



## Optimizing Hydropower: Comparative Analysis of Machine Learning Models for Energy Production Forecasting



Research Report submitted in partial fulfillment of the requirements for the **Postgraduate Diploma in Data Analytics** at The Independent Institute of Education, IIE MSA.

---

***Andiswa Nyongwana***

St10314980

*Supervisor: Fezile Matsebula*



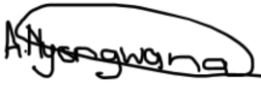
## **Abstract**

This study investigates predictive modeling for hydropower generation, focusing on comparing the performance of Random Forest model, KNN model and an ensemble model. The background stems from the need to optimize energy production in hydropower plants by understanding environmental influences. The primary argument is that robust predictive models can aid operational decision-making by highlighting key factors impacting output. The research examined how well these models captured relationships between environmental variables and hydropower output, specifically analyzing the Hydro Water Generation as output. Data was collected from historical records of reservoir levels, climatic data and hydropower generation metrics. The Random Forest model was chosen for its ability to handle complex, non-linear data and the KNN model for its simplicity and baseline comparison. The findings indicated that Random Forest outperformed KNN, identifying "Woodstock dam" as the most influential factor, suggesting that reservoir levels are critical for accurate predictions.

**Key words:** hydropower, predictive modeling, Random Forest, ensemble model KNN, reservoir levels, environmental factors, machine learning, energy optimization.

## DECLARATION

I hereby declare that the **research report** submitted for the **Postgraduate Diploma in Data Analytics** to The Independent Institute of Education (The IIE) is my own work and has not previously been submitted to another University or Higher Education Institution for a postgraduate qualification.

Signature 

Date 4/11/2024

## Contents

Abstract .....	1
1. Introduction.....	5
2. Rationale.....	7
3. Problem Statement .....	8
4. Research Questions .....	10
5. Research Objectives and Hypotheses.....	10
6. Literature Review .....	11
6.1. Introduction.....	11
6.2. Overview of hydropower production .....	12
6.3. Application of machine learning in hydropower production.....	13
6.4. What are Random Forest and KNN models.....	14
6.5. Performance of machine learning models .....	15
6.6. Evaluation of models .....	16
6.7. Research Gaps .....	17
6.8. Data variables.....	18
6.9. Data variables commonly used for predicting hydropower .....	19
6.10. Key concepts .....	21
7. Research Methodology .....	23
7.1. Research design.....	23
7.2. Population .....	24
7.3. Sampling .....	24
7.4. Data Collection.....	24
7.5. Data analysis method .....	25
8.3. Overview of the KNN, Random Forest and Ensemble model results .....	31
8.4. Predicted vs the actual values .....	33
9. Discussion .....	36
10. Conclusion .....	41
11. References.....	43

## List of Figures

Figure 1: Summary of environmental conditions .....	27
Figure 2: Summary statistics for Driel Barrage .....	27
Figure 3: Summary statistics for Woodstock and Stekfontein .....	28
Figure 4: Summary statistics for Drakensberg hydropower station .....	29
Figure 5: correlation matrix for the data variables .....	30
Figure 6: Feature Importance in Random Forest Model.....	33
Figure 7: Scenario one: KNN .....	34
Figure 8: Scenario 2: Random Forest .....	35
Figure 9: Ensembled model .....	36

## List of Tables

Table 1: Data variables that affect hydropower production.....	19
Table 2: No. of null values in the dataset.....	26
Table 3: Performance matrices.....	38

## **1. Introduction**

Greenhouse gases (GHGs) are the primary cause of climate change, largely resulting from anthropogenic activities. Research of the impacts of GHGs as the driver of climate change started in the 1800s (History, 2023). The research showed that carbon dioxide is the main cause of climate change and later on it was revealed that other GHGs are Methane, Nitrous oxide, Hydrofluorocarbons, Perfluorocarbons, Sulphur hexafluoride and Nitrogen trifluoride (History, 2023). The biggest contributor to GHG emissions is the energy sector, which contributes about 30%, followed by transport which contributes 14% and relies on energy (Hannah et al., 2020). This creates a need to invest in renewable energy as an alternative to provide clean energy and contribute to climate change mitigation.

Renewable energy plays a crucial role in reducing reliance on environmentally harmful non-renewable energy sources and mitigating the adverse effects of GHG emissions. On the contrary, the full potential of renewable energy has not been utilized. Renewable energy sources contributes about 40%, with hydropower being the highest contributor of a renewable energy source (13.9%) (IEA, 2023). To maximize the potential of hydropower production, there is a need for good planning and a profound understanding of energy production dynamics. To fully leverage the potential of renewable energy while minimizing environmental impact, comprehensive strategies are imperative to optimize hydropower production. If no proper hydropower production planning is done, South Africa could miss its maximum potential to optimize its mitigation efforts and will not achieve its potential hydropower production which is 14,000 GWh/year (Wilhelm, n.d). This necessitates enhancing efficiency, increasing output and ensuring sustainable management practices, facilitated by technologies such as machine learning. Addressing these challenges through technological informed planning and implementation is essential to push the transition towards a more sustainable energy future. By doing so, these actions will be contributing to national and global efforts to combat climate change and promote environmental stewardship.

One of the ways to improve planning and efficiency of hydropower stations is through the application of modern technologies such as machine learning to predict hydropower generation and the factors that affect the production. Integrating machine learning into hydropower systems is a big step forward in finding sustainable energy solutions. Leveraging the potential of machine learning can significantly enhance the efficiency, reliability and environmental compatibility of hydropower generation (Ramarope et al., 2023).

According to IBM, Machine learning, a subset of artificial intelligence, involves the development of algorithms that can learn from and make predictions or decisions based on data. This technology has shown remarkable success in various industries, from healthcare to finance, due to its ability to analyze vast amounts of data and uncover intricate patterns that are often beyond human capacity (IBM, n,d). In the context of hydropower, Machine Learning can be employed to optimize operations, predict maintenance needs, manage input resources and mitigate environmental impacts(Ekanayake et al., 2021). The application of ML in hydropower encompasses several critical areas. For instance, predictive maintenance algorithms can forecast equipment failures, thereby reducing downtime and maintenance costs. Optimization algorithms can enhance the efficiency of turbines and generators, leading to more effective energy production (Jung et al., 2021). Additionally, ML models can improve water resource management by predicting inflows and optimizing reservoir operations, ensuring a balanced approach between energy production and ecological sustainability(Ekanayake et al., 2021). As a result, in this research two machine learning algorithms were developed, ensembled and compared on how well they work to predict hydropower production.

The research explored the benefits of integrating machine learning into hydropower systems, in terms of predicting hydropower production. The overall aim is to determine the most effective machine learning algorithm for predicting hydropower production accurately. By reviewing the performance of two different algorithms, the research seek to identify the most suitable model that can be utilized for hydropower forecasting. Understanding which algorithm performs best will support and accelerate the ability to

implement efficient and reliable predictive models for hydropower production, thereby optimizing energy generation and resource allocation.

## **2. Rationale**

The rationale for this research lies in the critical need to enhance the efficiency and reliability of hydropower generation using the latest technological tools and algorithms such as machine learning, particularly at the Drakensberg hydropower station.

The Drakensberg hydropower station is the second largest hydropower producer in South Africa with an installed capacity of 1332 MW (Prinsloo and Burkhardt, 2019). However, the plant is currently running 25% below capacity (Prinsloo and Burkhardt, 2019). This underperformance underscores the imperative need for better planning and understanding of the factors affecting hydropower production, such as environmental conditions and maintenance issues. Utilizing advanced technology in this context can ensure maximum hydropower production capacity, which is crucial for mitigating climate change.

As climate change continues to impact weather patterns and water availability, accurate prediction of hydropower production becomes increasingly vital for effective energy management and sustainability (Ekanayake et al., 2021). However, existing prediction methods may lack precision and fail to capture the complex interplay between hydropower output and environmental factors such as rainfall, river flow rate and reservoir levels (Condemi et al., 2021). Furthermore, there is also limited research that has been conducted in South Africa on the application of machine learning in hydropower production planning and maintenance (Ramarope, 2023).

By leveraging machine learning algorithms, advanced prediction models tailored specifically to the unique characteristics of the Drakensberg hydropower station can be developed. These models have the potential to significantly improve the accuracy and timeliness of hydropower production forecasts, enabling operators to optimize energy generation schedules and resource allocation in response to changing environmental



conditions(Ramarope et al., 2023). Additionally, by analyzing the relationship between hydropower output and various environmental factors, insights into the underlying mechanisms driving hydropower generation can be uncovered, thus informing more effective decision-making processes and risk management strategies(Ramarope et al., 2023).

This research will also contribute to the resilience and sustainability of hydropower operations at the Drakensberg hydropower station and beyond. By enhancing the ability to predict hydropower production using machine learning, the transition towards a more reliable, efficient and environmentally friendly energy infrastructure is being supported, thus advancing the broader goals of climate mitigation and renewable energy adoption. The research will also support the achievement of Sustainable Development Goals (SDGs), such as SDG 6: Clean water and Sanitation, SDG 7: Affordable and clean Energy, SDG 9: Industry, innovation and infrastructure and SDG 13: Climate Action.

### **3. Problem Statement**

Renewable energy is important in reducing dependence on environmentally harmful non-renewable energy sources and mitigating the adverse effects of greenhouse gas emissions (Ekanayake et al., 2021). Effective production of hydropower, a significant renewable energy source, requires thorough planning and a deep understanding of energy production dynamics. This understanding may assist countries such as South Africa, who face several challenges in the development and implementation of hydropower systems to address these challenges.

South Africa heavily relies on fossil fuels, particularly coal, for electricity generation, contributing to high levels of greenhouse gas emissions and environmental pollution in the global south. Despite having potential sites for hydropower, there has been limited development of both small- and large-scale hydropower projects. The existing hydropower capacity is underutilized due to a general lack of knowledge and infrastructure for hydropower and hydrokinetic systems (Niebuhr et al., 2019).

Moreover, South Africa's status as a water-scarce country complicates hydropower development. The highly protected water infrastructure and ownership complexities of canals and other water delivery systems further hinder project implementation. The regulatory and legislative barriers, along with lengthy and complex approval processes, add to the challenges. Furthermore, misconceptions about hydropower, such as the belief that it consumes or pollutes water resources and a lack of understanding about the potential and benefits of hydropower and mechanical systems, limit their acceptance and implementation (Niebuhr et al., 2019). On the other hand, economic and technical challenges, including high import costs for hydropower turbines and low local availability, drive up capital costs. Additionally, the low velocities in most canal sections result in low efficiency for hydropower systems, making it difficult to achieve economically viable energy production without significant modifications. Maintenance and operational issues, such as debris buildup, further affect the reliability and sustainability of hydropower installations (Niebuhr et al., 2019).

To address the above-mentioned issues hydropower prediction systems can be developed to predict hydropower production in different project sites and assess the feasibility of hydropower production and how much hydropower can be produced. Furthermore, to maximize the potential of hydropower while minimizing its environmental impact, comprehensive strategies are essential. These strategies should aim to enhance efficiency, increase output and ensure sustainable management practices through advanced technologies such as machine learning (Ramarope et al., 2023). Machine learning models have demonstrated significant promise in predicting hydropower generation by analyzing meteorological variables and historical data (José et al., 2022). By leveraging machine learning, it is possible to optimize energy production, schedule maintenance effectively and manage resources efficiently (José et al., 2022). This approach not only supports the transition to a more sustainable energy future but also contributes to global efforts to combat climate change and promote environmental stewardship.

In response to these challenges, the proposed study will design and evaluate a machine learning model to predict hydropower production based on historical data and relevant environmental factors. The model will consider variables such as rainfall, river flow rate, reservoir levels and temperature to predict hydropower output. The objective is to identify and evaluate a predictive model that assists hydropower plant operators in optimizing energy production, scheduling maintenance and managing resources effectively. This research will specifically contribute to the planning and maintenance of hydropower stations in South Africa, ultimately enhancing hydropower production and planning efficiency to contribute to 62 million residents of South Africa having access to clean, renewable and reliable energy.

#### **4. Research Questions**

- a. What is the comparative performance of two different machine learning algorithms (random forest and K-nearest neighbors) in predicting hydropower production at the Drakensberg hydropower station?
- b. How does hydropower output correlate with environmental factors such as rainfall, river flow rate and reservoir levels at the Drakensberg hydropower station?
- c. What are the key factors influencing hydropower generation at the Drakensberg hydropower station and what is their relative importance in predictive models?

#### **5. Research Objectives and Hypotheses**

##### **a. Hypothesis:**

Null Hypothesis (H<sub>0</sub>): No environmental factors have a significant impact on the hydropower generation capacity.

Alternative Hypothesis (H<sub>1</sub>): At least one of the environmental factors has a significant impact on the hydropower generation capacity.

### **b. Objectives:**

1. To assess the performance of 2 different machine learning algorithms ((random forest and K-nearest neighbors) in predicting hydropower production.
2. To analyze the relationship between hydropower output and various environmental factors such as rainfall, river flow rate and reservoir levels.
3. To identify the key factors that influence hydropower generation and determine their relative importance in the predictive models.

## **6. Literature Review**

### **6.1. Introduction**

Globally hydropower contributes 75% of renewable energy sources (U.S Energy Information Administration, 2023). Hydropower is the generation of electricity using the energy of flowing water (U.S Energy Information Administration, 2023). South Africa possesses vast water resources, including its shoreline however, the country does not fully utilize these resources for hydropower production. While South Africa has the option to utilize its water catchments and seawater for hydropower generation, the 2010-2030 Integrated Resource Plan focuses solely on run-of-river hydropower. (Bekker et al., 2022) highlights the need for proper planning in South Africa for hydropower production, particularly at the municipal level.

Globally hydropower accounts for about 3% of energy consumption. This is a clear indication of the underutilization of hydropower (Ramarope et al., 2023). (Ramarope et al., 2023), explains that the low generation and use of hydropower also reflects a lack of policy frameworks, which creates enabling conditions for the promotion of renewable energy. Some factors, such as natural disasters, outside the government's control, may also impact hydropower generation.

The dynamic nature of hydropower production, influenced by environmental variables such as water inflows, precipitation patterns and temperature, necessitates advanced

predictive tools (Ekanayake et al., 2021). Accurate forecasting of hydropower generation is vital for improving operational efficiency, ensuring resource sustainability and responding to fluctuating energy demands. Traditional methods often fail to capture the nonlinear relationships and complex interactions among the numerous variables affecting hydropower systems (Aksoy, 2021). Consequently, there is a growing interest in leveraging machine learning (ML) techniques, which have shown remarkable success across various industries, to address these limitations in hydropower forecasting (José et al., 2022). ML, a subset of Artificial Intelligence, provides powerful methods for analyzing large datasets and identifying complex patterns that traditional techniques often miss (José et al., 2022). In the hydropower sector, ML has been successfully applied to predictive maintenance, turbine optimization and energy output forecasting (Jung et al., 2021) (Ramarope et al., 2023). Among the commonly used ML models, Random Forest is an ensemble of decision trees capable of handling complex data structures (Sahin and Ozbay Karakus, 2024), while K-Nearest Neighbors (KNN) is particularly effective for analyzing simpler datasets by identifying patterns based on neighboring data points (Sahin and Ozbay Karakus, 2024).

## **6.2. Overview of hydropower production**

South Africa renewable energy sector is guided by the White Paper. The objectives of the White Paper are to:

- Ensure that an equitable level of national resources are invested in renewable technologies.
- Direct public resources to the implementation of renewable energy technologies
- Introduce suitable fiscal incentives for renewable energy production and consumption.
- Create an enabling investment climate for the development of the renewable energy sector.

The white paper does not mention the use of technology such as artificial intelligence to optimize renewable energy production. This highlights the gap for artificial intelligence in addressing renewable energy barriers.

Hydropower generation, a cornerstone of renewable energy, depends upon the intricate dynamics of the water cycle, necessitating the need to understand its fundamental processes. The water cycle unfolds through stages shaped by solar energy: surface heating prompts evaporation, leading to condensation and eventually rainfall, which refills water bodies and continues the cycle (U.S Energy Information Administration, 2023). To reliably forecast hydropower production, various data variables are essential, with temperature and rainfall being particularly important due to their significant impact on water availability and river flow. To accurately model hydropower generation, researchers have carefully analyzed different input variables and their effects. River flow is a crucial factor, influencing turbine performance and showing the connection between weather conditions and climate changes ((Ramarope et al., 2023), (Aksoy, 2021)). Additionally, factors such as wind speed, temperature, and humidity significantly affect the water cycle, highlighting the complex relationship between environmental conditions and hydropower output. Interestingly, although precipitation usually increases river flow, its direct impact on hydropower generation can vary depending on location and time, as shown by detailed correlation analyses that was conducted in a study by (Ekanayake et al., 2021).

### **6.3. Application of machine learning in hydropower production**

The integration of machine learning techniques into the domain of hydropower generation has emerged as a promising avenue for enhancing operational efficiency and optimizing energy production (Aksoy, 2021). However, in South Africa, research on the application of machine learning to predict hydropower generation is limited (Ramarope, 2023). Globally, the application of machine learning in hydropower generation is met with varying perceptions. While some may view it as a complex and abstract concept, others recognize its potential to revolutionize energy production processes by leveraging data-driven insights (Jose et al., 2022). Machine learning algorithms are fundamental in hydropower generation, serving as guiding principles that lead researchers through the intricate

process of analyzing data and making decisions. These algorithms enable predictive modeling, optimizing turbine efficiency and forecasting energy output based on historical and real-time data (Jose et al., 2022).

In hydropower generation, machine learning algorithms function as powerful tools for data analysis and prediction. They are deployed to interpret vast datasets encompassing factors such as water flow rates, weather patterns and energy production metrics (Aksoy, 2021). The purpose of utilizing machine learning in this context is to derive actionable insights that facilitate informed decision-making and enhance operational efficiency. Machine learning models employed in hydropower generation can be classified as scholarly theories. These models are constructed based on systematic analysis of data and aim to provide predictions and optimizations for energy production (Condemni et al., 2021) processes. They encompass a range of techniques, including regression analysis, neural networks, decision trees and ensemble learning algorithms. Machine learning theories in hydropower generation are dynamic and adaptive, continually evolving through iterative learning processes. They are characterized by their ability to analyze complex datasets, identify patterns and generate accurate predictions.

#### **6.4. What are Random Forest and KNN models**

According to (Dutta et al., 2021) random forest is a supervised machine learning classifier that consists of a collection of decision trees each with a unique independent vector and uses the most common class of  $x$  as the input. To create each tree in the forest, a series of steps were followed. First, in the bootstrap test,  $N$  records are randomly sampled with replacement from the original data to create  $N$  training datasets. This is the initial step in developing the tree. If there are  $M$  input variables, a smaller number  $m < M$  is randomly selected at each node. The best split among these  $m$  variables is used to divide the node. The value of  $m$  remains constant during the forest development. Each tree is extended to its maximum extent. Multiple trees are then generated in the forest, with the number of trees pre-determined by the parameter  $N$  tree.

KNN has shown suitability for hydropower generation prediction due to its straightforward, instance-based learning approach. This method is particularly useful when handling datasets where relationships between features and target variables are non-linear and complex. KNN's non-parametric nature means it does not assume any prior distribution about the data, making it flexible for hydropower forecasting where inputs can be highly variable and influenced by fluctuating environmental conditions(Gao et al., 2019)

Past studies have demonstrated positive outcomes when applying KNN for hydropower prediction. For instance, research involving the optimization of hydropower generation through machine learning approaches highlighted that KNN performed well in terms of accuracy for short-term energy forecasts(Gao et al., 2019).

Machine learning models like K-Nearest Neighbors (KNN) and Random Forest have demonstrated strong predictive capabilities in various studies involving energy forecasting, including hydropower generation. KNN is particularly effective for scenarios where simple yet robust prediction methods are needed, benefiting from its non-parametric nature and ability to adapt to complex, non-linear relationships in data (Gao et al., 2019) (Hanoon et al., 2023). On the other hand, Random Forest, which builds multiple decision trees and averages their results, is valued for its high accuracy, resistance to overfitting and effective handling of missing data and noisy features(Ekanayake et al., 2021, Mlambo and Mhlanga, 2022). These models have proven beneficial in hydropower prediction by leveraging historical meteorological and hydrological data, optimizing operational strategies and minimizing forecast errors (Ramarope et al., 2023) . KNN's simplicity makes it a good choice for straightforward, short-term predictions, while Random Forest's ensemble nature allows for more complex, long-term forecasting and sensitivity analysis. The application of these models helps enhance operational decision-making, contributing to more efficient energy management and planning(Hanoon et al., 2023)

### **6.5. Performance of machine learning models**

In a study conducted in South Africa to predict hydropower generation (Ramarope et al., 2023) collected the lowest and highest temperature variations, minimum humidity, dew



point temperature, average pressure and median temperature as inputs variables for the a neural network and random forest model. In the abovementioned study the neural network multilayer perceptron model had 10 layers and it performed well. Using the trial-and-error method, the intermediate layer conducted necessary computations to forecast the flow. The optimization algorithm was developed utilizing operational data, which was further employed for calibrating and validating the forecasting models. On the other hand (Ekanayake et al., 2021) predicted hydropower generation using Gaussian process regression (GPR) and support vector regression (SVR), multiple linear regression (MLR) and power regression (PR). The Gaussian regression model outperformed the other models.

The models that used machine learning to predict hydropower generation using the neural networks algorithms performed better compared to other models ((Condemi et al., 2021), (Ekanayake et al., 2021), (Aksoy, 2021)). In addition to the neural networks performing well, support vector machine models seemed to also have better hydropower generation predictions (Aksoy, 2021). Furthermore, a study conducted in Turkey on the use of machine learning to predict hydropower production highlighted that models together with feature reduction mechanisms such as component analysis and feature grouping techniques can improve the performance of prediction models (Sahin and Ozbay Karakus, 2024). In a study conducted in South Korea, neural networks were used to develop a model that predicted hydropower generation, using temperature, humidity and precipitation. The model performed well and predicted that hydropower production will decrease in the future based climatic conditions (Jung et al., 2021).

## **6.6. Evaluation of models**

The model performances are evaluated using statistical methods such as mean absolute error (MAE), root mean square error (RMSE)/ Mean square error (MSE), mean squared error (MSE) and mean fundamental percentage error (MAPE). These metrics are used to assess the accuracy and precision of the model's predictions. Additionally, the correlation coefficient (R-value) is used to validate the accuracy of the model's output. The model's performance can also be compared across different phases, such as training, validation

and testing, to ensure its generalizability and ability to predict future hydropower generation accurately.

### **6.7. Research Gaps**

There are multiple gaps that exist in prediction of hydropower generation. These gaps encompass the absence of research on the analysis of feature selection's impact on hydropower production value. There is also a gap in the comparative study of machine learning methods for estimating hydro energy value and the optimization of hyperparameters in hydropower artificial intelligence models. Furthermore, there is a need for more research on accuracy and performance evaluation criteria for different AI models used in hydropower, as well as the impact of hyperparameter tuning on the accuracy and performance of machine learning and deep learning models in estimating energy production in hydropower (Aksoy, 2021).

(Ekanayake et al., 2021) study did not take into consideration the riverflow and wind speed which affects the turbine. This may be one of the reasons for the other 5 models to underperform. On the other hand (Ramarope et al., 2023) recommends that future research should focus on studying how precipitation patterns might change in the future and how hydropower plants can adapt their power generation strategies to cope with climate change. This research is crucial for ensuring the resilience and sustainability of hydropower systems in the face of environmental shifts. (Ramarope et al., 2023), also acknowledged the difficulty of projecting time series data over a long period, particularly when dealing with small datasets. (Ramarope et al., 2023) identified the challenge of using limited dataset and time series analysis. The study uses historical meteorological data from 2001 to 2019. While this provides a basis for training and validating the ML models, it may not have captured long-term trends or variations in hydropower generation. Additionally, the time series analysis conducted in the study is limited to a small dataset, which may not fully capture the complexity and variability of hydropower generation. In contrast, (Ekanayake et al., 2021) in their study started off by using monthly data, later on to improve the model performance the dataset was reduced and

quarterly data was used. On the other hand, (Aksoy, 2021) used hourly data. However, the data used by (Aksoy, 2021) was directly from the generator.

Another significant challenge in forecasting hydropower production lies in the scarcity of studies addressing the estimation of power generation in extensive areas, such as entire river systems. Consequently, there exists a necessity for further research aimed at developing more accurate methodologies for predicting the power output of hydropower plants across large geographical regions (Aksoy, 2021). Additionally, (Ekanayake et al., 2021) highlights the insufficient exploration of techniques to enhance the prediction of hydro-power capacity through simplification methods. One such method is principal component analysis. Further studies are required to compare the effectiveness of various simplification techniques against not employing any simplification, facilitating the identification of optimal approaches for predicting hydro-power capacity. Other feature reduction techniques such as feature selection or autoencoders could be explored to further improve the prediction accuracy (Condemi et al., 2021). An example of the feature that could be included is energy demand and reservoir levels (Condemi et al., 2021).

In a study conducted by (Ekanayake et al., 2021) one of the recommendations was the importance of combine rainfall data from multiple locations to increase the prediction capability of a model. (Jung et al., 2021), mentioned that there is lack of accurate runoff prediction for small hydropower station and there is need to combine more than one hydropower model when predicting river runoff, which impacts hydropower production. Similarly, (Condemi et al., 2021) recommends evaluating the performance of already developed prediction model in different geographical regions.

## **6.8. Data variables**

Independent variables will be water flow rates, temperature, humidity, water levels and river flow because they impact hydropower generation (Ramarope, 2023). The dependent variable will be hydropower generated. Predictions of hydropower offer insights into future energy production trends and the underlying factors influencing them.

## 6.9. Data variables commonly used for predicting hydropower

Hydropower generation is reliant on the water cycle, making it crucial to comprehend its fundamental stages. Therefore, it is important to understand the three steps of the water cycle.

The water cycle encompasses several processes:

- i. Energy from the sun heats the surfaces of rivers, lakes, and oceans, causing water to evaporate.
- ii. The evaporated water vapor then condenses to form clouds.
- iii. Subsequently, the condensed water precipitates as rain or snow.
- iv. This precipitation gathers in streams and rivers, eventually flowing back into lakes and oceans, where it evaporates once more, thus initiating the cycle anew (U.S Energy Information Administration, 2023)

Below is a table summarizing the data used to estimate the hydropower production. The most commonly used data are temperature and rainfall because they affect water availability and water flow in the river. Feature selection is very important to ensure good performance of a model.

Most of the data is sourced from literature reviews, hydropower stations, river basin authorities and meteorological offices. On the other hand (Ramarope et al., 2023) data was obtain from a model called NASA satellite.

**Table 1: Data variables that affect hydropower production**

Input variables	How does it affect hydropower generation
1. River flow	River flow affects turbine. Water flow is one of the most important variables because it is essential in understanding how meteorological variables and climate change will influence hydropower production(Ramarope et al., 2023, Aksoy, 2021).
2. Wind speed	The primary source of hydraulic power for producing hydropower is kinetic energy, which depends on water availability to influence the fluid's velocity (Condemi et al., 2021).

3. Temperature	Temperature affects water availability through evaporation rate. (Condemi et al., 2021)
4. Evaporation/ 5. humidity	Evaporation presents the loss of water in the reservoir and it is directly influenced by temperature.  Humidity is an important factor that affects the hydrological cycle and can influence the amount of runoff (Jung et al., 2021).
6. Precipitation	Precipitation in the form of rainfall and snowpack. Rainfall affects water levels in the river. However, in (Ekanayake et al., 2021), in one location the correlation analysis showed that rainfall had no impact on hydropower generation.
7. Generator data	Generator (Gen.) and turbine data can also be used to predict hydropower generation.  Gen. Bed Temperature Front (°C), Gen. Bed Temp. Back (°C), Gen. Winding Temperature _L1 (°C), Gen. Winding Temperature _L2 (°C), Gen. Winding Temperature _L3 (°C), Gen oil Cooling Temp. Input (°C), Wingspan (%), Turbine Forced Pipe Pressure (Bar) , Turbine Regulator Pressure (Bar), Turbine Regulator Temperature (°C), Turbine Loading Pool Water Level Front (m) , Turbine Loading Pool Water Level Back (m) , Turbine Tail Water Level (m) and Transformator Oil Temperature (°C) (Aksoy, 2021)
8. Solar radiation	Solar radiation influences the water cycle and the melting of snowpack, which affects water availability for hydro-power generation (Condemi et al., 2021)
9. Slope and soil qualities	Basin slope, curve number, soil database, and digital elevation model can be used to create a detailed model (Jung et al., 2021).

### iii. Feature Selection:

The following features were identified as the import important factors that affect hydropower production

- a) hydropower production
- b) Rainfall, temperature, humidity
- c) Reservoir levels

(Jung et al., 2021), (Aksoy, 2021), (Ramarope et al., 2023) and (Ekanayake et al., 2021)

### 6.10. Key concepts

- Hydropower generation: The process of generating electricity from flowing water, typically using turbines to convert the kinetic energy of water into electrical energy.
- Machine learning: A subset of artificial intelligence that focuses on the development of algorithms and models that enable computers to learn from and make predictions or decisions based on data without explicit programming.
- Prediction models: Mathematical representations or algorithms that utilize historical data to forecast future outcomes, in this case, predicting hydropower production based on environmental variables.
- Environmental factors: Variables such as rainfall, river flow rate, reservoir levels, and temperature that influence hydropower generation and are considered inputs to the prediction models.
- Data collection and analysis: The process of gathering and examining historical data on hydropower production and environmental variables to identify patterns, trends, and relationships that can inform the development of prediction models.
- Model evaluation: Assessing the performance and accuracy of prediction models using metrics such as mean squared error, root mean squared error, or coefficient of determination to determine their effectiveness in forecasting hydropower production.
- Optimization: The process of refining prediction models and adjusting parameters to improve their accuracy and reliability in predicting hydropower generation.
- KNN: non-parametric, supervised machine learning algorithm used for classification and regression tasks. In regression, KNN predicts the value of a target variable by finding the average of the  $k$  nearest data points (neighbors) in

the feature space. The algorithm works by calculating the distance (often Euclidean) between data points and selecting the closest ones as the basis for the prediction. The simplicity of KNN lies in its reliance on the idea that similar input features should result in similar output responses, making it easy to implement and interpret (Barrash et al., 2019).

- **Random Forest** - A Random Forest is an ensemble machine learning algorithm used for classification and regression tasks, composed of multiple decision trees trained on different random subsets of the data and features. Each tree in the forest makes a prediction, and the final output is determined by majority voting for classification or averaging for regression, which helps improve accuracy and prevent overfitting. By aggregating the results of diverse trees, Random Forest models are robust, handle complex, non-linear data relationships well, and provide insights into feature importance, though they can be less interpretable than single decision trees (Sessa et al., 2021).
- **R-Squared** – Is a Coefficient of Determination in regression, R-squared measures the proportion of variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, where 1 indicates that the model perfectly explains the variance in the data, and 0 means that the model explains none of the variance. This metric is important for models like KNN and Random Forest in regression tasks because it provides a general sense of how well the model captures the underlying trends in the data. An R-squared value close to 1 suggests that the model accurately predicts the target variable, while a lower value indicates a less effective model. However, R-squared alone doesn't account for overfitting; therefore, it's essential to pair it with other metrics for a complete performance assessment(Ekanayake et al., 2021).
- **Mean Absolute Error (MAE)** Mean Absolute Error is a metric that calculates the average of the absolute differences between the predicted and actual values. It is straightforward and provides a direct interpretation of the model's predictive

accuracy in terms of the average prediction error. MAE is important for evaluating KNN and Random Forest regression models because it shows how far the model's predictions deviate from actual observations. Unlike metrics that square the error (e.g., Mean Squared Error), MAE is more robust to outliers, making it a good choice when you want an error measure that reflects real-world prediction differences without excessive influence from extreme values(Ekanayake et al., 2021).

- **Root Mean Squared Error (RMSE)** Root Mean Squared Error is the square root of the average of squared differences between the predicted and actual values. It penalizes larger errors more than smaller ones due to squaring the errors before averaging, making it particularly useful when large errors are undesirable. RMSE is important for KNN and Random Forest because it provides insight into how well the model is predicting, emphasizing significant deviations. The lower the RMSE, the better the model's performance, as it indicates that the predictions are closer to the actual values. This metric is more sensitive to outliers than MAE, which can be both a strength and a weakness depending on the application(Ekanayake et al., 2021).

## **7. Research Methodology**

### **7.1. Research design**

The research followed a quantitative research methodology with a predictive approach to predict hydropower generation. An experimental design was used to identify the impact of environmental and climatic factors on hydropower production at the Drakensberg hydropower Station. The study involved collecting and analyzing secondary on various environmental and climatic variables, applying machine learning models to predict hydropower output and assessing the effects of these factors on the accuracy and reliability of the predictions. Two machine learning models (Random Forest and K-nearest neighbors) were developed and their accuracy was compared.

The study adopted a deductive approach, starting with a hypothesis about the relationships between environmental/climatic variables and hydropower output. Data was



collected and analyzed to test the hypothesis, three machine learning models (KNN, Random Forest and an Ensembled model of KNN and Random Forest) were applied and tested to predict hydropower generation and assess the environmental/climatic factors on the accuracy and reliability of the models.

## **7.2. Population**

The population is the entire dataset of historical data related to hydropower production, temperature, humidity, river flow rate, generation capacity and other relevant factors at the Drakensberg hydropower station. This dataset encompasses all available observation from 1974, when the Drakensberg hydropower station started operating, representing the complete set of data from which the 5 year data sample will be drawn for analysis.

## **7.3. Sampling**

Eskom can only provide datasets with a maximum timeline of 5 years (2020 – 2024). As a result, the study used the 5-year datasets provided by Eskom.

## **7.4. Data Collection**

Secondary data over 5 years was collected.

Historical data was collected from:

- a) hydropower production - Eskom 5-year data portal (Historical data on daily, monthly, or yearly hydropower production)

Source: <https://www.eskom.co.za/dataportal/data-request-form/>

- b) Rainfall, temperature, humidity levels – South Africa Weather Services

Source:

<https://www.visualcrossing.com/weather/weather-data-services>

- c) Reservoir levels – Department of Water and Sanitation dams' data portal

Source:

<https://www.visualcrossing.com/weather/weather-data-services>

d) River flow– The data has been requested from Umgeni Waters (The pumped storage scheme is built between the upper Braamhoek dam and the lower Bedford dam.

Source: - [Generic Drought Dashboards](#)

## **7.5. Data analysis method**

### **Data Preprocessing**

- There were no missing values in the data they will be handled. However, outliers were handled through imputation by replacing them with mean values.
- A statistical analysis was conducted including calculating the central points and exploratory analysis. Correlation and descriptive analysis will also be conducted.
- The data was standardized to ensure that data brings all features to a similar scale, especially for KNN and not for Random Forest
- The data was split into training data, validation data and model evaluation data.

Hypothesis testing was conducted to determine if there are significant relationships between hydropower production and other variables (rainfall, temperature, reservoir levels and river flow rate). Some of the hypothesis tests included t-tests to validate the relationships identified in exploratory analysis. The data was then visualized using bar charts, heat maps, line graphs and box and whisker plot.

### **Model Selection**

Random Forest and KNN was used to develop hydropower generation prediction models.

v. Model Training and evaluation:

The Drakensberg data will be divided into two parts. One for training the model using python and the other will be used for evaluating the model. Evaluation metrics such as mean squared error, root mean squared error.

## **8. Results**

## 8.1. Exploratory Data Analysis

According to Table 2, the data had no null values. There were 1491 datasets with 8 characteristics. The data features are shown in Table 1 below.

**Table 2: No. of null values in the dataset**

```
Null values in each column:
Driel Barrage           0
Woodstock Dam           0
Sterkfontein Dam        0
Hydro Water Generation  0
Pumped Water Generation 0
Drakensberg Hydropower  0
temp                    0
humidity                 0
precip                   0
dtype: int64

Columns with null values:
Series([], dtype: int64)
```

## 8.2. Statistical Analysis

Figure 1 displays the summary statistics for the environmental conditions, temperature, humidity and precipitation. The temperature has a mean of approximately 15.96°C, a median of 16.30°C, a minimum of 2.60°C and a maximum of 27.70°C values. Humidity shows a mean of 62.99%, a median of 64.30%, a minimum of 15.70% and a maximum of 98.20%, indicating a consistent distribution. Precipitation, with a mean of 3.41%, a median of 0.15%, a minimum of 0.00% and a maximum of 73.14%, demonstrates the highest variability, suggesting occasional extreme rainfall events.

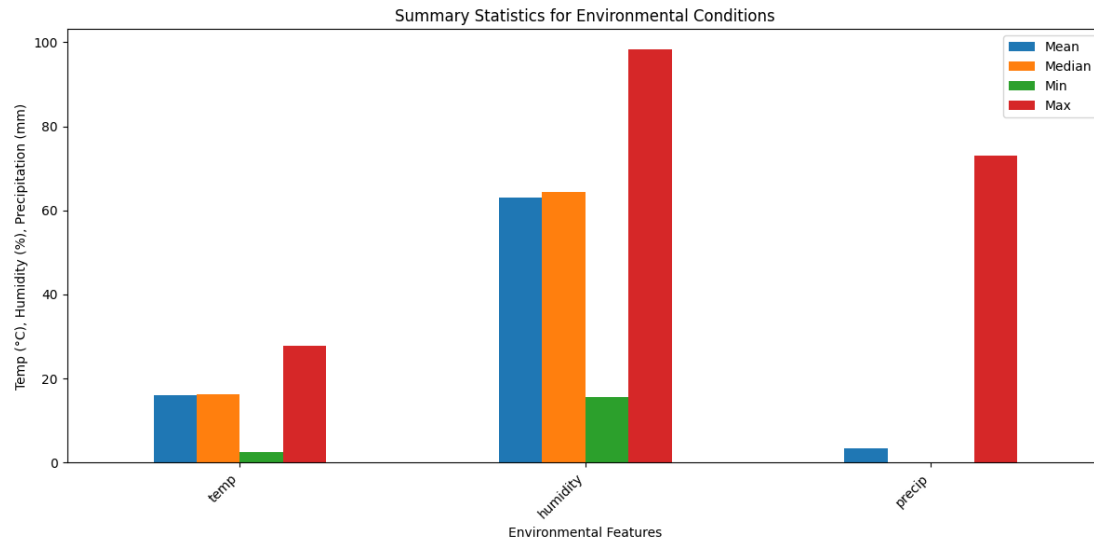


Figure 1: Summary of environmental conditions

Figure 2 shows that the mean for the Driel barrage dam levels is approximately  $4.28\text{mm}^3$ , with a median  $4.26\text{mm}^3$ , indicating a consistent central tendency. The minimum value recorded is  $3.92\text{mm}^3$ , while the maximum is  $4.76\text{mm}^3$ , highlighting a relatively narrow range. This suggests that the Driel Barrage levels are stable with minimal variability over the observed period.

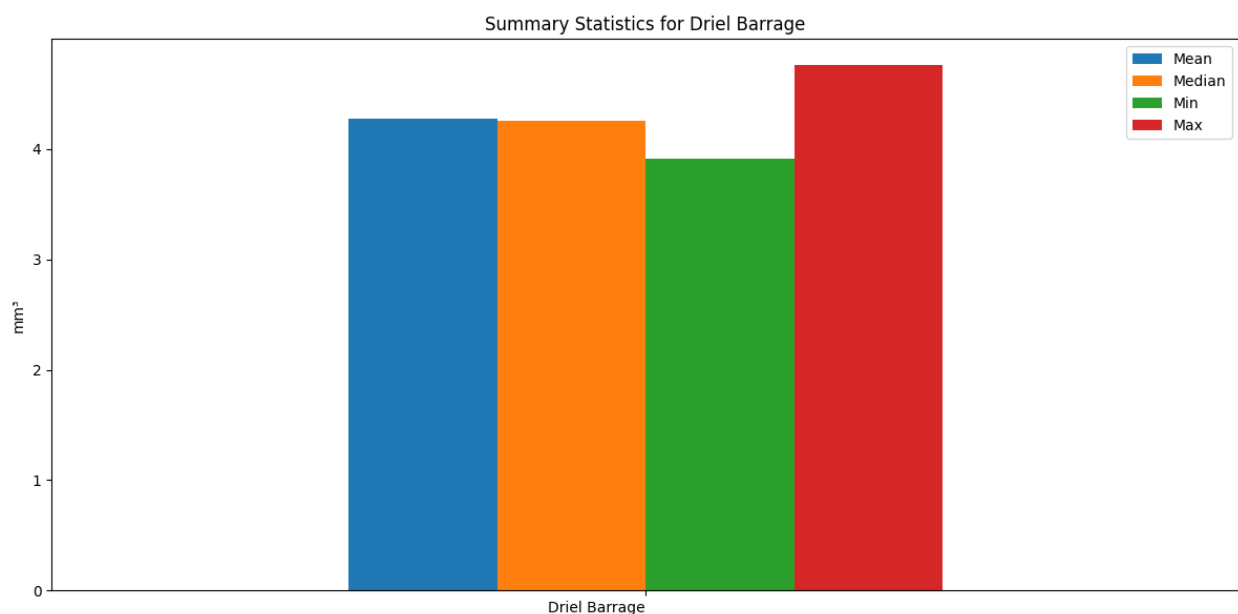


Figure 2: Summary statistics for Driel Barrage

Figure 3 shows that Sterkfontein Dam has significantly higher values across all summary statistics, with a mean of around  $2941.17 \text{ mm}^3$ , and a maximum value near  $3317.79 \text{ mm}^3$ , indicating it holds a larger capacity compared to Woodstock Dam. Woodstock Dam displays more consistent values with a mean of  $381.28 \text{ mm}^3$ , and a maximum of  $413.15 \text{ mm}^3$ , suggesting less variability. The overall data highlights that Sterkfontein Dam plays a more substantial role in water storage and potentially hydropower support.

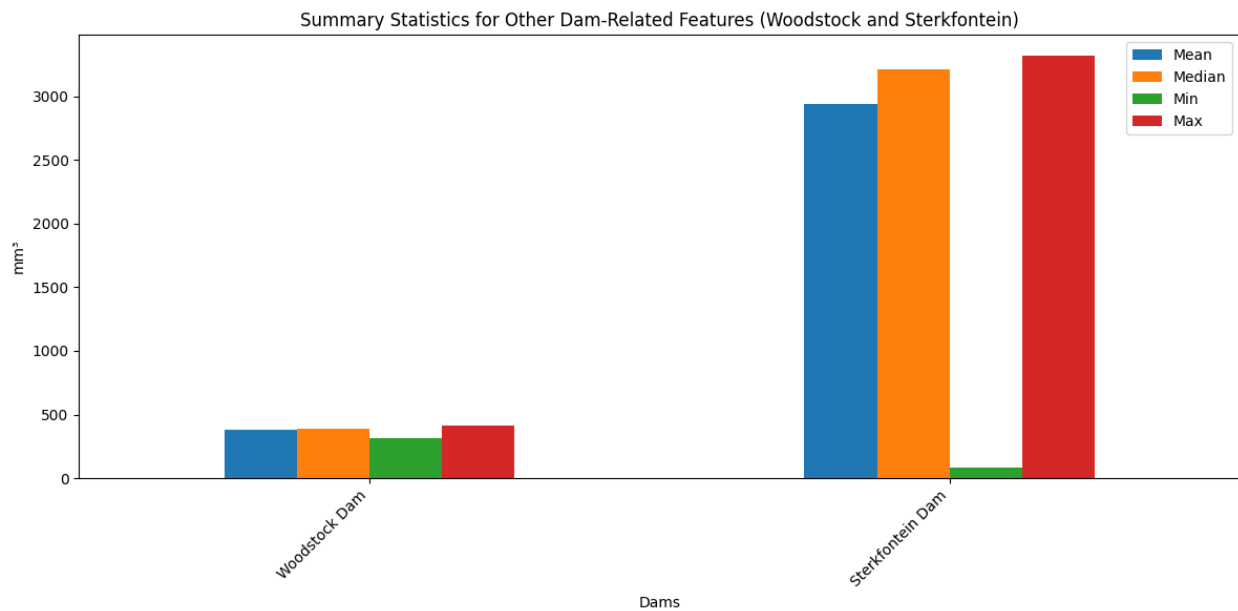


Figure 3: Summary statistics for Woodstock and Stekfontein

Figure 4 shows that Pumped Water Generation has the highest range, with a maximum value of approximately 1148.87 and a mean around 530.15, indicating significant variability. "Hydro Water Generation" has a mean of 221.15 but a much smaller maximum value of about 604.78, suggesting more moderate variability. Drakensberg hydropower displays the lowest range, with a mean of 81.74 Gen Unit per day and a maximum near 97.27 Gen Unit per day, reflecting relatively consistent production.

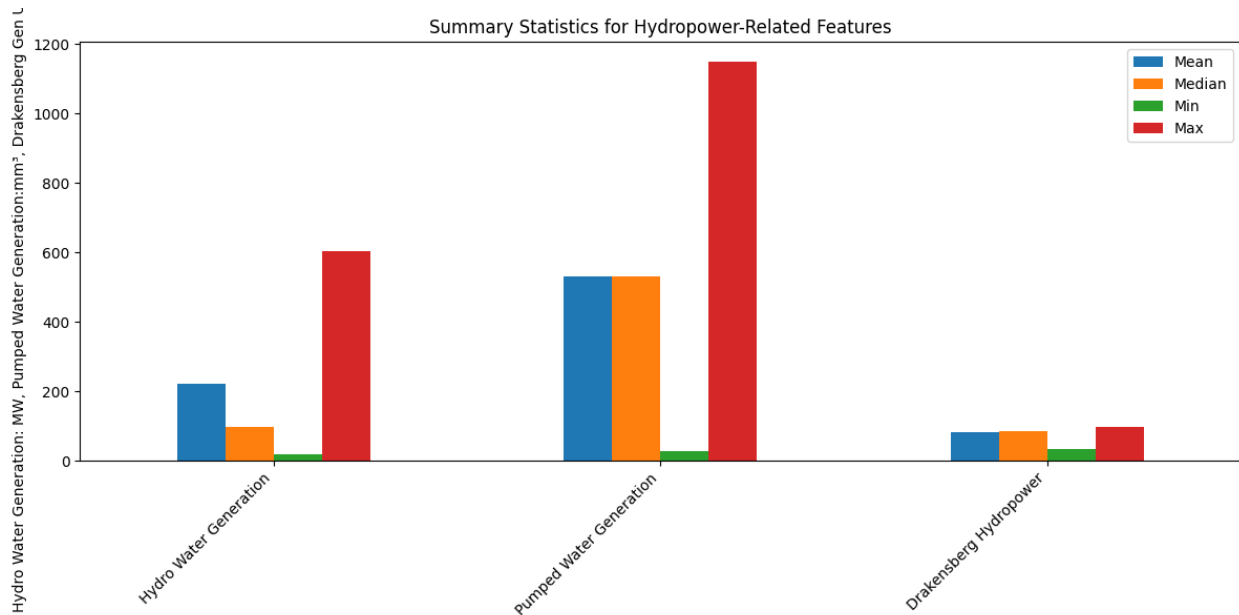


Figure 4: Summary statistics for Drakensberg hydropower station

The correlation matrix (Figure 5) shown below highlights the relationships among the features and their correlation with the target variable, Drakensberg hydropower. Notably, Driel Barrage has a moderate positive correlation (0.24) with the target, suggesting it may contribute useful information to predictive models. Conversely, Sterkfontein Dam shows a weak negative correlation (-0.14), indicating that its predictive power for the target might be limited. Features such as temperature and humidity display weak correlations with the target (-0.13 and -0.099, respectively), suggesting that their individual influence may not be significant. Additionally, the correlation between temperature and humidity is moderate (0.42), which indicates that these variables may share some overlapping information.

Considering potential redundancies, Woodstock Dam and Sterkfontein Dam exhibit a moderate positive correlation (0.51) with each other, suggesting that one of these features could potentially be removed to avoid multicollinearity. Similarly, humidity and precipitation show a notable correlation (0.42), indicating a shared relationship that may not add distinct value to the model if both are retained. Removing or combining features that are highly correlated can simplify the model, reduce overfitting and improve overall

interpretability. Ultimately, features with minimal or redundant influence on the target variable can be evaluated further for potential exclusion to enhance model efficiency.

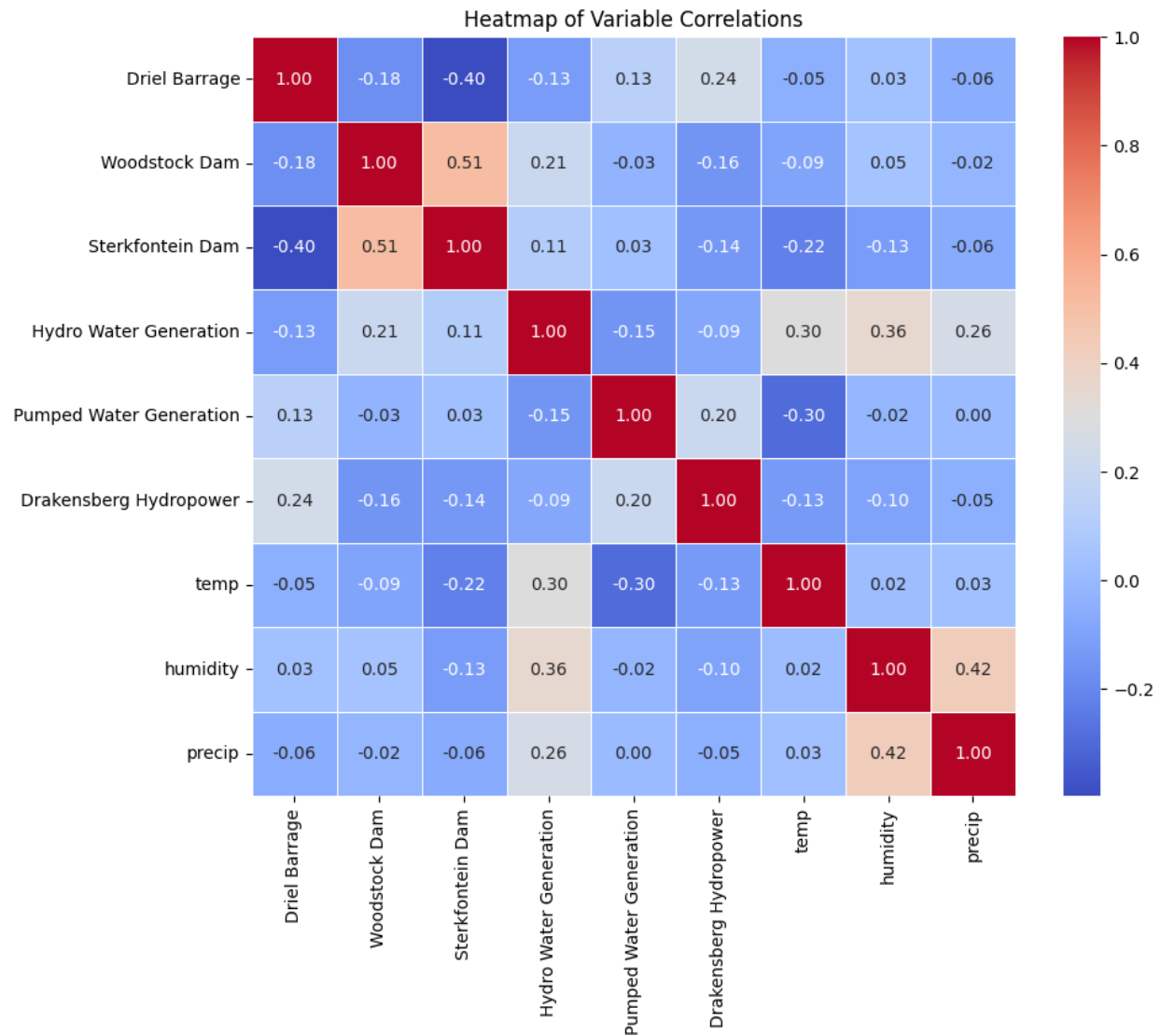


Figure 5: correlation matrix for the data variables

### **8.3. Overview of the KNN, Random Forest and Ensemble model results**

Below is detailed description of the different scenarios shown in Table 3, which were conducted to improve the performance of the model. The summary of the scenarios is shown in Table 3.

#### **Scenario One:**

In scenario one, the KNN model was applied to predict hydropower generation using a dataset where the target variable is 'Hydro Water Generation.' The data underwent several preprocessing steps, including scaling with MinMaxScaler and feature selection with SelectKBest based on univariate linear regression tests. This aimed to enhance model performance by focusing on the most relevant features.

The model used a grid search to optimize hyperparameters, selecting from a range of values for 'n\_neighbors,' 'weights,' and 'metric.' The best parameters found were 'metric': 'manhattan,' 'n\_neighbors': 6, and 'weights': 'distance.' The model's performance was evaluated using cross-validation (5-folds), yielding a Cross-Validation Root Mean Squared Error (RMSE) of approximately 126.70, Mean Absolute Error (MAE) of 80.12 and  $R^2$  of 0.61. These metrics suggest a reasonable prediction accuracy and model fit to the training data, indicating the model's ability to generalize somewhat effectively to unseen data.

The final evaluation on the test set showed an MAE of 76.97, an RMSE of 124.24, and an  $R^2$  of 0.63. These results are consistent with the cross-validation performance, supporting the model's efficacy under the selected hyperparameters and preprocessing techniques.

#### **Scenario 2:**

In this scenario, a Random Forest Regressor was employed to predict 'Hydro Water Generation'. Initially, the data was divided into an 80/20 split for training and testing. The



objective was to optimize the model's performance through hyperparameter tuning, which was achieved using a Randomized Search CV. This method tested 10 different combinations of parameters over 5-fold cross-validation, facilitating a comprehensive search across a broad parameter space to find the most effective model configuration.

The Randomized Search identified the optimal parameters as having 200 trees (`n_estimators`), no limit on tree depth (`max_depth`), a minimum of 2 samples required to split a node (`min_samples_split`) and a minimum of 2 samples per leaf (`min_samples_leaf`). This particular configuration allowed the model to capture complex patterns and interactions within the data, significantly enhancing its predictive power.

The performance of the Random Forest model showed considerable improvement, as evidenced by an  $R^2$  of 0.72 on the test dataset, indicating that the model could explain approximately 72% of the variance in hydropower generation. This was an indication of the model's improved accuracy and reliability. The important features were Woodstock dam and Sterkfontein Dam with scores of 0.29 and 0.21, respectively.

Additionally, the model underwent cross-validation to ensure its stability and generalizability across different subsets of data. The cross-validation results were very consistent with the test results, with an  $R^2$  of 0.72, a Mean Absolute Error (MAE) of 64.38, and a Root Mean Squared Error (RMSE) of 109.85. These metrics confirmed that the model not only performed well on the training data but also maintained its performance on new, unseen data, demonstrating excellent generalizability.

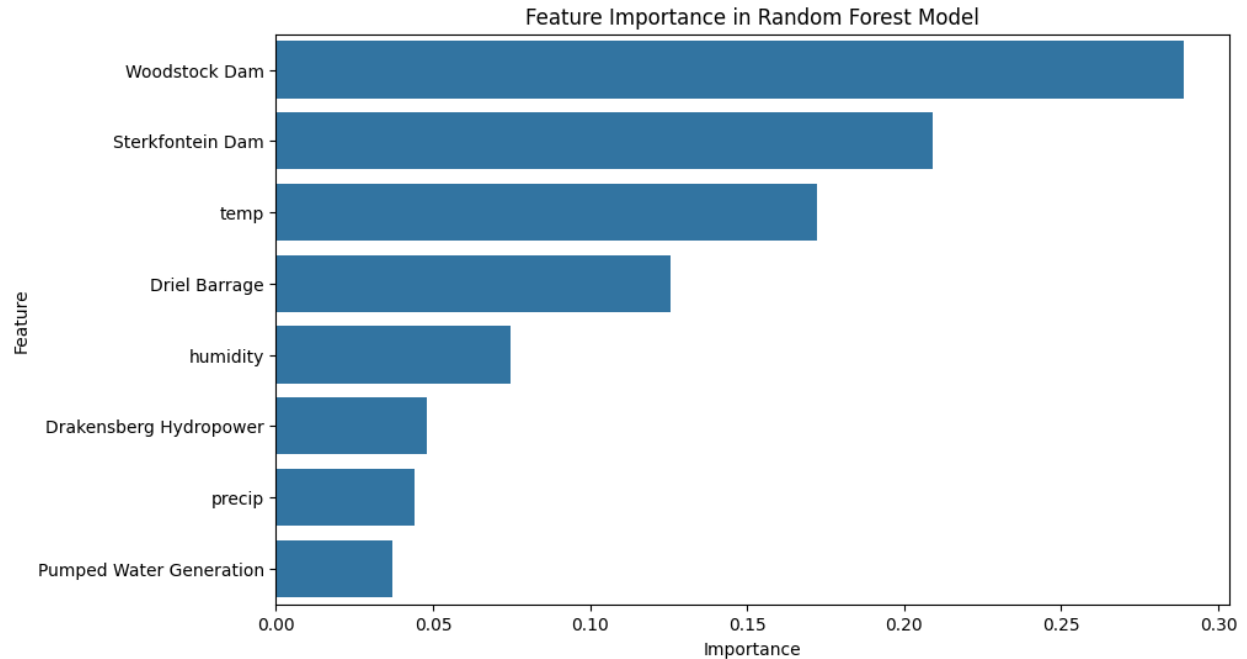


Figure 6: Feature Importance in Random Forest Model

### Scenario 3:

In scenario 3, two machine learning models—K-Nearest Neighbors (KNN) and Random Forest (RF) were ensembled and utilized to predict 'Hydro Water Generation'. An ensemble model was created by averaging the predictions from both KNN and RF models, leveraging the diverse strengths of each to enhance prediction reliability. This ensemble approach yielded a well-balanced performance with an MAE of 68.18, RMSE of 109.44, and an  $R^2$  of 0.714. This ensemble model managed to strike a balance between the two individual models, showcasing a robust method that potentially reduces individual model biases and variance, leading to more dependable predictions.

#### 8.4. Predicted vs the actual values

##### Scenario on: KNN

The scatter shows comparison between predicted and actual values of hydro water generation using the KNN model in Scenario One.

From the distribution of points, we observe that many predictions are clustered around the lower range of actual values, suggesting the model performs more consistently at this scale. However, as the actual values increase, the points tend to scatter more widely from the perfect fit line, indicating the model's predictions become less accurate for higher values of hydro water generation.

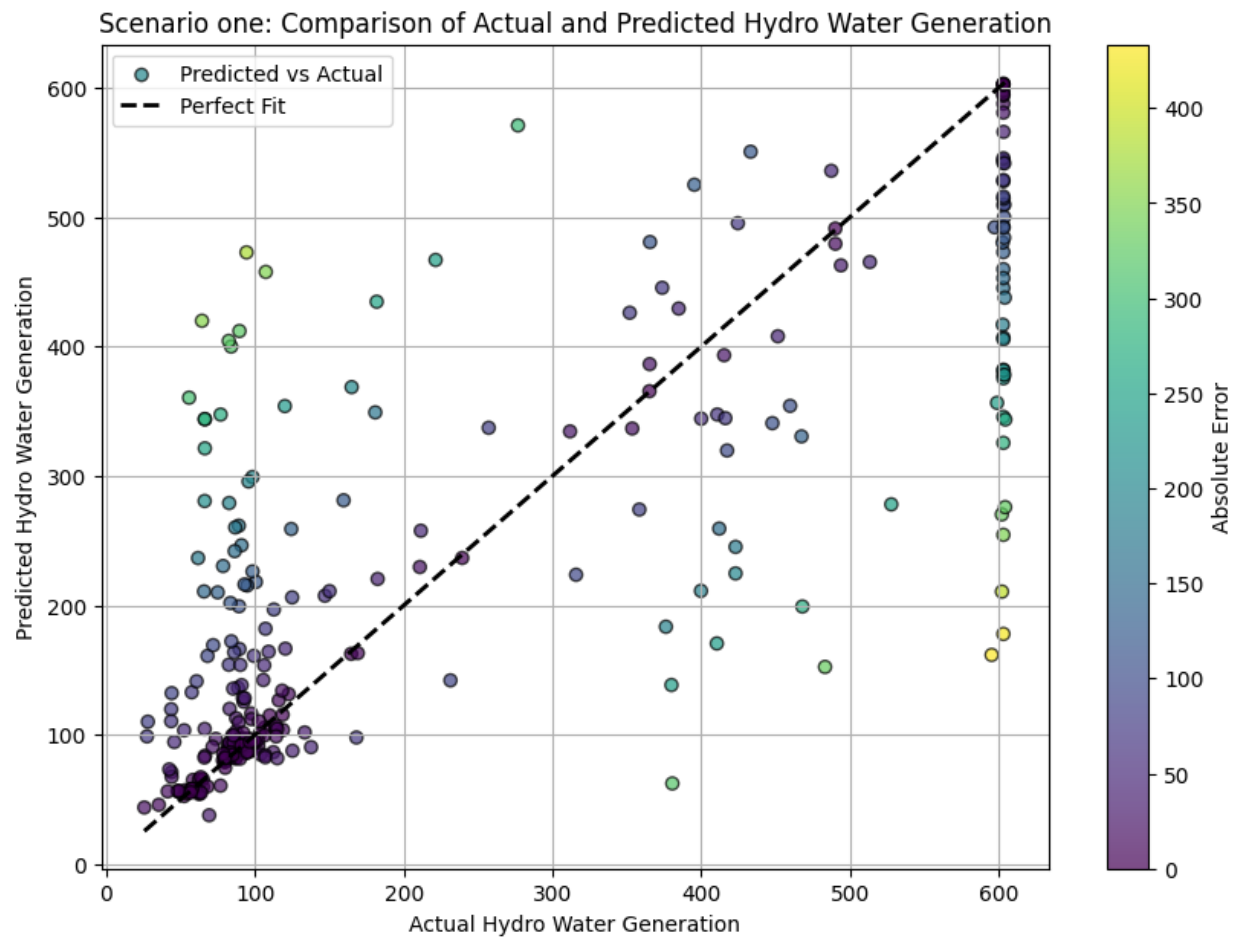


Figure 7: Scenario one: KNN

### Scenario 2: Random Forest

The scatter plot from Scenario Two visualizes the performance of a Random Forest model used to predict hydro water generation. Most of the data points cluster closely around this perfect fit line, particularly at the lower range of hydro water generation values, indicating that the Random Forest model generally predicts with high accuracy. However, as the

actual values increase, there is a slight spread in the data points from the line, showing that the model's predictions become less precise with higher values. Although some points, especially those colored yellow, indicate significant errors, these are relatively few and do not exhibit a systematic pattern.

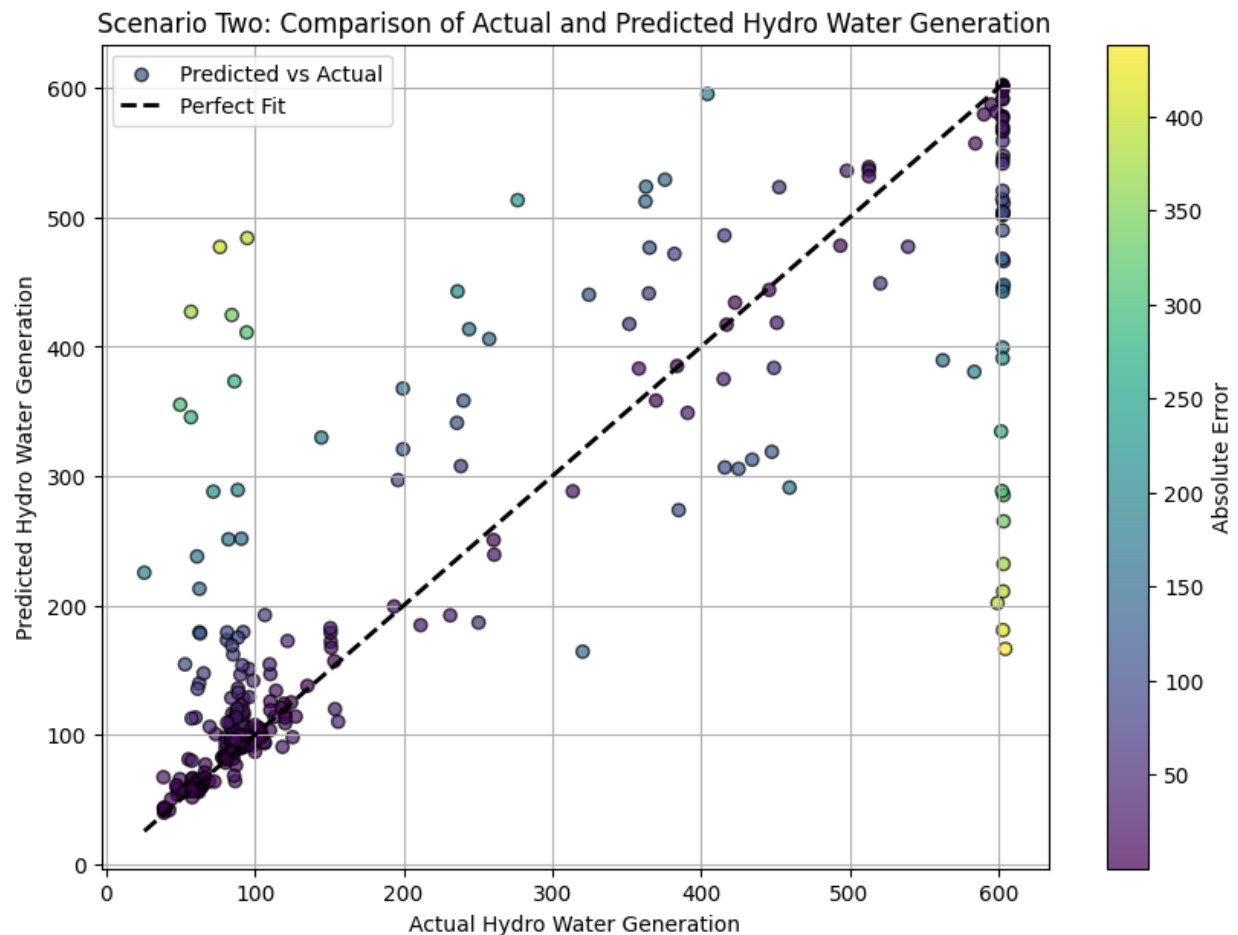


Figure 8: Scenario 2: Random Forest

### Scenario 3: Ensemble model

The scatter plot from Scenario Three showcases the performance of an Ensemble Model in predicting hydro water generation. The graph shows that the Ensemble Model performs effectively, especially at lower hydro water generation values, as evidenced by the clustering of points around the perfect fit line in this range. The consistent proximity of these points to the line across most value ranges suggests that the model generally

predicts with accuracy, maintaining a balance between underestimating and overestimating the actual values.

However, the model's performance exhibits some variability, particularly at higher values where the predictions diverge more significantly from the perfect fit line. This increased scatter, alongside the appearance of yellow-colored points, signals higher prediction errors at these higher values.

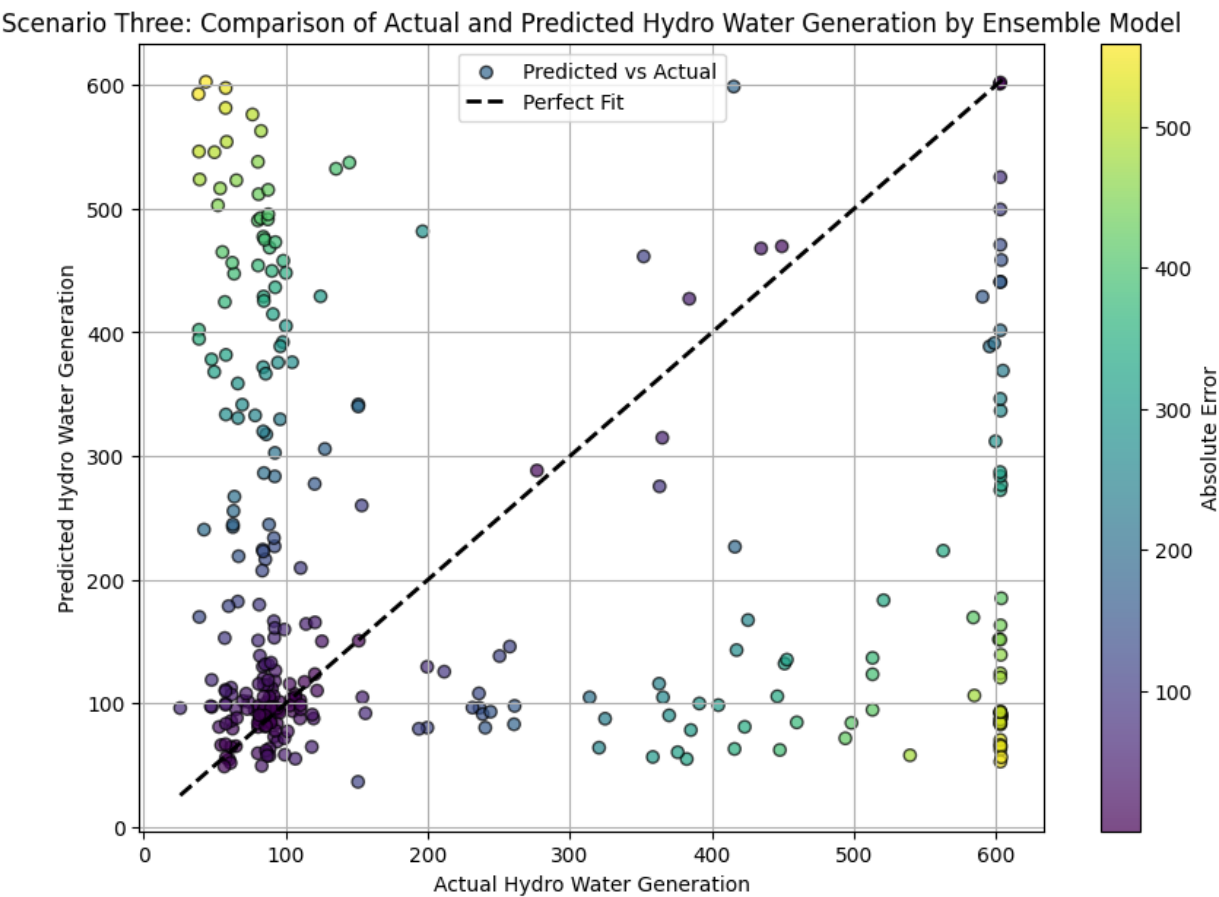


Figure 9: Ensembled model

## 9. Discussion

### Performance matrix of KNN model, Random Forest Model and an Ensemble model

The performance metrics presented in Table 4 compare three different predictive models: K-nearest neighbors (KNN), Random Forest and an Ensemble Model combining the

predictions of the previous two. These metrics are divided into evaluation (via cross-validation) and test results, which provides insights into how each model generalizes to unseen data.

In the evaluation phase, both the Random Forest and the Ensemble Model exhibit a coefficient of determination ( $R^2$ ) of 0.72, indicating that they can explain 72% of the variance in the target variable, which is substantially better than the KNN model's 0.61. This indicates that Random Forest and the Ensemble Model are more effective in capturing the relationships and patterns within the dataset. However, when considering the Mean Absolute Error (MAE), the Random Forest model shows better precision with the lowest MAE of 64.38, compared to the Ensemble Model's 68.09 and KNN's 80.21. This implies that while the Ensemble Model matches Random Forest in  $R^2$ , it does not predict individual values as accurately as the Random Forest model.

The test phase supports these findings, with the Random Forest and Ensemble Model maintaining an  $R^2$  of 0.72, demonstrating robustness and consistency in predicting new data. The MAE results in the test phase mirror those of the evaluation phase, where the Random Forest model achieves the lowest error (60.15), indicating that it not only predicts with high consistency but also with relatively better accuracy across individual predictions compared to the other models.

The Random Forest model consistently outperforms the KNN and Ensemble Models in both predictive accuracy and reliability across unseen data. The Ensemble Model, despite its decent performance, does not significantly enhance the results provided by the Random Forest alone, suggesting that the method of combining predictions might require refinement to better exploit the strengths of the underlying models. Meanwhile, the KNN model lags behind, possibly due to its simpler assumptions about the data, highlighting a need for parameter adjustments or more sophisticated feature engineering. This analysis strongly supports the use of Random Forest for this dataset due to its superior performance on both evaluation and testing grounds.

**Table 3: Performance matrices**

	Scenario 1: KNN	Scenario 2: Random Forest	Scenario 3: Ensemble Model
<b>Evaluation</b>			
R <sup>2</sup>	0.61	0.72	0.72
MAE	80.21	64.38	68.09
<b>Test</b>			
R <sup>2</sup>	0.63	0.72	0.72
MAE	76.97	60.15	68.82

In Scenario One, the underperformance of the KNN model in predicting higher values of hydro water generation can be attributed to several factors inherent to the nature of the KNN algorithm. KNN relies heavily on the local neighborhood of data points, which means its performance is directly influenced by the density and distribution of those points. At lower values where data points are densely packed, the model can make more reliable predictions. However, at higher values, the sparser distribution leads to less accurate predictions as the algorithm struggles to find a sufficient number of nearby points to form a robust average. This issue could be exacerbated by any outliers in the data, which are more influential in KNN compared to other algorithms, potentially skewing predictions significantly.

For Scenario Two, the Random Forest model exhibits strong performance at lower hydro generation values largely due to its ensemble nature, which aggregates predictions from multiple decision trees to improve accuracy and robustness. Each tree in the forest is built on a subset of data features, allowing the model to capture various aspects of the data effectively, particularly at lower ranges where these features may exhibit more consistent patterns. However, as the actual values increase, the model's performance begins to waver slightly. This slight spread in predictions could stem from the model's limited ability to extrapolate to higher value ranges not well-represented in the training data. Moreover, Random Forest models can sometimes overfit to noise in the training data, particularly if not properly tuned with parameters like tree depth or the number of trees, which might lead to less precise predictions at these higher ranges.

A study conducted by (Hanoon et al., 2023) on hydropower prediction highlights the importance of sensitivity and uncertainty analysis by removing outliers and calculating the 95PPU range, where predictions are expected to fall and assessing model reliability using a d-factor to measure the prediction band width. In this research project, similar methods were employed, addressing sensitivity and uncertainty by identifying outliers and replacing them with the median to enhance model stability and accuracy. This strategy aimed to ensure that predictive models like KNN and Random Forest could perform reliably when exposed to different datasets. The training, evaluation and validation of these models were structured into 3 scenarios, as presented in Table 4. The comparative analysis revealed that Random Forest did relatively better than KNN and the ensemble model. This result aligns with the findings of (Mlambo and Mhlanga, 2022) who noted the robustness of Random Forest in predicting hydropower production in a study conducted in South Africa.

As shown in Table 4, KNN was the least performing model. The performance of KNN was enhanced by improved the number of  $n_{estimators}$  and conducting a 5-fold cross-validation to ensure the model's ability to generalize across datasets. (José et al., 2022) supported this approach, emphasizing the importance of cross-validation for mitigating overfitting in machine learning applications related to improve machine learning used for predicting hydropower production.

In scenario two, the Random Forest model was improved through several adjustments to enhance its predictive power and generalizability. The initial Random Forest Regressor was set with 100 estimators and a random state for consistency. To address potential overfitting and improve performance on unseen data, a Regularized Random Forest was introduced, with constraints such as  $max\_depth=10$  and  $min\_samples\_split=2$ . These parameters helped control the complexity of the model by preventing the trees from growing excessively deep and overfitting the training data. The Random Forest model had an accuracy ( $R^2$ ) of 0.72 and an MAE (mean absolute error) of 64.38 when evaluating



it with cross-validation and it kept a consistent  $R^2$  of 0.72 with an even better MAE of 60.15 in the final testing.

However, as shown in figure 8, the actual values increase, the model's performance begins to waver slightly. This slight spread in predictions could be from the model's limited ability to extrapolate to higher value ranges not well-represented in the training data. Moreover, Random Forest models can sometimes overfit to noise in the training data, particularly if not properly tuned with parameters like tree depth or the number of trees, which might lead to less precise predictions at these higher ranges (Barreñada et al., 2024). This variability is also visible in the ensemble model, in figure 8, which could be improved by expanding the dataset, adding more relevant environmental features or incorporating more advanced predictive algorithms. Research by (Ekanayake et al., 2021) and (Gao et al., 2019) also emphasized that the proper handling of temporal data and feature scaling is crucial for accurate hydro-energy forecasting, which supports the findings in this analysis.

To enhance the predictive performance further, implementing more advanced models such as Artificial Neural Networks and Support Vector Regression could be beneficial. These models have shown strong results in energy forecasting studies, as they can capture complex, non-linear relationships more effectively (Ramarope et al., 2023) and (José et al., 2022). Additionally, incorporating time series analysis could provide more insights into temporal dependencies and patterns that influence hydropower production (Ekanayake et al., 2021)

Regarding the research questions, the comparative analysis showed that the Random Forest algorithm outperformed the K-Nearest Neighbors (KNN) model in predicting hydropower production. The Random Forest model achieved higher R-squared values, lower Mean Absolute Error (MAE) and Mean Squared Error (MSE) across most scenarios, indicating better predictive power. KNN, although improved with standardization, did not reach the accuracy levels of Random Forest. Therefore, Random Forest was determined to be better for this type of predictive task. The analysis revealed

that hydropower output showed significant positive correlations with environmental factors like reservoir levels and rainfall. Reservoir levels, particularly "Woodstock Dam and Sterkfontein Dam", were identified as a key factor influencing hydropower generation. Random Forest's feature importance analysis ranked these environmental factors as the most influential in predicting hydropower output. The null hypothesis was rejected because the dam levels and rainfall have an impact on the hydropower generation capacity.

## **10. Conclusion**

The primary goal of this research—finding an effective predictive model for hydropower generation—was partially addressed. Random Forest models with biweekly data preprocessing were confirmed to provide better predictive accuracy compared to KNN but the accuracy is very low. setting a path forward for both practical applications and future investigations. Additional work could explore the effects of varying data frequencies and more complex feature interactions to further refine these models and adapt them to other contexts. The research findings demonstrate that integrating biweekly data aggregation with Random Forest models yields improved predictive capabilities for hydropower forecasting. This approach effectively stabilized input features, reduced noise and allowed the model to capture some meaningful temporal patterns, resulting in an improved accuracy. The study underscores the importance of data stability, feature scaling and robust preprocessing methods in achieving optimal model performance. However increasing datasets may have produced better results. The null hypothesis ( $H_0$ ), stating that no environmental factors significantly impact hydropower generation capacity, was rejected. Instead, the alternative hypothesis ( $H_1$ ) was supported, as findings showed that variables such as reservoir levels and rainfall influenced hydropower output. These variables, particularly the levels, had a notable positive correlation with hydropower generation, validating their importance as key predictive features. This conclusion aligns with the feature importance rankings provided by Random Forest models, confirming the influence of these environmental factors.

Future research could expand on these findings by incorporating additional climatic variables. Increasing the dataset size or integrating external data sources would further enhance the generalizability and robustness of the model. Advanced ensemble techniques, including stacked models that combine the strengths of different algorithms, could also be explored for improved predictive outcomes. Real-time sensor data integration would enhance model adaptability, providing more accurate and responsive forecasts in dynamic environmental conditions.

The contributions of this study to renewable energy and artificial intelligence are substantial. It addresses key research gaps by comparing machine learning models, specifically Random Forest and KNN, in hydropower production prediction. The systematic analysis of how different features like temperature, humidity, rainfall and river flow affect hydropower output provides valuable insights into model accuracy. The focus on hyperparameter optimization for machine learning models, which has been limited in past research, is an essential step in improving prediction reliability. This work establishes benchmarks for model performance, aiding decision-making in the selection of appropriate AI techniques for hydropower systems.

Despite these contributions, the research acknowledges certain limitations. Model uncertainty remains an inherent challenge due to environmental variability, which may result in discrepancies between predictions and actual outcomes. Additionally, while humidity was included as a variable, the exclusion of certain data points like the evaporation rate might affect the comprehensiveness of the model. The dataset was limited to a five-year period, which have could restrict the model's ability to project long-term trends. However, this study still offers valuable insights into the operational efficiency of the Drakensberg Hydropower Station and supports broader renewable energy optimization efforts.

## 11. References

- AKSOY, B. 2021. Estimation of Energy Produced in Hydroelectric Power Plant Industrial Automation Using Deep Learning and Hybrid Machine Learning Techniques. *Electric Power Components and Systems*, 49, 213-232.
- BARRASH, S., SHEN, Y. & GIANNAKIS, G. B. Scalable and adaptive KNN for regression over graphs. 2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2019. IEEE, 241-245.
- BARREÑADA, L., DHIMAN, P., TIMMERMAN, D., BOULESTEIX, A.-L. & VAN CALSTER, B. 2024. Understanding random forests and overfitting: a visualization and simulation study. *arXiv preprint arXiv:2402.18612*.
- BEKKER, A., VAN DIJK, M. & NIEBUHR, C. 2022. A review of low head hydropower at wastewater treatment works and development of an evaluation framework for South Africa. *Renewable and Sustainable Energy Reviews*, 159, 112216.
- CONDEMI, C., CASILLAS-PÉREZ, D., MASTROENI, L., JIMÉNEZ-FERNÁNDEZ, S. & SALCEDO-SANZ, S. 2021. Hydro-power production capacity prediction based on machine learning regression techniques. *Knowledge-Based Systems*, 222, 107012.
- DUTTA, P., PAUL, S. & KUMAR, A. 2021. Comparative analysis of various supervised machine learning techniques for diagnosis of COVID-19. *Electronic devices, circuits, and systems for biomedical applications*. Elsevier.
- EKANAYAKE, P., WICKRAMASINGHE, L., JAYASINGHE, J. M. J. W. & RATHNAYAKE, U. 2021. Regression-Based Prediction of Power Generation at Samanalawewa Hydropower Plant in Sri Lanka Using Machine Learning. *Mathematical Problems in Engineering*, 2021, 4913824.
- GAO, X., SHAN, C., HU, C., NIU, Z. & LIU, Z. 2019. An Adaptive Ensemble Machine Learning Model for Intrusion Detection. *IEEE Access*, 7, 82512-82521.
- HANNAH, R., PABLO, R. & MAX, R. 2020. *Breakdown of carbon dioxide, methane and nitrous oxide emissions by sector* [Online]. Published online at OurWorldInData.org. Available: <https://ourworldindata.org/emissions-by-sector> [Accessed].
- HANOON, M. S., AHMED, A. N., RAZZAQ, A., OUDAH, A. Y., ALKHAYYAT, A., HUANG, Y. F. & EL-SHAFIE, A. 2023. Prediction of hydropower generation via machine learning algorithms at three Gorges Dam, China. *Ain Shams Engineering Journal*, 14, 101919.
- HISTORY. 2023. *Climate Change History* [Online]. HISTORY.COM EDITORS. Available: <https://www.history.com/topics/natural-disasters-and-environment/history-of-climate-change> [Accessed].
- IBM. n.d. *What is machine learning (ML)?* [Online]. Available: <https://www.ibm.com/topics/machine-learning> [Accessed].
- IEA. 2023. *Hydroelectricity* [Online]. Available: <https://www.iea.org/energy-system/renewables/hydroelectricity> [Accessed].
- JOSÉ, SANTOS, M., MODESTO DE ABREU, T., PRADO, L., MIRANDA, D., JULIO, R., VIANA, P., FONSECA, M., BORTONI, E. & BASTOS, G. 2022. Hydropower

- Operation Optimization Using Machine Learning: A Systematic Review. *AI*, 3, 78-99.
- JUNG, J., HAN, H., KIM, K. & KIM, H. S. 2021. Machine Learning-Based Small Hydropower Potential Prediction under Climate Change. *Energies*, 14, 3643.
- MLAMBO, F. & MHLANGA, D. 2022. Artificial Intelligence and Machine Learning for Energy in South Africa. *AfricaGrowth Agenda*, 19, 20-23.
- NIEBUHR, C. M., VAN DIJK, M., NEARY, V. S. & BHAGWAN, J. N. 2019. A review of hydrokinetic turbines and enhancement techniques for canal installations: Technology, applicability and potential. *Renewable and Sustainable Energy Reviews*, 113, 109240.
- PRINSLOO, L. & BURKHARDT, P. 2019. Eskom's Ingula hydro-power plant running at 25% below capacity [Online]. News24. Available: <https://www.businesslive.co.za/bd/national/2019-03-27-eskoms-ingula-hydro-power-plant-running-at-25-below-capacity/> [Accessed].
- RAMAROPE, S. I., FATOBA, O. S. & JEN, T.-C. 2023. Hydro-power generation forecast in South Africa based on Machine Learning (ML) models. *Scientific African*, 22, e01981.
- SAHIN, M. E. & OZBAY KARAKUS, M. 2024. Smart hydropower management: utilizing machine learning and deep learning method to enhance dam's energy generation efficiency. *Neural Computing and Applications*, 1-17.
- SESSA, V., ASSOUMOU, E., BOSSY, M. & SIMÕES, S. G. 2021. Analyzing the applicability of random forest-based models for the forecast of run-of-river hydropower generation. *Clean Technologies*, 3, 858-880.
- U.S ENERGY INFORMATION ADMINISTRATION 2023. Hydropower explained.
- WILHELM, K. n.d. *Enlight the rainbow nation: South Africa is the most industrialized nation in Africa with an abundant supply of natural resources. Hydropower* [Online]. ANDRITZ Available: <https://www.andritz.com//hydro-en/hydronews/hydropower-africa/southafrica> [Accessed].

Annex:

Link to colab code:

[https://colab.research.google.com/drive/1cDUBh\\_Q0rgnLVWaqzP8djYELyhaWX7ya?usp=sharing](https://colab.research.google.com/drive/1cDUBh_Q0rgnLVWaqzP8djYELyhaWX7ya?usp=sharing)



**Zertifikat  
Certificat**

**Certificado  
Certificate**

Promouvoir les plus hauts standards éthiques dans la protection des participants à la recherche biomédicale  
Promoting the highest ethical standards in the protection of biomedical research participants

**Certificat de formation - Training Certificate**

Ce document atteste que - this document certifies that



**Andiswa Nyongwana**

a complété avec succès - has successfully completed

**Regulatory framework of South-Africa (2024)**

du programme de formation TRREE en évaluation éthique de la recherche  
of the TRREE training programme in research ethics evaluation

Release Date: 2024/06/10  
CD : rNYCQk9to

Professeur Dominique Sprumont  
Coordinateur TRREE Coordinator



Programmes de formation continue (2 crédits)  
Continuing Education Programs (2 credits)

Federatio  
Pharmaceutica  
Helvetica

**FPH**

Programmes de formation  
postgraduate continues

Programmes de formation continue  
Continuing Education Programs

(RSEV - 30230217)

Ce programme est soutenu par - This program is supported by :  
European and Developing Countries Clinical Trials Partnership (EDCTP) ([www.edctp.org](http://www.edctp.org)) - Swiss National Science Foundation ([www.snf.ch](http://www.snf.ch)) - Canadian Institutes of Health Research (<http://www.cihr-irsc.gc.ca/fr/2091.html>) -  
Swiss Academy of Medical Sciences (SAMSGAMSW) ([www.samsa.ch](http://www.samsa.ch)) - Commission for Research Partnerships with Developing Countries ([www.crdp.ch](http://www.crdp.ch))