learning approaches compare in terms of performance?

Addressing and answering these questions can provide valuable and actionable insights for the Tanzanian government and Ministry of Water, helping to improve the efficiency of water pump maintenance. By accurately predicting the functionality of water pumps, resources can be used more efficiently too, ultimately leading to improved clean water access for the population.

The remainder of this paper is structured as follows. Section II provides a literature review of related work on predicting water pump functionality and the use of machine learning techniques for similar problems. Section III describes the methodology used in this study, including data analysis, preprocessing, and the classification models used. Section IV

presents the results of the classification models and compares their performance using several evaluation metrics. Section V discusses the implications of these findings, the limitations of the study, and potential future research directions. Finally, Section VI concludes the paper by summarising the main contributions and highlights the potential impact of this research on improving clean water access in Tanzania.

II. Literature Review

The prediction of functionality and condition of equipment, including water pumps, has been thoroughly investigated in recent years. Researchers have investigated using various machine learning techniques to tackle this problem, such as decision-tree methods like Random Forest [3] and Gradient Boosting Decision Trees (GBDTs), as well as deep learning approaches like TabNet [4].

One of the most popular approaches for handling tabular data, such as the Tanzanian water pump dataset, is the use of tree-based ensemble methods. Random Forest is a popular ensemble learning method and has been widely used for both regression and classification tasks due to its accuracy and robustness in handling mixed data types, outliers, and noise [3]. Xu et al. [5] and Xiong et al. [6] used Random Forest to predict water quality and develop a Water Quality Index for groundwater, respectively. They demonstrated how effective this approach is for prediction tasks. For the Tanzanian water pump dataset specifically, Pathak and Shalini [7] found that Random Forest achieved an accuracy of 77.35% and an F1 score of 0.621, although they noted that this was not as effective as using GBDT approaches or deep neural networks.

GBDTs have also proven to be effective for modelling tabular data, and implementations like XGBoost [8] and CatBoost [9] have shown to improve accuracy even further. XGBoost is an efficient and scalable implementation of gradient boosting. A recent study has shown how it can be used for accurate predictive maintenance in geothermal power plants [10], and another study has shown it to be useful in predicting conditions of compressor and water pump systems in ammonia production [11]. These studies show the potential of using XGBoost in handling large scale and sparse data structures to provide robust and accurate predictions.

CatBoost is another GBDT algorithm, which has gained attention for its ability to handle categorical features directly, reducing the need for extensive data pre-processing. Xiao, Wang, Liu, Gao, and Wu [12] developed a predictive model using CatBoost to predict faults in pavements, which showed better performance than other approaches such as Random Forest. Considering the high amount of categorical variables in the Tanzanian water pump dataset, CatBoost might be an effective approach to take.

In recent years, deep learning approaches have emerged as an alternative to tree-based methods for tabular data modelling. TabNet, a deep neural network designed specifically for tabular data [4], combines the benefits of tree-based algorithms, such as feature selection and interpretability, and deep neural networks, such as end-to-end-learning. Pathak and Shalini [7] applied TabNet to the Tanzanian water pump dataset and experimented with different loss functions to handle class imbalance. They achieved an accuracy of 83.6% using focal loss, and an F1 score of 0.697, outperforming every other approach.

Despite the growing interest in using deep learning approaches, there is an ongoing debate on whether neural networks actually outperform GBDTs on tabular data. McElfresh et al. [13] compared 19 algorithms across 176 datasets and found that the performance difference is often negligible, and that hyperparameter tuning is more important than choosing an algorithm in most cases. Shwartz-Ziv and Armon [14] argue that deep neural networks are not always the best choice for tabular data, and that XGBoost can outperform deep learning models while also being easier to optimise.

III. Methodology

A. Data Analysis

Before pre-processing the data, it is necessary to perform exploratory data analysis (EDA) to uncover insights and potential issues with the data, such as outliers or missing data. This allows us to understand the dataset better and make informed decisions for the modelling process.

To start, we began familiarising ourselves with the dataset by checking for duplicate records, examining the shape of the data, and identifying the types of variables, which included numerical and categorical. We found that the dataset contained 41 features, consisting of 6 numerical features and 35 categorical features. There were no duplicated records.

Descriptive statistics of the numerical features revealed that there are no missing values, but there are a lot of zero values. For example, there are a lot of zero values in the *longitude* feature, but it is impossible for any area in Tanzania to have a longitude of zero. This suggests that missing values have been imputed with zero, and this would require further investigation. Some features, such as *amount_tsh* and *num_private* consisted mostly of zero values, suggesting that they either should be discarded or investigated further. Additionally, there appeared to be outliers in the features *amount_tsh*, *gps_height*, and *population*, as they had unusually high maximum values compared to the third quartile values.

It also became apparent that some features should be treated as categorical variables instead of numerical, such as *id, region_code, district_code*, and *num_private*.

We found that only categorical variables contained missing values - out of the 35 categorical features only eight contained missing values. The feature for scheme name contained over 28,000 missing values, which is more than half, whereas the feature for waterpoint name only had two missing values. It became clear that a different approach for handling missing values would be needed for each feature.

Further analysis of the categorical features found that funder, waterpoint name, subvillage, and scheme name features had high cardinality, suggesting that they might not be suitable for the model without modification. Furthermore, the feature

recorded by only contained one unique value, meaning that the feature can be discarded as it contains no useful information.

Initial countplots for each categorical feature revealed that certain features appeared to be very similar to each other. For example, *quantity* and *quantity group* were the exact same. There were three features for waterpoint type and they also seemed to contain the same information, except they differed in terms of granularity, where one feature contained the most values, and the others combined less frequent values to decrease granularity. The same was the case for extraction type, water quality, water quantity, and management features. For the model, we only need one feature from each, so we would need to decide which ones to keep.

Further investigation of the *date_recorded* feature revealed that the dataset spans from 2002 to 2013, and the majority of records are from 2011, 2012, and 2013. The *installer* and *funder* features, despite their high cardinality, might be important for the model as certain funders and installers had more non-functioning water pumps than others.

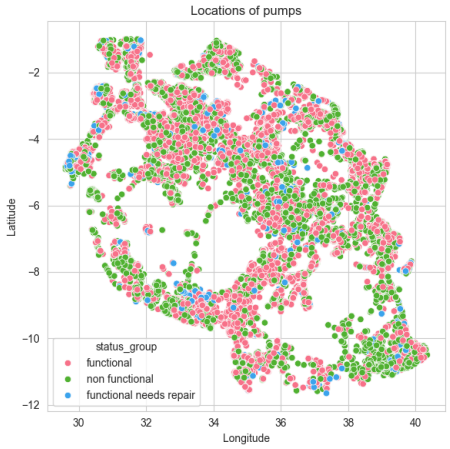

Fig. 1. Scatter plot to show the location and functionality of water pumps.

Analysis of the *longitude, latitude*, and *gps_height* features confirmed that zero values and negative values are indeed correct and should be treated as missing values. Furthermore, plotting the longitude and latitude values showed no clear relationship between pump functionality and location initially (see Figure 1).

The *construction_year* feature showed that older pumps are more likely to be non-functional (see Figure 2), making it a potentially useful feature, but missing values would need to be dealt with appropriately. This also suggests that creating a new feature for age might be better for model performance than using *construction_year*.

Further inspection of *gps_height* revealed there is very little difference between the pump functionality statuses. Figure 3 shows that each status followed a very similar trend. Therefore, it is likely that this feature should be dropped.

Examining the *public_meeting* and *permit* features revealed that pumps with a public meeting have a higher chance of being functional, whereas having a permit does not seem

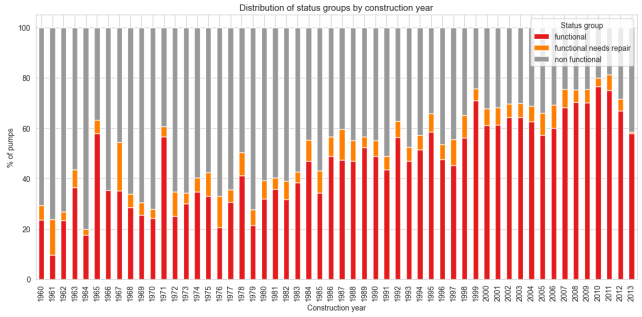

Fig. 2. Stacked bar plot to show distribution of pump functionality by construction year.

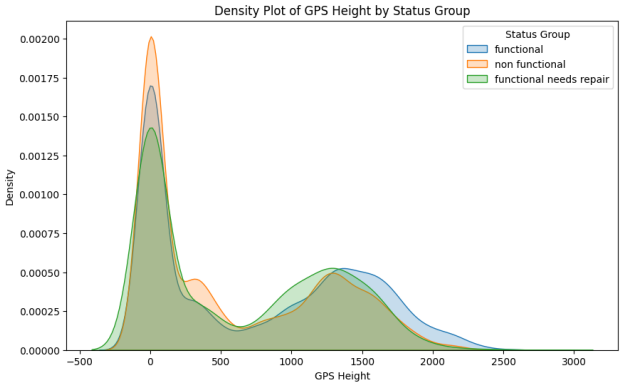

Fig. 3. Density plot of gps_height by water pump functionality status.

to have a significant impact on pump functionality. Certain extration type classes, such as *motorpump* and *other*, have a higher proportion of non-functional pumps compared to functional pumps. Pumps with good water quality have the highest percentage of functional pumps, while pumps with salty water have more non-functional pumps. Pumps with *dry* quantity have a very high percentage of non-functional pumps.

Inspection of *region* and *population* found that some regions only had zero values, indicating that there was nobody living in the entire region. This will need to be dealt with in the pre-processing stage.

These findings from EDA have provided important insights into the dataset, including potential issues and important features. They will inform pre-processing, including feature engineering, and feature selection decisions in the subsequent stages.

B. Pre-processing

We began pre-processing by dropping redundant columns from the dataset. The following 22 columns were dropped after results from data analysis found that they are likely to not be useful to the model: *id, date_recorded, longitude, latitude, gps_height, population, scheme_name, recorded_by, waterpoint_type_group, source, source_class, quantity_group, water_quality, payment, extraction_type, extraction_type_class, management, management_group, ward,*

subvillage, wpt_name, and *num_private*. This leaves 19 features for the model to be trained on.

Missing values for every other categorical feature, except for *permit* since a value of zero indicates false, we imputed missing values with "unknown".

We decided to drop *longitude* and *latitude* due to the high cardinality of each feature. We kept other geographical features such as *region* and *lga*, which had much lower cardinality and therefore more suitable for modelling.

Initially, we planned to create a new feature *age*, which would be calculated by subtracting *construction_year* from the current year. However, since data analysis found there to be over 20,000 zero values, we decided to modify *construction_year* instead by grouping the values into bins based on the decade. For zero values, we imputed these with "unknown" and this became one of the bins.

Since the feature *amount_tsh* mostly consisted of zero values, we decided to modify it to become a boolean variable, where true represents zero and false represents non-zero values. This modification should make the feature much more suitable for modelling, as the cardinality has been reduced and the issue of the many zero values has been dealt with.

For the features *installer* and *funder* we decided to keep the top 100 values and then group the remaining values into "other". This reduced the cardinality drastically for both, while also keeping the value count for "other" in an acceptable range.

We decided to impute the zero values in *population* using the median values for each region. Some regions did not have any data for population, indicating that those regions had no one living there. This is incorrect, so we calculated the overall median population for Tanzania, and also the median population per region. Regions with all zero values had values imputed using the overall median, and regions with a few missing values were imputed with the per region median. Median was chosen over mean as population data are typically skewed. Furthermore, we grouped values into 20 bins to reduce cardinality for modelling.

Following data analysis on *gps_height* (see Figure 3), we decided to drop this feature.

C. Classification

After pre-processing the data, we selected the following models to train: Random Forest, K-Nearest Neighbours (KNN), XGBoost, CatBoost, and TabNet.

KNN is a non-parametric algorithm that classifies new data points based on their proximity to training data points [15]. It is simple to implement and it also can handle multi-class classification problems.

Random Forest is an ensemble learning method that constructs multiple decision trees and then combines their predictions to make the final classification [3]. It is robust and is able to handle data with high dimensionality.

XGBoost and CatBoost are both gradient boosting algorithms. XGBoost is an optimised distributed gradient boosting library that is designed to be highly efficient, flexible, and portable [8]. CatBoost, on the other hand, is known for its ability to handle categorical features automatically [9]. Both algorithms have shown excellent performance in various different machine learning competitions and real-world applications.

TabNet is a deep learning architecture that is designed for tabular data [4]. It uses sequential attention to choose which features to reason from at each decision step, which allows for iterpretability and better learning since it can learn to focus on the most relevant features for the task at hand.

We chose these models to compare the performance of traditional machine learning algorithms (KNN and Random Forest) with the state-of-the-art gradient boosting algorithms (XGBoost and CatBoost), and a deep learning approach designed specifically for tabular data (TabNet). This allows us to assess which type of model is most suitable for predicting water pump functionality based on the Tanzania dataset.

To evaluate the performance of each model, we used standard classification metrics such as precision, recall, F1 score, and accuracy. These metrics give us a comprehensive view of each model's performance, considering both the model's ability to correctly identify each class (precision and recall) and how correct it is overall (accuracy).

By comparing the results of these models, we aim to identify the most effective approach for predicting water pump functionality in Tanzania based on the available data. The results of this comparison will be presented and discussed in the following sections.

IV. Results

The performance of each model was evaluated using precision, recall, F1 score, and accuracy. Table I presents the results of the classification models.

TABLE I
CLASSIFICATION RESULTS

Model	Precision	Recall	F1-score	Accuracy
KNN	0.66	0.60	0.62	0.74
Random Forest	0.68	0.63	0.65	0.79
XGBoost	0.74	0.63	0.65	0.79
CatBoost	0.74	0.62	0.65	0.78
TabNet	0.74	0.57	0.59	0.76

The results show that the gradient boosting algorithms, XGBoost and CatBoost, achieved the highest precision score of 0.74, suggesting that these models had the highest proportion of correctly predicted positive instances across all instances predicted as positive. However, their recall scores were slightly lower than Random Forest, suggesting that Random Forest was able to correctly identify a higher proportion of actual positive instances.

In terms of F1 score, which is the harmonic mean of precision and recall, XGBoost, CatBoost, and Random Forest all achieved a score of 0.65, suggesting that there is a balance between precision and recall. TabNet, despite being designed specifically for tabular data, had the lowest F1 score of 0.59, suggesting that it may not be the best choice for this dataset in particular.

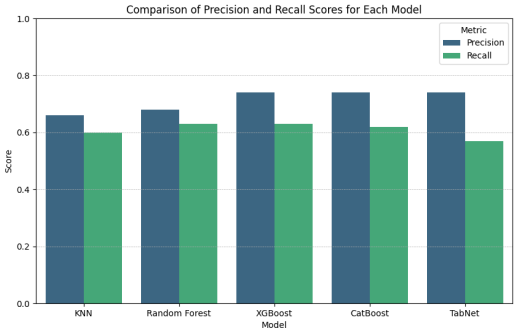

Fig. 4. Comparison of precision and recall scores for each model.

The accuracy scores show that Random Forest and XGBoost achieved the highest overall accuracy of 0.79, followed closely by CatBoost with 0.78. KNN and TabNet had lower accuracy scores of 0.74 and 0.76, respectively.

Figure 4 shows a comparison of the precision and recall scores for each model. It is evident that XGBoost and CatBoost have the highest precision, while Random Forest has a slightly higher recall. KNN and TabNet have lower scores for both metrics.

Overall, the results suggest that gradient boosting algorithms, particularly XGBoost and CatBoost, are the most effective for predicting water pump functionality in Tanzania based on the available data. Random Forest also performed well, and had a slightly higher recall score than the gradient boosting algorithms. Although KNN and TabNet still provided reasonable results, they did not perform as well as the other models in this study.

V. Discussion

The results of this study show how effective different machine learning algorithms are in predicting the functionality of water pumps in Tanzania, based on various geographical, technical, and management features. The gradient boosting algorithms, XGBoost and CatBoost, consistently performed well across all the evaluation metrics, and Random Forest also showed strong performance, especially in terms of recall.

The strong performance of XGBoost and CatBoost is likely to be due to their ability to handle complex, non-linear relationships between features and the target variable. They are known for how robust they are, as well as their ability to handle data with high dimensionality [8], [9]. Additionally, CatBoost automatically handles categorical features, which likely helped with its strong performance, given the high number of categorical variables in the dataset.

Random Forest's slightly higher recall score suggests that it may be better at identifying non-functional pumps, which could actually be helpful in a real-world scenario where identifying and repairing non-functional pumps is a priority. However, its low precision score indicates that it may also have a higher rate of false positives compared to the gradient boosting algorithms.

The lower performance of KNN and TabNet compared to the other models could be due to several factors. KNN's performance is heavily dependent on the choice of K and the distance metric used [15]. It is possible that further tuning of the hyperparameters could improve its performance. Despite TabNet being designed for tabular data, its worse performance suggests that its attention-based architecture may not be the best fit for this dataset in particular.

It is important to note that while the gradient boosting algorithms and Random Forest performed well, there is still room for improvement. The highest accuracy achieved was 0.79, which suggests that the models still misclassify a significant portion of the pumps. This is likely to be mostly due to limitations in the data, such as missing or inaccurate records, as well as the presence of any complex relationships that are difficult for the models to capture.

Future research could look into several avenues to build upon the findings of this study. Firstly, investigating the performance of other machine learning algorithms, such as support vector machines or deep learning architectures like convolutional neural networks (CNNs) could provide more insights into the most effective approaches for this task. Secondly, using time series data, such as historical records of pump failures and maintenance, could allow for the development of models that predict pump functionality over time, allowing for proactive maintenance scheduling. Thirdly. the models could be further refined and improved by conducting field studies to validate the models' predictions and to gather additional data on the factors that influence pump functionality. Furthermore, future research could look into the generalisability of these findings to other countries or regions that face similar water issues. This could potentially lead to the development of a more widely applicable framework for predicting water pump functionality. Finally, future researchers could collaborate with domain experts and policymakers in this area to translate these findings into insights that are actionable and decision-making tools that could help to maximise the impact of this research on improving clean water access in Tanzania and beyond.

VI. Conclusion

This study demonstrates the potential of machine learning algorithms, particularly gradient boosting algorithms like XGBoost and CatBoost, as well as Random Forest, KNN, and TabNet, for predicting water pump functionality in Tanzania. We used the Pump it Up: Data Mining the Water Table dataset for Tanzanian water pumps, which contained a wide variety of features, to develop models that achieved an accuracy of up to 0.79 in predicting the functional status of water pumps. These findings show how important it is to use data analysis and modelling to address challenges with infrastructure in developing countries.

The insights gained from this research could have significant implications for improving the clean water access in Tanzania. Now that the Tanzanian government and the Ministry of Water can accurately predict which pumps are likely to be non-functional, they can allocate resources for maintenance and

repair more efficiently. This approach could lead to more functional water pumps, which would ultimately improve the lives of millions of people who rely on them for their daily water.

Moreover, the methodology and findings of this study could potentially be extended to other countries that face similar water issues. The development of a more widely applicable model or framework for predicting water pump functionality could help to improve infrastructure investments and maintenance in other regions and even in other countries.

However, it is important to note the limitations of this study and the need for further research. While our models achieved promising results, there is still room for improvement in terms of accuracy and generalisability. Future research should focus on refining the models by incorporating more data sources, such as time series data on pump failures and maintenance, as well as exploring the use of other machine learning algorithms and ensemble methods. Furthermore, it would be beneficial to collaborate with people that have expert domain knowledge for this area, including policymakers, so that the models' predictions can be validated and their findings translated into actionable insights. This will be very important in order to maximise the impact of this research.

In conclusion, this study demonstrates the potential of machine learning in addressing infrastructure challenges and improving clean water access in Tanzania. We can develop more effective and efficient strategies for countries like this to maintain and repair water pumps by using this approach. Ultimately, this will positively contribute to the wellbeing and development of communities across the country and, hopefully in the future, in other countries too. As we continue to improve and expand upon approaches like this, we move closer to the goal of universal access to clean water.

REFERENCES

[1] worldometer, "Tanzania population (2024) - worldometers," Worldometers.info, 2024. [Online]. Available: https://www.worldometers.info/world-population/tanzania-population/

[2] DrivenData, "Pump it up: Data mining the water table," DrivenData. [Online]. Available: https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/23/

[3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[4] S. O. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," *arXiv (Cornell University)*, 08 2019.

[5] J. Xu, Z. Xu, J. Kuang, C. Lin, L. Xiao, X. Huang, and Y. Zhang, "An alternative to laboratory testing: Random forest-based water quality prediction framework for inland and nearshore water bodies," *Water*, vol. 13, p. 3262, 11 2021. [Online]. Available: https://www.mdpi.com/2073-4441/13/22/3262/pdf

[6] Y. Xiong, T. Zhang, X. Sun, W. Yuan, M. Gao, J. Wu, and H. Zhao, "Groundwater quality assessment based on the random forest water quality index—taking karamay city as an example," *Sustainability*, vol. 15, pp. 14 477–14 477, 10 2023.

[7] K. Pathak and L. Shalini, "Pump it up: Predict water pump status using attentive tabular learning," *arXiv (Cornell University)*, 01 2023.

[8] T. Chen and C. Guestrin, "Xgboost: a scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, p. 785–794, 2016. [Online]. Available: https://dl.acm.org/doi/10.1145/2939672.2939785

[9] L. O. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," *arXiv (Cornell University)*, 06 2017.

[10] R. Nanavaty, "Exploring autoencoders and xgboost for predictive maintenance in geothermal power plants," *49th Workshop of Geothermal Reservoir Engineering, Stanford University*, 2024.

[11] K. Salim, R. S. A. Hebri, and S. Besma, "Classification predictive maintenance using xgboost with genetic algorithm," *Revue d'Intelligence Artificielle*, vol. 36, pp. 833–845, 12 2022.

[12] W. Xiao, C. Wang, J. Liu, M. Gao, and J. Wu, "Optimizing faulting prediction for rigid pavements using a hybrid shap-tpe-catboost model," *Applied sciences*, vol. 13, pp. 12 862–12 862, 11 2023.

[13] D. McElfresh, S. Khandagale, J. Valverde, V. P. C, B. Feuer, C. Hegde, G. Ramakrishnan, M. Goldblum, and C. White, "When do neural nets outperform boosted trees on tabular data?" *arXiv.org*, 10 2023. [Online]. Available: https://arxiv.org/abs/2305.02997

[14] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 05 2022.

[15] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, pp. 175–185, 08 1992.