NOVA IMS Information Management School

# MDSAA

Master Degree Program in

**Data Science and Advanced Analytics**

# Forecasting Renewable Energy Production in Madeira Island

Gonçalo Lagos Ferreira

Project Work

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**
Universidade Nova de Lisboa

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**
Universidade Nova de Lisboa

# Forecasting Renewable Energy Production in Madeira Island

by

Gonçalo Lagos Ferreira

Project Work presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Business Analytics.

## Supervised by

Vítor Manuel Pereira Duarte dos Santos, PhD, NOVA Information Management School

Maria Anastasiadou, MSc, NOVA Information Management School

July 2025

# STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, July 2025.

Gonçalo Ferreira

# DEDICATION

To the most important people in my growth and every life achievement — my mom and dad. I would not be here without your unwavering hard work and countless sacrifices. Every milestone I've reached is more yours than mine, because these were only made possible through your unconditional love, tireless effort, and support from the very moment I was born until today. No words or actions could ever truly express the depth of my gratitude and appreciation for all you have done for me. From the bottom of my heart, thank you.

To my other two girls who stood by my side through every moment of this journey , the best sister and girlfriend I could have ever asked for, Sara and Carolina. Completing my master's thesis while working was incredibly challenging, but your constant help, support, and patience made what felt impossible, possible. You cooked for me, cared for me on my worst days, and took on every domestic task to ensure I had the time and space to focus on this project. This achievement is as much yours as it is mine. Thank you, from the bottom of my heart.

To my grandmothers, Luísa and Teresa, thank you for the love, joy, and wisdom you've given me throughout my life. Your presence and guidance have meant more than words can express.

To my godparents, Tininha and João , thank you for all the camaraderie, laughter, and advice. I've learned so much from you over the years, and I wouldn't be where I am today without your support.

To all my other relatives not individually mentioned — uncles, aunts, and cousins — I've been fortunate to never lack love or friendship, and for that, I thank you all.

To my friend Susana, who always supported me, especially throughout this journey at NOVA IMS. Although you were a year ahead, your friendship and guidance were invaluable. Without your help, I would have never ventured into the world of Data Science. Truly, thank you.

To the new friends I made on this journey at NOVA IMS — Alex, Hugo, Filipe, and Sebastião. You were, without a doubt, the best groupmates I've ever had. I learned so much from you, both technically and personally. Working with you not only made me a better data scientist, but also a better person. Thank you.

To my lifelong friend Francisco, whom I've known , and who has stood by me for as long as I can remember. Thank you for all the inconditional friendship.

To all my dear friends, whose names are too many to list. I've been truly fortunate to have friends who supported me and made me laugh when I needed it most. Thank you for always being there.

# ACKNOWLEDGEMENTS

To my supervisor, Professor Vítor Santos, I would like to express my sincere gratitude for your guidance and support throughout the most critical stages of this project. Your expertise and highly valuable suggestions made a fundamental difference in shaping the quality, strategy and outcome of this work. Thank you.

To my co-supervisor, Professor Maria Anastasiadou, I am deeply grateful for all the knowledge and insights you shared at every stage of this project. Your expertise in research and energy-related topics has helped me grow into a more competent student with each of our sessions. Thank you for your unwavering support, the generous time you dedicated to our weekly meetings, and for your constant availability to assist me whenever needed.

To Empresa de Eletricidade da Madeira, I express my heartfelt gratitude for exceeding my expectations by providing essential, diverse and high-quality data that was central to this project.

I would like to sincerely thank Eng. José Estêvão Abreu, Director of Regional Energy, whose support was instrumental in shaping the research strategy and bridging the gap between my abstract ideas and possible real-world applications. Your guidance and thoughtful advice were truly invaluable.

I am also deeply grateful to Eng. Agostinho Figueira, Director of Studies and Planning of EEM, for generously taking the time to meet with me, discuss potential research directions, and facilitate access to the data that made this work possible.

To all the professors at NOVA IMS who supported me throughout this journey. Their unwavering willingness to provide guidance, even beyond scheduled class hours, did not go unnoticed. I am truly grateful for their dedication, generosity, and commitment to the academic success of all students.

# ABSTRACT

The global shift toward sustainability has become increasingly urgent as reliance on fossil fuels accelerates climate change, pollutes ecosystems, and depletes finite resources. Yet, modern society depends on these energy sources for essential needs, given their stable nature. Balancing the demand for reliable power with the need for sustainable solutions is one of today's greatest challenges. Renewable energy offers a cleaner alternative, as it neither harms the environment nor relies on finite resources. However, its primary limitation is weather related intermittency, which poses a significant challenge for Madeira Island, a Portuguese region that remains more dependent on fossil fuels than its mainland. This increased reliance emerges from the island's status as an isolated electrical system, where the need for a stable and reliable energy supply is critical. In response, this research aimed to develop a practical and effective forecasting framework to support accurate renewable energy predictions and contribute to Madeira's decarbonization efforts. This case study leveraged solar, wind, hydropower, and biomass energy data, provided by Empresa de Eletricidade da Madeira, the region's main utility company, and complemented by weather data retrieved from Visual Crossing Weather API, to develop a clear and practical forecasting framework. The proposed solution consisted of day-ahead forecasts of XGBoost for wind and hydropower energy sources, and Random Forest for solar and biomass energy sources, all with grid search defined parameters, delivering the most accurate day-ahead forecasts of this research. To assess this solution's performance, error metrics and visualization methods (line chart and SHAP values) were employed. Although performance varied across energy sources, the model achieved strong results for hydropower, acceptable outcomes for wind, moderate but cautious results for biomass, and underwhelming performance for solar. Nonetheless, this research delivered a comprehensive and applicable forecasting framework, not only for Madeira Island but also for other regions aiming to apply time series techniques to enhance their reliance on renewable energy. In doing so, it contributed to the United Nations' 7th Sustainable Development Goal: ensuring access to affordable and clean energy for all.



# KEYWORDS

Renewable Energy Production; Energy Analysis; Machine Learning; Time Series Analysis


## Sustainable Development Goals (SDG):

# INDEX


1. Introduction ..... 1
   1.1 Context and Problem Identification ..... 1
   1.2 Objectives ..... 6
   1.3 Study Outcomes and contributions ..... 7
2. Literature review ..... 8
   2.1 AI applications in forecasting renewable energy production ..... 8
   2.2 Literature Review Methodology ..... 9
      2.2.1. Data Selection ..... 11
      2.2.2. Results and analysis ..... 15
         2.2.2.1. Keyword co-occurrence. ..... 19
      2.2.3. Discussion ..... 22
         2.2.3.1. Common approaches ..... 23
         2.2.3.2. Solar energy applications ..... 24
         2.2.3.3. Wind energy applications ..... 26
         2.2.3.4. Solar and wind energy applications. ..... 27
         2.2.3.5. Hydropower applications ..... 27
   2.3 Summary ..... 28
3. Methodology ..... 32
   3.1 Overview ..... 32
   3.2 Exploration Phase ..... 33
   3.3 Analytical Phase ..... 33
   3.4 Conclusive Phase ..... 36
4. Empirical Study ..... 37
   4.1 Data Understanding ..... 38
      4.1.1. Energy Production Dataset ..... 38
      4.1.2. Weather ..... 43
   4.2 Data Preparation ..... 47
      4.2.1. Merging all datasets ..... 47
      4.2.2. Data Transformation ..... 48
      4.2.3. Feature Selection ..... 53
   4.3 Modeling ..... 55
      4.3.1. First forecasting round ..... 56

4.3.2. Grid Search and Blocked-Cross Validation round ..... 58
4.3.3. Final forecasting Round ..... 60
4.4 Discussion ..... 62
4.4.1. Solar energy ..... 64
4.4.2. Wind energy ..... 64
4.4.3. Hydropower ..... 66
4.4.4. Biomass energy ..... 67
5. Conclusion ..... 68
5.1 Synthesis of the developed case study ..... 68
5.2 Limitations and recommendations for future works ..... 69
Bibliographical References ..... 71
Appendix A ..... 77
Appendix B ..... 80
Appendix C ..... 85

# LIST OF FIGURES

Figure 1.1- Global fossil fuel consumption (Our World in Data, 2024) ..... 2
Figure 1.2- Annual temperature anomaly, from 1850 until 2023 (Berkeley Earth, 2024). ..... 2
Figure 1.3 - Electricity generation per energy source in the Autonomous Region of Madeira (APREN, 2019) ..... 5
Figure 2.1- Neural Network (Campos et al., 2024) ..... 9
Figure 2.2- PRISMA Methodology ..... 10
Figure 2.3- PRISMA flowchart ..... 14
Figure 2.4- Keywords Co-occurrence Network ..... 20
Figure 2.5- Keywords Co-occurrence by year ..... 21
Figure 3.1- Methodological path ..... 33
Figure 4.1- Monthly evolution of renewable vs non-renewable energy production ..... 40
Figure 4.2- Yearly evolution of renewable vs non-renewable energy installed capacity ..... 40
Figure 4.3- Monthly percentage contribution of renewable sources ..... 41
Figure 4.4- Distribution of energy production by renewable sources ..... 41
Figure 4.5- Selected weather stations on Madeira Island ..... 44
Figure 4.6- Sine and Cosine transformations (Kapoor, 2024). ..... 50
Figure 4.7- Spearman correlation heatmap ..... 51
Figure 4.8- Min Max Scaler (adapted (Dralus et al., 2023)) ..... 53
Figure 4.9- Rationale of BCV (Cabello-López et al., 2023). ..... 58
Figure 4.10- Actual vs. forecasted renewable energy generation, in MWh. ..... 63
Figure 4.11- Beeswarm plots showing feature impacts for each RES-model combination ..... 64
Figure 4.12-Madeira Island's wind rose ..... 66
Figure C.1 – FW_Funchal boxplots ..... 85
Figure C.2 - LPMA boxplots ..... 86

# LIST OF TABLES

Table 1.1- Main types of Renewable Energy Sources (adapted from (*United Nations- Climate Action*, 2024) ..... 3
Table 2.1- Established Inclusion Criteria ..... 11
Table 2.2- Initial Results ..... 12
Table 2.3- Journal details and their Scimago Ranks ..... 15
Table 2.4- Number of citations and the country of origin for each article's data. ..... 17
Table 2.5- Keywords co-occurrence and the total link strengths ..... 19
Table 2.6- Principal Keywords of each cluster and respective colour ..... 21
Table 2.7- PRISMA methodology selected articles ..... 29
Table 3.1- Phases of CRISP-DM methodology and tasks to do in each Phase (Wirth & Hipp, 2000) ..... 34
Table 4.1- Information about selected weather stations ..... 44
Table 4.2- Variables with missing values ..... 45
Table 4.3- Motives to discard variables ..... 46
Table 4.4- Train test split performed ..... 49
Table 4.5- Feature engineered time variables ..... 49
Table 4.6- Selected features per source. ..... 54
Table 4.7- Description of each error metric (adapted from (Campos et al., 2024)). ..... 55
Table 4.8- First round error metrics results, in MWh ..... 57
Table 4.9- Second round results, in MWh ..... 59
Table 4.10- Final round results ..... 61
Table 4.11- Error metrics of the proposed hybrid solution ..... 63
Table A.1 - Metadata of features provided by EEM ..... 77
Table B.1- FW_Funchal weather variables' metadata ..... 80
Table B.2 - LPMA weather variables' metadata ..... 82

# LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| **ACF** | Autocorrelation Function |
| **AI** | Artificial Intelligence |
| **ARIMA** | Autoregressive Integrated Moving Average |
| **BCV** | Blocked Cross-Validation |
| **CNN** | Convolutional neural networks |
| **CRISP-DM** | Cross-industry standard process for data mining |
| **DL** | Deep Learning |
| **DST** | Daylights Saving Time |
| **EEM** | Empresa de Eletricidade da Madeira |
| **EU** | European Union |
| **FW_Funchal** | FW1564 Funchal PT |
| **LSTM** | Long Short-Term Memory |
| **MAE** | Mean Absolute Error |
| **ML** | Machine Learning |
| **MLP** | Multilayer Perceptron |
| **MW** | Megawatts |
| **MWh** | Megawatts- hour |
| **nRMSE** | normalized Root Mean Squared Error |
| **PACF** | Partial Autocorrelation Function |
| **PRISMA** | Preferred Reporting Items for Systematic Reviews and Meta-Analyses |
| **RES** | Renewable Energy Sources |
| **RF** | Random Forest |
| **RIS** | Research Information Systems |
| **RMSE** | Root Mean Squared Error |

| | |
|---|---|
| **RQ** | Research Question |
| **SARIMA** | Seasonal Auto-Regressive Integrated Moving Average |
| **SARIMAX** | Seasonal Autoregressive Integrated Moving Average with Exogenous Factors |
| **SHAP** | Shapley Additive Explanations |
| **SLR** | Systematic Literature Review |
| **XGBoost** | Extreme Gradient Boosting |

# 1. INTRODUCTION

Currently, there is widespread conversation about progress towards sustainability. The growing reliance on fossil fuels is posing a serious threat to the health of our planet, accelerating climate change, polluting ecosystems, and depleting finite resources (Shahzad, 2015). At the same time, humanity depends on energy to meet its most essential needs, from powering homes and transportation to supporting food production and healthcare systems (Deshmukh et al., 2023). Balancing the urgent need for sustainable energy solutions with society's demand for reliable power has become one of the greatest challenges of our time. It is believed that the way to a greener and more environmentally friendly world is by embracing renewable energy, instead of fossil fuels. Renewable energy is not harmful for our planet, nor does its resources deplete. However, as will be see more in depth, there are some constraints associated with these sources, predominantly their intermittence (Deshmukh et al., 2023).

This research intended to attenuate this intermittency, by producing high quality and accurate renewable energy source's forecasts, to apply to real-world scenarios. Ultimately, this would enable a better integration of these renewable sources.

To achieve these high-quality forecasts, this thesis embodied a Work Project, focusing on Madeira Island, using Empresa de Eletricidade da Madeira's data. However, it is expected that the knowledge produced in this research can contribute to an overall increase in the reliance on renewable energy in other locations.

## 1.1 CONTEXT AND PROBLEM IDENTIFICATION

Transitioning towards a sustainable world is a major topic nowadays (*European Green Deal*, 2019). Ever since the first industrial revolution took place, in Great Britain, around 1750s, the planet Earth has been experiencing unprecedented environmental transformations (Frisken, 1971). This phenomenon is called climate change. Climate change refers to long term shifts in temperature and weather patterns (*United Nations- Climate Action*, 2024). Climate change is a consequence of the utilization of fossil fuels, which were key factors to human industrialization and advancements, ever since the first industrial revolution.

Fossil fuels are carbon based non-renewable energy sources, originated by non-renewable wastes, that take millions of years to form. The primary fossil fuels are coal, natural gas and oil. These are widely used due to their relatively cheap price, consistent and stable energy supply, and to their energy efficiency, as these contain a large amount of energy per unit (Bi et al., 2024).

Fossil fuels have played a vital role in several areas, such as industrialization (enabling mass production), transportation (development of automobiles, airplanes and railways) and electricity generation. All these reasons sustain why fossil fuels provide around 80% of the

world's energy consumption (Bi et al., 2024). Ever since fossil fuels first began to drive these developments, their consumption has never decreased.

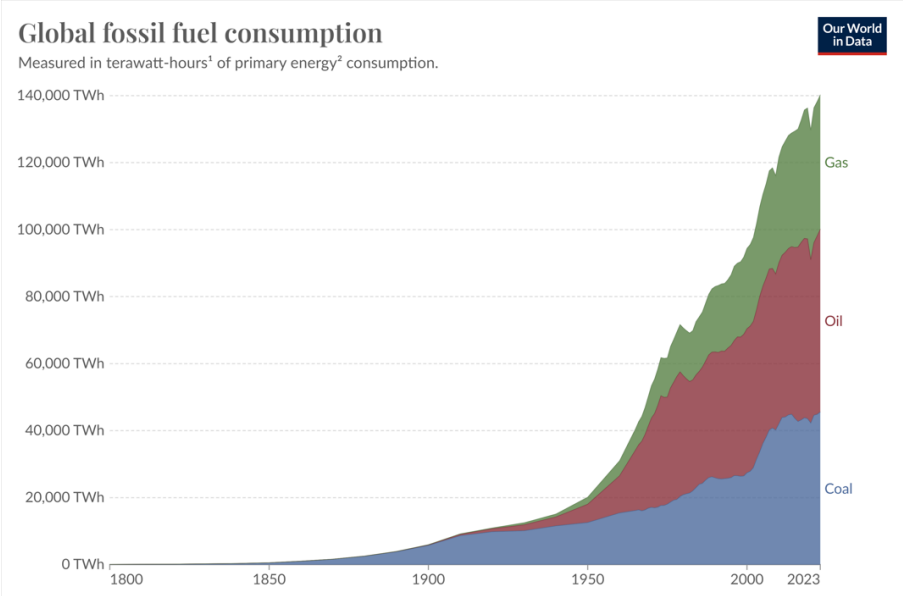


Figure 1.1- Global fossil fuel consumption (Our World in Data, 2024)

Nonetheless, fossil fuels need to be burnt to generate energy, and when they burn, they release large amounts of harmful gases, with the most noteworthy being Carbon Dioxide. This gas is the main responsible for climate change (Shahzad, 2015). The consequences of climate change are devastating, such as increase of temperature, endangerment of the lives of species on our planet, profound weather changes, melting of polar ice caps, sea level rise, acidic rain, among many others (*United Nations- Climate Action*, 2024), (Lemoine, 2018). Figure 1.2 shows that the annual temperature deviation has been recording the same positive trend as fossil fuels' consumption.

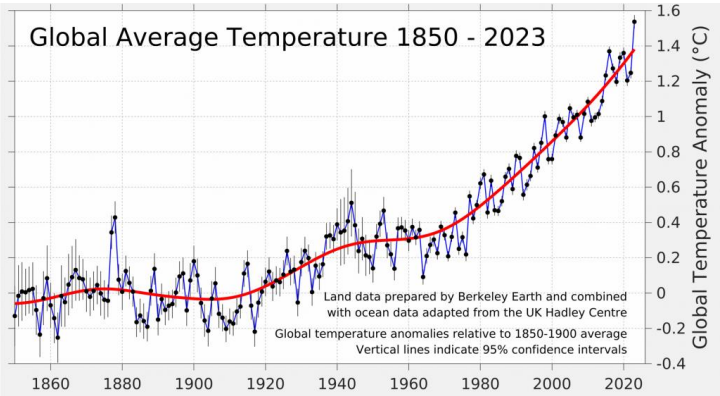


Figure 1.2- Annual temperature anomaly, from 1850 until 2023 (Berkeley Earth, 2024).

Beyond the environmental impacts, there is also a limitation in the quantity of fossil fuels available for use. As previously mentioned, fossil fuels are non-renewable energy sources, meaning they can be permanently exhausted by its consumption (Deshmukh et al., 2023).

Energy is an absolute must for ensuring that people can fulfill their primary needs, such as cooking, heating and transportation. In addition, energy demand is increasing day by day, due to world population growth. In that capacity, the need to grow the energy supply also arises, minding the need to protect our planet and to prevent the devastating effects of climate change. Therefore, it is imperative to explore alternative energy options, to meet the population's energy demands, in a more environmentally friendly manner (Deshmukh et al., 2023).

The increased urge of cleaner and limitless energy sources as led to the increased use of renewable energy sources (RES) over the last years (Chen et al., 2021).

Renewable energy sources are forms of energy derived from natural sources, which are replenished at a higher rate than they are consumed. The feature that makes a sustainable energy system stand out is the ability to ensure the fulfillment of the needs without diminishing the existing resources, unlike fossil fuels. The energy produced via these sources can be directly integrated within power grids or stored in batteries for later utilization (Deshmukh et al., 2023).

These sources are deemed as one of the best ways to decarbonize the energy market and reduce dependence on fossil fuels. Table 1.1 provides a brief breakdown and description of the main types of RES:

Table 1.1- Main types of Renewable Energy Sources (adapted from (*United Nations- Climate Action*, 2024)

| **Type of Renewable Energy Source** | **Description** |
|---|---|
| 1. Hydropower | Gravitational potential energy of water is converted into electrical energy through a hydraulic turbine. |
| 2. Wind energy | Kinetic energy of wind is converted into electricity by wind turbines. |
| 3. Solar energy | The sun's energy is turned into electricity via panels/ solar heaters. |
| 4. Biomass energy | Energy via the combustion of plants and animal remains. |

| | |
|---|---|
| 5. Geothermal energy | Heat energy trapped underneath the earth's crust, converted into electricity via steam turbines. |
| 6. Ocean energy | Thermal and tidal energy from the ocean, converted into electricity by turbines and other systems. |

As demonstrated, renewable energy is derived from natural limitless sources. The implementation of these sources is significant for Earth's green development. In addition, due to its nature, these are easy to capture. Compared with fossil fuels, RES have much lesser negative impact, and do not release any harmful gases (Deshmukh et al., 2023)

Regardless of the benefits offered by RES, these accounted only for 29.1% of the global electricity generation, in 2022, while fossil fuels' consumption continued to rise (*Our World in Data- Renewable Energy*, 2024).

Despite all the advantages enabled by renewable energy, there are several barriers, such as high initial costs, equipment stability, and most importantly, weather parameter's unpredictability. RES are profoundly influenced by the experienced weather conditions. For instance, wind energy cannot be generated without wind, and the output of solar energy will be lower on a cloudy day. Although the weather predictions are more accurate nowadays these are still not perfect. For this reason, RES do not provide a steady energy supply into the grid, being deemed as intermittent (Deshmukh et al., 2023).

Energy is vital to every aspect of life, which is why humanity needs a constant and reliable energy supply, primarily derived from fossil fuels. However, this non-renewable source is detrimental to our planet and will ultimately be depleted (Shahzad, 2015).

Madeira Island, a small Portuguese island located in the North Atlantic Ocean, is not an exception to this energy reality. Despite all the efforts that have been conducted, only 40% of Madeira's energy production comes from RES, mostly due to climate variability, but also to the fact that it operates as an isolated energy system, while having to ensure a reliable energy supply to its inhabitants (José Martins Silva, 2014). In comparison, 75% of Portugal's mainland energy production comes from renewable sources.

Madeira would benefit from priorly knowing what the energy production, per RES would be, to foster the dependence on these clean sources, and to better integrate these into the power grids. These forecasts can be achieved via time series analysis.

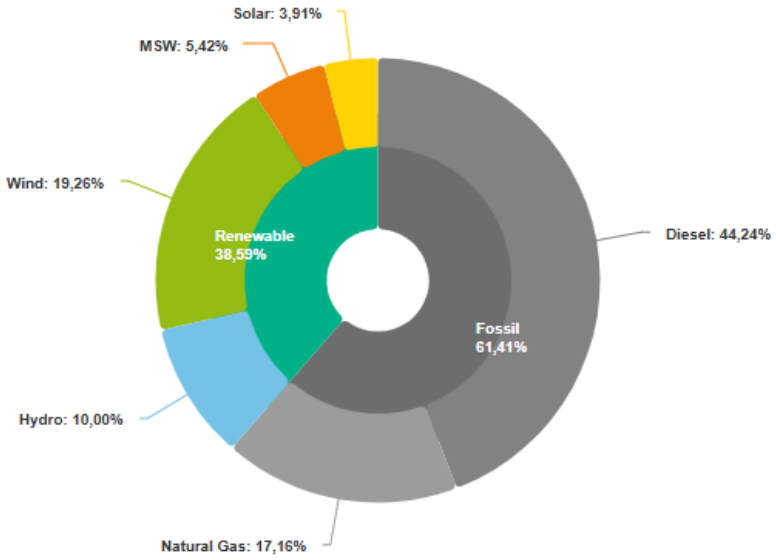


Figure 1.3 - Electricity generation per energy source in the Autonomous Region of Madeira (APREN, 2019)

A forecast is the development of unknown future information in a certain time horizon, using available information (Sweeney et al., 2020). More precisely, a forecast is defined as a single value to be expected in the future, given a particular model's criteria. Most forecasting problems rely on time series data, which consist in a sequence of chronological observations on a variable to study (Box et al., 2016). A proper model is fitted to the variable to be predicted, and the parameters are estimated using the past data. The process of developing an appropriate time series model for a variable of interest is referred to as time series analysis (Adhikari & Agrawal, 2013) . Time series analysis enjoys the benefit of being able to identify the patterns of a variable across time, and based on it, output data, which can be used as a basis for decision-making (Velicer & Fava, 2003). Due of this, time series analysis is widely applied in several fields, such as economics, medicine and energy (Zhang & Kline, 2007).

One of the applications of time series analysis in the field of energy, is forecasting renewable energy sources' production, as this production is an instantaneous response to weather phenomena, and installed capacity, throughout time (Sweeney et al., 2020). Accurate forecasts enable a better integration of RES within the power grids, since they minimise the use of electricity reserves, and better balance the difference between electricity demand and supply (Teixeira et al., 2024). Even though we have been experiencing more accurate forecasts in recent years, mainly due to improved weather forecasting methods and advances in data analysis' methods, there are still limitations on the generation of forecasts (Matrenin et al., 2022), (Teixeira et al., 2024).On several occasions, these forecasts provide low reliability, mostly due to the weather parameters' unpredictability (González Ordiano et al., 2018).

This research intended to mitigate the unreliability associated to RES's production, by computing accurate renewable energy production forecasts. In this research, a dataset from Madeira Island was used to produce accurate forecasts. Although this research focused on

Madeira Island as a case study, it intends to contribute to an overall improvement in the quality of renewable energy production's forecasts.

However, at this point, it is not clear which is the best time series analysis approach to forecast this production balance, given the input data. This unclearness leads to a research gap, when it comes to having readily available information and the decision of how to use it in the most valuable way.

This gap leads to the research question (RQ): "What are the best approaches to accurately forecast renewable energy production, in Madeira?".

## 1.2 OBJECTIVES

To answer the research question, and to find out the best time series approaches, out of the existing ones, it is necessary to assess thoroughly a RES' dataset. Given the need for local data, this project used data supplied by Empresa de Eletricidade da Madeira (EEM).

EEM is the main utilities company in Madeira, responsible for producing, transporting, distributing and commercializing energy within the region. EEM's renewable energy production comes from different sources, namely hydropower, wind energy, solar energy and biomass energy (Empresa de Eletricidade Da Madeira, 2024). Strikingly like the rest of the world, EEM is looking to reduce its reliance on fossil fuels and foster renewable energy production, while ensuring a steady energy supply like Madeira's inhabitants. To achieve this, high quality forecasts of the renewable energy produced, are necessary.

This thesis embodied a Work Project, using EEM's energy related data. This data consisted of the production of RES (hydropower, biomass, solar and wind), per every 15 minutes, from 1st of January 2021, until 31st of August 2024. In addition, this dataset was enriched with the evolution of installed capacity, per energy source, throughout the considered years.

The main research objective of this work project was to propose an efficient method to forecast renewable energy sources' production in Madeira. Ultimately, the higher quality forecasts intend to enable a better integration of renewable energy within power grids, not just in Madeira, but in all locations that aim to use the findings from this research, to produce renewable energy forecasts. To structure the quantitative analysis and achieve the research objective, the CRISP-DM framework was utilized.

To achieve the main goal, the following intermediate objectives were defined:

- Study the most relevant concepts in the field of renewable energy.
- Make a systematic literature review on forecasting and time series analysis in the field of energy.

- Use the acquired knowledge to select the most suitable time series algorithms, i.e. the ones that will produce the most accurate energy production forecasts.
- Assess the performance of the different algorithms, with Madeira's data, given the existing input data.
- Make useful and sustained recommendations, which would enable EEM to increase its reliance on renewable energy, by using the developed forecasts as a basis for decision making.

These intermediate objectives were deemed to be the ones necessary to reach the best forecasting method, which would foster the utilization of renewable energy, to the detriment of fossil fuels.

## 1.3 Study Outcomes and contributions

The main goal of this project was to contribute to a more sustainable world, one with efficient energy management, and to be reliant on RES. By focusing on achieving high quality renewable energy forecasts, this research intended to gather new and unknown, but also valuable insights (*European Green Deal*, 2019).

More concretely, this project aimed to comply with the United Nations' 7$^{th}$ sustainable development goal, by enabling enhanced energy performance, thus ensuring affordable and clean energy for all (Environment, 2017).

Only with accurate forecasts can we foster and optimize this clean energy production, to the detriment of fossil fuels. By very thoroughly studying the energy field and comparing different models' outputs with the given input variables, this thesis aimed to generate evidence-based guidelines, not only to Madeira, but also to researchers on which path to follow, when trying to accurately forecast renewable energy sources' production, to optimize the usage of these clean and greener energy.

This research deepens the knowledge of forecasting the production of renewable energy, per source, by demonstrating the promising potential of Data Science's applications and time series' frameworks in the domain of energy. It was intended that, by working on this real case scenario, with Madeira's data, all key industry players got motivated to implement this solution and framework. Ultimately, this could be an opportunity to contribute to a greener world.

This document is organised as follows: section 2 discusses the literature review and concepts of this study; section 3 discusses the methodology used in this study; section 4 presents the results and discussion of the experiments, and finally, section 5 concludes the document.

# 2. LITERATURE REVIEW

The increasing volume of data generated annually, combined with the sophistication of new Artificial Intelligence (AI) applications and the growing emphasis on transitioning to RES production, has highlighted the need to understand the most popular time series models and the latest developments in the field.

Combined with these phenomena, to be able to answer the previously mentioned research question, “What are the best approaches to accurately forecast renewable energy production, in Madeira?", and contribute to a more sustainable world, a literature review needed to be conducted.

A literature review is a methodological study that employs database searches to retrieve research findings, aiming to gain the most comprehensive insights possible on a specific topic or theme (Knopf, 2006).

With all this in mind, this literature review aimed to establish a strong foundation in the field of energy and time series’ framework, focusing on what are the most relevant techniques to accurately forecast renewable energy production per source. Conducting this literature review was a crucial step towards achieving the study’s final objective, which was to answer the research question.

Additionally, the literature review served as the first step in the research process, as it enabled the fulfilment of the first two intermediate objectives, which were “Study the most relevant concepts in the field of renewable energy”, and “ Make a systematic literature review on forecasting, and time series analysis in the field of energy”.

## 2.1 AI APPLICATIONS IN FORECASTING RENEWABLE ENERGY PRODUCTION

Artificial Intelligence (AI) is a branch of computer science dedicated to developing systems that can carry out tasks usually associated with human intelligence, including understanding visual information and recognizing speech. AI can be divided into various subfields (Campos et al., 2024).

Machine learning is a branch of AI that involves creating algorithms capable of learning from data and making decisions based on patterns found in that data. Rather than relying on explicitly programmed instructions, the system learns to perform tasks by being trained with examples. This training usually involves structured and labeled data, where both inputs and expected outcomes are provided. After training, the algorithm can generate predictions based on new input data (Campos et al., 2024).

Deep learning is a specialized area within AI that uses mathematical models built on artificial neural networks to mimic how the human brain works. These networks usually include an input layer, several hidden layers, and an output layer. Neurons in each layer are connected

through links, each carrying a weight and receiving a value from a neuron in the preceding layer. This flow continues through the layers until the output layer produces the result (Campos et al., 2024). Figure 2.1 presents an example of such network:

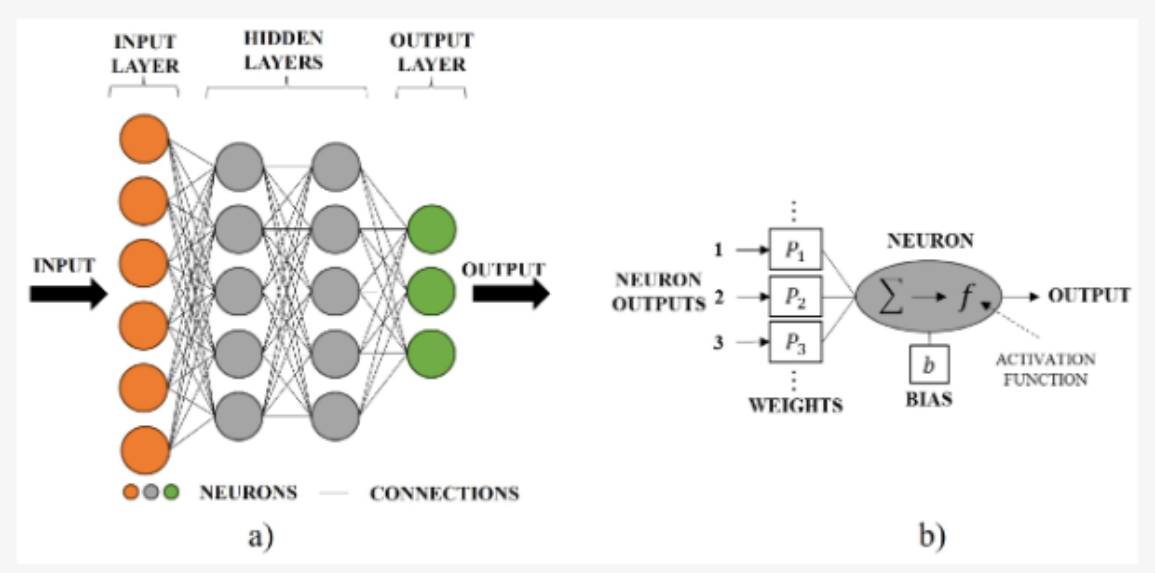


Figure 2.1- Neural Network (Campos et al., 2024)

Machine learning and deep learning both make use of autoregressive models, which operate on the principle that future values of a variable can be predicted using its past values. These models are commonly applied in time series analysis, as they take into account short-term fluctuations, long-term trends, and seasonal patterns (Szostek et al., 2024).

All these approaches, traditional machine learning, deep learning, and autoregressive models are effective frameworks for forecasting renewable energy production using historical data. By learning from past patterns, these methods enable accurate predictions of energy production, aiding in the precise assessment of the potential and efficiency of renewable technologies (Szostek et al., 2024).

Each of these approaches has its own advantages and limitations, yet all are widely applied in real-world scenarios. This chapter delved deeper into these methods, exploring their characteristics and practical applications.

## 2.2 Literature Review Methodology

In this research, a systematic literature review (SLR) was performed, following the PRISMA Guidelines, which are a set of standardized recommendations to help researchers establishing the state of knowledge on specific topics (PRISMA, 2020).

To achieve the primary objective of this research—proposing an efficient method for forecasting renewable energy sources' production in Madeira—it was essential to conduct a

SLR. Following the PRISMA guidelines, this review ensures the retrieval of the most relevant articles on the subject, providing a solid foundation for the study.

This SLR was expected to provide insights into the following topics, which were essential to fulfil the research objective:

- The current state of the art in forecasting RES production, including data preparation techniques, possible approaches, and what machine learning, deep learning, and autoregressive models are being used.

- Additional variables, beyond energy production, that enhance model interpretability.

To ensure that this SLR was built on the most relevant articles addressing the above topics, a targeted query using specific keywords related to renewable energy and time series was employed to refine the search. The expected outcome was to identify the articles that provide the most valuable insights into the outlined topics.

The SLR execution was combined with data visualization techniques, to quantify the most relevant keywords and concepts. These analyses were conducted with the help of VOSviewer, which is a bibliometric visualization software, very useful for constructing bibliometric networks (*VOSviewer*, 2024).

Given the complementation of data visualization techniques, this SLR underwent 3 phases: (1)- Data Selection, (2)- Results and Analysis and (3)- Discussion.

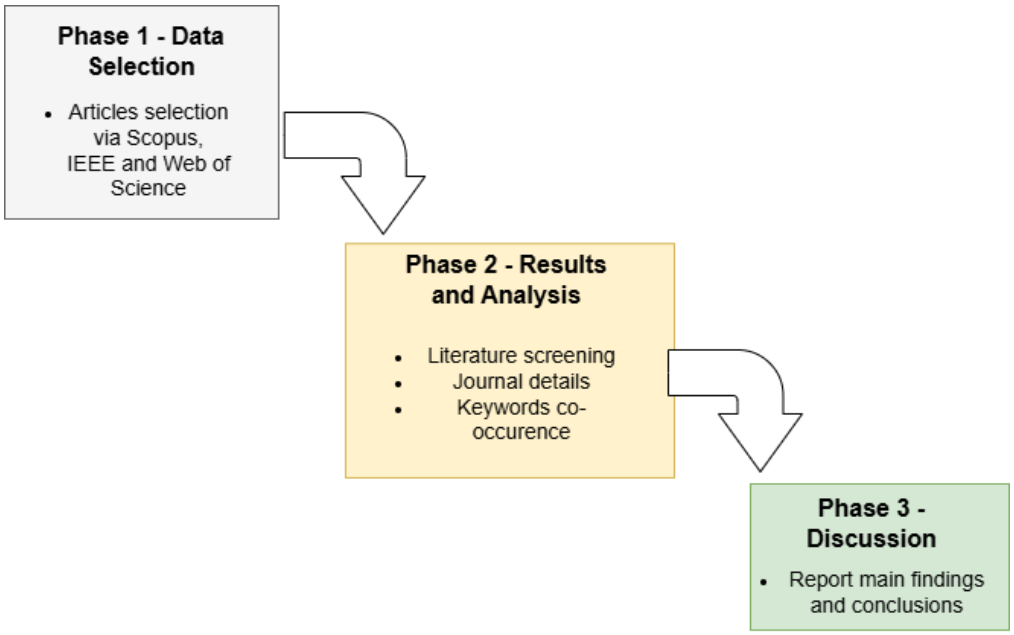


Figure 2.2- PRISMA Methodology

### 2.2.1. Data Selection

The goal of a data selection process is to find the best articles for a literature review. For this research, the selected databases were Web of Science, Scopus and IEEE, to retrieve the articles that sustained this literature review.

The databases output articles that meet a specified criteria and allow them to be exported the in several different formats. In this case, the results were exported in the format RIS, because this is a format that incorporates metadata from the articles in question. In the current research the criteria were to find out the relevant articles that target scientific works related to the usage of AI and ML techniques in renewable energy production forecasting. Therefore, the following Boolean search string was built:

("*Renewable Energy Production*" OR "*Energy Generation*" OR "*Energy Production*" OR "*Renewable Sources*" OR "*Energy Analysis*") AND (“*Machine Learning*" OR "*Forecasting*" OR "*Time Series Analysis*”)

The first step consisted of exporting all articles in these three databases that met the enunciated Boolean criteria.

Additionally, the RIS format is compatible with Mendeley Reference Manager software, which was used in this literature review. In this data selection phase, the following tasks needed to be performed:

**1)** Export the articles that meet the specified criteria across the 3 databases to the format RIS.
**2)** Use Mendeley Reference Manager for further processing, namely deleting the duplicated articles.
**3)** Remove articles that are not relevant to the study, based on title and abstract screening.
**4)** Read the full articles and remove from the study the ones that are not relevant.

Apart from addressing the query, to refine the search and get the most useful articles, Table 2.1 states inclusion criteria that were established for the three used databases:

Table 2.1- Established Inclusion Criteria

| Inclusion Criteria | Inserted Criteria |
|---|---|
| Year range | 2019-2024 |
| Language | Only English |
| Type of access | Only Open Access |
| Document type | Only Articles |
| Region of case study | EU only |

| Type of case study | Forecasting RES' production |
|---|---|

Even though all three databases offered the option to filter results by subject area, this research did not utilize that feature, given the high number of subject areas available.

To answer the research question, this research chose to use only case studies within the EU, as Madeira belongs to this organization, thus ensuring compliance with European legislation. Additionally, the query retrieved many articles primarily centred on topics such as fault detection or the economic aspects of energy applications. While these areas are valuable, they fell outside the central scope of this literature review, which aimed to find forecasting applications within renewable energy. Hence, only forecasting RES production case studies were considered.

Even though hydropower is the largest source of renewable energy in the electricity sector, no EU forecasting case studies of hydropower were found with this query (*United Nations-Climate Action*, 2024) . The search yielded case studies exclusively focused on wind and solar energy forecasting.

Since this project worked with data related to the production of nearly every renewable energy source (except for geothermal) , two additional articles outside of the query were added to this research, as they were not found during the PRISMA process. These were related to forecasting hydropower case studies.

A study in 2022 analysed scientific articles, which were available in the SCOPUS database from 2020 to 2022, associated to the creation of forecasting models, per renewable energy source. Most publications were related to wind energy predictions (56.11% of all analysed articles), followed by solar energy (40.46%), whereas water energy had the lowest number of articles (4.96%), according to the study. The author also stated that this data indicated a research gap, and that further research was necessary on this source (Krechowicz et al., 2022). This project was able to sustain this gap on the literature, and the fact that solar and wind energy's forecasting applications are the most common ones. Additionally, no biomass energy application was found.

Table 2.2- Initial Results

| **Database** | **Number of retrieved articles** |
|---|---|
| Scopus | 1243 |
| Web of Science | 1040 |
| IEEE | 200 |
| **Total via databases** | **2483** |

| | |
|---|---|
| Articles added via other sources | 2 |
| **Total** | **2485** |

It is important to note that the database results were obtained on November 23, 2024. Using the same combination of expression and inclusion criteria on a different date may yield different numbers, as new articles are published daily.

As previously noted, the articles were exported in RIS format for further processing inside Mendeley, with the most comprehensive metadata available for each article extracted from all three repositories.

After importing the RIS files from each of the three databases into Mendeley, this tool identified 883 sets of duplicated articles, based on having the same DOI (Digital Object Identifier), which is a unique combination of numbers, letters and symbols to uniquely identify an article (*University of Illinois Chicago*, 2024).

Some sets of the 883 articles had the same article repeated more than twice, meaning that in total, 942 articles were removed. After this, there were 1543 unique articles in the database, to perform further screening upon.

Furthermore, title screening led to the removal of 1374 articles, leaving 169 in the database. The removed articles did not meet the criteria previously outlined, namely for being either case studies outside the EU, fault detection studies, studies predicting different target variables (such as energy demand or price), or for being studies on electric cars or buildings, among others.

Moreover, abstract screening was conducted, resulting in the removal of 81 articles for the same reasons presented in the previous paragraph, along with the exclusion of entries without abstracts (possibly due to export errors). After this, 88 articles remained in the database.

Lastly, all 88 remaining articles were subject to full text screening, from which 72 were removed, based on similar reasons as before. Figure 2.2 sums up the articles' selection process, as well as provides more detailed insights into the reasons why 72 articles were removed:

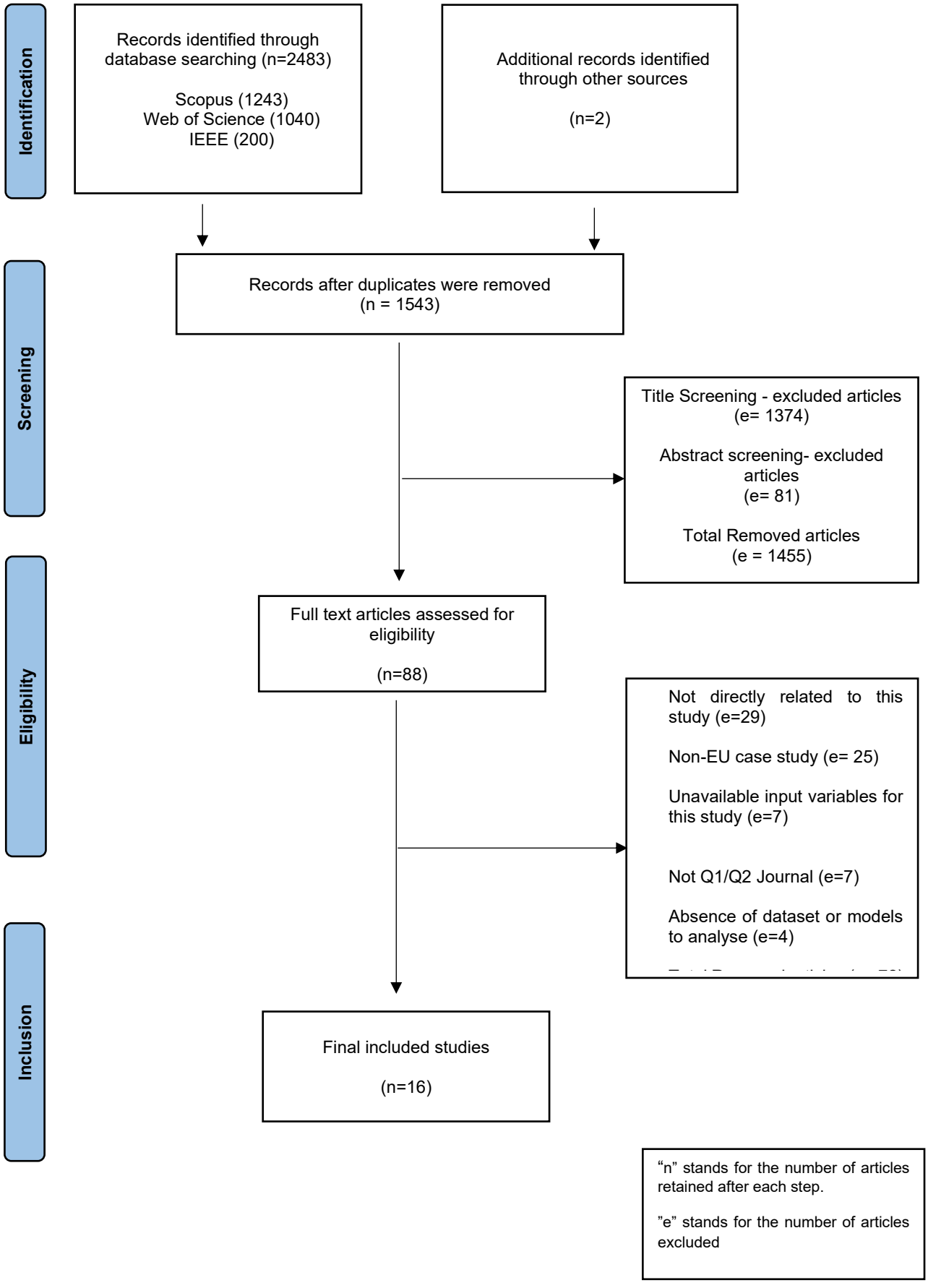


Figure 2.3- PRISMA flowchart

### 2.2.2. Results and analysis

After the filtering process, this study examined 16 articles published across 9 different journals, majority of which were published in Energies Journal (#8), one of the most respectful Journals in the field. The remaining, each with one article were published in Sustainability Switzerland, Forecasting, International Journal of Photoenergy, Energy Economics, Computers, Materials & Continua, Energy, Applied Energy and Sensors.

All journals were at least Quartile 1 or Quartile 2 ranked for all its subject areas. There were 5 journals which were Quartile 1 or 2 ranked, depending on the considered category. Three journals were Quartile 1 ranked across all its subject areas, and only one journal was Quartile 2 ranked for all its categories.

The Journal Publishers are from three different countries, based mostly in Switzerland (4), followed by Netherlands (3) and USA (2). The leading publisher is Multidisciplinary Digital Publishing Institute (MDPI) (4), followed by Elsevier (3). John Wiley & Sons and Tech Science Press, each with one, represented the remaining publishers of the final 16 articles.

The most common journal subject areas were Renewable Energy, Sustainability and the Environment (3), followed by Energy Engineering and Power Technology (2) and Computer Science Applications (2). Lastly, both General Energy and Computer Science: Information Systems are represented once.

This information is summarised in table 2.3:

Table 2.3- Journal details and their Scimago Ranks

| Journals | Scimago Rank | Number of Articles | Publisher | Country | Journal Subject Area |
|---|---|---|---|---|---|
| Energies | Q1/Q2 | 8 | MDPI | Switzerland | Energy Engineering and Power Technology |
| Energy Economics | Q1 | 1 | Elsevier | Netherlands | General Energy |
| Energy | Q1 | 1 | Elsevier | Netherlands | Renewable Energy, Sustainability and the Environment |

| | | | | | |
|---|---|---|---|---|---|
| Applied Energy | Q1 | 1 | Elsevier | Netherlands | Renewable Energy, Sustainability and the Environment |
| Computers, Materials & Continua | Q1/Q2 | 1 | Tech Science Press | USA | Computer Science Applications |
| International Journal of Photoenergy | Q2 | 1 | John Wiley & Sons | USA | Renewable Energy, Sustainability and the Environment |
| Forecasting | Q1/Q2 | 1 | MDPI | Switzerland | Computer Science Applications |
| Sensors | Q1/Q2 | 1 | MDPI | Switzerland | Computer Science: Information Systems |
| Sustainability (Switzerland) | Q1/Q2 | 1 | MDPI | Switzerland | Energy Engineering and Power Technology |

The analysis of the most-cited publications helped this research to detect the most important research articles in the literature. The most cited information was retrieved via Scopus and Web of Science databases. This study identified publications that were cited between 178 and 0 times.

The top five found publications were from the following authors: Khan, Walker and Zeiler, with 178 citations (the most cited one), followed by Gómez et al., with 48 citations, followed by Fara et al., with 40 citations, followed by Bochenek et al., with 34 citations and lastly, Almonacid-Olleros et al., with 23 citations.

Concerning the country of case study, more studies were performed with data from Poland (4), followed by Spain (3) and Italy (2). Bulgaria, Estonia, Germany, Netherlands, Portugal, Romania and Slovenia had each a case study in this sample.

Table 2.4- Number of citations and the country of origin for each article's data.

| Article | Number of Citations | Country of Case Study |
|---|---|---|
| 1. Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach | 178 | Netherlands |
| 2. Photovoltaic Power Prediction Using Artificial Neural Networks and Numerical Weather Data | 48 | Italy |
| 3. Forecasting of Energy Production for Photovoltaic Systems Based on ARIMA and ANN Advanced Models | 40 | Romania |
| 4. Day-Ahead Wind Power Forecasting in Poland Based on Numerical Weather Prediction | 34 | Poland |
| 5. A New Architecture Based on IoT and Machine Learning | 23 | Spain |
| 6. Short-Term Wind Energy Forecasting Using Deep Learning-Based Predictive Analytics | 15 | Estonia |
| 7. Forecasting solar energy production in Spain: A comparison of univariate and multivariate models at the national level | 13 | Spain |

| | | |
|---|---|---|
| 8. Prediction of Electricity Generation Using Onshore Wind and Solar Energy in Germany | 5 | Germany |
| 9. A Hybrid Method for the Run-Of-The-River Hydroelectric Power Plant Energy Forecast: HYPE Hydrological | 5 | Slovenia |
| 10. High Penetration of Renewable Energy Sources and Power Market Formation for Countries in Energy Transition: Assessment via Price Analysis and Energy Forecasting | 4 | Bulgaria |
| 11. Investigation of Load, Solar and Wind Generation as Target Variables in LSTM Time Series Forecasting, Using Exogenous Weather Variables | 4 | Spain |
| 12. Application of Artificial Intelligence Algorithms in Multilayer Perceptron and Elman Networks to Predict Photovoltaic Power Plant Generation | 3 | Poland |
| 13. Short-Term Forecast of Photovoltaic Solar Energy Production Using LSTM | 3 | Portugal |
| 14. Analysis of the Effectiveness of ARIMA, SARIMA, and SVR Models in Time Series Forecasting: A Case Study of Wind Farm Energy Production | 2 | Poland |
| 15. Forecasting photovoltaic production with neural | | |

| | | | |
|---|---|---|---|
| | networks and weather features | 0 | Italy |
| 16. | Forecasting Electricity Production in a Small Hydropower Plant (SHP) Using Artificial Intelligence (AI) | 0 | Poland |

#### 2.2.2.1. Keyword co-occurrence.

The co- occurrence of keywords was performed using VOSviewer. The analysis was done using the full counting method, with a minimum threshold of two co-occurrences. Of the 84 terms, 21 met the threshold, which are listed in table 2.5:

Table 2.5- Keywords co-occurrence and the total link strengths

| Keyword | Occurrences | Total link strength |
|---|---|---|
| Machine Learning | 5 | 22 |
| Prediction | 4 | 20 |
| Deep Learning | 3 | 14 |
| Output | 2 | 13 |
| Performance | 2 | 13 |
| Regression | 2 | 12 |
| Solar Energy Forecasting | 2 | 12 |
| Neural Networks | 2 | 9 |
| Solar Energy | 2 | 9 |
| Time Series Forecasting | 2 | 9 |
| Artificial Neural Networks | 2 | 8 |
| Neural-networks | 2 | 8 |
| Solar | 2 | 8 |
| Seasonal Decomposition | 2 | 4 |

| | | |
|---|---|---|
| ANN | 2 | 3 |
| Energy Forecast | 2 | 3 |
| Generation | 2 | 3 |
| LSTM | 2 | 3 |
| Neural- Network | 2 | 3 |
| Photovoltaic Systems | 2 | 3 |
| Renewable Energy | 2 | 1 |

The top five most frequently encountered keywords were “Machine Learning” (5 occurrences, 22 total link strength), “Prediction” (4 occurrences, 20 total link strength), “Deep Learning” (3 occurrences, 14 total link strength), “Output” (2 occurrences, 13 total link strength) and “Performance” (2 occurrences, 13 total link strength).

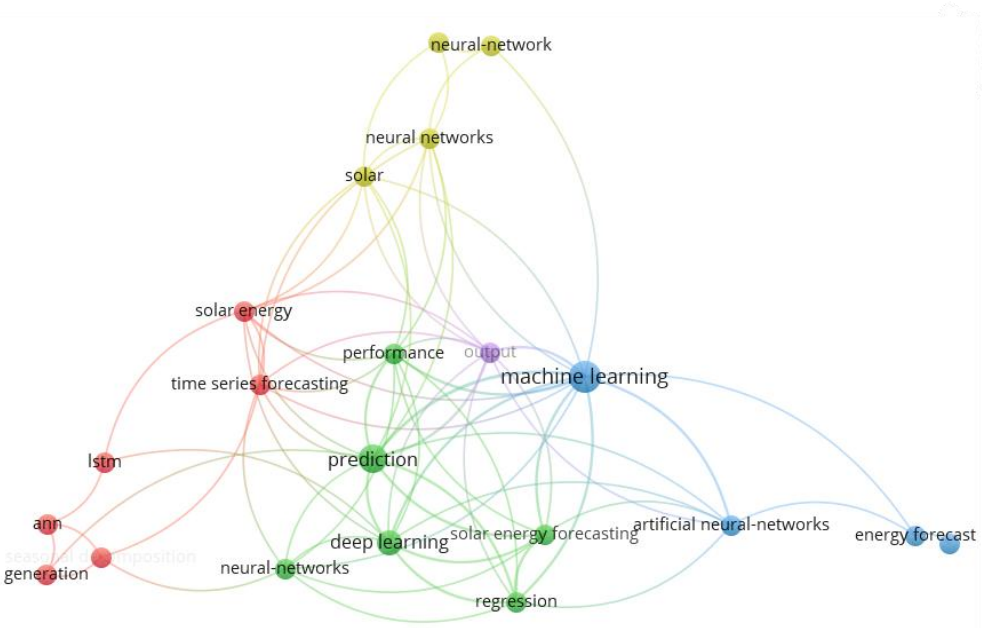


Figure 2.4- Keywords Co-occurrence Network

As depicted in Figure 2.4, the keywords co-occurrence analysis showed five clusters, with 21 keywords, 74 links and 90 total link strength. The five clusters were distinguished by colours, each featuring the following key nodes, as per table 2.6:

Table 2.6- Principal Keywords of each cluster and respective colour

| Principal Keywords | Colour |
|---|---|
| Solar Energy | Red |
| Prediction | Green |
| Machine Learning | Blue |
| Solar | Yellow |
| Output | Purple |

The keyword co-occurrence showed that clusters did not exhibit distinct separation between them. The cluster that displayed more distance was the yellow one (“Solar”). The purple cluster (“output”) was composed by only this principal keyword. The three biggest nodes, red (“Solar Energy”), green (“Prediction”) and blue (“Machine Learning”) had multiple links between them.

The keyword co-occurrence network, by year overlay visualization, revealed that “Artificial Neural Networks” were prominently discussed in 2022, as they were strongly linked to the 2022 colour. Additionally, “Deep Learning” and “Machine Learning” methodologies gained increasing attention starting in 2023, as reflected by their brighter tones. The most recent discussed keywords were “solar energy”, “time series forecasting” and “lstm”.

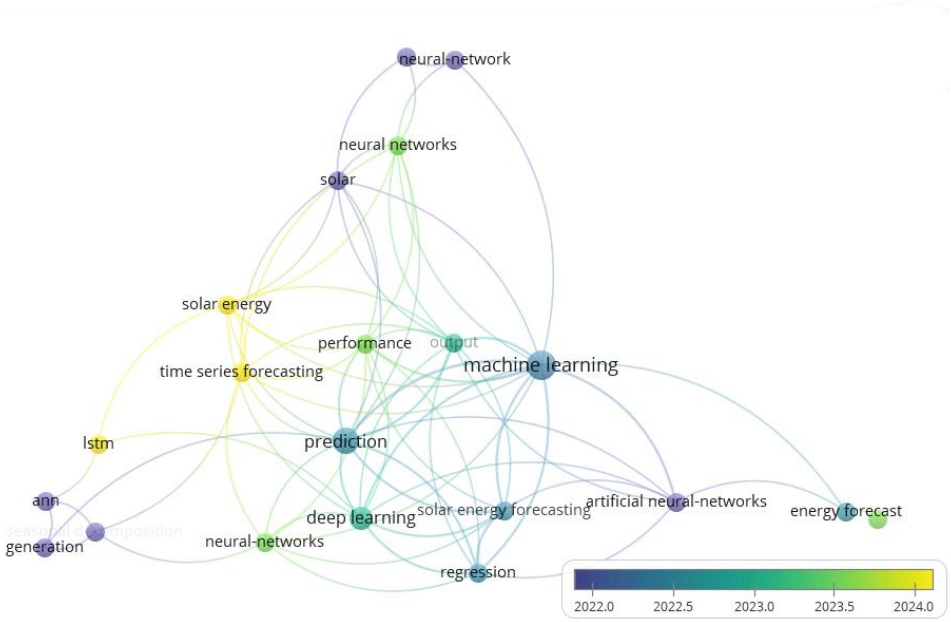


Figure 2.5- Keywords Co-occurrence by year

Advancing to the next subject, Table 2.7 presented an analysis of the occurrence and total link strength between AI-related keywords and specific renewable energy-related terms, based on the data from Table 2.5, which consisted of only keywords which met the threshold of at least two occurrences. The "Occurrences" column, in this table, indicated how many times each

renewable energy keyword from Table 2.5 appeared in connection with AI-related keywords within the same dataset. Specifically, the AI-related keywords considered in this analysis included: “Machine Learning”, “Prediction”, “Deep Learning”, “Output”, “Performance”, “Regression”, “Neural Networks”, “Time Series Forecasting”, “Artificial Neural Networks”, “Neural-Networks”, ”ANN”, “Generation”, “LSTM”, and “Neural-Network”. Meanwhile, the "Total Link Strength" column quantified the strength of these relationships, representing the summed weighted connections between AI-related and renewable energy-related keywords as calculated in the VOSviewer network.

Table 2.7- AI related Keywords co-occurrence with renewable energy, plus its total link strengths.

| Keyword | Occurrences | Total link strength |
|---|---|---|
| Solar Energy Forecasting | 12 | 206 |
| Solar Energy | 8 | 175 |
| Solar | 6 | 134 |
| Seasonal Decomposition | 4 | 39 |
| Photovoltaic Systems | 2 | 36 |
| Energy Forecast | 2 | 18 |

The results indicated that "Solar Energy Forecasting" had the highest co-occurrence (12 occurrences) and total link strength (206), suggesting that it was the most frequently and strongly linked renewable energy keyword in AI-related research. Similarly, "Solar Energy" (8 occurrences, 175 link strength) and "Solar" (6 occurrences, 134 link strength) were also highly relevant in AI applications, highlighting the strong focus of AI research on solar energy systems. On the other hand, "Photovoltaic Systems" (2 occurrences, 36 link strength) and "Energy Forecast" (2 occurrences, 18 link strength) exhibited much lower values, indicating that while they were linked to AI-related keywords, their presence and influence in AI-driven research was comparatively weaker.

### 2.2.3. Discussion

The aim of the SLR was to highlight and detect the literature on ML, DL and Time Series’ frameworks which were applied in forecasting renewable energy sources’ production.

All the identified literature was classified according to the type of forecasted renewable energy source, which resulted in four main groups: Solar energy applications (1), Wind energy applications (2), solar and wind energy applications (3) and Hydropower applications (4).

Regardless of the type of source, common approaches shared across all groups were presented first, followed by a detailed description of each of the four main groups.

#### 2.2.3.1. Common approaches

The PRISMA methodology and bibliometric analysis allowed for the identification of the most common approaches for integrating AI and ML techniques in renewable energy forecasting. By examining keyword co-occurrences and link strengths, the bibliometric analysis helped this research to highlight the most interconnected concepts within the field. These included various ML architectures such as DL, time series forecasting, and artificial neural networks, which frequently co-occured with renewable energy terms, particularly in the context of solar energy forecasting.

Data concerning RES production was collected on different granularities, with 15 minutes and 1 hour being the most common time granularities. In some cases, this data was later aggregated into a less granular time, such as hour or day. Additionally, whenever possible, authors tried to collect weather conditions at the time of production. To that extent, on those case studies, the authors tried collecting renewable energy production data over a long period of time, to mitigate weather variability (Cabello-López et al., 2023). Data was retrieved via either national repositories or weather APIs, such as Visual Crossing Weather (Dralus et al., 2023).

Missing values existed on several articles. The most common approach to deal with these was using linear interpolation, which is an imputation method that averages the closest points, since this involved time series data (Khan et al., 2022). Additionally, outliers are values significantly different from the others, which can disrupt the model creation process and reduce the quality of the forecast (Maciejewski et al., 2024). These were handled via interquartile range or isolation forest.

Regarding data normalization, the Min-Max Scaler was the most used method across the articles reviewed. The train-test split was consistently performed with at least 70% of the data allocated for training to minimize overfitting. Additionally, cross-validation, a widely recommended technique to prevent overfitting and reduce biased results, was also employed. The most used cross-validation method was K-fold. However, one study utilized Blocked-Cross-validation instead, as it preserves time dependency in the data (Cabello-López et al., 2023).

In most studies, authors compared the forecasting accuracy of DL models, ML models and autoregressive models. Although DL models have higher computational expense, these require less human intervention (Almonacid-Olleros et al., 2020). Furthermore, DL models tend to understand better non-linear relationships. This indicates that DL models might be better equipped to forecasting renewable energy production, given that weather unpredictability is a constraint to forecasting applications (Cabello-López et al., 2023).

Nonetheless, DL models are very sensitive to hyperparameterization, so several settings of the same model had to be experimented. The choice of parameters was given via Grid Search, to reduce bias, or via mentioned parameters within the Literature.

Traditional ML models have had good results with time series data and better learning times than DL ones. However, these applications have disadvantages, such as the need for human intervention and inability of recognizing non-linear complex patterns (Cabello-López et al., 2023).

Autoregressive models also need human intervention, as these require for the data to be stationary, and not all autoregressive models handle seasonality. Seasonal Trend Decomposition was applied throughout the studies to assess these assumptions. Additionally, to select the most appropriate number of lag variables, ACF and PACF plots need to be used (Shabbir et al., 2022). Despite the need for human intervention, autoregressive models are appreciated, due to their simplicity and nature of dependencies (Fara et al., 2021)

The most used error metrics to assess the models' quality were Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), RMSE (Root Mean Squared Error), R2 (R-Squared) and nRMSE (normalised Root Mean Squared Error).

When authors tested models with and without weather data, they found that models consistently performed worse without it. While hyperparameterization often resulted in lower error metrics, the most significant reductions in errors were achieved through feature-engineered variables and the inclusion of weather data. Additionally, the literature highlighted the importance of incorporating weather forecasts into renewable energy predictions. Studies that successfully integrated future weather forecasts with renewable energy production data were able to generate more reliable predictions.

Moreover, the most experimented forecasts were short-term, either 1 hour or 1 day ahead, but other horizons were experimented. The results showed that the errors increase with the forecasting horizon. Notably, a study stated that day ahead is the most experimented horizon in the literature, for RES. This is likely because many operators require forecasts having this horizon as part of their day-ahead operations, whether due to market participation or operational planning, reflecting the importance of reliable forecasts minding this horizon (Cabello-López et al., 2023).

#### 2.2.3.2. Solar energy applications

The final sample included eight articles focused on solar energy forecasting, making it the most well-represented category. This was supported by the bibliometric analysis, which revealed that solar energy-related keywords exhibited the strongest and most recurrent links with AI-related terms, highlighting the prominence of AI applications in source, despite being influenced by location and weather conditions (Almonacid-Olleros et al., 2020).

The most used explanatory variables for solar energy were installed capacity, number of photovoltaic panels, solar irradiance, sunshine duration, ambient temperature, daylight status, humidity, rainfall, precipitation time, wind speed, atmospheric pressure and cloudiness. Most studies employed pearson or spearman correlation to evaluate the impact of weather variables on solar energy production. Variables such as ambient temperature and solar irradiance showed a positive correlation with solar energy production. In contrast, factors like humidity, precipitation, rainfall, wind speed, and cloudiness were negatively correlated with solar energy production.

Missing values were found on several solar energy datasets. These missing values were deleted, whenever they were found at night, because there is no solar energy production at night, or whenever they were related to energy production problems (Goutte et al., 2024).

The ML models used in solar energy applications were: Multilayer Perceptron (MLP), which is an Artificial Neural Network (ANN), Random Forest, Linear Regression, K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Support Vector Regression (SVR) and Extreme Gradient Boosting (XGBoost).

On the other hand, the deep learning models used were Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN), whereas the Autoregressive ones were ARIMA and SARIMA.

Ensemble Learning is a ML technique, where multiple models are trained, and their output is combined. A study found that LSTM and ANN were the most used algorithms in this type of application and achieved promising results by using an Ensemble Learning model, with LSTM and ANN as baselines and XGBoost as the meta learner (W. Khan et al., 2022).

Since these case studies use time series data, the literature has recognized the importance of creating time variables to improve a model's performance, such as "Hour of the day", "Month" or "Season". An article has talked about the importance of applying sine and cosine transformations to the time created variable, to reveal cyclical patterns throughout the day and year (Goutte et al., 2024).

Numerous combinations of models, hyperparameterization techniques, and time horizons were explored in these studies. A common approach involved dividing the data by season and generating forecasts separately for each. Studies employing this method observed that errors were lower during the summer and spring months, while higher errors were typically reported in winter and fall.

To improve model interpretability, SHAP was used to assess each feature's impact on the forecast. SHAP values show how much each feature contributes to a prediction by breaking it down into the sum of individual feature effects (W. Khan et al., 2022).

The identified limitations included, in some cases, the failure to account for the microclimatic conditions specific to each energy plant, relying instead on broadly available weather data. Additionally, some studies did not incorporate weather forecasts aligned with the energy forecast horizons, which could have enhanced the accuracy of the predictions (Cabello-López et al., 2023; Goutte et al., 2024).

#### 2.2.3.3. Wind energy applications

This literature review yielded three case studies solely focused on forecasting wind energy production. Wind energy is affected by several factors, such as weather, season, time of the day, geography, and mostly due to wind speed. Hence, this instability brings new challenges to electric power grids, such as reliability, flexibility and power quality (Shabbir et al., 2022).

The literature pointed to the importance of incorporating reliable wind speed forecasts and installed capacity to produce accurate wind forecasts (Bochenek et al., 2021).

The autoregressive models used were ARIMA and SARIMA, which produced promising results, despite seasonality being found within wind data (Szostek et al., 2024).

The ML models used were Linear Regression, Support Vector Machine (SVM), Support Vector Regression (SVR), Random Forest, Multilayer Perceptron (MLP), Extreme Gradient Boosting (XGBoost), Autoregressive neural Network (AR-NN), Non-Linear Autoregressive Neural Network (NAR-NN) and Adaptive Neuro- Fuzzy Inference system (ANFIS). The Deep Learning used models were Deep Learning Neural Networks (DNN) and Long Short-Term Memory (LSTM).

Unlike solar energy cases, errors were smaller in winter and higher during summer. In addition, even though wind speed is an important input feature, the forecast errors were high under very high winds, and during high production periods (Bochenek et al., 2021). Tree based models, such as Random Forest performed better under high wind energy output.

Only one study on this sample compared ML, autoregressive models and deep learning models, and found that deep learning models produced the best results, as they understood better the nonlinear behaviour of wind (Shabbir et al., 2022).

Several time horizons were explored in these articles. However, experiments in wind energy applications generally covered longer time horizons compared to those in solar energy studies, with one study even extending to monthly forecasts (Szostek et al., 2024).

Limitations pointed in the literature were the lack of experimentation with other weather variables, such as humidity, wind direction, temperature and atmospheric pressure (Bochenek et al., 2021).

#### 2.2.3.4. Solar and wind energy applications.

Three articles conducted case studies focusing on the two most extensively studied sources in forecasting applications - solar and wind energy. The electricity generated from RES is dependent on the source, which, in the case of solar and wind power, is inherently intermittent (Koeva et al., 2023).

The most used input variables were installed capacity, temperature, atmospheric pressure, precipitation, humidity, cloudiness, wind speed and solar radiation. Correlation analysis was conducted between the two target variables and the input variables. Similarly to the solar energy studies, solar radiation and temperature were positively correlated with solar energy production, while cloudiness showed a negative correlation. For wind energy, wind speed, precipitation, and cloudiness were positively correlated with energy production, whereas atmospheric pressure and temperature exhibited negative correlations.

Wind energy production increased during colder seasons, such as winter, due to higher wind speeds. In contrast, solar energy production peaked during warmer seasons, particularly in summer, when sunlight was most abundant (Koeva et al., 2023).

All the articles included in this group used separated datasets for wind and solar energy, rather than combining both datasets. Moreover, a study stated that wind energy proved to be more challenging to forecast, than solar energy. This was due to weaker seasonal and cyclic components of wind energy (Shering et al., 2024)

Despite the existing weather fluctuations, the addition of weather data helped in producing more accurate forecasts (Shering et al., 2024).

The only autoregressive model used was SARIMAX. In addition, no traditional ML model was used. In contrary, DL models used were Long Short-Term memory (LSTM), and Transformer-based models, such as Informer, Transformer, Autoformer and FEDformer. LSTM and Informer produced the most promising results.

#### 2.2.3.5. Hydropower applications

This group of hydropower applications was composed by the two articles added via additional sources. Hydropower, like other RES, has certain weather randomness. For this reason, there is high probability of forecast deviation due to the nonlinear behaviour (Maciejewski et al., 2024).

These two case studies were run-of-the-river hydroelectricity. However, the same methodologies can be implemented to different hydropowerplants, provided they have the same seasonality and strong dependence on precipitation (Ogliari et al., 2020).

The input variables used in these articles were installed capacity, precipitation, air temperature and wind speed. One article performed day ahead predictions, while another one experimented with 12 hours and 24 hours ahead.

The ML models used were Artificial neural Networks (ANN), such as Multilayer Perceptron (MLP) and Radial Basis Function Networks (RBF), Random Forest and Gradient Boosting Decision Trees (GBDT). The tree-based models were the ones that yielded the lowest error metrics. No DL or Autoregressive models were used in these case studies.

Forecasts had higher error, when hydropower production was higher. Additionally, forecasts were unable to accurately predict productions in periods of short intense precipitation. Hydropower production was higher during winter, and lower during summer (Maciejewski et al., 2024).

## 2.3 SUMMARY

This research conducted a thorough SLR on the application of ML and AI techniques in RES' case studies. A detailed examination of various approaches and methodologies was carried out, focusing on how AI and ML were integrated into forecasting models across different renewable energy sources.

Conducting a SLR was the first step in answering the RQ: "What are the best approaches to accurately forecast renewable energy production in Madeira?". This SLR addressed the first two intermediate objectives, which were "Study the most relevant concepts in the field of renewable energy", and "Make a systematic literature review on forecasting, and time series analysis in the field of Energy".

The analysis was structured into thematic clusters, each representing a specific renewable energy domain, which were solar energy applications, wind energy applications, solar and wind energy applications and hydropower applications. Within each cluster, the key methodologies, model architectures, and evaluation metrics were explored, providing insights into how time series approaches change, depending on the characteristics of each renewable energy source. Solar was revealed to be the most studied RES in the literature, followed by wind. Hydropower, despite being a key RES, emerged as a notable research gap due to the limited number of studies available. Consequently, two additional articles had to be included externally to address this limitation.

Regardless of the specific RES, the literature revealed several common findings. Notably, accurate weather data was crucial, whether it was retrieved via public repositories or weather APIs. Additionally, DL models consistently delivered higher forecasting accuracy, while traditional ML models offered lower computational costs. Additionally, day-ahead forecasts revealed essential for effective operational scheduling.

Table 2.8 provides a structured summary of the models and methods used in each of the 16 considered articles, offering a comparative view of the techniques applied in renewable energy forecasting. This table helped in identifying common practices, variations in methodology, and trends in AI adoption within renewable energy forecasting.

Table 2.7- PRISMA methodology selected articles

| Article | Models and Methods Used |
|---|---|
| **Solar energy applications** | |
| 1. A New Architecture Based on IoT and Machine Learning Paradigms in Photovoltaic Systems to Nowcast Output Energy | LSTM, 3CNN+2LSTM, Linear Regression, KNN, SVM and Random Forest |
| 2. Application of Artificial Intelligence Algorithms in Multilayer Perceptron and Elman Networks to Predict Photovoltaic Power Plant Generation | Multilayer Perceptron |
| 3. Forecasting of Energy Production for Photovoltaic Systems Based on ARIMA and ANN Advanced Models | ARIMA and Artificial Neural Networks |
| 4. Forecasting photovoltaic production with neural networks and weather features | XGBoost, LightGBM, Linear Regression and Artificial Neural Networks |
| 5. Forecasting solar energy production in Spain: A comparison of univariate and multivariate models at the national level | CNN, LSTM, XGboost, Linear Regression and SVR |
| 6. Photovoltaic Power Prediction Using Artificial Neural Networks and Numerical Weather Data | Artificial Neural Networks |
| 7. Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach | Ensemble Learning, using LSTM and ANN as baselines for a meta-learner |

| | |
|---|---|
| 8. Short-Term Forecast of Photovoltaic Solar Energy Production Using LSTM | LSTM |
| **Wind energy applications** | |
| 9. Analysis of the Effectiveness of ARIMA, SARIMA, and SVR Models in Time Series Forecasting: A Case Study of Wind Farm Energy Production | ARIMA and SARIMA |
| 10. Day-Ahead Wind Power Forecasting in Poland Based on Numerical Weather Prediction | Random Forest, XGBoost, ANN and DNN |
| 11. Short-Term Wind Energy Forecasting Using Deep Learning-Based Predictive Analytics | SVM, Linear Regression, LSTM, ARIMA, Tree-based regression, AR, ANFIS and NAR |
| **Solar and wind energy applications** | |
| 12. High Penetration of Renewable Energy Sources and Power Market Formation for Countries in Energy Transition: Assessment via Price Analysis and Energy Forecasting | SARIMAX |
| 13. Investigation of Load, Solar and Wind Generation as Target Variables in LSTM Time Series Forecasting, Using Exogenous Weather Variables | LSTM |
| 14. Prediction of Electricity Generation Using Onshore Wind and Solar Energy in Germany | LSTM and 5 transformer-based models |
| **Hydropower applications** | |
| 15. A Hybrid Method for the Run-Of-The-River Hydroelectric Power Plant | Artificial Neural Networks |

| | |
|---|---|
| Energy Forecast: HYPE Hydrological Model and Neural Network | |
| 16. Forecasting Electricity Production in a Small Hydropower Plant (SHP) Using Artificial Intelligence (AI) | Multilayer Perceptron and Radial Basis Function Network |

# 3. METHODOLOGY

The performed SLR was the first step towards answering the RQ "What are the best approaches to accurately forecast renewable energy production, in Madeira?".

This SLR, along with all the articles examined, aimed to equip this research with the necessary domain knowledge and skills to produce reliable forecasts, ultimately contributing to a reduction in Madeira Island's reliance on fossil fuels. Through this process, the research developed a comprehensive understanding of the current state of the art in time series approaches for forecasting renewable energy production. It explored data preparation techniques, the application of various models, variables to be used, time horizons and how these approaches differ depending on the type of renewable energy source.

With the conclusion of the SLR, the first two intermediate objectives of this research were fulfilled - "Study the most relevant concepts in the field of renewable energy", and "Make a systematic literature review on forecasting, and time series analysis in the field of Energy". The remaining intermediate objectives were:

- Use the acquired knowledge to select the most suitable time series algorithms, i.e. the ones that will produce the most accurate energy production forecasts.
- Assess the performance of the different algorithms, with Madeira's data, given the existing input data.
- Make useful and sustained recommendations, which would enable EEM to increase its reliance on renewable energy, by using the developed forecasts as a basis for decision making.

Before tackling the remaining intermediate objectives, it was essential to first establish a clear Methodology.

## 3.1 OVERVIEW

The upcoming section is called Methodology and provided a detailed outline of the methodology path which was established for this study.

The purpose of a methodology is to provide a structured framework for understanding the research topic, guiding the process from the initial exploration to the final analysis while defining the course of action to be followed.

The methodology of this project was structured into three distinct phases: Exploratory, Analytical, and Conclusive. As outlined in the introduction, the CRISP-DM framework guided the quantitative approach used to generate these research's forecasts. While CRISP-DM constituted only one part of the overall methodology, it played a key role in shaping the

analytical process. A more detailed explanation of this framework is provided in the Analytical phase. Each of these phases was divided into specific steps.

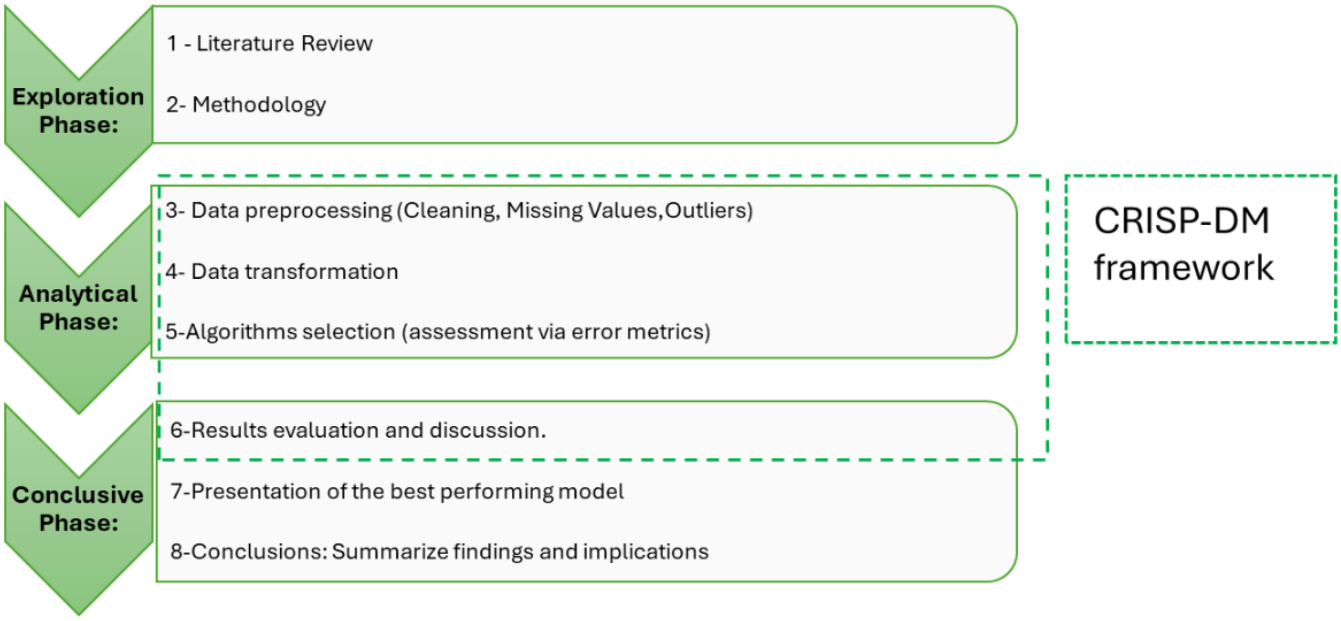


Figure 3.1- Methodological path

## 3.2 Exploration Phase

The first phase started with a SLR, following the PRISMA guidelines, to gain insights into the current state of the art methods for forecasting renewable energy production. This literature review highlighted the critical role of weather data, which could be obtained via weather APIs, such as Visual Crossing Weather API. Additionally, the literature also revealed the superior performance of both DL and traditional ML models, as well as the importance of day-ahead forecasts in operational grid planning.

The last step of this phase was to improve and refine the methodology. The goal of the methodology in this project, was to outline the steps moving forward, so that the insights gained in the SLR, were effectively applied to the renewable energy production dataset, provided by EEM. This enabled the generation of the more accurate forecasts.

## 3.3 Analytical Phase

This phase adopted a quantitative approach, inspired by the Cross Industry Standard Process for Data Mining, also known as CRISP-DM. The CRISP-DM methodology is the most widely used framework for data mining, analytics, and data science projects. It provides a structured, industry-standard approach that guides the execution of data mining projects across various sectors, ensuring consistency, efficiency, and reliability (Wirth & Hipp, 2000).

The CRISP-DM methodology is composed by 6 phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. This methodology facilitates the translation of business problems into data mining tasks, by recommending suitable data treatment and transformation techniques.

Moreover, it provides standardized metrics for assessing the effectiveness of the results. Although originally designed for data mining, this methodology can also be applied to a forecasting case study, as its core principles remain highly relevant to this project.

Table 3.1 outlines each step of the CRISP-DM methodology, detailing the specific tasks involved in each phase as adapted to the context of this research. It is important to note that this study concluded at the evaluation phase, as deployment fell outside the scope of this project.

Table 3.1- Phases of CRISP-DM methodology and tasks to do in each Phase (Wirth & Hipp, 2000)

| **Phase of CRISP-DM** | **Brief description** | **Tasks to do in each Phase of CRISP-DM, in this project** |
|---|---|---|
| 1. Business Understanding | Understand from a business perspective the objectives and requirements of the project | • Identify Research Gap, Question, Objective and intermediate objectives. |
| 2. Data Understanding | Identify, collect and analyze the datasets that can help in achieving the project goals. | • Complement EEM's dataset with Madeira's relevant weather data. |
| 3. Data Preparation | The most time-consuming stage of the project, on which data cleaning is performed (addressing missing values, inconsistencies within the data, and outlier treatment, if applicable). If needed, feature engineering and | • Clean missing values and outliers.<br>• Create time variables from the existing data. |

| | | |
|---|---|---|
| | selection are also performed on this stage. The goal of this stage is to prepare the final dataset. | • Apply sine and cosine transformations to time variables.<br>• Create one dataset per source and perform feature selection. |
| 4. Modeling | This stage determines the modelling technique, as well as the specific parameters, to be applied. | • Apply traditional ML and DL Models and tune them.<br>• Assess results, via error metrics.<br>• Use the best solution to forecast RES' production. |
| 5. Evaluation | This stage aims to look broadly at which model best meets the business and what to do next. | • Discuss the quality and applicability of the obtained results. |
| 6. Deployment | Development of a plan that defines how the final data mining results will be implemented and utilized. | • Share gained insights through the developed code files and the Thesis document. Implementation |

| | |
|---|---|
| | falls outside the scope of this project. |

## 3.4 CONCLUSIVE PHASE

In the final phase, the findings from the Analytical phase were critically analysed and discussed, with a specific emphasis on the capability to generate accurate forecasts.

After addressing the final three intermediate objectives, this research aimed to provide a definite answer to the RQ: ″What are the best approaches to accurately forecast renewable energy production in Madeira*?*″. The goal was for this answer to take the form of a comprehensive framework, incorporating the most effective time series methodologies.

This framework sought to mitigate the unpredictability of renewable energy production caused by weather variability by demonstrating a model capable of generating highly accurate forecasts.

Ultimately, this framework aimed to contribute to reducing Madeira's reliance on fossil fuels while also supporting the United Nations' 7th Sustainable Development Goal (Affordable and Clean Energy). By providing evidence-based guidelines on the best time series approaches, applicable across regions, this study was designed to foster the adoption of renewable energy solutions, paving the way for a more sustainable and environmentally friendly future.

# 4. EMPIRICAL STUDY

The central purpose of this project was to propose an efficient method to forecast the production of renewable energy sources in Madeira Island, hence reducing the dependency on fossil fuels.

To achieve this, this research leveraged relevant findings identified in the SLR and applied them to a real-world dataset from Madeira Island, provided by EEM.

The dataset comprised 15-minute interval records of energy production, including both renewable sources, hydropower, biomass, solar and wind, and non-renewable, fossil fuel and natural gas, which were collected from 1st of January 2021 until 31st August 2024. Additionally, the dataset included information on the evolution of yearly installed capacity for all energy sources, over the same period.

An important insight drawn from the literature is that the day-ahead forecasting horizon is the most extensively studied, highlighting its critical role in helping grid operators schedule generation and maintain system stability (Cabello-López et al., 2023). This, in turn, facilitates the integration of intermittent RES into the grid. Considering this and aligned with the research goal of reducing Madeira Island's reliance on fossil fuels, this study focused on producing day-ahead forecasts, given their utmost relevance and significance in the context of RES systems. Consequently, the data was later aggregated into a daily format, in the data preparation phase.

Moreover, given the well-established importance of weather data in enabling accurate forecasts, this research adopted a similar approach to that presented in one of the articles read in the SLR. Accordingly, observed weather data from two weather stations in Madeira were obtained from the Visual Crossing Weather API, over the same period of the production data (from 1st of January 2021 until 31st August 2024) to complement the existing dataset (Dralus et al., 2023).

In addition, the literature indicated that DL models were particularly well-suited in handling the nonlinear patterns inherent in weather-related data. However, these models typically require large datasets and are more computationally intensive compared to traditional ML approaches. Taking this into account, and guided by the findings in the literature, this research selected a Long Short-Term Memory (LSTM) network, which is a DL model, alongside several ML models: XGBoost, Random Forest, and Multilayer Perceptron (MLP), to compute day ahead RES production forecasts. Traditional time series models, such as SARIMA, were excluded due to their requirement for extensive manual configuration and their relatively poor performance when applied to forecasting RES.

Given that the selected models are considered black-box models, which are models whose internal mechanisms are not easily interpretable, this research employed the SHAP (Shapley Additive exPlanations) framework to enhance their interpretability (W. Khan et al., 2022).

These datasets followed throughout the project all steps of the CRISP-DM methodology (except deployment), in the correct order. It is important to note that this project aimed at merging both datasets, but the steps of CRISP-DM had to be followed first, in the correct order, before merging.

The first step of the CRISP-DM model is called business understanding, which was concluded by establishing clear research objectives and performing a SLR. Chapter 4 started in the second step, data understanding, and went until the last step, called evaluation, which was detailed in the discussion section.

The tool used to carry out all these steps was Python using Jupyter Notebooks, developed and run in Google Colab Pro with a GPU runtime for the modeling stage. Additionally, to enhance clarity for the reader, all RES types in the graphs followed a consistent colour scheme: solar in red, wind in orange, hydropower in blue, and biomass in green.

## 4.1 DATA UNDERSTANDING

As mentioned during the Methodology section, data understanding refers to the CRISP-DM stage, in which relevant datasets are identified, collected and analyzed, to support the project's objective. In this case, the objective was to propose an efficient method for forecasting renewable energy production in Madeira Island. With this definition in mind, the relevant datasets were identified and collected, namely the energy production datasets provided by EEM and observed weather data from Madeira, retrieved via the Visual Crossing Weather API.

The final step at this stage was to explore both datasets, to find possible errors and inconsistencies, and later treat them. It is important to note that some specific and minor data corrections were performed in this stage, for the simple reason that they could have a negative impact on data exploration. This exploration process was detailed in the following sections.

### 4.1.1. Energy Production Dataset

The energy production dataset was provided in an Excel file consisting of five tabs. Four of these tabs corresponded to individual years of data, while the fifth contained information on installed capacity by energy source per year. The objective was to unify them all into a single dataset. However, this project needed to follow the data understanding phase first, to ensure any potential issues were first identified.

These five tabs were initially imported into Python for exploration purposes. Following a preliminary inspection, it was noted that in total (including all years) the dataset contained

128 540 rows, each representing a 15-minute interval. The index was the timestamp, indicating the respective interval. A preliminary verification was done to see if there were any timestamps missing, but the length of the dataset matched the number of 15-minute intervals during this period.

Furthermore, a total of 12 variables were identified across all five tabs as listed below, with 6 of those representing energy production of a source – solar, wind, hydropower, biomass, fossil fuel and natural gas- and the other 6 representing its capacity – solar energy capacity, wind energy capacity, hydropower energy capacity, fossil energy capacity and natural gas energy capacity, with all these capacities in yearly values. For more information about these variables, refer to Table A.1 in Appendix A.

The energy sources' production variables were measured in megawatts-hour (MWh), which indicated the total amount of energy generated by a system over a certain period. In contrast, energy capacity variables came in megawatts (MW), since these represent the maximum rate at which a system can generate energy at a given moment. For example, if a system has a capacity of 10MW, and operates at full capacity for one hour, it will generate 10 MWh of energy (*PK Energy*, 2024).

The next step was to assess the existence of missing values. Missing values are data points that were not recorded or are unavailable for a dataset. These are important to assess, as most models are unable to handle missing values directly (Maciejewski et al., 2024). In case of existence of missing values, these should be either deleted or filled in with an appropriate technique (Emmanuel et al., 2021). The only variable containing missing values was hydropower, with 2 880 missing entries corresponding to the entire month of November 2021. This issue was addressed during the preprocessing stage.

Additionally, some timestamp irregularities were identified, motivated by daylight saving time (DST) adjustments. Specifically, each year contained four duplicated rows at 1:00 AM at the end of October (31/10/2021, 30/10/2022 and 31/10/2023) and four missing timestamps at 1:00 AM at the end of March (28/03/2021, 27/03/2022, 28/03/2023 and 31/03/2024). This pattern corresponded to the DST schedule, where clocks are set forward in March, skipping 1:00 AM, and set back in October, repeating 1:00 AM.

After identifying these inconsistency issues, the next step involved assessing the underlying patterns within the dataset. To facilitate this, data visualization techniques were employed, as they play a crucial role in the exploratory data analysis phase of machine learning projects. Visualizations enable a faster and more intuitive understanding of complex data structures compared to raw numerical outputs, allowing patterns, trends, and anomalies to be identified more efficiently and effectively (Midway, 2020). The following visualizations were chosen as the most relevant for this chapter.

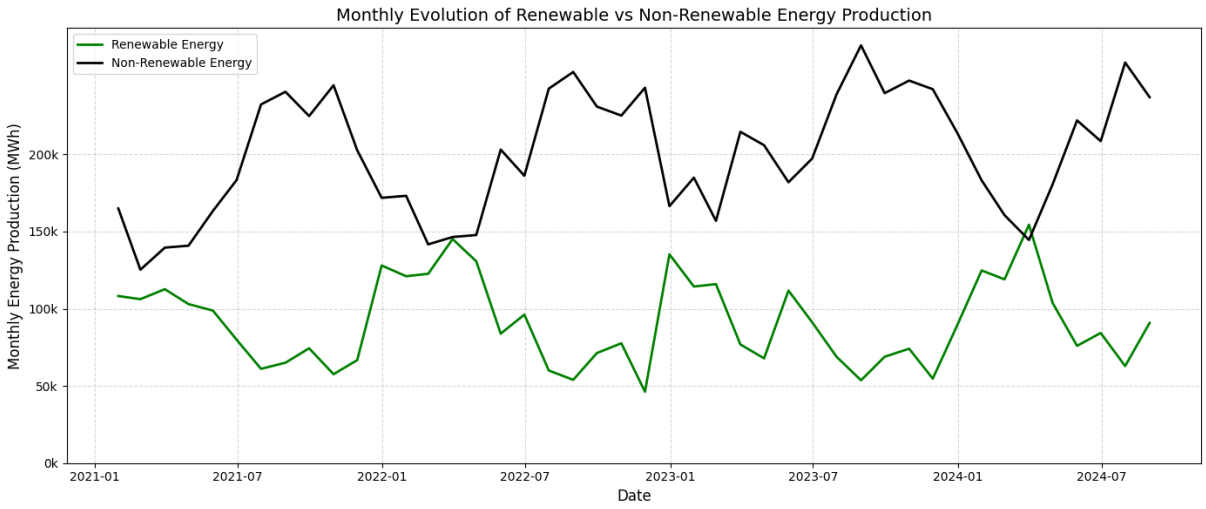


Figure 4.1- Monthly evolution of renewable vs non-renewable energy production

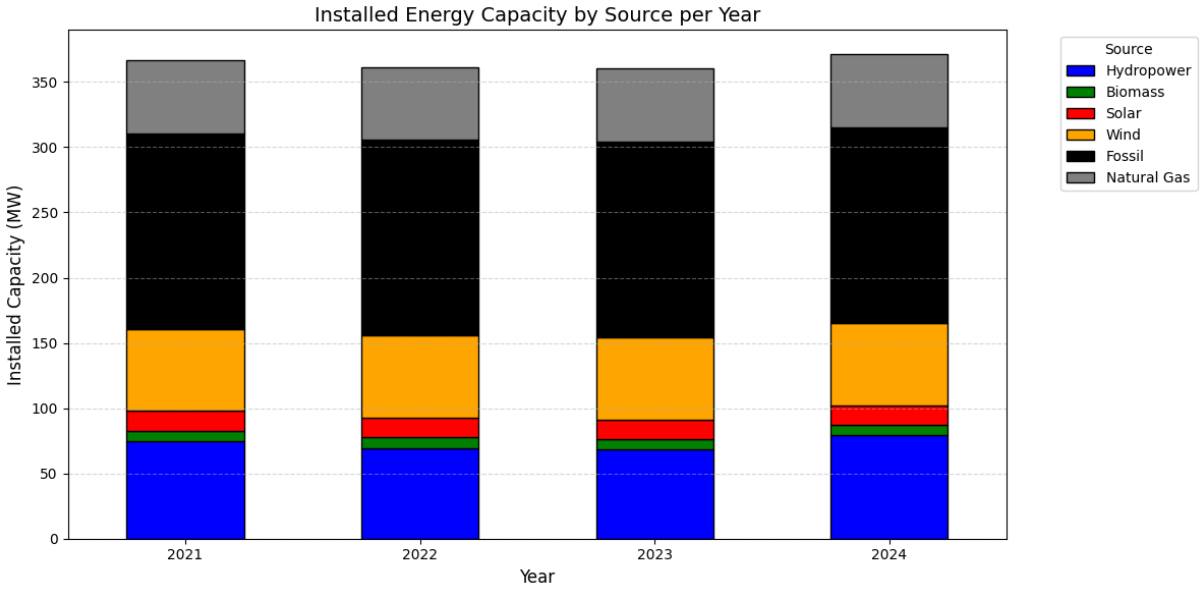


Figure 4.2- Yearly evolution of renewable vs non-renewable energy installed capacity

Figure 4.1 supported the observation introduced in Chapter 1 of this project, which was that Madeira Island consistently exhibited higher levels of non-renewable energy production compared to renewable energy production. The only exception occurred in March 2024, when renewable energy production briefly surpassed non-renewable output. The graph also highlighted the already mentioned seasonal variability of renewable energy generation, characterized by recurring peaks and troughs.

Figure 4.2 illustrates that non-renewable sources continued to dominate the installed energy capacity in the region throughout the analysed period, despite a slight increase in renewable energy capacity in 2024. Among renewable sources, hydropower and wind represented the largest shares of installed capacity. Solar had the 3rd largest installed capacity in the region, whereas biomass had the lowest installed capacity. Notably, the installed capacity remained

constant across most sources over the years, with hydropower being the only one to show variation.

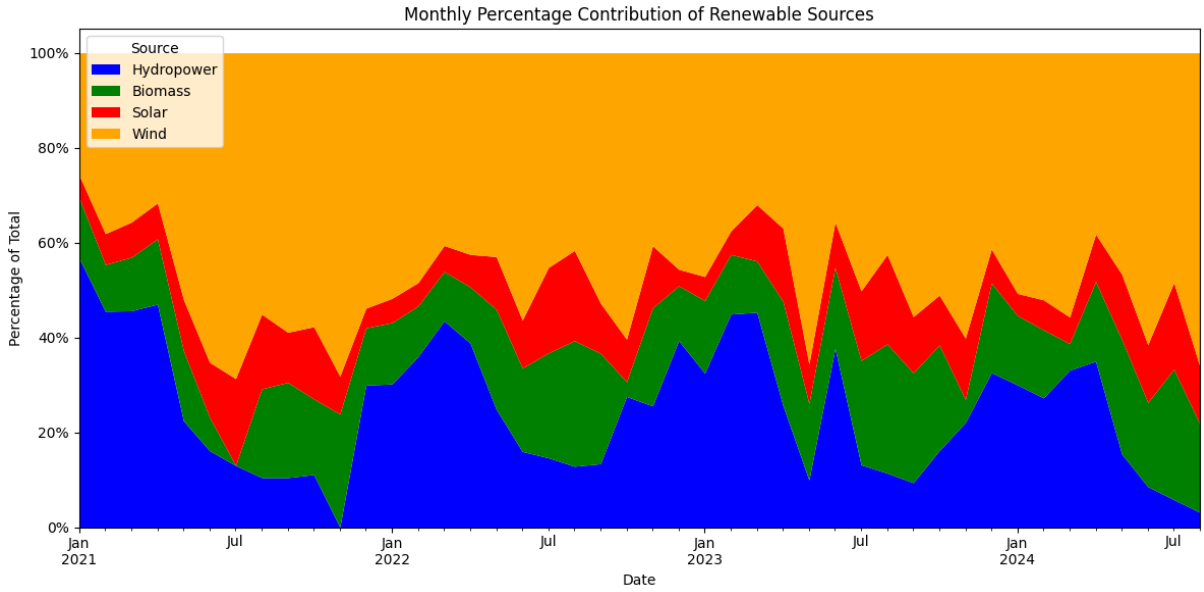


Figure 4.3- Monthly percentage contribution of renewable sources

Figure 4.3, presented in a stacked area format, illustrated the evolving relative contributions of each renewable energy source over time. Wind energy was consistently the dominant contributor to renewable production, despite not having the highest installed capacity. In contrast, hydropower, while holding the largest installed capacity among renewables, produced the second highest share of energy. All four sources exhibited clear seasonal variability, reflecting the influence of environmental conditions on renewable energy output.

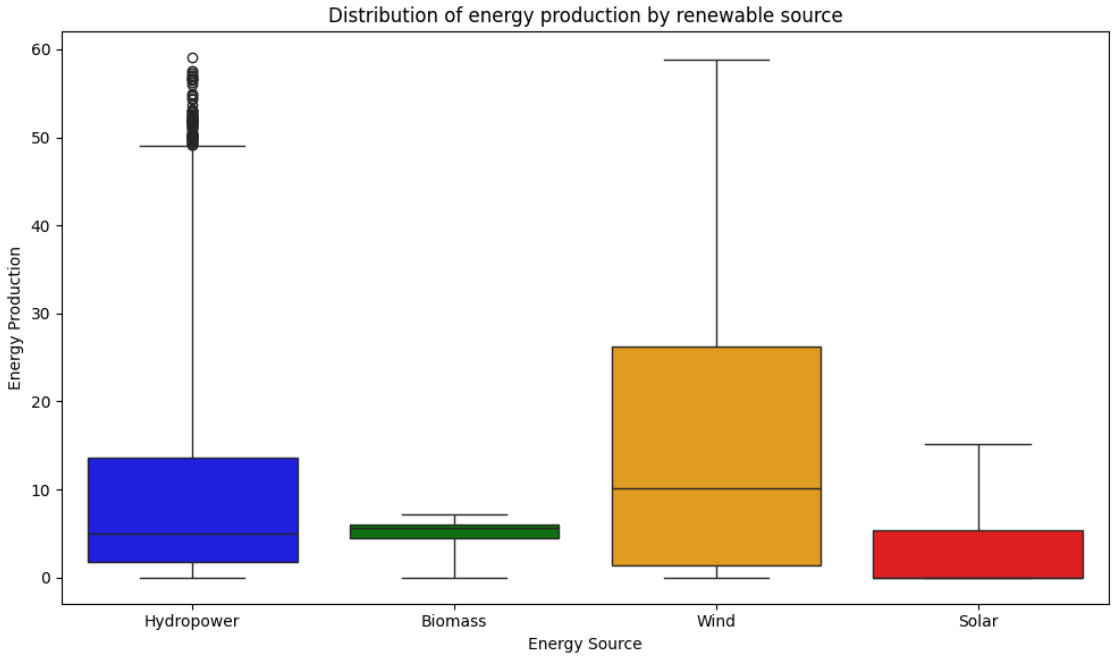


Figure 4.4- Distribution of  energy production by renewable sources

Outliers are data points that significantly deviate from the rest of the dataset and can disrupt analysis, thus reducing forecast accuracy (Maciejewski et al., 2024). Additionally, assessing the data distribution is essential, as non-normal distributions can introduce bias into model predictions (Cabello-López et al., 2023). Boxplots provide a clear and effective method for identifying both outliers and distributional skewness, making them a valuable tool in forecasting projects. This is precisely the role of Figure 4.4 in this analysis.

A boxplot is a statistical graph composed of several elements that summarize the distribution of a dataset. The coloured box represents the middle 50% of the data, bounded by the 25th percentile (Q1) at the lower edge and by the 75th percentile (Q3) at the upper edge. The line within the box indicates the median (Q2), dividing the data into two halves. Thus, 50% of the data points lie below the median (either within the box or along the lower whisker), and the remaining 50% lie above it.

The length of the box is known as the Interquartile Range (IQR), calculated as the difference between Q3 and Q1. The whiskers extend from the box to the data points that fall within a specified range: the lower whisker ends at $Q1 - (whis \times IQR)$ and the upper whisker at $Q3 + (whis \times IQR)$. In this analysis, the *whis* value was set to 3, a more conservative threshold chosen to accommodate the natural variability caused by extreme weather events. Data points falling outside the whisker bounds are classified as outliers, indicating observations that deviate significantly from the central distribution of the variable (Perinetti, 2019).

Additionally, a boxplot also infers about the skewness of a variable. Skewness is a measure of the asymmetry in a data distribution. A distribution is right skewed when most values are concentrated on the lower end, with a long tail extending to the right, indicating a few unusually high values. In contrast, a left-skewed distribution has most values on the higher end, with a long tail extending to the left, reflecting a few unusually low values that pull the mean below the median (A. A. Khan et al., 2021).

Bearing all this in mind, the analysis of the boxplot revealed that only hydropower exhibited outliers, while the other renewable sources remained within the whisker-defined range. However, the number of outliers in hydropower was minimal, amounting to just 0.08% of the total data points, hence not representing a significant portion of the dataset. This low percentage indicated that, despite some extreme values, hydropower data remained largely consistent within the expected variability.

Additionally, the boxplots revealed that the production data was recorded using an extrapolation-based approach. This was evident because the upper bounds of the boxplots reached the installed capacity values for all energy sources, something that is not physically possible within a 15-minute interval. Since installed capacity was measured in MW it represented the maximum energy that can be produced over the course of an hour. Therefore, for a 15-minute period, actual production could not approach these limits unless it

had been extrapolated. This was something noted to address during the data preparation stage (Mytton & Ashtine, 2022).

Considering the distribution, hydropower and wind exhibited substantial amplitude and clear right-skewness, as indicated by the extended upper whiskers and relatively low medians. This suggested that most production periods output low to moderate values, but both sources could reach very high peaks. Solar also demonstrated right-skewness, with a large concentration of values near zero due to the absence of solar generation during nighttime. In contrast, biomass displayed very low variability and a highly compact distribution, reflecting its consistent and controlled nature (*United Nations- Climate Action*, 2024). Notably, the right-skewness observed in the data was addressed during the preprocessing stage to ensure statistical assumptions were met in subsequent analyses.

Based on the data understanding process, several issues were identified and appropriately addressed during data preparation, including missing values in hydropower data, rows affected by DST, the use of an extrapolated-based approach, and the skewness present in the four target variables.

### 4.1.2. Weather

The database source chosen to retrieve weather data from Madeira Island was Visual Crossing Weather API. Obtaining this weather data was an essential step of this research, as weather variables are important input features to compute accurate renewable energy production (Castillo-Rojas & Salinas, 2024).

However, the energy production data provided by EEM was aggregated for the entire island and did not differentiate between individual production sites. In contrast, the Visual Crossing Weather API offers data from three distinct weather stations across the island. To better capture the island's varied weather conditions, this study selected two of those stations, which were FW1564 Funchal PT and LPMA. These locations were chosen specifically because weather patterns differ significantly across the island, and the goal was to aggregate them during the preprocessing stage, by computing the mean of both, to have a more complete look on the island’s different weather impact into renewable energy production.

The locations of both weather stations on the island are shown in figure 4.5, highlighted with yellow circles. Additional details about the weather stations are provided in Table 4.2:

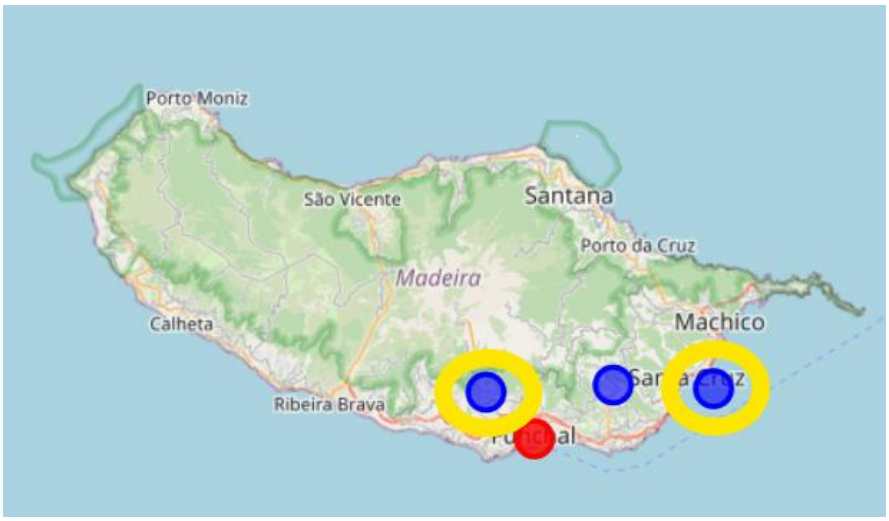


Figure 4.5- Selected weather stations on Madeira Island

Table 4.1- Information about selected weather stations

| Station | ID | Latitude | Longitude |
|---|---|---|---|
| FW1564 Funchal PT | F1564 | 32.678 | -16.945 |
| LPMA | LPMA | 32.680 | -16.770 |

After selecting these two weather stations, all available weather variables were downloaded for the period from 1st of January 2021, to 31st of August 2024, aligning with the energy production dataset.

An initial inspection revealed that the API provided a total of 28 distinct variables and 32 315 data records. Unlike the energy production data, the weather data was recorded in an hourly format. The downloaded variables were the following, for both weather stations: date_time, latitude, longitude, datetime, datetimeEpoch , temp , feelslike , humidity, dew, precip, precipprob, snow, snowdepth, preciptype, windgust, windspeed, winddir, pressure, visibility, cloudcover, solarradiation, solarenergy, uvindex, conditions, icon, stations, source and tzoffset. A detailed explanation of each variable can be consulted on the weather data documentation of Visual Crossing Weather API (*Visual Crossing API Weather Data Documentation*, 2023). Not all variables were relevant or of sufficient quality for this work, due to issues such as imbalance or lack of interpretability, as explained later. However, Appendix B includes two tables, one for FW_Funchal (B.1) and another for LPMA (B.2), with metadata for the features that were considered during data preparation.

After this initial inspection, a more detailed inspection was performed to both weather stations to find possible incoherences, such as missing values, inconsistencies, skewness

among any other possible situation. The first one related to the existence of missing values in the dataset, as per table 4.2:

Table 4.2- Variables with missing values

| Variable | Missing values FW_Funchal | Percentage of missing values FW_Funchal | Missing values LPMA | Percentage of missing values LPMA |
|---|---|---|---|---|
| preciptype | 30 622 | 95.29% | 30 981 | 96.41% |
| windgust | 5940 | 18.48% | 5940 | 18.48% |
| visibility | 3 | 0.01% | 3 | 0.01% |
| solarradiation | 11 | 0.03% | 12 | 0.04% |
| solarenergy | 11 | 0.03% | 12 | 0.04% |
| uvindex | 11 | 0.03% | 12 | 0.04% |
| tzoffset | 28 690 | 89.28% | 28 690 | 89.28% |

As previously mentioned, missing values pose challenges and should either be removed or imputed using appropriate techniques. However, imputation itself can lead to issues such as inaccurate results or degraded model performance. The higher the proportion of missing data in a variable, the more caution is required when selecting an imputation method. While there is no universal rule for handling missing data, it is generally advisable to discard variables with more than 20% missing values (*Centraal Bureau Voor de Statistiek*, 2025). For this reason, preciptype and tzoffset were deleted for both weather stations.

The remaining missing values were handled during the preprocessing stage, with particular attention given to the windgust variable, as its proportion of missing data was close to the defined threshold.

As in the previous chapter, both weather datasets exhibited missing timestamps at the end of March (28/03/2021, 27/03/2022, 26/03/2023 and 31/03/2024) and duplicated timestamps at the end of October (31/10/2021, 30/10/2022 and 29/10/2023) , due to the DST transitions. However, in both weather datasets, the DST adjustments in 2023 were recorded on different days as the ones recorded in the energy production dataset, which required additional attention when merging the datasets.

Additionally, other variables were discarded for various reasons identified during the exploratory analysis, as summarized in the following table:

Table 4.3- Motives to discard variables

| Motive to discard | Variables |
|---|---|
| No relevant information | datetimeEpoch and datetime |
| Little to no variation | latitude, longitude, snow and snowdepth |
| Text Variables (can't merge with mean) | Stations, icon, conditions and source |

The final stage of this data understanding involved the assessment of potential outliers and skewness on the input variables. Once again, boxplots were used for visual representation, and these can be found in Appendix C, Figures C.1 (FW_Funchal) and C.2 (LPMA). These boxplots provided an overview of the distribution and variability of the weather-related variables in this study.

Most features, such as temp, feelslike, humidity, and windspeed, displayed relatively symmetrical distributions, suggesting low or no skewness. However, certain variables, particularly precip, windgust, solarradiation, and solarenergy, exhibited right- skewness, as indicated by longer upper whiskers and medians positioned toward the lower end of the boxes. Outliers were present significantly in only four variables, with precip (4.66%), solarenergy (3%), solarradiation (3.48%) and visibility (1.43%) showing a concentration of extreme values.

Winddir was not assessed for outliers, as it was inherently a cyclical variable. While mathematical outliers and distribution could be computed, they would lack meaningful interpretation in this context and require different treatment than linear variables. This was accounted for during data preparation.

The data understanding process applied to the weather datasets enabled this study to identify key issues to address during data preparation, including mismatched DST rows, the need to merge datasets using the mean, missing values across variables, and skewness in the data distributions.

## 4.2 DATA PREPARATION

A thorough and comprehensive data understanding enabled this research to identify patterns within the retrieved datasets and, more importantly, to detect necessary corrections and transformations. These adjustments were essential to ensure the highest possible forecasting accuracy. The treatment of inconsistencies and errors was carried out during the third phase of the CRISP-DM model, known as data preparation.

Data preparation is the stage of the CRISP-DM process where the data is transformed into a format that maximizes the results obtained by subsequent models. This stage involves correcting all inconsistencies identified during the data understanding phase, converting the data into a format suitable for forecasting models, performing feature engineering ( deriving new variables from existing ones), and selecting the most relevant variables through feature selection techniques.

Since this research had a well-defined objective, to propose an efficient and reliable method for forecasting renewable energy production on Madeira Island, all actions taken during this stage were guided by that goal. To ensure the use of appropriate tools and methodologies, the research drew inspiration from the findings of the conducted SLR.

Based on the data understanding phase, the necessary actions for completing the data preparation were organized into the following categories:

**1)** Merging all datasets
**2)** Data transformation
**3)** Feature selection

### 4.2.1. Merging all datasets

The initial intermediate objective of this stage was to merge the two weather datasets (FW_Funchal and LPMA) by averaging their values and subsequently combine the result with the energy production dataset, to assess relationship between input and output variables. Before merging all datasets, it was necessary to:

- Perform sine and cosine transformations to winddir, as this was a cyclical feature.
- Aggregate the energy production dataset into an hourly format to ensure alignment with weather data.
- Remove conflicting DST entries across all datasets to ensure temporal consistency.

Sine and cosine transformations are common technique used to bypass the problem of representing cyclical variables in ML models. This problem also affected winddir, as it was cyclical in nature (0° is the same as 360°) and using it directly could mislead models that assume a linear progression. The solution to this was applying sine and cosine

transformations, creating two new features: wind_dir_sin and wind_dir_cos. This approach preserved winddir in a continuous and cyclical format, making it more suitable for ML algorithms (Goutte et al., 2024).

To apply this technique, winddir was first converted from degrees to radians. Then, two new features (wind_dir_sin and wind_dir_cos) were created by applying the sine and cosine functions, respectively, to the transformed values (Goutte et al., 2024). The original winddir variable was subsequently removed, as its raw form no longer added value to the project. Although this was a feature engineering step, it needed to be done before merging both datasets. This technique was applied to other cyclical variables during feature engineering.

Subsequently, FW_Funchal and LPMA were merged into a single weather dataset by averaging the corresponding variables, as both datasets contained only numerical data.

The energy production dataset was aggregated to an hourly format. Prior to this, all 15-minute values were multiplied by 0.25 to correct the extrapolation-based reporting, as each value represented a projected hourly rate (Mytton & Ashtine, 2022). Without this adjustment, summing the values would have produced unrealistically figures exceeding installed capacity. After correction, energy production values were summed hourly, while installed capacity retained the first value of each hour, as it changes only annually.

The final step involved removing the conflicting DSTs row from both datasets. In 2023, the DST shift occurred on March 26$^{th}$, in the energy production dataset and on March 28$^{th}$, in the weather dataset. After resolving this discrepancy, the datasets were merged into a single dataset comprising 32 131 rows and 30 columns. However, due to missing data questions, the aggregation to a daily format, matching the forecast horizon, could not be performed at that point.

### 4.2.2. Data Transformation

The train-test split is essential in ML to ensure unbiased model development. The model learns from the training set, while the validation set supports performance monitoring and hyperparameter tuning without data leakage. The untouched test set then offers a more realistic measure of generalization. In this study, the data was split into 70% training, 15% validation, and 15% test, aligning with recommendations from the literature. It is important to note that all transformations performed to the train dataset were also applied to validation and test (Shering et al., 2024).

Table 4.4- Train test split performed

| Dataset | Start date | End date | Percentage |
|---|---|---|---|
| Train | 01/01/2021 00h | 27/07/2023 06h | 70% |
| Validation | 27/07/2023 07h | 13/02/2024 02h | 15% |
| Test | 13/02/2024 03h | 31/08/2024 23h | 15% |

After this step, this research used Feature Engineering, creating the following time dependent variables:

Table 4.5- Feature engineered time variables

| Feature Name | Description |
|---|---|
| Day of the Month | Numeric day of the month (1-31) |
| Month | Numeric month of the year (1-12) |
| Season | Season of the year, defined as:<br>Winter (Dec 21 – Mar 19)<br>Spring (Mar 20 – Jun 20)<br>Summer (Jun 21 – Sep 21)<br>Autumn (Sep 22 – Dec 20) |
| Day of week | Numeric day of the week, with Monday = 0 and Sunday = 6. |
| Is weekend | Binary variable, indicating weekend (1) or weekday (0). |

These engineered variables captured cyclical time patterns, aiding the model in understanding temporal effects on renewable energy production, particularly through Season and Month. The variables Day of Week and Is Weekend highlighted variations in labour activity, which could impact biomass energy due to its link with factory-based combustion processes (United Nations – Climate Action, 2024).

Next, the Season variable was transformed using label encoding, which replaces a categorical value with a unique numerical label (Hancock & Khoshgoftaar, 2020). Hence, each season was assigned a numerical label as follows: winter = 0, spring = 1, summer = 2 and autumn = 3.

Even though variables like Day of the Month, Month, Season and Day of week were not measured in degrees, they still exhibited a cyclical nature. Therefore, sine and cosine transformations were applied to these variables as well. Unlike winddir where cycle length corresponded to 360° (or 2π), these time-based variables had different cycle lengths reflecting their specific periods.

For instance, there are 7 days on a week, so Day of week had a cycle length of 7. Consequently, the formula to apply here was adjusted accordingly, as shown in figure 4.8. Similarly to winddir, the original forms of these transformed variables were also deleted, as they contained no relevant information for this project after the transformation.

$$\left(\sin\left(\frac{2\pi \times \text{Value}}{\text{Cycle Length}}\right)\right) \quad \left(\cos\left(\frac{2\pi \times \text{Value}}{\text{Cycle Length}}\right)\right)$$

Figure 4.6- Sine and Cosine transformations (Kapoor, 2024).

With time variables now reflecting their cyclical nature, spearman correlation was applied to assess statistical dependence between features. This non-parametric measure ranges from -1 to 1, with values near -1 or 1 indicating strong inverse or direct relationships, respectively (Campos et al., 2024).

While other correlation types existed, spearman was most suitable for this research, as it captures consistent trends even in non-linear relationships, common in renewable energy data (Shabbir et al., 2022). Although it could have been applied during the data understanding phase, it would have been ineffective before transforming the time variables into their sine and cosine representations, which was necessary to reflect their cyclical nature.

Correlation analysis here primarily supported decisions on handling missing values, but nonetheless allowed to obtain some insights into the target variables' relationship with the input data:

- **Hydropower –** high positive correlation with Season_cos (0.62), Month_sin (0.57), Month_cos (0.35), high negative correlation with temp (-0.52) , feelslike (-0.52) and moderate negative correlation with dew (-0.40).

- **Biomass –** Only moderate positive correlation with Season_sin (0.29).

- **Solar -** very high correlation with solarradiation and solarenergy (both 0.94), moderate correlation with temp and feelslike (both 0.44) and with humidity (-0.36).

- **Wind –** high correlation with windgust (0.60) and moderate correlation with windspeed (0.35).

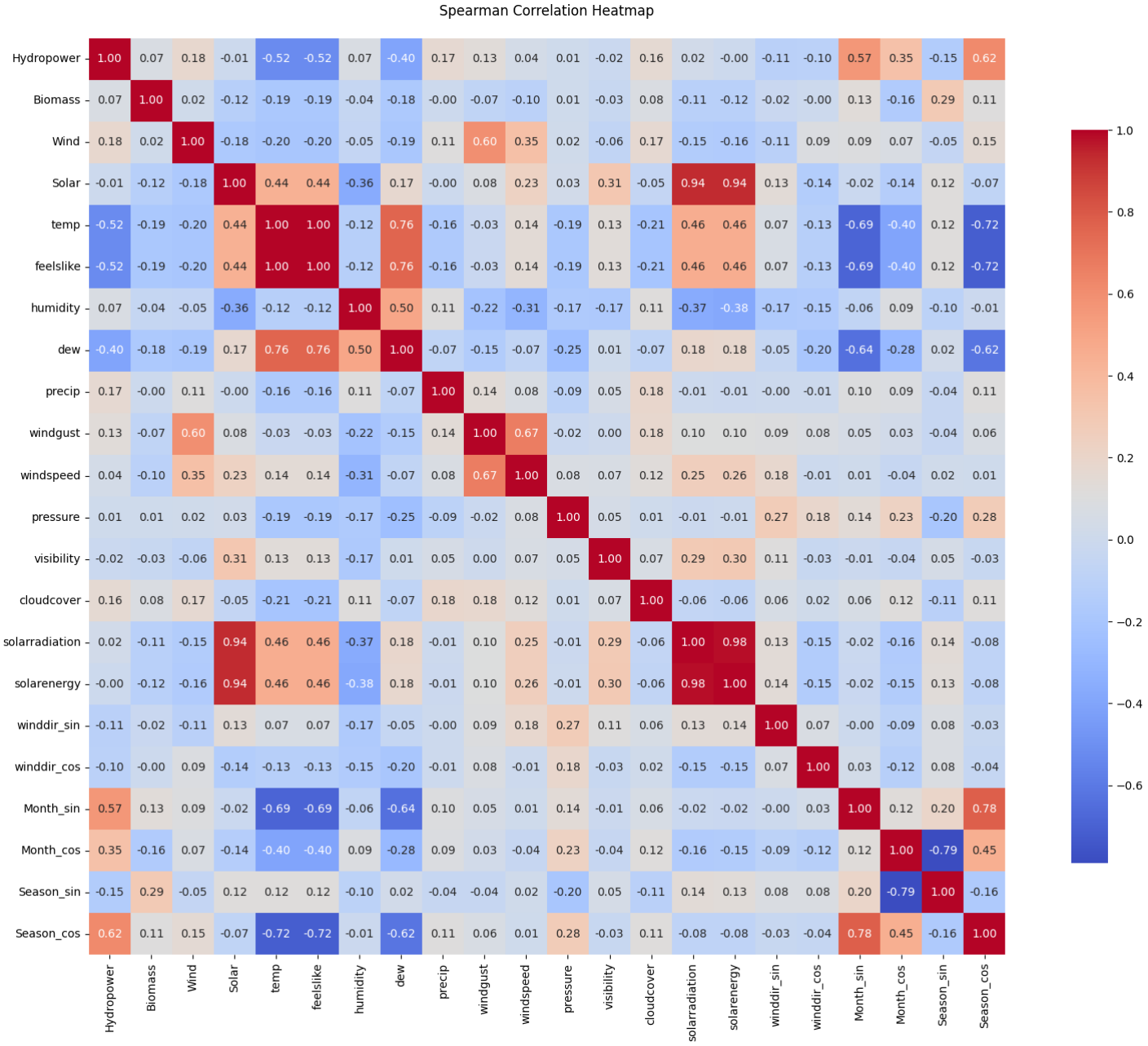


Figure 4.7- Spearman correlation heatmap

For readability, not all variables were included in the heatmap, namely, installed capacities due to lack of variance, non-RES variables as they were not central to the study, and day-related variables, which were excluded in favour of more relevant time-based features.

Furthermore, the next step was addressing the situation of the missing values, which were minimal for visibility, solarradiation, solarenergy and uvindex, but substantial for hydropower and windgust.

The imputation method used for the missing values of visibility, solarradiation, solarenergy and uvindex was linear interpolation**.** Linear interpolation imputes missing values by computing a weighted average of the surrounding points. The use of this technique was appropriate, as there were few missing values in these variables, and in time-dependent datasets, like this one, values tend to be like nearby values (W. Khan et al., 2022), (Shering et al., 2024) .

On the other hand, windgust and hydropower were critical variables for this research, with hydropower being one of the target variables, and windgust emerging as a key feature, due to its high correlation with wind. Given their importance and the high proportion of missing values, a more robust imputation method was employed, which was Random Forest imputation.

Although primarily a ML model, Random Forest is also effective for imputing missing values. It captures complex nonlinear relationships, handles large datasets well, and often outperforms other imputation methods (Tang & Ishwaran, 2017). These strengths made it a robust choice for this task. As granularity is important for accuracy, the dataset was only resampled to a daily format after imputation.

For hydropower, imputation features were selected based on correlation analysis and included temp, dew, Season_cos, Month_sin, and Month_cos. The model was trained on the available data and evaluated using hydropower data from December 2021, a temporally close month. It achieved a MAE of 2.40 MWh and an R-squared value of 0.86. Given this strong performance, no further refinements were deemed necessary, and the model was used to impute all missing values for November 2021.

For windgust, the selected features were wind and windspeed, given the correlation analysis. Unlike hydropower, windgust showed no clear pattern in its missing data. The model was trained on the training dataset and achieved a MAE of 2.42 and an R-squared value of 0.93. Given this strong performance, no further adjustments were necessary, and the model was used to impute the remaining missing values.

After addressing all data consistency and integrity issues, the dataset was aggregated to a daily format to meet prediction requirements. Variables provided by EEM were processed using the previously described method: production values were summed, installed capacity was taken from the first occurrence of each day, and weather variables, including sine and cosine features, were averaged. The only exception was precip, which was summed. This aggregation resulted in 938 days for train, and 204 and 201 for validation and test, respectively.

Once the data was converted to a daily format, lag and moving average features were created to capture temporal patterns. Lag features use past values (e.g., windgust_lag1 for yesterday's windgust) to help predict future outputs, based on the assumption that historical data holds predictive value (Surakhi et al., 2021). Moving averages or rolling means smooth short-term fluctuations to highlight underlying trends. This is particularly useful, for example, in forecasting wind energy (Su et al., 2022).

Accordingly, lag (1, 3 and 7 days) and rolling mean features (3 and 7 days) were generated for the four target variables (hydropower, wind, solar, biomass), as well as for humidity, solarradiation, windgust, and windspeed.

The final step in data transformation was normalization, ensuring all variables contributed equally during model training (Campos et al., 2024). To reduce skewness observed during the data understanding phase, a square root transformation was first applied, as normalization is sensitive to data range (Bartlett, 1936). This transformation mitigates skewness by compressing large values and stabilizing distributions, thereby enhancing forecast accuracy, as identified during the data understanding phase. It was applied to all features except to cyclical variables, which include negative values and are incompatible with square root transformation. Outliers were not removed in this research, as they amounted to a low percentage of total data points.

Consistent with the approaches in the SLR, this project used the MinMax Scaler for normalization. It scales all values to the [0, 1] range, where 0 represents the minimum and 1 the maximum of each variable, using the standard formula. This transformation was saved via joblib, so we could inverse- transform future predictions, hence reporting errors back in MWh (*Joblib*, 2021).

$$x' = \frac{x - \min}{\max - \min}$$

Figure 4.8- Min Max Scaler (adapted (Dralus et al., 2023))

Lastly, the dataset was partitioned by renewable source, as recommended in the SLR, since feature-selection algorithms produce different results depending on the target variable (Koeva et al., 2023).

### 4.2.3. Feature Selection

Feature selection is the phase of the data preparation, in which each technique applies its own mechanisms to identify the variables with the highest predictive power, thereby maximizing forecasting accuracy (Cabello-López et al., 2023).

The feature selection methods discussed in the SLR were not suitable for this project. As a result, alternative methods were identified from other sources, which were deemed more relevant to the project's needs.

The used feature selection methods were spearman correlation, variance threshold, XGBoost feature importance and mutual information, always minding the obtained domain knowledge.

Spearman correlation helps identify and remove one feature from pairs that are highly correlated, although it requires understanding of causal relationships, so no strict threshold could be defined to discard features.

The variance threshold method discards features with very low variance, typically removing those with zero variance, which was the threshold defined by this research (Kamalov et al., 2025).

Mutual information evaluates the dependency between each feature and the target variable. This research discarded immediately features that were not in the top 10 features with the highest scores based on the method (Wang et al., 2024).

Lastly, XGBoost feature importance ranks features by their contribution to model performance during training. Once again, features not among the top 10 were immediately discarded (Wang et al., 2024).

All these methods complemented each other and were combined to enhance feature selection. Based on the results and defined thresholds, the features selected to be used during the modeling stage were:

Table 4.6- Selected features per source.

| Renewable Energy Source | Selected features |
|---|---|
| Solar | solarenergy, cloudcover, Month_cos, precip, Season_sin, Solar_roll3, Solar_lag1, Solar_roll3, Solar_roll7 and solarradiation_lag7 |
| Wind | windgust, windspeed, Wind_lag1, winddir_cos, windir_sin and windgust_lag1 |
| Hydropower | Hydropower_lag1, Hydropower_roll3, Hydropower_roll7, Hydropower_lag3, Hydropower_lag7, Month_sin, temp, Season_cos, precip and solarenergy |

| | |
|---|---|
| Biomass | Biomass_lag1, Biomass_roll3, Biomass_roll7, Biomass_lag3, Biomass_lag7 and Month_sin |

## 4.3 MODELING

The modeling stage involves applying selected models to the data to evaluate the effectiveness of the defined methodology. In this specific case study, it served to assess whether the proposed solution could effectively address the RQ: "What are the best approaches to accurately forecast renewable energy production in Madeira", hence presenting an efficient method to forecast RES production in Madeira Island.

For these research's day ahead forecasts, four different models were selected, based on the SLR: LSTM (Long-short-term-memory), XGBoost (Extreme Gradient Boosting), Random Forest and MLP (Multilayer Perceptron).

Both LSTM and MLP are types of artificial neural networks: as a deep learning, LSTM is a gated recurrent architecture that maintains an internal state to capture long-term dependencies (Campos et al., 2024). By contrast, MLP is a purely feed-forward ANN that processes each input independently, without built-in memory (Gómez et al., 2020).

XGBoost together with Random Forest are tree-based models. XGBoost is a scalable, regularized gradient-boosted tree algorithm that builds trees sequentially, with each new tree correcting the residual errors of its predecessors, to efficiently minimize a loss function. On the other hand, Random Forest trains many decision trees in parallel on bootstrapped samples and averages their predictions to reduce variance and guard against overfitting (Bochenek et al., 2021).

Each of the four algorithms is driven by a unique rationale. Their performance was compared using the three metrics most frequently reported in the literature: mean absolute error (MAE), root mean squared error (RMSE), and R-squared.

Table 4.7- Description of each error metric (adapted from (Campos et al., 2024)).

| Error metric | Description | Formula |
|---|---|---|
| MAE | Average of the absolute values of errors | $\frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \hat{y_i}\rvert$ |

| | | |
|---|---|---|
| RMSE | Average magnitude of model's errors | $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$ |
| R-squared | Proportion of target variance explained by the model | $1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ |

with $y_i$=actual value and $\hat{y}_i$=predicted value

The modelling process underwent three rounds, as described on the next sections. RMSE and MAE were selected as the primary performance metrics, as they give an error metric in real terms, with special focus in RMSE, as it effectively captures the magnitude of prediction errors and penalizes larger deviations more heavily.

### 4.3.1. First forecasting round

The first round consisted of using mostly default of the selected models (with some exceptions), by training each model on the train dataset, and assessing their performance, on the validation dataset, via error metrics. The idea here was to avoid further adjusting underperforming models and focus only on the ones with promising day-ahead forecasts. Thereby, models that showed poor performance did not proceed to the next round.

The LSTM model was configured with two hidden layers containing 64 and 32 units, respectively, and incorporated dropout layers (0.2 and 0.1) to mitigate overfitting. The model was trained using the Adam optimizer for 50 epochs, and a lookback of 3, a commonly adopted configuration in time series forecasting cases. Similarly, the MLP model employed a two-layer architecture with 64 and 32 neurons, while all other parameters were left at their default values to maintain simplicity. For XGBoost, a maximum tree depth of 5 and a learning rate of 0.1 were selected to balance model complexity and stability. The Random Forest model retained all default parameters. All these settings, for all models, were chosen based on the established practices in the literature (Dralus et al., 2023), (Almonacid-Olleros et al., 2020), (Campos et al., 2024), (Bochenek et al., 2021).

Table 4.8- First round error metrics results, in MWh

| RES | Model | MAE | RMSE | R-squared |
|---|---|---|---|---|
| Solar | LSTM | 15.6 | 19 | 0.21 |
| | XGB | 11.3 | 14.6 | 0.55 |
| | RF | 12.9 | 15.8 | 0.33 |
| | MLP | 12.1 | 14.9 | 0.53 |
| Wind | LSTM | 194.3 | 256.9 | 0.25 |
| | XGB | 133.9 | 184 | 0.61 |
| | RF | 147.1 | 195 | 0.65 |
| | MLP | 127.3 | 170.9 | 0.66 |
| Hydropower | LSTM | 63.05 | 111.2 | 0.51 |
| | XGB | 45.2 | 72.8 | 0.79 |
| | RF | 37 | 54.2 | 0.91 |
| | MLP | 44.7 | 77.4 | 0.76 |
| Biomass | LSTM | 16.4 | 27.4 | 0.68 |
| | XGB | 12 | 28.1 | 0.66 |
| | RF | 10.7 | 21.8 | 0.60 |
| | MLP | 10.1 | 19.1 | 0.84 |

All metrics were computed in MWh, to get a clear understanding of the model performance. This reverse scaling of the results was performed by importing the saved joblib scaler file, and by squaring the result to reverse the square root transformation. This procedure of reverse scaling was applied in all forecasting rounds.

Overall, solar and wind energy proved to be the most challenging sources to forecast, likely due to their higher variability and dependence on weather conditions. In contrast,

hydropower and biomass exhibited more stable patterns, leading to more accurate predictions and stronger model performance.

Given that the primary goal of this project was to generate the most accurate day-ahead forecasts as possible, the best model for each source was the one that yielded lower MAE and RMSE. Given this, all models, except LSTM, excelled in forecasting at least one energy source, with XGBoost performing best for solar, Random Forest leading in hydropower prediction, and MLP achieving the highest accuracy for biomass and wind.

In contrast to most studies reviewed in the literature review, the LSTM model yielded the poorest forecasting performance and struggled to capture the nonlinear relationships inherent in weather and RES variables (Cabello-López et al., 2023). This was probably because LSTM models need to be trained on a lot of data, and this research only had 938 days of training. Nonetheless, given the poor results, LSTM did not move on to the next round.

### 4.3.2. Grid Search and Blocked-Cross Validation round

In this phase, Grid Search was used to optimize each model by evaluating a range of parameter combinations. The goal was to identify the configuration that resulted in the lowest RMSE, a metric selected for its sensitivity to large prediction errors (Cabello-López et al., 2023).

Moreover, this study employed cross-validation, a robust procedure to mitigate overfitting and biased results. However, given the presence of temporal dependencies in the data, standard cross-validation was not appropriate, as it neglects the sequential nature of time series. To address this constraint, Blocked Cross-Validation (BCV) was applied, since it preserves the temporal order, by avoiding data leakage. BCV evaluates the model across multiple temporally consistent, resulting in more reliable performance estimates and improved forecasting accuracy (Cabello-López et al., 2023).

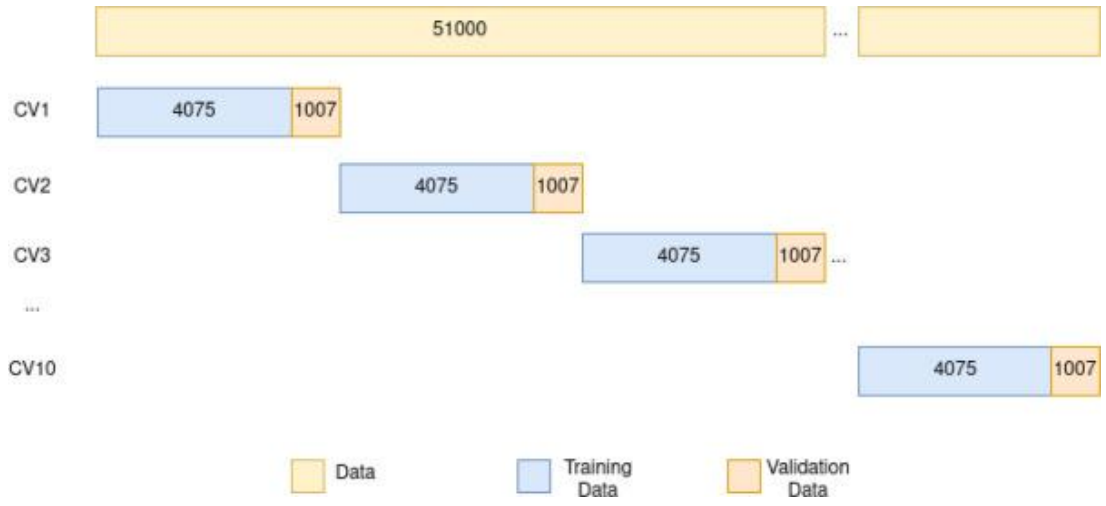


Figure 4.9- Rationale of BCV (Cabello-López et al., 2023).

In BCV, the dataset is split into consecutive blocks. Each block is used once as the validation set, while the blocks before are used as the training set. This preserves the order of the data and avoids training on future information (Cabello-López et al., 2023).

The results obtained using the best hyperparameters from the grid search for each RES, evaluated using 3-fold Blocked Cross Validation on the validation set, are summarized below:

Table 4.9- Second round results, in MWh

| RES | Model | MAE | CV RMSE | R-squared | Optimal Parameters by Best Model per RES |
|---|---|---|---|---|---|
| Solar | XGB | 11 | 14.3 | 0.57 | n_estimators = 150<br>max_depth = 10<br>min_samples_split = 5 |
| | RF | 10.9 | 13.3 | 0.58 | |
| | MLP | 12 | 14.7 | 0.53 | |
| Wind | XGB | 129.1 | 223.7 | 0.65 | hidden_layer_sizes = (128,64)<br>activation = relu<br>alpha = 0.001 |
| | RF | 132.2 | 224.2 | 0.63 | |
| | MLP | 128.1 | 191.8 | 0.65 | |
| Hydropower | XGB | 44.4 | 75.2 | 0.80 | n_estimators = 100<br>max_depth = 3<br>learning_rate = 0.1 |
| | RF | 46 | 68.4 | 0.78 | |
| | MLP | 50.2 | 75.5 | 0.71 | |
| Biomass | XGB | 16.9 | 31.7 | 0.77 | hidden_layer_sizes = (128,64)<br>activation = relu<br>alpha = 0.001 |
| | RF | 11.2 | 31.3 | 0.77 | |
| | MLP | 8.7 | 20.9 | 0.84 | |

Random Forest outperformed XGBoost as the top predictor for solar. XGBoost, however, surpassed Random Forest as the most effective model for hydropower forecasting. MLP continued to be the best at forecasting biomass and wind energy production.

Overall, all models experienced a slight increase in day-ahead accuracy, which was positive, given the more realistic and strict evaluation provided by the methodology. However, the MLP model showed a notably larger increase in error for hydropower, indicating a tendency towards overfitting.

### 4.3.3. Final forecasting Round

For the final round, only XGBoost and Random Forest were considered. The MLP models were excluded due to signs of overfitting and, more importantly, challenges related to interpretability. Although all three models are typically considered black-box models—characterized by limited interpretability due to their complex internal structures and numerous parameter combinations (Hassija et al., 2024) , tools like SHAP (SHapley Additive exPlanations) help mitigate this issue (W. Khan et al., 2022).

SHAP improves model transparency by quantifying how each feature contributed to a specific prediction (W. Khan et al., 2022). However, its implementation for neural networks, such as MLP, is more complex and computationally intensive compared to tree-based models, for which SHAP is both fast and exact (*Shap Documentation*, 2018). Due to this added difficulty in interpretability, the MLP model was ultimately excluded from the final analysis.

In the final evaluation round, XGBoost and Random Forest were assessed for day-ahead forecasts, using the optimal parameters identified in the previous round, for each source. For this step, the models were trained on a combined dataset comprising both the train and validation sets, allowing them to learn from a larger volume of data and thereby improving forecast accuracy.

Crucially, concatenation was done before applying square root and Min-Max scaling to avoid distorting variable scales. This ensured consistent transformation by treating the train and validation sets as a single unit.

The final XGBoost and Random Forest models were subsequently trained on this fully prepared final train dataset and evaluated on the unseen test dataset. This approach ensured a fair and unbiased assessment of the models, simulating the closest to a real-world application scenario, in this project, and avoiding any risk of data leakage. XGBoost revealed to be the best performing model for wind and hydropower, and Random Forest for solar and biomass.

The resulting performance metrics are presented below:

Table 4.10- Final round results

| RES | Model | MAE | RMSE | R-squared | Optimal Parameters by Best Model per RES |
|---|---|---|---|---|---|
| Solar | XGB | 12.9 | 15.8 | 0.32 | n_estimators = 150<br>max_depth = 10<br>min_samples_split = 5 |
| | RF | 12.8 | 15.7 | 0.33 | |
| Wind | XGB | 140.9 | 185.4 | 0.69 | n_estimators = 100<br>max_depth = 3<br>learning_rate = 0.1 |
| | RF | 146.8 | 194.3 | 0.65 | |
| Hydropower | XGB | 39 | 53.8 | 0.910 | n_estimators = 100<br>max_depth = 3<br>learning_rate = 0.1 |
| | RF | 35.1 | 55.3 | 0.905 | |
| Biomass | XGB | 12.26 | 27.6 | 0.35 | n_estimators = 100<br>max_depth = 5<br>min_samples_split = 5 |
| | RF | 10.4 | 21.1 | 0.62 | |

XGBoost emerged as the top-performing model for wind energy and hydropower, while Random Forest achieved better results for solar and biomass energy. However, it's important to highlight that both solar and biomass experienced a significant drop in performance during this final evaluation phase. In contrast, wind and hydropower showed improvements.

As the test set contained unseen data, evaluation was limited to a single pass without cross-validation to avoid selection bias and overly optimistic estimates (Cawley & Talbot, 2010). Preserving test set independence ensured a realistic assessment. Notably, hydropower showed a substantial error reduction, indicating strong generalization (Cawley & Talbot, 2010).

## 4.4 DISCUSSION

XGBoost emerged as the top-performing model for the most produced RES in this study , wind and hydropower. In contrast, the Random Forest models demonstrated superior performance for day-ahead forecasting of solar and biomass.

Based on the findings, the proposed forecasting strategy for Madeira Island was a hybrid one, as it involved utilizing XGBoost models for day-ahead predictions of wind and hydropower generation, while Random Forest models proved more effective for forecasting solar and biomass output, thereby optimizing accuracy. This solution, combined with the data preparation steps performed, is the answer to the central research question: “What are the best approaches to accurately forecast renewable energy production in Madeira?”.

Although this study focused on Madeira, the results suggested that this methodology could be adapted to other regions aiming to align with the United Nations’ 7th Sustainable Development Goal: ensuring access to affordable, reliable, sustainable, and modern energy for all.

This solution offered practical real-world benefits, having undergone rigorous assessment, as depicted in this project. While the models were trained using observed weather data, they can also be supplied with day-ahead weather forecasts to support operational energy planning.

It is important to highlight that the full training dataset spanned from January 1, 2021, to February 13, 2024, while the test dataset covered the period from February 14 to August 31, 2024. As a result, seasonal patterns may have influenced model performance, depending on the characteristics of each RES.

To better understand the strengths and weaknesses of each model in this hybrid solution, a thorough analysis of the error metric trends was essential. This analysis was complemented by line charts that compare forecasted values against actual observations, offering visual insight into model performance over time.

Furthermore, to enhance model interpretability, SHAP values were employed to bypass the black box model problem and visualized using beeswarm plots.

A SHAP beeswarm plot visualizes how features influence model predictions across instances. Each dot shows a data point's SHAP value (impact on output) and colour (feature value, red for high, blue for low). Features are ranked by importance, and the direction of dots indicates whether high or low values increase or decrease predictions, offering insight into both feature relevance and Effect (Ponce-Bobadilla et al., 2024).

The following section presented the key evaluation elements, error metrics, line charts, and beeswarm plots, followed by a detailed analysis of each energy source, grounded in the performance metrics and visualizations of their respective models.

Table 4.11- Error metrics of the proposed hybrid solution

| RES | | MAE | RMSE | R-Squared |
|---|---|---|---|---|
| Solar | (RF) | 12.8 | 15.7 | 0.33 |
| Wind | (XGB) | 140.9 | 185.4 | 0.69 |
| Hydropower | (XGB) | 39 | 53.8 | 0.91 |
| Biomass | (RF) | 10.4 | 21.1 | 0.62 |

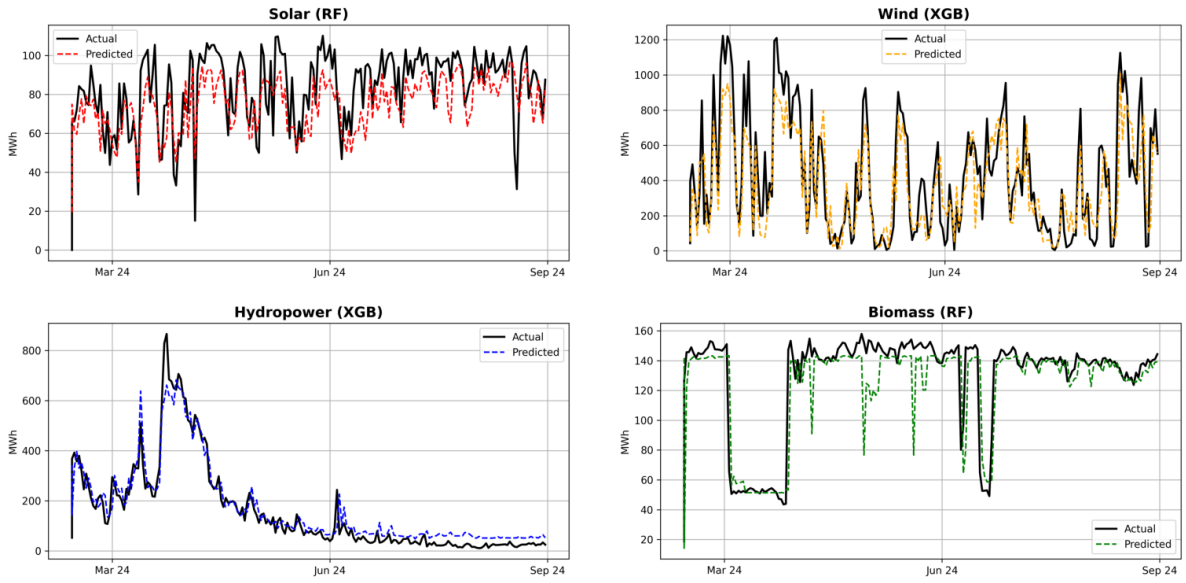


Figure 4.10- Actual vs. forecasted renewable energy generation, in MWh.

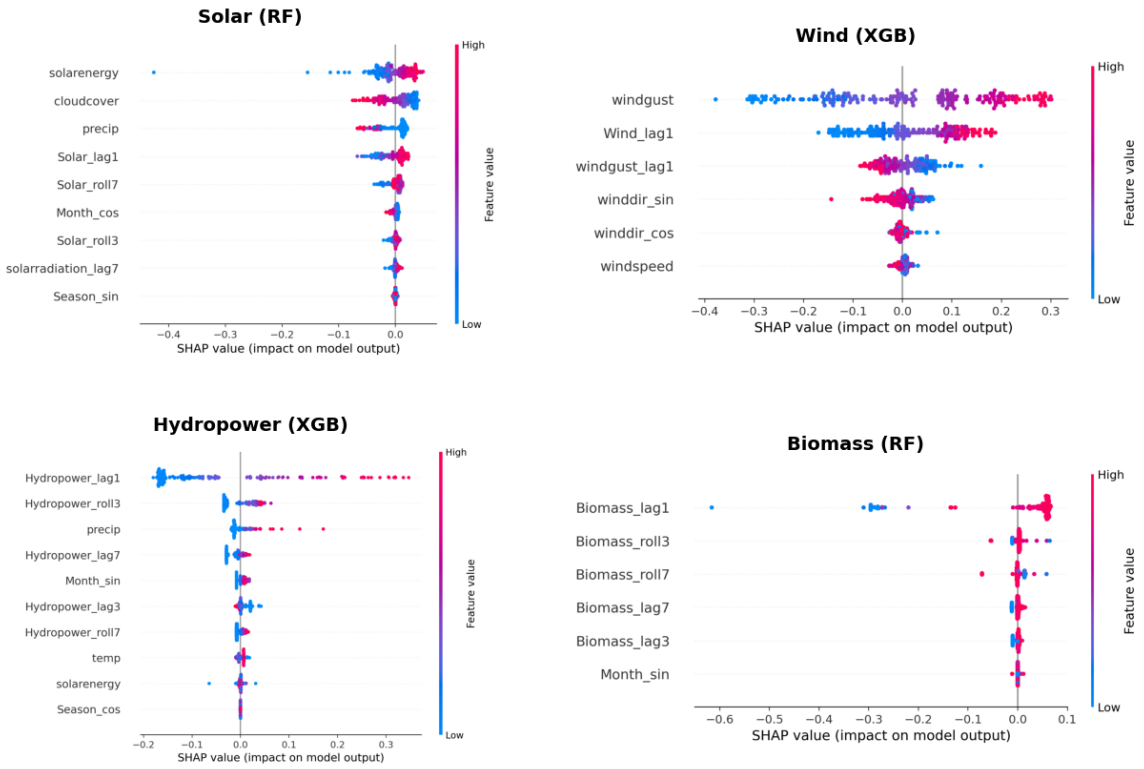


Figure 4.11- Beeswarm plots showing feature impacts for each RES-model combination

### 4.4.1. Solar energy

Although the solar model exhibited only a modest change in MAE and RMSE between modelling iterations, the R-squared value dropped significantly, from 0.59 to 0.33. This suggested that while the model still captured broad temporal trends, it underestimated consistently output during peak production periods, revealing a limited ability to reflect the full variability inherent in solar energy production.

This behaviour is consistent with established findings in the literature, which highlighted that solar energy forecasting tends to perform more effectively on an hourly scale as such intervals reflect more accurately the dynamic nature of weather-related variables, rather than on the day-ahead forecasts employed on this project (Szostek et al., 2024), (Almonacid-Olleros et al., 2020).

The beeswarm plot revealed that solarenergy, cloudcover, precip, and solar_lag1 were the most influential features in the solar energy model. High values of solarenergy (most impactful variable) and solar_lag1 were associated with higher forecasts, reflecting their strong positive contribution to predicted solar output. In contrast, high levels of cloudcover and precip corresponded to lower predictions, highlighting their negative impact on solar production.

### 4.4.2. Wind energy

Wind energy showed a modest improvement in model performance over the course of this modeling stage. The resulting metrics were generally satisfactory, and as illustrated in the line

chart, the model effectively captured the overall trends and fluctuations during the test period.

However, it struggled to accurately predict during the highest production peaks, which negatively impacted its performance metrics. This limitation was consistent with findings in the literature, which noted that forecasting wind energy during periods of peak production remains a persistent challenge in similar case studies (Bochenek et al., 2021). Unlike what was found in the literature, in this study, wind energy proved to be easier to predict than solar (Shering et al., 2024). This was probably because wind energy is considered easier to forecast than solar for day-ahead and longer-term horizons, with one study even attempting month-ahead forecasts, though solar tends to be more predictable at shorter timescales, like hourly forecasts (Szostek et al., 2024).

The beeswarm plot for the wind model showed that windgust, wind_lag1, winddir_sin, and winddir_cos were the most influential features in predicting wind power output. Among them, windgust stood out as the most impactful one, with higher values strongly increasing the predicted output, thereby capturing the model's sensitivity to sudden bursts of wind. Wind_lag1 also played a significant role, indicating that the model leveraged recent wind generation patterns to inform current forecasts. This suggested that when production was high on one day, it was likely to remain elevated on the following day, probably due to the persistence of wind conditions.

Moreover, the directional components winddir_sin and winddir_cos contributed more subtly, yet consistently, indicating that wind direction, while not the dominant driver, still played a meaningful role in shaping wind energy output. However, since these features were transformed trigonometrical representations of wind direction rather than its raw form, and their relationship with output was nonlinear, it was not possible to directly infer which specific wind directions resulted in higher production based solely on the SHAP beeswarm plot. To overcome this limitation, a wind rose visualization was generated to identify the wind directions most strongly associated with elevated wind energy output.

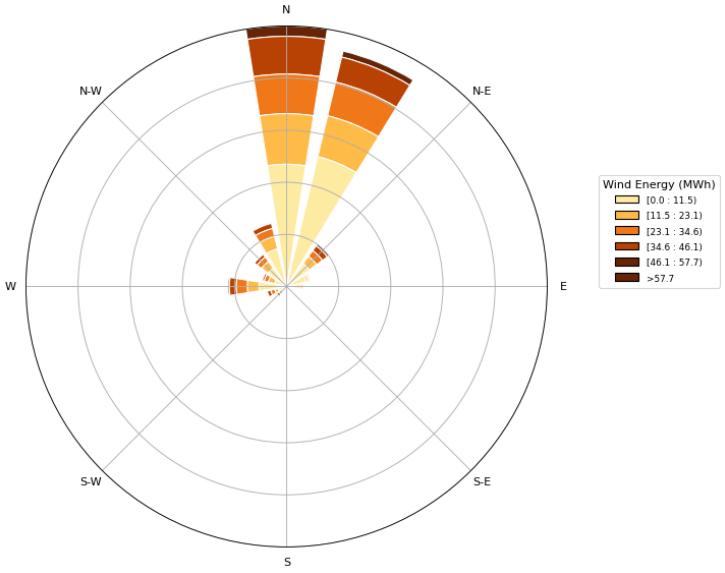


Figure 4.12-Madeira Island's wind rose

The wind rose plot indicated that wind most frequently originated from the north and northeast, as shown by the broader sectors in those directions, since sector size reflected how often the wind blew from each angle. The colour scale represents wind energy output, with darker shades indicating higher production. While wind arrived from various directions and some less frequent sectors showed relatively high output, the maximum energy levels (marked by the darkest shades) occurred exclusively when wind came from the north and northeast. This highlights that winds from those directions were not only the most frequent but also the most productive, yielding the highest levels of wind energy generation in Madeira Island.

Overall, the model appeared to be well-tuned to key meteorological drivers of wind variability, using both magnitude and direction to guide predictions.

### 4.4.3. Hydropower

Hydropower, powered by XGBoost, emerged as the most accurate among all RES, achieving an impressive R-squared of 0.91. This indicates that the model captured hydropower output fluctuations exceptionally well, a conclusion that is reinforced by the near-perfect alignment between the predicted (dashed) and actual values in the time series plot.

While these results were impressive, they should be interpreted with some caution. The test dataset was primarily comprised of spring and summer months, which, as identified during the data understanding stage, mean lower hydropower production. Despite this seasonal bias, the model also demonstrated strong performance during validation phases that included peak production periods, suggesting that it generalized well across different seasons.

The SHAP beeswarm plot for this model showed that the most influential features were hydropower_lag1, hydropower_roll3, precip  and hydropower_lag7. Out of these four, three

were lag-based features, indicating that the model relied heavily on recent output values to predict current production. This pattern is consistent with the nature of hydropower generation, as, while renewable, it is less directly dependent on immediate weather conditions, unlike solar or wind energy. Instead, it is influenced by reservoir storage levels and by water release policies that control flow for energy production and other purposes (Ogliari et al., 2020). To this end, hydropower_lag1 emerged as the most influential feature, indicating that high production on one day often followed high production the day before. Other lag-based features showed similar trends, though with lower sensitivity. Notably, precip also played a key role, particularly during rainy days, when increased rainfall contributed to higher hydropower output.

Unlike the previously discussed sources (wind and solar), it was difficult to draw direct parallels between the hydropower model's results and findings in the existing literature. As noted earlier, there is a clear research gap when it comes to time series case studies focused on hydropower, making it the least explored RES (Krechowicz et al., 2022). This lack of comparable studies limited the ability of this study to benchmark its model's performance or validate the relevance of key features against established findings in the literature.

Nonetheless, like the two additional hydropower studies, this project confirmed that precip contributed to improved forecast accuracy. Although not the most influential feature, it still played a meaningful role in hydropower prediction, despite this source's dependence to water release policies (Ogliari et al., 2020).

### 4.4.4. Biomass energy

The biomass energy model, built using Random Forest, performed reasonably well in terms of error metrics. However, given that it experienced decreases in model performance on the final round, it indicated overfitting, which meant that the model struggled in finding relevant patterns that justified biomass production. This was likely because biomass is not driven by weather conditions, but rather by operational factors that go beyond the day or weekend effects, which were the basis of some newly created features in this study (*United Nations-Climate Action*, 2024).

Moreover, the absence of forecasting case studies on biomass during the SLR, reinforced the notion that biomass is primarily dependent on operational planning. Nonetheless, the model revealed a moderate ability to explain variation in output. The time series plot showed that the model tracked overall trends accurately during stable periods but tended to underestimate sudden drops, which could have been caused by operational anomalies.

The SHAP beeswarm plot revealed that the most influential features were all lag-based, with a strong focus on biomass_lag1. Unlike the other studied RES, weather variables played no role here, suggesting that output was largely independent of immediate environmental conditions and only dependent on operational decisions.

# 5. CONCLUSION

## 5.1 SYNTHESIS OF THE DEVELOPED CASE STUDY

This project aimed to support the United Nations' 7th Sustainable Development Goal by enhancing the performance of renewable energy systems, thereby contributing to affordable and clean energy for all (Environment, 2017).

Specifically, this was achieved through day-ahead forecasting of renewable energy production on Madeira Island, using data on solar, wind, hydropower and biomass energy production, provided by Empresa de Eletricidade da Madeira, from January 1st, 2021, until August 31st, 2024.

Due to its insular nature, Madeira remains heavily reliant on fossil fuels, reinforcing the critical need for accurate renewable energy forecasting. This research was guided by its primary objective of developing an efficient and accurate forecasting method to address a clear gap in the existing literature, hence answering the defined research question, "What are the best approaches to accurately forecast renewable energy production, in Madeira?". To answer the RQ, intermediate objectives had to be established. Although focused on a regional case study, the findings were expected to be applicable to other locations with similar energy challenges.

Bearing all this in mind, a Systematic Literature Review was conducted following PRISMA guidelines to identify the most relevant studies offering insights into data preparation, modeling frameworks, and techniques used in RES forecasting case studies. This step was crucial for drawing key findings, which were later applied in this project, such as effective data transformation methods, the importance of incorporating weather data and where to source it, the use of both traditional machine learning and deep learning models, and the value of day-ahead forecasting for each energy source.

While a variety of literature was available for solar and wind energy forecasting, the review revealed a notable research gap, the limited number of studies focused on forecasting hydropower production. This gap constrained the ability to benchmark and compare the study's hydropower forecasting results against existing work.

All these findings were applied using the CRISP-DM framework, aiming to produce high quality day-ahead forecasts. CRISP-DM guided the project from business understanding through to evaluation, with each step adding value to this project. Business understanding highlighted the relevance of adding weather data to EEM's dataset, which was achieved via Visual Crossing Weather API for two distinct weather stations in Madeira Island. Data understanding allowed to note all situations that needed to be addressed, to achieve high quality forecast during the data preparation stage.

Finally, after multiple modeling iterations, the research produced a final solution with real-world applicability. This solution was presented during the evaluation phase of the CRISP-DM framework. It took the form of a hybrid approach: XGBoost was used for forecasting wind and hydropower generation, while Random Forest was applied to solar and biomass, with model parameters optimized through Grid Search. The results showed strong performance for hydropower, acceptable accuracy for wind, and a reasonable outcome for biomass, though with some caution advised. However, the model for solar energy yielded disappointing results.

Although the proposed solution did not achieve high accuracy across all RES, it demonstrated real-world applicability. By integrating the models with day-ahead weather forecasts, the solution can effectively support operational grid decision-making and contribute to more efficient renewable energy management.

To conclude, this project successfully addressed the intermediate objectives and ultimately answered the research question that guided this study by presenting a final hybrid solution for forecasting RES in Madeira Island. Importantly, the solution included not only the selection of the models, but also the comprehensive data preparation and transformation processes. The application of various techniques, such as sine and cosine transformations, imputation of missing values, square root transformation, feature engineering and feature selection, played a critical role in significantly improving the overall model performance. These methods are both relevant and transferable to future forecasting efforts. When combined with the limitations and recommendations identified in this study, they offer a strong foundation for improving model performance in subsequent renewable energy research.

## 5.2 LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

Even though autoregressive models were generally not favoured in most of the literature reviewed, they could have been included for performance comparison. However, due to time constraints and the level of manual intervention required, they were not implemented in this project.

The LSTM model used in this research underperformed compared to results reported in the literature, where LSTM often demonstrated strong forecasting capabilities. This discrepancy was likely due to the relatively small dataset, given that it was aggregated at a daily level. To improve results, future work could consider extending the time span of the dataset or using higher-frequency data. Additionally, deep learning models like LSTM are highly sensitive to their architecture, which was not extensively optimized in this study.

Several studies highlighted the strong performance of ensemble learning methods. However, due to time and computational limitations, this approach could not be explored in the current research.

Another important limitation was the amount of available data. The literature commonly recommended using at least 70% of the data for training. This constraint made it unfeasible

to reserve one full year for validation and another for testing, which is an important aspect when attempting to capture the cyclical and seasonal patterns in renewable energy generation.

The energy production data used in this research was aggregated, which made it impossible to precisely locate production levels with the specific weather conditions at the time of generation. Additionally, the merged weather data was limited to only two weather stations, which does not fully capture the variability of weather conditions across the entire island, limiting the generalization of the findings. Furthermore, the study relied on observed weather data, but testing the models with actual forecasted weather data would have provided a more realistic assessment of their operational performance.

This study did not remove outliers, as they constituted a low percentage of the data. However, their exclusion could be considered in future research.

Finally, to enhance the forecasting accuracy for each energy source, incorporating additional operational scheduling data and technical specifications of each production unit would provide a more realistic representation, particularly for biomass, where such details are crucial. For solar energy, adopting shorter forecasting horizons, such as 12-hour or hour-ahead predictions, could significantly improve model performance by capturing rapid fluctuations more effectively.

# BIBLIOGRAPHICAL REFERENCES


Adhikari, R., & Agrawal, R. K. (2013). *An Introductory Study on Time Series Modeling and Forecasting*. https://doi.org/10.48550/arXiv.1302.6613

Almonacid-Olleros, G., Almonacid, G., Fernandez-Carrasco, J. I., Espinilla-Estevez, M., & Medina-Quero, J. (2020). A New Architecture Based on IoT and Machine Learning Paradigms in Photovoltaic Systems to Nowcast Output Energy. *SENSORS*, *20*(15). https://doi.org/10.3390/s20154224

APREN. (2019). *APREN - Produção*. https://www.apren.pt/pt/energias-renovaveis/producao

Bartlett, M. S. (1936). The Square Root Transformation in Analysis of Variance. In *Source: Supplement to the Journal of the Royal Statistical Society* (Vol. 3, Issue 1). https://www.jstor.org/stable/2983678

Berkeley Earth. (2024). *Annual Temperature Anomaly*. https://berkeleyearth.org/global-temperature-report-for-2023/

Bi, Z., Guo, R., & Khan, R. (2024). Renewable Adoption, Energy Reliance, and CO2 Emissions: A Comparison of Developed and Developing Economies. *Energies* , *17*(13). https://doi.org/10.3390/en17133111

Bochenek, B., Jurasz, J., Jaczewski, A., Stachura, G., Sekula, P., Strzyzewski, T., Wdowikowski, M., & Figurski, M. (2021). Day-Ahead Wind Power Forecasting in Poland Based on Numerical Weather Prediction. *ENERGIES*, *14*(8). https://doi.org/10.3390/en14082164

Box, G., Jenkins, G., Reinsel, G., & Ljung, G. (2016). *Time Series Analysis* (5th ed.). Wiley.

Cabello-López, T., Carranza-García, M., Riquelme, J. C., & García-Gutiérrez, J. (2023). Forecasting solar energy production in Spain: A comparison of univariate and multivariate models at the national level. *APPLIED ENERGY*, *350*. https://doi.org/10.1016/j.apenergy.2023.121645

Campos, F. D., Sousa, T. C., & Barbosa, R. S. (2024). Short-Term Forecast of Photovoltaic Solar Energy Production Using LSTM. *ENERGIES*, *17*(11). https://doi.org/10.3390/en17112582

Castillo-Rojas, W., & Salinas, J. P. (2024). Forecasting Models Applied in Solar Photovoltaic and Wind Energy: A Systematic Mapping Study. *IEEE Access*, *12*, 151092–151111. https://doi.org/10.1109/ACCESS.2024.3471073

Cawley, G. C., & Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. In *Journal of Machine Learning Research* (Vol. 11).

*Centraal Bureau voor de Statistiek*. (2025). https://www.cbs.nl/nl-nl/maatschappij/natuur-en-milieu/indexen-en-trends--trim--/trim-frequently-asked-questions/which-proportion-of-missing-values-in-the-data-is-allowed-

Chen, C., Pinar, M., & Stengos, T. (2021). Determinants of renewable energy consumption: Importance of democratic institutions. *Renewable Energy*. https://doi.org/10.1016/j.renene.2021.07.030

Deshmukh, M. K. G., Sameeroddin, M., Abdul, D., & Abdul Sattar, M. (2023). Renewable energy in the 21st century: A review. *Materials Today: Proceedings*, *80*, 1756–1759. https://doi.org/10.1016/j.matpr.2021.05.501

Dralus, G., Mazur, D., Kusznier, J., & Dralus, J. (2023). Application of Artificial Intelligence Algorithms in Multilayer Perceptron and Elman Networks to Predict Photovoltaic Power Plant Generation. *ENERGIES*, *16*(18). https://doi.org/10.3390/en16186697

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00516-9

*Empresa de Eletricidade da Madeira*. (2024). https://www.eem.pt/pt/conteudo/quem-somos/

Environment, U. N. (2017). *GOAL 7: Affordable and clean energy*. https://www.unep.org/topics/sustainable-development-goals/why-do-sustainable-development-goals-matter/goal-7-affordable

*European Green Deal*. (2019). https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_pt

Fara, L., Diaconu, A., Craciunescu, D., & Fara, S. (2021). Forecasting of Energy Production for Photovoltaic Systems Based on ARIMA and ANN Advanced Models. *INTERNATIONAL JOURNAL OF PHOTOENERGY*, *2021*. https://doi.org/10.1155/2021/6777488

Frisken, W. R. (1971). Extended industrial revolution and climate change. *Eos, Transactions American Geophysical Union*. https://doi.org/10.1029/EO052i007p00500

Gómez, J. L., Martínez, A. O., Pastoriza, F. T., Garrido, L. F., Alvarez, E. G., & García, J. A. O. (2020). Photovoltaic Power Prediction Using Artificial Neural Networks and Numerical Weather Data. *SUSTAINABILITY*, *12*(24). https://doi.org/10.3390/su122410295

González Ordiano, J., Waczowicz, S., Hagenmeyer, V., & Mikut, R. (2018). Energy forecasting tools and services. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (Vol. 8, Issue 2). Wiley-Blackwell. https://doi.org/10.1002/widm.1235

Goutte, S., Klotzner, K., Le, H. V, & von Mettenheim, H. J. (2024). Forecasting photovoltaic production with neural networks and weather features. *ENERGY ECONOMICS*, *139*. https://doi.org/10.1016/j.eneco.2024.107884

Hancock, J. T., & Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00305-w

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., & Hussain, A. (2024). Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. In *Cognitive Computation* (Vol. 16, Issue 1, pp. 45–74). Springer. https://doi.org/10.1007/s12559-023-10179-8

*Joblib*. (2021). https://joblib.readthedocs.io/en/stable/index.html

José Martins Silva, É. DA. (2014). *INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA Aumento de Energia Renovável na Ilha da Madeira-Proposta Integrada para o Setor dos Transportes e da Geração de Energia Elétrica*.

Kamalov, F., Sulieman, H., Alzaatreh, A., Emarly, M., Chamlal, H., & Safaraliev, M. (2025). Mathematical Methods in Feature Selection: A Review. In *Mathematics* (Vol. 13, Issue 6). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/math13060996

Khan, A. A., Cheema, S. A., Hussain, Z., & Abdel-Salam, G. A. (2021). Measuring skewness: We do not assume much. *Scientia Iranica*, *28*(6E), 3525–3537. https://doi.org/10.24200/sci.2020.52306.2649

Khan, W., Walker, S., & Zeiler, W. (2022). Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach. *ENERGY*, *240*. https://doi.org/10.1016/j.energy.2021.122812

Knopf, J. W. (2006). *Doing a Literature Review*. *39*(1), 127–132. https://doi.org/10.1017/S1049096506060264

Koeva, D., Kutkarska, R., & Zinoviev, V. (2023). High Penetration of Renewable Energy Sources and Power Market Formation for Countries in Energy Transition: Assessment via Price Analysis and Energy Forecasting. *ENERGIES*, *16*(23). https://doi.org/10.3390/en16237788

Krechowicz, A., Krechowicz, M., & Poczeta, K. (2022). Machine Learning Approaches to Predict Electricity Production from Renewable Energy Sources. In *Energies* (Vol. 15, Issue 23). MDPI. https://doi.org/10.3390/en15239146

Lemoine, D. (2018). *ESTIMATING THE CONSEQUENCES OF CLIMATE CHANGE FROM VARIATION IN WEATHER* (NBER Working Paper Series). https://www.nber.org/papers/w25008

Maciejewski, D., Mudryk, K., & Sporysz, M. (2024). Forecasting Electricity Production in a Small Hydropower Plant (SHP) Using Artificial Intelligence (AI). *ENERGIES*, *17*(24). https://doi.org/10.3390/en17246401

Matrenin, P. V, Sh Atabaeva, L., & Sergeev, N. N. (2022). *Limitations and Perspectives of Short-Term Renewable Energy Generation Forecasting Methods*. https://doi.org/10.1109/SIBIRCON56155.2022.10017051

Midway, S. R. (2020). Principles of Effective Data Visualization. In *Patterns* (Vol. 1, Issue 9). Cell Press. https://doi.org/10.1016/j.patter.2020.100141

Mytton, D., & Ashtine, M. (2022). Sources of data center energy estimates: A comprehensive review. In *Joule* (Vol. 6, Issue 9, pp. 2032–2056). Cell Press. https://doi.org/10.1016/j.joule.2022.07.011

Ogliari, E., Nespoli, A., Mussetta, M., Pretto, S., Zimbardo, A., Bonfanti, N., & Aufiero, M. (2020). A Hybrid Method for the Run-Of-The-River Hydroelectric Power Plant Energy Forecast: HYPE Hydrological Model and Neural Network. *FORECASTING*, *2*(4), 410–428. https://doi.org/10.3390/forecast2040022

Our World in Data. (2024). *Global fossil fuel consumption*. https://ourworldindata.org/fossil-fuels

*Our World in Data- Renewable energy*. (2024). https://ourworldindata.org/renewable-energy

Perinetti, G. (2019). StaTips Part VII: Anatomy of a Boxplot. *South European Journal of Orthodontics and Dentofacial Research*, *6*(2). https://doi.org/10.5937/sejodr6-23903

*PK Energy*. (2024). https://pknergypower.com/difference-between-mw-and-mwh/?utm_source=chatgpt.com

Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S., & Stodtmann, S. (2024). Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and Translational Science*, *17*(11). https://doi.org/10.1111/cts.70056

Shabbir, N., Kütt, L., Jawad, M., Husev, O., Rehman, A. U., Gardezi, A. A., Shafiq, M., & Choi, J. G. (2022). Short-Term Wind Energy Forecasting Using Deep Learning-Based Predictive Analytics. *CMC-COMPUTERS MATERIALS & CONTINUA*, *72*(1), 1017–1033. https://doi.org/10.32604/cmc.2022.024576

Shahzad, U. (2015). ITEE Journal The Need For Renewable Energy Sources. *Information Technology & Electrical Engineering*, *4*, 16–19. http://www.iteejournal.org/archive/vol4no4/v4n4_4.pdf

*Shap documentation*. (2018). https://shap.readthedocs.io/en/latest/generated/shap.KernelExplainer.html

Shering, T., Alonso, E., & Apostolopoulou, D. (2024). Investigation of Load, Solar and Wind Generation as Target Variables in LSTM Time Series Forecasting, Using Exogenous Weather Variables. *ENERGIES*, *17*(8). https://doi.org/10.3390/en17081827

Su, Y., Cui, C., & Qu, H. (2022). Self-Attentive Moving Average for Time Series Prediction. *Applied Sciences (Switzerland)*, *12*(7). https://doi.org/10.3390/app12073602

Surakhi, O., Zaidan, M. A., Fung, P. L., Motlagh, N. H., Serhan, S., Alkhanafseh, M., Ghoniem, R. M., & Hussein, T. (2021). Time-lag selection for time-series forecasting using neural network and heuristic algorithm. *Electronics (Switzerland)*, *10*(20). https://doi.org/10.3390/electronics10202518

Sweeney, C., Bessa, R. J., Browell, J., & Pinson, P. (2020). The future of forecasting for renewable energy. In *Wiley Interdisciplinary Reviews: Energy and Environment* (Vol. 9, Issue 2). John Wiley and Sons Ltd. https://doi.org/10.1002/wene.365

Szostek, K., Mazur, D., Dralus, G., & Kusznier, J. (2024). Analysis of the Effectiveness of ARIMA, SARIMA, and SVR Models in Time Series Forecasting: A Case Study of Wind Farm Energy Production. *ENERGIES*, *17*(19). https://doi.org/10.3390/en17194803

Tang, F., & Ishwaran, H. (2017). *Random Forest Missing Data Algorithms*. https://doi.org/10.48550/arXiv.1701.05305

Teixeira, R., Cerveira, A., Pires, E. J. S., & Baptista, J. (2024). Advancing Renewable Energy Forecasting: A Comprehensive Review of Renewable Energy Forecasting Methods. In *Energies* (Vol. 17, Issue 14). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/en17143480

*United Nations- Climate Action*. (2024). https://www.un.org/en/climatechange/what-is-renewable-energy

*University of Illinois Chicago*. (2024, September 20). https://ask.library.uic.edu/faq/345899

Velicer, W. F., & Fava, J. L. (2003). Time Series Analysis. In *Handbook of Psychology* (pp. 581–606). Wiley. https://doi.org/10.1002/0471264385.wei0223

*Visual Crossing API weather data documentation*. (2023, March 23). https://www.visualcrossing.com/resources/documentation/weather-data/weather-data-documentation/

*VOSviewer*. (2024). https://www.vosviewer.com/

Wang, H., Liang, Q., Hancock, J. T., & Khoshgoftaar, T. M. (2024). Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. *Journal of Big Data*, *11*(1). https://doi.org/10.1186/s40537-024-00905-w

Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining*.

Zhang, G. P., & Kline, D. M. (2007). Quarterly Time-Series Forecasting With Neural Networks. *IEEE Transactions on Neural Networks*, *18*(6), 1800–1814. https://doi.org/10.1109/TNN.2007.896859

# APPENDIX A

The following table provides more detailed explanation concerning the metadata of all features received from EEM, namely a description of the data, unit and the range of values

Table A.1 - Metadata of features provided by EEM

| Variable | Description | Unit | Values range |
|---|---|---|---|
| Date | Day of the row, in the format DD/MM/YYYY | date | [01/01/2021; 31/08/2024] |
| Hour | Energy production recorded every 15 minutes, in the format DD/MM/YYYY HH:mm | date | [01/01/2021 00:00; 31/08/2024 23:00] |
| Fossil Fuel | Fossil fuel energy production recorded at 15-minute intervals | MWh | [ 0 ; 117.4 ] |
| Natural Gas | Natural Gas energy production recorded at 15-minute intervals | MWh | [ 0 ; 49.6 ] |
| Hydropower | Hydropower energy production recorded at 15-minute intervals | MWh | [ 0 ; 59 ] |
| Biomass | Biomass energy production recorded at 15-minute intervals | MWh | [ 0 ; 7.25 ] |

| | | | |
|---|---|---|---|
| Wind | Wind energy production recorded at 15-minute intervals | MWh | [ 0 ; 58.8 ] |
| Solar | Solar energy production recorded at 15-minute intervals | MWh | [ 0 ; 15.1] |
| Fossil Energy Capacity | Installed fossil fuel energy capacity in Madeira Island from 2021 to 2025 | MW | [ 149.8 ] |
| Natural Gas Energy Capacity | Installed natural gas energy capacity in Madeira Island from 2021 to 2025 | MW | [ 56 ] |
| Hydropower Energy Capacity | Installed Hydropower energy capacity in Madeira Island from 2021 to 2025 | MW | [ 68.3 ; 80.4 ] |
| Biomass Energy Capacity | Installed Biomass energy capacity in Madeira Island from 2021 to 2025 | MW | [ 8 ] |
| Wind Energy Capacity | Installed wind energy capacity in Madeira Island from 2021 to 2025 | MW | [ 63.1 ] |
| Solar Energy Capacity | Installed solar energy capacity in | MW | [ 15 ] |

| | |
|---|---|
| | Madeira Island from 2021 to 2025 |

# APPENDIX B

The following tables provide more detailed explanations concerning the metadata of the weather variables retrieved from Visual Crossing Weather API, which were used for this project, namely a description of the data, unit and the range of values.

Table B.1- FW_Funchal weather variables' metadata

| Variable | Description | Unit | Values range |
|---|---|---|---|
| date_time | The timestamp, in an hourly format, in the format DD/MM/YYYY HH:mm | date | [01/01/2021 00:00; 31/08/2024 23:00] |
| temp | Temperature recorded at the corresponding weather station, | °C | [ 10.1 ; 36.5 ] |
| feelslike | Temperature felt at the corresponding weather station, | °C | [ 10.1 ; 34.6 ] |
| humidity | Relative Humidity recorded at the corresponding weather station, | % | [ 13.1 ; 99.7 ] |
| dew | Measured temperature at which air is saturated with water vapor and condensation occurs | °C | [ -4.9 ; 23 ] |
| precip | Precipitation recorded at the corresponding weather station | mm | [ 0; 113.95 ] |

| | | | |
|---|---|---|---|
| precipprob | Precipitation chance recorded | % | 0 or 100 |
| windgust | Brief or sudden increase in wind speed | km/h | [ 4.7 ; 122] |
| windspeed | Rate at which air moves from one place to another | km/h | [ 0; 56.6 ] |
| winddir | Installed Hydropower energy capacity in Madeira Island from 2021 to 2025 | degrees | [ 0 ; 359 ] |
| pressure | Sea Level Pressure | mb | [ 997 ; 1036 ] |
| visibility | Maximum distance at which objects can be clearly seen in daylight | km | [ 1.9 ; 21.4 ] |
| cloudcover | Portion of the sky covered by clouds | km | [ 0; 100 ] |
| solarradiation | Power of sunlight received | $W/m^2$ | [ 0 ; 1046 ] |
| solarenergy | Measures the amount of solar radiation | $MJ/ m^2$ | [ 0 ; 3.8 ] |

| | | | |
|---|---|---|---|
| | received over a 1 square meter area | | |
| uvindex | The UV index | Numerical indicator | [ 0 ; 10 ] |

Table B.2 - LPMA weather variables' metadata

| Variable | Description | Unit | Values range |
|---|---|---|---|
| date_time | The timestamp, in an hourly format, in the format DD/MM/YYYY HH:mm | date | [01/01/2021 00:00; 31/08/2024 23:00] |
| temp | Temperature recorded at the corresponding weather station, | °C | [ 10 ; 36 ] |
| feelslike | Temperature felt at the corresponding weather station, | °C | [ 6.6 ; 34.7 ] |
| humidity | Relative Humidity recorded at the corresponding weather station, | % | [ 6.8 ; 100 ] |
| dew | Measured temperature at which air is saturated with | °C | [ -13 ; 24 ] |

| | | | |
|---|---|---|---|
| | water vapor and condensation occurs | | |
| precip | Precipitation recorded at the corresponding weather station | mm | [ 0; 106 ] |
| precipprob | Precipitation chance recorded | % | 0 or 100 |
| windgust | Brief or sudden increase in wind speed | km/h | [ 4.3 ; 108 ] |
| windspeed | Rate at which air moves from one place to another | km/h | [ 0; 61.2 ] |
| winddir | Installed Hydropower energy capacity in Madeira Island from 2021 to 2025 | degrees | [ 0 ; 359 ] |
| pressure | Sea Level Pressure | mb | [ 997 ; 1037 ] |
| visibility | Maximum distance at which objects can be clearly seen in daylight | km | [ 1 ; 19.5 ] |
| cloudcover | Portion of the sky covered by clouds | km | [ 0; 100 ] |

| | | | |
|---|---|---|---|
| solarradiation | Power of sunlight received | $W/m^2$ | [ 0 ; 1046 ] |
| solarenergy | Measures the amount of solar radiation received over a 1 square meter area | $MJ/ m^2$ | [ 0 ; 3.8 ] |
| uvindex | The UV index | Numerical indicator | [ 0 ; 10 ] |

# APPENDIX C

The boxplots for the FW_Funchal weather station revealed key patterns in the distribution and variability of its weather features. Temperature-related variables, including temp and feelslike, displayed relatively symmetrical distributions with moderate spread, suggesting stable and consistent conditions. Dew and humidity also followed a compact distribution, though slightly skewed toward lower values.

In contrast, variables such as precip, windgust, solarradiation, and solarenergy showed clear right skewness, with long upper whiskers and medians positioned toward the bottom of the boxes. These patterns suggested that while most values were low, occasional spikes occurred, reflecting intermittent but intense events like heavy rainfall, strong gusts, or bursts of solar intensity.

Precip stood out with a dense cluster of outliers, capturing extreme precipitation events. Windgust and windspeed also exhibited substantial variability, especially windgust, with both having several upper outliers. Visibility and uvindex were generally well-distributed, although uvindex included a few high-end outliers, possibly corresponding to episodes of extreme UV exposure. Pressure and cloudcover showed relatively tight and symmetrical distributions. Overall, these boxplots provided a clear view of both the typical weather conditions and the occasional extremes experienced at FW_Funchal.

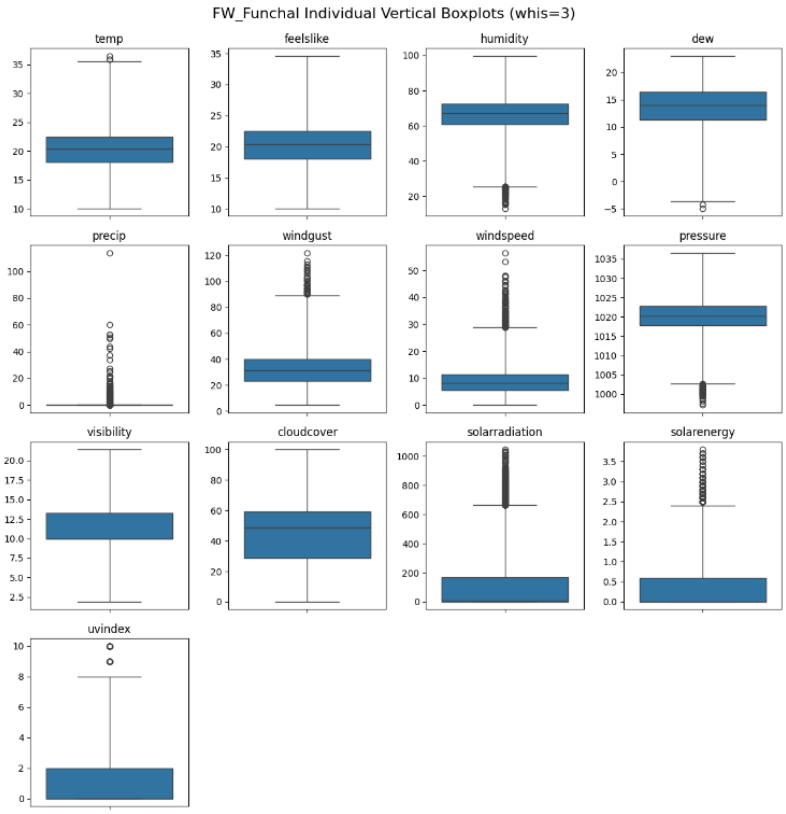


Figure C.1 – FW_Funchal boxplots

The boxplots for the LPMA weather station revealed distribution patterns that in many ways resembled those observed at FW_Funchal. Temperature-related features, including temp and feelslike showed symmetrical distributions with moderate spread, suggesting stable thermal conditions, like the trends seen in FW_Funchal. Dew and humidity were also consistently distributed, though slightly more concentrated toward higher values. As with FW_Funchal, variables such as precip, windgust, solarradiation, and solarenergy exhibited clear right skewness, with long upper whiskers and lower-positioned medians, pointing to occasional but pronounced spikes in rainfall, wind gusts, and solar input.

Outliers were notably present in precip and visibility, again resembling patterns from FW_Funchal, where extreme precipitation and low-visibility episodes occurred sporadically. Windgust continued to show a wider spread compared to windspeed, with many upper outliers, indicating frequent gust intensity fluctuation. Solar-related variables, solar radiation and solarenergy, presented similar outlier concentrations and skewness, reflecting shared regional exposure to variable sunlight. Pressure and cloudcover remained tightly distributed and nearly symmetrical, and uvindex, while mostly low, showed a few elevated values, as observed in FW_Funchal. These consistencies suggested that both stations, while distinct were influenced by similar climatic dynamics, particularly regarding extremes in windgust, precipitation, and solar exposure.

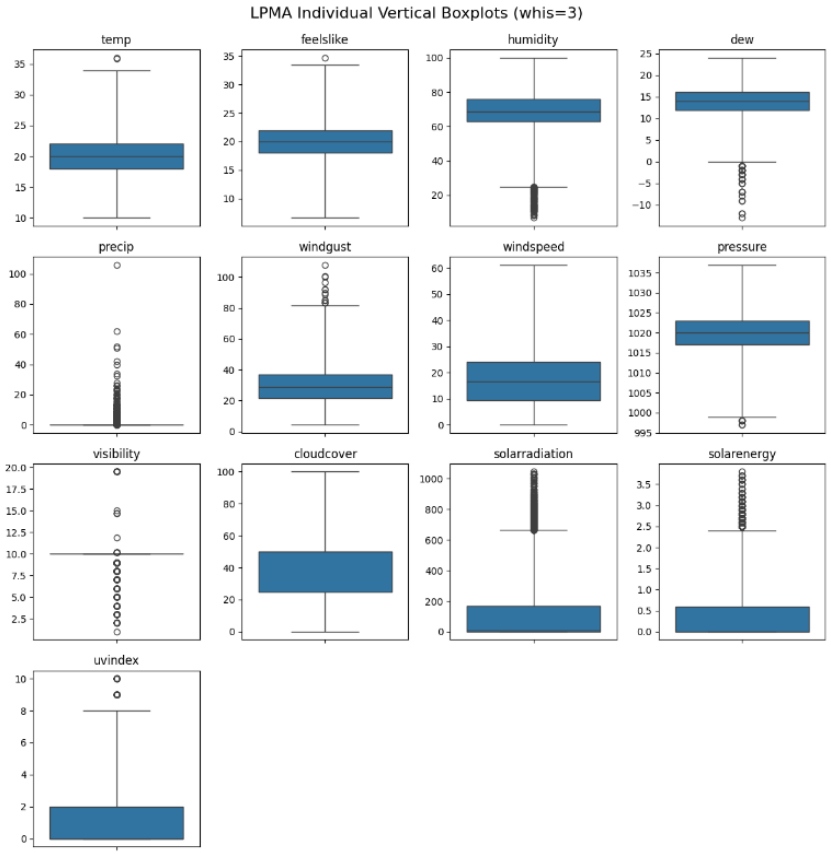


Figure C.2 - LPMA boxplots


**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**

Universidade Nova de Lisboa