FEDERAL UNIVERSITY OF CEARÁ

TECHNOLOGY CENTER

MECHANICAL ENGINEERING DEPARTMENT

POSTGRADUATE PROGRAM IN MECHANICAL ENGINEERING

NADJA GOMES DE OLIVEIRA

EVALUATION OF MACHINE LEARNING MODELS FOR SOLAR IRRADIANCE PREDICTION (GHI and DNI): A CASE STUDY IN PETROLINA, PE, BRAZIL

FORTALEZA

2022

NADJA GOMES DE OLIVEIRA

EVALUATION OF MACHINE LEARNING MODELS FOR SOLAR IRRADIANCE PREDICTION (GHI and DNI): A CASE STUDY IN PETROLINA, PE, BRAZIL

Master's dissertation presented to the Postgraduation Program in Mechanical Engineering from the Federal University of Ceará, as a partial requirement to obtaining the title of master's in Mechanical Engineering. Concentration area: Processes, Equipment and Systems for Renewable Energy.

Advisor: Prof. Dr. Paulo Alexandre Costa Rocha

FORTALEZA

2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

G615e Gomes de Oliveira, Nadja.
Evaluation of machine learning models for solar irradiance prediction (ghi and dni): a case study in Petrolina, PE, Brazil / Nadja Gomes de Oliveira. – 2022.
61 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia Mecânica, Fortaleza, 2022.
Orientação: Prof. Dr. Paulo Alexandre Costa Rocha.

1. machine learning. 2. global solar irradiance. 3. direct normal irradiance. 4. intra-hour forecasting. 5. caret R package. I. Título.

CDD 620.1

NADJA GOMES DE OLIVEIRA

EVALUATION OF MACHINE LEARNING MODELS FOR SOLAR IRRADIANCE PREDICTION (GHI and DNI): A CASE STUDY IN PETROLINA, PE, BRAZIL

Master's dissertation presented to the Postgraduation Program in Mechanical Engineering from Federal University of Ceará, as a partial requirement to obtaining the title of master's in Mechanical Engineering. Concentration area: Processes, Equipment and Systems for Renewable Energy.

Approved in: 29/07/2022

EXAMINATION BOARD

_____

Prof. Dr. Paulo Alexandre Costa Rocha (Advisor)
Universidade Federal do Ceará (UFC)

_____

Prof. Dr. André Valente Bueno
Universidade Federal do Ceará (UFC)

_____

Prof. Dr. Matheus Pereira Porto
Universidade Federal de Minas Gerais (UFMG)

To God.

To my parents Juraci and Jeane.

ACKNOWLEDGMENTS

To FUNCAP, for the financial support with the scholarship grant.

To Prof. Dr. Paulo Alexandre Rocha, for his excellent guidance.

To the professors participating in the examining board for their time, valuable contributions, and suggestions.

To the National Spatial Research Institution (INPE) and the National Organization System of Environmental Data (SONDA) for the data base used in the dissertation and for their academic contribution for students and professors.

To my parents (Jeane and Juraci) and my grandmother (Mundinha) for all the incentive and support during my life.

ABSTRACT

This work uses the SONDA network irradiance data to forecast global horizontal and direct normal irradiances (GHI and DNI) intra-hourly applying 5min and 60min forecast window resolution and five different time horizons (5min, 30min, 60min, 6 hours and 12 hours) during the period of four years for a solarimetric and anemometric station in the northeast of Brazil, Petrolina/PE. Five different machine learning models were tested, namely: Multivariate Adaptive Regression Splines (MARS), Least Absolute Shrinkage and Selection Operator (LASSO), k-nearest neighbors (kNN), Extreme Gradient Boosting (XGBoost) and an ensemble combination to form a final forecast (Ensemble with Ridge Regression). Their performance was compared using the RMSE and forecast skill (FS) relative to the smart persistence model. Results show that the machine learning models achieve significant forecast improvements over the reference model using only endogenous features. In addition, the Ensemble with Ridge Regression and XGBoost models have rarely been used for very short-term solar forecasting according to the literature. This framework can be used to select appropriate machine learning approaches for very short-term solar power forecasting and the simulation results can be used as a baseline for comparison. The XGBoost's forecast skill model was not the winner in all time horizons and resolutions, but it is among the best results for GHI and DNI, with normalized variables. The XGBoost model prevails when the time resolution of 5 min is chosen, not considering other error metrics, such as MBE. It is worth to mention, for the time resolution of 5 min, that the XGBoost model has the best FS results in 66.66% of the time comparing to all the six results for GHI and DNI with raw and normalized variables. For the time resolution of 60 min, the MARS model has the best forecast skill's results, dominating around 66.66% of all the outputs, including GHI and DNI for raw and normalized variables. Also, kNN is the Machine Learning model with the best outputs of MBE, proving that the model is more accurate and does not have huge estimations variations comparing to the other models.

Keywords: Machine learning, global solar irradiance, direct normal irradiance, intra-hour forecasting, Caret R package.

RESUMO

Este trabalho usa os dados de irradiância da rede SONDA para prever irradiância global horizontal e normal direta (GHI e DNI) intra-hora, aplicando 5 min e 60 min como resolução do intervalo de previsão e cinco horizontes de tempo diferentes (5min, 30min, 60min, 6 horas e 12 horas), durante o período de quatro anos em uma estação solarimétrica e anemométrica no nordeste do Brasil, Petrolina/PE. Cinco modelos diferentes de aprendizado de máquina foram testados: Multivariate Adaptive Regression Splines (MARS), Least Absolute Shrinkage and Selection Operator (LASSO), k-nearest neighbors (kNN), Extreme Gradient Boosting (XGBoost) e a combinacão das previsões de diversos modelos para formar um resultado final (Ensemble com Regressão Ridge). Seu desempenho foi comparado usando o RMSE e a habilidade de previsão (FS) em relação ao modelo de persistência inteligente. Os resultados mostram que os modelos de aprendizado de máquina alcançam melhorias significativas de previsão em relação ao modelo de referência usando apenas variáveis endógenas. Além disso, os modelos Ensemble com Regressão Ridge e XGBoost raramente têm sido usados para previsão solar de muito curto prazo de acordo com a literatura. Essa estrutura pode ser usada para selecionar abordagens de aprendizado de máquina apropriadas para previsão de energia solar de muito curto prazo e os resultados da simulação podem ser usados como linha de base para comparação. A habilidade de previsão do modelo XGBoost não foi o vencedor em todos os horizontes temporais e resoluções, mas está entre os melhores resultados para GHI e DNI, com variáveis normalizadas. O modelo XGBoost prevalece quando a resolução temporal de 5 min é escolhida, não considerando outras métricas de erro, como MBE. Vale ressaltar, para a resolução temporal de 5 min, o modelo XGBoost apresenta os melhores resultados da habilidade de previsão em 66,66% das vezes comparando com todos os seis resultados para GHI e DNI com variáveis brutas e normalizadas. Para a resolução temporal de 60 min, o modelo MARS apresenta os melhores resultados para FS, dominando cerca de 66,66% de todas as saídas, incluindo GHI e DNI para variáveis brutas e normalizadas. Além disso, kNN é o modelo de aprendizado de máquina com os melhores resultados do MBE, comprovando que o modelo é mais preciso e não possui grandes variações de estimativas em relação aos demais modelos.

Palavras-chave: Aprendizado de máquina, irradiância solar global, irradiância normal direta, previsão intra-hora, pacote Caret R.

LIST OF FIGURES

LIST OF GRAPHICS

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

A	Anemometric
ABSOLAR	Brazilian Association of Photovoltaic Solar Energy
ANN	Artificial Neural Network
ARIMA	Autoregressive Moving Averages
BEN	National Energy Balance
BNEF	Bloomberg New Energy Finance
CE	Ceará
CSP	Concentrated Solar Thermal Power
DNI	Direct Normal Irradiance
EPA	Energy Planning Agency
FS	Forecast Skill
G	Solar Irradiance
GHI	Global Horizontal Irradiance
IEA	International Energy Agency
INPE	National Institute For Space Research
kNN	K-Nearest Neighbors' Algorithm
ktDNI	Clear Sky Index with Direct Normal Irradiance
ktGHI	Clear Sky Index with Global Horizontal Irradiance
LASSO	Least Absolute Shrinkage and Selection Operator
MAE	Mean Absolute Error
MARS	Multivariate Adaptive Regression Splines
MBE	Mean Bias Error
nMAE	Normalized Mean Absolute Error
nMBE	Normalized Mean Bias Error
NWP	Numerical Weather Prediction
PE	Pernambuco
PV	Photovoltaic
RES	Renewable Energy Sources
RMSE	Root Mean Square Error
rRMSE	Relative Mean Squared Error

S	Solarimetric
s	Skill
SA	Solarimetric And Anemometric
SONDA	National Data Organization System
XGBoost	Extreme Gradient Boosting
ZCIT	Intertropical Convegence Zone

LIST OF SYMBOLS

%	Percentage
°	Degrees
GW	Giga Watt
GWp	Giga Watt Peak
$Iclr$	Clear-Sky Irradiance
k_t	Clearness Index
kWh	Kilowatt Hour
$l(\cdot)$	Loss Function
m^2	Square Metre
W	Watt
δ	Solar Declination
θz	Zenith Angle
φ	Latitude Values
ω	Clockwise Angle
$\Omega(f_t)$	Penalty Function

SUMMARY

1 INTRODUCTION

According to Brazil (2022), the solar energy opened the year 2022 with immense potential. The Federal Government sanctioned the law n°14,300, which creates the legal framework for distributed generation from renewable sources in Brazil, considered a strategic step for the legal security of the market and consumers. With it, generating and consuming their own clean, renewable, and competitive electricity becomes a right of every citizen, small business, and rural producer in the country. Since the beginning of the 2000s, there has been an important advance towards non-hydro renewable sources. These clean and competitive sources are based on renewable resources widely abundant in several Brazilian regions, especially in the northeast, which has the best solar and wind resources in the country.

The global shift towards renewable energy sources (RES) has driven the development of photovoltaic (PV) panels. For example, the costs of producing electricity from PV panels have dropped significantly, while simultaneously increasing the energy conversion efficiency. More specifically, the levelized cost of electricity of largescale PV panels has decreased by 73% between the years of 2010 and 2017 (IRENA, 2018). The decreased cost and increased efficiency have made PV panels a competitive alternative as a RES in many countries (Bessa and Andrade, 2017). However, since PV panel energy output depend on weather conditions such as cloud cover and solar irradiance, the energy output of the PV panels is unstable. To understand and manage the output variability is of interest for several actors in the energy market. In the short-term (0-5 hours), a transmission system operator is interested in the energy output from PV panels to find the adequate balance for the whole grid. The profitability of these operations relies on the ability to forecast the fluctuating solar PV panel energy output accurately.

One of the subfields of artificial intelligence - machine learning - has been used for solar irradiation studies as verified in Diagne et al. (2013), Qing et al. (2018) and Marquez et al. (2018). ML techniques have the improved computational capacity and a higher availability of quality data that made this technique very useful for forecasting solar energy. Nowadays, PV power forecasting based on the AI algorithm is a very popular research area because of its strong self-learning and self-adaptation ability. In the literature (Kaushika et al., 2014), the PV array generation sequence, weather type, irradiance intensity, and temperature are adopted to build the backpropagation (BP)

neural network prediction model. But this method requires a large number of historical power data and massive calculation (Ye et al., 2022).

According to Pedro et al., (2018), works related to the production of very short-term probabilistic solar functions using different types of techniques are relatively recent. As with punctual deterministic changes, most of these works were dedicated to GHI and are based on endogenous predictors. Endogenous use solar variables to forecast (GHI, DNI). On the other hand, exogenous uses other variables for prediction (rainfall, wind speed, etc.). The knowledge of solar irradiance components on the surface has a great importance and interest of the scientific community, mainly because Brazil is a country of great territorial extent (DE SOUZA JUNIOR et al., 2020).

In this study, solar predictions of global horizontal irradiance and direct normal irradiance were made for horizons of 5, 30 and 60 minutes, 6 and 12 hours a posteriori through the application of machine learning models in data sets of four years collected by the SONDA Network station data base in Petrolina/PE, Brazil.

The objective is to evaluate whether the use of endogenous attributes related to GHI and DNI, the use of information from the irradiance values of past instants, as well as whether the use of certain filters (zenith angle), caret package and clear-sky index in the data set provide an improvement in the accuracy of the machine learning algorithms used, namely: Multivariate Adaptive Regression Splines (MARS), Least Absolute Shrinkage and Selection Operator (LASSO), k-nearest neighbors (kNN), Extreme Gradient Boosting (XGBoost) and an ensemble combination to form a final forecast (Ensemble with Ridge Regression). The performance evaluation of the models was carried out by calculating the Root Mean Square Error (RMSE), Root Mean Square Error Normalized (nRMSE), Mean Error by Bias (MBE), Mean Absolute Error (MAE), Mean Absolute Error Normalized (nMAE) and Forecast Skill (FS).

2 OBJECTIVES

2.1 General Objective

The main objective of this dissertation is to evaluate the performance of different forecasting techniques with machine learning models for global horizontal and direct normal irradiance in the locality of Petrolina/PE situated in the Northeast of Brazil. This work covers five machine learning models such as MARS, LASSO, XGBoost, kNN and Ensemble with Ridge Regression, hypothesizing that the use of endogenous predictors can produce equal or superior results compared to results using sky images with exogenous predictor.

2. 2 Specific Objectives

1. Build the databases analyzed by carrying out collections from 01/01/2013 to 31/12/2016 in Petrolina/PE, where they included global solar and direct normal irradiance values with the zenith angle <= 85.
2. Assess the percentage of importance of the zenith angle as predictor.
3. Implement the machine learning algorithms used in the R programming language.
4. Compare the forecasting performance across two different time resolutions (5 min and 60 min).

2.3 Research Question

Based on our introductory discussion, the problem can be summarized by the following research question:

- How good machine learning techniques perform in very short-term for 5 min, 30 min, 60 min, 6 hours and 12 hours' time horizons and a time resolution of 5 min and 60 min forecasting of Global Horizontal and Direct Normal irradiance output using only endogenous predictors in the Northeast of Brazil?

3 LITERATURE REVIEW

3.1 Solar Energy Forecasting

For the challenges of this millennium, solar energy is one of the most promising energy alternatives and it has taken an increasingly important part, which will continue to rise, driven by carbon peaking and carbon neutrality strategic goals. As reported by the International Energy Agency (IEA), photovoltaic solar energy can represent a third of the global electricity production of the world by 2060 (IEA, 2011). In 2016, the Bloomberg New Energy Finance (BNEF) studies indicate that photovoltaic solar energy will represent more than 25% of the global electrical matrix by 2040, therefore, in less than 25 years. The Brazilian Energy Planning Agency´s (EPA) Energy Expansion Plan (EEP), from 2019 to 2029, indicates that renewable sources will remain a high priority, targeting 48% of Brazil's energy matrix by 2029. In a tropical country like Brazil, this potential is even more possible and viable. According to the projections of BNEF, photovoltaic energy will represent around 32% of the Brazilian electricity matrix in 2040, with an installed capacity between 110 and 126 GWac.

As a result, photovoltaics has the potential to be the largest source of electricity in the world in the long term, due to the abundance and distribution of the solar resource on the planet, constant reduction of technology costs and improvements in efficiency of materials and conversion. Therefore, the technical potential of photovoltaic energy in Brazil is enormous, greater than the sum of the technical potential of all other energy sources in the country. According to Ye et al. (2022), due to the intermittence and volatility of sunlight, photovoltaic (PV) power generation is more erratic than conventional power which results in some problems of the grid: frequency instability (Liu et al., 2020; Murty and Kumar, 2020), dispatch difficulty (Peng et al., 2020; Tummala, 2020), and voltage and current surges (Bozorg et al., 2020; Yang et al., 2021b). Hence, accurately forecasting the power generation of the PV system is one of the major issues of PV system’s engineering practice to settle the aforementioned problems (Huang et al., 2021; Yang et al., 2021).

The focus of solar forecasting is to provide a basis for plant scheduling and planning transactions in the electricity market in order to balance the supply and demand of power generation and ensure reliable operation. These changes are used by utilities, transmission system operators, energy service providers, energy traders and independent energy producers in their scheduling,

dispatch and energy demand. Furthermore, different methodologies for solar irradiance predictions have been proposed for various time horizons. Statistical models with measured irradiance on the spot are suitable for a very short time scale ranging from 5 min to 6 h. Predictions based on motion vectors of the satellite image clouds show good performance over a time interval of 30 min to 6 h (Lorenz and Heinemann, 2012). For forecasting horizons of about 6 h onwards, changes based on numerical weather prediction (NWP) models are generally more accurate (Inness and Dorling, 2012; Maini and Agrawal, 2006; Muselli et al., 1998).

Accurate solar forecasts over several time horizons are required so that Independent System Operators (ISOs) or equivalent grid balancing authorities are able to successfully integrate increased levels of solar power production while maintaining reliability. Solar forecasts on multiple time horizons become increasingly important as solar penetration grows for the purposes of grid regulation, load-following production, power scheduling and unit commitment. Short-term, intra-hour solar forecasts are particularly useful for power plant operations, grid balancing, real-time unit dispatching, automatic generation control (AGC) and trading. Forecasts for longer time horizons are of interest to utilities and ISOs for unit commitment, scheduling and for improving balance area control performance. Ultimately, a spectrum of solar forecasts is required to address the planning, operational and balancing needs of both the distribution and the transmission grids.

Solar forecasting is therefore an enabling technology for the integration of ever increasing level of solar penetration into the grid because it improves the quality of the energy delivered to the grid and reduces the ancillary costs associated with weather dependency. The combination of these two factors (better energy quality through information that is capable of lowering integration and operational costs) has been the driving motivation for the development of a complex field of research that aims at producing better solar forecasting capabilities for the solar resource at the ground level and for the power output from different solar technologies that depend on the variable irradiance at the ground level. Solar, wind, and load forecasting have become integral parts of the so-called ‘smart grid concept’ (Inman et al., 2013).

To date, high-fidelity, robust solar forecast systems that work for widely different microclimates remain evasive. The problem is of great complexity due to the non-linear and chaotic effect of cloud motion on solar irradiance at the ground level. However, a number of promising approaches have been developed in the past few years, and the incipient research field of solar meteorology for renewable generation has grown considerably by aggregating diverse areas of

knowledge such as atmospheric physics, solar instrumentation, machine learning, forecasting theory and remote sensing in its quest to better predictive skills. This work presents an overview of the forecasting methods for solar resourcing and solar power generation, as well as the theoretical basis for the most promising methods, and a discussion on their effectiveness for operational use (Inman et al., 2013).

According to the modeling means of prediction, the prevailing PV power prediction methods are broadly divided into three categories, namely, physical, statistical, and artificial intelligence (AI) forecasting technologies (Yang et al., 2021). The PV power forecasting technologies face different challenges. First, it is difficult for physical forecasting technology to obtain accurate future weather forecast information and determine output characteristic model parameters. Second, statistical forecasting technology is not demanding for geographical location and other information of PV systems but requires masses of historical data to deduce statistics laws. As for AI forecasting technology, it is easy to trap in the local optimum because of internal defects of the AI algorithm (Ye et al., 2022).

3.1.1 Climate Context of the Northeast of Brazil

This section presents a brief description of the typical climatic conditions of the Northeast region in Brazil and its unique characteristics. The expected large-scale integration of solar energy with existing energy supply structures - regulated by national authorities – is expected to significantly increase the importance of meteorological and climate information due to its strong impact on planning and operation of energy generation and distribution systems. The availability and variability of the solar energy's resource is intrinsically associated with weather conditions and climate of a region. This is because weather systems cause changes in cloudiness and concentrations of gases and aerosols, affecting the radiative processes that attenuate the solar radiation along its path through the atmosphere.

According to the data from the Brazilian Solar Energy Atlas of INPE - National Institute for Space Research (Pereira, 2006), Brazil has an excellent solar resource, which varies between 1,500 and 2,350 kWh/m²/year. It is a well distributed resource around the country, higher than in countries such as Germany (900 to 1,250 kWh/m²/year), France (900 to 1,650 kWh/m²/year) and even Spain (1,200 to 1,850 kWh/m²/year). The states with the highest levels of solar radiation in

Brazil are Bahia, Piauí, Paraíba, Rio Grande do Norte, Ceará, Tocantins, Goiás, Minas Gerais and São Paulo.

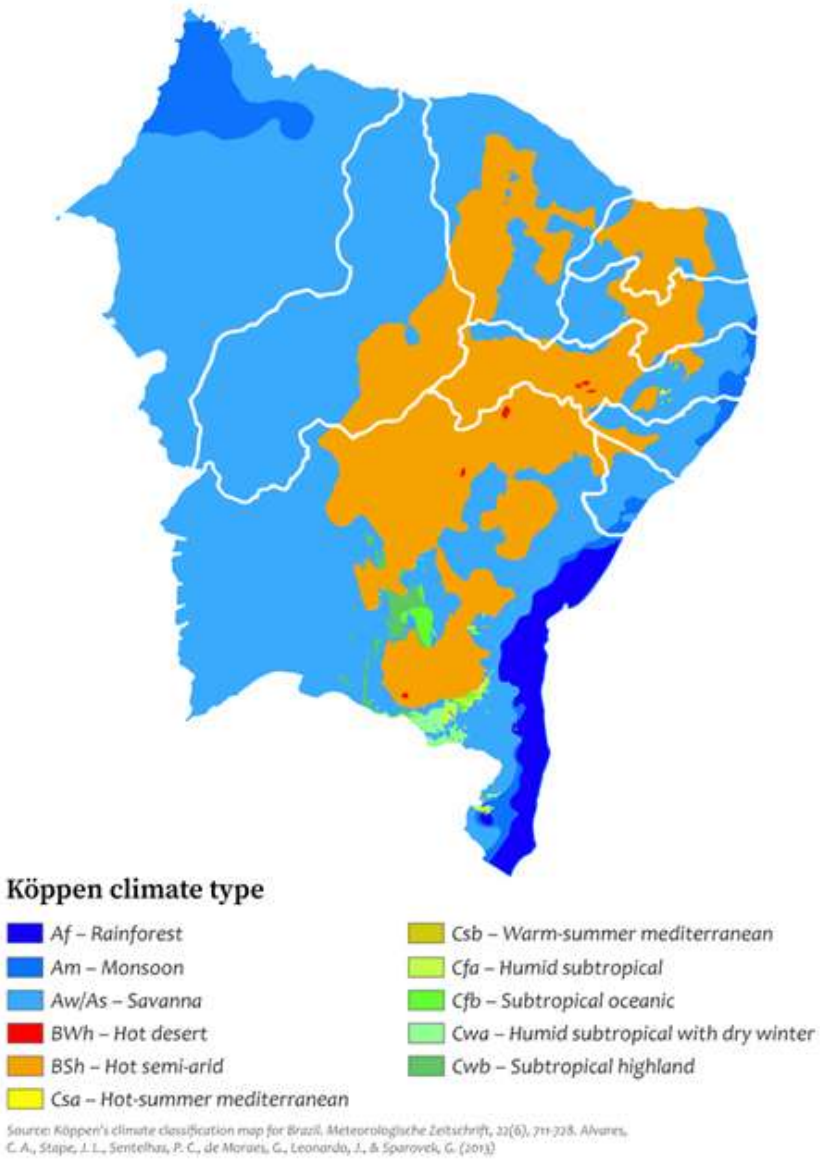
Brazil's climate is diverse because of several factors, such as the territorial extension, the geographical region and the dynamics of the air masses. The Brazilian's geographical region has a direct influence on the weather and climate conditions of a region. Higher points tend to be colder, in addition to creating favorable conditions for the formation of cloudiness through condensation by air lifting on the slopes. The atmospheric dynamics has a solid importance because it acts directly on temperature and precipitation, causing the regional climate differences. Figure 1 illustrates the distribution of climates characteristics in the Northeast of the Brazilian territory according to Köppen (Vianello and Alves, 2013). It can be noted that the largest parts of the Northeast of Brazil present the savanna and the semiarid climate.

The Northeast Region of Brazil, characterized as a semiarid region, presents scarce rainfall, and it is frequently affected by long periods of drought (Althoff et al., 2016; Awange et al., 2016). The annual accumulated precipitation does not exceed 500 mm in some areas of the Northeast semiarid; in contrast, there are areas like the coastland where the annual rainfall is more than 1500 mm. The Northeast region has areas with characteristics quite distinct from each other, its southern portion being under the influence of semi-stationary frontal systems, prefrontal systems, local convection sea and land breezes on the coast.

In the coastal plane, which runs from Rio Grande do Norte to the south of Bahia, the main mechanisms are the breeze activity together with the maximum convergence of trades and disturbances east ripples. Eventually, the displacement of the ZCIT - Intertropical Convergence Zone (Zona de Convergência Intertropical) on the east coast of the Northeast. According to Tuohy et al. 2015, the process of solar forecasting for various time horizons, methods and applications has many similarities to wind forecasting, but solar output is strongly linked to cloud cover. In general, the stratiform clouds and shallow convective clouds are the most frequent in the Northeast of Brazil, but the associated rainfall is not as abundant as precipitation caused by deep convective clouds.

It is also seen that a strong signal of shallow convective clouds modulates rainfall over the coastal areas of Brazil Northeast and adjacent ocean (Palharini et al., 2017). Stratiform clouds have an important effect on climate as they cover about 34% of the ocean and 18% of the land surface at any given time (Heymsfield, 1993).

Figure 1 – Climate classification for Brazil Northeast according to Köppen.



Source: Vianello and Alves, 2013.

3.2 Predictors

Concentrated solar thermal power (CSP) and photovoltaic (PV) power plants are the two main means of generating electricity from the solar resource: Concentrated thermal solar power systems convert heat from direct solar irradiance into steam that is used in a power cycle, such as

the Rankine cycle, to generate electricity. However, photovoltaic devices generate electricity by taking advantage of both the direct and diffuse components. Thus, direct solar irradiance predictions are of greater relevance in the context of solar thermal power systems, while global solar irradiance predictions are of greater importance in applications involving the use of photovoltaic devices (BENALI et al., 2019).

Direct normal solar irradiance forecasts show lower accuracy than those performed for global solar irradiance according to the work of Marquez and Coimbra (2011) and Pedro and Coimbra (2018) because of the high error metrics results for RMSE. In the first work, for a forecast horizon of 24 hours a posteriori, values of the relative mean squared error (rRMSE) were found for the prediction of GHI ranging from 14.8% to 19.3% and from 28.1% to 35% for DNI. In the second work, where forecasts were carried out in a horizon of up to 30 min ahead without the use of sky images, there was a reduction in the RMSE that varied between 8% and 20.4%, and 10.3% and 26.6% for the set of GHI and DNI tests, respectively.

In addition, other authors (TRAPERO et al., 2015) applied models based on time series analysis: exponential smoothing in state space (DONG et al., 2013), integrated autoregressive moving averages (ARIMA) (BOX, 1994) and a dynamic harmonic regression (YOUNG et al., 1999). The best results presented an rRMSE of 29.66% for GHI forecasts and 46.79% for DNI in horizons of 1 hour up to 1 day ahead. Again, by evaluating these results, there is an indication that DNI predictions are less reliable than the those made for GHI.

Benalli et al. (2019) performed predictions for the GHI, DNI and diffuse DHI components at intervals of 1 hour up to 6 hours ahead where rRMSE values were found in the range of 19.65% to 27.78%, 34.11% to 49.08 % and 35.08% to 49.14% for the predictions of the GHI, DNI and diffuse components, respectively, thus illustrating that diffuse solar irradiance predictions are even less reliable than the ones for DNI. The work in question focused on making GHI predictions.

Table 1 – Previous works incorporating machine learning and ensemble models for GHI and DNI forecast. Only the best results for FS (%) and RMSE (W/m²) are listed.

Study	Location	Time Horizon	Time Resolution	Parameters	Proposed method	Benchmark method	GHI		DNI	
							FS	RMSE	FS	RMSE
Urraca et al. (2016)	Spain	1h	30 min	Extraterrestrial irradiance, solar azimuth angle, solar elevation angle, solar hour angle, and the cosine of solar zenith angle	Random Forest	Pers	16.7	92.47	-	-
Pedro and Coimbra (2018)	Folsom (USA)	5 min	5 min	Pyranometer measurements of GHI and DNI and sky images	GB with images	SP	13.3	32.7	14.3	58.2
		30min					23.6	34.4	29.6	71.3
Hassan et al. (2017)	Africa	1h	Hourly/Daily	Solar elevation angle, global solar irradiance, diffuse fraction, global clearness index, normal clearness index, extraterrestrial horizontal irradiance, diffuse irradiance, daily global clearness index and the sunshine fraction	GB, Bagging, RF	MLP,SVR, DT	-	88.75	-	-
Kumari and Toshniwal (2021)	New Delhi (India)	1h	-	The latest hourly value of meteorological parameters (temperature, relative humidity, pressure, wind speed and direction), time information (hour of the day) and clear-sky index	XGBF-DNN	SP, SVR, RF	40.2	51.35	-	-

Source: elaborated by the author.

Table 1 – Previous works incorporating machine learning and ensemble models for GHI and DNI forecast. Only the best results for FS (%) and RMSE (W/m²) are listed.

Study	Location	Time Horizon	Time Resolution	Parameters	Proposed method	Benchmark method	GHI		DNI	
							FS	RMSE	FS	RMSE
Yang et al. (2020)	Chengde (China)	1h	15 min	Satellite Images	Combination of NWP-statistical methods-ANNs (FY-4A)	SP	12.2	180.93	0.4	278.61
Huertas-Tatoa et al. (2020)	Jean (Spain)	15 min-6h	15 min	Satellite-based model, WRF-Solar, SP and CIADCast	SVMs Radial General	Satellite-based model, WRF-Solar, SP and CIADCast)	16.19 (average)	29.19 (normalized)	13.33 (average)	45.15 (normalized)
	Lisbon (Portugal)						16.21 (average)	41.94 (normalized)	14.46 (average)	73.99 (normalized)
	Madrid (Spain)						15.04 (average)	32.89 (normalized)	10.19 (average)	56.19 (normalized)
	Seville (Spain)						19.14 (average)	27.67 (normalized)	17.09 (average)	42.54 (normalized)

Source: elaborated by the author.

According to Table 1, Pedro and Coimbra (2018), Kumari and Toshniwal (2021), Yang et al. (2020) and Huertas-Tatoa et al. (2020) used the Smart Persistence (SP) as a Benchmark Method in previous studies for time horizons between 5 min to 6 hours. Pedro and Coimbra (2018), Yang et al. (2020) and Huertas-Tatoa et al. (2020) apply satellite images as parameters, obtaining FS results for GHI ranging from 12.2% to 23.6% and RMSE between 32.7 W/m² and 180.93 W/m². Nevertheless, other studies use images as parameters, such as Urraca et al. (2016), Kumari and Toshniwal (2021) and have similar or superior FS results for GHI, varying from 16.7 % to 51.9% and RMSE between 51.35 W/m² and 106.13 W/m².

The best FS results were obtained by Kumari and Toshniwal (2021) in the location of India, although the smallest RMSE were achieved by Huertas-Tatoa et al. (2020) and Pedro e Coimbra (2018) in locations such as USA, Spain, and Portugal.

3.2.1 Global Horizontal Irradiance and Direct Normal Irradiance

The information set out in this section about irradiance concepts is important for our understanding and use in solar irradiance forecast. All physical, chemical, physicochemical and biological phenomena responsible for maintaining life in the Earth's atmosphere system are directly or indirectly linked to the amount of solar irradiance incident on the planet.

Part of the solar radiation is scattered, and part is absorbed by particles and molecules present in the air, such as water vapor, carbon dioxide, ozone, and nitrous compounds. Solar irradiance, when crossing the atmosphere, undergoes complex interactions with atmospheric constituents through the processes of absorption and scattering of incident radiation. Even though the atmosphere is very transparent, it is estimated that only 25% of the incident radiation at the top of the atmosphere reaches the Earth's surface without suffering any interference from atmospheric constituents.

The remaining 75% are absorbed, reflected back to space or scattered and, in this case, normally reach the surface in a direction different from the direction of incidence at the top of the atmosphere (LIOU, 2002). Solar radiation is treated as the total amount of energy emitted by the Sun whereas, solar irradiance refers to the amount of solar radiation received from the Sun per unit area. Solar irradiance (G) is the rate at which radiant energy is incident on a surface per unit area, usually given in W/m^2. This is constituted by diffuse and direct solar radiation, being influenced by some factors such as solar elevation, optical depth conditions and degree of cloudiness (ALVES, 1981).

The Direct Normal Irradiance (DNI) is the solar irradiance generated by radiant energy from the Sun without its scattering in the atmosphere. When the sky is clear, direct irradiance corresponds to 60 to 87% of global irradiance (LESTRADE et al. 1990). In the presence of cloudiness, solar irradiance decreases, as cloudiness and solar elevation are first-order factors in determining the variation of solar irradiance at the surface (KONDRATYEV, 1969).

Diffuse irradiance is the solar radiation received from the sun after changes in its direction by the dispersion of the atmosphere. The diffuse radiation incident on an inclined plane is composed by both the irradiance reflected from the ground and the diffuse radiation from the sky. Global Horizontal Irradiance (GHI) is the total solar irradiance incident on a horizontal surface,

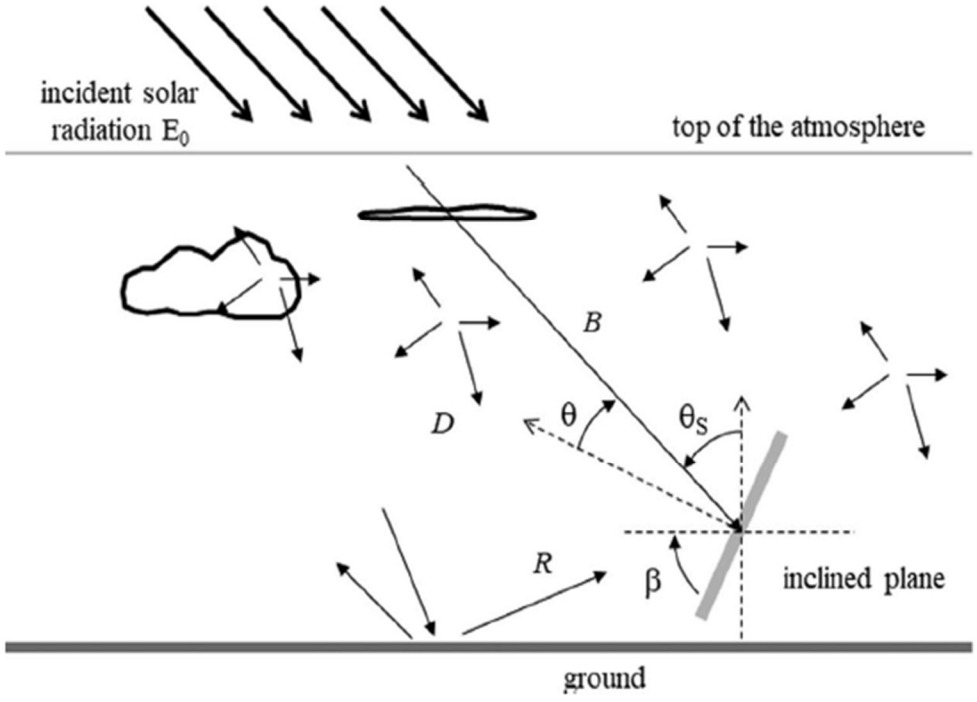
which is the sum of direct normal irradiance (DNI), diffuse horizontal irradiance and irradiance reflected on the ground (DUFFIE, BECKMAN, 2013).

3.2.2 Zenith angle

The solar zenith angle (θ_z) represents the angle formed between the vertical at the observation point and the direction of the line that connects the same point on the Earth's surface to the Sun. It can be calculated by knowing the location's latitude values (φ), the solar declination (δ) and the solar clockwise angle (ω). The zenith angle is equal to 90° when the Sun is on the horizon at sunrise or sunset (PEREIRA et al., 2017).

Figure 1 illustrates the irradiance as it hits the atmosphere, as well as the types of solar radiation described above. A schematic view of the direct beam (B), diffuse (D) and reflected (R) components of the radiation received by an inclined plane on the ground whose inclination is β can also be seen. θz is the solar zenith angle and θ is the angle formed by the direction of the Sun and the direction normal to the plane of incidence.

Figure 2 – Solar irradiance components.



Source: Wald (2021).

In the present dissertation, the main interest is on global horizontal irradiance (GHI) and direct normal irradiance (DNI), the global horizontal irradiance being the sum of the direct solar irradiance multiplied by the cosine of the zenith angle and the diffuse solar irradiance.

3.2.3 Clearness and Clear Sky Indices

Duffie and Beckman (2013) affirm that another common practice in solar energy is to work with the clear sky index k_t instead of the original solar irradiance time series. The clear-sky index is defined as:

$$k_t(t) = \frac{I(t)}{I^{clr}(t)} \tag{1}$$

where I is the solar irradiance, GHI or DNI, and I^{clr} is the clear-sky irradiance. In this dissertation it was computed following the algorithm given by Marquez and Coimbra (2008).

Another evolution of the clearness index concept was made possible with the development of proficient clear sky radiation modeling (Bird and Hulstrom, 1981). The main use of the clear-sky index is the removal of diurnal and seasonal signals from a given set of radiation data to apply advanced analysis techniques (Woyte et al., 2007), or to calculate power fluctuations (Lave et al., 2011a). This method has been used in more modern assessments of solar variability for solar energy purposes (Lave and Kleissl, 2010; Lave et al., 2011a; Hoff and Perez, 2010) and as an input and output of machine learning-based predictions of solar radiation (Sfetsos and Coonick, 2000; Yang et al., 2012; Benghanem and Mellit, 2010).

Its main use has been in classifying cloud types (Calbo' et al., 2001; Pages et al., 2003), in numerical climate modeling based on predictions (Mathiesen and Kleissl, 2011) and in calculating derived irradiance estimates from satellites (Zarzalejo et al., 2009). Algorithms that use endogenous variables as input data are applied in much of the research that performs solar irradiance predictions using machine learning methods (PEDRO et al., 2018), (MEJIA et al., 2018), (MUNKHAMMAR et al., 2018), which are related to solar irradiance values for previous moments, as well as the current one. The efficiency of these prediction learning models increases when they are applied to historical series with a stationary behavior.

The clearness index, k_t, is a common parameter derived from GHI, which reduces the potential for the non-stationary introduction of statistical approaches to the diurnal cycle of irradiance and seasonality (VOYANT et al., 2015). The k_t value for any given time and specific location on Earth is defined as the ratio of global solar irradiance measured at ground level G and its counterpart estimated at the top of the atmosphere G_0 (Liu and Jordan, 1960), as indicated by Equation 3.

$$k_t = G/G_0 \tag{2}$$

The k_t index is also known in the literature for expressing the cloudy condition of the sky (DAL PAI; ESCOBEDO, 2015). Low values of k_t indicate a large presence of clouds, or low global solar radiation compared to extraterrestrial radiation. High k_t values indicate clear sky conditions or little cloudiness. In other words, the clear-sky index serves as an indication of atmospheric conditions, showing more clearly the variations in global radiation as a function of climate. For a better detailing of these parameters, Duffie & Beckman (2013) is recommended.

3.3 Mathematical Models

3.3.1 Persistence

Smart Persistence is an improved version of Persistence which assumes the sky conditions will remain constant (instead of irradiance itself). The forecasting algorithm predicts the current solar radiation as the product of the current clear sky ratio (k_t) and the clear sky radiation (I_{cs}) in the predicted point. It is widely used as the baseline for validating more complex models and is fairly accurate in short-term horizons (TATO and BRITO, 2018).

The persistence model is often used as a reference for determining the FS and is a useful baseline model for short term forecasts. It is convenient to know if a forecast model provides better results than any trivial reference model, which in our case is the persistence model. The persistence model considers that the solar radiation at $t + 1$ is equal to the solar radiation at t. It assumes that the atmospheric conditions are stationary. It is also called a naive predictor.

$$G_{t+1} = G_t \tag{3}$$

Its accuracy decreases with the time horizon and is generally not adequate for more than 1 h. An improved version of this model is the smart persistence model. To consider the fact that the apparent position of the Sun is not identical between t and $t + 1$, the persistence model is corrected with a clear-sky ratio term and is then called smart persistence.

3.3.2 The LASSO Regression method

Ridge and LASSO regression are some of the simplest techniques to reduce model complexity and prevent over-fitting that may result from simple linear regression models by adding a penalty and shrinking the beta coefficients. It completely relies on the L1 penalty, which can reduce the coefficients' sizes so small that they can get to 0, leading to automatic feature selection (features with a 0 coefficient do not influence a model). Since λ (the "strength" of the penalty) can and should be tuned, a stronger(larger) penalty leads to more coefficients pushed to zero.

The LASSO is a relatively recent alternative to ridge and not only helps in reducing overfitting, but it can help in feature selection. The cost function for LASSO (Least Absolute Shrinkage and Selection Operator) regression can be written:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j| = RSS + \lambda \sum_{j=1}^{p}|\beta_j| \quad (4)$$

The LASSO coefficients, $\hat{\beta}_\lambda^L$, minimize the quantity and as with ridge regression, but the LASSO can shrink the coefficient estimates towards zero. In this sense, in the case of the LASSO, the $\ell 1$ penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large. Hence, much like the best subset selection, the LASSO automatically performs variable selection. As a result, models generated from the LASSO are generally much easier to interpret than those produced by ridge regression (James et al., 2013). For further understanding on the shrinkage and selection procedures of the LASSO, we refer the readers to consult Efron et al., 2004.

3.3.3 The k-Nearest Neighbours (kNN) method

The kNN model uses the predictors introduced above to forecast GHI and DNI. This model is established on the similarity of the predictors at the forecasting issuing time to the predictors

computed with the training data set. The kNN algorithm starts by computing the Euclidean distance for a new data set (i.e. testing or validation) and the features in the training set. This operation yields a distance vector for each feature. These are then combined into a single vector using a weighted sum, denoted as Ds, where the subscript s indicates the set of features used in the calculations. The algorithm proceeds to extract the k instances in the training data with the lowest distance. To each instance there is an associated time stamp $\{\tau_1, \ldots, \tau_2\}$ in the training set. k forecasts are then computed using the GHI or DNI training data subsequent to these time stamps:

$$\hat{f}i(t+\Delta t) = \langle kt \rangle_{[t-\Delta t,t]}\, x\, \langle I^{clr} \rangle_{[t,t+\Delta t]}\,,\ \mathrm{i} = 1, \ldots, \mathrm{k} \tag{5}$$

from which the final point forecast is calculated as:

$$\hat{I}(t+\Delta t) = \frac{\sum_{i=1}^{k} \alpha i \hat{f}i\,(t+\Delta t)}{\sum_{i=1}^{k} \alpha i} \tag{6}$$

where the weights α are the function of the distance Ds

$$\alpha \mathrm{i} = \left(\frac{1-Ds,i}{maxDs-minDs}\right)^{n},\ \mathrm{i} = 1, \ldots, \mathrm{k} \tag{7}$$

and n is an adjustable positive integer parameter. The algorithm summarized above depends on several parameters:

1. The number of nearest neighbors, k ∈ {1,2, … , max k}, where maxk = 150 in this case;
2. The set of features S, i.e., which features are used in the search for the nearest neighbors;
3. The weights in the weighted sum Ds denoted as ωi;
4. The exponent n ∈ {1,2, … , 5} for the weights ai in Eq. (7);

The optimal model is determined by minimizing the forecast error for the validation data set:

$$argmin_{k,S,w,n} \sqrt{\frac{1}{n}\sum_{i}^{n} \left(\hat{I}\,(ti+\Delta t, k, S, \omega, n) - I(ti+\ \Delta t)\right)^2} \tag{8}$$

Further details about the optimization procedure and the respective optimal kNN models for GHI and DNI can be found in Pedro and Coimbra (2018).

3.3.4 The gradient boosting method

According to Friedman (2002) and Hastie et al. (2009), boosting is a general approach that can be applied to many statistical learning methods for regression or classification. For regression problems, given a training data set, the goal is to find a function f(x) such that a specified loss function is minimized. Boosting approximates f(x) by an additive expansion of the form:

$$\hat{f}(x) \sum_{m=0}^{M} \beta_m \, h(x, \theta_m) \tag{9}$$

where the functions $h(x, \theta_m)$ are simply functions of x parameterized by θ_m. $h(x, \theta_m)$ called "base learners" or "weak learners" (Friedman, 2002).

The expansion coefficients β_m and the parameters θ_m are fit to the training data in a forward "stage wise" manner (i.e., without adjusting the previous expansion coefficients and parameters of the base learners that have already been added). Here, we restrict the application of boosting to the context of regression trees (i.e., the base learner $h(x, \theta)$ is a tree $T(\theta)$). For that purpose, boosting builds an ensemble of trees iteratively to optimize a loss function ψ: the squared loss function $\psi\big(y, f(x)\big) = (y - f(x))^2$, in this case.

3.3.5 The XGBoost method

XGBoost is an algorithm based on a sequential ensemble of decision trees, in which weak learners learn together to build a strong learner. Equation 10 shows the algorithm for the XGBoost method given by Munawar et al. (2019). Since the loss function *l(·)* for calculating residual is hard to optimize, the cost function $L^{(t)}$ is introduced as follows (Chen et al., 2016):

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, y'^{(t-1)}_i + f_i(x_i)) + \Omega(f_t) \tag{10}$$

where y_i is the index and t is time, y_i is the actual data and $y'^{(t)}_i$ is the forecasted value, $f_i(x_i)$ is the model being updated itera-tively. $\Omega(f_t)$ is the penalty function and l(·) is the loss function.

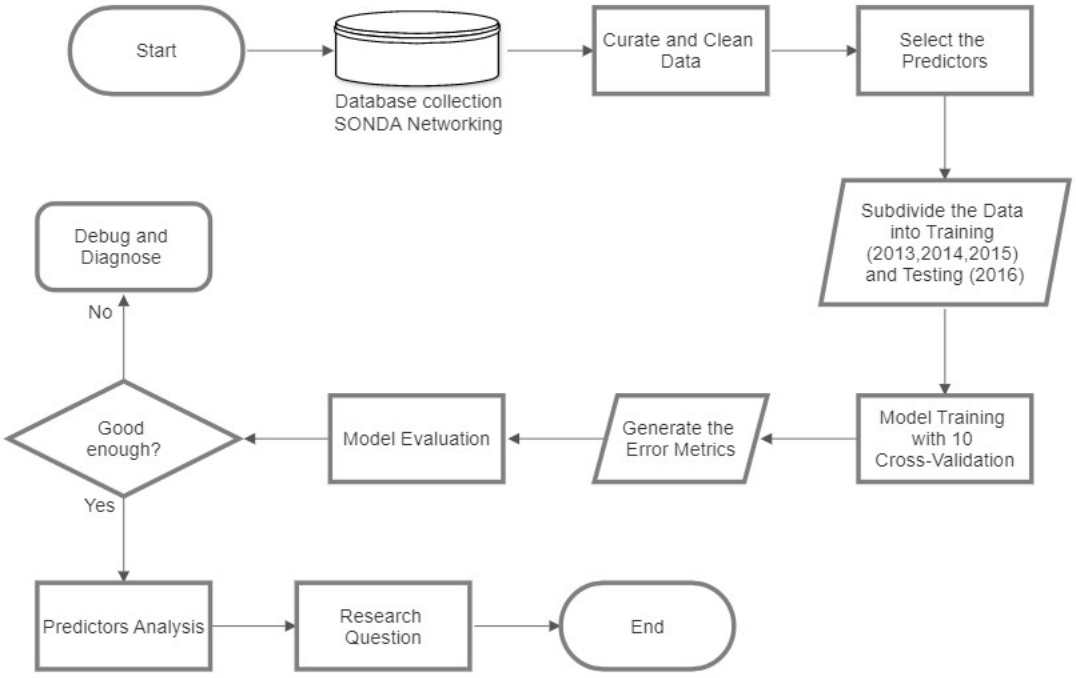
4 METHODOLOGY

This methodology section has been divided into 5 subsections in order to facilitate its explanation. In the first stage, the data will be collected from SONDA network station, encompassing the Northeast region of Brazil, as well as the period of time to be analyzed. In a second step, data filtering and the choice of parameters that will be used in the subsequent steps occurs.

In the third step, the algorithms of each machine learning model will be trained, extending the time horizons on 5 min intervals. The time horizons ranged from 5 min to 60 min; so, a time horizon of 30 min indicates a 30-min-ahead forecasting process.

In the fourth step, we generated the results of the error metrics for each machine learning model. To illustrate the main steps, Figure 3 shows the stages of the dissertation methodology.

Figure 3 – Methodology flowchart.



Source: Elaborated by the author.

4.1 SONDA Data Network

All the data were obtained from SONDA project (National Data Organization System, www.sonda.ccst.inpe.br). This data network was born from a project by the National Institute for Space Research (INPE) for the implementation of physical infrastructure and human resources aimed at surveying and improving the database of solar and wind energy resources in Brazil. The SONDA data collection network has measurement stations distributed throughout the Brazilian territory, covering 20 cities. Stations can be solarimetric (S), anemometric (A) or solarimetric and anemometric (SA). For the location of Petrolina/CE the type is SA. Each station class measures a set of variables that may differ depending on each station's configuration.

4.2 Predictors and Data Pre-processing

In this work, 83 predictors are considered, namely: time, year, day, min, zenith (filtering only angles with less or equal than 85 degrees), GHI and DNI irradiance average (W/m²) for raw and normalized values and target values from 7 previous time steps (taken from every 5-minute and 60-minute intervals). According to Larson (2019), we define night as the period when the solar zenith angle (θz) is greater than 85 degrees and daytime when the solar zenith angle (θz) is less than 85 degrees.

For GHI and DNI, 38 irradiance variables are modified according to the Clearness Index. That is, for the execution of the 5 forecasting models, the k_t values. According to Kuhn and Kjell (2013), transformations of predictor variables may be needed for several reasons. A few modeling techniques may have strict requirements, such as the predictors having an ordinary scale. In other circumstances, creating a good model can be complex owing to specific characteristics of the data (e.g., outliers).

The most straightforward and common data transformation is to center and scale the predictor variables. To center a predictor variable, the average predictor value is subtracted from all the values. As a result of centering, the predictor has a zero mean. Similarly, to scale the data, each value of the predictor variable is divided by its standard deviation. Scaling the data coerce the values to have a common standard deviation of one. These manipulations are generally used to improve the numerical stability of some calculations.

The caret packaged was applied in R, created by Kuhn (2008), which has the ability to transform, center, scale, or impute values, as well as to apply the spatial sign transformation and feature extraction to manage this series of transformations to multiple data sets. For each ML model the pre-processing was tested separately with and without the clear-sky index for GHI and DNI for 5 min and 60 min time resolution and 5 min, 30 min, 60 min, 6 hours and 12 hours' time horizon. For the LASSO method, no predictor pre-processing was available in the library. However, with the XGBoost model, the pre-processing with center and scale increased the FS results for the normalized variables: GHI and DNI with clear-sky index, although decreased for the raw variables: GHI and DNI.

The kNN model was the most benefited with center and scale pre-processing implementations, as with the normalized variables the FS results increase significantly and decreased for the raw variables. Therefore, we recommend centering and scaling the predictors for normalized variables prior to building XGBoost and kNN models. Comparing with the work of Coimbra et al. (2018), in this study the models selected used only endogenous inputs for generating the forecasts, including the zenith angle as a new auxiliary variable. In other words, the only inputs of the models are the past solar irradiance data, so we obtained continuous and workable time series of GHI and DNI by applying the following rules to the data:

- Remove all of the data for a solar zenith angle inferior or equal to 85° to avoid the side effects of including the low accuracy of the solar measurements before sunrise and after sunset. Thus, the time series obtained do not contain null night values;
- The backward average for the clear-sky index time series;
- The lagged 5-min average values for the 5-min and 60-min clear-sky index time series;
- The lagged 60-min average values for the 60-min and 12 hours clear-sky index time series;

4.3 Observed Database

In this dissertation, the database includes the 83 predictors already mentioned and 197,894 observations. The forecasting models are trained for solar irradiance measurements over the given period during daylight hours ($\theta_{zenith} <= 85°$) (specifically, GHI and DNI) obtained in Petrolina, PE, Brazil, 09° 04' 08" S and 40° 19' 11" W, Northeast of the country.

The raw 1-min data was quality controlled to remove physically impossible values, averaged into 5 min, 30 min, 60 min, 6 hours and 12 hours' time horizons of GHI and DNI directly from the raw data for four consecutive years: January 2013 to December 2016, and divided into three data sets: training, validation and testing. These four years are the ones with less missing values from the SONDA data base related to Petrolina's solarimetric and anemometric station.

In forecasts referring to the complete database, from the four years (2013 to 2016), three years (2013, 2014 and 2015) were chosen from the database, that is, 132,124 observations, which are used in the prediction model. This procedure is adopted to reduce the computational volume calculations and, consequently, the execution time. For the resolution of 5 min, the forecast horizons are 5, 30 and 60 minutes, Lag 1 represents the first step in the temporal resolution, which is equal to 5 min, Lag 6 is equal to 30 min and Lag 12 is equal to 60 min. For the resolution of 60 min, the forecast horizons are 60 min, 6 and 12 hours. In this case, the first step is Lag 1, which is equal to 60 min, Lag 6 is equal to 6 hours and Lag 12 is equal to 12 hours.

Each time horizon is composed with 7 lag times, which means that the time of horizon of 5min and 60min goes from lag 1 to lag 7, time horizon of 30 min and 6 hours goes from lag 6 to lag 12 and time horizon of 60 min and 12 hours goes from lag 12 to lag 18. For the first dataset, denoted as training or historical dataset, the radiation itself is used to be predicted using endogenous models. The second dataset; denoted as optimization dataset, is used in the optimization algorithm to determine the several free parameters (explained below) in the forecasting model. The third dataset; the independent testing set, is used to assess the performance of the forecasting model.

The three data sets were constructed by grouping disjointed subsets for each month, thus ensuring that all data sets are well representative of the irradiance data over the whole period. The second dataset, denoted the optimization dataset, is used in the optimization algorithm to determine the various free parameters (explained below) in the prediction model. The third set of data, the independent test set, is used to assess the performance of the prediction model. The three datasets were constructed by grouping disjoint subsets for each year, thus ensuring that all datasets are well representative of the irradiance data over the entire period. The models are validated by separating the data between the groups of training (with cross-validation) and testing. The models are implemented in R programming language under the RStudio development environment.

It is important to highlight that the present work focuses on the methodology used for the evaluation of different forecasting methods for very-short term of solar irradiance, observing the

effects of very long training data (several years) and the applicability of the methods to a wide range of solar variability microclimates may be done for future works.

4.4 Cross Validation

For the six models considered, except the persistence model, the data base from the last year, 2016, was used for test and the three first years (2013, 2014, 2015) for training. In training data, it is used 5-fold cross validation. This procedure is performed on all databases considered.

4.5 Error Metrics

The error metrics used to evaluate the performance of the applied machine learning models are presented in this section.

4.5.1. Deterministic error metrics

When the goal is to measure the performance of a model for regression problems where one tries to predict a numeric value, the residuals are important sources of information. Residuals are computed as the observed value minus the predicted value (i.e., $y_i - \hat{y}_i$).

4.5.2 RMSE (Root Mean Squared Error)

The root mean squared error (RMSE) is commonly used to evaluate models. RMSE is interpreted as how far, on average, the residuals are from zero and it emphasizes the larger errors.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{I}_i - I_i)^2} \tag{11}$$

with N representing the number of samples of the testing set.

4.5.3 nRMSE (Normalized Root Mean Squared Error)

The Normalized Root Mean Squared Error is the ratio of the RMSE to the mean of the actual values of the variable. It is usually presented as a percentage. This error metric provides rating ranges for the prediction, namely: % nRMSE < 10% excellent, 10% < % nRMSE < 20% good, 20% < % nRMSE < 30% reasonable and % nRMSE ≥ 30% bad, as per suggested by Li et al. (2013). It is calculated by Equation 13.

$$nRMSE = \frac{RMSE}{\bar{I}} \tag{12}$$

with $\bar{I}$ as the mean value of the GHI or DNI variable.

4.5.4 MAE (Mean Absolute Error)

The Mean Absolute Error (MAE) gives the average magnitude of forecast errors and calculates the mean of the absolute differences between the predicted value of GHI or DNI, $\hat{y}_i$, and the real value of GHI and DNI, y_i, as indicated in Equation (14):

$$MAE = \frac{1}{n} \sum_{j=1}^{n} |y_i - \hat{y}_i| \tag{13}$$

4.5.5 nMAE (Normalized Mean Absolute Error)

The Normalized Mean Absolute Error is the ratio between the MAE and the mean of the actual values of the variable calculated by Equation (15).

$$nMAE = \frac{MAE}{\bar{I}} \tag{14}$$

with $\bar{I}$ as the mean value of the GHI or DNI variable.

4.5.6 MBE (Mean Bias Error)

The bias or MBE is the average forecast error representing the systematic error of a forecast model to under or over forecast. As described below, a postprocessing of model output is useful to significantly reduce the bias.

$$MBE = \frac{1}{n}\sum_{j=1}^{n}(y_i - \hat{y}_i) \tag{15}$$

4.5.7 FS (Forecast Skill)

Another metric, used to evaluate the improvement relative to the baseline model (here the persistence model), is the forecast skill (FS) which is, according to Dazhi Yang (2019), the best parameter to compare forecast models at the moment and is given by:

$$s = \left(1 - \frac{RMSE_m}{RMSE_0}\right) x\ 100[\%] \tag{16}$$

where $RMSE_0$ is the RMSE for the persistence model and $RMSE_m$ is the RMSE for the models used in the work (here the MARS, LASSO, kNN, XGBoost or Ensemble with Ridge models).

5 RESULTS

5.1 Forecasting results

Five different machine learning models were applied to the testing set to evaluate their performance in an independent data set. The error metrics RMSE, nRMSE, R-squared, MAE, nMAE, MBE and the FS for all the models are listed in Attachments in Tables 4 to 10 for GHI and DNI of raw and normalized variables, respectively.

The best machine learning models for GHI, ktGHI, DNI and ktDNI including each time resolution (5min and 60 min) and time horizon (5min, 30min, 60 min, 6h and 12h) varied between the five models (Tables 2 and 3). According to the literature (STRAVOS et al., 2019), this could be explained by the NFL theorem: "averaged over all optimization problems, without re-sampling all optimization algorithms perform equally well" (Wolpert, 1996). Besides optimization, the NFL theorem has been successfully used to tackle important theoretical issues pertaining supervised learning in machine learning systems. The NFL theorem has become a suite of theorems which has given significant results in various scientific fields.

Table 2 – Best forecast skill results for the GHI and *kt*GHI forecast for the testing set with time resolution of 5 min and 60 min. RMSE values are in W/m² and the skill s is in percentage. t+5min, t+30min, t+60min, t+6h, t+12h are the time horizons for 5 min, 30min, 60min, 6 hours and 12 hours, respectively.

Global Horizontal Irradiance		Error Metrics			Global Horizontal Irradiance with Clear-Sky Index		Error Metrics		
Time Resolution	Time Horizon	Models	RMSE	s	Time Resolution	Time Horizon	Models	RMSE	s
5 min	t + 5min	XGBoost	67.72	28.53%	5 min	t + 5min	XGBoost	67.60	28.57%
60 min	t + 60min	MARS	58.76	66.79%	60 min	t + 60min	MARS	128.62	36.89%
5 min	t + 30min	Ensemble_Ridge	96.06	39.27%	5 min	t + 30min	XGBoost	97.22	38.69%
60 min	t + 6h	LASSO	130.83	70.52%	60 min	t + 6h	MARS	158.25	66.07%
5 min	t + 60min	Ensemble_Ridge	106.14	51.18%	5 min	t + 60min	XGBoost	107.9	50.49%
60 min	t + 12h	kNN	132.54	73.33%	60 min	t + 12h	MARS	164.75	66.87%

Source: elaborated by the author.

Table 3 – Best forecast skill results for the DNI and ktDNI forecast for the testing set with time resolution of 5 min and 60 min. RMSE values are in W/m² and the skill s is in percentage. t+5min, t+30min, t+60min, t+6h, t+12h are the time horizons for 5 min, 30min, 60min, 6 hours and 12 hours, respectively.

Direct Normal Irradiance		Error Metrics		
Time Resolution	Time Horizon	Models	RMSE	s
5 min	t + 5min	XGBoost	98.84	31.19%
60 min	t + 60min	MARS	129.01	20.83%
5 min	t + 30min	Ensemble_Ridge	157.38	28.23%
60 min	t + 6h	LASSO	261.80	32.27%
5 min	t + 60min	MARS	186.82	26.66%
60 min	t + 12h	kNN	275.48	41.39%

Direct Normal Irradiance with Clear-Sky Index		Error Metrics		
Time Resolution	Time Horizon	Models	RMSE	s
5 min	t + 5min	XGBoost	98.87	31.13%
60 min	t + 60min	MARS	139.38	19.03%
5 min	t + 30min	XGBoost	157.81	28.07%
60 min	t + 6h	MARS	264.65	32.57%
5 min	t + 60min	XGBoost	189.18	25.86%
60 min	t + 12h	MARS	282.07	39.66%

Source: elaborated by the author.

According to Table 2 and 3, the extreme gradient boosting, and MARS models prevail for GHI and DNI, with normalized variables when the time resolution of 5 min and 60 min are chosen, respectively. It is worth to mention, for the time resolution of 5 min, that the XGBoost model has the FS's best results in 66.66% of the cases comparing to all the twelve results for GHI and DNI with raw and normalized variables. Although, for the time resolution of 60 min, the MARS model has the FS's best results in 66.66% of the cases for GHI and DNI with raw and normalized variables.

The results reveal that GHI and DNI with Clear-Sky Index have a prevalent model for each different time horizon, which it does not happen with GHI and DNI with raw variables. For the time resolution of 5 min and using raw variables, the RMSE for all the time horizons (5 min, 30 min, 60 min) ranges between 67.72 and 118.96 W/m² for GHI, whereas for DNI, the RMSE ranges from 98.84 to 203.39 W/m² (Graphic 1 and Graphic 2). A reduction in the RMSE translates into a significant FS that ranges between 26.19% and 51.18%, and between 20.15% and 31.19% for the GHI and DNI with raw variables testing set, respectively (Graphic 3 and Graphic 4).

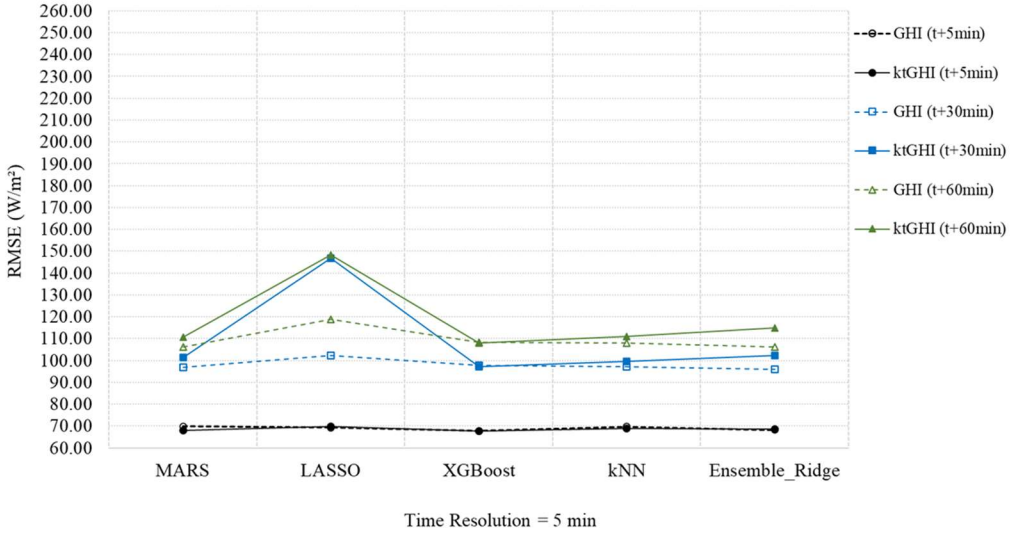
Also, for the normalized variables, time resolution of 5 min, the RMSE for all the time horizons (5 min, 30 min, 60 min) ranges between 67.60 and 148.27 W/m² for *kt*GHI, whereas for *kt*DNI, the RMSE ranges from 98.87 to 203.13 W/m² (Graphic 1 and Graphic 2). The reduction in RMSE implies into a FS that range between 7.44% and 50.49%, and between 20.39% and 31.13%

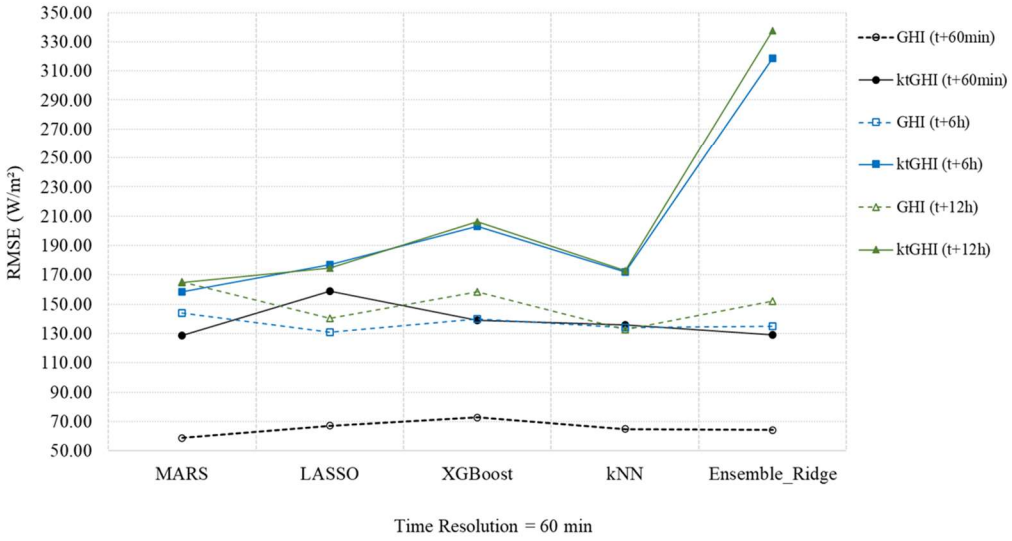
for the *kt*GHI and *kt*DNI with normalized variables testing set, respectively (Graphic 3 and Graphic 4).

For the time resolution of 60 min and raw variable, the RMSE for all the time horizons (60 min, 6 h, 12 h) ranges between 58.76 and 165.23 W/m² for GHI, whereas for DNI, the RMSE ranges from 129.01 to 350.15 W/m² (Graphic 1 and Graphic 2). A reduction in RMSE translates into a significant FS that ranges between 58.88% and 73.32%, and between 13.36% and 41.39% for the GHI and DNI with raw variables testing set, respectively (Graphic 3 and Graphic 4). Also, for the normalized variables, time resolution of 60 min, the RMSE for all the time horizons (60 min, 6 h, 12 h) ranges between 128.62 and 337.74 W/m² for *kt*GHI, whereas for *kt*DNI, the RMSE range from 139.38 to 343.95 W/m² (Graphic 1 and Graphic 2). The reduction in the RMSE implies into a FS that range between 22.07% and 66.87%, and between 11.02% and 39.66% for the *kt*GHI and *kt*DNI with normalized variables testing set, respectively (Graphic 3 and Graphic 4).

According to Graphic 1, for GHI with raw and normalized variables, the LASSO and Ensemble with Ridge model have the highest RMSE results, for time resolution of 5 min and 60 min, respectively. It can be noticed that the lowest RMSE results for GHI and DNI, raw and normalized variables, are for the time horizons of 5 min.

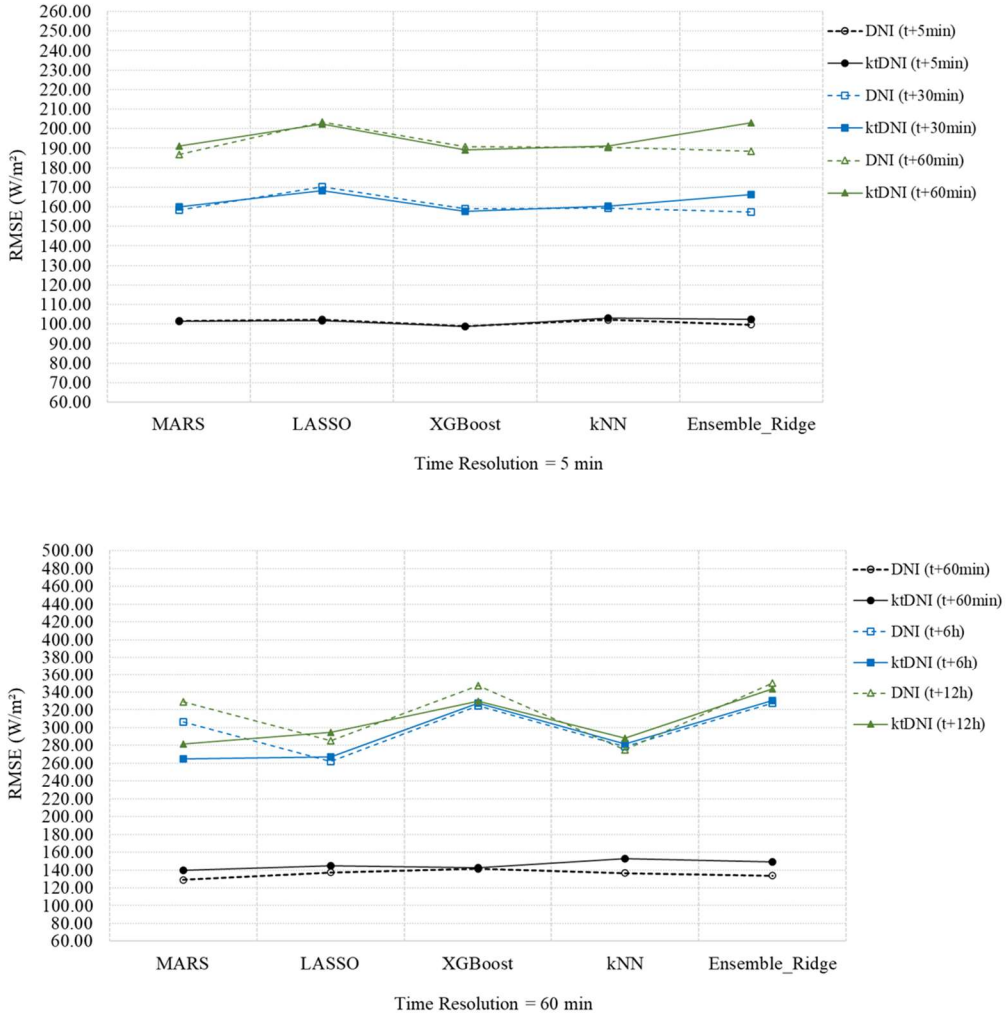
Graphic 1 – RMSE for GHI and ktGHI forecasts (testing set). t+5min, t+30min, t+60min, t+6h, t+12h are the time horizons for 5 min, 30min, 60min, 6 hours and 12 hours, respectively. RMSE values are in W/m².





Source: elaborated by the author.

Graphic 2 – RMSE for DNI and ktDNI forecasts (testing set). t+5min, t+30min, t+60min, t+6h, t+12h are the time horizons for 5 min, 30min, 60min, 6 hours and 12 hours, respectively. RMSE values are in W/m².

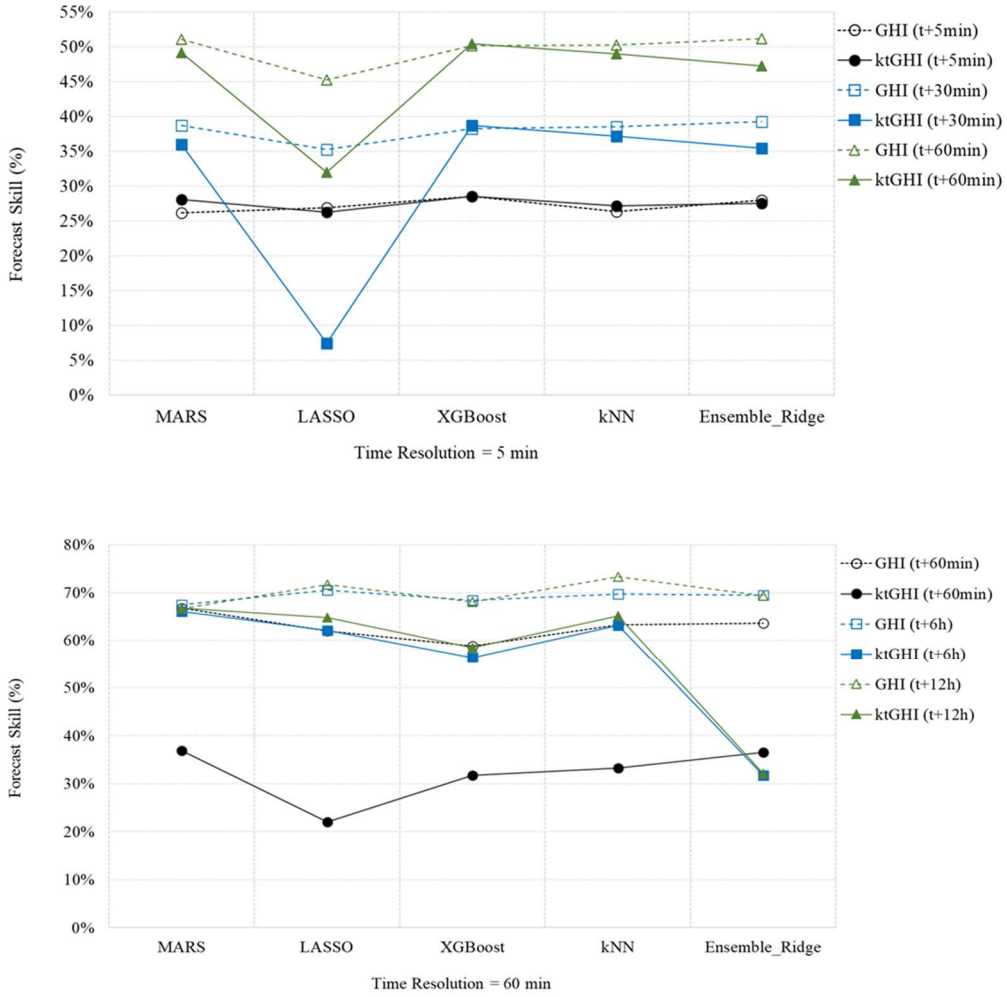


Source: elaborated by the author.

As we can see in Graphic 3, the FS results for GHI using the Clear-Sky Index and time resolution of 5 min, has a deterioration of 27.83% and 13.32% for the LASSO model, for the time horizons of 30 min and 60 min, respectively. For the time resolution of 60 min, the use of Clear-Sky Index for the GHI decreases the FS between 1.45% and 39.98% for all time horizons (Graphic 3). On the other hand, when using the Clear-Sky Index, time resolution of 5 min, the FS's outcome for DNI has a maximum upgrade of 0.85% for the LASSO model, and a drop of 4.01% and 5.57% for the Ensemble with Ridge model can be notice, for the time horizons of 30 min and 60 min, respectively (Graphic 4).

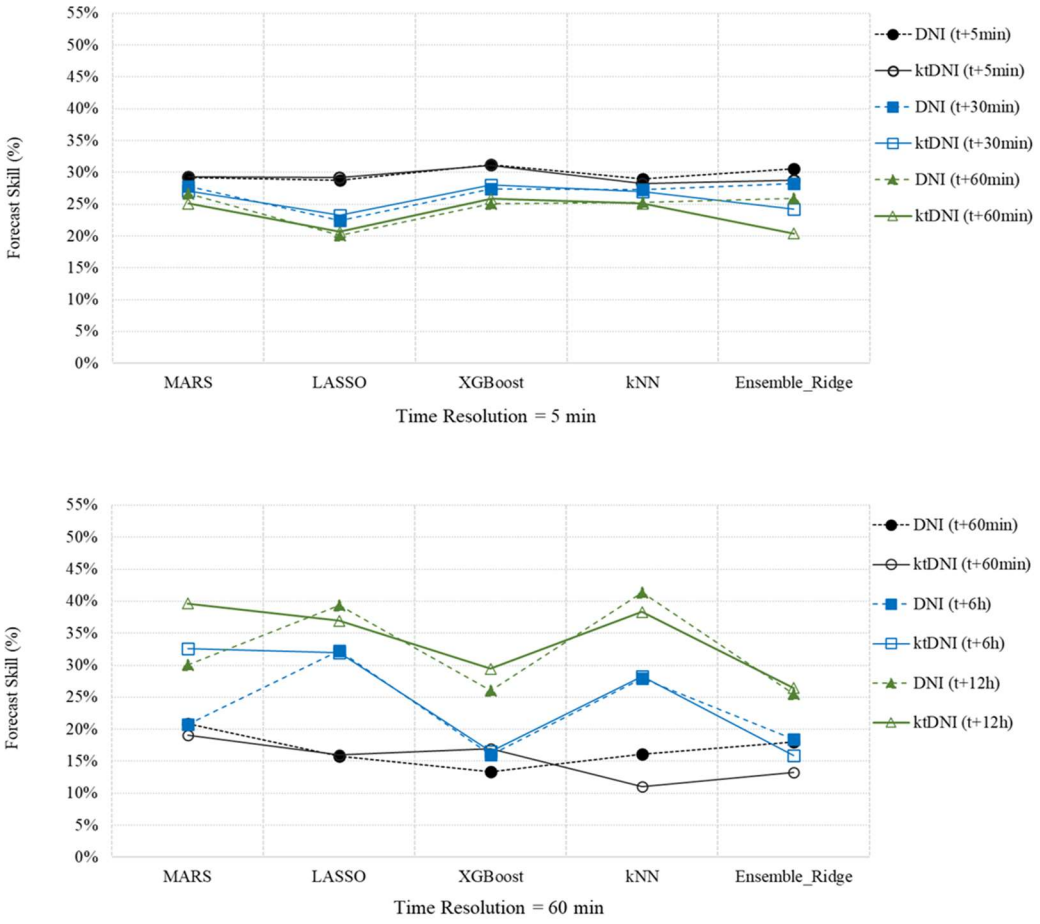
For the time resolution of 60 min, the use of Clear-Sky Index for the DNI increases the FS between 0.35% and 5.12% in all time horizons and decreases for the MARS models in 11.78% and 9.60%, for the time horizons of 30 min and 60 min, respectively (Graphic 4). The FS results for GHI and DNI show a different behavior, as for the first one, the FS increases when the time horizon grows and for the Direct Normal Irradiance the pattern is the opposite, the FS decreases when the time horizon grows (Graphic 3 and Graphic 4).

Graphic 3 – Forecast skill for GHI and ktGHI forecasts models (testing set), with time resolution of 5min and 60 min, respectively t+5min, t+30min, t+60min, t+6h, t+12h are the time horizons for 5 min, 30min, 60min, 6 hours and 12 hours, respectively. Forecast skill values are in percentage.



Source: elaborated by the author.

Graphic 4 – Forecast skill for DNI and kDNI forecasts (testing set), with time resolution of 5min and 60 min, respectively t+5min, t+30min, t+60min, t+6h, t+12h are the time horizons for 5 min, 30min, 60min, 6 hours and 12 hours, respectively. Forecast skill values are in percentage.

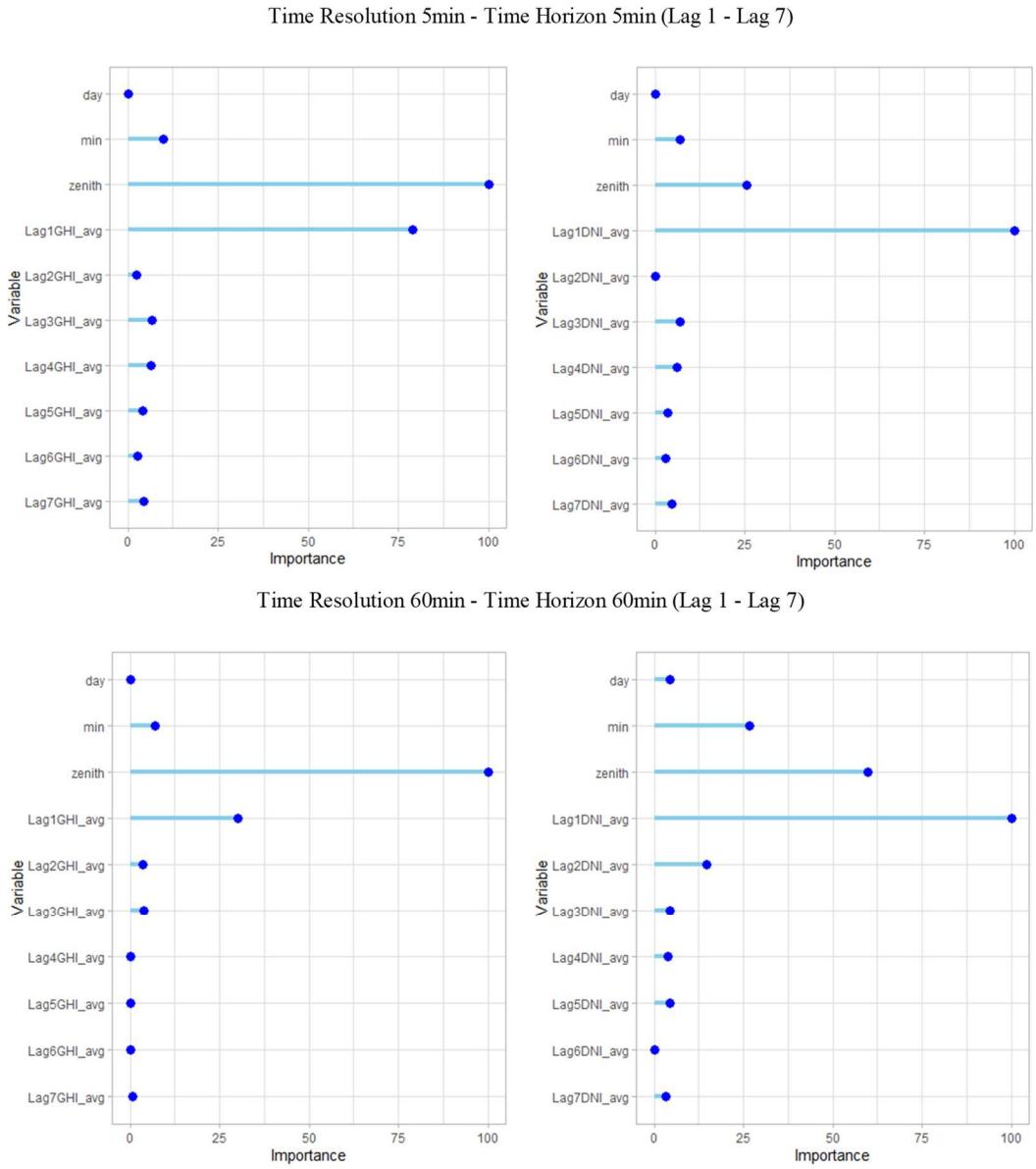


Source: elaborated by the author.

Figure 4 and Figure 5 show the most important independent variables for GHI and DNI for raw and normalized variables in the LASSO method. It's clearly that the importance of the zenith angle decreases significantly with DNI for raw and normalized variables. In addition, the use of 7 lag times is enough with the time resolution of 5 minutes and 60 minutes (Figure 4 and 5), as the model decreases the level of importance until maximum of 25% of the other lags for GHI and DNI with raw and normalized variables. For DNI with raw and normalized variables, the level of importance is very significantly for Lag1, reaching 100% for the time resolution of 5min and 60min and time horizon of 5min, 30min, 60min and 6 hours.

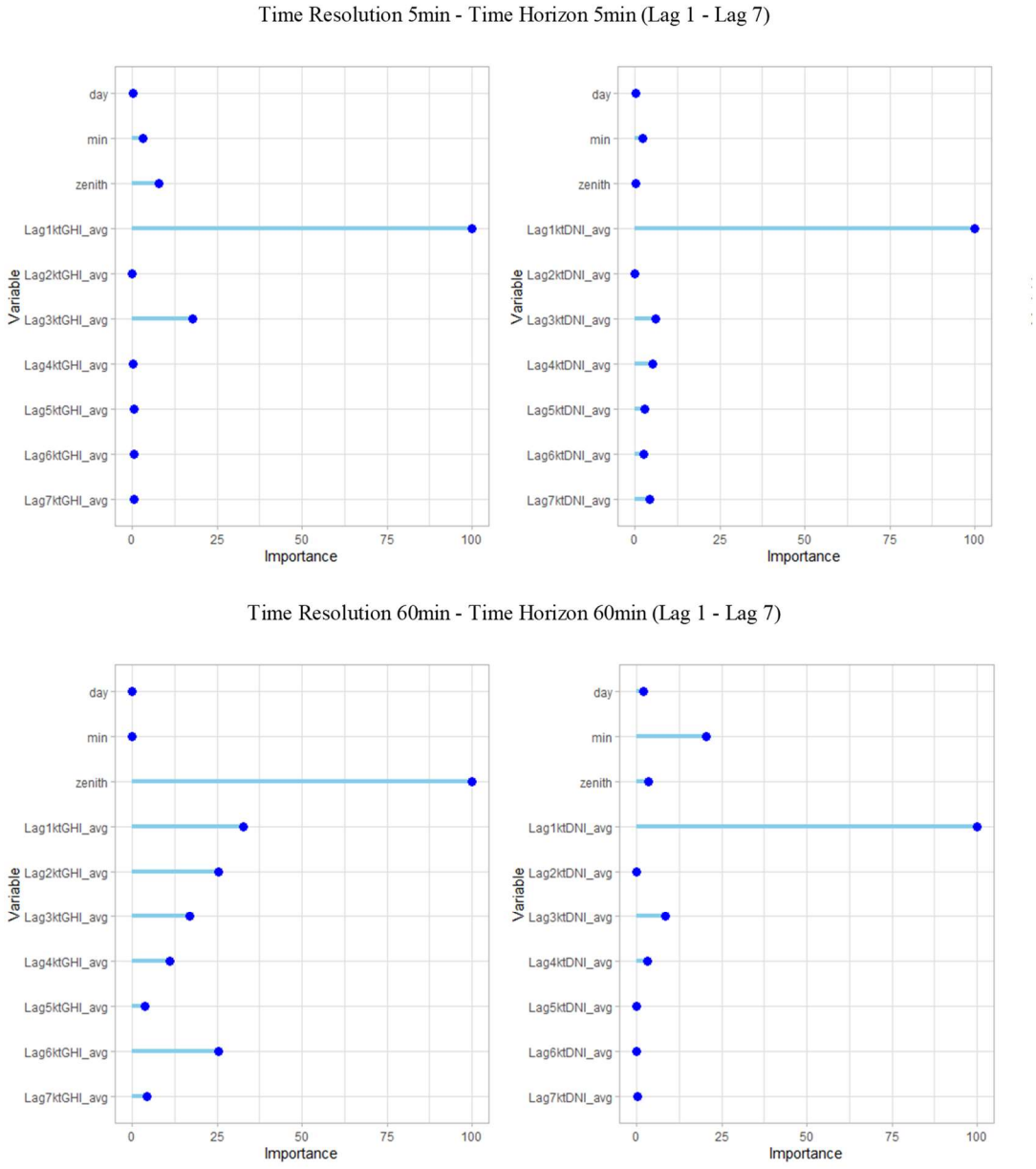
It can be seen from Figure 4 that the zenith angle account for a direct contribution in the model for GHI. Through the analysis of model feature contribution, the effectiveness of the feature selection is verified, and it is confirmed that different algorithms have different emphasis on the raw and normalized variables.

Figure 4 – Variable importance (in percentage) using LASSO for GHI and DNI for 5min and 60min time resolution, respectively.



Source: elaborated by the author.

Figure 5 – Variable importance (in percentage) using LASSO for ktGHI and ktDNI for 5min and 60min time resolution, respectively.

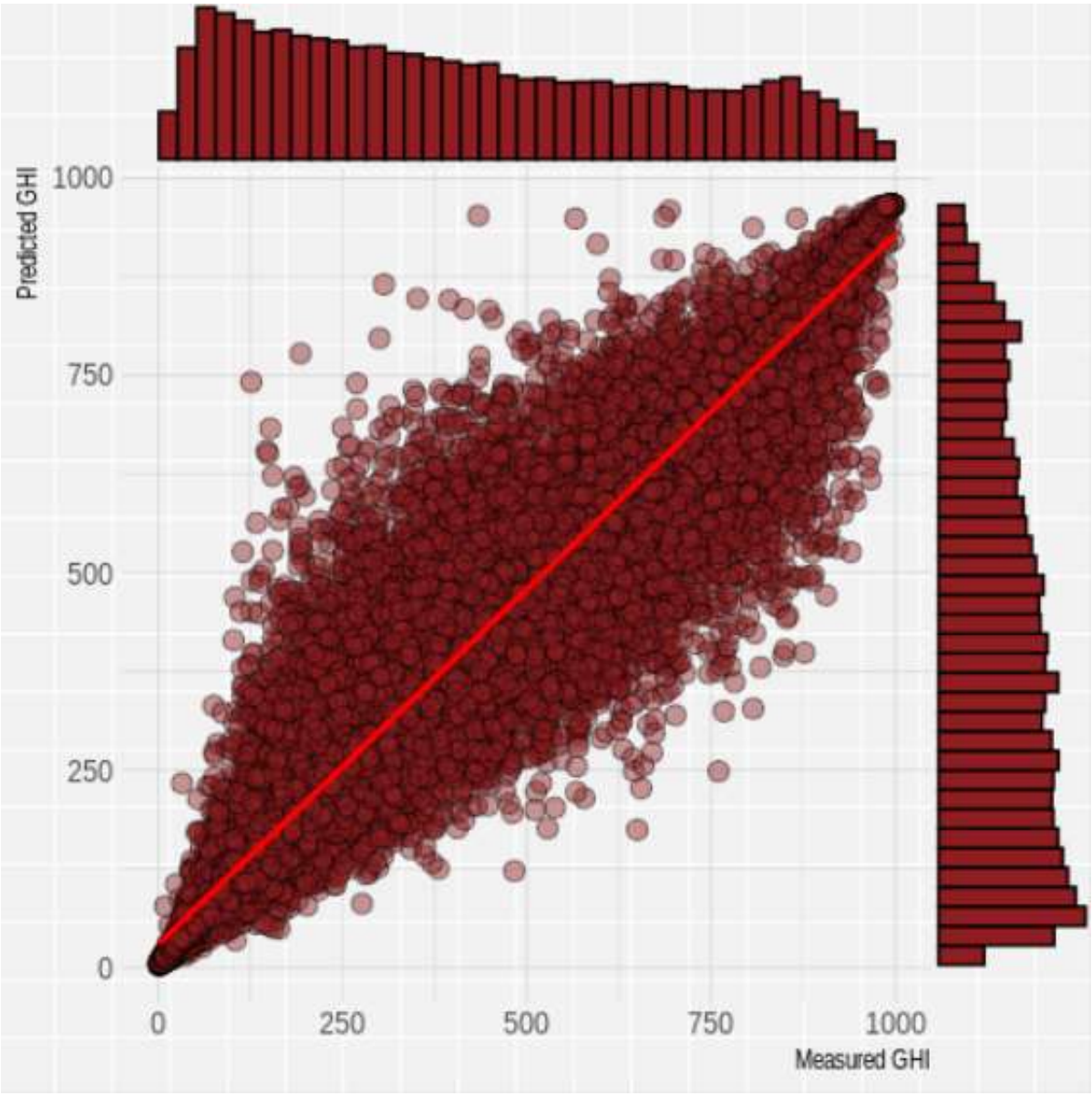


Source: elaborated by the author.

To verify the predicted and measured forecasting of GHI and DNI for the model with the best results, XGBoost, in Figure 6 the predicted GHI histograms have a different shape as the scatter plot also indicates that the model tends to overestimate when the measure is small and underestimate it when it is larger. Although, we could notice that there is not a significant difference

in the histograms for DNI, but the model tends to underestimate when the measure is larger, as shown in Figure 7. This confirms that the proposed transformation did not statistically distort the response variable.

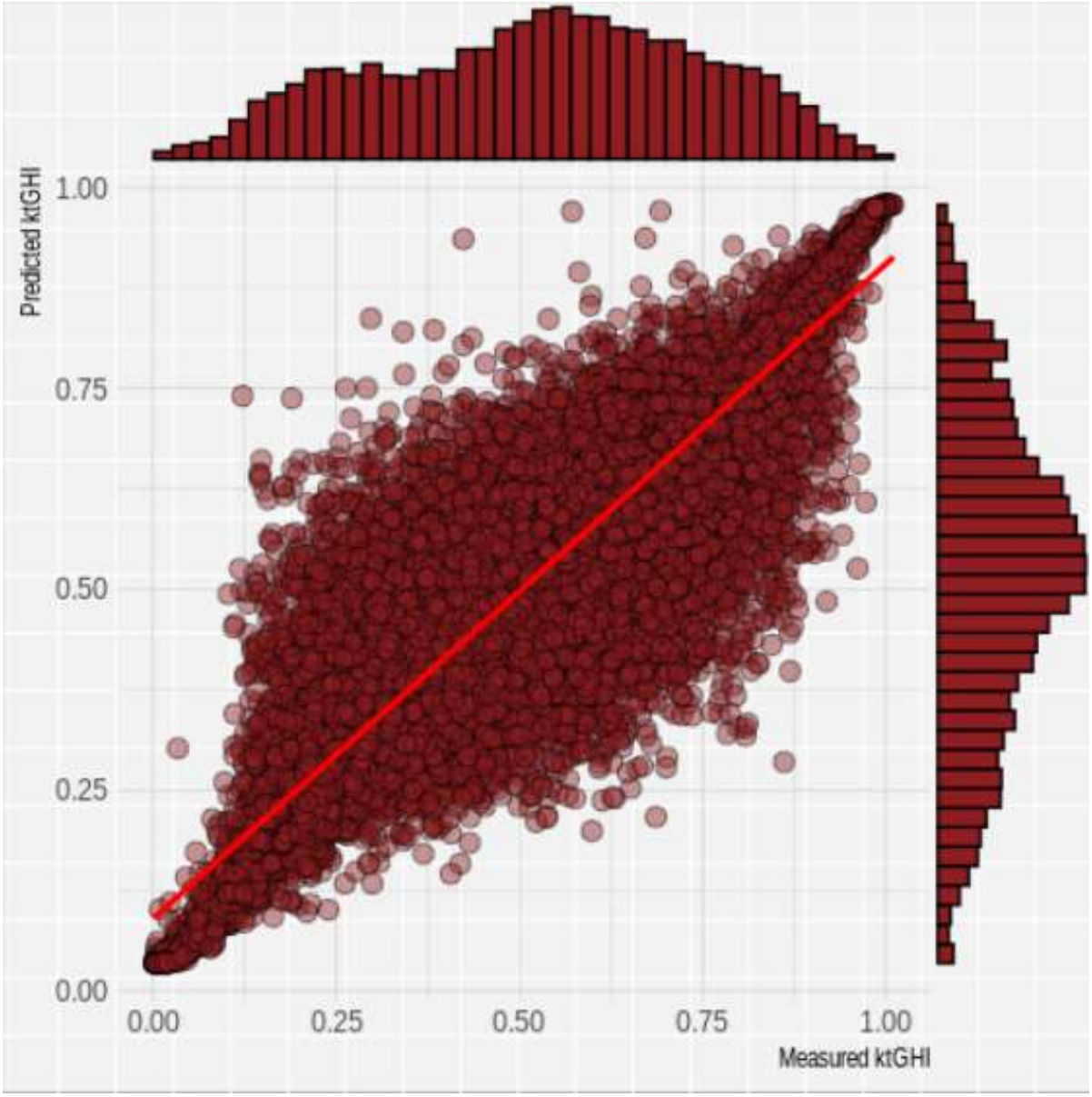
Figure 6 – Scatter plot using XGBoost for GHI for raw (a) and normalized (b) variables respectively.
(a)



Source: elaborated by the author.

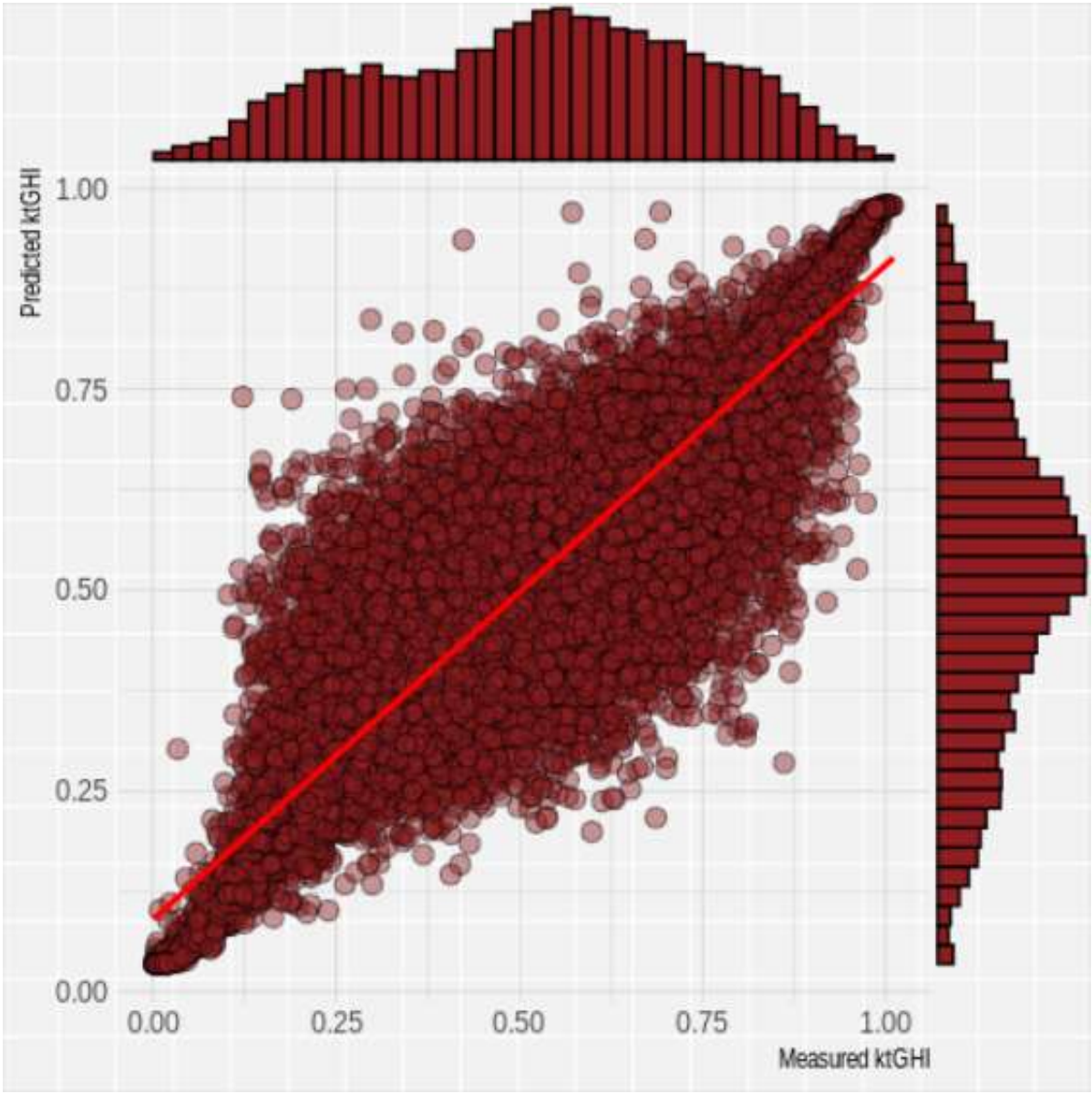
Figure 6 – Scatter plot using XGBoost for GHI for raw (a) and normalized (b) variables respectively.

(b)



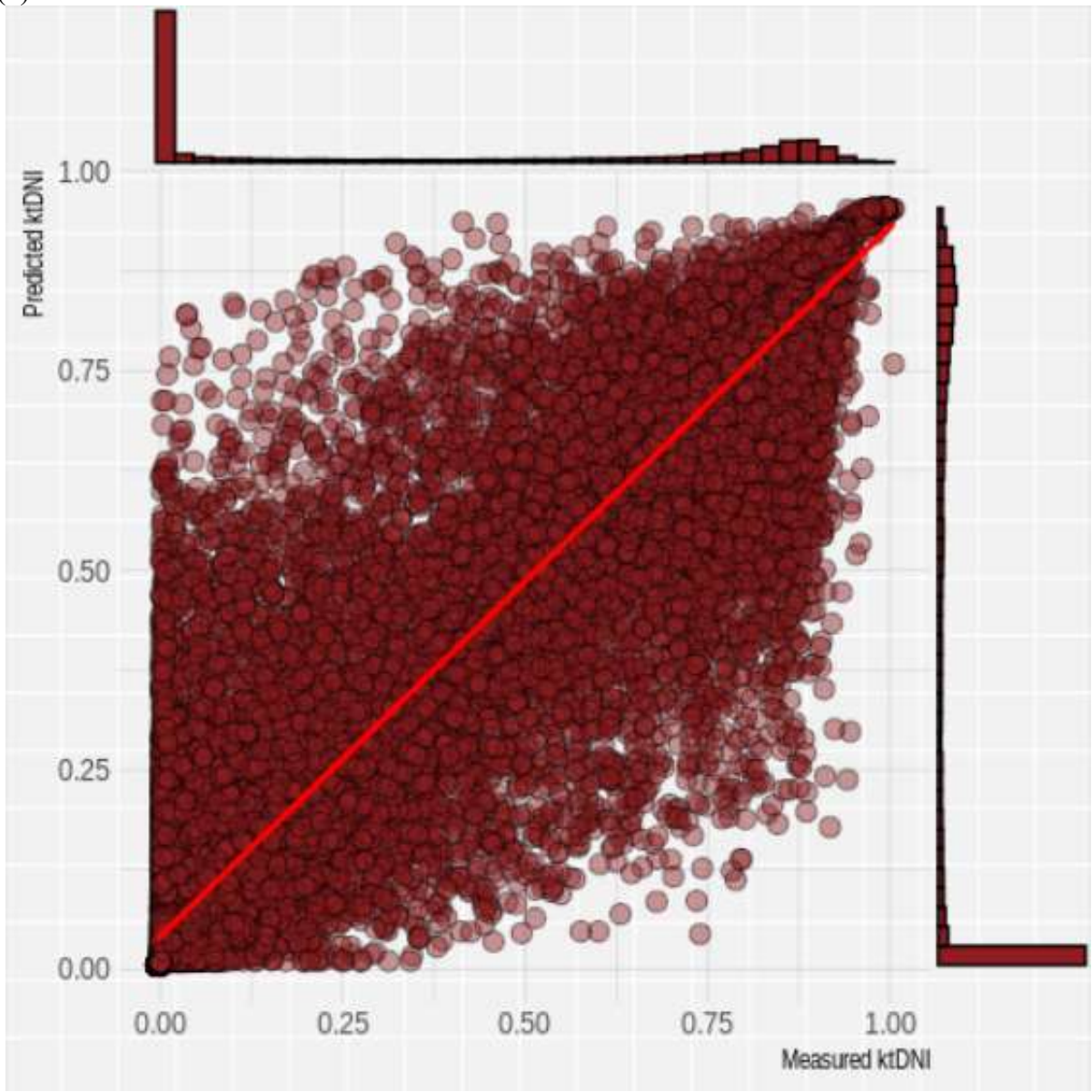
Source: elaborated by the author.

Figure 7 – Scatter plot using XGBoost for DNI for raw (a) and normalized (b) variables respectively.
(a)



Source: elaborated by the author.

Figure 7 – Scatter plot using XGBoost for DNI for raw (a) and normalized (b) variables respectively.
(b)



Source: elaborated by the author.

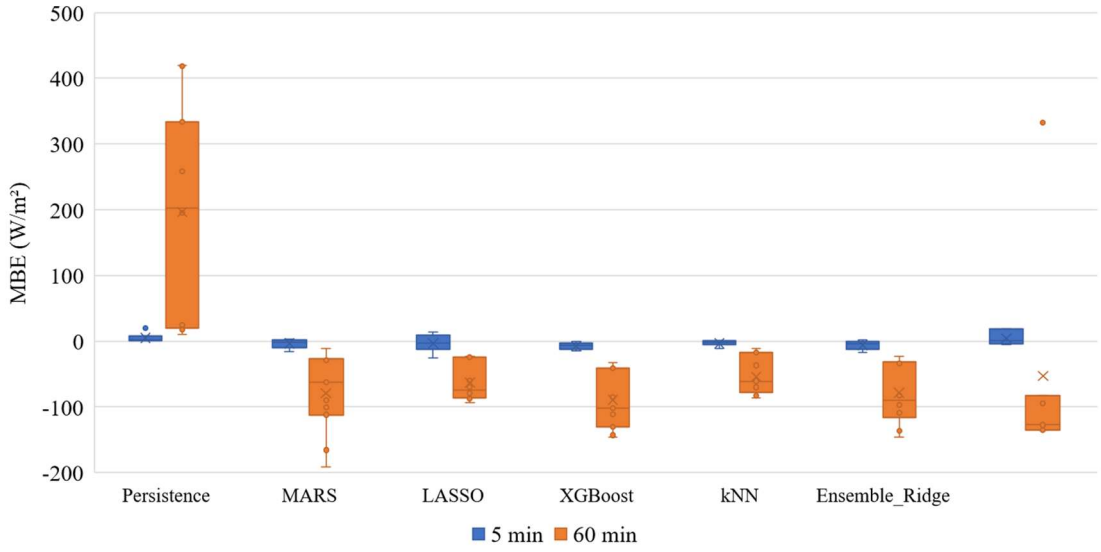
5.1.1 Overview of Error Metrics results

To further evaluate the forecast, normalized root mean squared error (nRMSE), root mean squared error (RMSE), normalized mean absolute error (nMAE), and normalized mean bias error (MBE) were computed over the given period during daylight hours (θ_{zenith} <= 85°) for the following time horizons: 5 min, 30 min, 60 min, 6 hours and 12 hours. According to Graphic 5,

Graphic 6, Graphic 7 and Graphic 8, the results for GHI and DNI with raw and normalized variables show a stable behavior for the nRMSE, RMSE and nMAE error metrics. The negative results for the MBE indicate that all models underestimate the power output for Global Horizontal Irradiance and Direct Normal Irradiance forecasts. Contrarywise, positive results for MBE indicate that the models are overestimating the power output.

According to Table 4, it notices that the time resolution of 60 min has higher negative outputs for MBE than the time resolution of 5 min, which proves an underestimation of the results, even with higher results for the FS. However, the positive result for MBE in the persistence model proves an overestimation in the results.

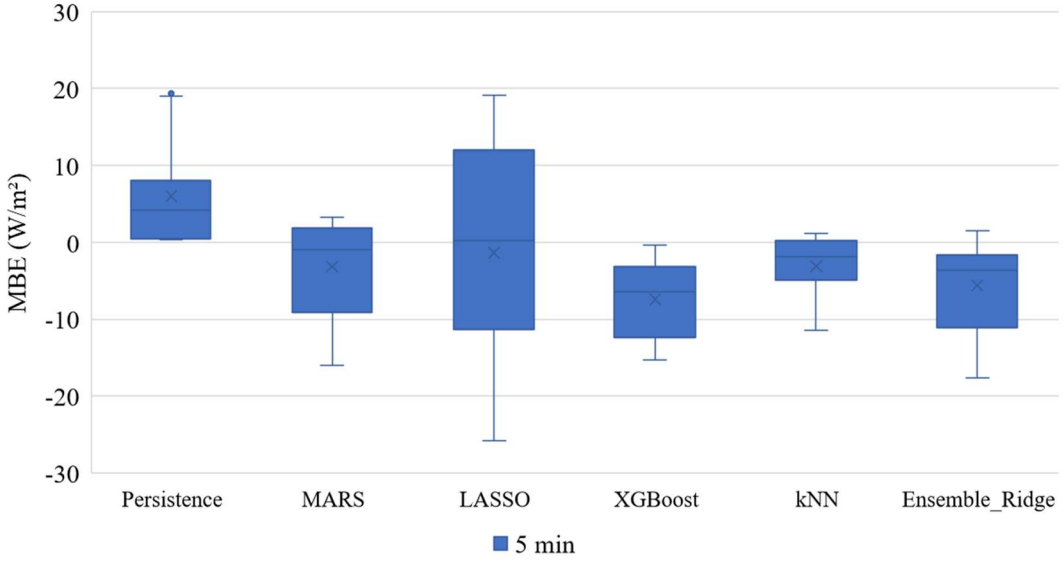
Table 4 – Boxplot of Mean Bias Error (MBE) for each forecast models, comparing the time resolution of 5 min and 60 min. MBE values are W/m².



Source: elaborated by the author.

kNN is the Machine Learning model with the best outputs for MBE, proving that the model is more accurate and does not have huge variations on the estimations.

Table 5 – Boxplot of Mean Bias Error (MBE) of each forecast model for the time resolution of 5 min. MBE values are in W/m².



Source: elaborated by the author.

Table 6 – Minimum and maximum results for MBE (W/m²), RMSE (W/m²) and FS (%) for GHI and DNI, raw and normalized variables, for the forecasts models (testing set), with time resolution of 5 min and 60 min. Forecast skill (FS) values are in percentage.

Time Resolution	Predictor	RMSE (W/m²)		MBE (W/m²)		FS (%)	
		Min.	*Max.*	*Min.*	*Max.*	*Min.*	*Max.*
5 min	GHI	67.72	118.96	-15.22	-1.96	26.19	51.18
	kt GHI	67.60	148.27	-25.72	-2.63	7.44	50.49
	DNI	98.84	203.39	-6.37	13.34	20.15	31.19
	kt DNI	98.87	203.13	-5.05	19.08	20.39	31.13
60 min	GHI	58.76	165.23	-112.16	-10.80	58.88	73.33
	kt GHI	128.62	337.74	-111.75	-28.90	22.07	66.87
	DNI	129.01	350.15	-191.23	-16.39	13.36	41.39
	kt DNI	139.38	343.95	-135.59	-17.28	11.02	39.66

Source: elaborated by the author.

The MBE outcomes, time resolution of 5 min, varies from -15.22 to -1.96 W/m², and from -25.72 to -2.63 W/m² for GHI with raw and normalized variables, respectively (Table 6). It also ranges from -6.37 to 13.34 W/m², and from -5.05 to -19.08 W/m², for DNI with raw and normalized

variables, respectively (Table 6). However, the MBE outcomes, time resolution of 60 min, varies from -112.16 to -10.80 W/m², and from -111.75 to -28.90 W/m² for GHI with raw and normalized variables, respectively (Table 6). It also ranges from -191.23 to -16.39 W/m², and from -113.59 to -17.28 W/m², for DNI with raw and normalized variables, respectively (Table 6).

The kNN model and the XGBoost are the models with the best results when taking the MBE error metric in consideration: 75% and 66.66% of the results, respectively, for the time resolution of 5 min, including GHI and DNI. For the time resolution of 60 min, the MARS model has the best results of MBE, dominating around 58.33% of all the outputs and the kNN model has 66.66%, including GHI and DNI for raw and normalized variables.

5.1.2 Overview of Machine Learning models results

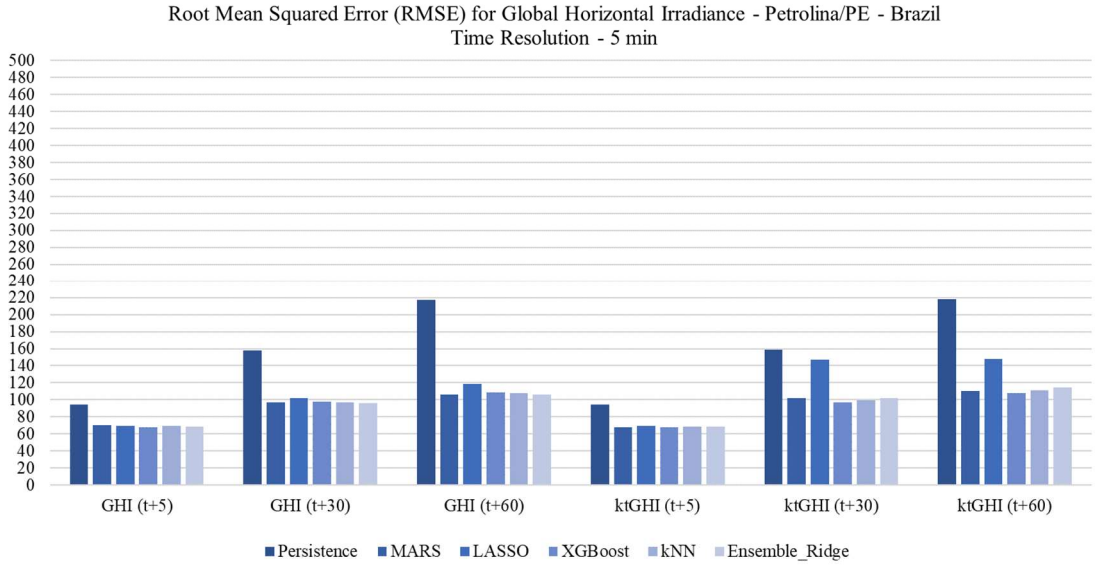
All the five techniques (MARS, LASSO, XGBoost, kNN and Ensemble with Ridge for raw and normalized variables) clearly outperform the persistence model. This allegation is reinforced by the RMSE decrease results, between 7% and 50.49% for 5 min time resolution, and between -11.02% and 73.33% for 60 min time resolution, shown in Graphic 5 and Graphic 6. Furthermore, the results indicate that, whatever the machine learning technique, the inclusion of clear-sky index does not bring a clear improvement for all the models using the data from Petrolina/PE in Brazil.

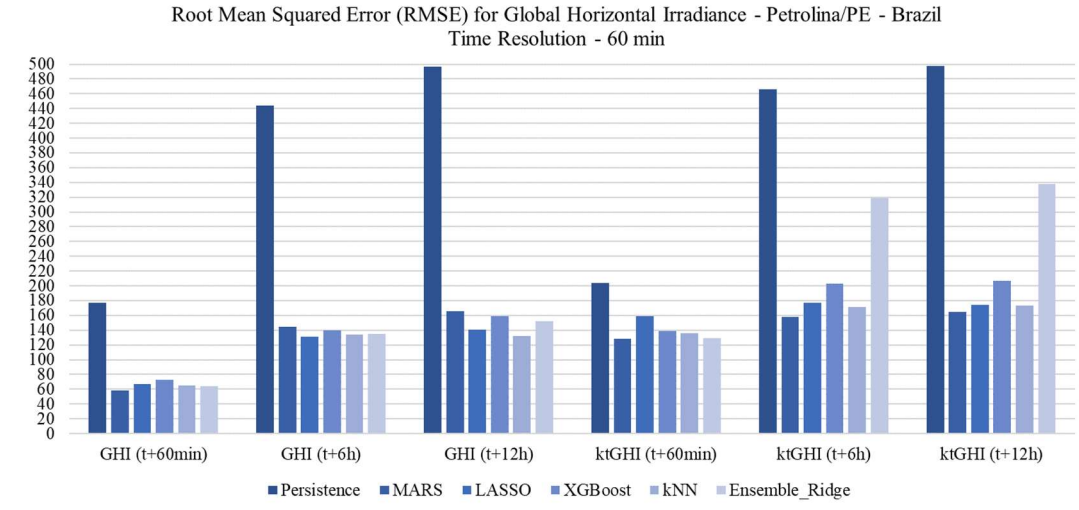
The proposed dissertation indicates that the use of endogenous and linear regression models can achieve a maximum FS result for GHI of 10.98% (time resolution of 5 min) and 26.59% (time resolution of 60 min) higher than a recent study from Kumari and Toshniwal (2021), which use extreme gradient boosting model and deep neural network for 1 hour time horizon. Comparing with the study of Pedro and Coimbra (2018), using the same time resolutions of 5 min and time horizon of 5 min, the present dissertation had an improvement of 15.23% and 16.89%, for GHI and DNI, respectively and an increase of the RMSE results in 35.01 W/m² and 40.64 W/m², GHI and DNI, respectively. Also, for the time horizon of 30 min and time resolution of 5 min, the improvement was 15.67% for GHI and a decrease of -1.03% for DNI, and an increase of the RMSE results in 61.66 W/m² and 86.08 W/m², GHI and DNI, respectively correlating to Pedro and Coimbra (2018).

Contrasting the study of Hassan et al. (2017), this dissertation achieved a RMSE 29.99 W/m² lower, for GHI with a time horizon and resolution of 60 min. Correlating the results of the

present study with Yang et al. (2020), which used a time horizon of 60 min and a time resolution of 15 min for GHI and DNI, for the same time horizon of 60 min, but a resolution of 60 min, the results for the FS are significantly higher and the RMSE results are 58.76 W/m² and 129.01 W/m² lower for GHI and DNI, respectively. Although, because the time horizon is not equals to one another, the comparison is very questionable.

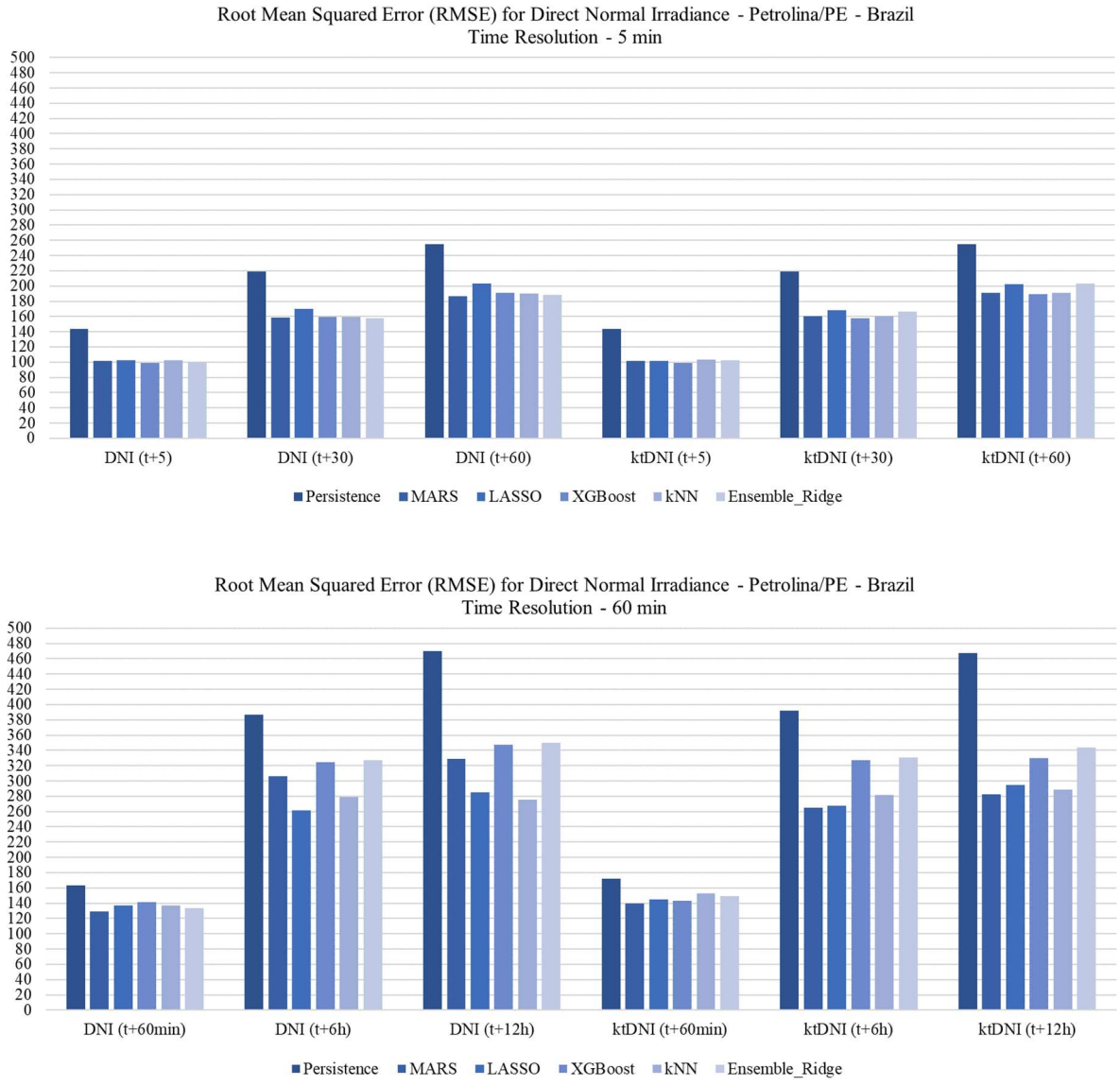
Graphic 5 – RMSE for GHI and ktGHI forecasts (testing set) compared with the Persistence Model. t+5min, t+30min, t+60min, t+6h, t+12h are the time horizons for 5 min, 30min, 60min, 6 hours and 12 hours, respectively. RMSE values are in W/m².





Source: elaborated by the author.

Graphic 6 – RMSE for DNI and ktDNI forecasts (testing set) compared with the Persistence Model. t+5min, t+30min, t+60min, t+6h, t+12h are the time horizons for 5 min, 30min, 60min, 6 hours and 12 hours, respectively. RMSE values are in W/m².



Source: elaborated by the author.

It is possible to notice that the accuracies of the models are not significantly sensitive to the presence of the clear sky index, , with few exceptions. In fact, there is a decrease in the error metrics of the methods, except for the MBE, which proves an overestimation of the higher results. Regarding the filtering of data, using center and scale in general, the pre-processing of the data

brought benefits to the performance of the models when compared to the case of using information from unfiltered data.

In terms of probabilistic forecasts, the results presented above clearly demonstrate the benefit of using endogenous features from a sensor station. This was demonstrated for both GHI and DNI for horizons shorter than 12 hours. The implementation of such forecasting models would be crucial to improving grid management and integrating intermittent energy sources more effectively into the grid. Indeed, probabilistic forecasts are important inputs for stochastic models of grid management (Hytowitz et al., 2015) (Olivares et al., 2015). Additionally, solar forecast accuracy is critical for optimal grid-connected storage management (Hanna et al., 2014). The presented work provides a methodology that can further these goals for the sub 12 hours window.

For future work, a good strategy would be to use the same models for different climates and locations of Brazil to compare also if the results are directly connected to each region or not and use the same as attributes for the methods.

6 CONCLUSIONS

In the present dissertation, predictions of the global horizontal (GHI) and direct normal DNI) irradiation were performed using data from 2013 to 2016, in the location of Petrolina/PE for time horizons of 5 min, 30 min, 60 min, 6 hours and 12 hours, time resolution of 5 min and 60min through the implementation of the following machine learning models: MARS, LASSO, XGBoost, *k*NN, Ensemble with Ridge and Persistence.

It was found that comparing with the time resolution of 5 min, the use of the time resolution of 60 min, increased the RMSE average between 19.9% and 108.8%, with the MBE average error metric increased between 10.3% to 30.8% with negative results, underestimating the forecasting outcomes, which proved that the most accurate time resolution is the 5 min one.

Boosting algorithm is the method that best suited the data under study for the FS results. The XGBoost's model was not the best one in all time horizons and resolutions, but it is persistently among the best results for GHI and DNI, with normalized variables. The extreme gradient boosting model prevails when the time resolution of 5 min is chosen, considering the FS results.

For the time resolution of 5 min, the XGBoost model has the FS's best results in 66.66% of the time comparing to all the six results for GHI and DNI with raw and normalized variables, for the time resolution of 60 min, the MARS model has the FS's best results in 66.66% of the time for GHI and DNI with raw and normalized variables.

kNN is the model with the best outputs of MBE, proving that the model is more accurate and does not have huge estimations variations. However, the model does not have the highest results for the FS, it is included in the best ones. The k-nearest neighbors algorithm showed the lowest results of MBE with 7 temporal horizons, indicating that it is the model with the lowest sensitivity to the presence of this variable.

All the five techniques (MARS, LASSO, XGBoost, kNN and Ensemble with Ridge for raw and normalized variables) clearly outperform the persistence model. Also, the results indicate that, whatever the machine learning technique, the inclusion of clear-sky index does not bring a clear improvement for all the models using the data from Petrolina/PE in Brazil.

REFERENCES

ALTHOFF, T. D.; MENEZES, R. S. C.; DE CARVALHO, A. L. Climate change impacts on the sustainability of the firewood harvest and vegetation and soil carbon stocks in a tropical dry forest in Santa Teresinha municipality, northeast Brazil," **Forest Ecology and Management**, vol. 360, pp. 367–375, 2016. DOI: https://doi.org/10.1016/j.foreco.2015.10.001

ANDRADE, J. R.; BESSA, R. J. Improving renewable energy forecasting with a grid of numerical weather predictions. **IEEE Transactions on Sustainable Energy**, v. 8, n. 4, p. 1571–1580, out. 2017. DOI: https://doi.org/10.1109/TSTE.2017.2694340

AWANGE, J. L.; MPELASOKA, F.; GONCALVES, R. M. When every drop counts: analysis of droughts in Brazil for the 1901-2013 period. **Science of The Total Environment**, v. 566-567, p. 1472–1488, out. 2016. DOI: https://doi.org/10.1016/j.scitotenv.2016.06.031

BENALI, L.; NOTTON, G.; FOUILLOY, A.; VOYANT, C.; DIZENE, R. Solar radiation forecasting using artificial neural network and random forest methods: application to normal beam, horizontal diffuse and global components. **Renewable energy**, v. 132, p. 871-884, 2019. DOI: https://doi.org/10.1016/j.renene.2018.08.044

PAI, E. D.; ESCOBEDO, J. F. Estimativa da radiação atmosférica em função dos índices radiométricos kt e kd para Botucatu-SP. **Energia na Agricultura**, v. 30, n. 2, p. 172, 15 dez. 2014. DOI: https://doi.org/10.17224/EnergAgric.2015v30n2p172-179

DONG, Z.; YANG, D.; REINDL, T.; WALSH, W. M. Short-term solar irradiance forecasting using exponential smoothing state space model. **Energy**, v. 55, p. 1104-1113, 2013. DOI: https://doi.org/10.1016/j.energy.2013.04.027

HANNA, R. et al. Energy dispatch schedule optimization for demand charge reduction using a photovoltaic-battery storage system with solar forecasting. **Solar Energy**, v. 103, p. 269–287, maio 2014. DOI: https://doi.org/10.1016/j.solener.2014.02.020

HASSAN, M. A. et al. Ultra-short-term exogenous forecasting of photovoltaic power production using genetically optimized non-linear auto-regressive recurrent neural networks. **Renewable Energy**, v. 171, p. 191–209, jun. 2021. DOI: https://doi.org/10.1016/j.renene.2021.02.103

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning, second edition**: data mining, inference, and prediction. New York: Springer, 2009.

HUERTAS-TATOA, J.; CENTENO BRITO, M. Using smart persistence and random forests to predict photovoltaic energy production. **Energies** 2019, 12, 100. DOI: https://doi.org/10.3390/en12010100

HYTOWITZ, R. B.; HEDMAN, K. W. Managing solar uncertainty in microgrid systems with stochastic unit commitment. **Electric Power Systems Research**, v. 119, p. 111–118, fev. 2015. DOI: https://doi.org/10.1016/j.epsr.2014.08.020

IRENA. **Renewable power generation costs in 2017**. Technical report, International Renewable Energy Agency, Abu Dhabi, January 2018.

INMAN, R. H.; PEDRO, H. T. C.; COIMBRA, C. F. M. Solar forecasting methods for renewable energy integration. **Progress in Energy and Combustion Science**, v. 39, n. 6, p. 535–576, dez. 2013.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning**. New York: Springer, 2013.

KUHN, M. Building predictive models in r using the caret package. **Journal of Statistical Software**, [S. l.], v. 28, n. 5, p. 1–26, 2008. https://doi.org/10.18637/jss.v028.i05

KUMARI, P.; TOSHNIWAL, D. Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance. **Journal of Cleaner Production**, p. 123285, ago. 2020. DOI: https://doi.org/10.1016/j.jclepro.2020.123285

LARSON, V. E. Forecasting solar irradiance with numerical weather prediction models. **Solar Energy Forecasting and Resource Assessment**, p. 299–318, 2013. DOI: https://doi.org/10.1016/B978-0-12-397177-7.00012-7

LARSON, D. P. **Data-driven forecasting for grid-connected solar power plants**. 2019. 97 p. Theses (Mechanical Engineering) - University of California San Diego, California, 2019.

OLIVARES, D. E. et al. Stochastic-predictive energy management system for isolated microgrids. **IEEE Transactions on Smart Grid**, v. 6, n. 6, p. 2681–2693, nov. 2015. DOI: https://doi.org/10.1109/TSG.2015.2469631

PEDRO, H. T. C.; INMAN, R. H.; COIMBRA, C. F. M. Mathematical methods for optimized solar forecasting. **Renewable Energy Forecasting**, p. 111–152, 2017. DOI: https://doi.org/10.1016/B978-0-08-100504-0.00004-4

PEDRO, H. T. C. et al. Assessment of machine learning techniques for deterministic and probabilistic intra-hour solar forecasts. **Renewable Energy**, v. 123, p. 191–203, 1 ago. 2018. DOI: https://doi.org/10.1016/j.renene.2018.02.006

PEREZ, R. et al. Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe. **Solar Energy**, v. 94, p. 305–326, ago. 2013. DOI: https://doi.org/10.1016/j.solener.2013.05.005

QING, X.; NIU, Y. Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. **Energy**, v. 148, p. 461–468, abr. 2018. DOI: https://doi.org/10.1016/j.energy.2018.01.177

SEBRAE. **Cadeia de valor da energia solar fotovoltaica no Brasil**. Available in: https://www.sebrae.com.br/Sebrae/Portal%20Sebrae/Anexos/estudo%20energia%20fotovolt%C3%A1ica%20-%20baixa.pdf. Access in: 10 aug. 2021.

SOUZA, R.C.; CAMARGO, M.E. **Análise e previsão de séries temporais**: os modelos ARIMA, 2. ed. 2004.

TRAPERO, J. R.; KOURENTZES, N.; MARTIN, A. Short-term solar irradiation forecasting based on dynamic harmonic regression. **Energy**, v. 84, p. 289–295, maio 2015. DOI: https://doi.org/10.1016/j.energy.2015.02.100

TUOHY, A. et al. Solar forecasting: methods, challenges, and performance. **IEEE Power and Energy Magazine**, v. 13, n. 6, p. 50–59, nov. 2015. DOI: https://doi.org/10.1109/MPE.2015.2461351

URRACA, R. et al. Smart baseline models for solar irradiation forecasting. **Energy Conversion and Management**, v. 108, p. 539–548, jan. 2016. DOI: https://doi.org/10.1016/j.enconman.2015.11.033

VOYANT, C. et al. Machine learning methods for solar radiation forecasting: A review. **Renewable Energy**, v. 105, p. 569–582, may 2017. DOI: http://doi.org/10.1016/j.renene.2016.12.095

WOLPERT, D. The lack of a priori distinctions between learning algorithms. **Neural Computation**. 1996. DOI: https://doi.org/10.1162/neco.1996.8.7.1341

YANG, D. Making reference solar forecasts with climatology, persistence, and their optimal convex combination. **Solar Energy**, 193, pp. 981-985, 2019. DOI: https://doi.org/10.1016/j.solener.2019.10.006

YANG, D. et al. History and trends in solar irradiance and PV power forecasting: a preliminary assessment and review using text mining. **Solar Energy**, v. 168, p. 60–101, jul. 2018. DOI: https://doi.org/10.1016/j.solener.2017.11.023

YANG, D. et al. Reconciling solar forecasts: geographical hierarchy. **Solar Energy**, v. 146, p. 276–286, abr. 2017. DOI: https://doi.org/10.1016/j.solener.2017.02.010

YANG, L.; YANG, X.; HUA, J.; WU, P.; LI, Z.; JIA, D. Very short-term surface solar irradiance forecasting based on fengyun-4 geostationary satellite. **Sensors** 2020, 20, 2606. DOI: https://doi.org/10.3390/s20092606

YE, H. et al. State-of-the-art solar energy forecasting approaches: critical potentials and challenges. **Frontiers in Energy Research**, v. 10, 15 mar. 2022. DOI:

YOUNG, P. C.; PEDREGAL, D. J.; WLODEK. Dynamic harmonic regression. **Journal of forecasting**, v. 18, n. 6, p. 369-394, 1999. DOI: https://doi.org/10.1002/(SICI)1099-131X(199911)18:6<369::AID-FOR748>3.0.CO;2-K

APPENDICES

APPENDIX A – Tables with error metrics for the GHI, ktGHI, DNI and ktDNI forecasts for the testing set with time resolution of 5 min.

Global Horizontal Irradiance		Error Metrics						
Time Horizon	Models	RMSE	nRMSE	R^2	MAE	nMAE	MBE	s
t + 5min	Persistence	94.75	0.21	0.89	52.69	0.12	0.41	
	MARS	69.93	0.16	0.93	39.89	0.09	-2.01	26.19%
	LASSO	69.30	0.16	0.93	40.59	0.09	-2.36	26.86%
	XGBoost	67.72	0.15	0.94	40.40	0.09	-6.38	28.53%
	kNN	69.79	0.16	0.93	39.26	0.09	-1.96	26.34%
	Ensemble_Ridge	68.24	0.15	0.94	40.55	0.09	-3.27	27.98%
t + 30min	Persistence	158.18	0.35	0.71	119.13	0.27	1.32	
	MARS	96.98	0.22	0.87	65.99	0.15	-6.68	38.69%
	LASSO	102.38	0.23	0.86	72.78	0.16	-7.59	35.27%
	XGBoost	97.66	0.22	0.87	66.39	0.15	-11.65	38.26%
	kNN	97.22	0.22	0.87	61.86	0.14	-3.65	38.54%
	Ensemble_Ridge	96.06	0.21	0.88	65.19	0.15	-8.45	39.27%
t + 60min	Persistence	217.44	0.49	0.51	179.73	0.4	8.19	
	MARS	106.26	0.24	0.85	75.92	0.17	-9.84	51.13%
	LASSO	118.96	0.27	0.81	89.49	0.2	-12.62	45.29%
	XGBoost	108.36	0.24	0.84	76.95	0.17	-15.22	50.16%
	kNN	108.11	0.24	0.84	72.94	0.16	-4.71	50.28%
	Ensemble_Ridge	106.14	0.24	0.85	75.4	0.17	-12.03	51.18%

Source: elaborated by the author.

Direct Normal Irradiance		Error Metrics						
Time Horizon	Models	RMSE	nRMSE	R^2	MAE	nMAE	MBE	s
t + 5min	Persistence	143.63	0.36	0.84	68.14	0.17	0.47	
	MARS	101.69	0.25	0.92	56.66	0.14	1.73	29.20%
	LASSO	102.33	0.25	0.91	57.62	0.14	2.89	28.76%
	XGBoost	98.84	0.24	0.92	48.89	0.12	-2.62	31.19%
	kNN	102.07	0.25	0.91	51.89	0.13	-1.70	28.94%
	Ensemble_Ridge	99.73	0.25	0.92	55.70	0.14	1.42	30.57%
t + 30min	Persistence	219.28	0.54	0.66	123.26	0.30	7.20	
	MARS	158.40	0.39	0.79	104.27	0.26	0.15	27.77%
	LASSO	170.14	0.42	0.76	123.69	0.31	8.93	22.41%
	XGBoost	159.17	0.39	0.79	101.99	0.25	-6.37	27.41%
	kNN	159.40	0.39	0.79	102.42	0.25	-1.49	27.31%
	Ensemble_Ridge	157.38	0.39	0.80	103.68	0.26	-3.02	28.23%
t + 60min	Persistence	254.72	0.63	0.56	152.50	0.38	19.37	
	MARS	186.82	0.46	0.71	135.81	0.34	1.94	26.66%
	LASSO	203.39	0.50	0.66	159.38	0.39	13.34	20.15%
	XGBoost	190.82	0.47	0.70	139.11	0.34	-2.47	25.08%
	kNN	190.36	0.47	0.70	135.12	0.33	0.25	25.27%
	Ensemble_Ridge	188.58	0.47	0.71	137.45	0.34	-1.37	25.96%

Source: elaborated by the author.

Global Horizontal Irradiance with Clear-Sky Index		Error Metrics						
Time Horizon	Models	RMSE	nRMSE	R^2	MAE	nMAE	MBE	s
t + 5min	Persistence	94.64	0.21	0.89	52.67	0.12	0.34	
	MARS	68.09	0.15	0.94	38.14	0.09	-2.63	28.06%
	LASSO	69.79	0.16	0.93	40.04	0.09	-2.97	26.26%
	XGBoost	67.60	0.15	0.94	40.04	0.09	-6.35	28.57%
	kNN	68.87	0.15	0.94	40.09	0.09	-4.93	27.22%
	Ensemble_Ridge	68.58	0.15	0.92	40.97	0.09	-4.44	27.54%
t + 30min	Persistence	158.56	0.35	0.71	119.37	0.27	0.87	
	MARS	101.51	0.23	0.86	71.40	0.16	-11.60	35.98%
	LASSO	146.76	0.33	0.72	118.02	0.26	-25.72	7.44%
	XGBoost	97.22	0.22	0.87	66.33	0.15	-12.61	38.69%
	kNN	99.61	0.22	0.86	67.15	0.15	-8.93	37.18%
	Ensemble_Ridge	102.4	0.23	0.82	72.51	0.16	-13.17	35.42%
t + 60min	Persistence	217.94	0.49	0.50	180.24	0.40	7.47	
	MARS	110.69	0.25	0.84	82.38	0.18	-15.99	49.21%
	LASSO	148.27	0.33	0.71	119.10	0.27	-25.61	31.97%
	XGBoost	107.90	0.24	0.84	76.93	0.17	-15.20	50.49%
	kNN	111.02	0.25	0.83	77.28	0.17	-11.44	49.06%
	Ensemble_Ridge	114.83	0.26	0.77	83.88	0.19	-17.67	47.31%

Source: elaborated by the author.

Direct Normal Irradiance with Clear-Sky Index		Error Metrics						
Time Horizon	Models	RMSE	nRMSE	R^2	MAE	nMAE	MBE	s
t + 5min	Persistence	143.56	0.35	0.84	68.13	0.17	0.43	
	MARS	101.55	0.25	0.92	54.11	0.13	1.06	29.26%
	LASSO	101.70	0.25	0.91	56.29	0.14	3.37	29.16%
	XGBoost	98.87	0.24	0.92	51.77	0.13	-0.41	31.13%
	kNN	103.04	0.25	0.91	53.75	0.13	0.17	28.22%
	Ensemble_Ridge	102.30	0.25	0.92	56.69	0.14	1.47	28.74%
t + 30min	Persistence	219.40	0.54	0.66	123.38	0.30	6.98	
	MARS	159.89	0.40	0.79	106.81	0.26	3.25	27.12%
	LASSO	168.36	0.42	0.77	121.92	0.30	12.96	23.26%
	XGBoost	157.81	0.39	0.79	101.76	0.25	-5.05	28.07%
	kNN	160.29	0.40	0.79	104.15	0.26	0.35	26.94%
	Ensemble_Ridge	166.25	0.41	0.78	109.22	0.27	-2.46	24.22%
t + 60min	Persistence	255.15	0.63	0.56	152.82	0.38	19.01	
	MARS	191.07	0.47	0.70	142.99	0.35	2.46	25.11%
	LASSO	202.44	0.50	0.67	158.93	0.39	19.08	20.66%
	XGBoost	189.18	0.47	0.71	136.93	0.34	-4.70	25.86%
	kNN	191.05	0.47	0.70	138.03	0.34	1.17	25.12%
	Ensemble_Ridge	203.13	0.50	0.67	145.73	0.36	-4.00	20.39%

Source: elaborated by the author.

APPENDIX B – Tables with error metrics for the GHI, ktGHI, DNI and ktDNI forecasts for the testing set with time resolution of 60 min.

Global Horizontal Irradiance		Error Metrics						
Time Horizon	Models	RMSE	nRMSE	R^2	MAE	nMAE	MBE	s
t + 60min	Persistence	176.93	0.43	0.64	154.51	0.38	16.97	
	MARS	58.76	0.14	0.96	39.26	0.10	-10.80	66.79%
	LASSO	67.14	0.16	0.95	48.68	0.12	-22.61	62.05%
	XGBoost	72.75	0.18	0.95	54.36	0.13	-33.31	58.88%
	kNN	64.81	0.16	0.95	44.57	0.11	-11.36	63.37%
	Ensemble_Ridge	64.25	0.16	0.95	47.04	0.11	-22.94	63.68%
t + 6h	Persistence	443.85	1.08	0.02	362	0.88	262.5	
	MARS	144.14	0.35	0.84	110.76	0.27	-89.75	67.52%
	LASSO	130.83	0.32	0.87	107.29	0.26	-69.88	70.52%
	XGBoost	139.99	0.34	0.85	109	0.27	-85.31	68.46%
	kNN	134.04	0.33	0.82	101.08	0.25	-60.8	69.80%
	Ensemble_Ridge	135.05	0.33	0.86	106.06	0.26	-82.66	69.57%
t + 12h	Persistence	496.93	1.21	0.08	421.11	1.03	419.93	
	MARS	165.23	0.40	0.84	134.26	0.33	-112.16	66.75%
	LASSO	140.31	0.34	0.84	117.65	0.29	-74.88	71.76%
	XGBoost	158.4	0.39	0.82	125.96	0.31	-101.92	68.12%
	kNN	132.54	0.32	0.82	103.77	0.25	-58.48	73.33%
	Ensemble_Ridge	151.94	0.37	0.84	122.65	0.30	-97.46	69.42%

Source: elaborated by the author.

Direct Normal Irradiance		Error Metrics						
Time Horizon	Models	RMSE	nRMSE	R^2	MAE	nMAE	MBE	s
t + 60min	Persistence	162.95	0.41	0.78	101.32	0.25	24.53	
	MARS	129.01	0.32	0.84	87.53	0.22	-24.20	20.83%
	LASSO	137.29	0.34	0.82	97.68	0.25	-23.95	15.75%
	XGBoost	141.18	0.35	0.82	98.43	0.25	-40.50	13.36%
	kNN	136.65	0.34	0.82	93.41	0.23	-16.39	16.14%
	Ensemble_Ridge	133.68	0.34	0.84	93.65	0.24	-33.58	17.96%
t + 6h	Persistence	386.53	0.97	0.15	260.91	0.66	202.76	
	MARS	306.16	0.77	0.36	253.61	0.64	-165.70	20.79%
	LASSO	261.8	0.66	0.40	220.86	0.55	-86.82	32.27%
	XGBoost	324.81	0.82	0.20	268.49	0.67	-143.04	15.97%
	kNN	278.56	0.70	0.33	220.51	0.55	-83.04	27.93%
	Ensemble_Ridge	327.29	0.82	0.18	268.24	0.67	-136.94	18.38%
t + 12h	Persistence	470.05	1.18	0.01	335.70	0.84	333.53	
	MARS	328.76	0.83	0.30	293.47	0.74	-191.23	30.06%
	LASSO	285.05	0.72	0.29	249.43	0.63	-93.59	39.36%
	XGBoost	347.48	0.87	0.10	289.62	0.73	-146.65	26.08%
	kNN	275.48	0.69	0.30	227.94	0.57	-70.20	41.39%
	Ensemble_Ridge	350.15	0.88	0.10	291.43	0.73	-146.28	25.51%

Source: elaborated by the author.

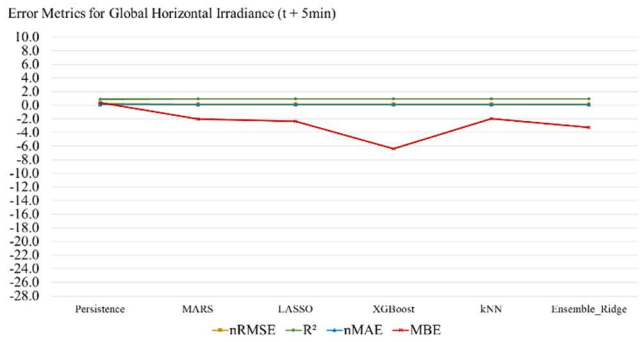
Global Horizontal Irradiance with Clear-Sky Index		Error Metrics						
Time Horizon	Models	RMSE	nRMSE	R^2	MAE	nMAE	MBE	s
t + 60min	Persistence	203.82	0.50	0.52	168.50	0.41	10.14	
	MARS	128.62	0.31	0.81	59.00	0.14	-28.9	36.89%
	LASSO	158.84	0.39	0.75	99.46	0.24	-59.95	22.07%
	XGBoost	139.06	0.34	0.79	67.53	0.16	-40.95	31.77%
	kNN	135.94	0.33	0.80	70.97	0.17	-36.98	33.30%
	Ensemble_Ridge	129.3	0.32	0.81	52.75	0.13	-32.15	36.56%
t + 6h	Persistence	466.42	1.14	0.00	387.36	0.94	258.08	
	MARS	158.25	0.39	0.76	97.65	0.24	-60.55	66.07%
	LASSO	176.72	0.43	0.70	116.17	0.28	-73.97	62.11%
	XGBoost	203.29	0.50	0.68	138.86	0.34	-110.77	56.42%
	kNN	171.76	0.42	0.74	113.40	0.28	-77.5	63.17%
	Ensemble_Ridge	318.54	0.78	0.35	126.81	0.31	-96.67	31.71%
t + 12h	Persistence	497.35	1.21	0.00	420.32	1.02	418.89	
	MARS	164.75	0.40	0.74	106.24	0.26	-62.21	66.87%
	LASSO	174.61	0.43	0.73	121.17	0.30	-78.94	64.89%
	XGBoost	206.29	0.50	0.66	144.97	0.35	-111.75	58.52%
	kNN	172.85	0.42	0.73	117.07	0.29	-76.43	65.24%
	Ensemble_Ridge	337.74	0.82	0.27	145.09	0.35	-109.48	32.09%

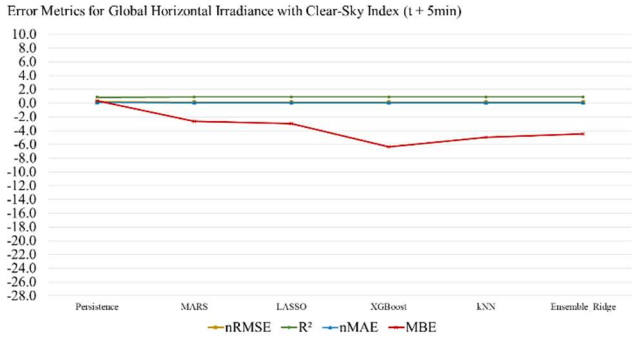
Source: elaborated by the author.

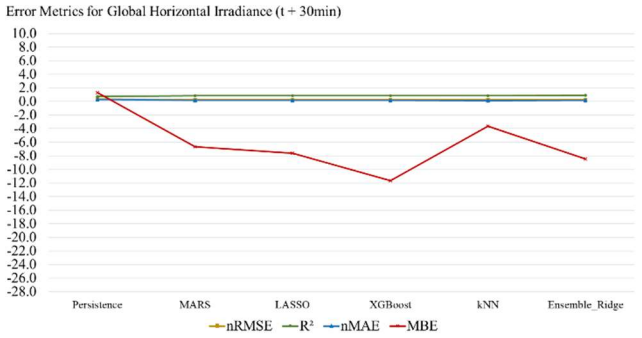
Direct Normal Irradiance with Clear-Sky Index		Error Metrics						
Time Horizon	Models	RMSE	nRMSE	R^2	MAE	nMAE	MBE	s
t + 60min	Persistence	172.13	0.43	0.75	106.37	0.27	19.91	
	MARS	139.38	0.35	0.82	94.21	0.24	-27.33	19.03%
	LASSO	144.64	0.36	0.80	98.86	0.25	-23.43	15.97%
	XGBoost	142.94	0.36	0.81	98.13	0.25	-36.42	16.96%
	kNN	153.16	0.38	0.77	107.39	0.27	-17.28	11.02%
	Ensemble_Ridge	149.33	0.38	0.78	99.99	0.25	-31.65	13.25%
t + 6h	Persistence	392.49	0.99	0.12	265.45	0.67	195.47	
	MARS	264.65	0.66	0.41	216.71	0.54	-100.50	32.57%
	LASSO	267.20	0.67	0.37	223.47	0.56	-87.94	31.92%
	XGBoost	327.62	0.82	0.17	260.52	0.65	-130.92	16.53%
	kNN	281.39	0.71	0.30	225.13	0.57	-86.97	28.31%
	Ensemble_Ridge	330.30	0.83	0.15	264.37	0.66	-127.62	15.85%
t + 12h	Persistence	467.47	1.17	0.00	333.80	0.84	332.81	
	MARS	282.07	0.71	0.38	249.18	0.63	-127.45	39.66%
	LASSO	294.77	0.74	0.24	263.63	0.66	-95.08	36.94%
	XGBoost	329.75	0.83	0.14	279.04	0.70	-135.59	29.46%
	kNN	288.21	0.72	0.25	239.58	0.60	-82.44	38.35%
	Ensemble_Ridge	343.95	0.86	0.08	289.99	0.73	-135.38	26.42%

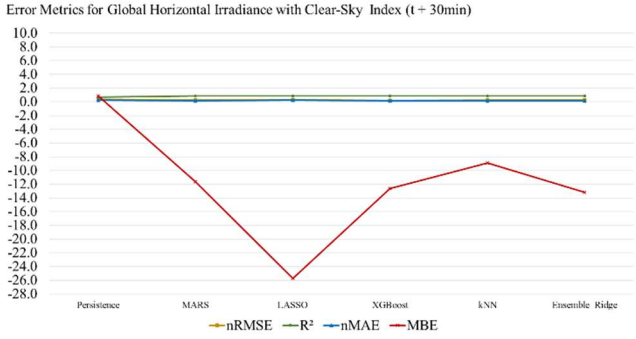
Source: elaborated by the author.

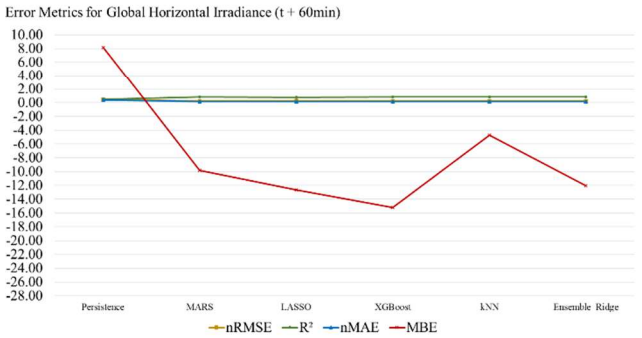
APPENDIX C – Error Metric Graphics for GHI, ktGHI, DNI and ktDNI forecasts (testing set), with a resolution of 5 min.

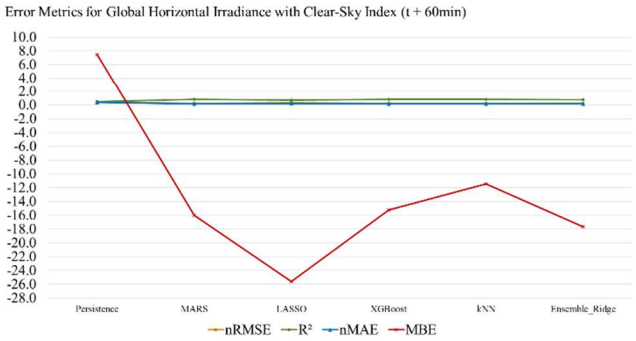




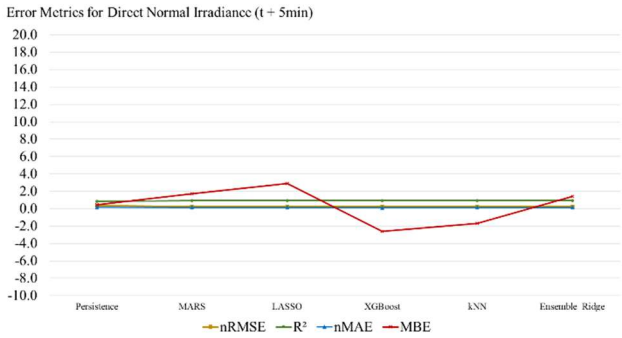


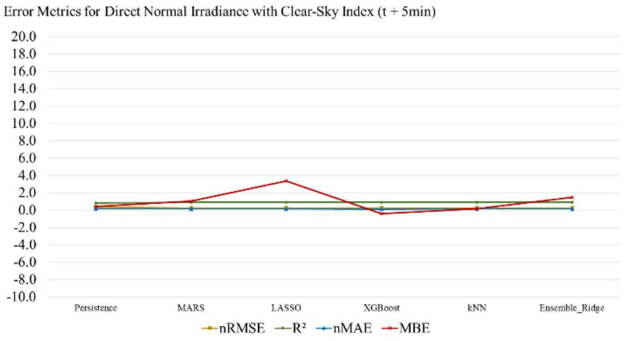


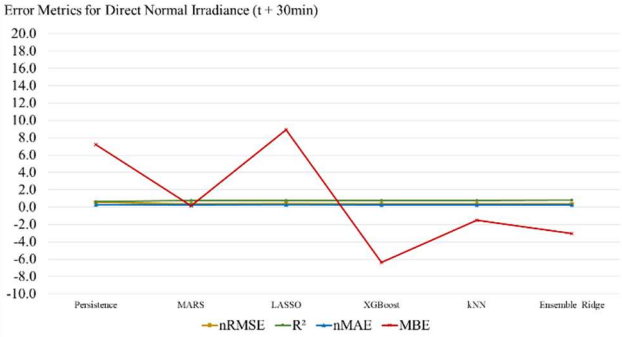


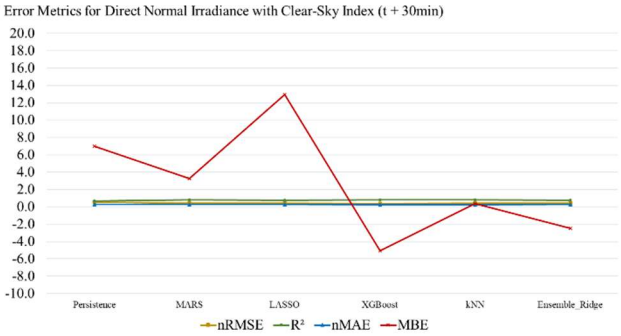


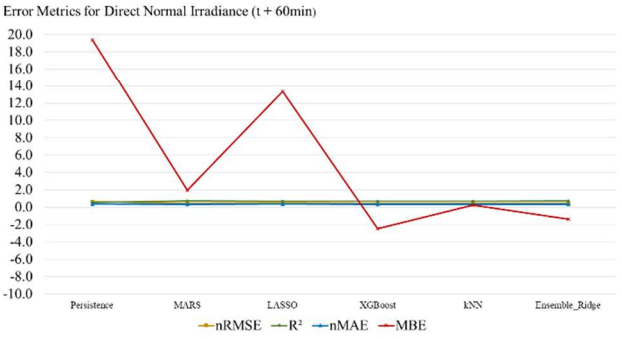
Source: elaborated by the author.

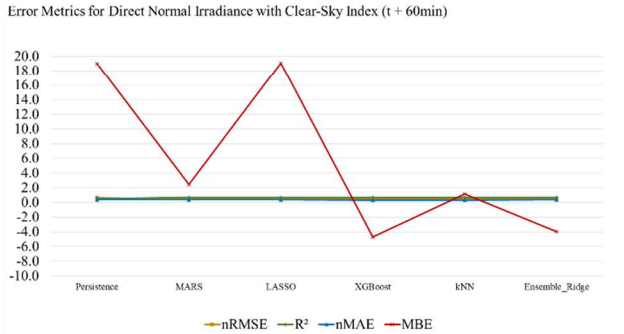






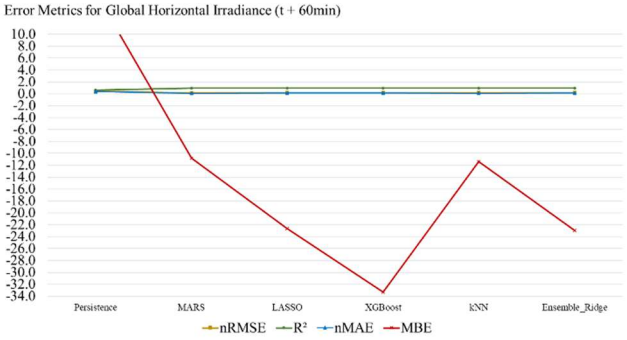


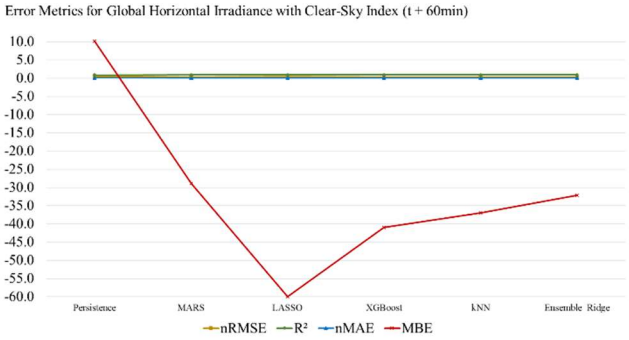


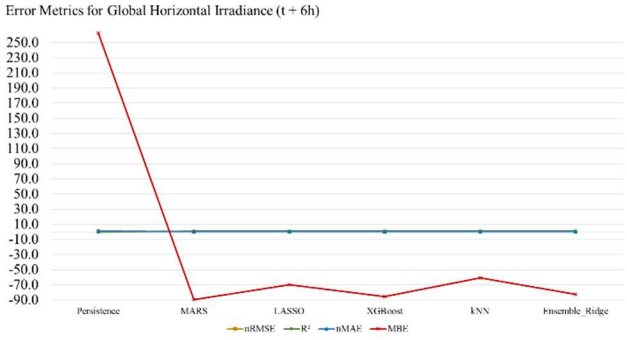


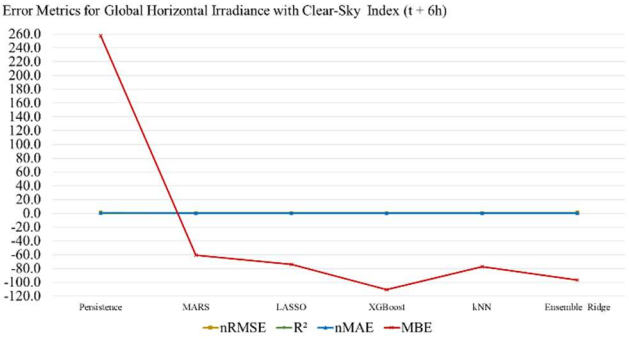
Source: elaborated by the author.

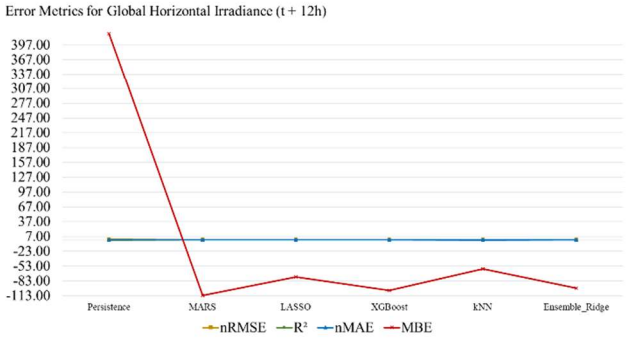
APPENDIX D – Error Metric Graphics for GHI, ktGHI, DNI and ktDNI forecasts (testing set), with a resolution of 60 min.

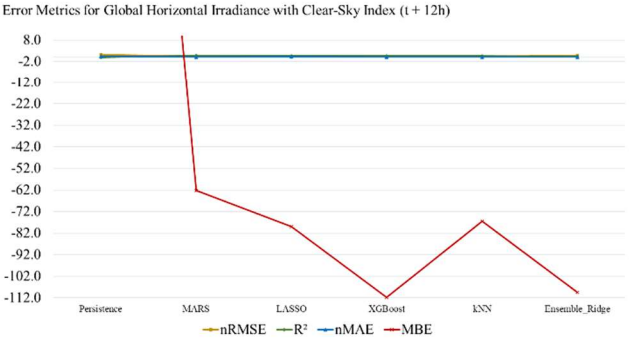




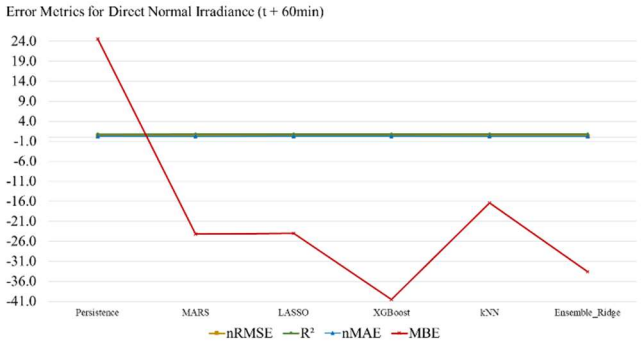


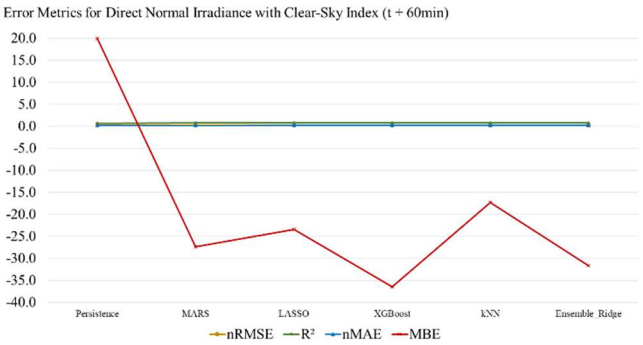


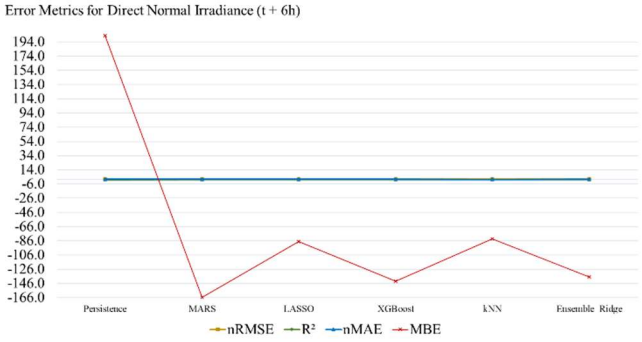


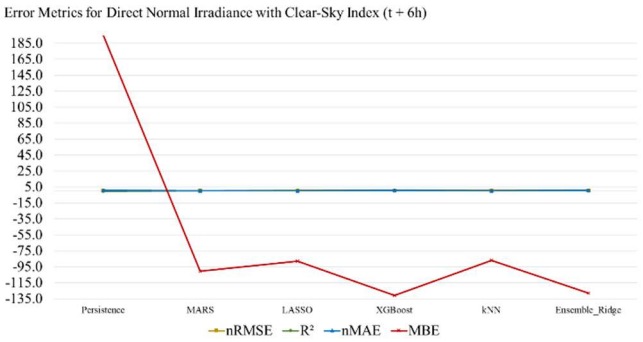


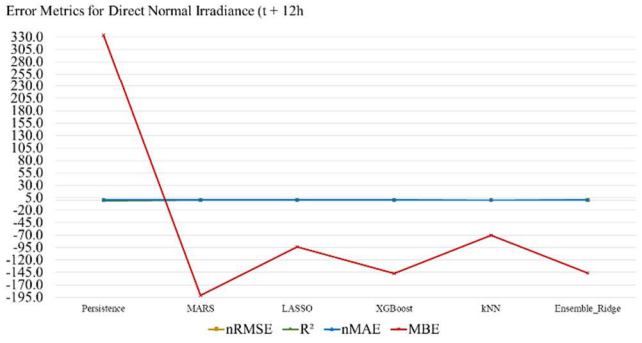
Source: elaborated by the author.

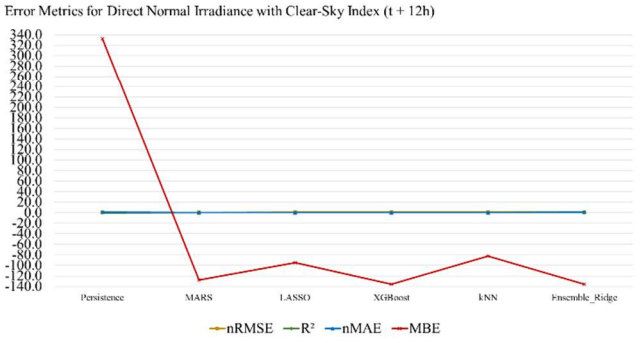












Source: elaborated by the author.