Brainae Journal of Business, Sciences and Technology
ISSN "2789-374X (print)" "2789-3758 (online)
Volume 26, Issue 7, March 2023
http://www. brainajournal.com info@brainae.org

SOLAR PHOTOVOLTAIC POWER OUTPUT PREDICTION WITH MACHINE LEARNING TECHNIQUES: CASE STUDY OF RWAMAGANA SOLAR POWER PLANT- RWANDA

[1]Nshimiyimana Aimé, [2]Benimana Raymond Celestin, [3]Tuyishime Silas, & [4]Hanyurwimfura Gerard;
[1]University of Rwanda-African Centre of Excellence in Data Science "UR-ACEDS"
+250 788 671098; nshimiyaime@gmail.com
[2]Indian Institute of Technology Delhi
+250 788 598765; raybeni02@gmail.com
[3]University of Rwanda-College of Science and Technology "UR-CST"
+250 788 552 792, tuysilas@gmail.com
[4]Indian Institute of Technology Delhi
+250 788 662 946, hanyurwagerard@gmail.com

[5]Co-Author: Dr. Kundan Kumar
University of Rwanda-African Center of Excellence in Data Science "UR-ACEDS"
+250 786183431; kundan.kumar011@gmail.com
Received: 20 February, 2023**; Accepted:** 15 March, 2023**; Published:**16 March 2023
https://doi.org/10.5281/zenodo.7746014

Abstract:

Solar Photovoltaic has been used for long due to potential shortage of fossil fuel energy, its effect on the environment, and the increase in energy consumption around the world. Solar PV power output is irregular in nature due to the intermittency and the other dependent factors, like wind velocity, irradiance, ambient temperature, humidity, etc. Due to uncertainty, it is challenging to predict power output from solar photovoltaic. The electric power system in Rwanda is facing a challenge of power outages that is caused by several reasons including the intermittence of energy from solar PV power plants which result to load shedding in different regions of the country. To attain a sustainable solution to this problem; prediction of the power output from solar PV can be utilized which may help the utilities to plan the scheduling of the power system around the country in a predictive approach. This can be achieved by intelligent algorithm which is used to analyze the data of the recent five years. For both metrics (i.e., R2-SCORE and RMSE) used to evaluate the three machine learning methods, the KNNR with parameter K=8 is the best model. It achieved R_SCORE of 0.978 as shown in Table 2. K in this case means that it uses 8 data points that are nearby the input of the concern. These algorithms provide the prediction model which can be considered as formula which is very useful to the utilities in decision making.

Key words: *Machine learning, solar photovoltaic, renewable energy, metrics, regression model*

Introduction:

The development of a country cannot be achieved without having sufficient energy, and electrical energy is in amongst others. Rwanda is a sub-Saharan African country and has set its targets for economic growth strategies involving having sufficient electrical energy. All sectors place a high priority on reliable and competitively priced energy, driving a substantial need to improve energy capacity and infrastructure. Both on-grid and off-grid connection approaches are promoted in Rwanda, so that electricity access may be accelerated to meet the national target which is to provide 100% of electricity access to the population by 2024 [1].

The electrical energy in Rwanda is generated from hydroelectric power plants, thermal power plants (methane gas, peat, Diesel), and solar photovoltaic (PV) power plants. The electric power system in Rwanda is facing a challenge of power outages that are caused by several reasons including the intermittence of energy from solar PV power plants which results to load shedding in different regions of the country. The issue of power reduction is due to the intermittence of solar-based power generation and it is addressed by

using thermal power plants regardless of their associated consequences like CO2 emissions in the atmosphere and the global shortage of their primary resources.

A sustainable solution can be achieved when renewable energy resources are used and exploited at the maximum level. Solar photovoltaic is among the available renewable energy resources in Rwanda and when the prediction of its output power is properly done; the power outages, load shedding, and sometimes blackout in the power network system, will therefore be minimized or eradicated.

The machine learning methods have been used to predict the output power [2] and this project of "Solar photovoltaic power output prediction with machine learning" will use the data from the existing solar PV power plant to make power plant output predictions with the help of linear, decision tree and KNeighbors regressor techniques.

These intelligent algorithms were used to analyze the data of Rwamagana power plan, and to build its associated prediction model. The results from this analysis are very useful and helpful for the decision-makers of the plant. The predicted results can inform the organization whether the plant is generating a profit, and they can also be used in checking the depreciation of the plant equipment after a certain running period.

Rwanda is a small country situated in east-central Africa below the equator, it is bordered by Uganda in the East, Tanzania in the North, Burundi in the South, and the Democratic Republic of Congo in the West. Understandably, it is a landlocked country. Rwanda covers roughly 25,000 (94%) square kilometers of land and 1,400 (5.3%) square kilometers of water. Rwanda's population was estimated at approximately 13,084,494 people according to the 2020 world meter index [3].

Rwanda has five provinces namely, Eastern Province, Western Province, Northern Province, Southern Province, and the City of Kigali. Rwamagana solar power plant which is taken as a case study of this project is situated in Eastern Province, Rwamagana District, Rubona Sector, Karambi Cell. Rwamagana Solar PV power plant is a grid-connected to medium voltage (MV) which feeds to Musha Substation. Power plant prediction will additionally help for improving the power supplied to the population, the grid reliability, and decrease the frequent power outages occurring in some areas of the country.

Problem statement:

As like in many African sub-Saharan countries, some localities of Rwanda have characterized by the power outage, cutoffs, blackout over the last five years due to the electrical energy shortage [4]. The study of Bimenyimana et al.[5] has shown that; in 2019 Rwanda had a crisis of electricity involving power outages (blackouts) in the grid-connected users. An increase of the population did not match or reflect the production of electrical energy (i.e., the demand is too high).

For instance; the rate of industries operating in the country has gone beyond the plan, and effort required to establish new power plant is not straight forward. Proper plan of budget, enough time for its establishment must be done. Power plant are so expensive to establish. It is too challenging to have immediate output power without having a computation function of it that can help in future planning.

Objectives of the Study:

The main objective is to make a Solar photovoltaic power output prediction with machine learning techniques, and the following objectives were achieved:

1. To perform a preliminary analysis of Rwamagana output power plant data.
2. To build predictive model of output power from Rwamagana power plant using the linear regression, decision tree and KNeighbors repressors.
3. To compare results from the used machine learning algorithms in predicting the output power.

Research Questions:

1. How Rwamagana power plant data can be described?
2. How the output power is predicted for a known input variable?
3. How to get the best algorithm among the used machine learning regressors?

Conceptual Review:

Solar Photovoltaic (PV) is a power plant constructed using semiconductor components or parts [6]. These semiconductor devices convert energy of the light into electricity. Silicon and other elements of

group IV elements are most commonly used semiconductors to conduct electricity in some conditions. A PV module is composed with series and parallel combination of solar cells. The number of series and parallel cells determines the voltage and current of the PV panel respectively. The PV module converts light photons from sunrays into electrical energy, this electrical energy is produced at the level of the cell which is the basic energy converting device for a PV module.

The model of the PV panel is converted into an equivalent circuit which consists of a current source and a single diode with series resistance (*Rs)* and a parallel resistance (*Rp).* A solar PV has 4 modules (i.e. Parameters of Solar PV Module, Module I-V Characteristic, and Maximum Power Point Tracking "MPPT").

Parameters of Solar PV Module are available under; open circuit voltage (Voc), Short circuit current (Isc), maximum power point current (Impp), maximum power point voltage (Vmpp), and maximum power (Pmpp). the open circuit voltage, and the short circuit current are the two important parameters, to consider when talking about IV characteristic. The open circuit voltage is the potential measured, between the cell's output terminals and no-load conditions. On the other hand, the short circuit current is the current that flows through the terminals when a short circuit is applied between them. In the two case there is no power transfer, i.e. in case of open circuit, output current (I_{out})= *0*, therefore output power (*P) = 0*, in case of short circuit the terminal voltage is zero *V = 0)*, which also makes *P = 0*.

Maximum Power Point Tracking (MPPT) is an automated system that varies the parameters of a photovoltaic module, so that it can yield the highest amount of power it can. The system does not change the module position to follow the sun light although mechanical trackers also exist[7]. MPPT depending on the inputs varies the duty cycle of a DC-DC converter, and therefore assists the load to get the maximum power possible from the source, without changing the source.

However, if both; an electronic and mechanical tracking system are used together, the performance may be improved. The performance of a PV depends of many parameters. Some of these, are related to the holding environment and location (i.e. irradiance of solar, temperature, current material lifetime, parasitic resistances, and etc.[6]. For better understanding how MPPT works, first we need to know that a maximum power point tracker (MPPT) which is used with the purpose of getting the maximum power possible from the module. Maximum power is conveyed from a source to load when the source circuit impedance is identical to the load impedance as the maximum power transfer theorem states [6].

Therefore, tracking maximum power point is nothing else other than trying to match the module's impedance with the load impedance. To achieve impedance matching appropriate equipment in the name of the converter is required, by adjusting the converter duty cycle the impedances should be controlled. The change of the duty cycle corresponds to the change in impedance as seen in the source also changed, till the impedance matching is attained. The amount and the direction of the variations are determined by an algorithm, which is finally called a maximum power point tracker (MPPT)[8][9].

Methods like ANN (Artificial Neural Network), Perturb and Observe (P&O), Improved Perturb and Observe, and many others were used for Maximum Power Point Tracking (MPPT). ANN stands for Artificial Neural Network and it is one type of the machine learning algorithm. Machine learning is a set of algorithms used to build models[10].

Machine learning is data-driven approach based on mathematics and computer programming where the concepts of probability, statistics, linear algebra, computer science, and algorithms are used to build intelligent models[11]. This may be regression or classification. The active power feed is measured using numeric continuous values, so machine learning recalls the use of regression models for modeling our data. There are several algorithms used for regression but only linear regression, K-nearest neighbor regressor (KNNR), and decision tree will be reported in our results.

Theoretical framework:

Researchers across all disciplines (i.e. science, technology, engineering, mathematics, etc.) have shown that effective research recalls the use of the literature review, theoretical and conceptual frameworks [12]. There is a strong correlation between these frameworks for the current successful research. This way of new researchers gives a better direction toward the research methodological decisions while providing clear and important results.

An indicator of its success is that there is an increase in researchers using these frameworks. A combination of literature and the two frameworks has a significant guide for the research using; qualitative data, quantitative data, or a mixture of them. This

approach provides a good way to answer research questions used like in traditional methods. This way of research can also cope with interviews, surveys, observations, artifacts, or other instruments used for the existing research.

A literature review is the fundamental part of the research where the research questions needs are specific, its problem, and it covers an investigation of the relevant topic within the same field and similar orientation f research. The framework is referred to as the basic structure to guide the research and it has two main types which are theoretical and conceptual frameworks. Theoretical frameworks involve a set of theories to support the research.

Conceptual Framework:

The conceptual frameworks use an abstract representation of model research by referring to the theoretical framework and literature review

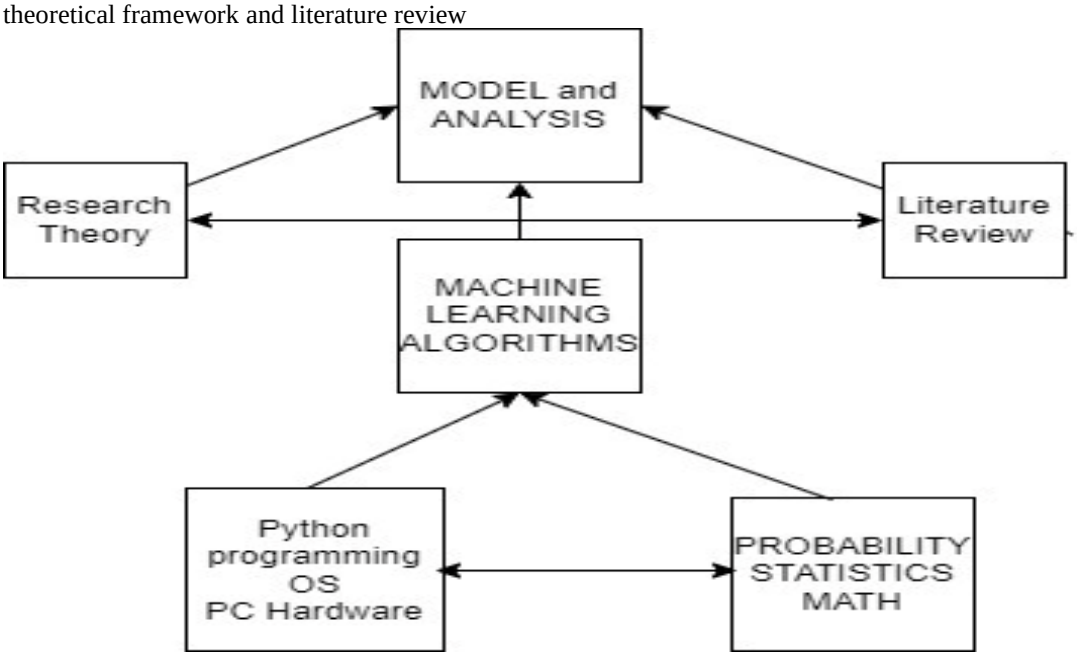


Figure 1. The research conceptual framework

Figure 1. shows that; computer programming, statistics, machine learning algorithms, theory, and literature of related work are input variables to achieve the analysis and modeling of the ongoing research.

Material and Methods:

This research is a mixture, i.e., quantitative and qualitative. Data from the power plant are numbers and being interested with the measured power output recall quantitative while there is another type of data from questionnaires/interviews that are assessed with the concept used in qualitative analysis.

The concepts required for the two types were followed properly. Surveys, tools for data collection for analysis, sampling strategies to choose respondents, etc. have all been done based on the research principles of quantitative and qualitative research. Secondary data was obtained from the Rwamagana power plant; each data entry was recorded within 15 min period in the last 4 years (i.e., 2016 -2019).

The research has used 63 individuals from GIGAWATT GL-BAL RWANDA, 123 from electricity sellers, and 247 from STECOMA which is an independent wood processing located at GISOZI (AGAKIRIRO); and this should result in a population of 433 individuals. Two techniques of sampling were used to choose the sample, that is random and stratified sampling. And the number of individuals to participate in the research was computed using Taro Yamane formula (Yamane, 1973) on Eq. [3].

$$n = \frac{N}{1+Ne^2} \quad (3)$$

n = sample size

N = population size

e = error (0.1) reliability level 99%, or

e = level of precision always set the value of 0.1

Thus, the sample size is

$$n = \frac{433}{1 + 433 * 0.1^2} = 81$$

The sample used was 81 individuals in the research.

Machine learning algorithms:

Linear regression is the simplest machine learning technique and its commonly used for building a predictive model based on data [13]. The nature of the relationship between the independent variable (x) and dependent variable(y) is linear. Eq. [1] is the general mathematical form of it.

$y = Ax + B + \epsilon$ (1)

The y is the dependent variable which is proportional to the variation of x. ϵ is the error term. The purpose of linear regression is to find the coefficients A and B that minimizes errors. The K-Nearest-Neighbors (KNN) is a non-parametric algorithm (i.e. no presumption required on the initial dataset), the output of new input is computed by the number of nearest neighbors (K); the majority among the neighborhood points determines the value/class of that input[14].

In classification, the majority of classes show the label of unknown input but the regression needs to use Euclidian distance. By changing the K parameter, the best prediction model is determined by the value of K with minimum errors. It is called K-Nearest-Neighbors Regressor (KNNR) when it is used for regression purposes. Regardless of other challenges, this algorithm is effortless.

A decision tree is a classification/regression machine learning that uses an iterative partitioning of the dataset starting from the root node while forming a tree structure representation graph [15]. The logic behind this approach is complex, and the CART (classification and regression) is a well-known and famous algorithm for its implementation.

The CART performs a successively binary partition from the root node. The node here means the attribute of the dataset and its selection depends on the calculations of impurity criteria for the samples of the same category or with an approximate value in the final nodes (i.e., leaves). The starting point of the partitioning is random, but it becomes fixed if the splitting founds minimum impurity.

Evaluation methods:

R2-SCORE and MSE are two metrics that can be used to control and evaluate regression machine learning models. The correlation level of predicted and observed values are measured by R2-SCORE [16].

The Mean Squared Error (MSE) is a commonly used regression metric it computes the average squared error between the actual and estimated values[17] but there is the possibility to use RMSE which is the mathematical squared root of MSE. The MSE metric is only challenged by the outlier, and Eq. [4] shows its mathematical form of RMSE.

$$\text{MSE} = \frac{1}{n}\sum_i^n (p_i - r_i)^2 \quad (2)$$

$$RMSE = \sqrt{MSE} \quad (3)$$

$$RMSE = \sqrt{(\frac{1}{n}\sum_i^n (p_i - r_i)^2)} \quad (4)$$

RMSE: root mean squared error,
MSE: mean squared error,
n: number of samples,
p_i: predicted value of the sample at i-index,
r_i: actual value of the sample at i-index.

Findings and Discussions

This part provides description of the process used to get results as shown on Figure 2.

Figure 2. Process of data analysis and modeling

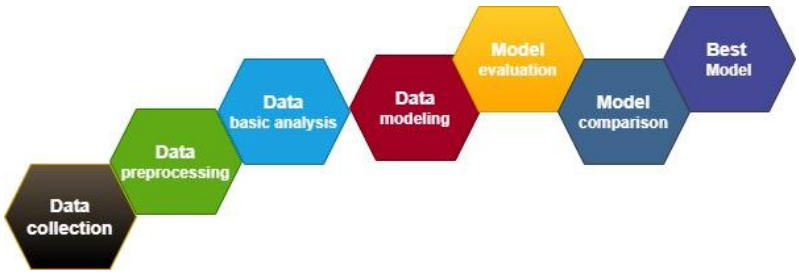


Figure 2 is the summary of activities of the research. Collection, preprocessing, analysis and modeling of data are main task of the research.

Data collection:

The secondary data were taken from Rwamagana solar PV plant and they are made by only 2019 has recorded 49967 observations, and the rest of the years had recorded 50000 observations for each year. This has resulted in 249967 records. These data have been organized into 5 columns which are timestamp, irradiation (W/m^2), wind speed (m/s), and the average active power feed (W). The active power feed is the response

variable of the analysis and other variables were taken as features. Each observation is measured within 15 minutes recorded in the database. Other data were taken from the 81 respondents thought questionnaires and interviews.

Data preprocessing:

A five years' data collection from 2016 to 2020 was preprocessed and analyzed using computer programming (i.e., Python).

Data analysis:

In the 249967 observations collected, the timestamp variable was removed from the dataset as it does not have a logical impact on the active power feed response. The zero output power observations have been removed because the amount of irradiation is too small (its values are closer to 0). The new dataset had 8290 observations that could pass through linear regression, decision tree, and KNNR machine learning algorithms for the active power feed prediction model development. The 81 respondent's data and the plant data are analyzed in the next section.

Table 1. Electricity cases versus users

	Electrical Energy Sellers				
	STECOMA		**GIGAWATT**	**TOTAL**	**%**
Power Outage	14	6	2	22	***27.2***
Blackout	2	1	3	6	***7.4***
Stability in electrical energy supply	25	20	8	53	***65.4***
TOTAL	41	27	13	81	
%	***50.6***	***33.3***	***16.0***		

Table 1. shows the frequency table between electricity partners and their opinion about power outage, blackout and their stability of electrical energy supply.

Figure 3. The plot of electricity cases versus users

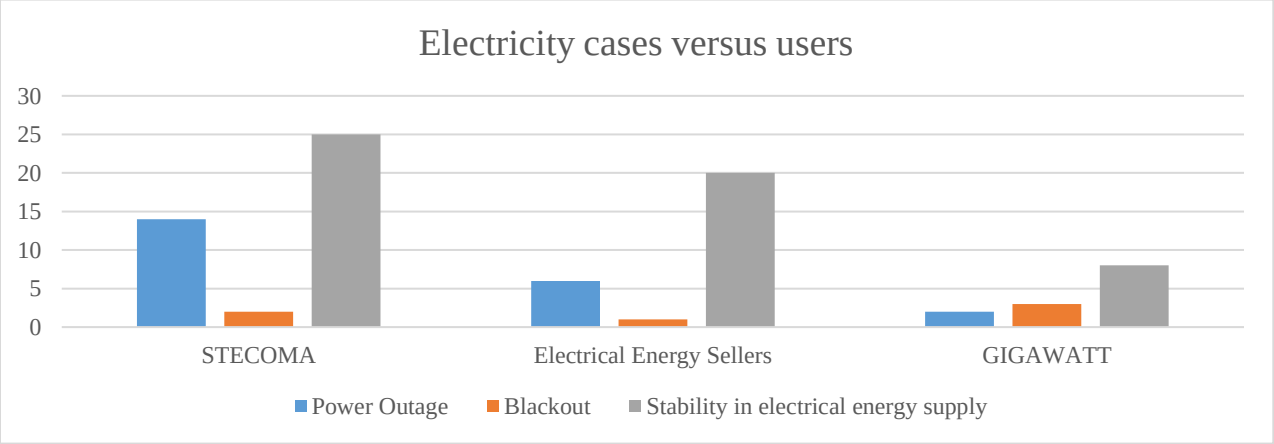


Figure 3. shows that; the sample used in the research is stable in electricity supply but, there is some power outage in the consumers of electricity.

Figure 4. Power average moth plot

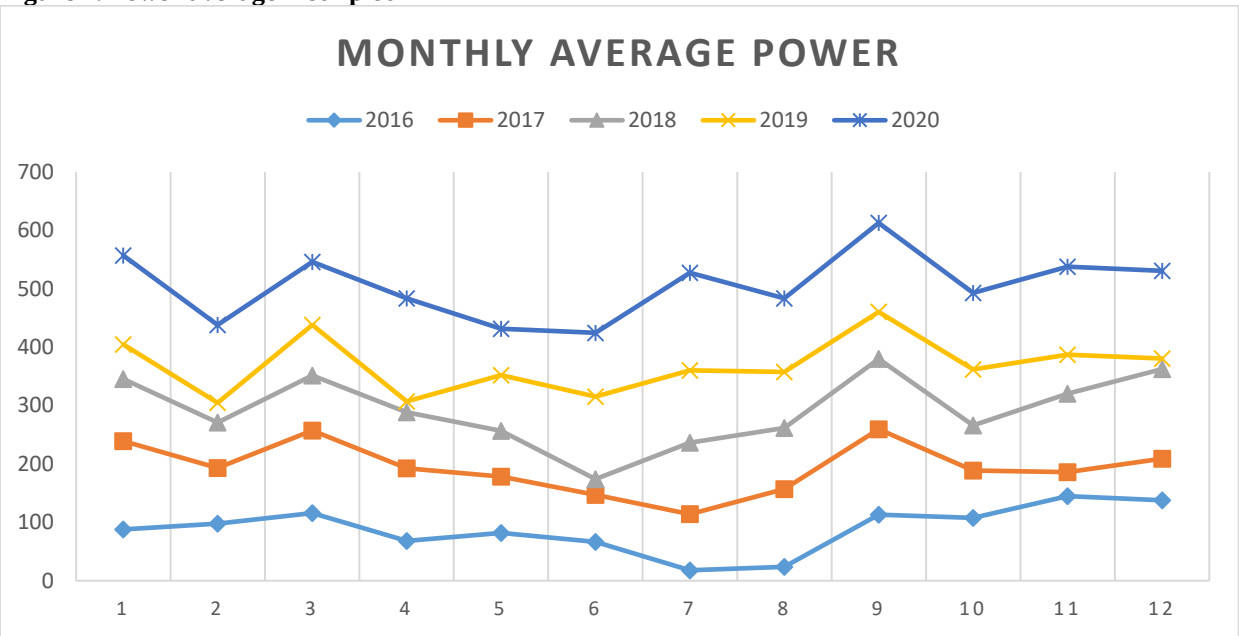


Figure 4. shows that; for a period of 5 years, there is a short decrease in the power in January, a short increase of power in February, a long decrease in power between March and June, an increase up to September, and another short decrease at the end of the year. September and march are good periods to produce high values of power output.

Figure 5. Solar irradiation versus active power output

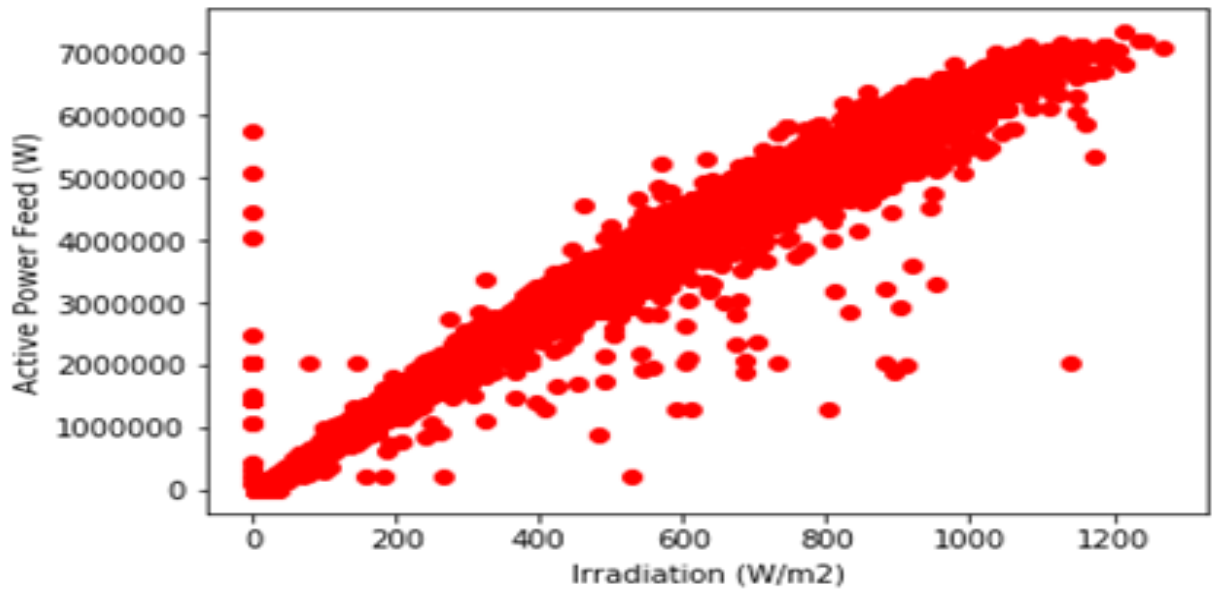
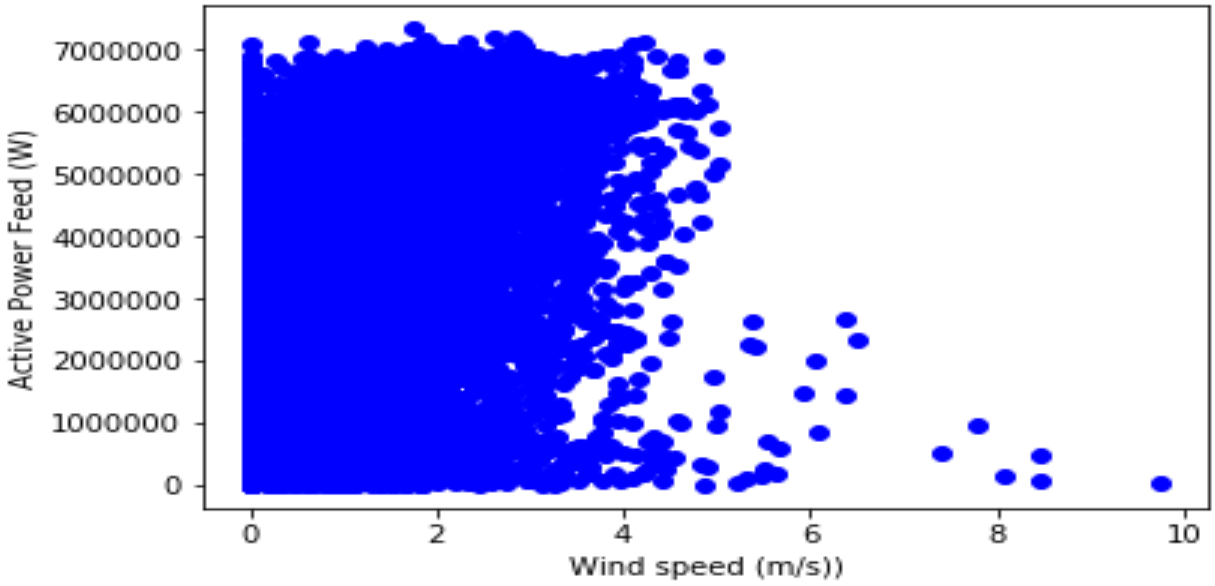


Figure 5. shows that there is a positive relationship between irradiation measured in W/m^2 and active power feed in Watts (W). This means that when the irradiation increases the average output power feed increases proportionally up to around 1100 W/m^2, and when the irradiation goes beyond 1100W/m^2 the average output power start to decrease slowly with an increase in irradiation.

Figure 6. Wind speed versus active power



Unlike the irradiation relationship (Figure 5) the wind speed and the active power feed (Figure 6) are inversely proportional from each other. An increase of one measurement corresponds to a decrease of the other measurements and vice versa. It is appeared that when the wind speed is greater than 3m/s the power start to decrease.

Figure 7. Temperature versus active power

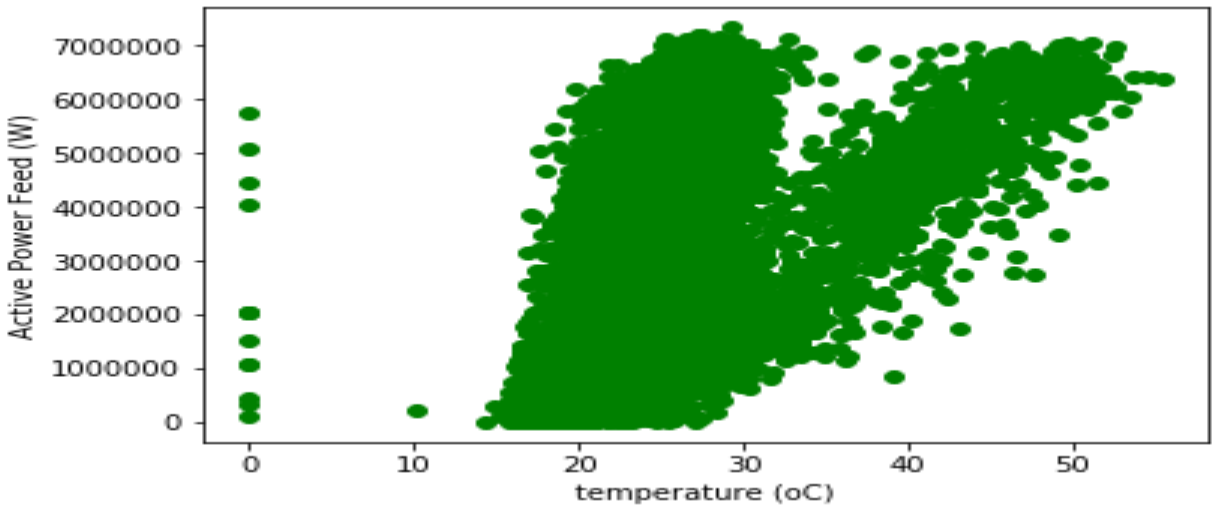


Figure 7. reveals that the minimum temperature to produce the power is around 13° C, and when it reaches about 25°C the power is at its maximum value. After 25°C the power starts to decrease due to the temperature coefficients of the solar panels, and then it continues to increase in a normal way.

Data modeling:

A total of 8290 observations were divided into training and testing sets. This had made 6217 observations (75% of data) to train each model, a 2073 observation (25% of data) which is the rest of the data was used to test each of the models for evaluation. The model evaluation performance is reported using R2-SCORE and root mean squared error (RMSE) for the training and testing phases.

Table 2. Comparison of models

	R2-SCORE		RMSE	
	Training	Testing	Training	Testing
LR	0.978	0.972	315482.584	347273.861
DTR (max_depth=4)	0.979	0.975	302616.601	324533.786
KNNR (K=8)	**0.985**	**0.978**	**80397.855**	**310842.101**

Best model:

For both metrics (i.e., R2-SCORE and RMSE) used to evaluate the three machine learning methods, the KNNR with parameter K=8 is the best model. It achieved R_SCORE of 0.978 as shown in Table 2. K in this case means that it uses 8 data points that are nearby the input of the concern.

Discussion:

The study was conducted successfully and objectives were achieved. They are some considerations in the findings. In the grid-connected electricity network; the preliminary analysis of the selected samples has indicated that there are few cases of the power outage and blackout electricity.

The blackout may occur but the end users (e.g. electricity seller's agents and other normal subscribers) cannot see its impact as it occurs in a very short time. Some industries/big companies, supplies of electricity, or other companies that have a direct concern to its provision may notice the blackout electricity cases. Most electricity users that are already connected are satisfied with the energy supply.

The three algorithms used to build the predictive model are all excellent with no overfitting scenarios. KNNR is the best in the comparison of the three algorithms used. For all the five years-data; March and September are the best seasons to produce a high value of active power output.

Solar irradiation is a very important input variable and it is proportional to the active power output. Electricity energy prediction is needed for other existing power plants, demands, and finding the correlation between energy produced and the demand; to support the national plan of electrical energy production.

References:

[1] R. Nyamvumba and M. Gakuba, "the Republic of Rwanda Energy, Water and Sanitation Ltd Expression of Interest for Scaling Up Renewable Energy Program (Srep) Financing for Energy Projects in Rwanda," no. April, 2014.

[2] S. Theocharides, G. Makrides, G. E. Georghiou, and A. Kyprianou, "Machine learning algorithms for photovoltaic system power output prediction," *2018 IEEE Int. Energy Conf. ENERGYCON 2018*, no. June, pp. 1–6, 2018, doi: 10.1109/ENERGYCON.2018.8398737.

[3] "C OVID-19 Investor Assessment for Ghana ,."

[4] J. Dzansi, S. L. Puller, B. Street, and B. Yebuah-Dwamena, "The Vicious Circle of Blackouts and Revenue Collection in Developing Economies: Evidence from Ghana *," no. October, 2018, [Online]. Available: https://brittanystreet.github.io/website/Dzansi_Puller_Street_Yebuah-Dwamena_Blackouts2018.pdf.

[5] S. Bimenyimana, G. N. O. Asemota, J. D. D. Niyonteze, C. Nsengimana, P. J. Ihirwe, and L. Li, "Photovoltaic solar technologies: Solution to affordable, sustainable, and reliable energy access for all in Rwanda," *Int. J. Photoenergy*, vol. 2019, 2019, doi: 10.1155/2019/5984206.

[6] V. K., "An Overview of Factors Affecting the Performance of Solar PV Systems," *Energy Scan*,

no. February, pp. 2–8, 2017, [Online]. Available: https://www.researchgate.net/publication/319165448.

[7] F. Rasool, M. Drieberg, N. Badruddin, B. Singh, and M. Singh, "Modeling of PV panels performance based on datasheet values for solar micro energy harvesting," *Int. Conf. Intell. Adv. Syst. ICIAS 2016*, 2017, doi: 10.1109/ICIAS.2016.7824072.

[8] T. Anuradha, P. D. Sundari, S. Padmanaban, P. Siano, and Z. Leonowicz, "Comparative analysis of common MPPT techniques for solar PV system with soft switched, interleaved isolated converter," *Conf. Proc. - 2017 17th IEEE Int. Conf. Environ. Electr. Eng. 2017 1st IEEE Ind. Commer. Power Syst. Eur. EEEIC / I CPS Eur. 2017*, no. November, 2017, doi: 10.1109/EEEIC.2017.7977885.

[9] M. Hlaili and H. Mechergui, "Comparison of Different MPPT Algorithms with a Proposed One Using a Power Estimator for Grid Connected PV Systems," *Int. J. Photoenergy*, vol. 2016, 2016, doi: 10.1155/2016/1728398.

[10] S. Angra and S. Ahuja, "Machine learning and its applications: A review," *Proc. 2017 Int. Conf. Big Data Anal. Comput. Intell. ICBDACI 2017*, no. March 2017, pp. 57–60, 2017, doi: 10.1109/ICBDACI.2017.8070809.

[11] S. Lamba, P. Saini, V. Kukreja, and B. Sharma, "Role of Mathematics in Machine Learning," *SSRN Electron. J.*, no. 04, pp. 2543–2548, 2021, doi: 10.2139/ssrn.3833931.

[12] J. A. Luft, S. Jeong, R. Idsardi, and G. Gardner, "Literature Reviews, Theoretical Frameworks, and Conceptual Frameworks: An Introduction for New Biology Education Researchers," *CBE Life Sci. Educ.*, vol. 21, no. 3, p. rm33, 2022, doi: 10.1187/cbe.21-05-0134.

[13] W. Y. Loh, "Classification and regression trees," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 1, no. 1, pp. 14–23, 2011, doi: 10.1002/widm.8.

[14] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A brief review of nearest neighbor algorithm for learning and classification," *2019 Int. Conf. Intell. Comput. Control Syst. ICCS 2019*, no. May, pp. 1255–1260, 2019, doi: 10.1109/ICCS45141.2019.9065747.

[15] B. Cunha, C. Droz, A. Zine, S. Foulard, and M. Ichchou, "A Review of Machine Learning Methods Applied to Structural Dynamics and Vibroacoustic," 2022, [Online]. Available: http://arxiv.org/abs/2204.06362.

[16] B. Sekeroglu, Y. K. Ever, K. Dimililer, and F. Al-Turjman, "Comparative Evaluation and Comprehensive Analysis of Machine Learning Models for Regression Problems," *Data Intell.*, vol. 4, no. 3, pp. 620–652, 2022, doi: 10.1162/dint_a_00155.

[17] V. Plevris, G. Solorzano, N. Bakas, and M. E. A. Ben Seghie, "Investigation of performance metrics in regression analysis and machine learning-based prediction models," *8th Eur. Congr. Comput. Methods Appl. Sci. Eng. 5–9 June 2022, Oslo, Norw.*, no. June, 2022, doi: 10.23967/eccomas.2022.155.