



UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

DOCTORADO EN CIENCIAS DE LA INGENIERÍA MECÁNICA ELÉCTRICA



TESIS

IMPUTATION OF MISSING DATA IN PHOTOVOLTAIC PANEL MONITORING SYSTEM

PRESENTADO POR:

SAUL HUAQUIPACO ENCINAS

PARA OPTAR EL GRADO ACADÉMICO DE:

DOCTOR EN CIENCIAS DE LA INGENIERÍA MECÁNICA ELÉCTRICA

PUNO, PERÚ

2022



UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

DOCTORADO EN CIENCIAS DE LA INGENIERÍA MECÁNICA ELÉCTRICA

TESIS

IMPUTATION OF MISSING DATA IN PHOTOVOLTAIC PANEL MONITORING SYSTEM

PRESENTADA POR:

SAUL HUAQUIPACO ENCINAS

PARA OPTAR EL GRADO ACADÉMICO DE:

DOCTOR EN CIENCIAS DE LA INGENIERÍA MECÁNICA ELÉCTRICA



APROBADA POR EL JURADO SIGUIENTE:

PRESIDENTE


.....
Dr. MATEO ALEJANDRO SALINAS MENA.

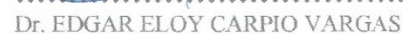
PRIMER MIEMBRO


.....
Dr. NORMAN JESUS BELTRAN CASTAÑON.

SEGUNDO MIEMBRO


.....
Dr. LEONIDAS VILCA CALLATA

ASESOR DE TESIS


.....
Dr. EDGAR ELOY CARPIO VARGAS

Puno, 10 de noviembre de 2022

ÁREA: Ciencias de la Ingeniería.

TEMA: Imputation of missing data in photovoltaic panel monitoring system.

LÍNEA: Mecánica Eléctrica.



DEDICATION

A ADRA y SAHR por estar a mi lado y ser el motivo



ACKNOWLEDGMENTS

A mis docentes del Doctorado en CIENCIAS DE LA INGENIERÍA MECÁNICA ELÉCTRICA por guiarme por la senda de la investigación.

Al Dr. Norman Jesús Beltrán Castañón por su incondicional y desinteresado apoyo.

Al Dr. José Emmanuel Cruz de la Cruz por su amistad y dirección en la senda de la investigación.

A mi asesor Dr. Edgar Eloy Carpio Vargas por el soporte brindado.

Al Consejo Nacional de Ciencia Tecnología e Innovación Tecnológica (CONCYTEC) y al Fondo Nacional de Desarrollo Científico, Tecnológico y de Innovación Tecnológica (FONDECYT)

“Este trabajo fue financiado por el CONCYTEC-FONDECYT en el marco de la convocatoria E041-01 [número de contrato N°180-2018-FONDECYT-BM-IADT-AV”.



TABLE OF CONTENTS

	Pág.
DEDICATION	i
ACKNOWLEDGMENT	ii
TABLE OF CONTENTS	iii
LIST OF TABLES	viii
LIST OF FIGURES	x
RESUMEN	xiii
ABSTRACT	xiv
INTRODUCTION	1

CHAPTER I

LITERATURE REVIEW

1.1. Theoretical Framework.....	3
1.1.1. Missing data imputation.....	3
1.1.2. Mean	3
1.1.3. Median	4
1.1.4. KNN.....	4
1.1.5. Frequency.....	5
1.1.6. International Electrotechnical Commission (IEC).....	5
1.1.7. IEC 60904-1:2020.....	6
1.1.8. IEC 61724-1:2021.....	6
1.1.9. Python.	7
1.1.10. Fog Computing	7
1.2. Research Background	8



CHAPTER II

STATEMENT OF THE PROBLEM

2.1. Significance of the Problem.....	16
2.2. Statement of the problem.....	16
2.3. Justification.....	17
2.4. Study objectives.....	18
2.4.1. General objective.....	18
2.4.2. Specific objectives.....	18
2.5. Hypotheses.....	18
2.5.1. General hypotheses.....	18
2.5.2. Specific hypotheses.....	18

CHAPTER III

MATERIALS AND METHODS

3.1. Place of study.....	19
3.2. Population.....	19
3.3. Research method.....	19
3.4. Detailed description of methods for specific objectives.....	19

CHAPTER IV

RESULTS

4.1. System description.....	21
4.1.1. Photovoltaic system (Solar Edge) with DC-DC converter.....	21
4.1.1.1. Photovoltaic panel.....	22
4.1.1.2. Single Phase Invertir SE3000H.....	23
4.1.1.3. Energy optimizer solar Edge P370.....	24
4.1.2. Photovoltaic system (String) with String inverter.....	25



4.1.2.1. Photovoltaic panel.....	27
4.1.2.2. String Inverter	27
4.2. Data acquisition system	29
4.2.1. IEC normative conditions	29
4.2.1.1. IEC 60904-1 normative conditions.....	29
4.2.1.2. IEC 61724 normative conditions	29
4.2.1.3. Monitoring System Classifications	31
4.2.1.4. Measured Parameters	32
4.2.1.5. Calculated Parameters.....	34
4.2.1.6. Traditional Performance Ratio.....	35
4.2.1.7. Temperature-Corrected Performance Ratios	35
4.2.2. Data acquisition system SFCR Solar Edge.....	35
4.2.2.1. Current and voltage transducers	37
4.2.2.2. Power meter	37
4.2.2.3. Micro plc logo 8.3.....	37
4.2.2.4. Rs 485 Modbus	37
4.2.3. Data acquisition system SFCR String.....	37
4.2.3.1. Current and voltage transducers	39
4.2.3.2. Power meter	39
4.2.3.3. Micro plc logo 8.3.....	39
4.2.3.4. RS 485 Modbus	39
4.3. Data Storage.....	39
4.3.1. FOG Computing	40
4.3.2. Server	41
4.3.3. LabVIEW	42
4.3.4. Graphical user interface GUI.....	42
4.4. Data processing.....	44



4.4.1. SFCR Solar Edge data processing	44
4.4.1.1. Data set SFCR Solar Edge	44
4.4.1.2. Data processing methodology SFCR Solar Edge	47
4.4.1.3. SFCR Solar Edge Score models comparison	62
4.4.1.4. SFCR Solar Edge MAE models comparison.....	63
4.4.1.5. SFCR Solar Edge MSE models comparison	64
4.4.1.6. SFCR Solar Edge determination coefficient models comparison	65
4.4.1.7. SFCR Solar Edge Adjusted determination coefficient models comparison	66
4.4.1.8. SFCR Solar Edge Training time models comparison.....	67
4.4.1.9. SFCR Solar Edge Test time models comparison	68
4.4.2. SFCR String data processing	69
4.4.2.1. Data set SFCR String	69
4.4.2.2. Data processing methodology SFCR String	71
4.4.2.3. SFCR String Score models comparison	86
4.4.2.4. SFCR String MAE models comparison	87
4.4.2.5. SFCR String MSE models comparison	88
4.4.2.6. SFCR String determination coefficient models comparison	89
4.4.2.7. SFCR String Adjusted determination coefficient models comparison.....	90
4.4.2.8. SFCR String Training time models comparison	91
4.4.2.9. SFCR String Test time models comparison	92
4.4.3. String VS Solar Edge data imputation models comparison.....	93
4.4.3.1. SFCR String vs Solar Edge Score models comparison	93
4.4.3.2. SFCR String vs Solar Edge MAE models comparison	94
4.4.3.3. SFCR String vs Solar Edge MSE models comparison.....	95
4.4.3.4. SFCR String vs Solar Edge determination coefficient models comparison	96



4.4.3.5. SFCR String vs Solar Edge Adjusted determination coefficient models comparison	97
4.4.3.6. SFCR String vs Solar Edge Training time models comparison	98
4.4.3.7. SFCR String vs Solar Edge Test time models comparison	99
CONCLUSIONS	100
BIBLIOGRAPHY	101



LIST OF TABLES

	Pág
Table 1 Photovoltaic panel Era Solar 370 data sheet.....	23
Table 2 Inverter data sheet SE3000H.	23
Table 3 Solar Edge P370 datasheet.....	25
Table 4 Talesun photovoltaic panel data sheet.	27
Table 5 Data sheet Inverter String SUNNY BOY.....	28
Table 6 Monitoring System Classifications.....	31
Table 7 Measured Parameters A.....	32
Table 8 Measured Parameters B.	33
Table 9 Calculated Parameters.	34
Table 10 Cloud computing vs Fog Computing.....	40
Table 11 Server features.	42
Table 12 Data set SFCR Solar Edge.....	44
Table 13 Data set with missing data SFCR Solar Edge.....	46
Table 14 SFCR Solar Edge: KNN=1001.....	49
Table 15 SFCR Solar Edge: KNN=101.....	51
Table 16 SFCR Solar Edge KNN=5.	53
Table 17 SFCR Solar Edge Mean.....	55
Table 18 SFCR SOLAR EDGE Median.....	57
Table 19 SFCR Solar Edge Frequent.....	59
Table 20 Data set completed example SFCR Solar Edge.....	61



Table 21 Data set SFCR String.....	69
Table 22 Data set with missing data SFCR String.....	70
Table 23 SFCR String KNN=1001.....	73
Table 24 SFCR String KNN=101.....	75
Table 25 SFCR String KNN=5.....	77
Table 26 SFCR String Mean.....	79
Table 27 SFCR String Median.....	81
Table 28 SFCR String Frequent.....	83
Table 29 Data set completed example SFCR String.....	85

LIST OF FIGURES

	Pág
Figure 1 SFCR installation with Solar Edge inverter and CC-CC optimizers.....	21
Figure 2 SFCR connection diagram with Solar Edge inverter with optimizers.	22
Figure 3 Solar Edge single phase inverter.	24
Figure 4 Solar Edge P370 DC-DC energy optimizer.	25
Figure 5 SFCR installation with String inverter.	26
Figure 6 Connection diagram SFCR with String inverter.	26
Figure 7 SUNNY BOY single-phase inverter.	28
Figure 8 Installation of the SFCR Solar Edge data acquisition system.....	36
Figure 9 SFCR Solar Edge data acquisition system diagram.	36
Figure 10 Installation of the SFCR String data acquisition system.....	38
Figure 11 SFCR String data acquisition system diagram.	38
Figure 12 Operating scheme of fog computing.	41
Figure 13 SFCR Solar Edge graphical user interface.	43
Figure 14 SFCR String graphical user interface.	43
Figure 15 Graphical scheme of missing data SFRC Solar Edge.	45
Figure 16 SFRC Solar Edge Data Amount.	45
Figure 17 Data processing methodology SFCR Solar Edge.....	47
Figure 18 Correlation of variables SFCR Solar Edge.....	48
Figure 19 SFCR Solar Edge Score models comparation.....	62
Figure 20 SFCR Solar Edge MAE models comparation.	63



Figure 21 SFCR Solar Edge MSE models comparison.....	64
Figure 22 SFCR Solar Edge determination coefficient models comparison.....	65
Figure 23 SFCR SolarEdge Adjusted deter. coefficient models comparison.	66
Figure 24 SFCR Solar Edge Training time models comparison.	67
Figure 25 SFCR Solar Edge Test time models comparison.	68
Figure 26 Graphical scheme of missing data SFRC String.	69
Figure 27 SFRC String Data Amount.....	70
Figure 28 Data processing methodology SFCR String.....	71
Figure 29 Correlation of variables SFCR String.	72
Figure 30 SFCR String Score models comparison.....	86
Figure 31 SFCR String MAE models comparison.	87
Figure 32 SFCR String MSE models comparison.....	88
Figure 33 SFCR String determination coefficient models comparison.....	89
Figure 34 SFCR String Adjusted determination coefficient models comparison.	90
Figure 35 SFCR String Training time models comparison.	91
Figure 36 SFCR String Test time models comparison.....	92
Figure 37 SFCR String vs Solar Edge Score models comparison.....	93
Figure 38 SFCR String vs Solar Edge MAE models comparison.	94
Figure 39 SFCR String vs Solar Edge MSE models comparison.	95
Figure 40 SFCR String vs Solar Edge deter. coefficient models comparison.....	96
Figure 41 SFCR String vs Solar Edge Adjusted deter. coefficient models comparison.	



Figure 42 SFCR String vs Solar Edge Training time models comparison..... 98

Figure 43 SFCR String vs Solar Edge Test time models comparison..... 99



RESUMEN

En la investigación científica la adquisición y procesamiento de datos tienen un rol fundamental, en los sistemas fotovoltaicos dada su naturaleza, este proceso presenta deficiencias por diversos factores como la dispersión de los módulos instalados, las condiciones climáticas o por la cantidad de información que se tienen que obtener, por lo que los procesos de adquisición, almacenamiento y procesamiento de datos son muy importantes. La presente investigación desarrolló un sistema de adquisición, almacenamiento y procesamiento de datos para sistemas fotovoltaicos, siguiendo la normativa europea IEC 60904 y IEC 61724 para la adquisición de datos, Fog Computing para el almacenamiento de la información y finalmente para el procesamiento se usó Aprendizaje Automático. Los resultados mostraron que el modelo basado en KNN obtuvo un SCORE de 99.08%, MAE de 25.3 y MSE de 93.16. Concluyendo que el Modelo basado en KNN es el más robusto para imputar datos en el monitoreo de sistemas fotovoltaicos.

Palabras clave: Imputación de datos, monitoreo de sistemas fotovoltaicos.



ABSTRACT:

In scientific research, data acquisition and processing play a fundamental role. In photovoltaic systems, given their nature, this process presents deficiencies due to various factors such as the dispersion of the installed modules, climatic conditions or the amount of information that must be obtained, so the processes of data acquisition, storage and processing are very important. The present research developed a data acquisition, storage and processing system for photovoltaic systems, following the European standards IEC 60904 and IEC 61724 for data acquisition, Fog Computing for information storage and finally Machine Learning was used for processing. The results showed that the KNN-based model obtained a SCORE of 99.08%, MAE of 25.3 and MSE of 93.16. Concluding that the KNN-based model is the most robust model for data imputation in PV system monitoring.

Keywords: Data imputation, photovoltaic monitoring system

INTRODUCCIÓN

In the age of energy for the public and private sectors of the countries, photovoltaic systems are becoming more prevalent and playing a significant role. For these systems to work efficiently, they must be turned into smart systems. For this, its operation can be monitored, taking not only the characteristics of the system, but also complementary ones such as irradiance and temperature. Therefore, the present work proposes the collection, storage and processing of information collected from photovoltaic systems connected to the grid in extreme conditions at more than 3800 meters above sea level. (Xu & Qiao, 2011) say this paper describes details of the design and instrumentation of smart photovoltaic modules, a wireless sensor network, and software for real-time sensing and control of a photovoltaic system with maximum power point tracking at module level. Field condition is monitored by voltage, current, irradiance, and temperature sensors distributed across the photovoltaic field. The sensory data are periodically sampled and transmitted to a base station. (Shariff et al., 2015) The traditional method is to gather the data and send it across wires. Given the price and technological restrictions of the wire, monitoring must frequently always be local to the plant being watched. It increases the system's capital and maintenance costs, which is another disadvantage. In this study, they created a wireless Zigbee monitoring system for photovoltaic installations connected to the grid. variables such as temperature and irradiation, PV power output and grid inverter power output are monitored. (Rezk et al., 2017) It exposes data acquisition systems (DAQS) as widely used in photovoltaic plants in order to evaluate the performance and then optimize it. The purpose of this research was to develop a cost-effective DAQS with Lab-VIEW. The developed monitoring system was used to continuously collect and monitor the electrical output parameters of a stand-alone PV system. Such parameters include; PV generated voltage, current and power. (Chouder et al., 2013) Using the LabVIEW real-time interface system, this study provides an in-depth analysis of the performance and dynamic behavior of solar systems. In order to make measurements and compare simulation results in real time, this program intends to integrate many measuring tools into a single system. The thorough monitoring and examination of PV systems is crucial. (Martínez-Cambolor, 2007) we study the problem of comparing the power of classification of different methods from the ROC Curve. On one hand, we propose a method based on the supremum measure and, on the other hand, we study the problem of comparing two or more ROC curves from the asymptotic properties of area under ROC curves (AUC). (Deng & Lumley, 2021) Multiple data imputation to deal with missing



data is gaining popularity over time. Although other multiple imputation approaches are well studied and have proven their validity, they have limitations when processing datasets with complex data structures. Their data imputation results depend on expert handling of the inherent relationships that exist between variables

CHAPTER I

LITERATURE REVIEW

1.1. Theoretical Framework

1.1.1. Missing data imputation

During the last decades, procedures have been developed that have better statistical properties than traditional options such as data elimination (listwise), observation matching (pairwise), the means method and hot deck. Multiple imputation (IM) algorithms can be applied using commercial and free access packages, but imputing information should not be understood as an end in itself. Its implications for secondary data analysis should be evaluated with caution, and this paper concludes that there is no ideal imputation method. Each situation is different, and the non-response rate and its spatial distribution change between surveys, so it is not advisable to adopt a priori the same imputation procedure for all variables, in all surveys. In the first part, the theory on which the imputation procedures used are based is analyzed, and in the second, eight alternative methods are applied to impute different income concepts for data from a household survey, and the sensitivity of the data is evaluated. poverty and inequality indices (Gini, Theil and Atkinson ($\epsilon = 2$)), to the imputation techniques used. It is shown that the poverty indices are sensitive to the imputation methods, while the information substitution procedure has less impact on the inequality indicators. (Medina & Galván, 2007).

1.1.2. Mean

The arithmetic mean of the observable values is used to replace the missing observation in the mean imputation method. Let $x(1), x(2), \dots, x(n)$ be the observed values of a dataset containing n observations, the estimate of a missing value, X_{mis} , using the mean imputation method is given by (Mohammed et al., 2021).

$$\begin{aligned}\hat{x}_{mis} &= \frac{1}{n_o} \sum_{i=1}^{n_o} x_i, \\ &= \frac{\sum_{i=1}^{n_o} x_i}{n_o},\end{aligned}$$

where n_o is the dataset's observed values' size. When the data is continuous normal, the mean imputation is suitable. One of the disadvantages of the mean imputation is that it undermines the overall variability in the data (Mohammed et al., 2021).

1.1.3. Median

In the interim, the middle ascription strategy replaces the lost esteem with the middle of the watched values in a dataset. Thus, using the median imputation the estimate of X_{mis} is (Mohammed et al., 2021).

$$\hat{x}_{mis} = \begin{cases} x_{(s)} & \text{if } n_o \text{ is odd} \\ \frac{x_{(s)} + x_{(s+1)}}{2} & \text{if } n_o \text{ is even} \end{cases},$$

where $X_{(s)}$ is the middle-observed value. The median imputation is suitable when the data is skewed or when outliers are present in a dataset.(Mohammed et al., 2021).

1.1.4. KNN

The missing observations are replaced with values from related records in the provided dataset using the k nearest neighbor imputation, or k-NN imputation. The distance function is typically used to determine the similarity. The Euclidean function is the most widely used distance function. The k-NN establishes a collection of k nearest neighbors and then replaces any missing observations for a given variable with the average of those of its neighbors. Thus, to estimate a missing observation, X_{mis} , using the k nearest neighbors imputation method, the algorithm is as follows:(Mohammed et al., 2021).

- Choose a suitable value of k, the number of nearest neighbors.
- Compute the distance between the missing observation on variable i and the other observed values using the Euclidean distance function given as

$$d(x_m, x_o) = \sqrt{\sum_{i=1}^n (x_{mi} - x_{oi})^2},$$

where X_{mi} is the value of variable i on the target observation, X_m , X_{oi} is the value of variable i on the other observed value, X_o , and $d(X_{mi}, X_{oi})$ is the distance between the target observation, X_m , and the observed value, X_o . (Mohammed et al., 2021).

- Choose k nearest observations,
- Calculate the weights of the k nearest values as.

$$w_j = \frac{1}{d(x, x_j^n)^2}.$$

- The estimate of the missing value is the weighted average of the k nearest neighbors which can be obtained as.

$$\hat{x}_{mis} = \frac{\sum_{j=1}^k w_j x_j^n}{W},$$

Where X_j^n , $j = 1, 2, \dots, k$ is the k nearest neighbors, w_j are weights of the k neighbors, and $W = \sum_{j=1}^k w_j$, the method is time consuming when the size of the data is large, and choosing the value of k is also difficult. (Mohammed et al., 2021).

1.1.5. Frequency

The frequency is a statistical measure that gives us information about the number of times an event is repeated when performing a certain number of occasional experiments. This measure is represented by the letters f_i . The letter f refers to the word frequency and the letter i refers to the i -th performance of the random experiment. (Mohammed et al., 2021)

1.1.6. International Electrotechnical Commission (IEC)

Our work helps to establish cheap infrastructure, access to efficient and sustainable energy, smart urbanization and transportation systems, reduce climate change, and improve environmental and human safety. IEC unites more than 170 nations and offers 20,000 professionals worldwide a global, impartial, and independent forum for

standardization. oversees four conformity evaluation systems whose participants attest that the equipment, facilities, services, and personnel are operating in accordance with specifications. With the help of conformity assessment and the over 10,000 IEC International Standards that EC publishes, governments can create a national quality infrastructure and businesses of all sizes may acquire and sell safe and dependable goods with confidence. constant in the majority of nations. IEC International Standards are utilized in testing and certification to confirm that the manufacturer's claims are being kept, and they serve as the foundation for risk and quality management. The 17 Sustainable Development Goals of the UN are directly supported by the activities of IEC. (IEC, 2021c).

1.1.7. IEC 60904-1:2020.

Measurement of photovoltaic current-voltage characteristics in photovoltaic devices, Part 1 Procedures for measuring the current-voltage characteristics (I-V curves) of photovoltaic (PV) devices in actual or simulated sunlight are outlined in IEC 60904-1:2020. These steps can be used with a single solar cell, a group of solar cells, or a PV module. With reference to (often but not only) the global reference spectral irradiance AM1.5 described in IEC 60904-3, this document is suitable to non-concentrating PV devices for usage in terrestrial situations. (IEC, 2021a)

1.1.8. IEC 61724-1:2021.

Performance of photovoltaic systems Part 1: Observation the International Standard IEC 61724-1:2021 contain, tools, and procedures for performance tracking and evaluation of photovoltaic (PV) systems are described in IEC 61724-1:2021. It also provides the foundation for other standards that rely on the information gathered. This paper provides guidelines for selecting monitoring systems and describes kinds of photovoltaic (PV) performance monitoring systems. The first edition, which was released in 2017, is canceled and replaced with this second version. In comparison to the previous edition, this edition features the following notable technological changes: (IEC, 2021b)

- Bifacial system monitoring is implemented.
- Updated irradiance sensor specifications.
- New technology-based soil measurement methods have been developed.

- Class C monitoring systems are taken out of service.
- A number of specifications, suggestions, and explanatory notes have been updated.

1.1.9. Python.

Python is an interpretative programming language that emphasizes objects. Classes, dynamic typing, highly high-level dynamic data types, exceptions, and modules are all covered. In addition to object-oriented programming, it supports a number of other programming paradigms, such as procedural and functional programming. Python offers a huge range of capabilities and a relatively straightforward syntax. It provides interfaces for numerous system calls, libraries, and window systems, and it may be modified in C or C++. It can also be utilized as an extension language for programs that demand a programmable interface. Finally, Python is portable: it runs on many Unix variants including Linux and macOS, and on Windows (Ren, 2021).

1.1.10. Fog Computing

Fog computing: This technology uses fog nodes, which are made up of routers, switches, and network gateways, to provide storage and compute capabilities. These device-equipped fog nodes are regarded as virtual nodes and contribute to the availability of network virtualization. This capability leads to the larger usage of fog processing in mobiles as well as in IoT devices (Anandakumar & Ramu, 2020).

1.2. Research Background

The various investigations carried out in relation to the proposed topic are presented below:

(Huaquipaco et al., 2022) Climatological factors influence the performance of grid-connected photovoltaic systems (PV). These factors vary according to the altitude above sea level. For this purpose, two PV of 3 kW each were installed, and their performance was measured under the IEC 62053 standard. Subsequently, the cross-validation of both models has performed, whose results showed that the DC-DC PV has a better result in 6.09% over the String PV model, so we conclude that the DC-DC Pv converter performs better at 3800 m.a.s.l.

(Huaquipaco et al., 2021) The current study suggests the collection, modeling, and prediction of a multivariate SFV utilizing a multiparametric regression model. Five regression models with machine learning are presented, three of which employ shrinkage regularization and two of which employ extreme gradient boosting. The test times are also always shorter. The results were validated so that they not only have mathematical significance, but are also real, showing that XGBoost with n estimators = 10 does not meet the five validation conditions, so this prediction model should not be considered.

(Killam et al., 2021) Understanding performance and degradation mechanisms is essential for enhancing overall dependability and lifespans, which in turn depends on accurate in-field characterization of photovoltaics. We make use of Suns-VOC, which is frequently used to measure individual solar cells in laboratories, and we talk about the challenges of adapting the method to outdoor systems. VOC, ideality factor, and pseudo fill factor all fall within 1% of the laboratory readings despite weather variations. The monitoring of the system's VOC at 0.05 to 0.1 suns, during periods of low power output, is also shown to offer a figure of merit that can reveal system damage at an early stage.

(Sun et al., 2020) A remote monitoring system for photovoltaic modules based on wireless sensor networks is intended to increase the management effectiveness of photovoltaic power plants because the working condition of photovoltaic modules cannot be tracked in real time and defective components cannot be specifically located and controlled. The most accurate time synchronization technique uses the Gaussian delay model, while the most effective reference broadcast synchronization approach uses the

least amount of energy during synchronization. The situation of the solar modules can be evaluated appropriately through the analysis and processing of the received data, allowing for the realization of photovoltaic power plant information management.

(Lazzaretti et al., 2020) Recent growth in the usage of photovoltaic energy is primarily attributable to new global regulations aimed at limiting the use of fossil fuels. In addition to being impacted by many types of faults, environmental factors have a significant impact on the efficiency of PV systems, which can result in significant energy loss during system operation. Additionally, utilizing the same MS, we suggest a recursive linear model that uses the irradiance and temperature of the PV panel as input signals and power as an output to detect system failures. For an Artificial Neural Network model, the accuracy of the classification stage is 95.44% using the same days and defects used in the detection module. The combined accuracy of detection and classification is 92.64%.

(Øgaard et al., 2020) In order to provide more dependable monitoring options for PV systems installed in these conditions, the goal of this effort is to reduce this instability. The metrics' fluctuation is reduced more drastically than with general low irradiance or clear sky filtering, and more data is kept in the relevant dataset. Comparisons of certain yield and performance indices based on machine learning modeling are the best performance indicators. The investigation identifies two ways to boost PV monitoring systems' reliability without spending more on hardware. First, choosing a proper performance metric will improve reliability. Second, rather than using the conventional literature thresholds, filters that explicitly target the source of the variability can be used to reduce the variability of the performance indicator.

(Xia et al., 2020) This research presents a revolutionary real-time monitoring method for photovoltaic generation. The remote monitoring of centralized or distributed solar systems is made possible by the Internet of Things when combined with cloud servers and terminal apps. The server then chooses three-phase current as the sample sets from the uploaded data to create a composite current characteristic combining wavelet packet energy and waveform parameter, and creates a fault diagnosis model based on the probabilistic neural network to assess the health status of the PV inverter online. This article outlines the construction of the user software at the application layer, the hardware design of the ZigBee and 4G modules, and composition of the diagnosis model for open circuit failure of the PV inverter through the cloud server.

(Huaquipaco et al., 2020) Not everyone has access to information through libraries or the Internet, especially in rural locations where it is doubtful that they will have access to energy networks or telecommunications infrastructure due to its dispersion. In Peru, 24.6% of the rural population still lacks access to electricity, according to the World Bank. This work suggests using photovoltaic energy to power a data server called the "Solar Library" that contains a wealth of information in book format, audios, videos, simulators, and more. Users can access this server via mobile devices wirelessly without installing any additional programs or applications. The implementation of a physical library is not very feasible due to high economic and logistic costs.

(Slapšak et al., 2019) A novel in situ measurement method has recently been created using tiny digital relative humidity sensors. The measurement method proved to be a useful tool for both long-term outdoor monitoring in the field as well as in situ monitoring of water concentration in solar modules exposed to accelerated test conditions in climatic chambers. In our concept, the RFID antenna and all necessary readout circuits are integrated with a 130 m thick polyimide foil onto which up to seven digital humidity sensors can be soldered. They can be positioned wherever in the PV module, either in front of or behind the solar cells, thanks to their incredibly small size and wireless design. The technologies were used in small modules that each included one full-size crystalline silicon solar cell.

(Sha et al., 2019) In this paper, a hot spot discovery and checking framework based on UAV is proposed. The framework employs UAV to carry warm imager and journey consequently to gather pictures and transmit them back to the ground station control framework. The real-time picture preparing of the ground station identifies hot spots and spares the assessment comes about to the nearby database, and can encode and yield the assessment video containing the hot spot area.

(Samara & Natsheh, 2019) A novel in situ measurement method has recently been created using tiny digital relative humidity sensors. The measurement method proved to be a useful tool for both long-term outdoor monitoring in the field as well as in situ monitoring of water concentration in solar modules exposed to accelerated test conditions in climatic chambers. In our concept, the RFID antenna and all necessary readout circuits are integrated with a 130 m thick polyimide foil onto which up to seven digital humidity sensors can be soldered. They can be positioned wherever in the PV module, either in

front of or behind the solar cells, thanks to their incredibly small size and wireless design. The technologies were used in small modules that each included one full-size crystalline silicon solar cell.

(Ma et al., 2019) Solar photovoltaic modules are used in building-integrated photovoltaic systems to generate electricity in place of conventional building materials. Photovoltaic modules cannot be monitored using traditional techniques since they are often mounted on building roofs or facades. The method suggested in this research differs from those from other studies in that it just calls for the current at the greatest power point. The suggested simulation of the current-voltage characteristic curves also obtains a reduced root mean square error value and demonstrates a superior capacity to reflect the current-voltage characteristics of the solar modules. The six characteristics of the solar module, which is employed in the building-integrated photovoltaic system, are also extracted using several ways.

(Krismadinata et al., 2019) A wireless monitoring system prototype is presented in this study. It was used to examine the electrical properties of two identically sized and typed solar panels. A copper pipe for water circulation is installed on the bottom side of one of the solar panels. For both solar modules, the experiment is carried out at the same time, location, and level of sunshine exposure. It is noted that the two solar modules exhibit the characteristics of open circuit voltage, short-circuit current, temperature on their top and bottom sides, and solar radiation. The gathered data is analyzed and visually presented. Wireless communication is used with the AT mega 8535 and PC. The measurements are verified against the instrument standard.

(Fazai et al., 2019) In this research, we take into account a machine learning strategy combined with statistical testing hypothesis for improved PV system defect detection performance. The presented approach uses a generalized likelihood ratio test (GLRT) chart to identify PV system problems and a Gaussian process regression (GPR) methodology as a modeling framework. Using both actual and simulated PV data, the proposed GPR-based GLRT technique is evaluated by tracking the critical PV system variables (current, voltage, and power). To assess the fault detection performance of the suggested technique, the calculation time, missed detection rate (MDR), and false alarm rate (FAR) are computed.

(Ortega et al., 2019) Power losses in photovoltaic systems account for around 15% to 20% of modern PV systems' performance ratio. PV module failures can occur for a number of causes, and because they are connected in series with the other modules in the string, a failure in one module may cause losses in the entire string. The only way to identify these errors is through routine monitoring. Individual module failures cannot be found using monitoring approaches that are focused on groups of modules. This paper suggests a method for measuring individual PV modules partially and recomposing their attributes using just tiny capacitors with capacitances in the range of tens of microfarads and without power electronics components.

(Pazhoohesh et al., 2019) An essential part of studying energy and buildings is data collecting. A building's energy consumption modeling and other control and management systems might be significantly impacted by errors and inconsistent data gathered from test environments. It gives a comparison of eight techniques for adding missing values to sensor data construction. The data set utilized in this study is made up of actual information that was gathered from our test subject, a living laboratory at Newcastle University. This method has been performed 1000 times in order to get more precise and reliable findings, and the average of 1000 simulations is presented in this work.

(Khan et al., 2019) Photovoltaic (PV) cell usually shows unpredictable results due to abrupt change in environment. Environmental factors including temperature, dust content, irradiance, and air mass directly affect how well PV cells perform. Monitoring PV and the associated environmental variables is crucial for comprehending and analyzing PV in real-world circumstances. According to reports, planned PV monitoring systems are now costly, complicated, and have a few numbers of applications, only a select handful of which are wireless. In this piece, we create and put into practice a basic, affordable, wireless PV monitoring solution. A real-time analysis and data logging of open circuit voltage, short circuit current, ambient temperature, cell temperature, and maximum power point are carried out by the proposed PVMS.

(Harrou et al., 2018) In this study, the creation of a ground-breaking fault detection and diagnosis system for monitoring solar systems' direct current side is reported. To achieve this, we suggest a statistical method that improves fault detection by combining the benefits of the one-diode model with those of the univariate and multivariate exponentially weighted moving average charts. These residuals, which are used as failure

indicators, record the discrepancy between measurements and MPP predictions for current, voltage, and power from the one-diode model. When a defect is found in the MEWMA chart, the type of fault is determined using the univariate EWMA chart based on current and voltage indicators. Using actual data from the grid-connected PV system built at the Renewable Energy Development, we tested this approach. Results show the capacity of the proposed strategy to monitor the DC side of PV systems and detect partial shading.

(Sabry et al., 2018) Photovoltaic (PV) parameters monitoring is very important for the implementation and optimum utilization of solar energy as electricity source. The design of a straightforward, reasonably priced, and low-consumption wireless PV monitoring system is proposed in this study, together with a driving program for logging the parameters of the PV system. Four different types of sensors are used in the circuit to handle four characteristics that are crucial for the real-time study and prediction of PV performance. Only a single pair of XBee RF modules is an active component; all other components, such as resistors and capacitors for PV current, voltage regulators for signal conditioning, temperature sensors, and irradiance sensors, are passive. The proposed system succeeds in providing real-time monitoring with lower cost and can be extended for more functions such as controlling tracking system and failure diagnosis

(Beránek et al., 2018) A cutting-edge technology for tracking the sun has been created. The system was designed to gather, analyze, and process data while measuring the key parameters and features of solar plants. A specialized data recorder called the BB box is deployed at the producing plants. The new monitoring system has tracked 65 solar power facilities totaling 175 MWp in the Czech Republic and other countries. The power generated by the constructed PV plants corresponds to the predictions made by the widely used program PVGIS throughout the preceding seven years of operation.

(CHASE, 2018) In the framework of environmental science and technology, this thesis introduces the PLACOT2AM in situ sensing platform. Through internet of things technology, the platform is an integrated system with sensors, data collecting, processing, and wireless communication. The platform contains an expert system that enables the analysis of environmental variable data collection while it is being collected. This enables the generation of knowledge to support society's decision-making about the extremes of environmental variability that might bring about health or productivity. In the second

iteration, the platform monitors super-extreme sun irradiance at an altitude and latitude below sea level in Belém, of 1321 W/m².

(Rezk et al., 2017) PV facilities frequently use data acquisition systems to gather all system data for purposes of analyzing and optimizing plant performance. The development of a cost-effective DAQS based on Lab-VIEW is the major goal of this endeavor. Additionally, it enables the sketching of PV panel properties under actual test conditions. Additionally, since the short circuit current of PV modules is directly proportional to the solar concentration, it is possible to measure the total amount of solar radiation by monitoring it. The suggested approach is regarded as a good way to get the system data necessary for performance analysis and improvement of PV plants. Field tests have shown that the monitoring data acquired are quite good. The computer screen's depiction of the web material is incredibly educational.

(Madeti & Singh, 2017) A PV plant's performance is monitored and/or evaluated using the photovoltaic monitoring system, which gathers and examines a variety of metrics being observed there. An efficient monitoring system is necessary for any PV system's dependable and steady functioning. Existing PV monitoring systems can only be used in large-scale PV projects due to their high cost and complexity. Numerous pieces of literature have reported on various elements of PV monitoring systems during the past ten years. This contains a thorough rundown of all the main PV monitoring assessment methodologies and how well they perform in comparison. Sensors and their operating principles, controllers utilized in data collecting systems, data transmission techniques, and data storage and analysis are the main features of PV monitoring systems that this study investigates.

(Yahyaoui & Segatto, 2017) For grid-connected solar plants, where every kilowatt-hour is essential since only kilowatt-hours that are fed into the grid are paid for, improving reliability and performance of photovoltaic plants are significant goals that boost the competitiveness of PV systems. For a PV plant linked to a single-phase grid, this research article uses two current and voltage indicators to assess and discriminate, in real-time, the faults associated to bypassed PV modules, open-circuit strings, and partial shade. The usefulness of the suggested strategy was demonstrated by studies testing the efficacy of these indicators utilizing a Control and Data Acquisition System.



(Mekki et al., 2016) This work introduces a defect detection approach for solar modules operating in partially shadowed environments. To estimate the output photovoltaic current and voltage under varying operating circumstances, an artificial neural network is used. Used were the measured data from Jijel University's Renewable Energy Laboratory (REL). Comparing the predicted current and voltage to the observed values reveals important details about the operation of the solar module under consideration. The efficiency of the suggested strategy has been demonstrated via the investigation of various shading patterns. The findings demonstrated that the developed approach properly determines the impact of shade on the solar module.

CHAPTER II

STATEMENT OF THE PROBLEM

2.1. Significance of the Problem

The quality of data at the time of developing research is fundamental, so the processes of acquisition, storage and processing of data are very important, currently there are methods to collect, save and process data, but they have deficiencies such as coverage, proprietary systems and reliability. Therefore, in the present work the following is questioned:

To what extent will the use of a reliable data acquisition, storage and processing system with data imputation techniques affect the monitoring of a photovoltaic system?

2.2. Statement of the problem

To what extent will the use of a reliable data acquisition, storage and processing system with data imputation techniques affect the monitoring of a photovoltaic system?

This question can be broken down into the following:

- To what extent will the use of a reliable data acquisition system affect the monitoring of a photovoltaic system?
- To what extent will the use of a reliable storage system affect the monitoring of a photovoltaic system?
- To what extent will the use of a data processing system with data imputation techniques affect the monitoring of a photovoltaic system?

2.3. Justification

In scientific research, the acquisition and processing of data have a fundamental role, in photovoltaic systems, given their nature, this process presents deficiencies due to various factors such as the dispersion of the installed modules, the climatic conditions and the amount of data that must be acquire, this work aims to ensure this acquisition and processing of data in order to have a better quality of data with which scientists can continue developing research.

This research arose from the need to pre-process the data obtained from the photovoltaic systems in order to apply machine learning techniques to carry out publishable research.

The research seeks to provide information that will be useful to the entire scientific community, especially in the area of solar photovoltaic energy, to improve knowledge about the scope of the problem in the sector and to be able to propose strategies to address it.

Since there are not enough local and national studies on the proposed topic and its consequences, the present work is convenient to consolidate a better knowledge on the methods of collection, processing and imputation of data from photovoltaic systems.

The present work has a methodological usefulness, since future research could be carried out using compatible methodologies, so that joint analyses, comparisons between specific time periods and evaluations of the interventions being carried out for the prevention of problems associated with the variables of this research would be possible.

In scientific research, the acquisition and processing of data have a fundamental role, in photovoltaic systems, given their nature, this process presents deficiencies due to various factors such as the dispersion of the installed modules, the climatic conditions and the amount of data that must be acquire, this work aims to ensure this acquisition and processing of data in order to have a better quality of data with which scientists can continue developing research.

2.4. Study objectives

2.4.1. General objective

Develop a reliable data acquisition, storage and processing system for photovoltaic systems using artificial intelligence for data imputation.

2.4.2. Specific objectives

- Acquire data from photovoltaic systems.
- Store data obtained from photovoltaic systems
- Process the data using artificial intelligence techniques for data imputation

2.5. Hypotheses

2.5.1. General hypotheses

A reliable system affects data acquisition, storage and processing with data imputation techniques in the monitoring of a photovoltaic system

2.5.2. Specific hypotheses

- A reliable system affects data acquisition in the monitoring of a photovoltaic system.
- A reliable system affects data storage in the monitoring of a photovoltaic system.
- A reliable system affects data processing with data imputation techniques in the monitoring of a photovoltaic system.

CHAPTER III

MATERIALS AND METHODS

3.1. Place of study

This research is developed in the city of Juliaca at 3800 m.a.s.l. At $15^{\circ} 24'40.7'' S$ $70^{\circ} 05'35.7'' W$ in the department of Puno in Peru, this area is particularly important for the development of research in photovoltaic systems because it has high solar radiation.

3.2. Population

Photovoltaic solar park of 6000 Watts located on the university campus of the Universidad Nacional de Juliaca located in the district of San Miguel, province of San Román department of Puno

3.3. Research method

- Applied: because knowledge and theories or basic research is used to solve an existing problem.
- Quantitative: because it generates data or numerical information that can be worked on in a statistical way.

3.4. Detailed description of methods for specific objectives

- a) Detailed description of the use of materials, equipment, supplies, among others



- Computer
- Internet
- Server.
- Power meter.
- Solar power plant.
- Printer
- Stationery.

b) **Description of variables to be analyzed in the specific objectives**

- Acquisition and monitoring of photovoltaic system data
- Imputation of data in photovoltaic system

c) **Inferential statistical test application**

Imputation of data, KNN, Mean, Median and Frequent as Models

CHAPTER IV

RESULTS

4.1. System description.

The research project consists of the implementation of photovoltaic systems connected to the grid (SFCR). These are photovoltaic systems connected to the grid, with a DC-DC converter and with a String inverter.

4.1.1. Photovoltaic system (Solar Edge) with DC-DC converter

The SFCR with DC-DC converter has 10 photovoltaic modules of 370 Wp each, 10 DC-DC converters for each module and a 3KW single-phase inverter for the entire system, as main elements. For the validation tests of the SFCR with a DC-DC converter, the photovoltaic modules were identified and the parameter tests performed on each one of them, this system is shown in Figure 1 and 2.



FIGURE 1
SFCR installation with Solar Edge inverter and CC-CC optimizers.

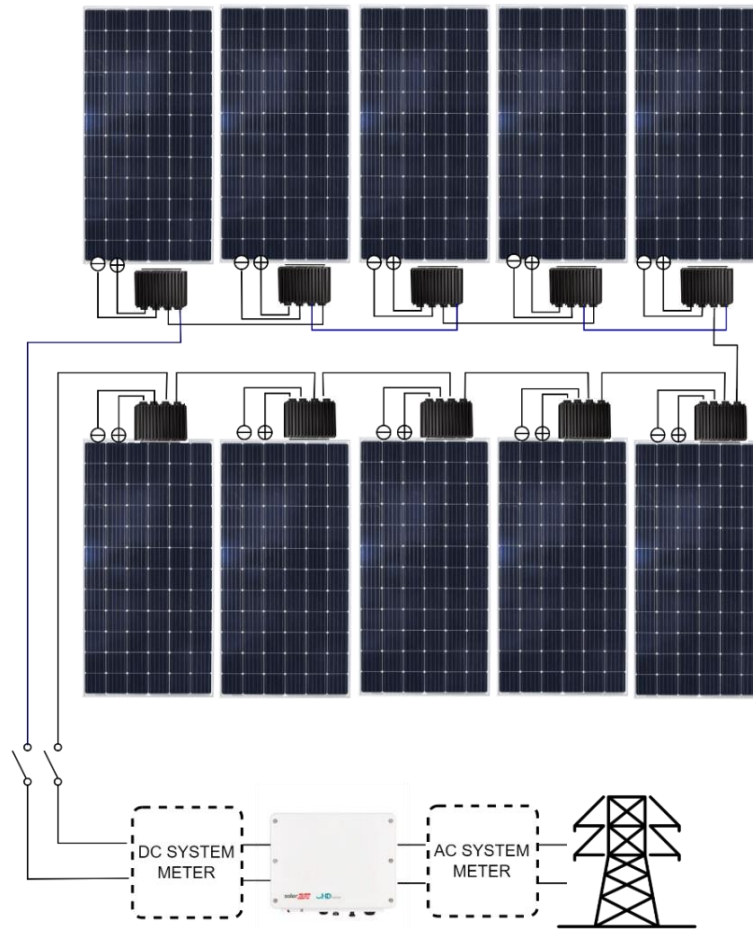


Figure 2
SFCR connection diagram with Solar Edge inverter with optimizers.

4.1.1.1. Photovoltaic panel.

The photovoltaic modules are of the 370Wp monocrystalline type, from the ERA SOLAR brand, with the ESPSC370 model see table 1, with an output tolerance of $\pm 3\%$. The operating temperature in the range of -40°C to $+85^{\circ}\text{C}$. Standard test conditions $1000\text{W}/\text{m}^2$, AM1.5; 25°C

Table 1
Photovoltaic panel Era Solar 370 data sheet.

ERA SOLAR 370	
Maximum Power (Pmax)	370W
Power Tolerance	+3%
Open Circuit Voltage (Voc)	48.3 V
Short Circuit Current (Isc)	9.95 A
Operating Voltage (Vmpp)	40.1 V
Operating Current (Impp)	9.23 A
Maximum System Voltage	1000 VDC
Nominal Operating Cell Temperature (NOCT)	45±2 C°
Module Weight	21.5 Kg
Module Dimensions	1956X992X35 mm
Application Class	Class A

4.1.1.2. Single Phase Inverter SE3000H

The SolarEdge SE3000H HD-Wave 3000W inverter see figure 3, is a grid connection inverter that allows you to get the most out of each solar panel individually, working together with the SolarEdge optimizers see table 2.

Table 2
Inverter data sheet SE3000H.

Single phase inverter SE3000H	
Operating Temperature Range	-40° to +60°
Compatible communication interfaces	RS485, Ethernet, Wi-Fi (optional), Input
Maximum DC power	4650W
Input voltage	480Vdc
Nominal DC input voltage	380Vdc
Input current	9Adc
Maximum inverter return	99.2%
Weighted European performance	98.8%
Power consumption at night	<2.5

	Output
Nominal AC output power	3000VA
Maximum AC output power	3000VA
AC output voltage (nominal)	220/230 Vac
AC output voltage range	184-264.5 Vac
AC frequency (nominal)	50/60 \pm 5
Maximum continuous output current	14 ^a
Total Harmonic Distortion (THD)	<3%
Power factor	1 adjustable -0.9 to 0.9



Figure 3

Solar Edge single phase inverter. Adapted from (SolarEdge, 2021) CCBy 2.0

4.1.1.3. Energy optimizer solar Edge P370

The SolarEdge P370 Optimizer see figure 4, is a necessary element in installations with a SolarEdge inverter. An optimizer must be incorporated for each solar panel that the series that we connect to the inverter has. This P370 model is suitable for panels of 60 or 72 cells and supports a power of up to 370W see table 3.

Table 3
Solar Edge P370 datasheet.

Power Optimizer	
Rated Input DC Power	370W
Absolute Maximum Input Voltage	60Vdc
MPPT Operating Range	8-60Vdc
Maximum Short Circuit Current (Isc)	11A _{dc}
Maximum DC Input Current	13.75A _{dc}
Maximum Efficiency	99.5%
Weighted Efficiency	98.8%
Maximum Output Current	15A _{dc}
Maximum Output Voltage	60Vdc
Maximum Allowed System Voltage	1000Vdc
Operating Temperature Range	-40C° - +85C°



Figure 4
Solar Edge P370 DC-DC energy optimizer. Adapted from (SolarEdge, 2021),CCBy 2.0

4.1.2. Photovoltaic system (String) with String inverter

The SFCR with a single-phase String-type inverter has 12 polycrystalline photovoltaic modules of 270 W_p, making a total of 3.24 W_p of photovoltaic generator power with a 3-kW String-type inverter. The system diagram is shown in Figure 5 and 6.



Figure 5
SFCR installation with String inverter.

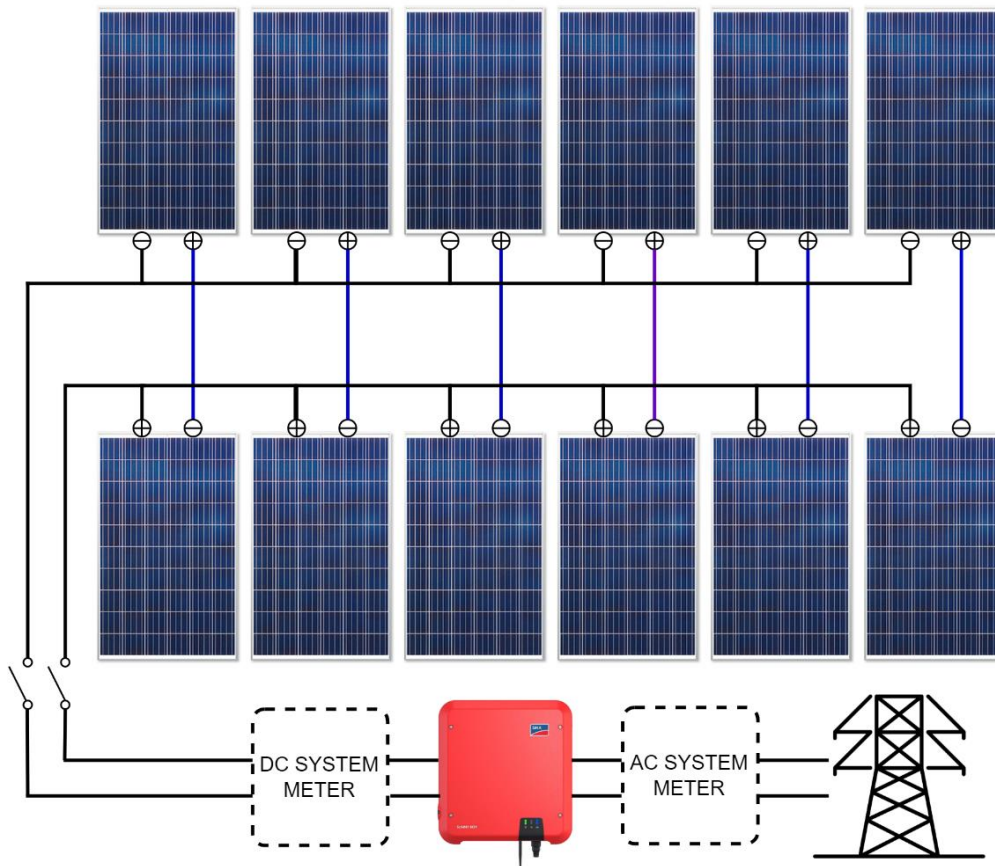


Figure 6
Connection diagram SFCR with String inverter.

4.1.2.1. Photovoltaic panel.

The photovoltaic modules are of the 270Wp polycrystalline type, brand TALESUN with model TP660P-270, with an output tolerance of $\pm 3\%$. Normal operating temperature in the range of $45\text{ }^{\circ}\text{C} \pm 2\text{ }^{\circ}\text{C}$ see table 4.

Table 4

Talesun photovoltaic panel data sheet.

Talesun Photovoltaic Module TP660P-270	
Maximum Power (Pmax)	270W
Power Tolerance	+3%
Open Circuit Voltage (Voc)	38.5V
Short Circuit Current (Isc)	9.09A
Operating Voltage (Vmpp)	31.3V
Operating Current (Impp)	8.63A
Maximum Series Fuse	20A
Maximum System Voltage	1000VDC
Nominal Operating Cell Temperature (NOCT)	$45\pm 2\text{ }^{\circ}\text{C}$
Module Weight	18.5Kg
Module Dimensions	1650X992X35mm
Application Class	Class A
STC: 1000W/m ² , AM1.5;25C°	

4.1.2.2. String Inverter

The Sunny Boy see figure 7, is a transformer less inverter, responsible for receiving the direct current generated by the panels and transforming it into single-phase alternating current that is used directly for consumption see table 5.

Table 5
Data sheet Inverter String SUNNY BOY.

Single phase inverter SUNNY BOY 3.0	
Operating Temperature Range	-25 °C to +60 °C
Compatible communication	Modbus (SMA, Sunspec),
Maximum inverter return	97%
Weighted European performance	96.4
Power consumption at night	5W
Input	
Maximum DC power	5500 Wp
Input voltage	600 V
MPP voltage range	110 V to 500 V
Input current	15A
Output	
Nominal AC output power	3000W
Power max. apparent AC	3000VA
AC output voltage (nominal)	220/230 Vac
AC output voltage range	184-264.5 Vac
AC frequency (nominal)	50/60 ± 5
Maximum continuous output current	16A
Power factor	1



Figure 7
SUNNY BOY single-phase inverter. Adapted from (AG, 2022), CCBy 2.0

4.2. Data acquisition system

The data acquisition of the photovoltaic systems was carried out in accordance with international standards IEC 60904-1 "Measurement of Photovoltaic Current-Voltage Characteristics" and the IEC 61724 "Photovoltaic System Performance Monitoring" Guidelines for Measurement, Data Exchange and Analysis"

4.2.1. IEC normative conditions

4.2.1.1. IEC 60904-1 normative conditions

The standard IEC 60904-1 "Photovoltaic devices Measurement of Photovoltaic Current-Voltage Characteristics for non-simulated environments has the following characteristics that were taken into account for the realization of this research (IEC, 2021a)

- Voltage and Current measured with ± 0.2 % Uncertainty in V_{oc} and I_{sc}
- Test and Reference device coplanar within $\pm 2^\circ$ and normal to the sun within $\pm 5^\circ$.
- Reference cell and test device temperature measured with ± 1 °C uncertainty.
- Spectral mismatch error correction if matched reference cell not used (IEC 60904).
- If reference cell > 2 °C from calibration temperature corrections applied.

4.2.1.2. IEC 61724 normative conditions

In carrying out this research, the recommendations stipulated in the IEC 61724 standard "Photovoltaic System Performance Monitoring Guidelines for measurement data exchange and analysis" were applied considering the following items (IEC, 2021b):

Voltage and Current

- AC and or DC uncertainty including Instrumentation $< 1\%$ of reading

Power

- DC Calculated based upon instantaneous and not averaged readings or directly measured with wattmeter

- AC power accounts for power factor and harmonic distortion
- Uncertainty including Instrumentation < 2%

Appendix A

Data Acquisition - Irradiance, Temperature, Voltage, Current, Power (IEC, 2021b)

Sampling

- Parameters which vary directly with irradiance shall be sampled with 1 min or less interval. Parameters with larger time constants, an arbitrary interval may be specified between 1 min and 15 min. Special consideration for increasing the sampling frequency shall be given to any parameters which may change quickly as a function of system load.

Linearity

- The difference between measured and applied signal < ± 1 % of full-scale value at of 0, 20, 40, 60, 80, and 100 % of full scale.

Stability

- 100 % full scale dc signal applied for 6 h. Should the fluctuation of the input signal exceed ± 0.2 %, the results shall be compensated by using a voltmeter with an accuracy better than ± 0.2 %.

Integration

- Apply expected maximum sensor signal for > 6 h. Signal times interval within 1% of expected value. With shorted input integral less than 1% of maximum signal applied times interval.

4.2.1.3. Monitoring System Classifications

Monitoring level classification system according to IEC 61724 see table 6.

Table 6
Monitoring System Classifications.

	Class A	Class B	Class C
Description	Greatest precision	Medium- level precision	Basic precision
Typically, targeted PV system size	Utility- scale	Commercial scale	Residential and small commercial
Suitable applications			
System performance assessment	X	X	X
Documentation of a performance guarantee	X	X	
Forecasting performance	X	X	
Electricity network interaction assessment	X	X	
Monitoring integration of distributed generation, storage, & loads	X	X	
System losses analysis	X		
PV technology assessment	X		
PV system degradation measurement	X		

Note. Adapted from (IEC, 2021b)

4.2.1.4. Measured Parameters

Measured parameters classification system according to IEC 61724 see table 7 and 8.

Table 7
Measured Parameters A.

Category	Parameter	Symbol	Units	Required?		
				Class A	Class B	Class C
Irradiance	In-plane irradiance	G_i	$W \cdot m^{-2}$	√	√	√
	In-plane direct beam irradiance	$G_{i,b}$	$W \cdot m^{-2}$	for concentrator systems	for concentrator systems	for concentrator systems
	In-plane diffuse irradiance	$G_{i,d}$	$W \cdot m^{-2}$	for concentrator systems	For concentrator systems	
	Global horizontal irradiance	G_G	$W \cdot m^{-2}$	√		
	Diffuse horizontal irradiance	G_d	$W \cdot m^{-2}$			
Environmental Factors	Ambient air temperature	T_{amb}	°C	√	√	√
	PV module temperature	T_{mod}	°C	√	√	
	Soiling ratio	SR		√		
	Wind speed	WS	$m \cdot s^{-1}$	√	√	
	Wind direction	WD	degrees	√		

Note. Adapted from (IEC, 2021b).

Table 8
Measured Parameters B.

Category	Parameter	Symbol	Units	Required?		
				Class A	Class B	Class C
PV array output	PV array output voltage (DC)	V_A	V	X		
	PV array output current (DC)	I_A	A	X		
	PV array output power (DC)	P_A	kW	X	X	
Inverter output	Inverter output voltage (AC)	V_{inv}	V	X		
	Inverter output current (AC)	I_{inv}	A	X		
	Inverter output power (AC)	P_{inv}	kVA	X	X	X
	Inverter output power factor	λ_{inv}		X		
System output	Output voltage (AC)	V_{out}	V	X		
	Output current (AC)	I_{out}	A	X		
	Output power (AC)	P_{out}	kVAr	X	X	X
	System power factor			X		

Note. Adapted from (IEC, 2021b).

4.2.1.5. Calculated Parameters

Calculated parameters according to IEC 61724 see table 9.

Table 9
Calculated Parameters.

Parameter	Symbol	Unit
Irradiation		
In-plane irradiation	H_i	$\text{kWh}\cdot\text{m}^{-2}$
Electrical energy		
PV array output energy	E_A	kWh
Inverter output energy	E_{inv}	kWh
Energy output from pv system	E_{out}	kWh
Array power rating		
Array power rating (DC)	P_0	kWp
Yields and yield losses		
PV array energy yield	Y_A	$\text{kWh}\cdot\text{kWp}^{-1}$
Final system yield	Y_f	$\text{kWh}\cdot\text{kWp}^{-1}$
Reference yield	Y_r	$\text{kWh}\cdot\text{kWp}^{-1}$
Array capture loss	L_C	$\text{kWh}\cdot\text{kWp}^{-1}$
Balance of system (BOS) loss	L_{BOS}	$\text{kWh}\cdot\text{kWp}^{-1}$
Efficiencies		
Array efficiency	η_A	None
System efficiency	η_f	None
BOS efficiency	η_{BOS}	None

Note. Adapted from (IEC, 2021b), 2021

4.2.1.6. Traditional Performance Ratio

- Indicates the overall effect of losses on the system output
- Quotient of the system's final yield Y_f to its reference yield Y_r

$$\begin{aligned}
 PR &= Y_f / Y_r \\
 &= (E_{out} / P_0) / (H_i / G_{i,ref}) \\
 &= \left(\sum_k \frac{P_{out,k} \times \tau_k}{P_0} \right) / \left(\sum_k \frac{G_{i,k} \times \tau_k}{G_{i,ref}} \right) \quad \text{Units = h / h}
 \end{aligned}$$

Moving P_0 to the denominator sum expresses both numerator and denominator in units of energy:

$$PR = \left(\sum_k P_{out,k} \times \tau_k \right) / \left(\sum_k \frac{P_0 \times G_{i,k} \times \tau_k}{G_{i,ref}} \right) \quad \text{Units = kW-h / kW-h}$$

- Traditional P_R neglects array temperature, resulting in seasonal variation when calculated for time periods less than one year.

4.2.1.7. Temperature-Corrected Performance Ratios

Seasonal variation of the traditional P_R is removed by calculating a temperature-corrected performance ratio:

$$\begin{aligned}
 PR' &= \left(\sum_k P_{out,k} \times \tau_k \right) / \left(\sum_k \frac{(C_k \times P_0) \times G_{i,k} \times \tau_k}{G_{i,ref}} \right) \\
 C_k &= 1 + \gamma \times (T_{mod,k} - T_{ref}) \quad \text{Temp. correction to power}
 \end{aligned}$$

Using 25 °C as T_{ref} gives PR'_{STC} .

4.2.2. Data acquisition system SFCR Solar Edge

This is one of the investigated photovoltaic systems consisting of an arrangement of 10 monocrystalline photovoltaic panels out of 370 with energy optimizers CC-CC and a single-phase inverter, see figure 8 and 9.



Figure 8
Installation of the SFCR Solar Edge data acquisition system.

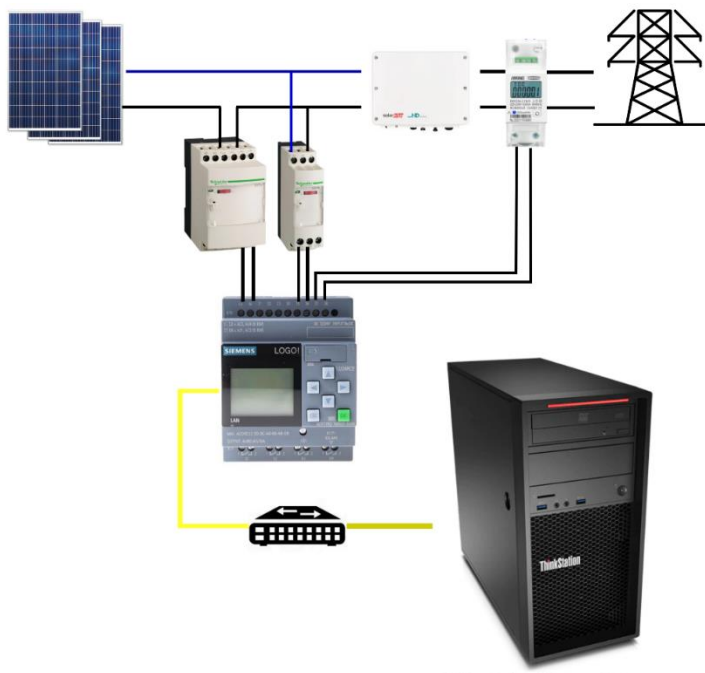


Figure 9
SFCR Solar Edge data acquisition system diagram.

4.2.2.1. Current and voltage transducers

SCHNEIDER brand ZELIO ANALOG converters in charge of measuring the voltage and current of the system between the solar panels and the inverter, information that is then sent to the PLC.

4.2.2.2. Power meter

HIKING TOMZN Meter of current, voltage, active power, reactive power, COS, frequency Total energy in positive and inverse KW/h with Standard IEC 62053-21 (IEC61036) MODBUS- RS 485 Service temperature limit: $-25^{\circ} + 70^{\circ}$ Accuracy Class 1.

4.2.2.3. Micro plc logo 8.3

SIEMENS LOGO PLC with 12 / 24RCE, logic module, PS / I / O device: 12 / 24VDC / relay, 8 DI (4AI) / 4DO, 400 block memory, expandable modular, Ethernet, integra. web server, data logger, standard microSD card for LOGO Soft Comfort V8

4.2.2.4. Rs 485 Modbus

It is a standard interface of the physical layer of communication, a method of signal transmission, the 1st level of the OSI (Open Systems Interconnection) model, transmitters and receivers exchange data through a twisted pair cable of rigid wires of 22 or 24 AWG, the maximum cable length used in RS-485 communications is 1200 meters at 100 Kbps with up to 247 peripherals

4.2.3. Data acquisition system SFCR String

This is the other of the investigated photovoltaic systems consists of an arrangement of 12 polycrystalline photovoltaic panels of 270W with a single-phase String inverter, see figure 10 and 11.



Figure 10
Installation of the SFCR String data acquisition system.

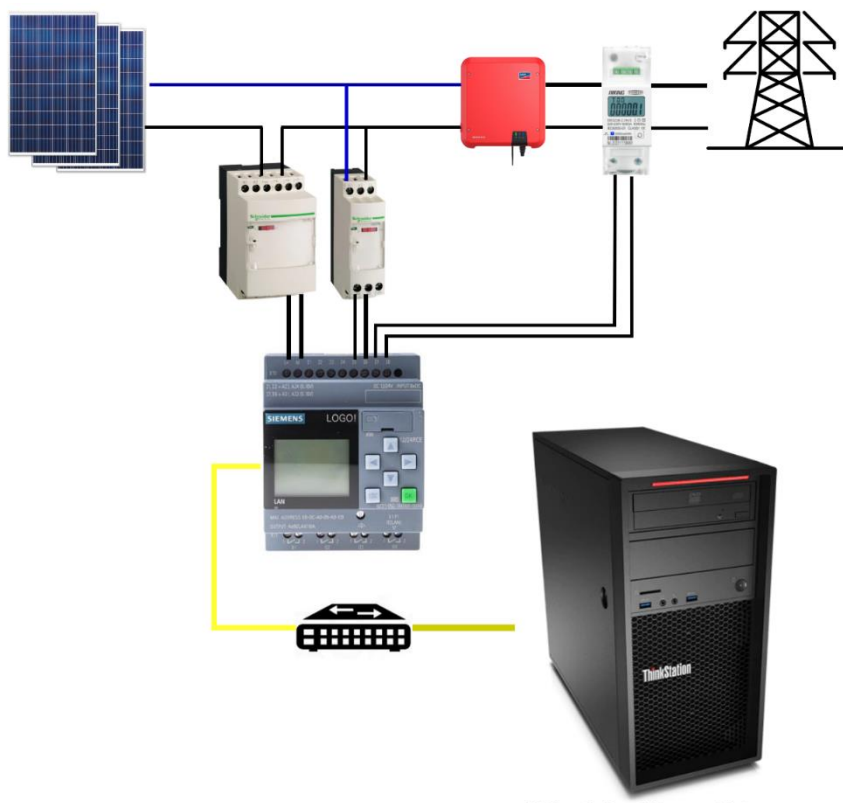


Figure 11
SFCR String data acquisition system diagram.

4.2.3.1. Current and voltage transducers

SCHNEIDER brand ZELIO ANALOG converters in charge of measuring the voltage and current of the system between the solar panels and the inverter, information that is then sent to the PLC.

4.2.3.2. Power meter

HIKING TOMZN Meter of current, voltage, active power, reactive power, COS, frequency Total energy in positive and inverse KW/h with Standard IEC 62053-21 (IEC61036) MODBUS- RS485 Service temperature limit: -25° + 70° Accuracy Class 1.

4.2.3.3. Micro plc logo 8.3

SIEMENS LOGO PLC with 12 / 24RCE, logic module, PS / I / O device: 12 / 24VDC / relay, 8 DI (4AI) / 4DO, 400 block memory, expandable modular, Ethernet, integra. web server, data logger, standard microSD card for LOGO Soft Comfort V8

4.2.3.4. RS 485 Modbus

It is a standard interface of the physical layer of communication, a method of signal transmission, the 1st level of the OSI (Open Systems Interconnection) model, transmitters and receivers exchange data through a twisted pair cable of rigid wires of 22 or 24 AWG, the maximum cable length used in RS-485 communications is 1200 meters at 100 Kbps with up to 247 peripherals

4.3. Data Storage

Data storage is another of the objectives of this research project, once the acquisition of these data is stored on a local server and processed there, this because it took time to process the information through fog computing taking into account Considering the following characteristics but above all mainly the low quality of internet service available at the data collection site, in the following table 10 we can see the comparison of the two proposed methodologies Cloud computing and Fog Computing.

Table 10
Cloud computing vs Fog Computing.

Requirements	Cloud Computing	Fog Computing
Latency	High	Low
Delay Jitter	High	Very Low
Location of Service	Within the Internet	At the edge of the local network
Distance between client and server	Multiple hops	One hope
Security	Undefined	Can be defined
Attack on data enroute	High probability	Very low probability
Location awareness	No	Yes
N° of server nodes	Few	Very large
Support for Mobility	Limited	Supported
Real time interactions	Supported	Supported
Type of last mile connectivity	Leased Line	Wireless
Response time	Several Minutes	Milliseconds
Architecture	Centralized	Distributed
continuously Internet access	High	Low

4.3.1. FOG Computing

As can be seen in the following figure 12, we have two photovoltaic systems, each one of them collects the information from the sensors through the PLC LOGO using the Modbus communication protocol, then the PLC sends the information to the server using the Ethernet communication protocol. through a network switch; Once the information is received, the server stores the information for further processing and monitoring by the researchers.

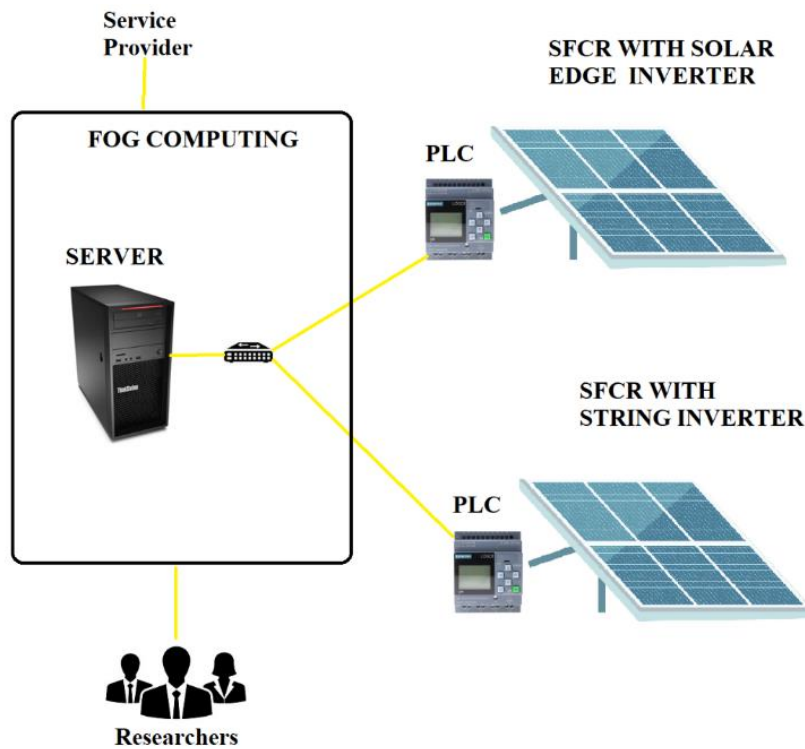


Figure 12
Operating scheme of fog computing.

4.3.2. Server

A computer was used as the central server of the system based on fog computing, it is in charge of collecting the information from the PLCs and also storing the information for further processing, the characteristics of the server are shown in the following table 11.

Table 11
Server features.

Item	Features
Computer	Lenovo P320 Intel
Processor	Intel® Core TM i7-7700 (3,6 Ghz, up to 4.2 Ghz with intel turbo boost, 8 MB cache, 4 cores)
RAM memory	32 GB DDR4 SDRAM 2400 MHZ e
Disk	HHD 1 TB SAT a de 7200 rpm + SSD 500gb
Case	Chassis in tower with 250W power supply
Optical drive	DVD+RW
Video	Nvidia Quadro P4000 8GB GDDR5 256-bit
Network	1 x 10/100/1000 mb / s gigabit ethernet (rj45) intel I219lm. Multimedia: Realtek ALC662
Ports and slots	6 x USB 3.1 gen 1 type-a 2 x USB 2.0 type-a 1 x de-9 serial monitor: 2 x display port 1 XVGA audio
Display	Monitor 32 "curved looor, 2560x1440, va, HDMI / DP / headphone refresh speed: 144hz, aspect ratio: 16: 9, brightness: 250cd / m2, contrast ratio: 2500: 1, view angle: 178° (h) / 178° (v), response time: 1ms, 100 - 240 vac
Peripherals	Keyboard and mouse
Operating system	Windows 10 pro (64 bits)

4.3.3. LabVIEW

The software in charge of collecting and storing the information is LabVIEW (Laboratory Virtual Instrument Engineering Workbench), which receives the information from the LOGO PLCs through the Modbus RS485 communication protocol, to later be decoded and stored in the central computer. The LabVIEW version used is 2018.

4.3.4. Graphical user interface GUI

In the following figure 13 and 14 we can see the graphical interfaces developed in the LabVIEW software to visualize and graph the data obtained from the photovoltaic systems, for the case of this study the following values were recorded: Date, Hour, AC

Current, AC Voltage, AC Power, AC Frequency, AC Apparent Power, AC Reactive Power, AC Power, Factor, DC Current, DC Voltage, DC Power.

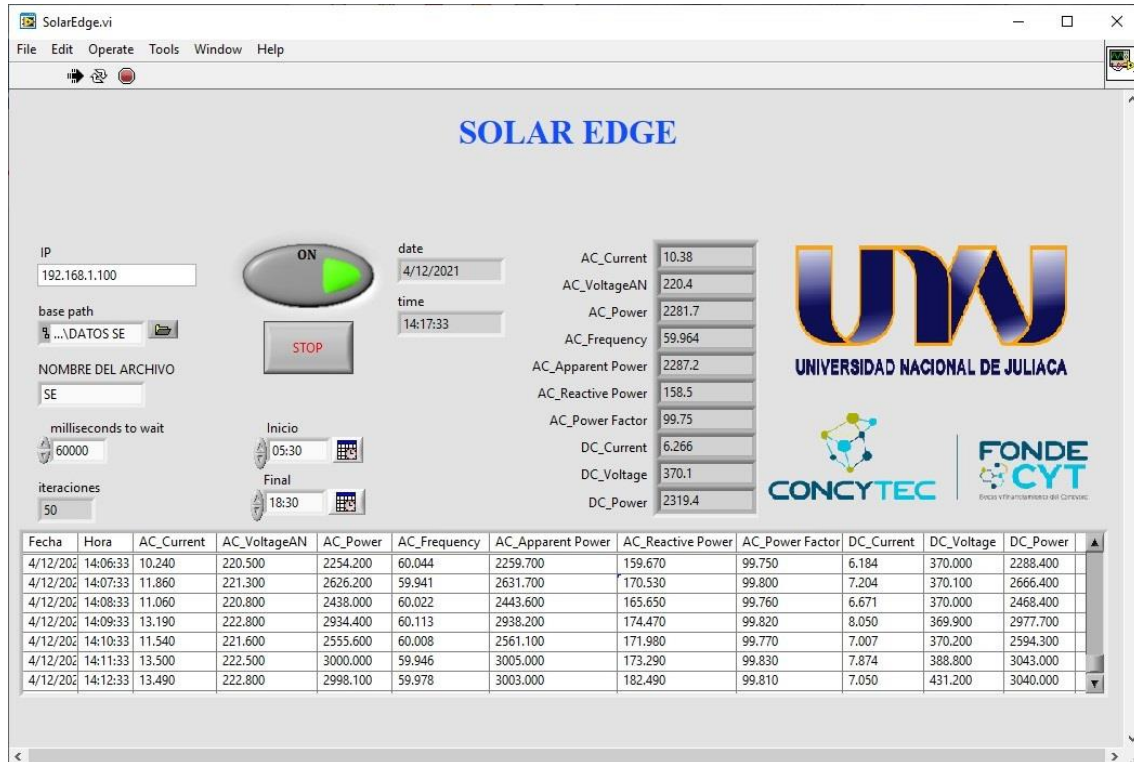


Figure 13
SFCR Solar Edge graphical user interface.

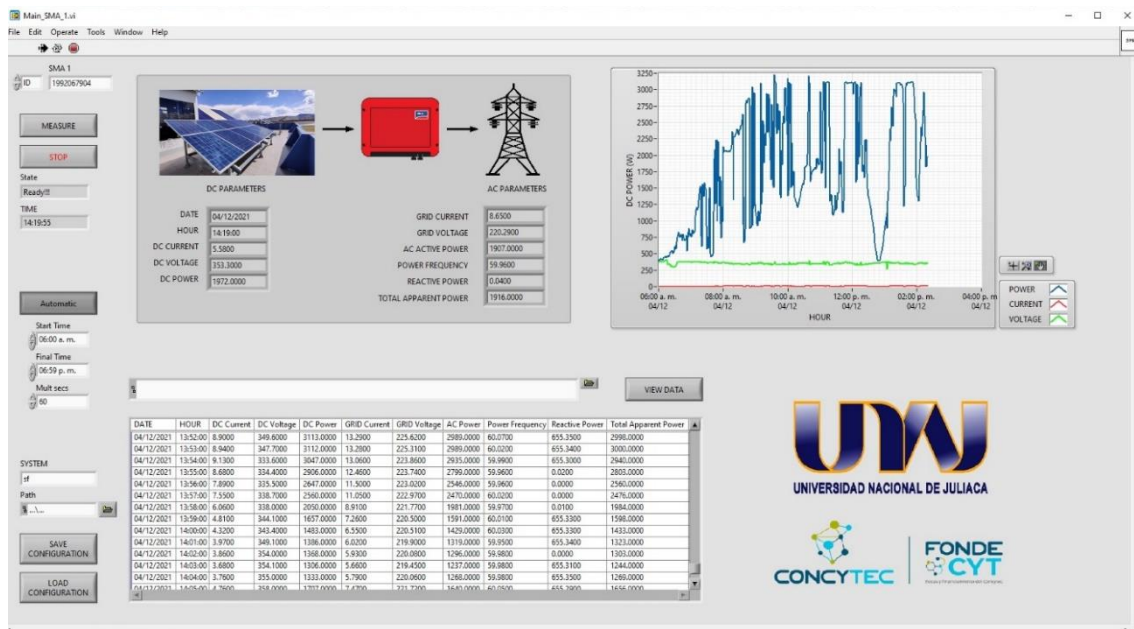


Figure 14
SFCR String graphical user interface.

4.4. Data processing.

The system records data every 60 seconds in accordance with the IEC 61724 standard, these data are grouped into the following 12 parameters: date, hour, ac current, ac voltage, ac power, ac frequency, ac apparent power, ac reactive power, ac power factor, dc current, dc voltage and dc power and the amount generated by each of the 2 systems is 5000 records, so to the day of the two systems we have about 10,000 records to process and by the end of the year we will have more than 3.6 million of records to process, so the most viable way to process so much information is to use artificial intelligence techniques to achieve results in the shortest possible time.

4.4.1. SFCR Solar Edge data processing

The methodology that will be applied to process the imputation of missing data will be artificial intelligence through machine Learning and Python following the KNN, Mean, median and frequent models.

4.4.1.1. Data set SFCR Solar Edge

Data set sample from SFCR Solar Edge see table 12; Graphical scheme of missing data SFRC Solar Edge see figure 15, Solar Edge Data Amount, see figure 16 and Data set with missing data SFCR Solar Edge see table 13.

Table 12
Data set SFCR Solar Edge.

	AC	AC	AC	AC	AC	AC	DC	DC	DC
	CURRENT	VOLTAJE	POWER	FRECUENCY	APARENT POWER	REACTIVE POWER	CURRENT	VOLTAGE	POWER
count	4769	4832	4819	4907	4858	4856	4861	4846	4859
mean	7.428035	218.8292	1640.146	60.0398	1647.841	150.5274	4.217523	369.6095	1660.418
std	4.855554	5.908173	1089.576	2.13503	1081.63	40.05224	2.686922	75.67641	1104.69
min	0	13.61	0	59.861	0	0	0	0	0
25%	2.87	216.6	635.4	59.969	638.65	135.5775	1.728	370	641.9
50%	7.1	219.7	1576.4	59.996	1561.65	154.47	4.252	370.1	1582.8
75%	12.98	221.6	2874.55	60.023	2876.975	176.4025	6.994	370.3	2916.85
max	13.84	228.7	3009	171.12	3016	216.71	8.206	445.7	3055

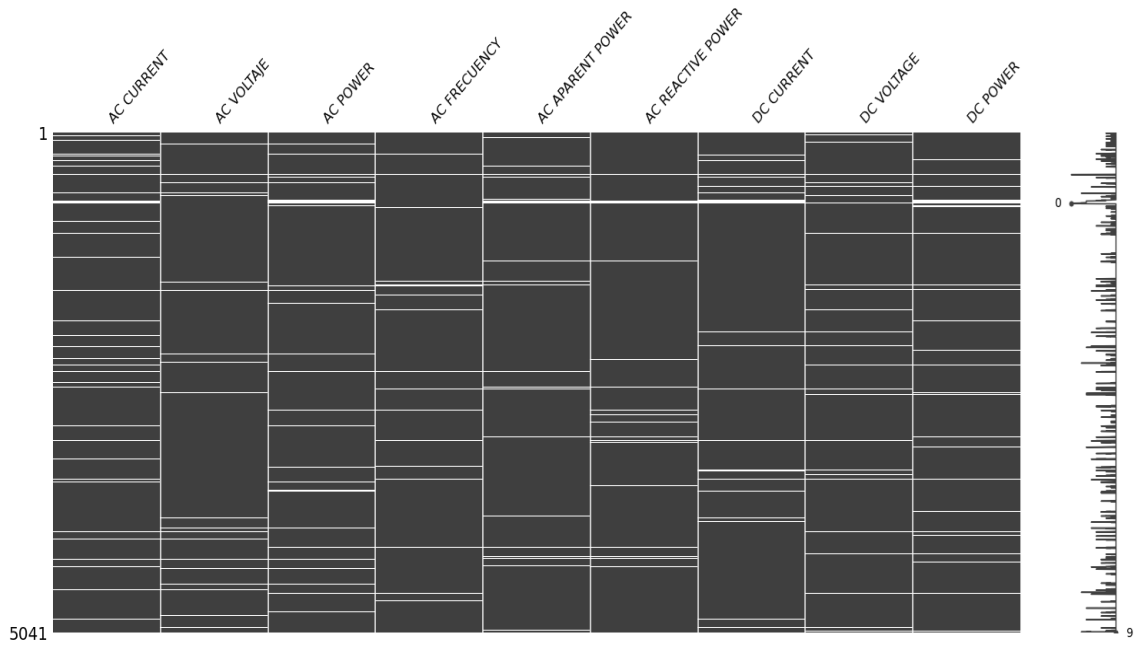


Figure 15
Graphical scheme of missing data SFRC Solar Edge.

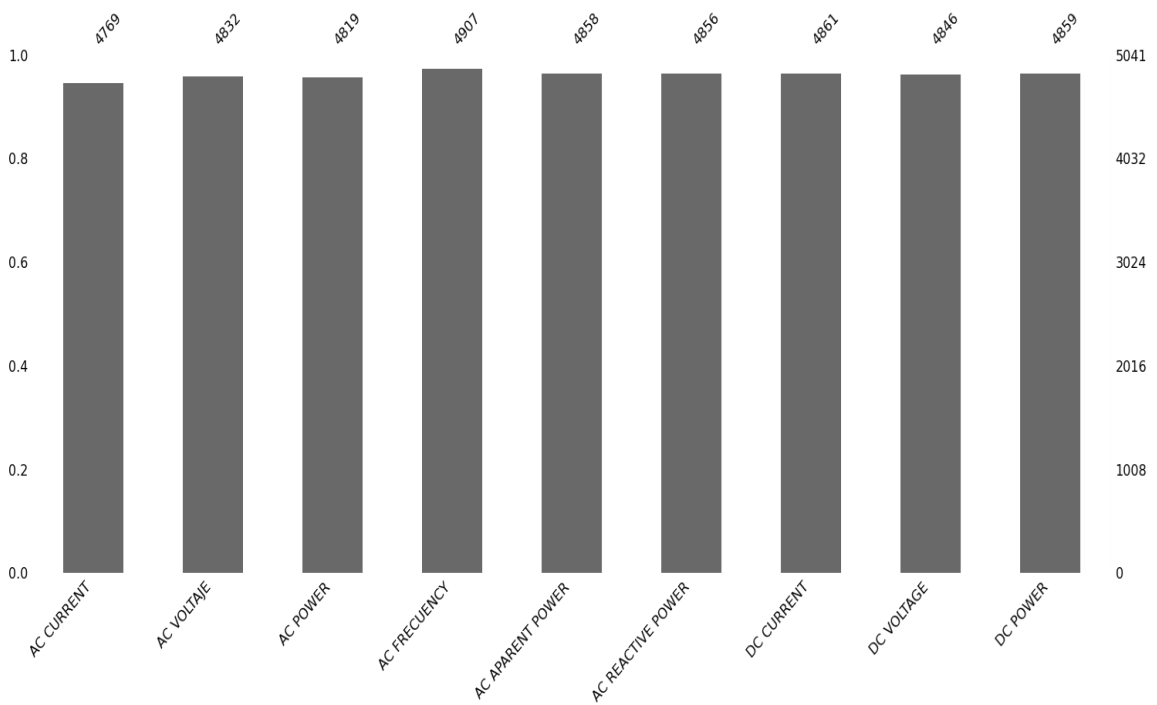


Figure 16
SFRC Solar Edge Data Amount.

Table 13
Data set with missing data SFCR Solar Edge.

	AC	AC	AC	AC	AC	AC	DC	DC	DC
	CURRENT	VOLTAJE	POWER	FRECUENCY	APARENT	REACTIVE	CURRENT	VOLTAGE	POWER
					POWER	POWER			
0	0.65	218.2	81.97	59.981	143.86	118.23	0.247	370.1	91.47
1	0.66	217.7	88.12	59.974	140.50	112.46	0.231	370.2	85.51
2	0.68	217.8	95.01	60.005	152.04	118.70	0.261	369.9	96.46
3	0.74	217.5	123.74	59.965	162.53	105.37	0.303	369.8	112.21
4	0.75	217.7	NaN	59.945	164.38	110.81	0.333	369.8	122.87
5	NaN	217.2	134.78	59.977	169.93	103.49	0.383	370.2	141.80
6	0.85	217.4	NaN	NaN	182.32	107.79	0.403	370.0	149.28
7	0.90	217.3	NaN	NaN	196.21	102.98	0.459	369.7	169.55
8	0.95	217.6	179.07	59.951	207.10	103.34	0.482	369.9	178.37
9	0.98	217.3	178.53	60.027	215.98	103.77	0.520	369.7	192.31

4.4.1.2. Data processing methodology SFCR Solar Edge

The figure 17 shows the process followed by the data imputation models applied in this investigation.

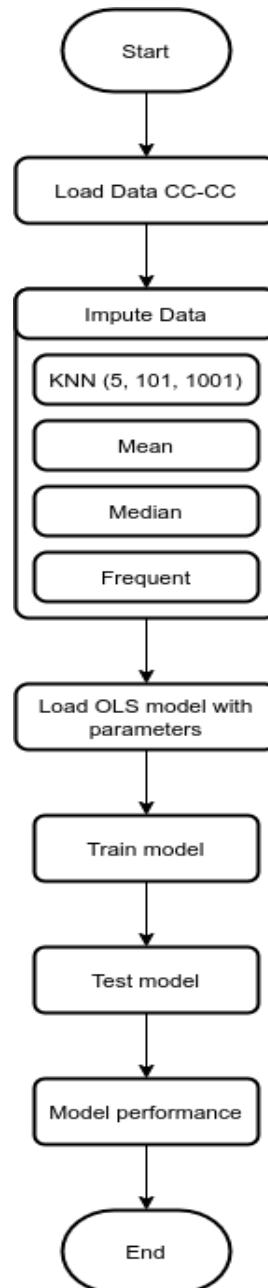


Figure 17
Data processing methodology SFCR Solar Edge.

Figure 18 shows the correlation between of variables of to SFCR Solar Edge

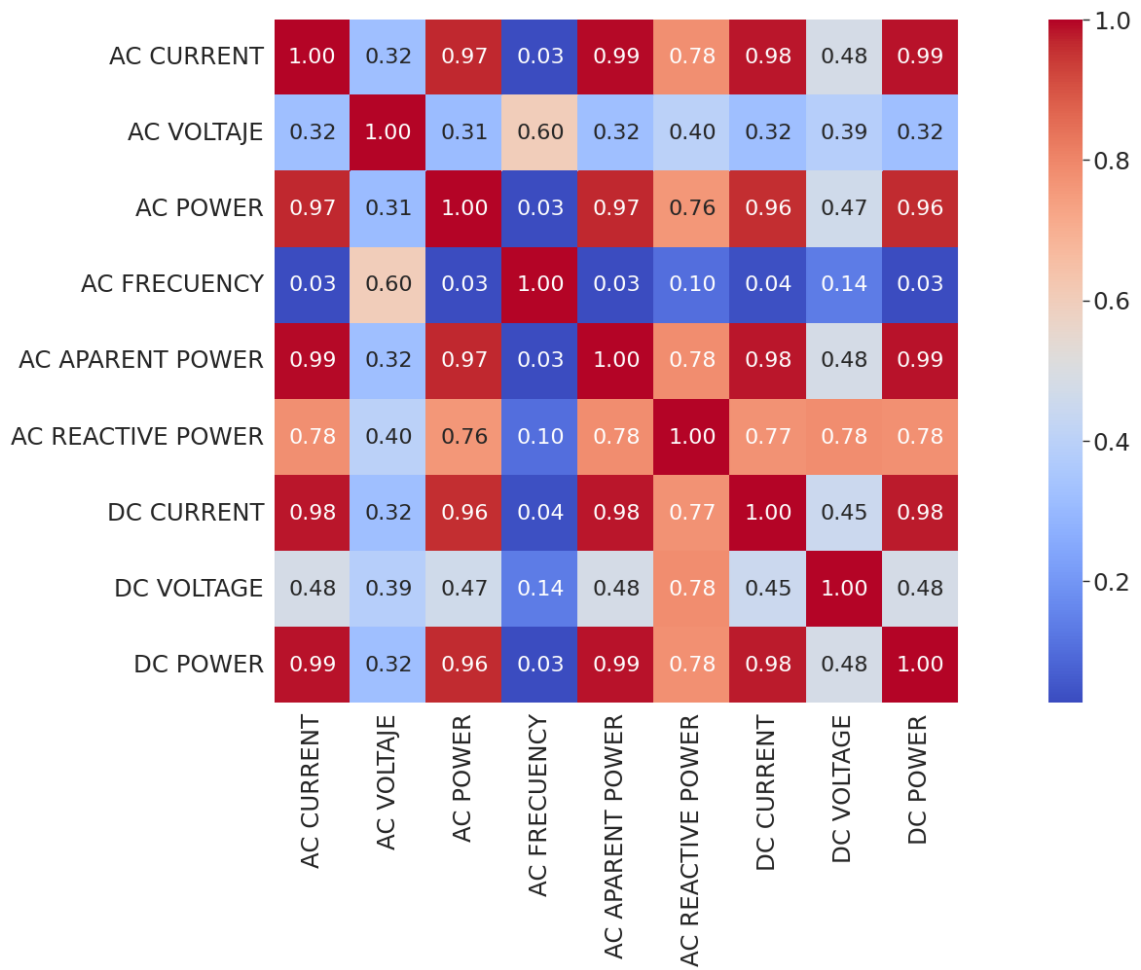


Figure 18
Correlation of variables SFCR Solar Edge.

SFCR Solar Edge: KNN=1001

The table 14 shows the results obtained after applying the KNN model with $k = 1001$.

Table 14
SFCR Solar Edge: KNN=1001.

	AC CURRENT	AC VOLTAJE	AC POWER	AC FRECUENCY	AC APARENT POWER	AC REACTIVE POWER	DC CURRENT	DC VOLTAGE	DC POWER
0	0.65	218.2	81.97	59.98	143.86	118.23	0.25	370.1	91.47
1	0.66	217.7	88.12	59.97	140.5	112.46	0.23	370.2	85.51
2	0.68	217.8	95.01	60.01	152.04	118.7	0.26	369.9	96.46
3	0.74	217.5	123.74	59.97	162.53	105.37	0.3	369.8	112.21
4	0.75	217.7	2862.94	59.95	164.38	110.81	0.33	369.8	122.87
5	1.05	217.2	134.78	59.98	169.93	103.49	0.38	370.2	141.8
6	0.85	217.4	223.2	60.04	182.32	107.79	0.4	370	149.28
7	0.9	217.3	140.09	60.01	196.21	102.98	0.46	369.7	169.55
8	0.95	217.6	179.07	59.95	207.1	103.34	0.48	369.9	178.37
9	0.98	217.3	178.53	60.03	215.98	103.77	0.52	369.7	192.31

Score Test of the Least Squares model

0.9386595049322922

MAE of the Least Squares model

93.62817880579244

MSE of the Least Squares model

269.1241702616536

Determination Coefficient of the Least Squares model

0.9386595049322922

Adjusted coefficient of determination of the Least Squares model

0.9385619842724071

MULTIPLE LINEAR REGRESSION MODEL DATA

Slopes or coefficients value "a":

[9.26174874e+01 -8.32522469e-01 1.05030860e+00 9.32052283e-02
2.62057606e-02 8.10689127e+01 1.26499099e-01 2.68279687e-01]

Intersection or coefficient value "b":

103.45637343261001

$$Y = 92.6174874 * X_1 - 0.832522469 * X_2 + 1.05030860 * X_3 + 0.0932052283 * X_4 + 0.0262057606 * X_5 + 81.0689127 * X_6 + 0.126499099 * X_7 + 0.268279687 * X_8 + 103.45637343261001$$

OLS training time

0:00:00.007283

OLS test time

0:00:00.002563

SFCR Solar Edge: KNN=101

The table 15 shows the results obtained after applying the KNN model with $k = 101$.

Table 15
SFCR Solar Edge: KNN=101.

	AC CURRENT	AC VOLTAJE	AC POWER	AC FRECUENCY	AC APARENT POWER	AC REACTIVE POWER	DC CURRENT	DC VOLTAGE	DC POWER
0	0.65	218.2	81.97	59.98	143.86	118.23	0.25	370.1	91.47
1	0.66	217.7	88.12	59.97	140.5	112.46	0.23	370.2	85.51
2	0.68	217.8	95.01	60.01	152.04	118.7	0.26	369.9	96.46
3	0.74	217.5	123.74	59.97	162.53	105.37	0.3	369.8	112.21
4	0.75	217.7	2864.15	59.95	164.38	110.81	0.33	369.8	122.87
5	1.08	217.2	134.78	59.98	169.93	103.49	0.38	370.2	141.8
6	0.85	217.4	229.23	59.99	182.32	107.79	0.4	370	149.28
7	0.9	217.3	138.59	59.99	196.21	102.98	0.46	369.7	169.55
8	0.95	217.6	179.07	59.95	207.1	103.34	0.48	369.9	178.37
9	0.98	217.3	178.53	60.03	215.98	103.77	0.52	369.7	192.31

Score Test of the Least Squares model

0.9385386951408857

MAE of the Least Squares model

93.78729380788776

MSE of the Least Squares model

269.38116093271987

Determination Coefficient of the Least Squares model

0.9385386951408857

Adjusted coefficient of determination of the Least Squares model

0.9384409824145596

MULTIPLE LINEAR REGRESSION MODEL DATA

Slopes or coefficients value "a"

[9.29349817e+01 -7.83117325e-01 8.48247247e-01 9.40946314e-02

2.78974176e-02 8.11662134e+01 1.24027184e-01 2.65644155e-01]

Intersection or coefficient value "b"

105.50809777048448

$$Y = 92.9349817 * X_1 - 0.783117325 * X_2 + 0.848247247 * X_3 + 0.0940946314 * X_4 + 0.0278974176 * X_5 + 81.1662134 * X_6 + 0.124027184 * X_7 + 0.265644155 * X_8 + 105.50809777048448$$

OLS Training time

0:00:00.007330

OLS test time

0:00:00.004010

SFCR Solar Edge KNN=5

The table 16 shows the results obtained after applying the KNN model with $k = 5$.

Table 16
SFCR Solar Edge KNN=5.

	AC CURRENT	AC VOLTAJE	AC POWER	AC FRECUENCY	AC APARENT POWER	AC REACTIVE POWER	DC CURRENT	DC VOLTAGE	DC POWER
0	0.65	218.2	81.97	59.98	143.86	118.23	0.25	370.1	91.47
1	0.66	217.7	88.12	59.97	140.5	112.46	0.23	370.2	85.51
2	0.68	217.8	95.01	60.01	152.04	118.7	0.26	369.9	96.46
3	0.74	217.5	123.74	59.97	162.53	105.37	0.3	369.8	112.21
4	0.75	217.7	2869.59	59.95	164.38	110.81	0.33	369.8	122.87
5	0.79	217.2	134.78	59.98	169.93	103.49	0.38	370.2	141.8
6	0.85	217.4	213.44	59.98	182.32	107.79	0.4	370	149.28
7	0.9	217.3	106.56	59.98	196.21	102.98	0.46	369.7	169.55
8	0.95	217.6	179.07	59.95	207.1	103.34	0.48	369.9	178.37
9	0.98	217.3	178.53	60.03	215.98	103.77	0.52	369.7	192.31

Score Test of the Least Squares model

0.937560722752234

MAE of the Least Squares model

95.36337947766249

MSE of the Least Squares model

271.65241402597877

Determination Coefficient of the Least Squares model

0.937560722752234

Adjusted coefficient of determination of the Least Squares model



0.9374614552208386

MULTIPLE LINEAR REGRESSION MODEL DATA

Slopes or coefficients value "a"

[8.40566122e+01 -6.40307836e-01 1.44376399e-02 1.07946102e-01
-6.31433903e-02 8.46376658e+01 1.63765089e-01 2.83619741e-01]

Intersection or coefficient value "b"

121.65555457934329

$$Y = 84.0566122 * X_1 - 0.640307836 * X_2 + 0.0144376399 * X_3 + 0.107946102$$
$$* X_4 - 0.0631433903 * X_5 + 84.6376658$$
$$* X_6 + 0.163765089 * X_7 + 0.283619741 * X_8$$
$$+ 121.65555457934329$$

OLS training time

0:00:00.002527

OLS test time

0:00:00.002815

SFCR Solar Edge Mean

The table 17 shows the results obtained after applying the Mean model.

Table 17
SFCR Solar Edge Mean.

	AC CURRENT	AC VOLTAJE	AC POWER	AC FRECUENCY	AC APARENT POWER	AC REACTIVE POWER	DC CURRENT	DC VOLTAGE	DC POWER
0	0.650000	218.2	81.970000	59.981000	143.86	118.23	0.247	370.1	91.47
1	0.660000	217.7	88.120000	59.974000	140.50	112.46	0.231	370.2	85.51
2	0.680000	217.8	95.010000	60.005000	152.04	118.70	0.261	369.9	96.46
3	0.740000	217.5	123.740000	59.965000	162.53	105.37	0.303	369.8	112.21
4	0.750000	217.7	1640.145555	59.945000	164.38	110.81	0.333	369.8	122.87
5	7.428035	217.2	134.780000	59.977000	169.93	103.49	0.383	370.2	141.80
6	0.850000	217.4	1640.145555	60.039798	182.32	107.79	0.403	370.0	149.28
7	0.900000	217.3	1640.145555	60.039798	196.21	102.98	0.459	369.7	169.55
8	0.950000	217.6	179.070000	59.951000	207.10	103.34	0.482	369.9	178.37
9	0.980000	217.3	178.530000	60.027000	215.98	103.77	0.520	369.7	192.31

Score Test of the Least Squares model

0.9608237737157986

MAE of the Least Squares model

87.59172845852068

MSE of the Least Squares model

209.71013783436953

Determination Coefficient of the Least Squares model

0.9608237737157986

Adjusted coefficient of determination of the Least Squares model

0.9607614903671751

MULTIPLE LINEAR REGRESSION MODEL DATA

Slopes or coefficients value "a"

[5.10006882e+01 2.28186787e+00 -2.77871835e+00 2.80818245e-01



3.63623312e-01 7.07597706e+01 -3.79545869e-04 2.89070033e-01]

Intersection or coefficient value "b"

-365.44480762886064

$$Y = 51.0006882 * X_1 + 2.28186787 * X_2 - 2.77871835 * X_3 + 0.280818245 * X_4 \\ - 0.363623312 * X_5 + 70.7597706 \\ * X_6 - 0.000379545869 * X_7 + 0.289070033 * X_8 \\ - 365.44480762886064$$

OLS Training time

0:00:00.008241

OLS test time

0:00:00.003755

SFCR SOLAR EDGE Median

The table 18 shows the results obtained after applying the Median model.

Table 18
SFCR SOLAR EDGE Median.

	AC CURRENT	AC VOLTAJE	AC POWER	AC FRECUENCY	AC APARENT POWER	AC REACTIVE POWER	DC CURRENT	DC VOLTAGE	DC POWER
0	0.65	218.2	81.97	59.981	143.86	118.23	0.247	370.1	91.47
1	0.66	217.7	88.12	59.974	140.50	112.46	0.231	370.2	85.51
2	0.68	217.8	95.01	60.005	152.04	118.70	0.261	369.9	96.46
3	0.74	217.5	123.74	59.965	162.53	105.37	0.303	369.8	112.21
4	0.75	217.7	1576.40	59.945	164.38	110.81	0.333	369.8	122.87
5	7.10	217.2	134.78	59.977	169.93	103.49	0.383	370.2	141.80
6	0.85	217.4	1576.40	59.996	182.32	107.79	0.403	370.0	149.28
7	0.90	217.3	1576.40	59.996	196.21	102.98	0.459	369.7	169.55
8	0.95	217.6	179.07	59.951	207.10	103.34	0.482	369.9	178.37
9	0.98	217.3	178.53	60.027	215.98	103.77	0.520	369.7	192.31

Score Test of the Least Squares model

0.9616751662816645

MAE of the Least Squares model

86.64285541510337

MSE of the Least Squares model

207.44449798088687

Determination Coefficient of the Least Squares model

0.9616751662816645

Adjusted coefficient of determination of the Least Squares model

0.9616142364983286



MULTIPLE LINEAR REGRESSION MODEL DATA

Slopes or coefficients value "a"

[5.01318185e+01 2.17193247e+00 -2.88629646e+00 2.76657692e-01
4.05013076e-01 7.25106952e+01 -6.77598585e-03 2.92151149e-01]

Intersection or coefficient value "b"

-338.5401667408278

$$Y = 50.1318185 * X_1 + 2.17193247 * X_2 - 2.88629646 * X_3 + 0.276657692 * X_4 \\ - 0.405013076 * X_5 + 72.5106952 \\ * X_6 - 0.00677598585 * X_7 + 0.292151149 * X_8 \\ - 338.5401667408278$$

OLS Training time

0:00:00.007574

OLS test time

0:00:00.004769

SFCR Solar Edge Frequent

The table 19 shows the results obtained after applying the Frequent model.

Table 19
SFCR Solar Edge Frequent.

	CURRENT	AC VOLTAJE	AC POWER	AC FRECUENCY	AC APARENT POWER	AC REACTIVE POWER	DC CURRENT	DC VOLTAGE	DC POWER
0	0.65	218.2	81.97	59.981	143.86	118.23	0.247	370.1	91.47
1	0.66	217.7	88.12	59.974	140.50	112.46	0.231	370.2	85.51
2	0.68	217.8	95.01	60.005	152.04	118.70	0.261	369.9	96.46
3	0.74	217.5	123.74	59.965	162.53	105.37	0.303	369.8	112.21
4	0.75	217.7	0.00	59.945	164.38	110.81	0.333	369.8	122.87
5	0.00	217.2	134.78	59.977	169.93	103.49	0.383	370.2	141.80
6	0.85	217.4	0.00	59.980	182.32	107.79	0.403	370.0	149.28
7	0.90	217.3	0.00	59.980	196.21	102.98	0.459	369.7	169.55
8	0.95	217.6	179.07	59.951	207.10	103.34	0.482	369.9	178.37
9	0.98	217.3	178.53	60.027	215.98	103.77	0.520	369.7	192.31

Score Test of the Least Squares model

0.9008141043796998

MAE of the Least Squares model

112.5108142366538

MSE of the Least Squares model

350.3529333006365

Determination Coefficient of the Least Squares model

0.9008141043796998

Adjusted coefficient of determination of the Least Squares model

0.9006564161513686



MULTIPLE LINEAR REGRESSION MODEL DATA

Slopes or coefficients value "a"

[40.59986732 2.01688293 -0.32026563 0.28296111 -1.04328913
93.17332918 0.54673333 0.29596882]

Intersection or coefficient value "b"

-492.7432097982205

$$Y = 40.59986732 * X_1 + 2.01688293 * X_2 - 0.32026563 * X_3 + \\ 0.28296111 * X_4 - 1.04328913 * X_5 + 93.17332918 * X_6 - 0.54673333 * \\ X_7 + 0.29596882 * X_8 - 492.7432097982205$$

OLS training time

0:00:00.006944

OLS test time

0:00:00.006578

The table 20 show sample data set completed example SFCR Solar Edge

Table 20

Data set completed example SFCR Solar Edge.

	AC CURRENT	AC VOLTAJE	AC POWER	AC FRECUENCY	AC APARENT POWER	AC REACTIVE POWER	DC CURRENT	DC VOLTAGE	DC POWER
count	5041	5041	5041	5041	5041	5041	5041	5041	5041
mean	7.439795	218.0938	1661.126	59.94348	1637.959	149.9048	4.214484	366.8904	1654.15
std	4.85212	10.65405	1091.441	3.185876	1082.406	40.82926	2.689414	80.69801	1095.419
min	0	0	0	0	0	0	0	0	0
25%	2.93	216.2	653.5	59.969	628.8	134.96	1.728	370	655.1
50%	7.08	219.5	1614.7	59.996	1538.3	154.01	4.249	370.1	1540.5
75%	12.98	221.5	2875.013	60.022	2873.1	176.28	6.994	370.3	2900.1
max	13.84	228.7	3009	171.12	3016	216.71	8.206	445.7	3055

4.4.1.3. SFCR Solar Edge Score models comparison

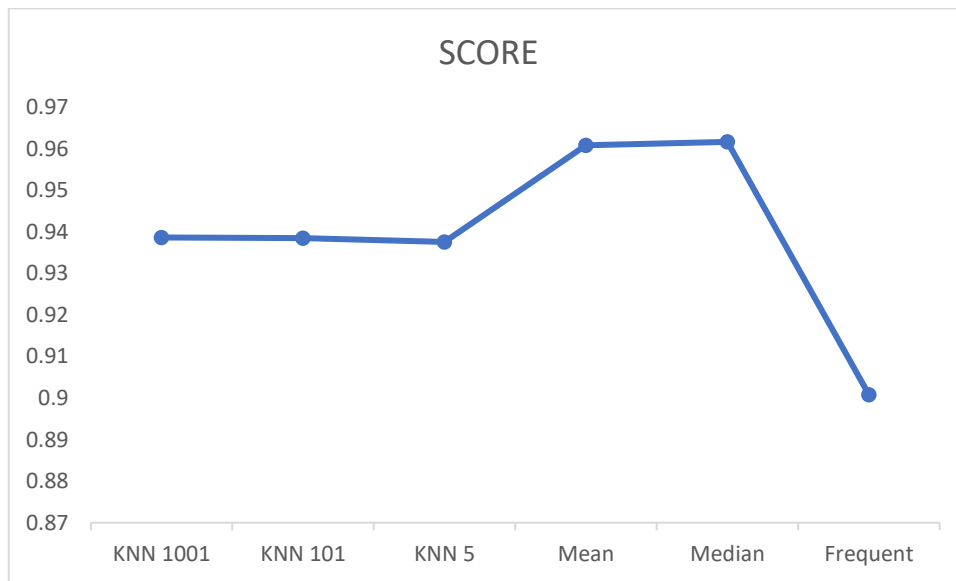


Figure 19
SFCR Solar Edge Score models comparison.

As it shows in the figure 19, the data imputation models that have obtained the best SCORE are the Median and Mean with 96.17% and 96.08% respectively; followed by the KNN model with $K = 1001$, $K = 101$ and $k = 5$ with 93.87%, 93.85% and 93.76% finally we have the Frequent model that reached only 90.08%

The SCORE refers to the score that the model has generated, it indicates the percentage of success when carrying out the imputation of data by applying two models.

4.4.1.4. SFCR Solar Edge MAE models comparison

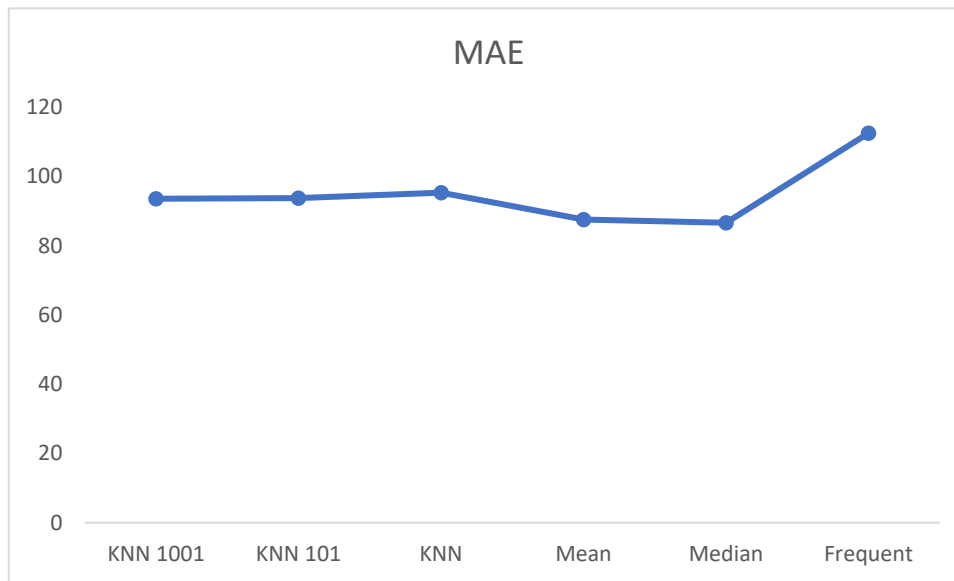


Figure 20
SFCR Solar Edge MAE models comparison.

As it shows in the figure 20, the models that have obtained the lowest MAE* are the Median and Mean with 86.64 and 87.59 respectively; followed by the KNN model with $K = 1001$, $K = 101$ and $k = 5$ with 93.63, 93.79 and 95.36 finally we have the Frequent model that reached 112.51

The Mean Absolute Error (MAE) is the mean of the differences between the target variable and the predictions without the sign, it does not vary much if there are extreme values in the data and the error is interpreted as units of the target variable and is determined by:

$$\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

4.4.1.5. SFCR Solar Edge MSE models comparison

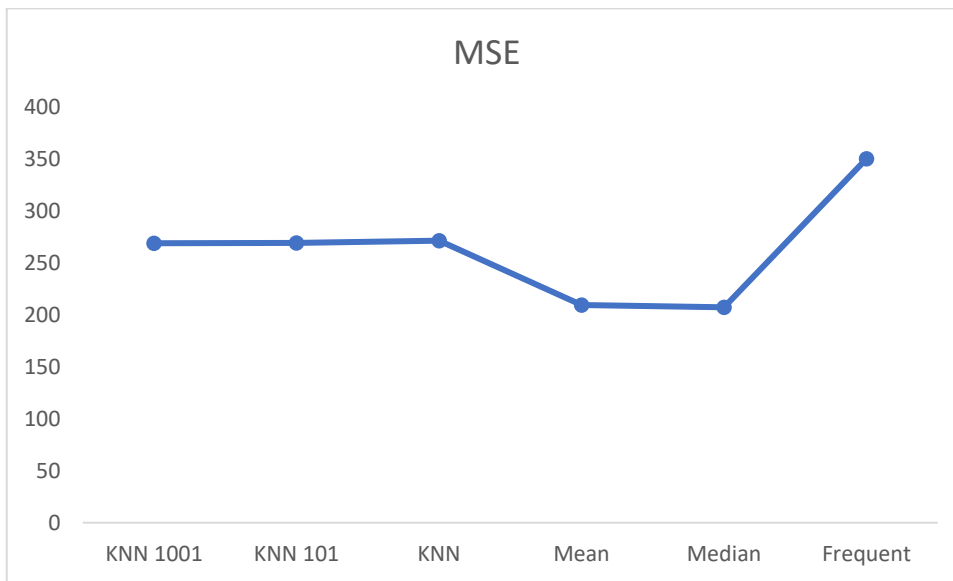


Figure 21
SFCR Solar Edge MSE models comparison.

As it shows in the figure 21, the models with the lowest MSE obtained are the Median and Mean with 207.44 and 209.71 respectively; followed by the KNN model with $K = 1001$, $K = 101$ and $k = 5$ with 269.12, 269.38 and 271.65 finally we have the Frequent model that reached 350.35

The Mean Square Error (MSE) as an estimator measures the average of the squared errors (the difference between the estimator and what is estimated), emphasizing outliers or extreme values, the error is interpreted as squared units and is determined by:

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

4.4.1.6. SFCR Solar Edge determination coefficient models comparison

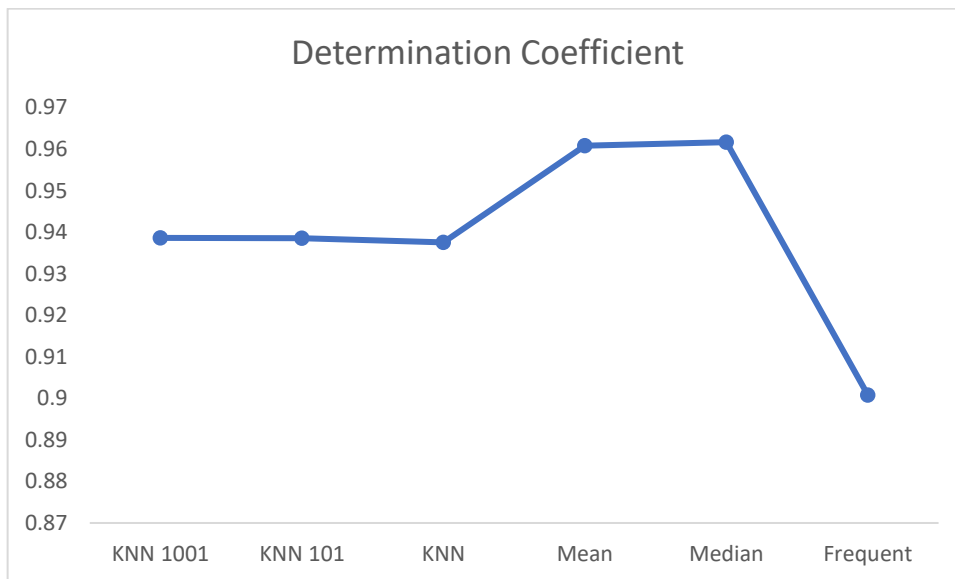


Figure 22

SFCR Solar Edge determination coefficient models comparison.

As it shows in the figure 22, the models that have obtained the highest coefficient of determination R^2 are the Median and Mean with 0.9617 and 0.9608 respectively; followed by the KNN model with $K = 1001$, $K = 101$ and $k = 5$ with 0.9387, 0.9385 and 0.9376 finally we have the Frequent model that reached 0.9008

The determination coefficient (R^2 or R squared) measures the portion of the variance of the objective variable that can be explained by the model, one of the ways to define R^2 is to take the correlation between the objective variable and the predictions, raised to the square and is defined by:

$$r = r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

R^2 It has a maximum value of 1 when the model explains all the variance, although it could have negative values.

4.4.1.7. SFCR Solar Edge Adjusted determination coefficient models comparison

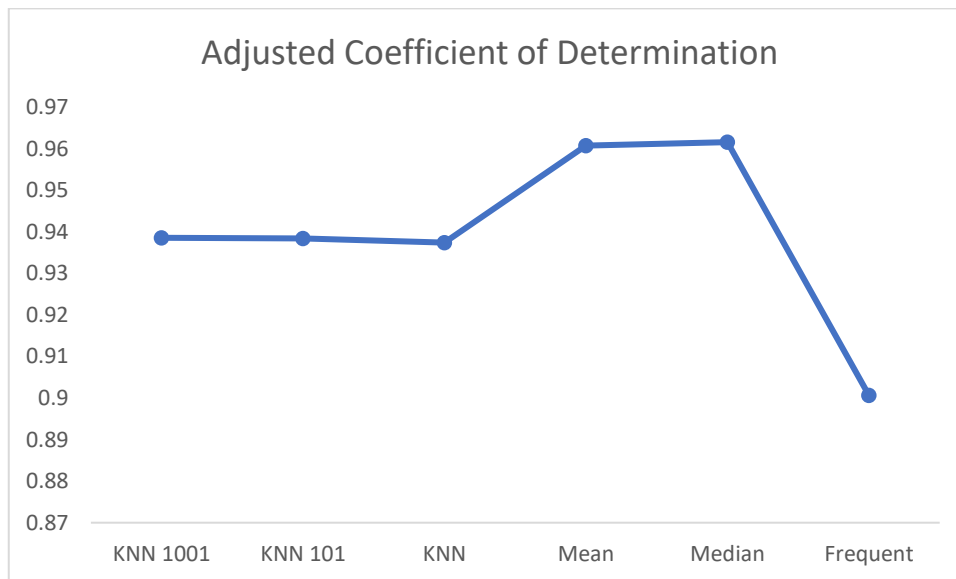


Figure 23

SFCR SolarEdge Adjusted deter. coefficient models comparison.

As it shows in the figure 23, the models that have obtained the highest adjusted coefficient of determination are the Median and Mean with 0.9616 and 0.9607 respectively; followed by the KNN model with K = 1001, K = 101 and k = 5 with 0.9386, 0.9384 and 0.9375 finally we have the Frequent model that reached 0.9007

Unlike the coefficient of determination, the adjusted coefficient of determination explains whether the model could be overfitting, so it is important to consider this metric that takes into account the complexity of the model and is determined by:

$$1 - \frac{(1 - R^2)(n - 1)}{(n - k - 1)}$$

4.4.1.8. SFCR Solar Edge Training time models comparison

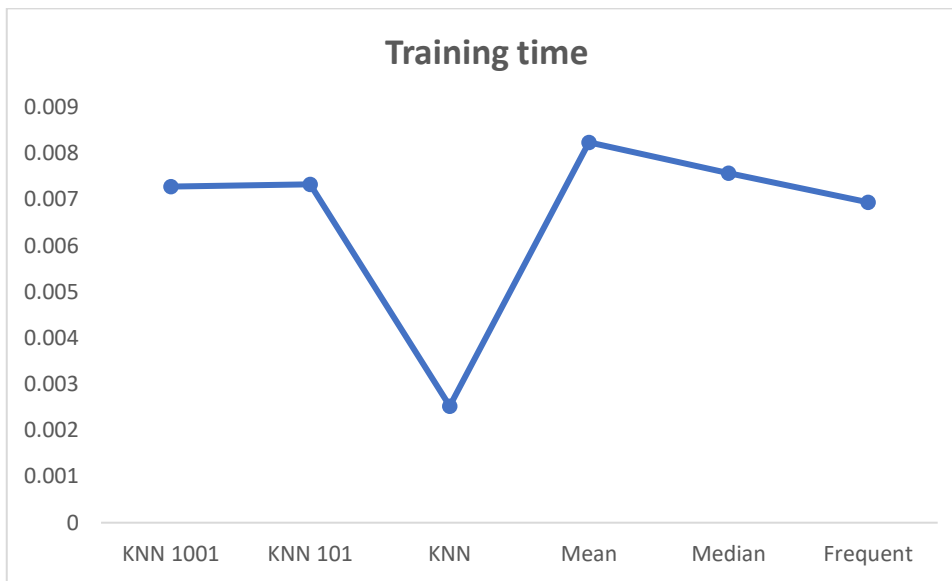


Figure 24
SFCR Solar Edge Training time models comparison.

In figure 24 we can see that the model for the imputation of KNN data with $K = 5$ is the one with the least processing time with 2,572ms followed by the Frequent model with 6,944 seconds, and then we have KNN with $K = 1001$ and $K = 101$ with 7.283ms and 7.33ms finally the models that took the longest are the median and mean with 7.574ms and 8.241ms

Training time is the time it takes the algorithm to obtain the model according to the conditions established from the training data.

4.4.1.9. SFCR Solar Edge Test time models comparison

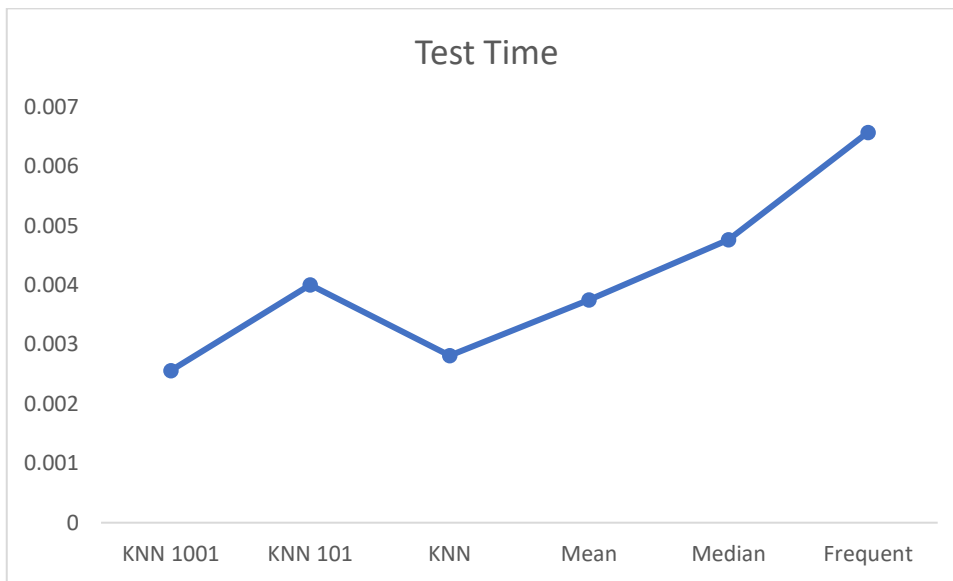


Figure 25
SFCR Solar Edge Test time models comparison.

In figure 25 we can see that the model for the imputation of KNN data with $K = 1001$ and $K = 5$ are the ones that took the least test time with 2,563ms and 2,815ms respectively; then they are followed by mean, KNN with $K = 101$ and they mediate with 3.755ms, 4.01ms and 4.769ms; Finally, the model that took the longest time was the frequent one with 6,578ms.

Test time, is the time it takes the algorithm to classify the new values according to the conditions established from the test or test data

4.4.2. SFCR String data processing

4.4.2.1. Data set SFCR String

Data set sample from SFCR String see table 21; Graphical scheme of missing data SFRC String see figure 26, String Data Amount, see figure 27 and Data set with missing data SFCR String see table 22

Table 21
Data set SFCR String.

	DC CURRENT	DC VOLTAGE	DC POWER	GRID CURRENT	GRID VOLTAJE	AC POWER	POWER FRECUENCY	REACTIVE POWER	APARENTE POWER
count	4706	4628	4697	4681	4656	4688	4702	4730	4789
mean	4.689535	349.931	1546.647	7.395768	220.5474	1484.699	60.22103	288.0823	1496.563
std	8.16199	60.58124	999.812	15.28159	3.279582	982.3839	12.3168	325.2566	980.7534
min	0.11	2.01	1.24	0.23	209.43	0	0	0	11
25%	1.7925	335.275	651	2.88	218.65	612	59.96	0	615
50%	4.105	349.3	1434	6.35	220.8	1395	59.99	0.04	1406
75%	7.46	364.2	2508	10.99	222.53	2431	60.02	655.33	2445
max	388.1	3101	3180	367	229.72	3028	655.32	655.35	3065

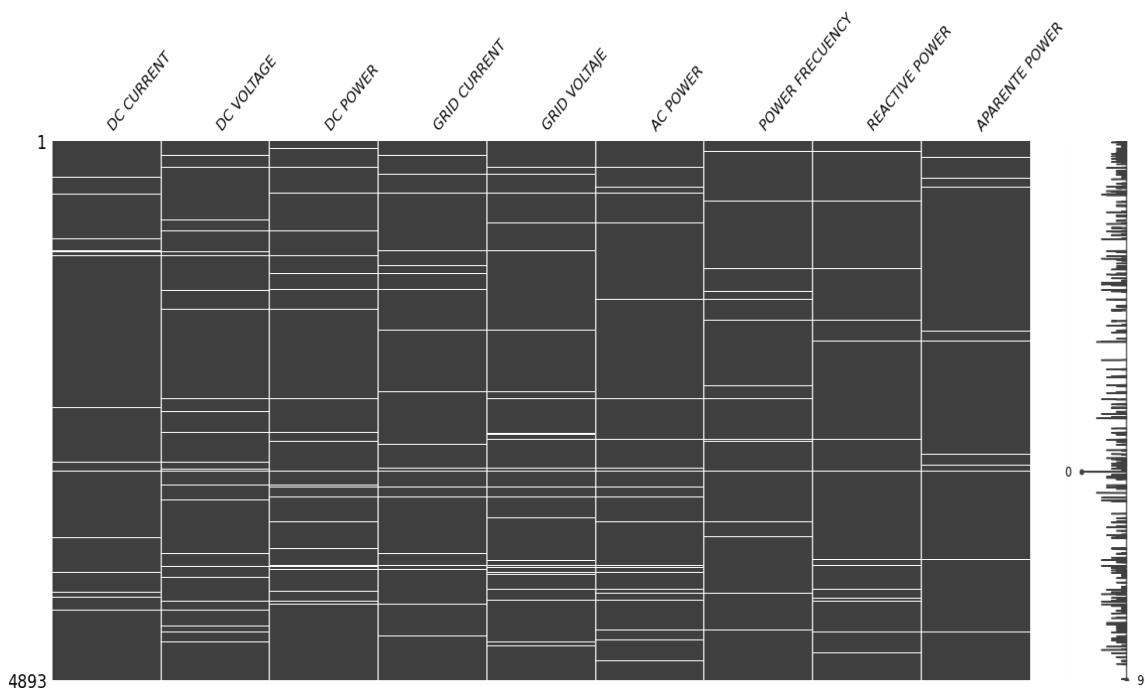


Figure 26
Graphical scheme of missing data SFRC String.

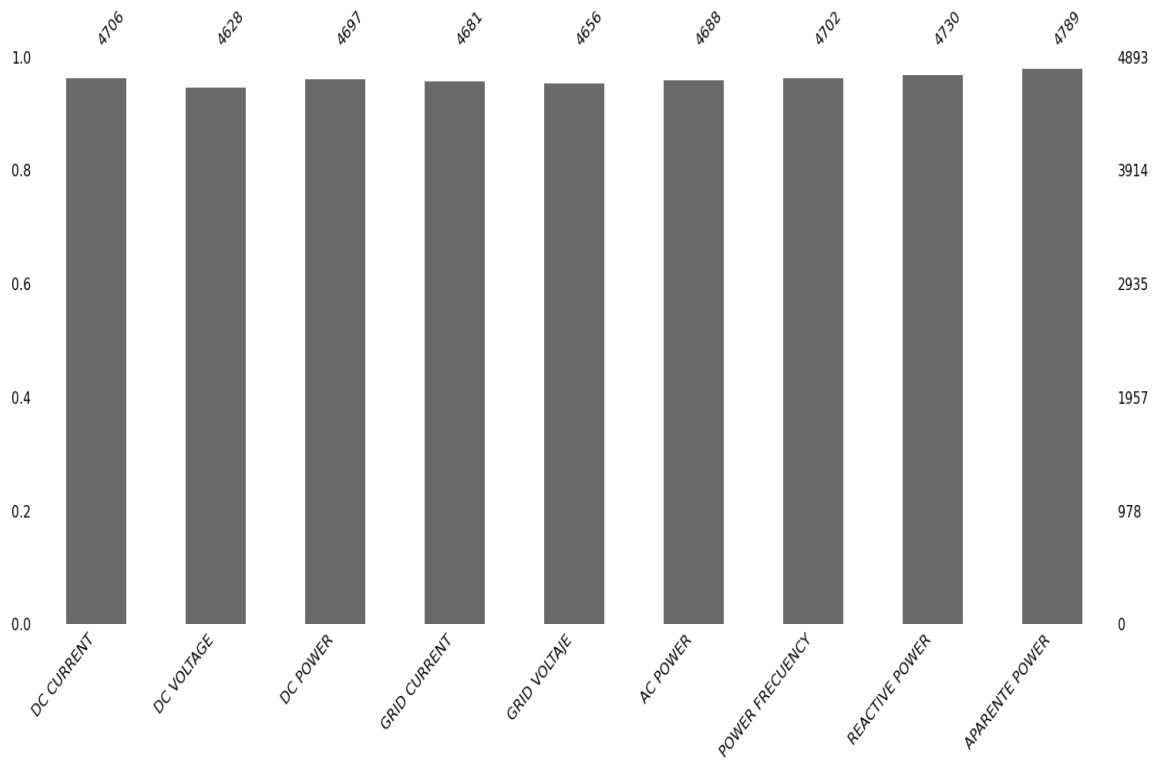


Figure 27
SFRC String Data Amount.

Table 22
Data set with missing data SFRC String.

	DC CURRENT	DC VOLTAGE	DC POWER	GRID CURRENT	GRID VOLTAJE	AC POWER	POWER FRECUENCY	REACTIVE POWER	APARENTE POWER
0	0.44	381.8	170.0	0.75	219.82	101.0	59.96	655.32	104.0
1	0.46	388.8	181.0	0.77	220.53	110.0	60.05	655.31	110.0
2	0.48	392.5	192.0	0.79	220.06	118.0	59.97	655.34	117.0
3	0.51	393.4	203.0	0.84	219.93	130.0	59.95	655.32	129.0
4	0.53	394.4	212.0	0.86	219.96	141.0	59.99	655.34	139.0
5	0.56	396.3	224.0	0.89	219.70	150.0	59.90	0.00	153.0
6	0.59	396.4	235.0	0.93	220.04	165.0	59.91	NaN	165.0
7	NaN	396.4	246.0	0.98	220.22	175.0	59.95	NaN	177.0
8	0.67	375.9	254.0	1.02	219.97	186.0	60.05	655.35	187.0
9	0.69	NaN	267.0	1.08	219.67	201.0	59.94	0.00	203.0

4.4.2.2. Data processing methodology SFCR String

The figure 28 shows the process followed by the data imputation models applied in this investigation.

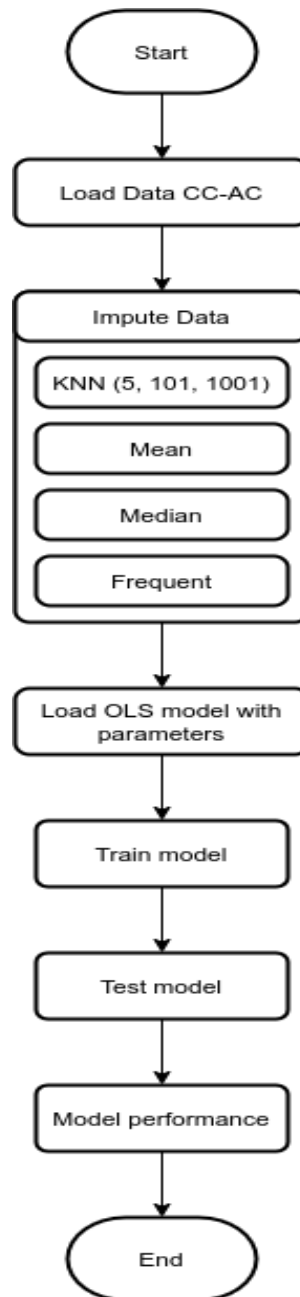


Figure 28
Data processing methodology SFCR String.

Figure 29 shows the correlation between of variables of to SFCR String

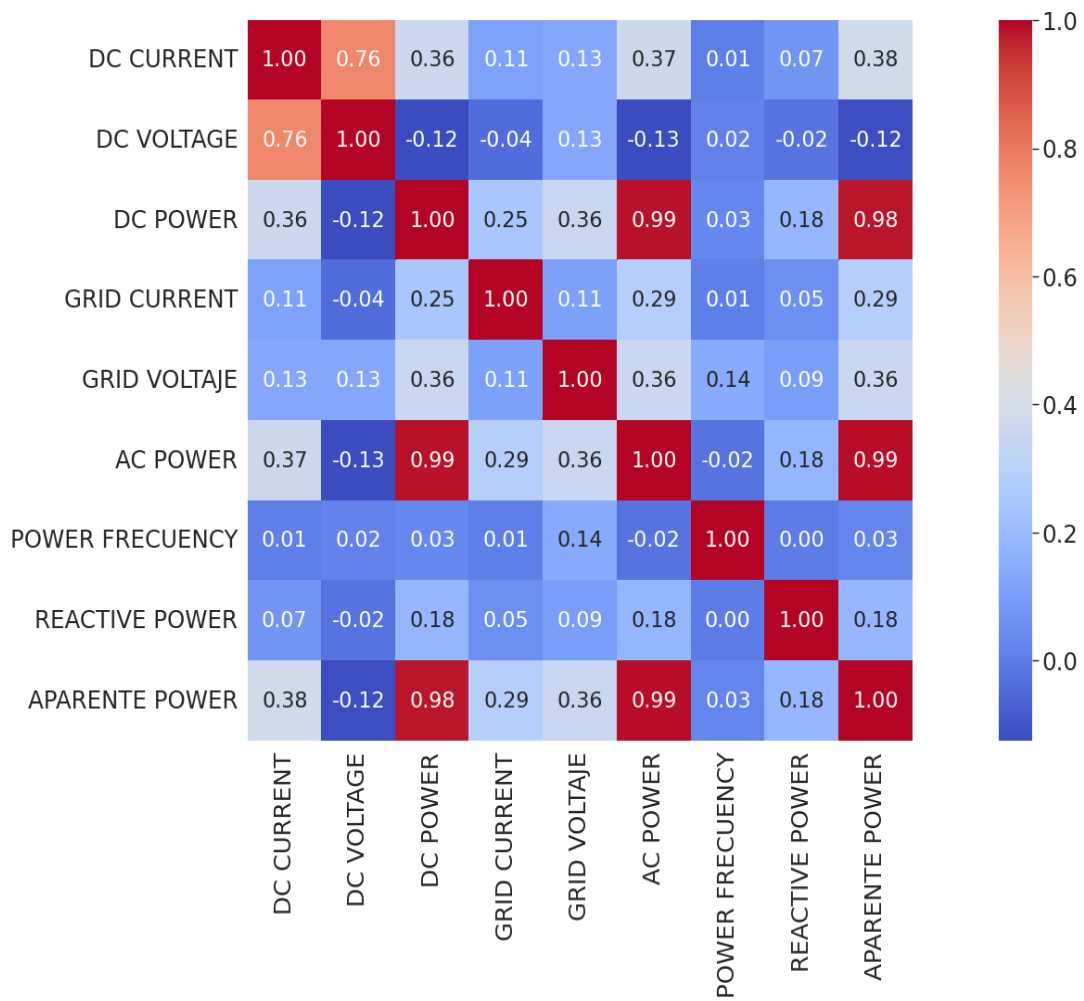


Figure 29
Correlation of variables SFCR String.

SFCR String KNN=1001

The table 23 shows the results obtained after applying the KNN model with $k = 1001$.

Table 23
SFCR String KNN=1001.

	DC CURRENT	DC VOLTAGE	DC POWER	GRID CURRENT	GRID VOLTAJE	AC POWER	POWER FRECUENCY	REACTIVE POWER	APARENTE POWER
0	0.44	381.8	170	0.75	219.82	101	59.96	655.32	104
1	0.46	388.8	181	0.77	220.53	110	60.05	655.31	110
2	0.48	392.5	192	0.79	220.06	118	59.97	655.34	117
3	0.51	393.4	203	0.84	219.93	130	59.95	655.32	129
4	0.53	394.4	212	0.86	219.96	141	59.99	655.34	139
5	0.56	396.3	224	0.89	219.7	150	59.9	0	153
6	0.59	396.4	235	0.93	220.04	165	59.91	260.51	165
7	0.64	396.4	246	0.98	220.22	175	59.95	161.66	177
8	0.67	375.9	254	1.02	219.97	186	60.05	655.35	187
9	0.69	368.5	267	1.08	219.67	201	59.94	0	203

Score Test of the Least Squares model

0.9908930537105121

MAE of the Least Squares model

25.300842267277332

MSE of the Least Squares model

93.1636282838906

Determination Coefficient of the Least Squares model

0.9908930537105121

Adjusted coefficient of determination of the Least Squares model

0.9908781365175727

MULTIPLE LINEAR REGRESSION MODEL DATA

Slopes or coefficients value "a"

[-0.22626666 -0.06027328 0.42887022 0.88917318 1.42996145 -3.80276292 -
0.00917298 0.55406415]

Intersection or coefficient value "b"

-76.0014904748391

$$Y = -0.22626666 * X_1 - 0.06027328 * X_2 + 0.42887022 * X_3 + 0.88917318 * X_4 \\ + 1.42996145 * X_5 - 3.80276292 \\ * X_6 - 0.00917298 * X_7 + 0.55406415 * X_8 - 76.0014904748391$$

OLS training time

0:00:00.005383

OLS test time

0:00:00.003301

SFCR String KNN=101

The table 24 shows the results obtained after applying the KNN model with $k = 101$.

Table 24
SFCR String KNN=101.

	DC CURRENT	DC VOLTAGE	DC POWER	GRID CURRENT	GRID VOLTAJE	AC POWER	POWER FRECUENCY	REACTIVE POWER	APARENTE POWER
0	0.44	381.8	170	0.75	219.82	101	59.96	655.32	104
1	0.46	388.8	181	0.77	220.53	110	60.05	655.31	110
2	0.48	392.5	192	0.79	220.06	118	59.97	655.34	117
3	0.51	393.4	203	0.84	219.93	130	59.95	655.32	129
4	0.53	394.4	212	0.86	219.96	141	59.99	655.34	139
5	0.56	396.3	224	0.89	219.7	150	59.9	0	153
6	0.59	396.4	235	0.93	220.04	165	59.91	267.97	165
7	0.63	396.4	246	0.98	220.22	175	59.95	147.22	177
8	0.67	375.9	254	1.02	219.97	186	60.05	655.35	187
9	0.69	369.11	267	1.08	219.67	201	59.94	0	203

Score Test of the Least Squares model

0.9907472392548805

MAE of the Least Squares model

25.528777202643436

MSE of the Least Squares model

93.91058903596821

Determination Coefficient of the Least Squares model

0.9907472392548805

Adjusted coefficient of determination of the Least Squares model

0.990732083217624

MULTIPLE LINEAR REGRESSION MODEL DATA

Slopes or coefficients value "a"

[-0.24373962 -0.05901015 0.42736763 0.94022836 1.43345759 -
3.79667806 -0.00897807 0.55533833]

Intersection or coefficient value "b"

-77.4394668320117

$$Y = -0.2437396 * X_1 - 0.05901015 * X_2 + 0.42736763 * X_3 + 0.94022836 * X_4 \\ + 1.43345759 * X_5 - 3.79667806 \\ * X_6 - 0.00897807 * X_7 + 0.55533833 * X_8 - 77.4394668320117$$

El tiempo de entrenamiento OLS es

0:00:00.006649

OLS test time

0:00:00.003501

SFCR String KNN=5

The table 25 shows the results obtained after applying the KNN model with $k = 5$.

Table 25
SFCR String KNN=5.

	DC CURRENT	DC VOLTAGE	DC POWER	GRID CURRENT	GRID VOLTAGE	AC POWER	POWER FREQUENCY	REACTIVE POWER	APARENTE POWER
0	0.44	381.8	170	0.75	219.82	101	59.96	655.32	104
1	0.46	388.8	181	0.77	220.53	110	60.05	655.31	110
2	0.48	392.5	192	0.79	220.06	118	59.97	655.34	117
3	0.51	393.4	203	0.84	219.93	130	59.95	655.32	129
4	0.53	394.4	212	0.86	219.96	141	59.99	655.34	139
5	0.56	396.3	224	0.89	219.7	150	59.9	0	153
6	0.59	396.4	235	0.93	220.04	165	59.91	462.21	165
7	0.61	396.4	246	0.98	220.22	175	59.95	114.2	177
8	0.67	375.9	254	1.02	219.97	186	60.05	655.35	187
9	0.69	374.41	267	1.08	219.67	201	59.94	0	203

Score Test of the Least Squares model

0.9893622026458794

MAE of the Least Squares model

28.10189705937704

MSE of the Least Squares model

100.76496039637755

Determination Coefficient of the Least Squares model

0.9893622026458794

Adjusted coefficient of determination of the Least Squares model

0.9893447779163886

MULTIPLE LINEAR REGRESSION MODEL DATA



Slopes or coefficients value "a"

[-0.12985192 -0.08724189 0.43227787 0.9634608 1.51729094 -
3.80509714 -0.00853119 0.54837856]

Intersection or coefficient value "b"

-83.95184630887456

$$Y = -0.12985192 * X_1 - 0.08724189 * X_2 + 0.43227787 * X_3 + 0.9634608 * X_4 \\ + 1.51729094 * X_5 - 3.80509714 \\ * X_6 - 0.00853119 * X_7 + 0.54837856 * X_8 - 83.95184630887456$$

OLS training time

0:00:00.003638

OLS test time

0:00:00.000899

SFCR String Mean

The table 26 shows the results obtained after applying the Mean model.

Table 26
SFCR String Mean.

	DC CURRENT	DC VOLTAGE	DC POWER	GRID CURRENT	GRID VOLTAJE	AC POWER	POWER FRECUENCY	REACTIVE POWER	APARENTE POWER
0	0.440000	381.800000	170.0	0.75	219.82	101.0	59.96	655.320000	104.0
1	0.460000	388.800000	181.0	0.77	220.53	110.0	60.05	655.310000	110.0
2	0.480000	392.500000	192.0	0.79	220.06	118.0	59.97	655.340000	117.0
3	0.510000	393.400000	203.0	0.84	219.93	130.0	59.95	655.320000	129.0
4	0.530000	394.400000	212.0	0.86	219.96	141.0	59.99	655.340000	139.0
5	0.560000	396.300000	224.0	0.89	219.70	150.0	59.90	0.000000	153.0
6	0.590000	396.400000	235.0	0.93	220.04	165.0	59.91	288.082336	165.0
7	4.689535	396.400000	246.0	0.98	220.22	175.0	59.95	288.082336	177.0
8	0.670000	375.900000	254.0	1.02	219.97	186.0	60.05	655.350000	187.0
9	0.690000	349.931033	267.0	1.08	219.67	201.0	59.94	0.000000	203.0

Score Test of the Least Squares model

0.9552629133000204

MAE of the Least Squares model

72.73616653151974

MSE of the Least Squares model

202.53113502026318

Determination Coefficient of the Least Squares model

0.9552629133000204

Adjusted coefficient of determination of the Least Squares model

0.9551896338787265



MULTIPLE LINEAR REGRESSION MODEL DATA

Slopes or coefficients value "a"

[1.03961346e+00 -2.02991098e-01 4.03067058e-01 1.39782217e-01

7.89579018e+00 -3.81557281e+00 5.19784926e-03 5.46213939e-01]

Intersection or coefficient value "b"

$$Y = 1.03961346 * X_1 - 0.202991098 * X_2 + 0.403067058 * X_3 + 0.139782217 * X_4 + 7.89579018 * X_5 - 3.81557281 * X_6 + 0.00519784926 * X_7 + 0.546213939 * X_8 - 1403.2696010439024$$

OLS training time

0:00:00.003408

OLS test time

0:00:00.002672

SFCR String Median

The table 27 shows the results obtained after applying the Median model.

Table 27
SFCR String Median.

	DC CURRENT	DC VOLTAGE	DC POWER	GRID CURRENT	GRID VOLTAJE	AC POWER	POWER FRECUENCY	REACTIVE POWER	APARENTE POWER
0	0.440	381.8	170.0	0.75	219.82	101.0	59.96	655.32	104.0
1	0.460	388.8	181.0	0.77	220.53	110.0	60.05	655.31	110.0
2	0.480	392.5	192.0	0.79	220.06	118.0	59.97	655.34	117.0
3	0.510	393.4	203.0	0.84	219.93	130.0	59.95	655.32	129.0
4	0.530	394.4	212.0	0.86	219.96	141.0	59.99	655.34	139.0
5	0.560	396.3	224.0	0.89	219.70	150.0	59.90	0.00	153.0
6	0.590	396.4	235.0	0.93	220.04	165.0	59.91	0.04	165.0
7	4.105	396.4	246.0	0.98	220.22	175.0	59.95	0.04	177.0
8	0.670	375.9	254.0	1.02	219.97	186.0	60.05	655.35	187.0
9	0.690	349.3	267.0	1.08	219.67	201.0	59.94	0.00	203.0

Score Test of the Least Squares model

0.9533700323510211

MAE of the Least Squares model

73.47655943785347

MSE of the Least Squares model

206.8277801404784

Determination Coefficient of the Least Squares model

0.9533700323510211

Adjusted coefficient of determination of the Least Squares model

0.9532936523876322



MULTIPLE LINEAR REGRESSION MODEL DATA

Slopes or coefficients value "a"

[9.57734225e-01 -1.97365504e-01 4.10357315e-01 6.62035535e-02

7.89244499e+00 -3.82315852e+00 3.82243232e-03 5.39457390e-01]

Intersection or coefficient value "b"

-1404.930391223308

$$Y = 0.957734225 * X_1 - 0.197365504 * X_2 + 0.410357315 * X_3 + 0.0662035535 * X_4 + 7.89244499 * X_5 - 3.82315852 * X_6 + 0.00382243232 * X_7 + 0.539457390 * X_8 - 1404.930391223308$$

OLS training time

0:00:00.006220

OLS test time

0:00:00.004142

SFCR String Frequent

The table 28 shows the results obtained after applying the Frequent model.

Table 28
SFCR String Frequent.

	DC CURRENT	DC VOLTAGE	DC POWER	GRID CURRENT	GRID VOLTAJE	AC POWER	POWER FRECUENCY	REACTIVE POWER	APARENTE POWER
0	0.44	381.8	170.0	0.75	219.82	101.0	59.96	655.32	104.0
1	0.46	388.8	181.0	0.77	220.53	110.0	60.05	655.31	110.0
2	0.48	392.5	192.0	0.79	220.06	118.0	59.97	655.34	117.0
3	0.51	393.4	203.0	0.84	219.93	130.0	59.95	655.32	129.0
4	0.53	394.4	212.0	0.86	219.96	141.0	59.99	655.34	139.0
5	0.56	396.3	224.0	0.89	219.70	150.0	59.90	0.00	153.0
6	0.59	396.4	235.0	0.93	220.04	165.0	59.91	0.00	165.0
7	0.12	396.4	246.0	0.98	220.22	175.0	59.95	0.00	177.0
8	0.67	375.9	254.0	1.02	219.97	186.0	60.05	655.35	187.0
9	0.69	365.6	267.0	1.08	219.67	201.0	59.94	0.00	203.0

Score Test of the Least Squares model

0.8860421040679903

MAE of the Least Squares model

118.68125592785127

MSE of the Least Squares model

337.2140056701279

Determination Coefficient of the Least Squares model

0.8860421040679903

Adjusted coefficient of determination of the Least Squares model

0.8858554408477904

MULTIPLE LINEAR REGRESSION MODEL DATA

Slopes or coefficients value "a"

[9.65132774 -1.16060437 0.3419143 2.22048602 12.6943602 -
3.65553804 -0.02474979 0.53702671]

Intersection or coefficient value "b"

-2046.1280568123734

$$Y = 9.65132774 * X_1 - 1.16060437 * X_2 + 0.3419143 * X_3 + 2.22048602 * X_4 \\ + 12.6943602 * X_5 - 3.65553804 \\ * X_6 - 0.02474979 * X_7 + 0.53702671 * X_8 - 2046.1280568123734$$

OLS training time

0:00:00.008956

OLS test time

0:00:00.004598

The table 29 show sample data set completed example SFCR String

Table 29

Data set completed example SFCR String.

	DC CURRENT	DC VOLTAGE	DC POWER	GRID CURRENT	GRID VOLTAJE	AC POWER	POWER FRECUENCY	REACTIVE POWER	APARENTE POWER
count	4893	4893	4893	4893	4893	4893	4893	4893	4893
mean	4.705902	349.3624	1538.668	7.410567	220.3644	1482.092	60.16317	288.0693	1497.502
std	8.032087	59.9527	998.5154	15.36621	7.095778	980.4984	12.19608	322.4166	981.9714
min	0	0	0	0	0	0	0	0	0
25%	1.81	335	648	2.84	218.66	610	59.97	0	612
50%	4.11	349.3	1412	6.35	220.79	1395	59.99	0.05	1406
75%	7.49	363.8	2506	11	222.52	2428	60.02	655.33	2445
max	388.1	3101	3180	367	229.72	3028	655.32	655.35	3065

4.4.2.3. SFCR String Score models comparison

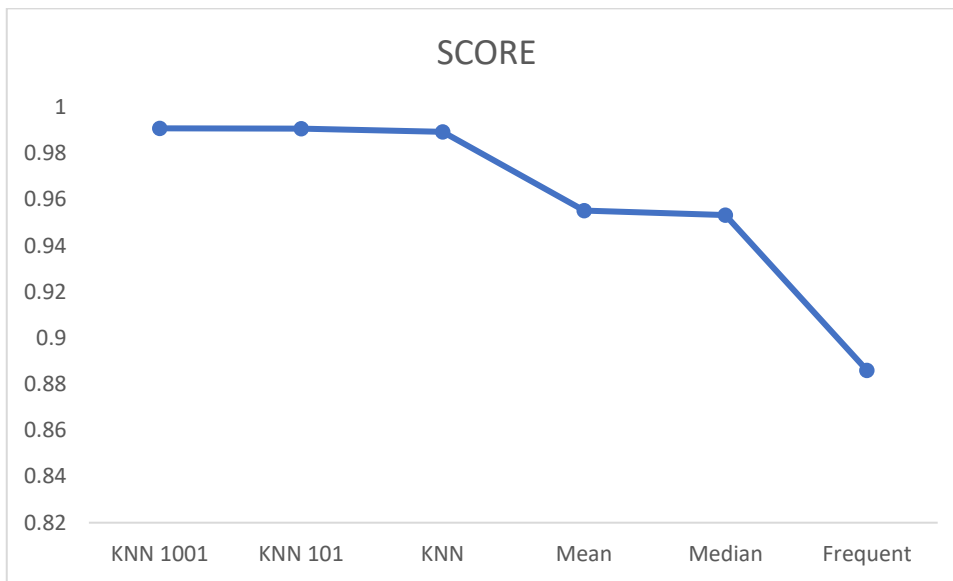


Figure 30
SFCR String Score models comparison.

In the figure 30, the data imputation models that have obtained the best SCORE are the KNN model with $K = 1001$, $K = 101$ and $k = 5$ with 99.09%, 99.07% and 98.93% followed by Mean and Median with 95.53% and 95.34% respectively; the finally we have the Frequent model that reached only 88.60%

The SCORE refers to the score that the model has generated, it indicates the percentage of success when carrying out the imputation of data by applying two models.

4.4.2.4. SFCR String MAE models comparison

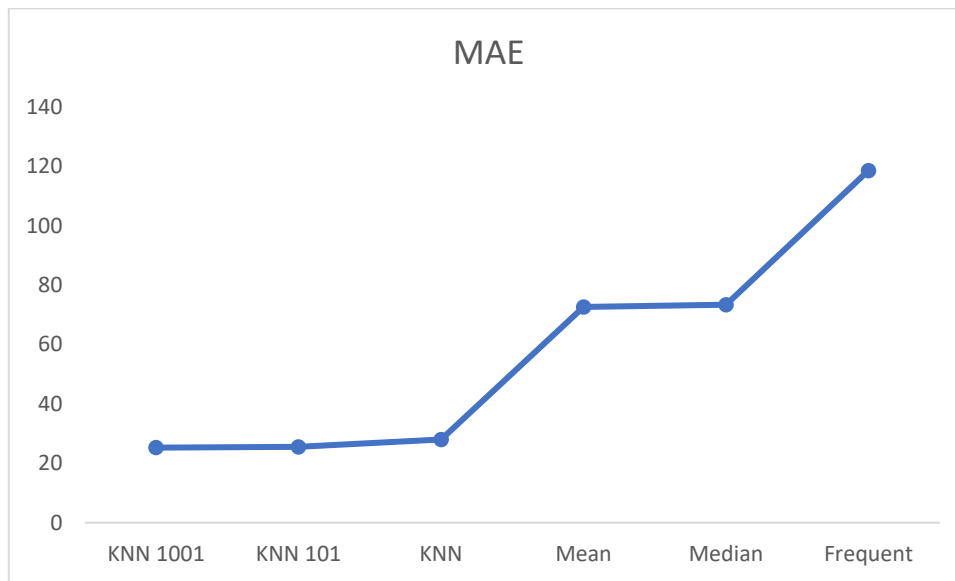


Figure 31
SFCR String MAE models comparison.

In the figure 31, the models that have obtained the lowest MAE are the KNN model with $K = 1001$, $K = 101$ and $k = 5$ with 25.30, 25.52 and 28.1 followed by Mean and Median with 72.74 and 73.48 respectively; the finally we have the Frequent model that reached only 118.68

The Mean Absolute Error (MAE) is the mean of the differences between the target variable and the predictions without the sign, it does not vary much if there are extreme values in the data and the error is interpreted as units of the target variable and is determined by:

$$\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

4.4.2.5. SFCR String MSE models comparison

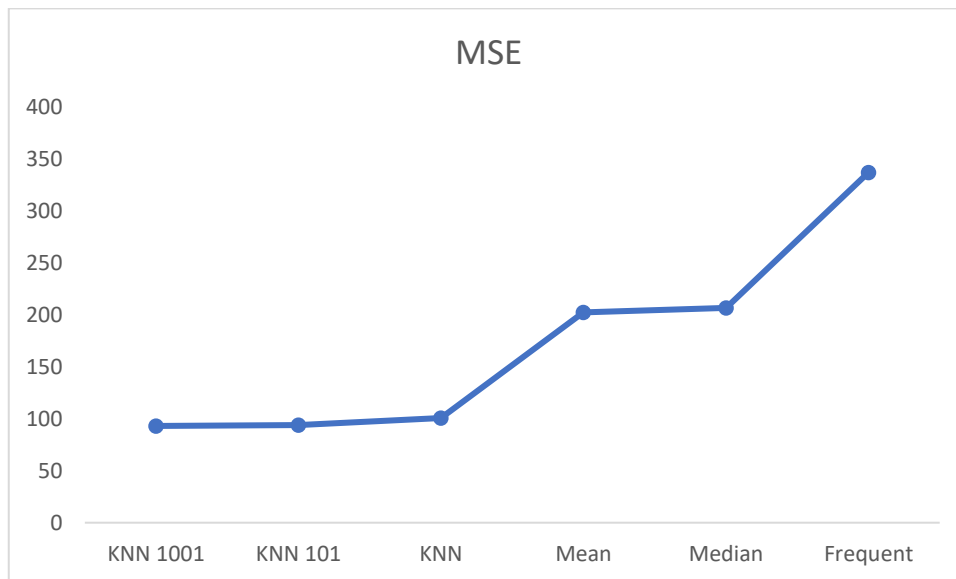


Figure 32
SFCR String MSE models comparison.

In the figure 32, the models with the lowest MSE are the KNN model with $K = 1001$, $K = 101$ and $k = 5$ with 93.16, 93.91 and 100.76 followed by Mean and Median with 202.53 and 206.82 respectively; the finally we have the Frequent model that reached only 337.21

The Mean Square Error (MSE) as an estimator measures the average of the squared errors (the difference between the estimator and what is estimated), emphasizing outliers or extreme values, the error is interpreted as squared units and is determined by:

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

4.4.2.6. SFCR String determination coefficient models comparison

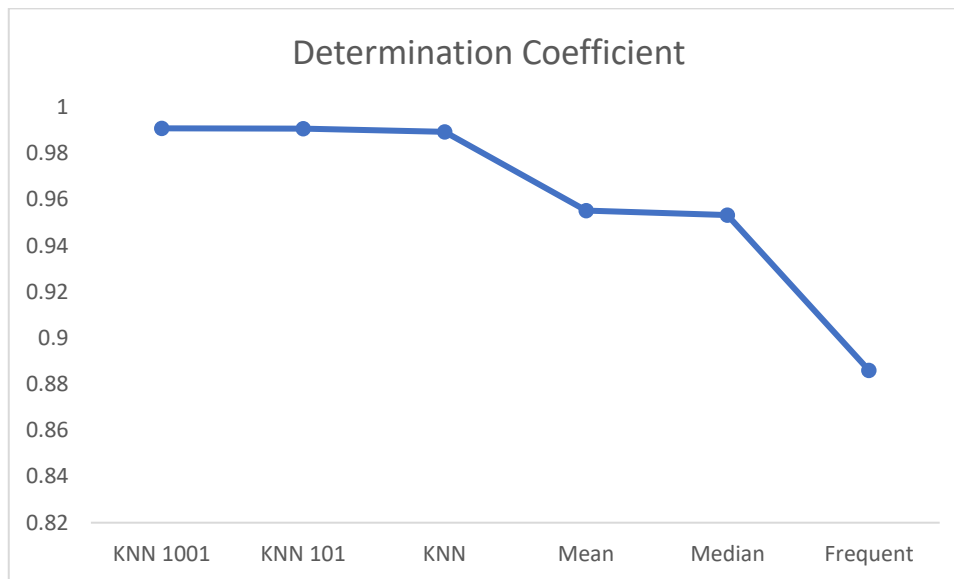


Figure 33

SFCR String determination coefficient models comparison.

As observed in the figure 33, the models that have obtained the highest coefficient of determination R^2 are the KNN model with $K = 1001$, $K = 101$ and $k = 5$ with 0.9909, 0.9907 and 0.9894 followed by Mean and Median with 0.9553 and 0.9534 respectively; the finally we have the Frequent model that reached 0.8860

The determination coefficient (R^2 or R squared) measures the portion of the variance of the objective variable that can be explained by the model, one of the ways to define R^2 is to take the correlation between the objective variable and the predictions, raised to the square and is defined by:

$$r = r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

R^2 : it has a maximum value of 1 when the model explains all the variance, although it could have negative values.

4.4.2.7. SFCR String Adjusted determination coefficient models comparison

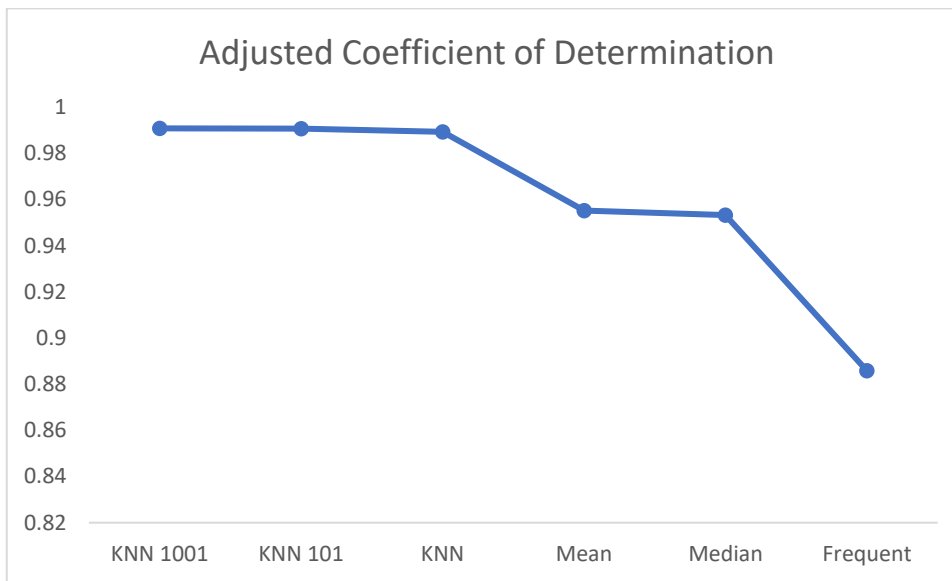


Figure 34

SFCR String Adjusted determination coefficient models comparison.

As can be seen in the figure 34, the models that have obtained the highest adjusted coefficient of determination are the KNN model with $K = 1001$, $K = 101$ and $k = 5$ with 0.9909, 0.9907 and 0.9893 followed by Mean and Median with 0.9552 and 0.9533 respectively; the finally we have the Frequent model that reached 0.8859

Unlike the coefficient of determination, the adjusted coefficient of determination explains whether the model could be overfitting, so it is important to consider this metric that takes into account the complexity of the model and is determined by:

$$1 - \frac{(1 - R^2)(n - 1)}{(n - k - 1)}$$

4.4.2.8. SFCR String Training time models comparison

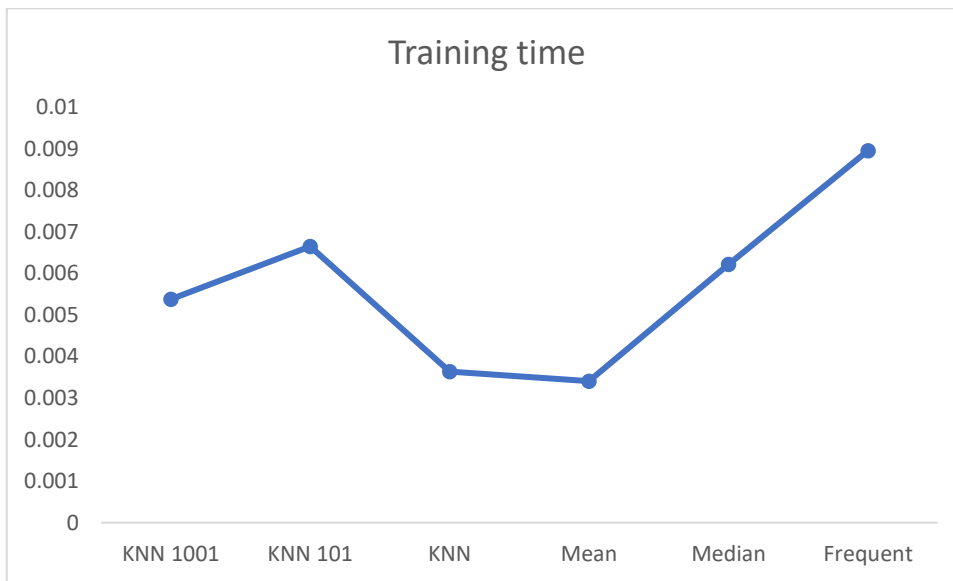


Figure 35
SFCR String Training time models comparison.

In figure 35 we can see that the model for the imputation of Mean is the one with the least processing time with 3.408ms followed by KNN data of $K = 5$ with 3.638ms and then we have KNN with $K = 1001$ of 5.383ms followed by Median model with 6.22ms and KNN of $K = 101$ with 6.649 finally the model that took the longest are the Frequent with 8.956ms.

Training time is the time it takes the algorithm to obtain the model according to the conditions established from the training data.

4.4.2.9. SFCR String Test time models comparison

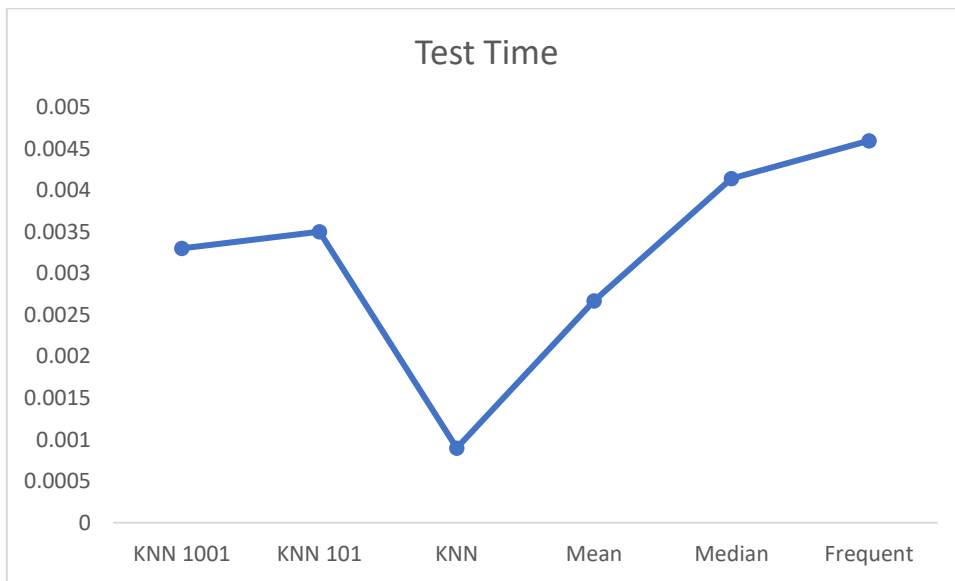


Figure 36
SFCR String Test time models comparison.

In figure 36 we can see that the model for the imputation of KNN data with $K = 5$ are the one that took the least test time with 0.899ms, followed by mean with 2.672ms and KNN with $K = 101$, $K=1001$ with 3.5ms and 3.3ms respectively; Finally, the models that took the longest time was the Median and frequent one with 4.142ms and 4.598ms.

Test time, is the time it takes the algorithm to classify the new values according to the conditions established from the test or test data

4.4.3. String VS Solar Edge data imputation models comparison

While it is true, the validation is looking for the numerical results that quantify the hypothetical relationships between variables in acceptable ranges as descriptions of the data; This technique is generally used to quantify prediction models. If we make a comparison between these techniques, we note that each one has advantages and disadvantages according to the type of data to be treated: some serve to validate models with data sizes or small samples, others for large sample sizes and others focus on model comparison for the most accurate estimation of metrics.

4.4.3.1. SFCR String vs Solar Edge Score models comparison

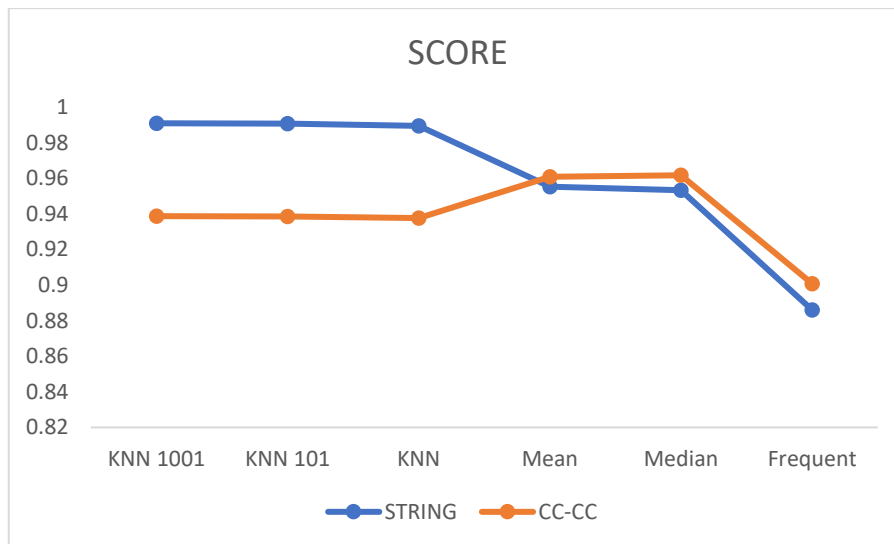


Figure 37

SFCR String vs Solar Edge Score models comparison.

In figure 37 we can see the SCORE of the models for data imputation in the String and Solar Edge (CC-CC) photovoltaic systems, hence we can see that for the SFCR String the best model is the KNN either with $K = 1001$, 101 or 5 since with this model its SCORE is holm oak from 98.93% to 99.01%, however the most optimal for the data imputation for the SFCR Solar Edge are Mean Y Median with a SCORE of 96.08% and 96 On the contrary, 16% in the Frequent model is the one with the lowest performance obtained with both photovoltaic systems with a SCORE of 88.60% for the SFCR String and 90.08% for the SFCR Solar Edge.

4.4.3.2. SFCR String vs Solar Edge MAE models comparison

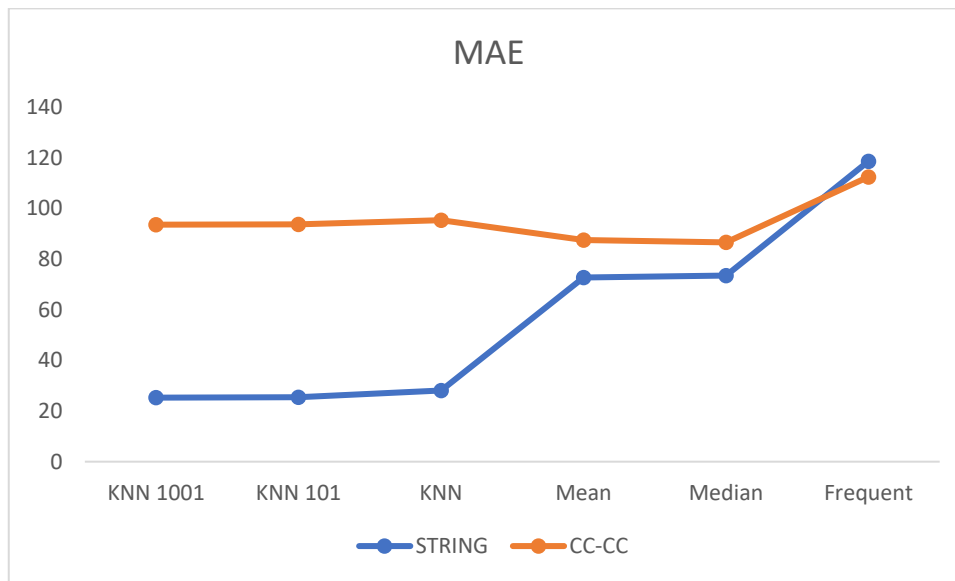


Figure 38
SFCR String vs Solar Edge MAE models comparison.

In figure 38 we can see the comparison of the MAE that the models have generated for the two photovoltaic systems connected to the String and Solar Edge network, it is revealed that for the SFCR String the models with KNN have the lowest error reaching a maximum 28.1, however for the SFCR Solar Edge the models that have presented the least error are Mean and Median reaching a maximum of 87.59; on the contrary, the Frequent model is the one with the worst performance for both systems, reaching a value of 118.68

4.4.3.3. SFCR String vs Solar Edge MSE models comparison

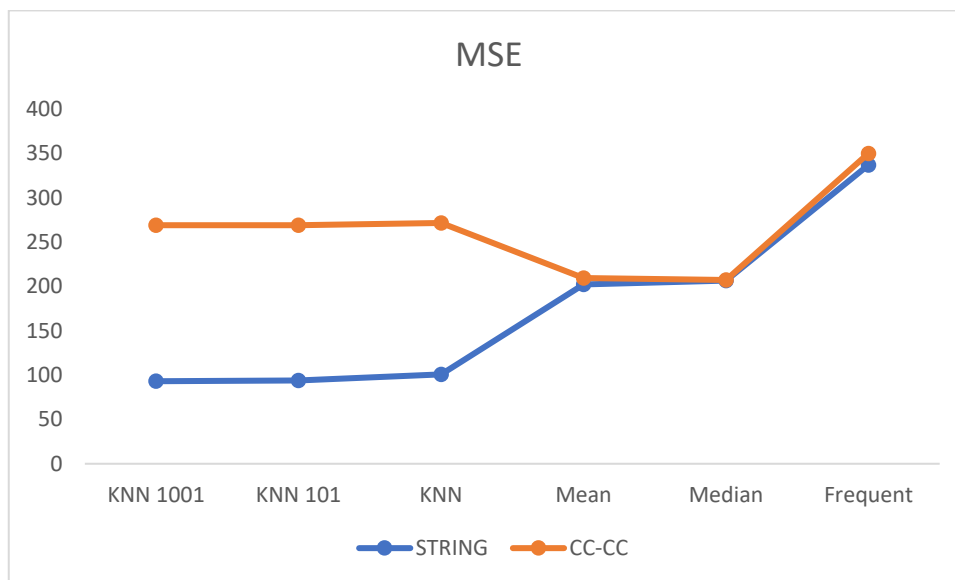


Figure 39
SFCR String vs Solar Edge MSE models comparison.

In figure 39 we can see the comparison of the MSE that the models have generated for the two photovoltaic systems connected to the String and Solar Edge network, it is revealed that for the SFCR String the models with KNN have the lowest error reaching a maximum 100.76, however, for the SFCR Solar Edge, the models that have presented the least error are Mean and Median, reaching a maximum of 209.71, on the contrary, the Frequent model is the one with the worst performance for both systems, reaching a value of 350.35.

4.4.3.4. SFCR String vs Solar Edge determination coefficient models comparison

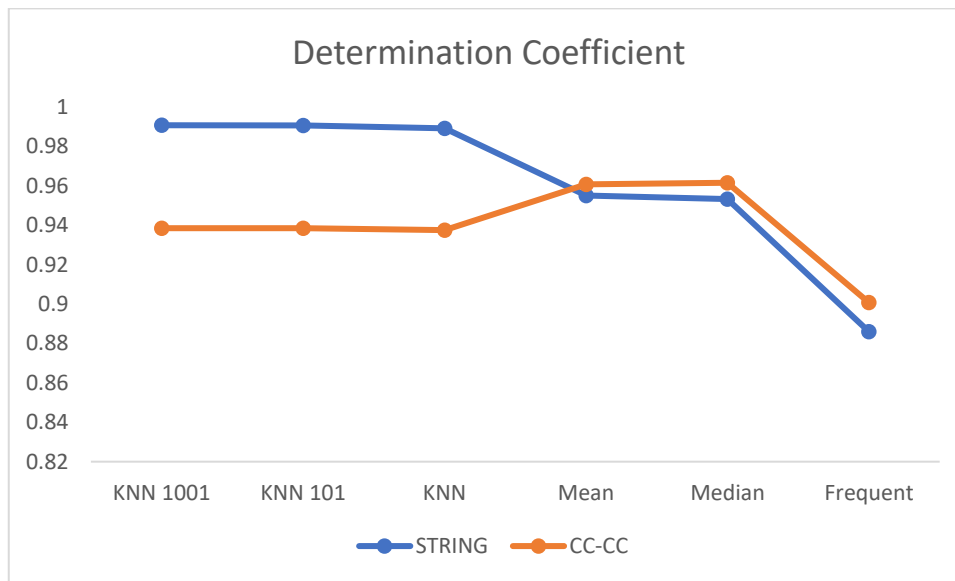


Figure 40
SFCR String vs Solar Edge deter. coefficient models comparison.

As can be seen in figure 40, the comparison of the determination coefficients that the models have generated for the two photovoltaic systems connected to the String and Solar Edge network reveals that for the SFCR String the models with KNN have the closest value to 1 low reaching a maximum of 0.9909, however for the SFCR Solar Edge the models that show a value closer to 1 are Mean and Median reaching a maximum of 0.9616, on the contrary, the Frequent model is the one with the worst performance for both systems reaching a single value of 0.900

4.4.3.5. SFCR String vs Solar Edge Adjusted determination coefficient models comparison

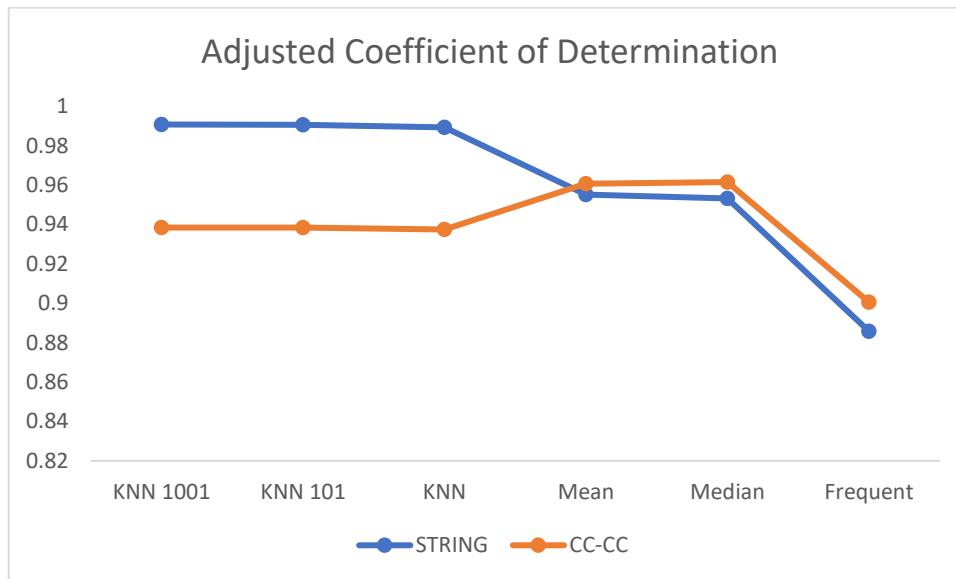


Figure 41
SFCR String vs Solar Edge Adjusted deter. coefficient models comparison.

As can be seen in figure 41, the comparison of the Adjusted coefficient of determination that the models have generated for the two photovoltaic systems connected to the String and Solar Edge network reveals that for the SFCR String the models with KNN have the value closer to 1 low reaching a maximum of 0.9909, however for the SFCR Solar Edge the models that show a value closer to 1 are Mean and Median reaching a maximum of 0.9616, on the contrary, the Frequent model is the one with the worst performance for both systems reaching a single value of 0.900

4.4.3.6. SFCR String vs Solar Edge Training time models comparison

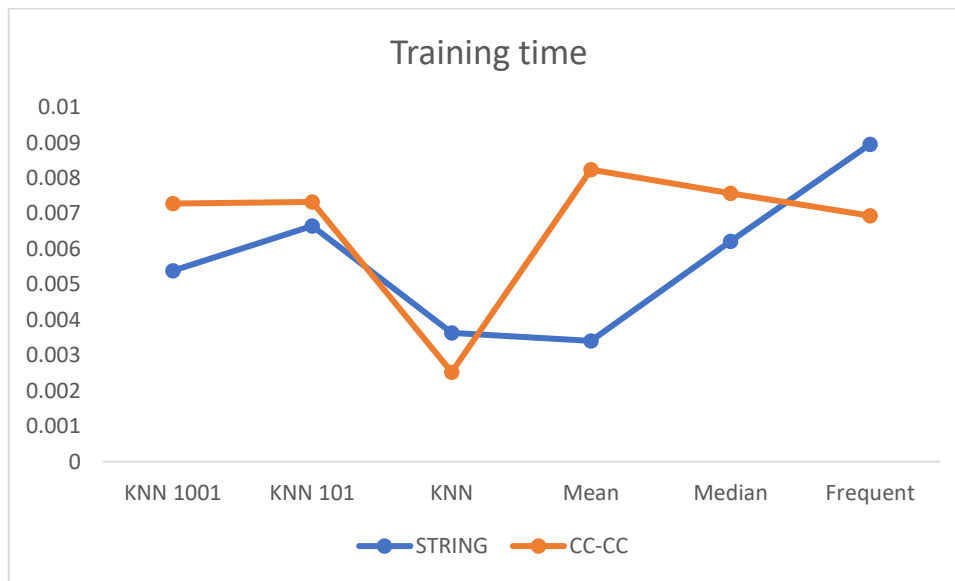


Figure 42
SFCR String vs Solar Edge Training time models comparison.

Figure 42 shows the performance of the models with respect to the training time that I take in each photovoltaic system connected to the String and Solar Edge network,

For the SFCR String the shortest training time was obtained in the Mean model with 3.408ms followed by the KNN model with $K = 5$ which took 3.638ms, however for the SFCR Solar Edge the one that took the least training time was the KNN model with $K = 5$ with 2.0527ms

We can see that the one that takes the most training time is the Frequent model with 8.956ms for the SFCR String and 6.944ms for the Solar Edge system.

4.4.3.7. SFCR String vs Solar Edge Test time models comparison

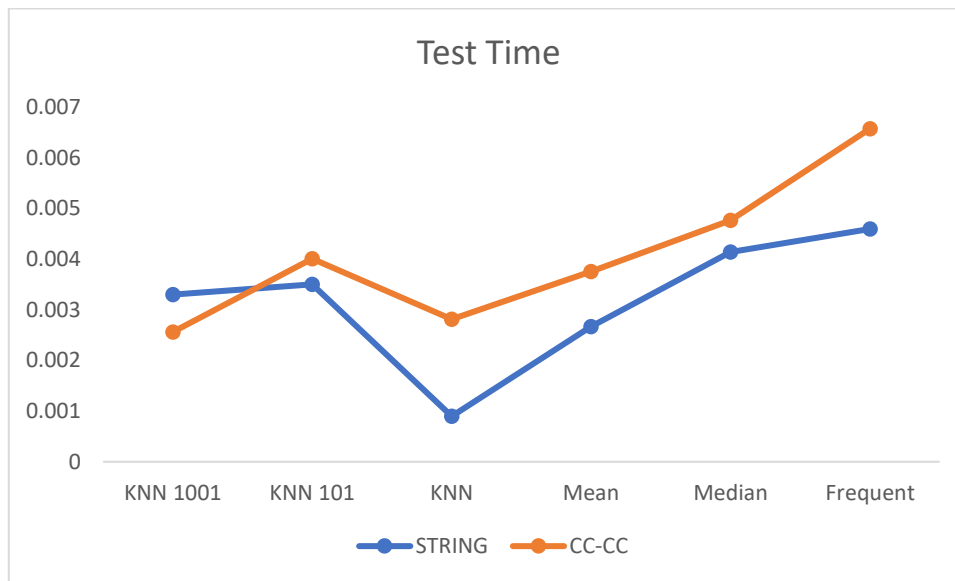


Figure 43
SFCR String vs Solar Edge Test time models comparison.

Figure 43 shows the performance of the models with respect to the test time that it took in each photovoltaic system connected to the String and Solar Edge network,

For the SFCR String, the shortest training time was obtained by the KNN model with $K = 5$ which took 0.899ms followed by Mean with 2.672ms, however for the SFCR Solar Edge the one with the shortest test time was the KNN model with $K = 1001$ with 2,563ms

We can note that the one that takes the longest training time is the Frequent model with 4,598ms for the SFCR String and 6,578ms for the Solar Edge system.

CONCLUSIONS

A data acquisition system was implemented complying with IEC 60904-1 and IEC 61724, Voltage and Current measured with $\pm 0.2\%$ Uncertainty in V_{oc} and I_{sc} , Uncertainty including Instrumentation $<2\%$. With a data reading and logging every 60 seconds.

The data storage is carried out through the fog computing model, achieving a lower latency in the connection, direct access and disregarding the internet because in the area where the project is located does not have stable internet access, the fog computing showed me with the best candidate for the advantages over cloud computing.

The processing for the imputation of missing data was carried out using machine learning as a tool and KNN, Mean, Median and Frequent as Models. KNN is shown as the best model for the imputation of missing data, reaching a SCORE of 99.08%, MAE of 25.3, MSE of 93.16, Coefficient of determination 0.9909, Adjusted coefficient of determination of 0.9907, Training time of 3.638ms and test time of 0.899ms.

BIBLIOGRAPHY

- AG, S. S. T. (2022). *Inversor fotovoltaico / SMA Solar*.
<https://www.sma.de/es/productos/inversor-fotovoltaico.html>
- Anandakumar, H., & Ramu, A. (2020). *Business Intelligence for Enterprise Internet of Things*.
- Beránek, V., Olšan, T., Libra, M., Poulek, V., Sedláček, J., Dang, M. Q., & Tyukhov, I. I. (2018). New monitoring system for photovoltaic power plants' management. *Energies*, 11(10). <https://doi.org/10.3390/en11102495>
- CHASE, O. A. (2018). Arquitetura de uma plataforma de sensoriamento autônoma no contexto de tecnologia ambiental para monitoramento in situ de parâmetros ambientais e da super irradiação solar. In *Universidade Federal do Pará*. UNIVERSIDADE FEDERAL DO PARÁ.
- Chouder, A., Silvestre, S., Taghezouit, B., & Karatepe, E. (2013). Monitoring, modelling and simulation of PV systems using LabVIEW. *Solar Energy*, 91, 337–349. <https://doi.org/10.1016/j.solener.2012.09.016>
- Deng, Y., & Lumley, T. (2021). *Multiple Imputation Through XGBoost*. 2012. <http://arxiv.org/abs/2106.01574>
- Fazai, R., Abodayeh, K., Mansouri, M., Trabelsi, M., Nounou, H., Nounou, M., & Georghiou, G. E. (2019). Machine learning-based statistical testing hypothesis for fault detection in photovoltaic systems. *Solar Energy*, 190(July), 405–413. <https://doi.org/10.1016/j.solener.2019.08.032>
- Harrou, F., Sun, Y., Taghezouit, B., Saidi, A., & Hamlati, M. E. (2018). Reliable fault detection and diagnosis of photovoltaic systems based on statistical monitoring approaches. *Renewable Energy*, 116, 22–37. <https://doi.org/10.1016/j.renene.2017.09.048>
- Huaquipaco, S., Beltran, N., Sarmiento, V., Pizarro, H., Cruz, J., Condori, R., Cutipa, J. R., Romero, C., & Achahuanco, N. (2020). Solar library. *Proceedings of the ISES Solar World Congress 2019 and IEA SHC International Conference on Solar Heating and Cooling for Buildings and Industry 2019*, 2507–2513.

- <https://doi.org/10.18086/swc.2019.52.01>
- Huaquipaco, S., Cruz, J., Beltran Castañon, N. J., Pineda, F., Romero, C., Chura Acero, J. F., & Mamani Machaca, W. (2021). *Modeling And Prediction Of A Multivariate Photovoltaic System, Using The Multiparametric Regression Model With Shrinkage Regularization And Extreme Gradient Boosting*.
<https://doi.org/10.18687/laccei2021.1.1.557>
- Huaquipaco, S., Macêdo, W. N., Pizarro, H., Condori, R., Ramos, J., Vera, O., Cruz, J., Mamani, W., & Beltran, N. (2022). Cross-validation of the operation of photovoltaic systems connected to the grid in extreme conditions of the highlands above 3800 meters above sea level. *International Journal of Renewable Energy Research*, 12(2), 950–959. <https://doi.org/10.20508/ijrer.v12i2.12570.g8490>
- IEC. (2021a). *IEC 60904-1:2020 | IEC Webstore | water management, smart city, rural electrification, solar power, solar panel, photovoltaic, PV, LVDC*.
<https://webstore.iec.ch/publication/32004>
- IEC. (2021b). *IEC 61724-1:2021 | IEC Webstore*.
<https://webstore.iec.ch/publication/65561>
- IEC. (2021c). *International Electrotechnical Commission*. <https://www.iec.ch/homepage>
- Khan, M., Iqbal, J., Ali, M., Muhmmad, A., Zahir, A., & Ali, N. (2019). Designing and implementation of energy-efficient wireless photovoltaic monitoring system. *Transactions on Emerging Telecommunications Technologies*, April, 1–11.
<https://doi.org/10.1002/ett.3685>
- Killam, A. C., Joseph, F., & Stuart, G. (2021). *Article Monitoring of Photovoltaic System Performance Using Outdoor Suns-V OC Monitoring of Photovoltaic System Performance Using Outdoor Suns-V OC*. 1–18.
<https://doi.org/10.1016/j.joule.2020.11.007>
- Krismadinata, Lapisa, R., & Asnil. (2019). A wireless monitoring system for comparison photovoltaic and photovoltaic thermal characteristics. *IOP Conference Series: Materials Science and Engineering*, 602(1). <https://doi.org/10.1088/1757-899X/602/1/012027>

- Lazzaretti, A. E., da Costa, C. H., Rodrigues, M. P., Yamada, G. D., Lexinoski, G., Moritz, G. L., Oroski, E., de Goes, R. E., Linhares, R. R., Stadzisz, P. C., Omori, J. S., & Dos Santos, R. B. (2020). A monitoring system for online fault detection and classification in photovoltaic plants. *Sensors (Switzerland)*, 20(17), 1–30. <https://doi.org/10.3390/s20174688>
- Ma, X., Li, M., Du, L., Qin, B., Wang, Y., Luo, X., & Li, G. (2019). Online extraction of physical parameters of photovoltaic modules in a building-integrated photovoltaic system. *Energy Conversion and Management*, 199(September), 112028. <https://doi.org/10.1016/j.enconman.2019.112028>
- Madeti, S. R., & Singh, S. N. (2017). Monitoring system for photovoltaic plants: A review. *Renewable and Sustainable Energy Reviews*, 67, 1180–1207. <https://doi.org/10.1016/j.rser.2016.09.088>
- Martínez-Camblor, P. (2007). Comparación de pruebas diagnósticas desde la curva ROC. *Revista Colombiana de Estadística*, 30(2), 163–176.
- Medina, F., & Galván, M. (2007). Imputación de datos: teoría y práctica. In *Estudios estadísticos y prospectivos* (Vol. 4). <https://doi.org/978-92-1-323101-2>
- Mekki, H., Mellit, A., & Salhi, H. (2016). Artificial neural network-based modelling and fault detection of partial shaded photovoltaic modules. *Simulation Modelling Practice and Theory*, 67, 1–13. <https://doi.org/10.1016/j.simpat.2016.05.005>
- Mohammed, M. B., Zulkafli, H. S., Adam, M. B., Ali, N., & Baba, I. A. (2021). Comparison of five imputation methods in handling missing data in a continuous frequency table. *AIP Conference Proceedings*, 2355(May). <https://doi.org/10.1063/5.0053286>
- Øgaard, M. B., Riise, H. N., Haug, H., Sartori, S., & Selj, J. H. (2020). Photovoltaic system monitoring for high latitude locations. *Solar Energy*, 207(May), 1045–1054. <https://doi.org/10.1016/j.solener.2020.07.043>
- Ortega, E., Aranguren, G., & Jimeno, J. C. (2019). New monitoring method to characterize individual modules in large photovoltaic systems. *Solar Energy*, 193(September), 906–914. <https://doi.org/10.1016/j.solener.2019.09.099>

- Pazhoohesh, M., Pourmirza, Z., & Walker, S. (2019). A Comparison of Methods for Missing Data Treatment in Building Sensor Data. *2019 IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE), July 2018*, 255–259.
- Ren, Y. (2021). Python Machine Learning : Machine Learning and Deep Learning With Python , Scikit-Learn, and TensorFlow 2. *International Journal of Knowledge-Based Organizations*, 11(1), 67–70.
- Rezk, H., Tyukhov, I., Al-Dhaifallah, M., & Tikhonov, A. (2017). Performance of data acquisition system for monitoring PV system parameters. *Measurement: Journal of the International Measurement Confederation*, 104, 204–211. <https://doi.org/10.1016/j.measurement.2017.02.050>
- Sabry, A. H., Hasan, W. Z. W., Ab. Kadir, M. Z. A., Radzi, M. A. M., & Shafie, S. (2018). Wireless monitoring prototype for photovoltaic parameters. *Indonesian Journal of Electrical Engineering and Computer Science*, 11(1), 9–17. <https://doi.org/10.11591/ijeecs.v11.i1.pp9-17>
- Samara, S., & Natsheh, E. (2019). Intelligent Real-Time Photovoltaic Panel Monitoring System Using Artificial Neural Networks. *IEEE Access*, 7, 50287–50299. <https://doi.org/10.1109/ACCESS.2019.2911250>
- Sha, W., Dai, C., & Jiang, L. (2019). Design of patrol monitoring and control system for hot spot of solar photovoltaic module. *Proceedings - 2019 International Conference on Intelligent Computing, Automation and Systems, ICICAS 2019*, 668–671. <https://doi.org/10.1109/ICICAS48597.2019.00145>
- Shariff, F., Rahim, N. A., & Hew, W. P. (2015). Zigbee-based data acquisition system for online monitoring of grid-connected photovoltaic system. *Expert Systems with Applications*, 42(3), 1730–1742. <https://doi.org/10.1016/j.eswa.2014.10.007>
- Slapšak, J., Mitterhofer, S., Topic, M., & Jankovec, M. (2019). Wireless System for in Situ Monitoring of Moisture Ingress in PV Modules. *IEEE Journal of Photovoltaics*, 9(5), 1316–1323. <https://doi.org/10.1109/JPHOTOV.2019.2918044>
- SolarEdge. (2021). *SolarEdge | Un líder mundial en soluciones smart energy*. <https://www.solaredge.com/es>



- Sun, X., Su, Y., Huang, Y., Tan, J., Yi, J., Hu, T., & Zhu, L. (2020). Design and development of a wireless sensor network time synchronization system for photovoltaic module monitoring. *International Journal of Distributed Sensor Networks*, 16(5). <https://doi.org/10.1177/1550147720927341>
- Xia, K., Ni, J., Ye, Y., Xu, P., & Wang, Y. (2020). A real-time monitoring system based on ZigBee and 4G communications for photovoltaic generation. *CSEE Journal of Power and Energy Systems*, 6(1), 52–63. <https://doi.org/10.17775/cseejpes.2019.01610>
- Xu, X., & Qiao, D. (2011). Remote monitoring and control of photovoltaic system using wireless sensor network. *2011 International Conference on Electric Information and Control Engineering, ICEICE 2011 - Proceedings*, 633–638. <https://doi.org/10.1109/ICEICE.2011.5778367>
- Yahyaoui, I., & Segatto, M. E. V. (2017). A practical technique for on-line monitoring of a photovoltaic plant connected to a single-phase grid. *Energy Conversion and Management*, 132, 198–206. <https://doi.org/10.1016/j.enconman.2016.11.031>