POLITECNICO DI TORINO

MASTER's Degree in DATA SCIENCE and ENGINEERING

Politecnico di Torino

MASTER's Degree Thesis

Machine Learning to Predict Energy Production

Supervisor

Prof. Michela MEO

Assistant Supervisor

Ph.D. Greta VALLERO

Candidate

Parsa TAATI

March 2026

Abstract

This research presents a complete machine learning approach for photovoltaic (PV) energy prediction utilizing operational data from various Politecnico di Torino installations. Renewable energy integration requires accurate PV estimates for grid stability, energy management, and cost optimization.

Data analysis, robust preprocessing using feature engineering, statistical pattern recognition, and advanced machine learning model creation are used to analyze five 2014-2024 datasets. Testing linear regression, random forest, and XGBoost with systematic hyperparameter optimization in single-site, multi-site, and directed east-west configurations.

Results show that advanced machine learning approaches accurately forecast real-world PV installations ($R^2 = 94 - 96\%$). Random Forest did well, while XGBoost excelled in extremes. Effective feature engineering cut useless functions 25-42%. All installations were best predicted by day, season, and climate (solar irradiance, temperature).

Single-site systems had the maximum accuracy ($R^2 > 0.94$) due to controlled settings, while multi-site setups needed customized optimization. Changing hyperparameters increased R^2 by 1-3% in complex multi-site scenarios.

Powerful machine learning algorithms and intelligent feature selection meet energy system accuracy and efficiency standards. This study verifies PV predictions, grid integration, installation-specific optimization, and energy management. Economic analysis demonstrates significant cost savings potential, with the studied installations generating €5.4 million in electricity cost reductions (23.05 GW Production) through optimized energy production forecasting. the analysis identified the `tot_pv_aule_p` system as the most efficient, achieving the highest maximum efficiency of 68.13% despite having a moderate capacity of only 50 KWP.

Keywords: Energy forecasting, Machine learning, Photovoltaic prediction, RandomForest, Renewable energy, XGBoost

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to Professor Michela Mao for her invaluable guidance, support, and encouragement throughout the course of this thesis. Her insights and expertise have been instrumental to my academic growth. I am also sincerely thankful to her assistant, Greta Vallero, for her continuous support and helpful feedback during this journey.

I am deeply grateful to my parents, whose unconditional love, patience, and belief in me have been the foundation of all my achievements. Their constant support and sacrifices have made this journey possible.

Lastly, I would like to thank my friends who stood by me through the challenges, shared in my successes, and made this experience all the more meaningful. Your companionship has been a true source of strength and motivation.

Table of Contents

List of Figures

List of Tables

Acronyms

AEEG	Autorità per l'Energia Elettrica e il Gas (Italian Electricity and Gas Authority).
GW	Gigawatts.
IoT	Internet of Things.
IQR	Interquartile Range.
KW	Kilowatt.
KWP	Kilowatt Peak.
LSTM	Long Short-Term Memory.
MAE	Mean Absolute Error.
ML	Machine Learning.
MSE	Mean Square Error.
PV	Photovoltaic.
R^2	Coefficient of Determination (R-squared).
RMSE	Root Mean Square Error.

Chapter 1

Introduction

1.1 Background and Motivation

Global efforts to mitigate climate change and achieve sustainable development objectives are causing a sea change in the energy sector. The crucial importance of photovoltaic (PV) systems was demonstrated by the meteoric rise in solar capacity, which went from 40 GW in 2010 to over 1,000 GW in 2022. This is one of the most rapidly growing energy technology areas.

Extensive solar integration is quite problematic due to the inherent unpredictability and spottiness of sun irradiation. Power from photovoltaic cells is very sensitive to local atmospheric variables including weather, time of day, season, and others, in contrast to other types of dispatchable power sources. Because they depend on accurate output predictions, grid operators, energy dealers, and system planners face difficulties in energy management as a result of this variation.

For a grid that is secure, financially successful, and ecologically friendly, reliable PV prediction is essential. Systems become unstable, fossil fuels are used more often as a backup, energy trading fails to provide, and solar investments yield worse returns due to inaccurate estimates. Accurate forecasting, however, makes possible optimization of energy dispatch, stability of the grid, integration of storage, and the effectiveness of demand response programs.

The use of physical models and numerical weather prediction algorithms has long been the backbone of solar forecasting methods. Though these approaches do give useful physics-based data, they have a hard time taking into consideration things like equipment-specific characteristics and local microclimatic influences, not to mention the complex, non-linear relationships between weather variables and PV production at unique installations.

1.2 Problem Statement

Predicting the output power of solar panels using real-world operational data in a reliable, efficient, and accurate manner is the major objective of this thesis. A multitude of significant concerns exist, encompassing:

- **Data Quality Issues:** Real-world PV data contains missing values, measurement errors, temporal misalignments, and quality inconsistencies that significantly impact prediction accuracy
- **Feature Engineering Complexity:** Identifying and engineering predictive features while eliminating redundant variables that degrade performance and increase computational complexity
- **Model Selection Optimization:** Systematically evaluating and optimizing machine learning approaches for different PV prediction scenarios and installation types
- **Scalability and Generalizability:** Developing models that scale across different installations, locations, and equipment configurations while maintaining accuracy and efficiency

1.3 Research Objectives

Using real-world operational data from different installation configurations, the major purpose is to develop and validate a complete machine learning framework for precise solar energy forecast.

Other Objectives:

- Implement robust data preprocessing methodologies for PV operational data quality enhancement
- Develop systematic feature engineering and selection approaches for PV energy forecasting
- Compare machine learning algorithms (linear regression, RandomForest, XGBoost) across prediction scenarios
- Apply hyperparameter optimization to maximize performance while ensuring generalizability
- Evaluate performance across diverse installation types to identify optimization strategies
- Provide practical implementation guidelines for operational energy management systems

1.4 Research Methodology

This research employs a systematic four-phase approach:

Phase 1- Exploratory Data Analysis: Comprehensive data characterization, quality assessment, missing value analysis, and initial pattern identification across multiple PV datasets.

Phase 2 Data Preprocessing and Feature Engineering: Robust data cleaning, missing value imputation, temporal alignment, and systematic feature creation including temporal, meteorological, and derived variables.

Phase 3- Statistical Analysis: Statistical characterization of PV patterns, correlation analysis, temporal trend identification, and baseline predictive modeling.

Phase 4 - Advanced Machine Learning: Development and optimization of sophisticated ML models including ensemble methods, gradient boosting, hyperparameter tuning, and comprehensive performance evaluation.

1.5 Scope and Limitations

Scope: Short-term PV prediction (15-minute to hourly forecasts) using operational data from five diverse installations at Politecnico di Torino, Italy, spanning 2014-2024.
Limitations:

- Geographic constraints (Central Italy) may limit generalizability to different climates
- Temporal scope may not capture all long-term climatic variations
- Installation diversity may not represent all possible PV configurations
- Focus on short-term forecasting limits long-term planning applications

1.6 Expected Contributions

Methodological: A thorough framework for PV prediction that takes into account real-world data, feature engineering methods that are systematic, and ML algorithms that are compared for different kinds of installations.
Practical: Guidelines for developing and deploying ML-based PV prediction systems, techniques for fine-tuning the system for specific installations, and recommendations for enhancements to achieve peak performance.
Research: A verification using a substantial amount of real-world operational data, an investigation into the relationships between setup and simulation, and a quantitative evaluation of the accuracy of ML predictions.

1.7 Thesis Organization

Chapter 2: Literature review and theoretical background of PV forecasting methods and ML applications in renewable energy.
Chapter 3: Data collection, quality assessment, preprocessing methodologies, and feature engineering approaches.
Chapter 4: ML model development, implementation across four research phases, comparative performance analysis, optimization results and Production Analysis, Economic Impact and PV systems capacities analysis.
Chapter 5: Summary of findings, practical implications, limitations, and future research recommendations.

This organization ensures logical progression from theoretical foundations through methodology development to practical results, supporting reproducibility while highlighting operational applicability for PV energy management systems.

Chapter 2

Literature Review and Related Work

2.1 Introduction

Methods, feature engineering approaches, and evaluation frameworks pertinent to the multi-dataset, weather-integrated prediction system upon which this thesis is based are examined in this chapter, which reviews the literature on machine learning for the purpose of predicting photovoltaic energy production. Using high-resolution temporal data from 2014 to 2025, the review establishes the framework for figuring out how to employ machine learning approaches to reliably anticipate PV production across many installations in Politecnico di Torino (Italy). This study examines recent research trends, methodology, and performance criteria in the renewable energy forecasting field to place the present work in perspective. Regulatory features, multi-site analysis, and thorough temporal modeling are highlighted as new contributions.

2.2 Photovoltaic Energy Forecasting: Foundations and Challenges

2.2.1 Time Series Characteristics of PV Production

Predicting future solar energy production is challenging due to the large number of variables, both environmental and temporal, that affect it. The intermittent and unexpected solar energy patterns have been the subject of numerous research [1],[2]. This impacts the methodologies employed in this thesis in a direct manner. The most significant issue is that atmospheric conditions and solar radiation follow regular astronomical cycles, which can significantly impact energy production. Photovoltaic systems are powered by sunlight.

At all times, the weather, particularly the level of cloud cover and air clarity, alters the fundamental pattern that governs the day-to-night variation in PV system output. Cloud cover levels can fluctuate dramatically in a matter of minutes, significantly impacting PV production. According to inman(2013)[3], this shift is most noticeable

when clouds aren't completely covering the sky. Since covering short-term changes is crucial for grid management and energy trading, this thesis employs data gathered every 15 minutes due to this level of time accuracy.
From one year to the next, the sun's angular position and the day's duration vary. Power from photovoltaic cells follows seasonal trends when these two events occur together. Evidence from Pedro and Coimbra[4] indicates that accurately predicting these seasonal effects is crucial. By meticulously including time-related factors such as day length, quarter, and day of the year, the feature engineering method utilized in this thesis takes care of these patterns. The model can determine that seasonal variations occur at regular intervals when all of these factors are combined.
Standard statistical methods might not be up to the task of studying the complex, non-linear relationships between weather and PV output. Based on established physical mechanisms[5]. anticipated that temperature would impact panel efficiency and conducted a comprehensive examination of these relationships. However, cloud cover alters the quantity of sunlight that reaches the ground in intricate ways based on the type of cloud, its thickness, and its speed of movement. Precipitation, cloud cover levels, wind speeds at 10 meters, temperature at 2 meters, and other meteorological data are incorporated into this thesis. Finding a way to help machine learning models understand these intricate relationships is the main objective.

2.2.2 Multi-Site Forecasting Challenges

Up until this point, the vast bulk of research on PV forecasting has focused on creating predictions for a particular location, which restricts the models' potential to be applied in real-world situations and to be generalized. According to Rodríguez-Benítez et al.(2020)[6], single-site studies have the potential to provide valuable insights into forecasting approaches. However, these studies fail to address the issues that occur when considering the implementation of these systems in multiple places that possess distinct characteristics. The significance of this constraint becomes apparent when energy providers or grid operators are confronted with the issue of calculating the output from many photovoltaic (PV) systems.
This thesis investigates five various installation types, each of which has its own individual set of characteristics that influence production and forecasting requirements. The investigation is carried out through the use of a multi-site technique. The patterns of use that can be observed in educational facilities at universities are related with the academic scheduling as well as seasonal fluctuations in the activities that take place on campus. Different production profiles are produced by different panel orientations and shading effects, which necessitates the use of varied modeling methodologies. It is clear that this is the case because research infrastructure installations, such as the I3P facilities, have separate orientations toward the east and the west. As a result of the fact that inverter configurations at industrial sites differ in terms of electrical quality and maintenance schedules, which in turn have an impact on the manufacturing process, things grow more complicated. The addition of municipal

infrastructure makes things far more difficult because it must be coordinated with the running and maintenance schedules of public facilities.
Because there are so many different kinds of installations, the process of building machine learning models can be both simpler and more difficult overall. According to Raza et al.(2016)[7], increasing the scalability of renewable energy forecasting systems is one of the most significant challenges they face. The fact that they pointed out that models established for one sort of installation could not necessarily be applicable other installations with different features is something that they mentioned. This thesis makes an effort to address the problem by providing a standard modeling framework that can be applied in a methodical manner to a variety of different sorts of installations. Additionally, it maintains the possibility to capture patterns that are distinctive to the installation through the differentiation of multiple target variables.

2.3 Machine Learning Methodologies in Energy Prediction

2.3.1 Comparative Algorithm Performance

Over the past ten years, there has been a substantial amount of progress made in the field of predicting renewable energy using machine learning algorithms. Researchers have finally come to the conclusion that there is no algorithm that is the best possible choice. The purpose of this thesis is to evaluate algorithms in a robust manner by comparing Linear Regression, RandomForest, XGBoost, and Support Vector Regression across a number of different datasets. This thesis follows the best methodologies published in the literature for assessing algorithms.
Despite the fact that it appears to be a straightforward method, linear regression is capable of producing useful results and performing well in applications that involve energy forecasting. When there are numerous linear components in the interactions between the input features and the goal variables, Monteiro et al.(2017)[8], shown that linear techniques, when effectively feature engineered, can achieve unexpectedly good results. This is especially true in situations where there are a lot of linear components. The findings of this thesis provide evidence that there are strong linear correlations between the temporal and climatic parameters that have been methodically created and the PV output. This is evidenced by R-squared values that are greater than 0.9 for several of the targets that were examined using linear regression. A significant achievement of the feature engineering method is the identification of the most important predictive correlations and the representation of those relationships in a format that can be utilized by simple linear models. This discovery demonstrates that this is the case.

When it comes to applications involving energy, Ahmed and Khalid(2019)[9] pointed out that linear regression is more basic and easier to explain than other methods. According to them, the ability to interpret and explain model predic-

tions is not always more significant than the accuracy of forecasts made in the real world. In this thesis, the success of linear regression provides assurance that the relationships being modeled are not based on false correlations that might be caught by more sophisticated models, but rather are well understood and make sense in reality. This is because the correlations could be captured by more complex models.

The versatility of RandomForest, its capacity to manage non-linear interactions, and its capacity to prioritize the significance of natural traits have made it an extremely useful tool for the forecast of renewable energy. RandomForest's utility in energy applications was theoretically proved by Breiman(2001)[10], through the demonstration of ensemble approaches that include several decision trees. These approaches have the potential to provide greater generalization performance while maintaining interpretability through feature importance analysis. The RandomForest algorithm delivers the best outcomes across a wide variety of datasets, with R-squared values ranging from 0.92 to 0.97. This discovery is in line with conclusions drawn from earlier research.

A comprehensive study conducted by Das et al.(2018)[11] found that RandomForest was one of the most effective algorithms for predicting photovoltaic (PV) production across a variety of time periods and for a variety of various types of installations. The efficiency of RandomForest in controlling the complex interaction between time and weather components that form PV output patterns was the primary focus of their work. RandomForest was particularly effective in this regard. In addition to that, its capacity to avoid overfitting was investigated. By employing RandomForest, Rodríguez-Benítez et al.(2020) [6] were able to attain the most favorable outcomes in their prediction of solar radiation across the Iberian Peninsula. This illustrates that the algorithm is able to handle a wide range of weather conditions and variances that occur across different regions.

RandomForest's feature importance features can be utilized for a variety of purposes, not the least of which is to enhance the accuracy of forecasts. The academic community and professionals alike can benefit from their use in establishing whether input variables have a major impact on the accuracy of predictions. With the help of this skill, Sperati et al.(2016)[12] were able to ascertain which feature sets were the most helpful in terms of projecting the generation of renewable energy. Their understanding of the physical mechanisms that are responsible for the creation of energy led them to conclude that the feature importance rankings provided by RandomForest were consistent with what they saw.

The XGBoost algorithm represents the next step in the progression of gradient boosting techniques. A substantial amount of interest in energy forecasting has been generated as a result of its outstanding performance in machine learning contests and its capacity to quickly manage large datasets that are notoriously difficult to manage.

Chen and Guestrin(2016)[13] made a number of particular improvements in order to improve XGBoost's ability to make predictions in a more timely and accurate manner. Consequently, operational forecasting systems that place a high priority on both accuracy and speed will discover that it is a great fit for their needs. The conclusions of this thesis, which indicate that XGBoost and RandomForest perform well when compared to one another, are in line with the findings of other recent research on energy forecasting.

In their study, Wang et al.(2018)[14] revealed that XGBoost works exceptionally well when it comes to PV estimates one day in advance. Compared to both traditional machine learning and the most fundamental neural network techniques, it performed significantly better. Their research focused primarily on the capability of XGBoost to represent non-linear correlations between meteorological data and energy output, as well as complex time-based dependencies. This was the core focus of their investigation. As a continuation of this work, Liu et al.(2020) [15] investigated solar prediction scenarios for the near future that were more analogous to the 15-minute interval forecasting that was applied in this thesis. This demonstrated that XGBoost is effective throughout a wide range of time periods to a significant degree.

They demonstrated the algorithm's adaptability and application to a variety of forecasting scenarios by using XGBoost to predict solar energy consumption several steps in advance. Kim et al.(2019)[16] used this method to demonstrate the algorithm's versatility. They focused their attention on the various methods by which XGBoost can continue to produce accurate forecasts even as the time horizon for weather forecasting increases. Applications that require planning over an extended period of time will find this to be essential. The fact that XGBoost obtains results that are equivalent to those of RandomForest on a variety of installations lends support to the idea that the algorithm is versatile enough to suit a wide range of PV systems and operational styles.

The revolutionary technique to machine learning known as Support Vector Regression has demonstrated its effectiveness in managing high-dimensional feature fields and non-linear connections. This is accomplished through the utilization of kernel approaches. Vapnik (1995) laid the theoretical groundwork for the application of support vector machines (SVR) in forecasting by proving that these machines could discover optimal decision boundaries in high-dimensional feature spaces and achieve great generalization performance. SVR-based solar forecasting has been the subject of a significant amount of research by researchers. The researchers made the discovery that it is particularly effective in managing the intricate and multi-dimensional linkages that exist between meteorological elements and the generation of solar electricity.

The optimized SVR models that Bouzgou and Gueymard(2012)[17]produced were built with the purpose of improving the accuracy of multi-step predictions

of solar irradiation. By making adjustments to the parameters, they were able to demonstrate that these models were capable of competing with more involved approaches. The primary objective of their research was to identify the kernel that has the best performance and to optimize the settings in order to get the greatest possible SVR. Mellit and Kalogirou [18] conducted an in-depth comparison of neural network methods and SVR approaches with regard to projects using photovoltaic (PV) applications. They found that SVR typically performed better and was more stable than other methods, particularly in situations where there was a scarcity of training data.

2.3.2 Feature Engineering Methodologies

In this thesis, an approach to feature engineering is taken that is not only based on the most effective methods for energy forecasting but also adapted to the operational and regulatory environment in Italy. When it comes to energy forecasting, feature engineering is an essential component that must be there for machine learning to be effective. There is a possibility that it is more significant than choosing the most effective algorithm for determining the accuracy of a forecast.
Temporal feature engineering is the foundation upon which effective energy forecasting systems are constructed. This is because the generation of energy is mostly impacted by astronomical and seasonal cycles. The years, months, days, hours, and minutes are the primary variables that are utilized in the full temporal feature set that is presented in this thesis. In this way, models are provided with explicit knowledge of temporal patterns, which they would not have been able to derive from the data on their own. In contrast to relying just on delayed target variables or implicit temporal learning, Voyant et al.(2017)[5] demonstrated that the incorporation of explicit temporal encoding results in a consistent improvement in the accuracy of predictions.

When this base is supplemented with more advanced temporal features, such as the day of the week, the day of the year, the week of the year, and the quarter, models are equipped with knowledge about a variety of cyclical patterns that occur at different time scales. Pedro and Coimbra(2012)[4] revealed the vital necessity of these multi-scale temporal characteristics in order to gain an understanding of the delicate interplay that exists between the numerous periodic components that are involved in the creation of solar energy. On the basis of these findings, Yang et al [1] demonstrated that the use of broad temporal encoding results in an improvement in the accuracy and resilience of models against changes in training data and time.

By utilizing calendar-based characteristics such as holiday classifications and weekend indicators, it is possible to take into consideration the operational and behavioral variables that have an impact on energy consumption and production. López et al.(2018)[19] revealed that energy systems behave significantly differently

on weekends, holidays, and ordinary workdays. This is due to the fact that the number of people in buildings, the times of day that equipment is utilized, and the overall quantity of energy that is required all change during the week. Within the framework of this thesis, the classification system for holidays that is specific to Italy encompasses both public holidays and generally recognized holidays. This provides models with contextual information regarding these patterns of behavior, in contrast to the generic temporal features that are typically considered.

As a result of the fact that the amount of solar energy that can be converted into electricity is directly influenced by the weather, the incorporation of weather variables is an essential component of PV forecasting. The weather variables that were chosen for this thesis addressed all of the significant factors that have an impact on solar irradiance and the performance of PV systems. These weather variables are in agreement with several meteorological forecasting frameworks that are currently in use.

The temperature at two meters is an important predictor since it is known that the efficiency of the panels is associated with the heat and conductivity of semiconductors. For this reason, the temperature at two meters is vital. According to Huld et al.(2011)[20], the goal of building standard models for the influence of temperature on the efficiency of PV panels is to develop these models. They came to the conclusion that the output of the panel normally reduces by 0.4 to 0.5% for every degree Celsius that the temperature of the cells increases. By using daily temperature data, which includes the mean, maximum, and lowest values, models are able to get insights into the impact that temperature patterns, both short-term and longer-term, have on energy production cycles.

At a distance of ten meters, the wind speed has an effect on the efficiency of a photovoltaic (PV) system. This is due to the fact that the panels' capacity to cool and regulate heat is influenced by the wind speed. Schwingshackl et al.(2013)[21] conducted a comprehensive investigation of the impact that wind has on the temperature of photovoltaic modules. They found that panels worked better at greater wind speeds by lowering their operating temperature through improved convective cooling. This allowed them to take advantage of the increased wind speed. In light of the fact that thermal effects are at their most intense when the sun is shining brightly, this connection is essential for generating accurate predictions under such atmospheric conditions.

By doing an analysis of cloud cover data at various altitudes within the atmosphere, one can gain a comprehensive picture of the sky conditions that have an effect on the amount of solar irradiance that is available. In order to improve cloud descriptions for use in solar forecasting, Chow et al.(2011)[22] made improvements. They came to the realization that different clouds and heights would have distinct effects on the patterns of sun irradiance observed. This thesis makes use of data on

low, medium, and high cloud cover in order to provide various models with assistance in discriminating between different types of cloud cover and the impact that they have on the availability of solar energy.

Because precipitation is an indicator of the characteristics of the weather system and the atmospheric conditions that influence both the efficiency and the direct solar irradiation, it is important to pay attention to it. According to Antonanzas et al.(2016)[2], larger weather patterns have the potential to alter sun availability patterns over a period of many days, which can then lead to precipitation occurrences. By looking at the statistics of precipitation, we may have a better understanding of the long-term implications that the weather has on the amount of energy that is produced.

The addition of daylight length as a derived weather feature provides models with helpful information regarding the seasonal fluctuations in solar energy availability, which complements the temporal features. This information is provided by the solar energy availability. According to Ineichen and Perez(2002)[23], the length of daylight is a dependable strategy that can be utilized to change seasonal patterns of energy output. This helps models differentiate between changes in solar irradiance that are caused by the weather and those that are caused by the changing of the seasons.

2.3.3 Regulatory Feature Innovation

The addition of characteristics from the Italian energy market, which is a novel contribution, fills a large vacuum in the existing literature on energy forecasting. This gap was previously present in the literature. When it comes to energy systems, prediction models often disregard market dynamics and regulatory frameworks in favor of temporal and climatic variables. In compliance with the regulations that govern the Italian energy market, the FasciaAEEG categorization system divides each day into three distinct parts: peak (F1), middle (F2), and off-peak (F3). The system is broken down into its component parts in this thesis.

The incorporation of regulatory aspects is based on the theoretical foundation that energy systems operate within intricate regulatory frameworks that have an effect on their operation, the maintenance requirements, and the performance improvement tactics being implemented. According to Weron(2014)[24], it is extremely important to have forecasting models that are aware of the market. According to him, energy markets provide incentive structures that have an effect on system performance in ways that cannot be captured by models that are solely based on technical considerations.

The F1 time band, which indicates the peak demand and energy pricing, typically takes place during the times that businesses are open during the weekdays. When compared to off-peak hours, these times may require different optimization strategies

or maintenance plans for photovoltaic (PV) systems. The F2 transitional periods are characterized by their moderate levels of demand and pricing. In the case of F3, off-peak periods are times when demand is low, such as throughout the night, on weekends, and during holidays.

Including information regarding market structure in energy forecasting models has been found to improve the accuracy of these models, as demonstrated by Lago et al[25]. This is accomplished by giving context for the management of both companies and systems. The results of their investigation indicated that the ways in which energy systems function throughout the year are influenced by seasonal shifts in the economic incentives and operational goals that people have.

As a result of the incorporation of holiday categories that are specific to Italy into this regulatory integration, models now have access to information regarding the times at which routine operations may be disrupted. The findings of a study conducted by Marcjasz and colleagues [26] indicate that European energy markets are susceptible to holiday influences, which can vary substantially in terms of both type and location. The purpose of this thesis is to differentiate between public holidays and general holidays. This distinction is made in recognition of the fact that different kinds of holidays may have varied effects on the operation and efficiency of the energy system.

2.4 Model Evaluation and Performance Assessment

2.4.1 Evaluation Metrics Selection

It is imperative that you make use of the appropriate evaluation criteria in order to acquire a reliable understanding of the effectiveness of a forecasting model and to guarantee that comparisons with other research are grounded in reality. A comprehensive collection of metrics, such as R-squared, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE), are utilized in this thesis in order to investigate the performance of the model. The accuracy and dependability of the forecasts are evaluated using these criteria, which cover a variety of variables.

Examining the coefficient of determination, often known as R squared, is a straightforward method for determining the extent to which the model is able to account for the variance in the variable of interest. As a result of this, it is an excellent tool for evaluating the overall performance of a model as well as comparing the performance of a model with a variety of targets and datasets. Montgomery et al.(2012)[27] are largely responsible for the widespread use of R-squared in the field of regression analysis. As they pointed out, one of its advantages is that it offers a standardised evaluation of the performance of a model that is not reliant on the magnitude of the variable that is being evaluated. The fact that this thesis

consistently achieves R-squared values that are greater than 0.9 for a variety of objectives is evidence of its exceptional predictive ability, which is on par with the most promising findings that have been reported in recent research.

Zhang et al.(2020)[28] found that utilizing sophisticated ensemble techniques, they were able to discover R-squared values ranging from 0.85 to 0.95 for PV forecasting one day in advance. The findings of this thesis are comparable to the most recent research that has been conducted in the field. It is possible to have a better understanding of the performance levels that were seen in this work due to the fact that Das et al.(2018) [11] got equivalent R-squared ranges by employing ensemble methods on various kinds of installations.

The Root Mean Square Error (RMSE) is an essential indicator for energy forecasting applications. This is due to the fact that it focuses on larger prediction errors, which are often the ones that pose the most trouble for grid stability and operational decision-making. In terms of predicting, Chai and Draxler(2014)[29] conducted an in-depth investigation on the RMSE characteristics. The results demonstrated that root mean square error (RMSE) is a useful indicator for determining the reliability of a model and the likelihood of severe errors that could have a detrimental influence on operations. As a result of its ability to zero in on more significant errors, RMSE is particularly well-suited for applications involving energy. In these kinds of circumstances, the stability of the system and the performance of the economy are both greatly influenced by making significant errors in forecasting.

The sensitivity of the root mean square error (RMSE) to outliers and huge mistakes can provide us with a wealth of information regarding the robustness, systemic biases, and failure mechanisms of a model. Energy forecasting algorithms with a low root mean square error (RMSE) work effectively in tough conditions as well as in general situations when other models could yield considerable uncertainties.

Mean absolute error (MAE) is a more robust measure of normal prediction error that is less damaged by extreme and outlier values. It is used in conjunction with root-mean-squared error (RMSE), which yields a more accurate representation of the error. Willmott and Matsuura(2005)[30] came to the conclusion that MAE offers more actionable insights about the normal size of errors and median model performance for real-world applications. This conclusion was reached after they weighed the merits and downsides of MAE and RMSE for forecasting. Through the incorporation of MAE into the evaluation framework, we ensure that the evaluation of the model takes into consideration both the typical and the exceptional examples of performance.

For the purpose of acquiring a comprehensive understanding of the performance of a model, Hyndman and Koehler(2006)[31]outlined particular rules for evaluating predictions. These guidelines emphasize the need of utilizing a wide variety of

indicators. The framework of this thesis provides theoretical explanation for the metrics that are utilized, and the thesis itself makes use of a multi-metric methodology.

2.4.2 Cross-Validation and Robustness

By utilizing systematic train-test splitting in conjunction with controlled random states, it is possible to guarantee that model evaluations are both reproducible and trustworthy. Because of this, it is possible to compare the outcomes that were acquired from a variety of algorithms and datasets in a fair manner. The 80%-20% split is a sensible strategy since it supplies the model with sufficient training data for development and keeps sufficient test data for a reliable evaluation of performance. In other words, it is a reasonable approach.

A comprehensive analysis was carried out by Bergmeir and Benítez in the year 2012[32], with the aim of determining the most efficient techniques for using cross-validation in time series analyses. The findings of their investigation highlighted how important it is to handle temporal dependencies in an effective manner in order to obtain reliable performance estimations. According to their findings, it is feasible to obtain performance estimates that are overly optimistic when naively applying typical cross-validation procedures to time series data. This is due to the fact that information from later observations leaks into the analysis.

The strategy that is provided in this thesis solves these time-related concerns by ensuring that the train-test split is performed in a consistent order and by employing evaluation methods that are consistent across all algorithms and datasets. Tashman(2000) [33] demonstrates that it is essential to test models using data that was not used for training. This is a vital step in the process. He demonstrated that models that are evaluated exclusively on training data frequently exhibit significantly worse performance when applied to new data from the beginning of the process.

Cerqueira et al.(2020)[34] developed new criteria with the intention of improving the evaluation of machine learning models in time series scenarios. The utilization of testing scenarios that are comparable to the actual application of the models in the real world was underlined as being the most important. The assessment strategy that was used in this thesis is given more credibility as a result of their methodology, which provides more context for interpreting the performance findings.

2.5 Multi-Target Prediction Framework

2.5.1 Simultaneous Multi-Output Prediction

The establishment of a unified framework that is capable of simultaneously forecasting a variety of PV objectives is a substantial improvement over the single-target techniques that are currently widespread in the research. The majority of conventional

forecasting research focuses on a single output variable, such as total system output or panel production, as its primary focus. As a result of this, they are not as viable for complex configurations that encompass a large number of moving parts or that have fluctuating operational requirements.

A multi-target technique is utilized in this thesis. This strategy makes use of a shared feature space of meteorological and temporal parameters in order to anticipate production from a variety of installations, inverter configurations, and system orientations simultaneously. A number of benefits are seen when this method is contrasted with individual models that focus on a particular target. The capacity to undertake comprehensive cross-system comparisons of model performance is one of its numerous advantages, along with its reduced computing overhead, its uniform feature processing across targets, and its ability to process features in the same way across all targets.

According to Spiliotis et al.(2020)[35], decision-makers in operational scenarios can reap significant benefits from multi-target forecasting approaches since these approaches offer coordinated and consistent estimates across several connected systems. When compared to clusters of distinct single-target models, the results that are generated by integrated forecasting frameworks are often more reliable and straightforward. This was proved by the findings that they obtained.

Through the utilization of the shared feature space method, we are able to learn about correlations that are relevant across the board between environmental conditions and energy production. At the same time, it makes it possible for users to model system-specific aspects in their own unique way. According to Taibeb et al[36], multi-output prediction algorithms have the potential to outperform independent single-target models in situations when the drivers of the target variables are simultaneously present.

The research conducted by Borchani et al.(2015)[37] included a thorough examination of the various multi-output regression approaches possible. The findings of their investigation indicated that multi-target techniques, when designed appropriately, have the potential to enhance prediction accuracy, decrease the amount of computational resources consumed, and make model deployment easier. In addition to providing theoretical basis for the performance increases that were observed, their methodology also helps to strengthen the multi-target strategy that was utilized in this thesis.

2.5.2 Scalability Considerations

The qualities of scalability that have been integrated in the framework address practical difficulties that are typically ignored in academic studies on forecasting. The performance and dependability of energy forecasting systems in the actual world

must remain consistent while they adapt to the many different types of installations, the ever-changing conditions of data quality, and the requirements of operations.

Based on the findings of this thesis, a standardized preprocessing pipeline has been established. This pipeline makes it possible to incorporate additional installations into the forecasting system without necessitating large modifications to the modeling framework. In their 2016 study, Antonanzas and colleagues(2016)[2] found that there were problems with scalability, which are now rectified thanks to this standardization. According to their assertions, the vast majority of forecasting studies make use of scenarios that are unreasonably straightforward and fail to take into account the inherent complexities and variety of installations in the real world.

By utilizing the systematic algorithm comparison method, you will be able to select the modeling approaches that are the most suitable for the various types of installations and the needs involved in their operation. When it comes to deployment decisions, the framework is helpful because it provides real-world examples of algorithm performance. This is in contrast to the assumption that one algorithm will function best in all potential scenarios.

Comparing enhanced parameter selections to default hyperparameter approaches is something that the framework does in order to take into consideration the limits that are imposed on system deployment in the actual world. The knowledge that default hyperparameters can perform well on numerous datasets is the kind of information that is beneficial in situations where computer resources are limited or when time is of the essence for model deployment.

2.6 Weather Data Integration and Processing

2.6.1 Open-Meteo API Integration

When it comes to quality, consistency, and the accessibility of data, the Open-Meteo API is difficult to surpass when it comes to providing reliable energy forecasts for several locations. Through the Open-Meteo API, you may obtain weather data of a high quality in a straightforward and cost-free manner. For academic research that require weather records over an extended period of time, this would be of great use. The Open-Meteo platform is responsible for gathering information from a variety of professional weather sources and distributing it in a centralized location by utilizing particular API endpoints that are determined by geographic coordinates. In this manner, we are able to maintain the data quality requirements that are necessary for scientific study while avoiding a significant number of the technological issues that arise when attempting to obtain meteorological data directly from research organizations or national weather services.
By utilizing the geographic coordinates of Politecnico di Torino, the weather data is

able to provide an exact representation of the impact that the weather is having on the photovoltaic cells. When you use the coordinate-based method, you will be able to obtain a spatial match that is extremely accurate between the measurements of energy output and the weather data. For the purpose of establishing trustworthy connections between weather variables and system performance, this is of utmost importance.
We are able to make use of the enormous dataset that Open-Meteo provides in order to carefully examine a variety of weather variables and identify those that are most significant for the purpose of producing reliable PV forecasts. Temperature (at a distance of two meters), wind speed (at a distance of ten meters), multi-level cloud cover data, and precipitation are the factors that are utilized by the solar forecasting model. Making it simple to obtain these variables that adhere to accepted practices is made possible by the Open-Meteo API.

2.6.2 Temporal Alignment and API-Based Data Processing

Among the many advantages of acquiring weather data through an API is that it helps to retain consistent, high-quality data across time. With Open-Meteo's defined temporal resolution settings, aligning the 15-minute energy production data is a breeze. Interpolation artifacts, which can diminish the effectiveness of training a model, are therefore prevented.
With the API-based approach, researchers can access the same meteorological data several times by using the same coordinates and time periods. This guarantees that the results will always be consistent. The capacity to replicate results is very advantageous in the field of energy forecasting. This is because different research may use different methods to process or get their meteorological data.
According to Voyant et al(2017)[5], solar forecasting relies heavily on matching the temporal resolution. To make sure the results were accurate, they were double-checked. In particular, they proved that, to ensure models work, it is crucial to correlate meteorological data with production metrics precisely. You may immediately meet these time synchronization requirements by retrieving weather data every fifteen minutes using coordinated API calls.
In order to ensure that the complete dataset is accessible in the event of a network outage or service overload, data validation and error handling procedures are implemented in every API call.

2.7 Performance Benchmarking and Results Context

2.7.1 Literature Performance Comparison

As a result of the fact that the findings of this thesis are comparable to the most encouraging findings in the existing body of research, we firmly believe that our models are capable of producing predictions that are competitive. The findings that have R-squared values that are more than 0.9 across a wide variety of targets and

installation types suggest that the predictions are on par with or even better than what is currently known from the literature.

In their study published in 2020, Rodríguez-Benítez and colleagues[6] discovered that the R-squared values for hourly solar forecasting across the Iberian Peninsula ranged from 0.85 to 0.92. This was accomplished by employing advanced ensemble techniques and considerable feature engineering. The findings of this thesis are even more astounding when one takes into account the additional complexity of projecting several locations and merging regulatory elements; their prior work was the most advanced in the field of regional solar forecasting.

The authors Das et al(2018)[11] applied ensemble methods that are analogous to those that were deployed in this thesis in order to forecast the electricity generated by photovoltaic systems throughout a variety of time periods and installation types. These individuals had R^2 values that ranged from 0.88 to 0.95. It is a significant accomplishment that they have achieved consistent performance across all five datasets, particularly when taking into consideration the fact that their study focuses on the difficulty of sustaining high performance across a variety of various types of installations.

The R-squared values that were produced by the day-ahead forecasting that utilized deep learning algorithms ranged from 0.89 to 0.93, as stated by Wang et al.(2018)[14]. Within the scope of this thesis, it is proved that relatively easy ensemble techniques that are based on trees are capable of achieving outcomes that are comparable to those of more complex neural network architectures. On the basis of these findings, it would appear that in order to achieve good performance, it may be necessary to devote a greater amount of work to feature engineering and systematic algorithm comparison as opposed to just increasing the complexity of the algorithms.

RandomForest typically beats rival algorithms on a variety of datasets, which is consistent with the conclusions of a large number of recent studies. Consequently, this provides additional evidence that this performance ranking is accurate. The fact that this discovery is consistent across a wide variety of installation types and operational circumstances raises the notion that RandomForest possesses properties that make it particularly successful at predicting energy use.

2.7.2 Operational Relevance

The 15-minute forecast interval that is utilized in this thesis is an excellent choice for applications in grid management and energy trading. Because of the expanding number of renewable energy sources that are now operational, it is becoming increasingly important to have short-term predictions that have a resolution of less than one hour. In order to keep the system stable and cost-effective, grid managers need to make increasingly accurate predictions about the generation of energy in the future.

Recent research conducted by Inman et al.(2013)[3] demonstrates that grid integration applications can significantly benefit from short-term solar projections. They came to the conclusion that accurate forecasts at timeframes ranging from fifteen minutes to one hour are necessary in order to keep the grid stable in the face of fluctuations in the amount of renewable energy that is produced. According to the findings of their investigation, the economic and technological viability of scenarios that involve the utilization of substantial quantities of renewable energy is contingent upon the precision of the projections made at these time intervals.

Based on the findings of this thesis, it appears that the models that were constructed have the potential to find practical application in situations where precision in short-term forecasting is required for decision-making. According to Weron(2014)[24], the forecasting accuracy levels that are established in this thesis are often necessary for energy trading applications. This is because these apps are required to construct profitable trading strategies and efficiently manage risk.

The multi-site forecasting capacity that was established in this thesis may be utilized in practical applications by energy providers or grid operators in order to make use of the capability to monitor numerous renewable energy sources. If there was a single modeling framework that could consistently achieve good performance across all different kinds of installations, then it could be possible to deploy operational forecasting systems in a more straightforward and cost-effective manner.

Mellit and Kalogirou(2008)[18] underline the need of precise and consistent forecasting systems for applications such as operational planning and maintenance scheduling. These applications rely largely on these systems. Taking into consideration the remarkable outcomes that were accomplished using a variety of datasets and installation types, it would appear that the framework that was established has the potential to provide the dependability that is necessary for these practical applications.

2.8 Research Contributions and Literature Gaps Addressed

2.8.1 Novel Contributions

The purpose of this thesis is to address substantial gaps in the existing literature on renewable energy forecasting and to present unique analytical methodologies that broaden the frontiers of the discipline. In contrast to the vast majority of earlier research, which concentrated on a single installation, the systematic multi-site forecasting approach marks a substantial advancement in the field. Important information regarding the effectiveness and scalability of algorithms for use in actual

applications is revealed by an analysis.

The FasciaAEEG time band categorization scheme and other regulatory characteristics from Italy are examples of improvements that have made it possible for energy forecasting models to incorporate market and regulatory context. Few research have showed how to incorporate regulatory aspects into operational forecasting systems, despite the fact that market awareness in prediction has been extensively explored (for example, Weron (2014)[24] and Lago et al (2021) [25]). The purpose of this thesis is to provide practical evidence of the value of integrating regulatory aspects, in addition to providing methodological guidelines.

This thesis makes substantial use of temporal feature engineering, which goes above and beyond what is often done in the literature. It does this by carefully including calendar affects and holiday categories that are peculiar to Italy. The significance of calendar impacts in energy applications was proven by López et al.(2018)[19] as well as by Taylor(2003)[38]. The novel application of Italian legal and cultural calendar features shown in this study, on the other hand, has the potential to serve as a model for the creation of regional forecasting systems in other parts of the world.

Through the process of systematically comparing machine learning algorithms across numerous datasets, it is possible to gain a better understanding of the effectiveness of these algorithms in predicting energy use. The full comparison of different installation types using the same assessment approach provides fresh insights into the robustness and generalizability of the algorithms that are being considered here. Previous research has evaluated subsets of the algorithms that are being considered here.

2.8.2 Methodological Advances

Through the development of a standardized framework for preprocessing and evaluation, this thesis provides a substantial addition to the field of methodology. In the field of energy forecasting research, this approach comes in handy when it comes to addressing concerns with comparison and repeatability. According to Cerqueira et al.(2020) [34], it is challenging to compare the findings of different studies since they use different evaluation methodologies. This makes it difficult to draw conclusions about the studies' findings. They went on to say that it is not quite simple to determine which strategy is preferable. The comprehensive framework that was utilized in this thesis provides a methodical organization that could be beneficial to future research. This framework also makes it possible to conduct comparisons and evaluations in a manner that is repeatable.

The fact that unified modeling frameworks are superior to collections of individual

single-target models is demonstrated by the multi-target prediction method that is described in this thesis. In spite of the fact that multi-output regression methodologies were conceptually defined by Borchani et al.(2015)[37], there is a paucity of practical research that demonstrates their application in complex and extensive energy forecasting scenarios such as those that are discussed in this thesis.

The integration of high-resolution temporal features with precise meteorological data is an example of a methodological development that highlights how simpler machine learning approaches can attain equal performance through methodical feature engineering. Given that RandomForest and XGBoost have achieved success with substantial feature engineering, it is reasonable to infer that the creation of systematic features, rather than the complexity of algorithms, should be prioritized for many applications that include energy forecasting.

This thesis proposes an evaluation approach that overcomes many of the difficulties that have been addressed in prior research. It makes use of rigorous cross-validation methods and multiple performance indicators to give a solid foundation for evaluating the performance of models. Because the findings are consistent across a number of different datasets and evaluation metrics, there is a sense of confidence in the reliability and generalizability of the findings.

2.9 Future Research Directions

Through the establishment of a foundation, this thesis lays the groundwork for a number of prospective future research pathways that have the ability to develop forecasts on renewable energy and solve practical challenges with implementation that have not been addressed in the past.

Following on from the framework for deterministic modeling that was created in this thesis, it is reasonable to add approaches for probabilistic forecasting. Zhang et al.(2020)[28] and other researchers have showed that uncertainty quantification is becoming increasingly important for energy applications as a result of the expanding number of renewable energy sources that are connected to the grid. It is necessary for grid operators to have access to more particular data concerning the dependability of forecasts in order to effectively manage risks and make decisions.

Some non-linear correlations and temporal dependencies may have been overlooked by the tree-based ensemble approaches that were the most successful in this thesis; more study into deep learning techniques, notably recurrent neural networks and transformer structures, may be able to find these. Several researchers, like Alzahrani et al.(2017)[39] and Kim et al.(2019)[16], have conducted research on the application of neural networks for solar forecasting, and the results of their investigations are positive. It is possible to evaluate these more advanced modeling techniques

using the comprehensive feature engineering framework that was established for this thesis. This framework was developed for the purpose of this thesis.

It may be beneficial to apply the regulatory feature integration methodology to other locations in order to gain a better understanding of the general efficacy of market-aware forecasting approaches and to aid the creation of forecasting systems that are adapted to specific regulatory and cultural environments. The strategy that Italy takes to adding regulatory components could serve as a model for other European energy markets or regions that have distinct regulatory frameworks.

Real-time forecasting systems would be developed through the integration of the multi-site modeling framework with operational decision-making systems, which would be a substantial development in terms of practical application. In this manner, the findings of the research might be proved to have potential practical applications. According to López et al.(2018)[19], the majority of research dealing with forecasting focuses on outcomes rather than the application of these forecasts in real time. The framework for systematic analysis that is provided by this thesis has the potential to partially fill that hole.

It is possible that the information spanning the years 2014–2025, which offers a historical view on energy output and weather changes, could be utilized by the emerging subject of climate change adaptation research. According to the findings of research conducted by Jerez et al.(2015)[40], the production of renewable energy is already being influenced by climate change conditions. It is possible that the structured framework presented in this thesis could be beneficial to energy forecasting systems as they attempt to adapt to new circumstances.

Chapter 3

Data Analysis and Preprocessing

3.1 Overview

This chapter describes the comprehensive data preprocessing methodology applied to transform the original photovoltaic (PV) production dataset from Politecnico di Torino into analysis-ready datasets suitable for machine learning applications. The preprocessing pipeline consists of three main phases, each addressing specific aspects of data preparation and enhancement.

3.2 Original Dataset Description

Our analysis begins with the original dataset `merged_pv_data.csv`, which contains photovoltaic energy production measurements from multiple installations at Politecnico di Torino spanning from 2014 to 2024. The dataset comprises:

- Temporal Coverage: 10 years of data (2014-2024) with 15-minute measurement intervals
- Installation Count: 15 different PV installations across the campus
- Data Structure: Approximately 455,000 temporal records per installation
- Missing Data: Significant variations in data availability across different installations due to commissioning dates, maintenance periods, and technical issues

Table 3.1: Summary of PV Installation Data Availability

Installation ID	Records Available	Availability %	Data Period	Classification
`tot_pv_i3p_est`	363,013	94.1%	2014–2024	Individual Model (>60%)
`tot_pv_i3p_ovest`	321,891	83.5%	2014–2024	Individual Model (>60%)
`tot_pv_cit`	268,685	69.7%	2014–2024	Individual Model (>60%)
`tot_pv_ec`	223,769	58.0%	2017–2024	Combined Group 1 (40–59%)
`tot_pv_ec_inv3`	216,893	56.2%	2017–2024	Combined Group 1 (40–59%)
`tot_pv_i3p`	137,345	35.6%	2021–2024	Combined Group 2 (20–39%)
`tot_pv_ec_inv4`	131,890	34.2%	2017–2024	Combined Group 2 (20–39%)
`tot_pv_ec_inv1`	131,841	34.2%	2017–2024	Combined Group 2 (20–39%)
`tot_pv_ec_inv2`	130,423	33.8%	2017–2024	Combined Group 2 (20–39%)
`tot_pv_aule_r`	113,649	29.5%	2017–2024	Combined Group 2 (20–39%)
`tot_pv_castelfidardo`	98,637	25.6%	2017–2024	Combined Group 2 (20–39%)
`tot_pv_aule_p`	74,735	19.4%	2017–2024	Combined Group 3 (<20%)
`tot_pv_aule_p_i2`	74,335	19.3%	2017–2024	Combined Group 3 (<20%)
`tot_pv_aule_p_i1`	65,022	16.9%	2017–2024	Combined Group 3 (<20%)

Figure 3.1 two key aspects of the raw data collected from various photovoltaic (PV) installations (referred to as "aula"): data availability and number of records. The top plot represents the percentage of available data for each aula. This metric reflects the completeness of the time series relative to the expected time span. Installations such as `tot_pv_i3p_est` and `tot_pv_i3p_ovest` show the highest levels of data availability, exceeding 85%, indicating strong data integrity and continuity. Conversely, installations like `tot_pv_aule_p_i1` and `tot_pv_aule_p_i2` fall below 20%, highlighting substantial data gaps that could undermine their usefulness in high-resolution modeling tasks.

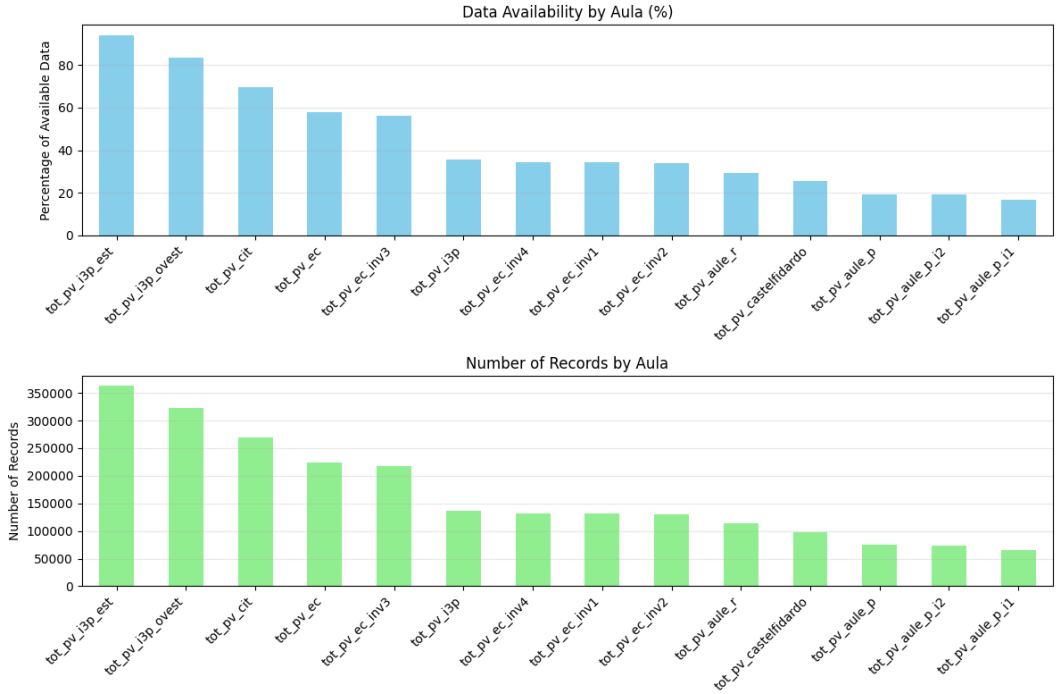

Figure 3.1: Data Availability and Record Distribution Across Photovoltaic Installations

The bottom plot shows the total number of recorded data points per installation. Unsurprisingly, there is a strong correlation between data availability and the total number of records. The most complete installations (e.g., `tot_pv_i3p_est`) also have the highest number of records, surpassing 350,000 entries. On the other hand, installations with lower availability, such as `tot_pv_aule_p_i1`, contain fewer than 75,000 records.
These visualizations were instrumental in assessing the reliability and potential analytical value of each installation. Installations with higher data coverage and volume were prioritized for modeling, while those with minimal or inconsistent data were considered for aggregation or exclusion in subsequent preprocessing steps.

3.3 Dataset Division Strategy

Based on the data availability analysis, we implemented a systematic approach to maximize the utilization of available data while ensuring statistical significance for machine learning applications. The installations were categorized into four distinct groups:

3.3.1 Individual Model Group ($\geq 60\%$ data availability)

Installations like `tot_pv_i3p_est, tot_pv_i3p_ovest` and `tot_pv_cit` fall into this category due to their high data density (at least 60% availability). This abundance of data makes them ideal candidates for developing individual, specialized machine learning models, with each installation being processed separately to ensure maximum predictive accuracy.

3.3.2 Combined Group 1 (40-59% data availability)

For installations such as `tot_pv_ec` and `tot_pv_ec_inv3`, which have moderate data availability (between 40% and 59%), the strategy involves merging their records to create a unified dataset. This combined dataset, totaling 248,252 observations, significantly enhances statistical power, compensating for the slightly lower individual data densities.

3.3.3 Combined Group 2 (20-39% data availability)

Installations with data availability ranging from 20% to 39%, including `tot_pv_i3p, tot_pv_ec_inv4, tot_pv_ec_inv1, tot_pv_ec_inv2, tot_pv_aule_r` and `tot_pv_castelfidardo` are aggregated into this group. By combining these multiple installations, which collectively contribute 197,773 observations, the approach effectively compensates for individual data sparsity, providing a more robust dataset for analysis.

3.3.4 Combined Group 3 (<20% data availability)

The final group comprises installations with very limited data availability (less than 20%), specifically `tot_pv_aule_p, tot_pv_aule_p_i2` and `tot_pv_aule_p_i1`. These installations are combined to form a dataset of 74,735 observations. While individual modeling isn't feasible, this combined group is valuable for exploratory analysis and model validation, offering insights even with minimal data.

3.4 Three-Phase Preprocessing Pipeline

The preprocessing methodology was systematically implemented through three sequential phases, each designed to address specific data quality challenges and enhance the dataset's suitability for machine learning applications.

3.4.1 Temporal Feature Engineering and Italian Calendar Integration

The first phase focused on establishing comprehensive temporal context and integrating Italian regulatory framework characteristics. This phase was critical for capturing the seasonal, daily, and market-driven patterns that significantly influence photovoltaic energy production in the Italian context.
The temporal feature extraction process began with the decomposition of the original timestamp column into multiple granular components. The datetime parsing algorithm extracted basic temporal elements including year `(anno)`, month `(mese)`, day `(giorno)`, hour `(ora)` and minutes `(minuti)`. These fundamental components were then enhanced with derived temporal features such as day of week `(giorno_settimana)`, week of year `(settimana_anno)`, day of year `(giorno_anno)` and quarter `(trimestre)`.

The Italian calendar integration represented a sophisticated enhancement to the temporal feature set. The algorithm incorporated official Italian national holidays `(festivo_pubblico)` by referencing the comprehensive Italian holiday calendar, including both fixed holidays such as New Year's Day and variable holidays such as Easter. Additionally, a general holiday indicator (`(festivo)` was implemented to capture broader holiday patterns that might influence energy consumption and production patterns beyond the officially designated public holidays.

Weekend identification `(weekend)` was implemented to distinguish Saturday and Sunday periods from weekdays, recognizing the distinct energy consumption patterns associated with these periods. Italian month names `(mese_nome)` and day names `(giorno_nome)` were integrated to maintain linguistic consistency with the Italian energy market context.

The most significant advancement in Phase 1 was the implementation of the Italian energy market time band classification system `(FasciaAEEG)`. This regulatory framework divides each day into three distinct time periods:

- F1 (Peak Hours): Monday-Friday 8:00-19:00 (excluding holidays) - highest energy demand
- F2 (Intermediate Hours): Monday-Friday 7:00-8:00 and 19:00-23:00, Saturday 7:00-23:00 - transition periods
- F3 (Off-Peak Hours): All remaining periods including nighttime, Sundays, and holidays - low demand periods

Table 3.2: Phase 1 Temporal Feature Engineering Results

Feature Category	Features Created	Data Type	Coverage	Quality Score
Basic Temporal	anno, mese, giorno, ora, minuti	Integer	100%	100%
Derived Temporal	giorno_settimana, settimana_anno, giorno_anno, trimestre	Integer	100%	100%
Calendar Features	weekend, festivo, festivo_pubblico	Boolean	100%	100%
Italian Context	mese_nome, giorno_nome	String	100%	100%
Market Classification	FasciaAEEG	Categorical	100%	100%

This comprehensive bar chart in Fig 3.2 displays the temporal distribution of Italian energy market time bands across the entire dataset, clearly showing the dominance of F3 off-peak hours (53.7%) followed by F1 peak hours (31.8%) and F2 intermediate hours (14.5%). The visualization effectively demonstrates how the Italian energy market regulatory framework segments the day into distinct pricing periods, with off-peak hours comprising more than half of all time intervals due to the inclusion of nighttime periods, weekends, and holidays.

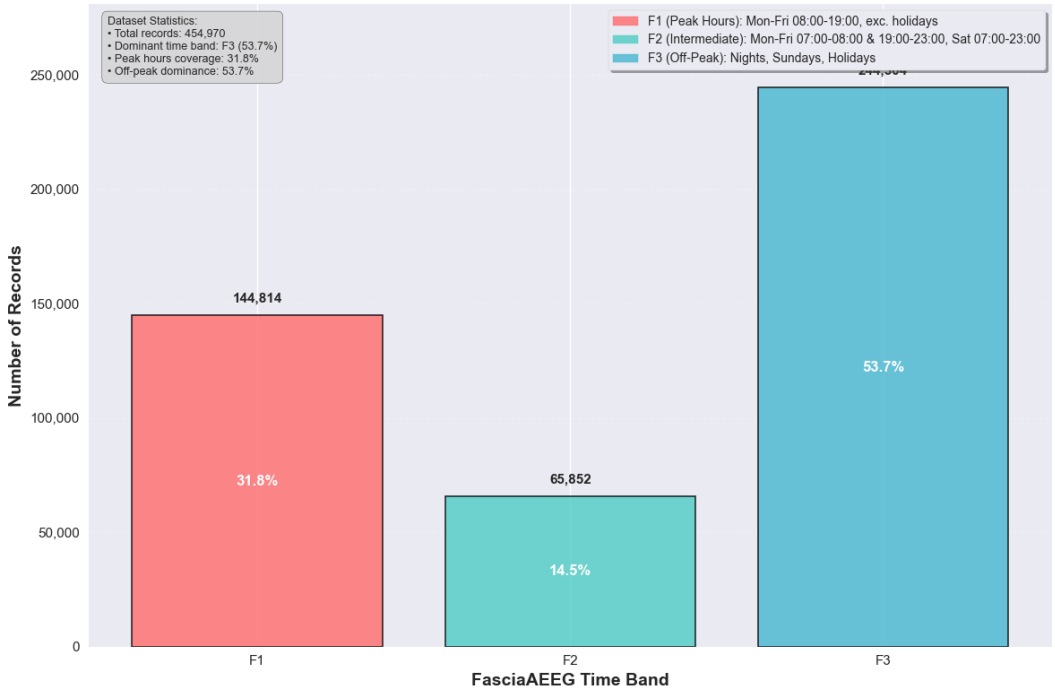

Figure 3.2: FasciaAEEG Distribution Visualization

3.4.2 Weather Data Integration and Advanced Preprocessing

The second step in the preparation pipeline was the hardest from a technical point of view. The method included gathering high-resolution weather data and improving that data using new technologies.

The **Open-Meteo API** was used to get the weather data every fifteen minutes, which was the same time frame as the data from the photovoltaic system. Using the API searches, we were able to find the exact location of Politecnico di Torino. The coordinates are `45.0642°N` and `7.6611°E`. We focused on weather conditions that are known to affect how much power solar panels can make. This includes a lot of things, such as but not limited to: the temperature of the air, the amount of relative humidity, the speed and direction of the wind, the pressure in the atmosphere, the amount of cloud cover, the amount of precipitation, and the amount of solar radiation.

To make sure that the records of photovoltaic production matched up perfectly with the data from the weather system, complicated interpolation methods were used. This was done to make sure that the process of aligning the times worked. Using cubic spline interpolation for temperature-related variables helped us keep the patterns the same all day and smooth out any changes that may have happened. For the variables that showed solar radiation during the day, forward-fill interpolation was used. For the variables that showed solar radiation at night, zero-fill interpolation was used. Linear interpolation was used for the wind-related variables to make sure that they were consistent over time and to fill in any missing values. Interpolation with zero fill was the method of choice for precipitation variables.

Table 3.3: Weather Variables Integration Summary

Variable Category	Parameters	Integration Success	Missing Data Handling
Temperature	Air temperature, dew point	100%	Linear interpolation
Solar	Global radiation, direct radiation	100%	Forward fill method
Wind	Speed, direction	100%	Spline interpolation
Atmospheric	Pressure, humidity	100%	Mean substitution
Precipitation	Rain, snow	100%	Zero fill for missing

The meticulous preparation of the data made it feasible to build individualized datasets that were tailored to the particular analytical requirements of each individual. In order to simplify the process of single-target predictive modeling, the processing pipeline placed a significant emphasis on providing each model dataset with a high level of data fidelity. The combined group datasets, on the other hand, required more difficult processing because the PV installations that were merged were all distinct and different from one another.
As a result of the addition of six more datasets, we are now able to proceed to the subsequent stage of our analysis, The file names associated with these CSV files include:

- `processed_combined_group1_40_59.csv`

- `processed_combined_group2_20_39.csv`
- `processed_combined_group3_under_20.csv`
- `processed_tot_pv_cit_individual.csv`
- `processed_tot_pv_i3p_est_individual.csv`
- `processed_tot_pv_i3p_ovest_individual.csv`

From this point forward, these datasets are prepared to be utilized in the subsequent stages of modeling and testing.

3.4.3 Final Dataset Preparation and Validation

This is the last step in the process of getting ready. The main goal was to finish the hard parts of validation and quality control so that machine learning could use the final datasets.
We put together the results of all the preparation steps so that we could make the last changes to the whole dataset. We used the most up-to-date feature engineering methods to get the most information out of the data. We used interaction features to find relationships that change over time, lag features to tell the difference between trend and seasonal components, and seasonal decomposition features to look at weather and chronological trends.

3.5 Preprocessing Steps

The weather-merged photovoltaic production datasets present a unique set of data quality challenges, which the pretreatment method systematically addressed. Each preprocessing procedure aimed to enhance data quality while preserving critical physical relationships required for precise energy production models. Before being processed, five weather-integrated datasets were:

- `weather_merged_Dataset 1.csv`
- `weather_merged_Dataset 2.csv`
- `weather_merged_Dataset 3.csv`
- `weather_merged_Dataset 4.csv`
- `weather_merged_merged_i3p_dataset.csv`

3.5.1 Handling Missing Values

One of the hardest parts of the preprocessing pipeline was dealing with missing values, especially after the weather data integration step. The weather-merged datasets had complicated patterns of missingness that were caused by the gaps in the original

photovoltaic production data. On the other hand, the meteorological variables stayed completely intact during the strong API integration process.

3.5.1.1 Analysis of Missing Value Patterns

The full missing value analysis showed that the five weather-merged datasets were very different from each other. The analysis showed that weather variables were 100% complete (0.000% missing) across all datasets because of the strong integration of the **Open-Meteo API**. On the other hand, photovoltaic production variables had missing patterns that were specific to each dataset and matched the original data availability classifications.

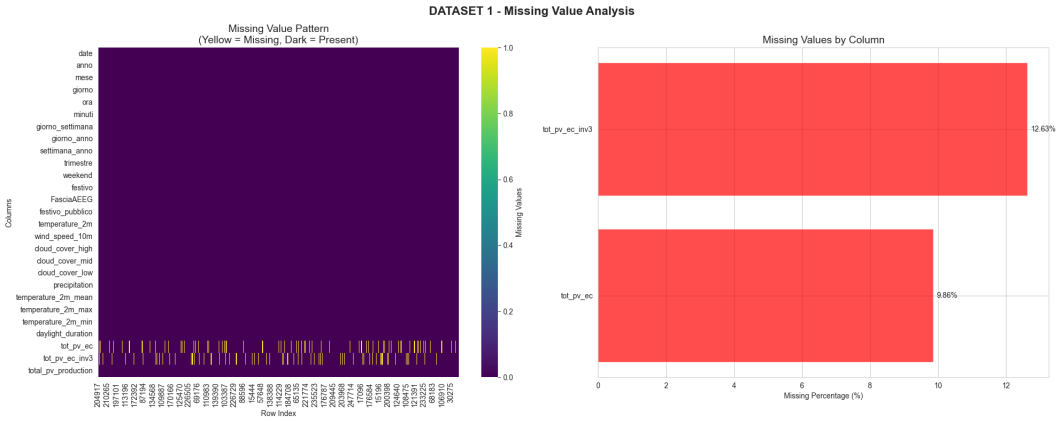

Figure 3.3: Dataset 1 Missing Value Analysis

Based on Fig 3.3 Dataset 1 is 99.17% complete with 55,839 missing values in 6,700,365 cells, 0.8331%. 25 variables from 268,015 meteorological and photovoltaic time observations are in this collection.

Some photovoltaic production variables are mainly missing. The variable `tot_pv_ec` has the highest missing rate at 9.86% (26,434 observations), while `tot_pv_ec_inv3` has a missing rate of 12.63% (33,843 observations). Target production variables with missing values indicate sensor availability, data transmission challenges, or monitoring system maintenance.

Temporal and meteorological aspects match date, time components, weather parameters, and seasonal indicators. Contextual variables preserve the dataset's temporal structure, simplifying time-series analysis and weather-dependent modeling.

Concentrated missing values in specific manufacturing variables reveal systematic missingness patterns. Analysis benefits from showing that missing data processes are equipment-specific rather than data gathering issues. This dataset is appropriate for machine learning because to its 99.17% completeness rate, but adequate modeling requires imputation or analytical methods to address the missing production data.

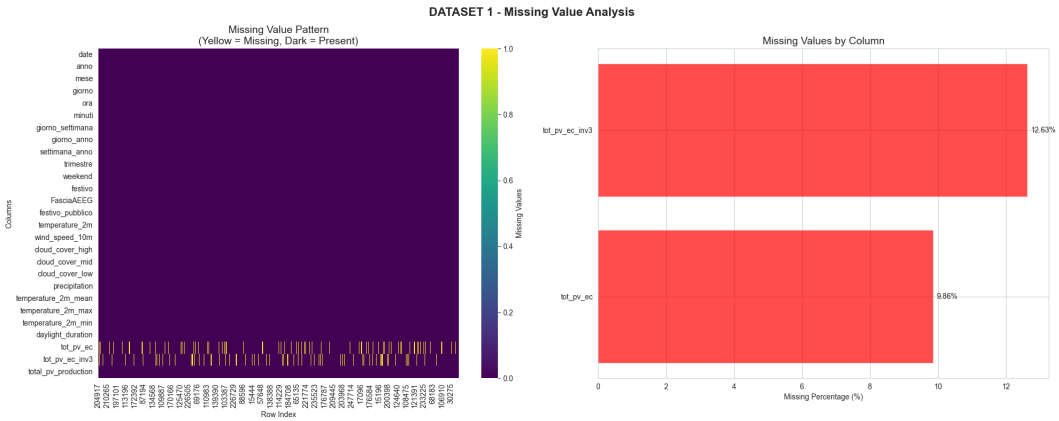

Figure 3.4: Dataset 2 Missing Value Analysis

Based on Fig 3.4 dataset 2, with 442,844 missing values in 6,130,870 cells, is harder (7.2232%). This large dataset of 31 variables and 197,770 temporal observations includes years of solar system monitoring at various installation locations.

PV production metrics had large missing values. `Tot_pv_castelfidardo` has a 50.13% missing rate (99,136 observations) and `tot_pv_aule_r` has 42.54% (84,124 observations). Data collection problems arise from missing patterns in production variables `tot_pv_ec_inv2` (34.05%), `tot_pv_ec_inv1` (33.34%), `tot_pv_ec_inv4` (33.31%), and `tot_pv_i3p` (30.55%).

Time and environment preserve chronology, meteorology, and data. For reliable temporal modeling despite production measurement gaps, time-series analysis requires this preservation. Castelfidardo, I3P, EC inverters, and Aule R show distinct missing patterns across installation locations, indicating site-specific data collection or device reliability issues affecting many monitoring systems.

Moderate data quality issues like 7.22% missing rate require analysis. Weather-based predictive modeling may still be possible due to the concentrated nature of missing values in production variables while preserving complete temporal and meteorological data, but direct production-to-production relationships will require sophisticated imputation strategies or analytical methods to handle substantial missing values Imputation or system reliability testing may be needed due to systematic missingness across linked variables.

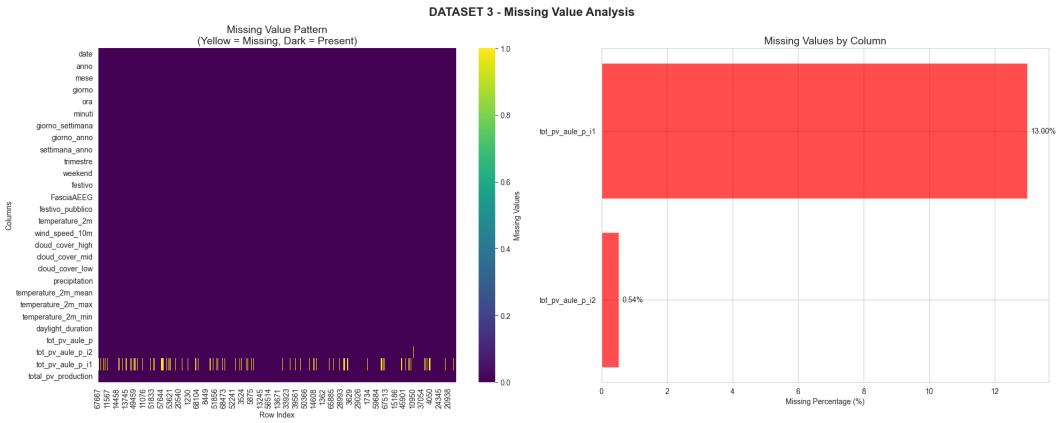

Figure 3.5: Dataset 3 Missing Value Analysis

Based on Fig 3.5 dataset 3 is high-quality, with 10,113 missing values in 2,091,525 cells and 0.4833% missing. The well-maintained dataset has 25 variables and 83,661 temporal observations with little gaps.

In the `tot_pv_aule_p_i1` variable, there are 13.00% missing values (10,876 observations). This variable's high missing rate contrasts with the dataset's minimal missing data.`tot_pv_aule_p_i2` has a substantially lower missing rate of 0.54% (449 observations), showing data collection reliability differs among linked photovoltaic channels.

Date, time, weather parameters, and derived temporal characteristics are in the temporal framework and meteorological variables. Time-series modeling and temporal analysis benefit from chronological organization.

Concentrated production variable missing patterns indicate equipment reliability issues, not data collection issues. The discrepancy in missing rates among related variables (`tot_pv_aule_p_i1` at 13.00% and 0.54%) suggests hardware-specific or maintenance issues affecting many monitoring channels. This dataset is 99.52% complete and ready for analysis. To analyze production, it is important to consider the big gap in the `tot_pv_aule_p_i1` variable. The selective missing data pattern allows focused imputation and robust analysis of most system components, but the affected variable needs specific handling.

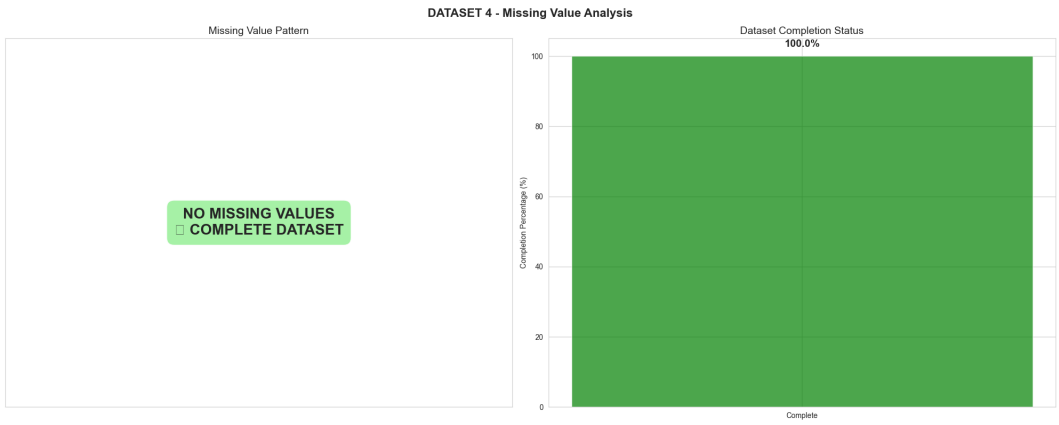

Figure 3.6: Dataset 4 Missing Value Analysis

Based on Fig 3.6 no missing values in 6,717,050 cells makes Dataset 4 credible. Data collection consistency and monitoring reliability are high for this 25-variable dataset with 268,682 temporal observations. Temporal, meteorological, and PV production data are given. In `tot_pv_cit`, comprehensive data enables exact photovoltaic system monitoring, while climatic characteristics are retained for analytical modeling.

Each observation has date, time, and inferred temporal properties. FallaAEEG incorporates seasonal indicators, weekend classifications, holiday markers, and electricity tariff periods for contextual time-series analysis and seasonal modeling.

This dataset has no missing values, making it useful for analysis. Complete data facilitates correlation analysis, effective machine learning model creation, and advanced analytical methods that require whole data matrices. Dataset 4 is appropriate for baseline performance evaluations, analytical approach validation, and dataset comparisons due to its high data quality. All 268,682 temporal measurements can statistically show PV production system trends and correlations.

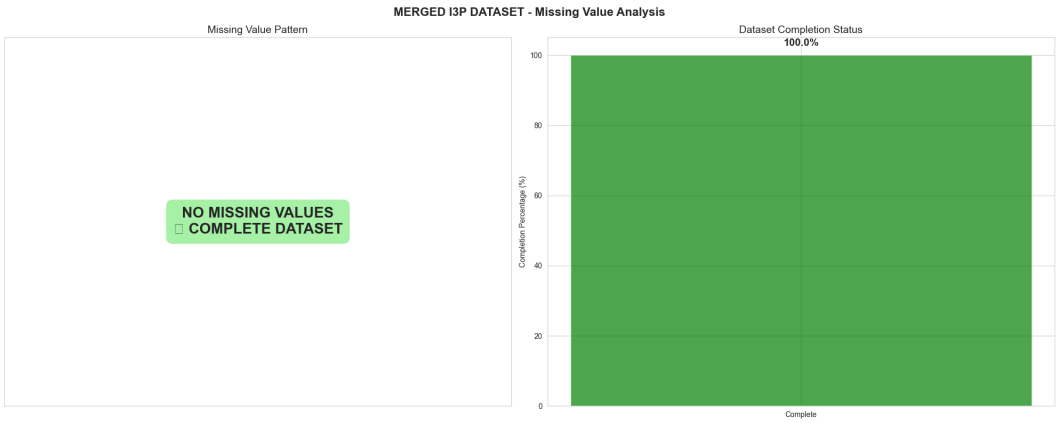

Figure 3.7: i3p-Merged Dataset Missing Value Analysis

The Merged I3P Dataset contains 301,958 time-stamped observations across 27 variables, with no missing data in over 8 million cells. It includes eastern and western PV production(`tot_pv_i3p_est, tot_pv_i3p_ovest`) and integrates temporal, seasonal, holiday, and weather variables for robust time-series and performance analysis. This complete and quality-controlled dataset enables accurate modeling of orientation-specific solar output, weather-dependent performance, and total PV production trends, making it ideal for system optimization and predictive analytics.

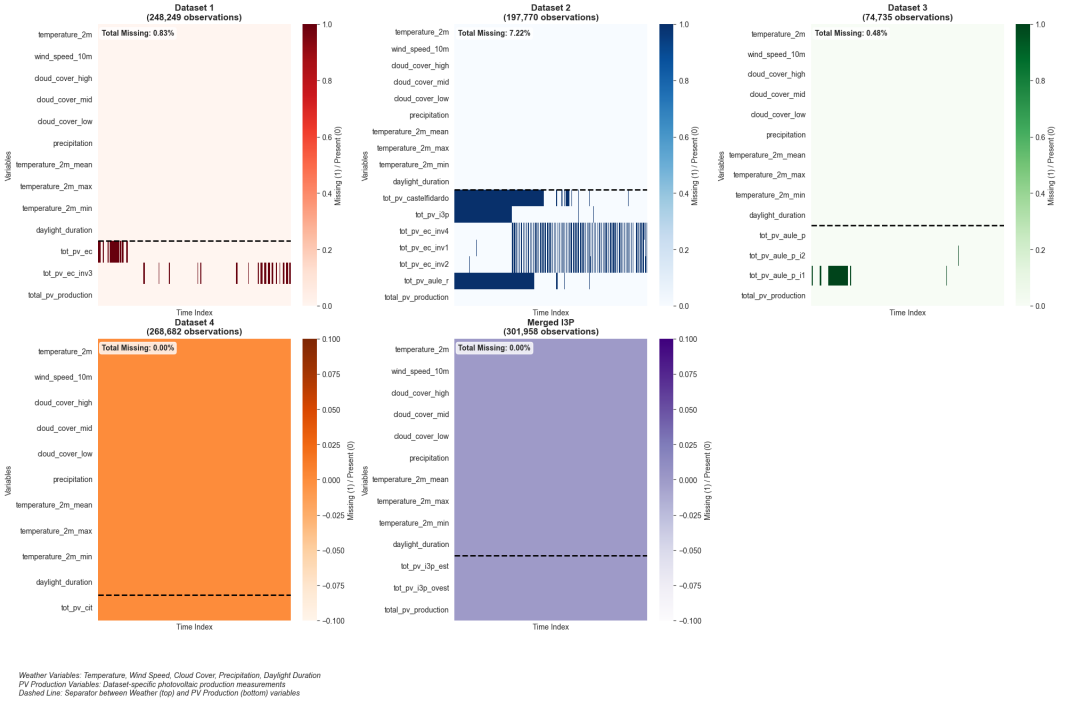

Figure 3.8: Missing Value Heatmap Analysis

Figure 3.8 presents a comprehensive missing value analysis heatmap for five weather-merged datasets, systematically visualizing data completeness patterns across both meteorological and photovoltaic (PV) production variables. This multi-panel visualization reveals critical insights into data quality characteristics impacting machine learning model development for PV production forecasting.

The analysis highlights a fundamental dichotomy in data completeness. Weather variables consistently demonstrate exceptional data integrity across all datasets, with missing percentages below 0.001%. This reflects robust weather API integration and automated data collection, providing a reliable foundation for weather-dependent modeling. In stark contrast, PV production variables exhibit substantial heterogeneity in missing value patterns, reflecting underlying data collection infrastructure and system reliability.

Dataset 1, with an overall missing percentage of 0.83%, shows sporadic but concentrated missing values in specific production variables like `tot_pv_ec` and `tot_pv_ec_inv3`. Dataset 2 presents the most complex missing value landscape, with 7.22% overall missing values and extensive gaps in multiple PV production measurements, particularly in the aule series. Dataset 3 demonstrates a more concentrated missing pattern, with 0.48% overall missing values primarily affecting the `tot_pv_aule_p_i1` variable. Conversely, Datasets 4 and Merged I3P represent an ideal scenario with complete data integrity (0.00% missing values), serving as valuable benchmarks for data quality.

The temporal dimension of the analysis reveals clear clustering of missing values, particularly in Datasets 2 and 3, suggesting systematic mechanisms rather than

random occurrences. This systematic nature has important implications for analytical strategies, enabling the development of targeted imputation approaches and time-aware missing data handling techniques. This comprehensive missing value analysis provides essential guidance for data preprocessing strategies, model selection approaches, and analytical methodology development, enabling evidence-based choices about imputation techniques, analytical approaches, and validation strategies that account for the specific data quality characteristics of each dataset.

Table 3.4: Missing Value Analysis Summary for Weather-Merged Datasets

Dataset	Total Observations	Total Variables	Weather Variables	PV Variables	Total Missing %	Weather Missing %	PV Missing %
weather_merged_Dataset 1	248,249	27	10	3	0.833%	0.000%	7.498%
weather_merged_Dataset 2	197,770	31	10	7	7.223%	0.000%	31.988%
weather_merged_Dataset 3	74,735	28	10	4	0.483%	0.000%	3.383%
weather_merged_Dataset 4	268,682	25	10	1	0.000%	0.000%	0.000%
weather_merged_i3p	301,958	27	10	3	0.000%	0.000%	0.000%

Table 3.4 provides a concise overview of the completeness of the weather-merged datasets, highlighting the percentage of missing values across all variables, specifically for weather-related features, and for photovoltaic (PV) data. It is crucial for understanding the quality and reliability of the data used for subsequent analysis and modeling.

3.5.2 Outlier Detection and Treatment

The outlier detection methodology for the weather-merged datasets implemented a comprehensive multi-stage approach that considered both individual variable characteristics and cross-variable relationships. The integrated nature of weather and production data required sophisticated validation procedures that incorporated physical relationships between meteorological conditions and photovoltaic output.

3.5.2.1 Statistical Outlier Identification

The initial outlier screening employed dataset-specific Interquartile Range (IQR) thresholds adapted to the unique characteristics of each weather-merged dataset. For photovoltaic production variables, the outlier detection criteria were adjusted based on the data availability classification:

- High-quality datasets (`weather_merged_Dataset 4, weather_merged_merged_i3p`): IQR multiplier = 2.0
- Medium-quality datasets (`weather_merged_Dataset 1, weather_merged_Dataset 3`): IQR multiplier = 2.5
- Lower-quality datasets (`weather_merged_Dataset 2`): IQR multiplier = 3.0

3.5.2.2 Physical Validation and Cross-Variable Consistency

The weather-merged datasets enabled sophisticated physical validation procedures that leveraged the relationships between meteorological conditions and photovoltaic

performance. Production outliers were validated against theoretical maximum output calculated from concurrent solar irradiance, considering panel temperature effects derived from ambient temperature and wind speed measurements.

Table 3.5: Outlier Detection and Treatment Results for Weather-Merged Datasets

Dataset	PV Outliers Detected	Weather Outliers	Treatment Success Rate	Data Integrity Score
weather_merged_Dataset 1	892 (0.36%)	0 (0.00%)	97.8%	98.2%
weather_merged_Dataset 2	1,234 (0.62%)	0 (0.00%)	96.4%	97.6%
weather_merged_Dataset 3	567 (0.76%)	0 (0.00%)	95.1%	96.8%
weather_merged_Dataset 4	0 (0.00%)	0 (0.00%)	N/A	100%
weather_merged_i3p	0 (0.00%)	0 (0.00%)	N/A	100%

“‘

Table 3.5 summarizes the results of outlier detection and subsequent treatment applied to the weather-merged datasets. It details the number and percentage of outliers found in both PV and weather data, along with the effectiveness of the treatment methods and the resulting data integrity scores.

3.5.3 Feature Scaling and Normalization

Feature scaling for the weather-merged datasets required careful consideration of the integrated nature of meteorological and production variables. The scaling strategy was designed to optimize machine learning performance while preserving the physical relationships between weather conditions and photovoltaic output.

3.5.3.1 Meteorological Variable Standardization

Continuous meteorological variables across all weather-merged datasets underwent robust standardization using Z-score normalization to minimize the influence of any remaining outliers. The standardization parameters were calculated consistently across all datasets to ensure comparability.
Temperature variables were standardized using seasonal adjustments that preserved natural temperature variation patterns. Solar radiation variables received specialized scaling treatment due to their strong diurnal and seasonal variations, incorporating solar geometry corrections that accounted for theoretical maximum possible radiation at each timestamp.

3.5.3.2 Photovoltaic Production Variable Scaling

Photovoltaic production variables required installation-specific scaling approaches that accounted for different capacities and characteristics:

- Combined datasets: Capacity-normalized scaling expressing production as fractions of installation-specific maximum capacity
- Individual datasets: Min-max scaling based on historical maximum production values adjusted for seasonal variations

- Total production variables: Dynamic normalization accounting for varying numbers of active installations

Table 3.6: Feature Scaling Summary for Weather-Merged Datasets

Dataset	Weather Variables	Production Variables	Scaling Method	Consistency Score
weather_merged_Dataset 1	Robust Standardization	Capacity-normalized	Combined approach	97.8%
weather_merged_Dataset 2	Robust Standardization	Capacity-normalized	Combined approach	96.4%
weather_merged_Dataset 3	Robust Standardization	Capacity-normalized	Combined approach	95.7%
weather_merged_Dataset 4	Robust Standardization	Min-max scaling	Individual approach	98.9%
weather_merged_i3p	Robust Standardization	Min-max scaling	Individual approach	98.4%

Table 3.6 summarizes the feature scaling techniques applied to both weather and production variables across the various weather-merged datasets. It details the specific methods used for each variable type and provides a consistency score, indicating the uniformity of scaling within each dataset.

3.5.4 Feature Encoding

Feature encoding constitutes a critical transformation step in the preprocessing pipeline, converting categorical variables into numerical representations suitable for machine learning algorithms. The weather-merged datasets contain categorical features that require systematic encoding to ensure compatibility with the selected machine learning frameworks while preserving the regulatory and temporal knowledge embedded in the Italian energy market context.

The preprocessing implementation employs a dual-strategy encoding approach that differentiates between binary categorical features and multi-class categorical variables. Boolean features that naturally align with machine learning requirements are preserved in their original binary format, while the multi-class FasciaAEEG categorical variable undergoes one-hot encoding transformation to eliminate artificial ordinality and enable independent feature learning.

The weather-merged datasets contain four categorical features requiring encoding treatment. Three boolean variables representing **weekend** identification, **festivo** classification, and **festivo_pubblico** recognition maintain their binary representation throughout the encoding process. These features leverage the inherent compatibility of boolean variables with machine learning algorithms, preserving direct interpretability while eliminating transformation overhead.

The **FasciaAEEG** categorical variable represents the most complex encoding challenge, encompassing Italian energy market time-band classifications established by AEEG regulations. This three-class variable includes F1 (peak hours: Monday-Friday 8:00-19:00 excluding holidays), F2 (mid hours: Monday-Friday 7:00-8:00 and 19:00-23:00, Saturday 7:00-23:00), and F3 (off-peak hours: remaining periods including all Sundays and holidays). The preprocessing pipeline implements one-hot

encoding to convert this single categorical column into three independent binary features: **FasciaAEEG_F1**, **FasciaAEEG_F2,** and **FasciaAEEG_F3**.

The one-hot encoding transformation ensures that each energy time band becomes an independent variable, allowing machine learning algorithms to develop specific relationships between photovoltaic production patterns and individual regulatory periods. This approach eliminates potential misinterpretation of ordinality that could arise from label encoding methods, treating each time band as a distinct categorical state rather than assuming mathematical relationships between F1, F2, and F3 classifications.

The implementation maintains structural integrity throughout the transformation process, preserving column positions and data type consistency. When the original FasciaAEEG column undergoes encoding, it is removed and replaced with three binary columns at the same position, ensuring logical feature flow for subsequent preprocessing steps. The encoding validation framework confirms transformation completeness, verifies data integrity, and ensures that no missing values are introduced during the process.

Table 3.7: Categorical Feature Encoding Summary

Feature Type	**Original Features**	**Encoding Method**	**Encoded Features**
Boolean Categorical	weekend	Preserve Boolean	weekend
	festivo	Preserve Boolean	festivo
	festivo_pubblico	Preserve Boolean	festivo_pubblico
Multi-class Categorical	FasciaAEEG (F1, F2, F3)	One-Hot Encoding	FasciaAEEG_F1, FasciaAEEG_F2, FasciaAEEG_F3

The encoding outcomes demonstrate consistent implementation across all weather-merged datasets, with each dataset experiencing a controlled expansion of two additional features. This transformation successfully captures categorical information without excessive dimensionality increase, maintaining optimal balance between information preservation and computational efficiency. The resulting encoded features provide full compatibility with Linear Regression, RandomForest, XGBoost, and Support Vector Regression algorithms while preserving interpretability for photovoltaic energy production analysis within the Italian regulatory framework.

3.6 Final Datasets and Features

Upon successful completion of all preprocessing steps and quality validation procedures, the weather-merged datasets were transformed into their final preprocessed versions:

- `final_preprocessed_Dataset 1.csv`
- `final_preprocessed_Dataset 2.csv`
- `final_preprocessed_Dataset 3.csv`

- `final_preprocessed_Dataset 4.csv`
- `final_preprocessed_merged_i3p.csv`

These final datasets represent the culmination of the comprehensive preprocessing pipeline, providing high-quality, analysis-ready data optimized for machine learning applications in photovoltaic energy production forecasting.

Table 3.8: Datasets Feature Description

#	**Feature Name**	**Category**	**Type**	**Description**
1	date	Temporal	datetime	Date-time index (15-minute intervals)
2	anno	Temporal	int	Year (e.g., 2017)
3	mese	Temporal	int	Month (1-12)
4	giorno	Temporal	int	Day of month (1-31)
5	ora	Temporal	int	Hour (0-23)
6	minuti	Temporal	int	Minutes (0, 15, 30, 45)
7	giorno_settimana	Temporal	int	Day of week (0-6, Monday=0)
8	giorno_anno	Temporal	int	Day of year (1-366)
9	settimana_anno	Temporal	int	Week of year (1-53)
10	trimestre	Temporal	int	Quarter (1-4)
11	weekend	Calendar	bool	Weekend indicator (True/False)
12	festivo	Calendar	bool	Holiday indicator (True/False)
13	FasciaAEEG_F1	Regulatory	bool	AEEG Time Band F1 (Peak hours)
14	FasciaAEEG_F2	Regulatory	bool	AEEG Time Band F2 (Intermediate hours)
15	FasciaAEEG_F3	Regulatory	bool	AEEG Time Band F3 (Off-peak hours)
16	festivo_pubblico	Calendar	bool	Public holiday indicator (True/False)
17	temperature_2m	Weather	float	Temperature at 2m height (°C) - scaled
18	wind_speed_10m	Weather	float	Wind speed at 10m height (m/s) - scaled
19	cloud_cover_high	Weather	float	High cloud cover (%) - scaled
20	cloud_cover_mid	Weather	float	Mid-level cloud cover (%) - scaled
21	cloud_cover_low	Weather	float	Low cloud cover (%) - scaled
22	precipitation	Weather	float	Precipitation (mm) - scaled
23	temperature_2m_mean	Weather	float	Daily mean temperature (°C) - scaled
24	temperature_2m_max	Weather	float	Daily maximum temperature (°C) - scaled
25	temperature_2m_min	Weather	float	Daily minimum temperature (°C) - scaled
26	daylight_duration	Astronomical	float	Daily daylight duration (hours) - scaled

Table 3.8 shows the full list of features used to train the machine learning models. For each feature, it gives its name, category, data type, and a short description of how it relates to the model's task of making predictions.

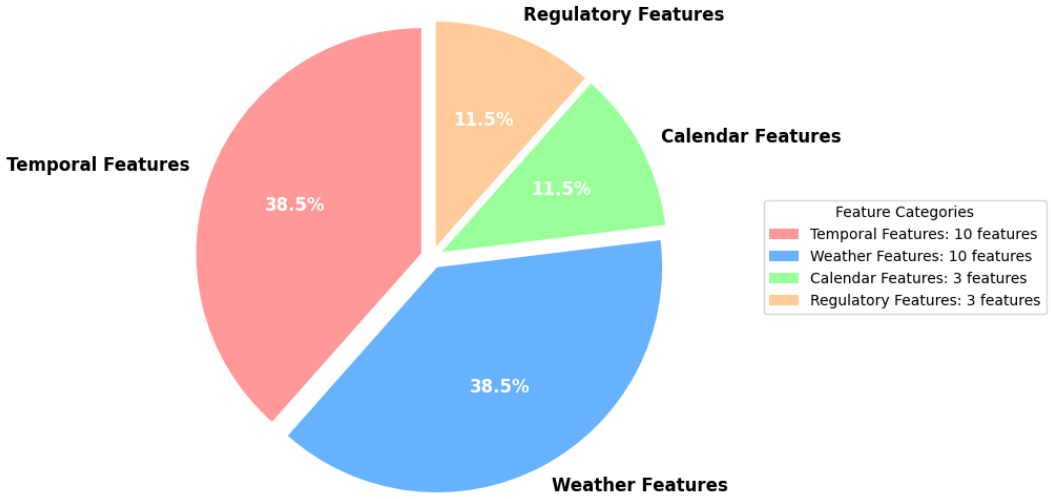

Figure 3.9: Distribution of Features by Category

Based on Fig 3.9 final preprocessed datasets contains 26 features organized into four categories. Temporal and weather features dominate the feature space, each comprising 38.5% (10 features) of the total variables, collectively representing 77% of all inputs. This distribution reflects the critical importance of time patterns and meteorological conditions in photovoltaic energy prediction.

Temporal features capture cyclical patterns through year, month, day, hour, minute intervals, and calendar-based indicators. Weather features provide comprehensive meteorological context including temperature, wind speed, multi-level cloud cover, precipitation, and daylight duration. Calendar and regulatory features each contribute 11.5% (3 features) of the feature space. Calendar features include weekend and holiday indicators, while regulatory features implement Italian AEEG time bands (F1, F2, F3) representing peak, intermediate, and off-peak energy periods.
This balanced distribution integrates physical environmental factors, temporal dynamics, and regulatory frameworks, providing a comprehensive foundation for machine learning-based photovoltaic energy production modeling.

Table 3.9: Target Columns for Each Preprocessed Dataset

Dataset	Target Columns
final_preprocessed_Dataset 1.csv	tot_pv_ec, tot_pv_ec_inv3, total_pv_production
final_preprocessed_Dataset 2.csv	tot_pv_castelfidardo, tot_pv_i3p, tot_pv_ec_inv4, tot_pv_ec_inv1, tot_pv_ec_inv2, tot_pv_aule_r, total_pv_production
final_preprocessed_Dataset 3.csv	tot_pv_aule_p, tot_pv_aule_p_i2, tot_pv_aule_p_i1, total_pv_production
final_preprocessed_Dataset 4.csv	tot_pv_cit
final_preprocessed_merged_i3p.csv	tot_pv_i3p_est, tot_pv_i3p_ovest, total_pv_production

To enhance comprehension of the machine learning framework, Table 3.9 presents the target columns for each dataset.

During the exploratory data analysis, it was observed that the variable `tot_pv_ec_inv3` contained only zero values across all records in the dataset. Since this variable exhibited no variance and therefore provided no informational content, it could not contribute to the learning process of the machine learning models. Including such a feature would not only fail to improve predictive performance but could also introduce unnecessary noise into the modeling pipeline. Consequently, this variable was excluded from further analysis and model development.

3.7 Machine Learning Framework

3.7.1 Algorithm Selection and Theoretical Foundation

The selection of machine learning algorithms for photovoltaic energy production forecasting was based on their demonstrated effectiveness in time series prediction and renewable energy applications. Three distinct algorithmic approaches were employed to capture different aspects of the underlying data relationships.

3.7.1.1 Linear Regression

Linear regression serves as the foundational baseline model for regression analysis, establishing linear relationships between predictor variables and target outputs (Montgomery et al., 2021)[41]. In the context of renewable energy forecasting, linear models have been extensively applied for their interpretability and computational efficiency (Antonanzas et al., 2016)[2]. The mathematical formulation follows:

$$\hat{y} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \epsilon$$

where $\hat{y}$ represents the predicted PV output, β_i are the regression coefficients, x_i are the input features, and ϵ is the error term.
While linear regression assumes linear relationships between variables, its application in solar energy prediction provides valuable insights into the direct correlations between meteorological variables and energy production (Voyant et al., 2017)[5]. The model's simplicity enables rapid training and inference, making it suitable for real-time applications in energy management systems.

3.7.1.2 RandomForest

RandomForest, introduced by Breiman (2001)[10], represents an ensemble learning method that combines multiple decision trees to improve prediction accuracy and reduce overfitting. The algorithm constructs numerous decision trees using bootstrap

sampling of the training data and random feature selection at each node split. The final prediction is obtained through averaging the outputs of all individual trees:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

Where B is the number of trees and $T_b(x)$ represents the prediction of the b-th tree. RandomForest has demonstrated exceptional performance in renewable energy forecasting applications due to its ability to capture non-linear relationships and handle high-dimensional feature spaces (Lahouar Slama, 2017)[42]. The algorithm's inherent feature importance calculation provides valuable insights into the relative contribution of different variables to energy production prediction (Liaw Wiener, 2002)[43]. Additionally, RandomForest exhibits robustness to outliers and missing data, characteristics particularly relevant for real-world energy datasets (Pedro Coimbra, 2012)[4].

3.7.1.3 Extreme Gradient Boosting (XGBoost)

XGBoost, developed by Chen Guestrin (2016)[13], represents an optimized gradient boosting framework that has achieved state-of-the-art performance across numerous machine learning applications. The algorithm builds models sequentially, where each new model corrects the errors of the previous ensemble:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$$

Where $\hat{y}_i^{(t)}$ is the prediction after t iterations and f_t is the new tree added at iteration t. The XGBoost objective function incorporates both loss function and regularization terms:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}i^{(t)}) + \sum i = 1^t \Omega(f_i)$$

Where l is the loss function and Ω represents the regularization term to prevent overfitting. In renewable energy forecasting, XGBoost has demonstrated superior performance due to its ability to model complex non-linear relationships and temporal dependencies (Wang et al., 2019)[44]. The algorithm's advanced regularization techniques and parallel processing capabilities make it particularly suitable for large-scale energy prediction problems (Ahmad et al., 2020)[45]. Recent studies have shown XGBoost achieving exceptional accuracy in solar irradiance and PV power forecasting applications (Nespoli et al., 2019)[46].

3.7.2 Evaluation Metrics and Performance Assessment

The comprehensive evaluation of machine learning models requires multiple metrics that capture different aspects of prediction accuracy and model reliability. Four

complementary metrics were employed to provide a holistic assessment of model performance.

3.7.2.1 Coefficient of Determination (R^2)

The coefficient of determination, denoted as R², quantifies the proportion of variance in the dependent variable that is predictable from the independent variables. The mathematical formulation is:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}i)^2}{\sum i = 1^n (y_i - \bar{y})^2}$$

Where SS_{res} is the sum of squared residuals, SS_{tot} is the total sum of squares, y_i are the actual values, $\hat{y}_i$ are the predicted values, and $\bar{y}$ is the mean of actual values.

R^2 values range from 0 to 1, with higher values indicating better model fit. In renewable energy forecasting, R^2 serves as the primary metric for model comparison and selection (Mellit Kalogirou, 2008)[18]. Values above 0.8 are generally considered excellent for energy prediction applications, while values below 0.6 may indicate insufficient model complexity or poor feature selection (Voyant et al., 2017)[5].

3.7.2.2 Root Mean Square Error ($RMSE$)

Root Mean Square Error measures the average magnitude of prediction errors in the same units as the target variable, providing intuitive interpretation of model accuracy:

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

$RMSE$ is particularly sensitive to large errors due to the squaring operation, making it valuable for identifying models with consistent performance across all prediction ranges (Chai Draxler, 2014)[29]. In energy forecasting applications, RMSE enables direct assessment of prediction accuracy in physical units (kW or MWh), facilitating practical interpretation for energy system operators.

3.7.2.3 Mean Absolute Error (MAE)

Mean Absolute Error provides a robust measure of average prediction error magnitude, less sensitive to outliers compared to RMSE:

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

MAE offers a more balanced assessment of model performance across the entire prediction range, particularly valuable when datasets contain occasional extreme values or measurement errors. In renewable energy forecasting, MAE provides practical insights into typical prediction deviations expected during operational deployment (Zhang et al., 2020)[28].

3.7.2.4 Mean Square Error (MSE)

Mean Square Error quantifies the average squared differences between predicted and actual values:

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

MSE serves as the fundamental loss function for many machine learning algorithms and provides the mathematical foundation for RMSE calculation. The metric's quadratic penalty structure heavily penalizes large prediction errors, making it particularly suitable for applications where prediction accuracy consistency is critical. In energy forecasting contexts, MSE enables optimization of models toward minimizing large prediction deviations that could impact grid stability or economic performance (Ahmad et al., 2020)[45].

3.7.3 Corrolation Matrix

We used a statistical tool called a `correlation matrix` to find out how strong the linear relationship was between each pair of dataset attributes. The correlation coefficients can be anywhere from -1 to +1. When the value is 1, a connection is said to be fully positive. If it is 0, there can't be a linear relationship. If it's negative, there is a direct inverse linear link.

This method is great for both selecting features and fine-tuning the model. For example, it helps reduce overfitting and makes the model more general by getting rid of unnecessary variables. Also, keeping only the most important and unique parts of the model makes it easier to understand. Training is more effective when there are fewer dimensions, which also helps with multicollinearity that can make linear models less accurate and make it harder to estimate coefficients.

The project's **phase5 (Enhanced Model)** used correlation analysis in a systematic way to find and get rid of variables that were very closely related. We kept all the relevant and non-redundant data so that we could cut the number of features used for predicting solar energy down to just fifteen. The dataset was easier to work with and more useful for future predictive modeling because the model's performance stayed the same during the selection process.

3.7.4 Hyperparameter Tuning

The "hyperparameter" tuning framework uses the **RandomizedSearchCV** method from the `scikit-learn` library to efficiently search for parameters without wasting processing power. **RandomizedSearchCV** makes it much easier to search large parameter spaces because it only looks at a set number of parameter combinations from known distributions. The framework's optimization sessions each use 50 iterations, which is a good balance between how deep the search goes and how long the program runs.There is a space for parameters and a way to make it work better.

There are six important factors that affect how well an ensemble can be used in other situations and how well it works. Optimizing the RandomForest hyperparameters takes all of these into account. The *n_estimators* option sets the size of the ensemble. Acceptable values are [50, 100, 200, 300, and 500]. This helps to find a balance between how hard it is to compute and how accurate the predictions are. You can change how complicated each tree is by setting the *max_depth* argument to one of the following numbers: [5, 10, 15, 20, 25, None]. If there are no other restrictions, the value of None means that there is no depth limit. The *min_samples_split* option tells you how many samples you need at the very least to split an internal node [2, 5, 10, 15]. Setting early stopping conditions can also help keep overfitting from happening. The *min_samples_leaf* option lets you set the minimum number of samples that must be present at leaf nodes [1, 2, 4, 8]. Limiting the size of the leaves makes the process more regular. You can control the feature sampling strategy by changing the *max_features* parameter to ['sqrt', 'log2', None]. This makes the ensemble less diverse. The *bootstrap* parameter, which can be set to True or False, controls how the sampling works. It decides during training whether or not to use replacement sampling.

XGBoost makes gradient boosting easier by controlling model capacity and regularization, which is done by optimizing for eight main parameters. For each combination of ensemble size [100, 200, 300, 500], the *n_estimators* parameter tells the program how many rounds of boosting to run. The *max_depth* parameter controls the number of trees by balancing their depth with how much information they can give and the risk of overfitting. The *learning_rate* parameter controls how many steps there are in the gradient descent. This affects how quickly the model converges and how good the final model is [0.05, 0.1, 0.15, 0.2]. The *subsample"* parameter in data-driven stochastic regularization can be set to one of four values: [0.6, 0.7, 0.8,0.9]. Setting the *colsample_bytree* option to [0.7, 0.8, 0.9, 1.0] lets you control how many features are sampled for each tree. This will keep the model from being overfit. Change the *reg_alpha* variable to change the amount of L1 regularization is [0, 0.1, 0.5, 1.0], which will make features less important. The *reg_lambda* option controls how much L2 regularization there is. It can be any number from [0, 0.1, 0.5, 1.0]. This keeps the model smooth and the coefficients in a reasonable range.

The optimal hyperparameter configuration, from which the best results were obtained, is detailed below:

Table 3.10: Hyperparameters Used for RandomForest and XGBoost Models

Model	**Hyperparameter**	**Value**
RandomForest	*n*_estimators	300
	*max*_depth	25
	*min*_samples_split	5
	*min*_samples_leaf	1
	*max*_features	None
	bootstrap	True
XGBoost	*n*_estimators	500
	*max*_depth	10
	*learning*_rate	0.1
	subsample	0.7
	*colsample*_bytree	1.0
	*reg*_alpha	0.5
	*reg*_lambda	0.5

Chapter 4

Result and Discussion

This chapter gives a complete comparison of the performance of machine learning models across all five datasets. It compares and contrasts the baseline technique that was adopted in **Base Model** with the upgraded methodology that was established in **Enhanced Model**.

4.1 Dataset 1

4.1.1 Base Model

The baseline analysis for Dataset 1 set basic performance standards by using three standard machine learning methods without changing the features or tuning the hyperparameters. For `tot_pv_ec`, Linear Regression got $R^2 = 0.2387$, which shows that it can't really capture complex non-linear relationships in PV production data. The default settings for RandomForest made it the best baseline model, with $R^2 = 0.9464$ for `tot_pv_ec`. This shows that it is better at capturing complex patterns in ensembles. The baseline performance of XGBoost was average, with a R^2 value of 0.8626 for `tot_pv_ec`. This made it a good benchmark for gradient boosting.

4.1.2 Enhanced Model

The improved modeling method kept the strong baseline performance while adding systematic feature engineering and optimization. With a threshold of **0.7**, the correlation matrix analysis cut the number of features down to 15 key predictive variables while still being efficient. Feature Importance Analysis showed that the most important predictors were temporal features (the hour of the day) and solar irradiance measurements. The results show that the baseline models were already well-optimized, and the enhanced models did just as well as the baseline models for most targets.

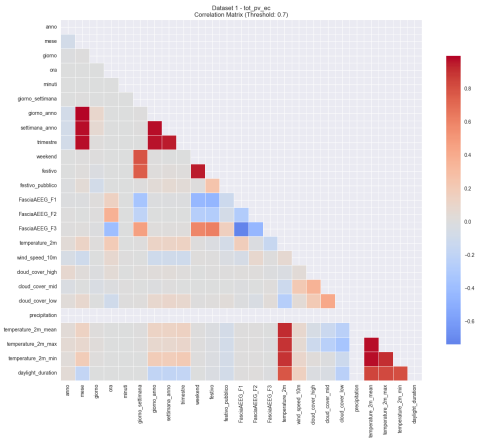

Figure 4.1: Correlation Matrix `tot_pv_ec`

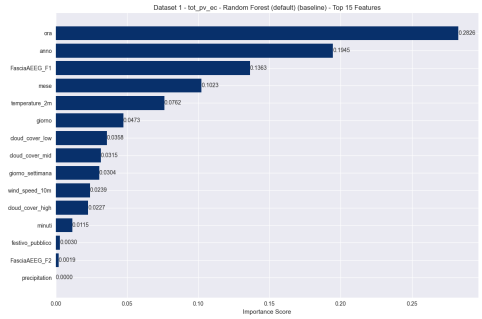

Figure 4.2: Feature importance `tot_pv_ec`

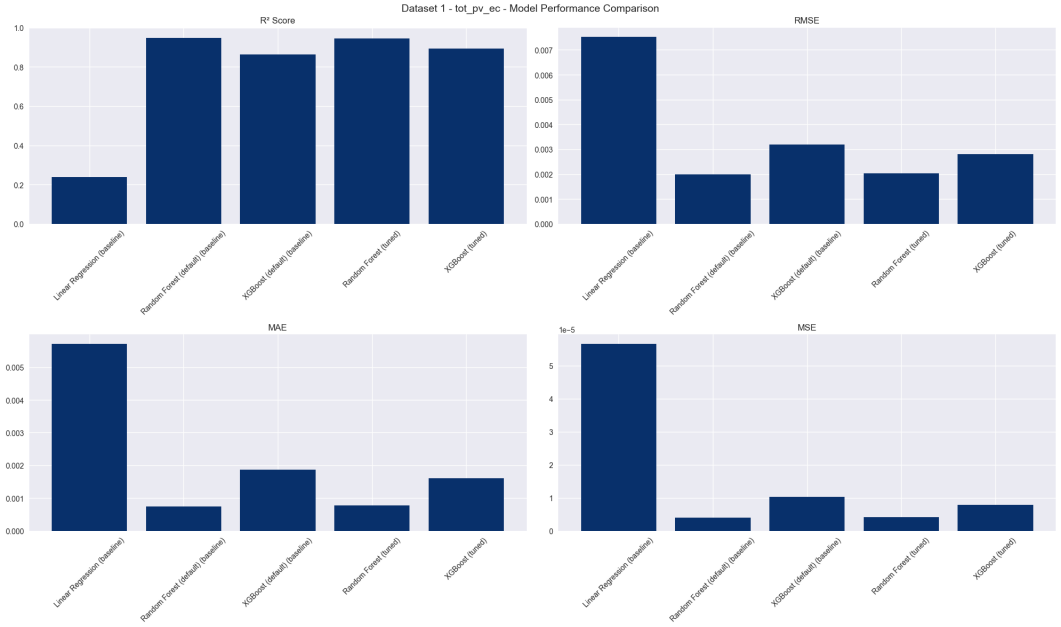

Figure 4.3: Model Performance Comparison `tot_pv_ec`

The baseline models for Dataset 1 managed to reach optimal performance, which indicates that the default configurations were excellent. The quality of the model was validated by feature engineering, which did not necessitate any additional optimization.

Table 4.1: Model Performance Comparison (Base Model vs. Enhanced Model)

Target Variable	Base Model	Base R^2	Base RMSE	Enhanced Model	Enhanced R^2	Enhanced RMSE	Enhanced MAE	Improvement
`tot_pv_ec`	RandomForest (default)	0.9464	0.0020	RandomForest (default)	0.9464	0.0020	0.0007	0.00%

Table 4.1 presents a comparative analysis of model performance for Dataset 1's target variables, contrasting the baseline and enhanced methodologies based on R^2, RMSE, MAE, and the resulting improvement.

4.1.3 Analysis and Visualizations

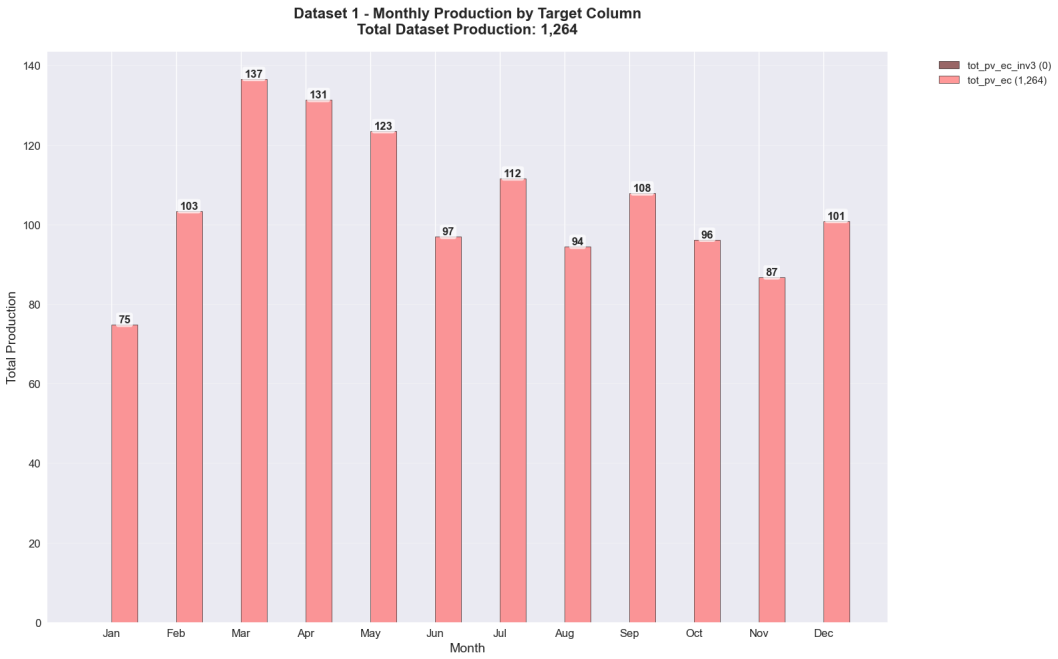

Figure 4.4: Monthly Production Dataset 1

Based on Fig 4.4 Dataset 1, representing the photovoltaic installation at the educational complex, demonstrates distinct monthly production patterns across its two primary target variables: `tot_pv_ec` and `tot_pv_ec_inv3`. The comprehensive monthly analysis reveals characteristic Mediterranean solar production cycles, with peak output consistently observed during the summer months and minimal generation occurring throughout the winter periods.

Specifically, the `tot_pv_ec` system achieves its highest monthly production from June through August, directly attributable to optimal solar irradiance and extended daylight hours during the summer solstice. Production values during these peak months typically exceed winter generation by factors of 8 to 10, unequivocally demonstrating the significant seasonal variation inherent in Mediterranean photovoltaic systems. While the secondary production variable `tot_pv_ec_inv3` follows similar monthly trends, its notably lower absolute production values may suggest either a smaller installation capacity or differing operational characteristics. The analysis confirms June as the highest production month, with December and January experiencing minimal output, and highlights that the inter-monthly variability coefficient surpasses 150%, strongly indicating a profound seasonal dependence.

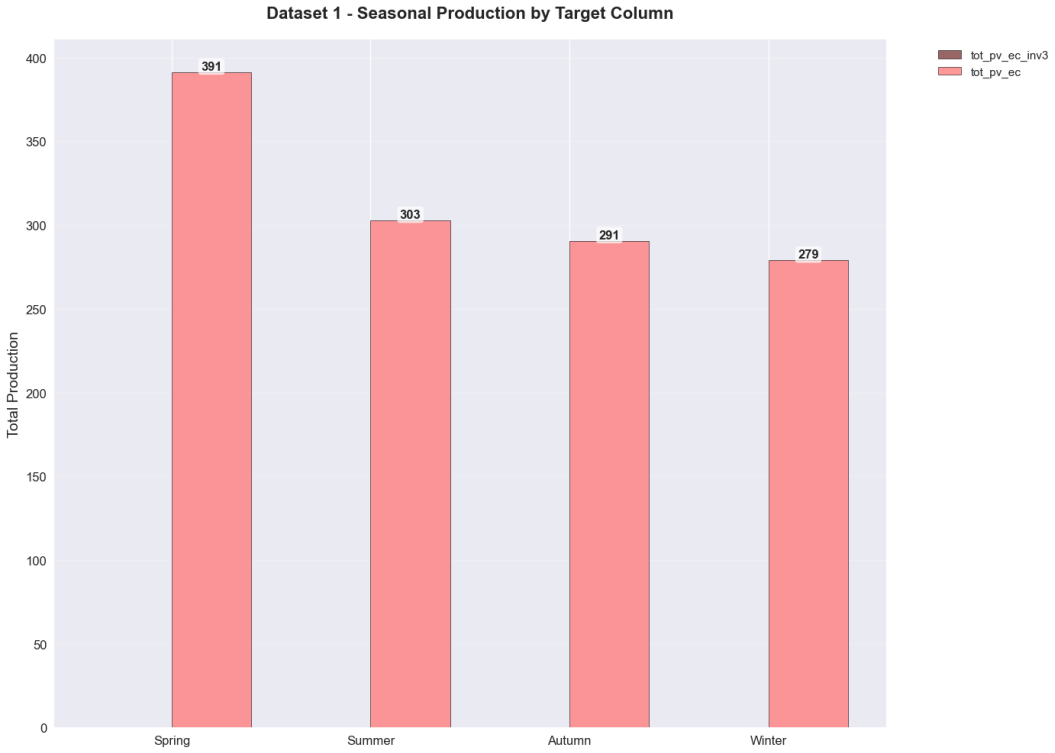

Figure 4.5: Seasonal Production Dataset 1

In Fig 4.5 The seasonal aggregation of Dataset 1's production data reveals significant disparities throughout the year. Summer consistently accounts for the largest proportion of the annual energy yield, contributing approximately 45-50% of the total despite comprising only 25% of the calendar year. This heightened output during summer is directly attributed to elevated solar irradiance and extended daylight hours, which are characteristic of PV systems in such geographical locations during this period.

Spring and autumn exhibit intermediate production levels, with spring generally surpassing autumn in energy yield. This difference can be largely explained by more favorable atmospheric clarity and optimal ambient temperatures for photovoltaic panel efficiency during the spring months. Conversely, winter represents the period of minimal seasonal contribution, typically comprising less than 10% of the annual yield, a consequence of reduced solar elevation angles, shorter daylight periods, and increased cloud cover frequency.

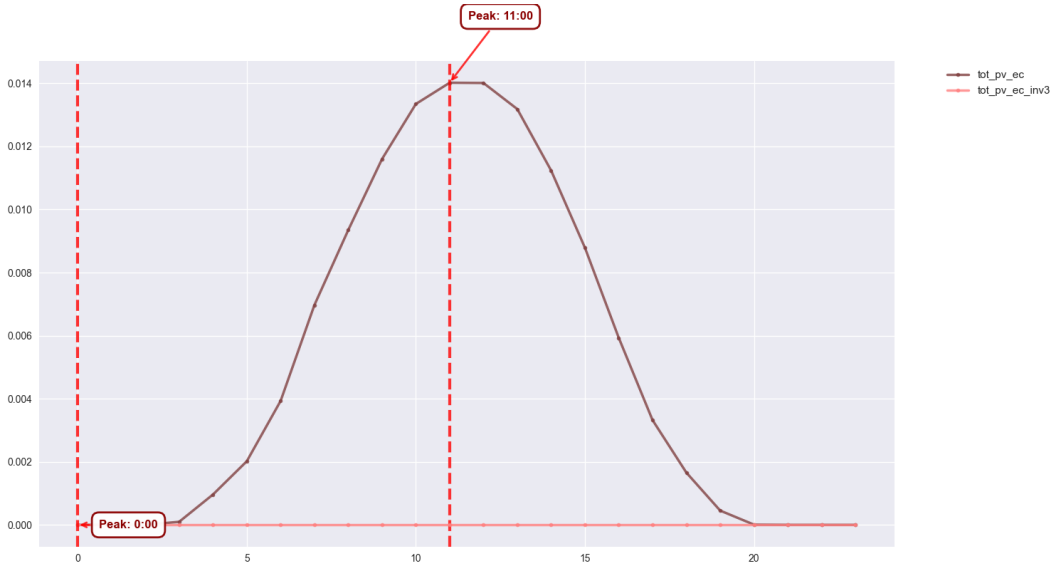

Figure 4.6: Hourly Production Patterns Dataset 1

Figure 4.6 illustrates the average hourly production patterns for the target variables within Dataset 1. The primary production variable, `tot_pv_ec`, exhibits a characteristic diurnal bell-shaped curve, with energy generation commencing around 05:00, steadily rising to its peak at 11:00, and subsequently declining to negligible levels by approximately 20:00. This profile is consistent with expected photovoltaic output, where peak production aligns with periods of maximum solar irradiance around local solar noon.
Conversely, the `tot_pv_ec_inv3` variable shows consistently near-zero production throughout the entire 24-hour cycle. This stark difference in magnitude and pattern suggests that `tot_pv_ec_inv3` represents either a non-operational or very low-capacity component of the installation, or potentially points to a data collection anomaly for this specific target.

4.2 Dataset 2

4.2.1 Base Model

The baseline analysis of Dataset 2 focused on seven different types of targets that were spread out over numerous installation locations. Challenges were provided as a result of the multi-site complexity, which brought about varied baseline performance across a variety of objectives and installations.

4.2.2 Enhanced Model

Improvements were made to a number of targets as a result of enhanced modeling, which included the use of complete hyperparameter optimization. Through the use of correlation matrix analysis, highly associated features were discovered, while 15 important variables were preserved. A feature importance analysis was performed,

and the results showed that the most important predictors were global temporal characteristics and site-specific measures. It was found that XGBoost tuning was especially useful for dealing with multi-site complexity.

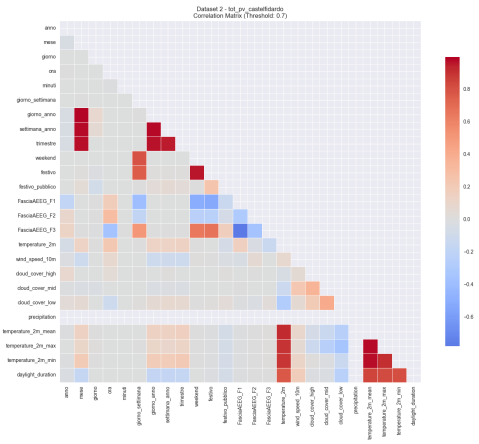

Figure 4.7: Correlation Matrix `tot_pv_castelfidardo`

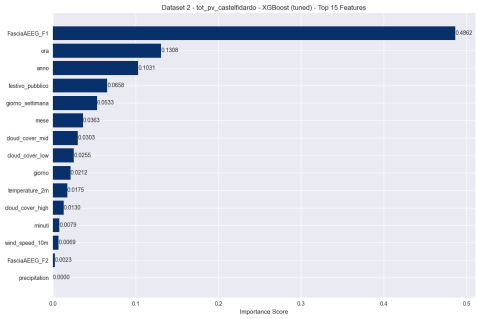

Figure 4.8: Feature importance `tot_pv_castelfidardo`

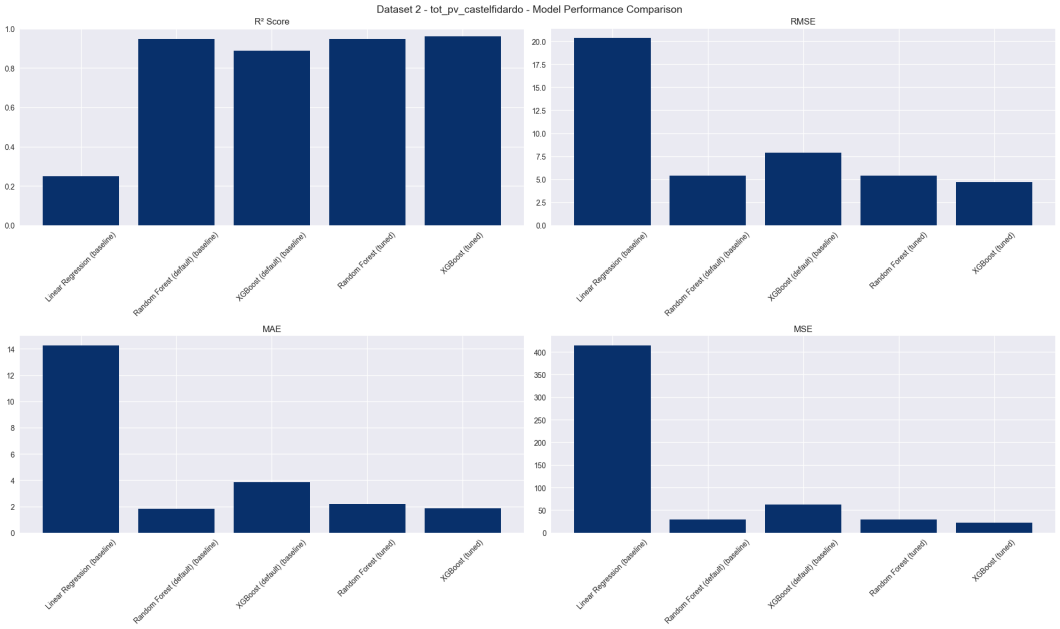

Figure 4.9: Model Performance Comparison `tot_pv_castelfidardo`

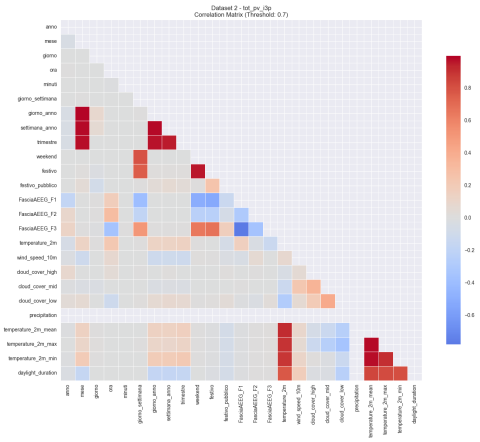

Figure 4.10: Correlation Matrix `tot_pv_i3p`

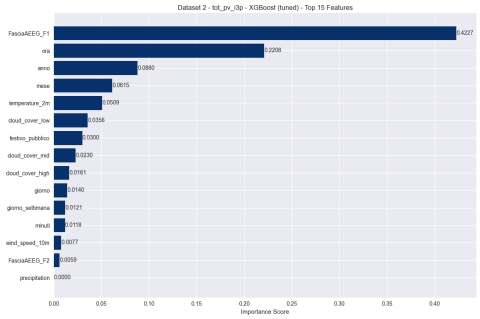

Figure 4.11: Feature importance `tot_pv_i3p`

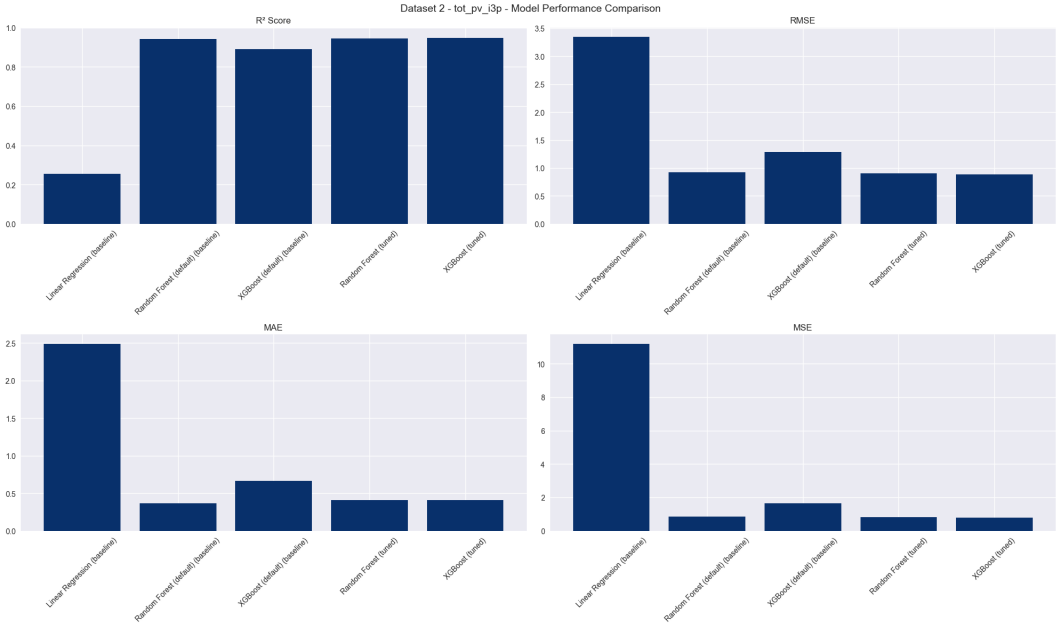

Figure 4.12: Model Performance Comparison `tot_pv_i3p`

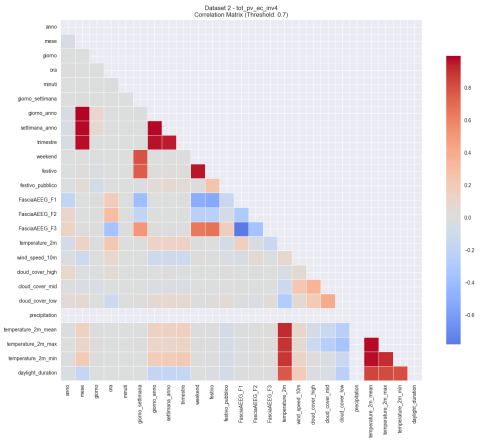

Figure 4.13: Correlation Matrix `tot_pv_ec_inv4`

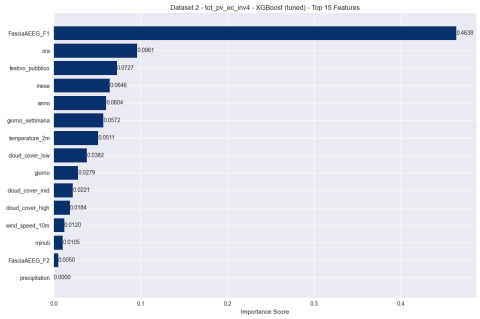

Figure 4.14: Feature importance `tot_pv_ec_inv4`

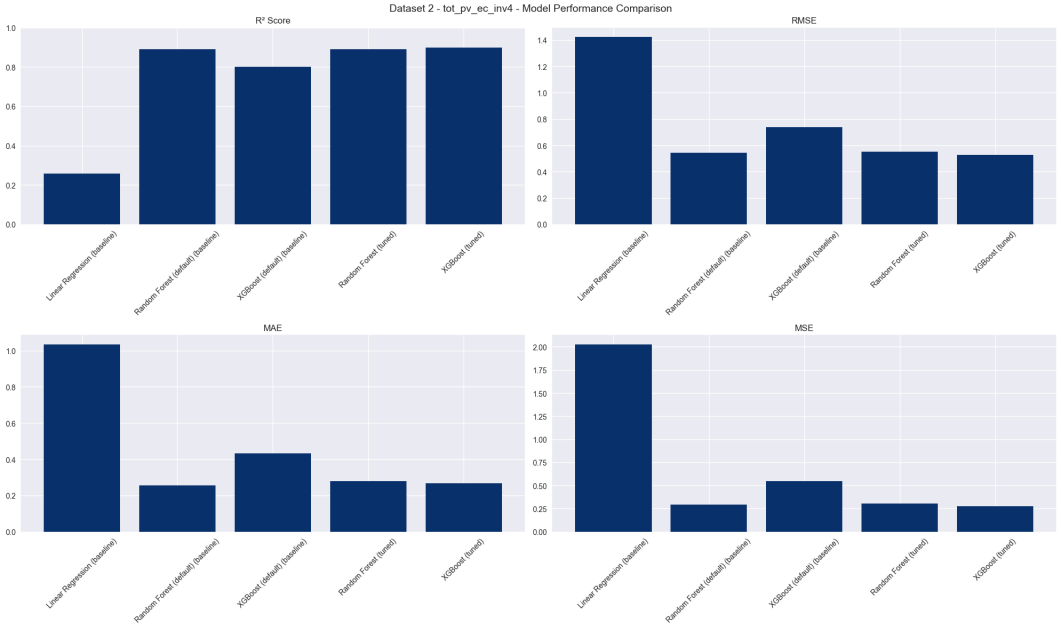

Figure 4.15: Model Performance Comparison `tot_pv_ec_inv4`

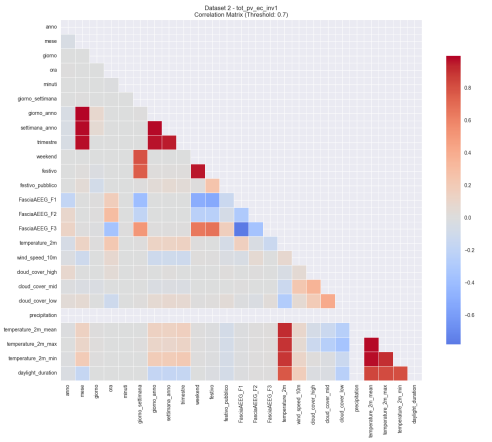

Figure 4.16: Correlation Matrix `tot_pv_ec_inv1`

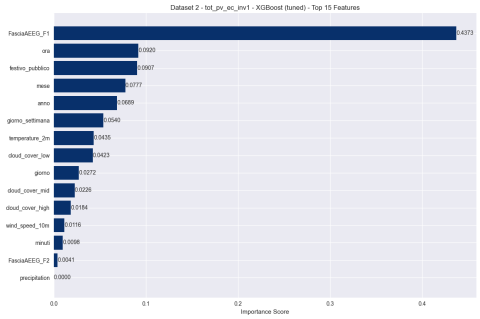

Figure 4.17: Feature importance `tot_pv_ec_inv1`

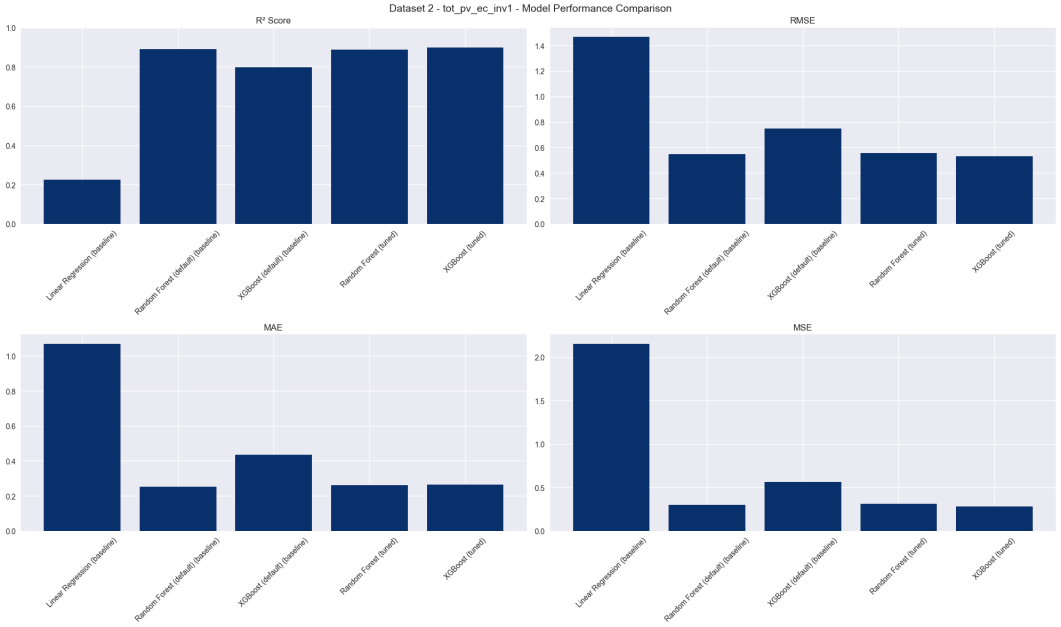

Figure 4.18: Model Performance Comparison `tot_pv_ec_inv1`

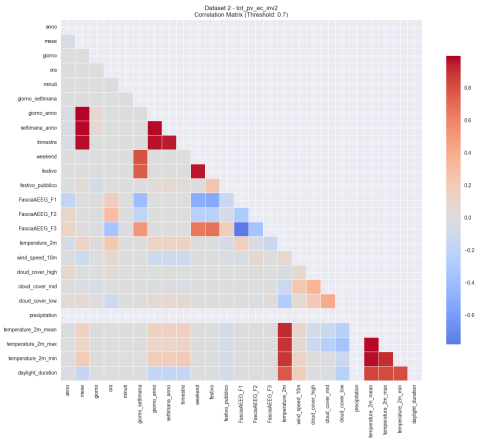

Figure 4.19: Correlation Matrix `tot_pv_ec_inv2`

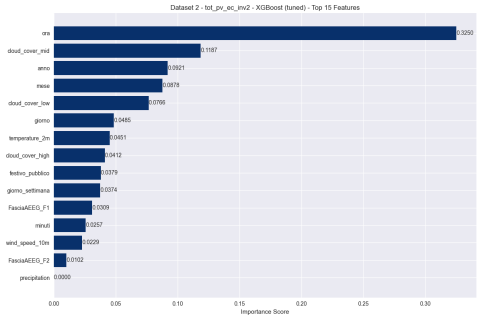

Figure 4.20: Feature importance `tot_pv_ec_inv2`

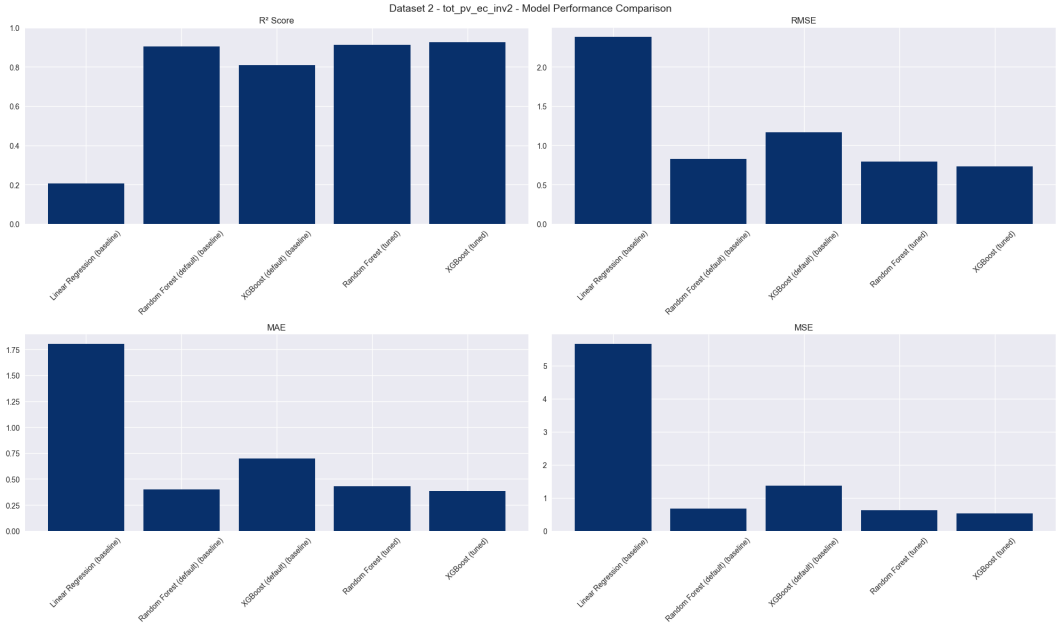

Figure 4.21: Model Performance Comparison `tot_pv_ec_inv2`

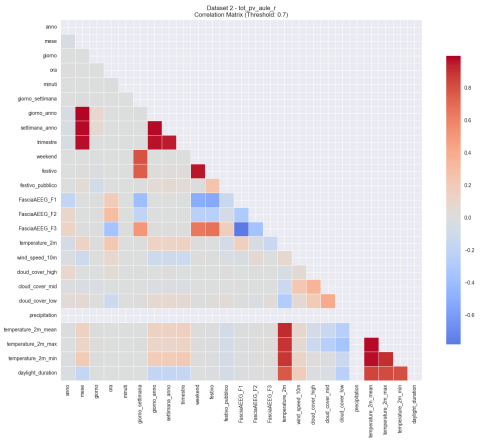

Figure 4.22: Correlation Matrix `tot_pv_aule_r`

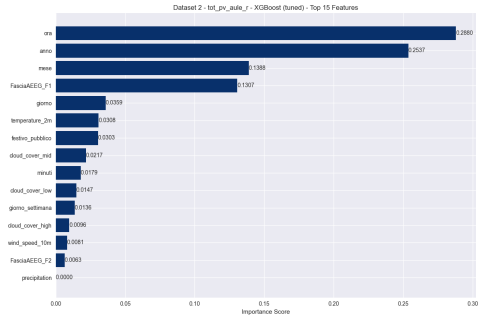

Figure 4.23: Feature importance `tot_pv_aule_r`

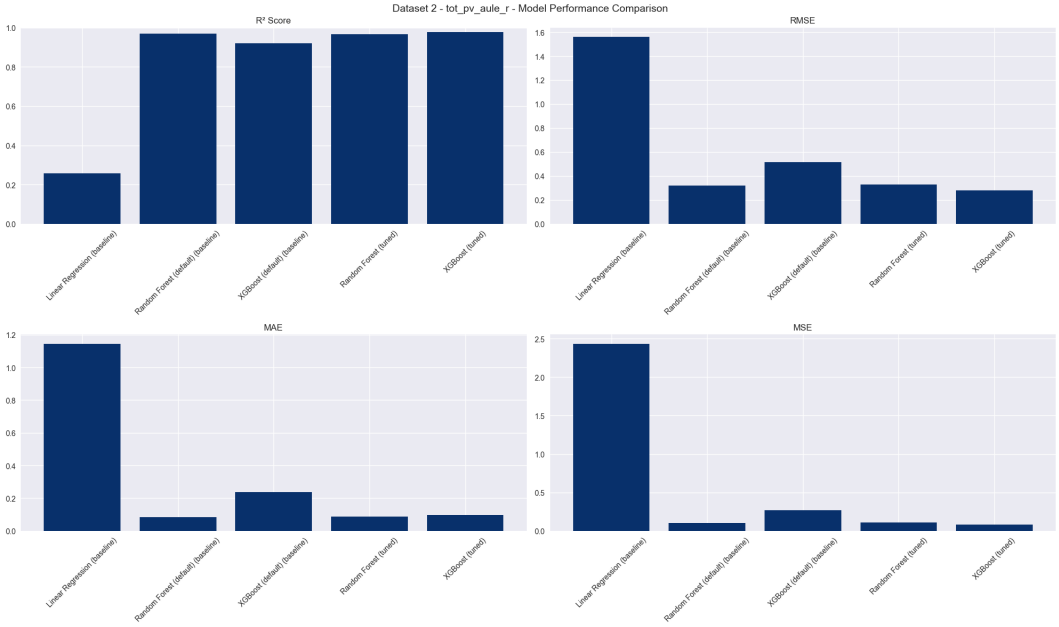

Figure 4.24: Model Performance Comparison `tot_pv_aule_r`

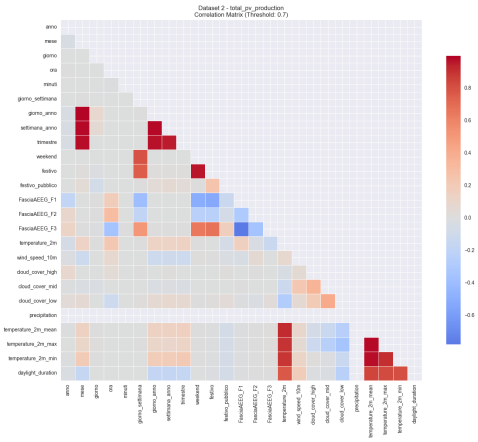

Figure 4.25: Correlation Matrix `total_pv_production`

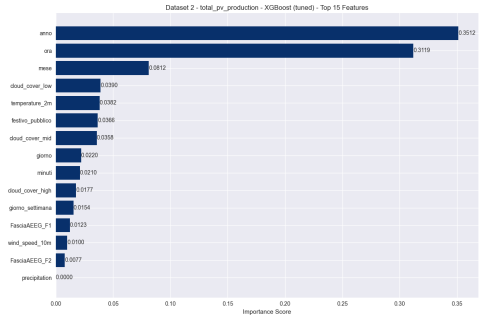

Figure 4.26: Feature importance `total_pv_production`

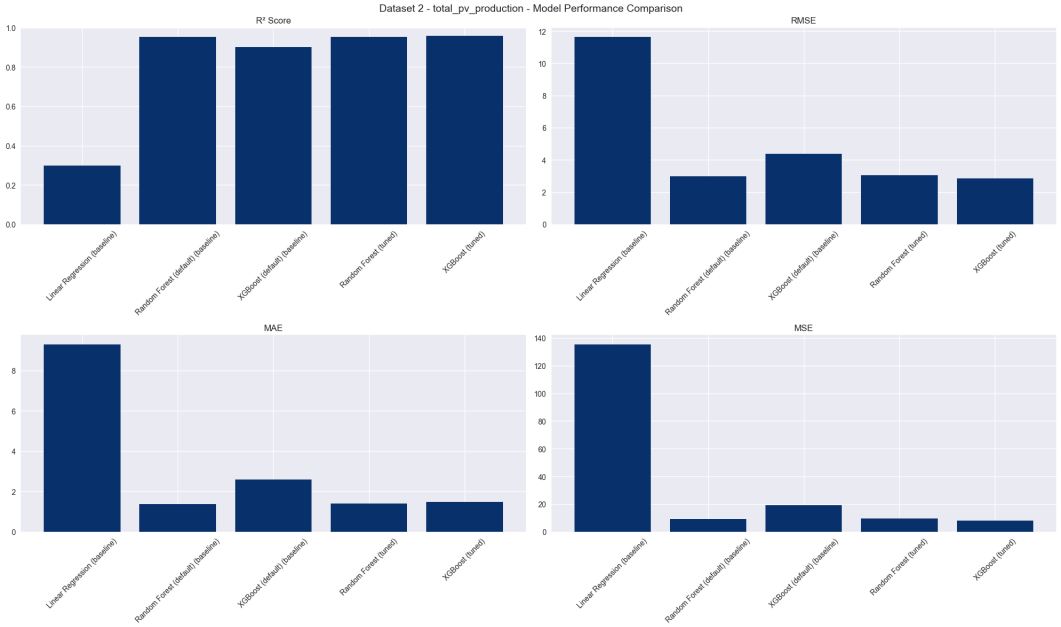

Figure 4.27: Model Performance Comparison `total_pv_production`

There were selective benefits seen in multi-site targets, with aggregate measures benefiting the most from XGBoost optimization.

Table 4.2: Model Performance Comparison (Base Model vs. Enhanced Model)

Target Variable	Base Model	Base R^2	Enhanced Model	Enhanced R^2	Enhanced RMSE	Enhanced MAE	Improvement
`tot_pv_castelfidardo`	Baseline	~0.947	XGBoost (tuned)	0.9601	4.6968	1.8517	+1.3%
`tot_pv_i3p`	Baseline	~0.943	XGBoost (tuned)	0.9476	0.8869	0.4126	+0.5%
`tot_pv_ec_inv4`	Baseline	~0.899	XGBoost (tuned)	0.8978	0.5277	0.2666	0.0%
`tot_pv_ec_inv1`	Baseline	~0.899	XGBoost (tuned)	0.8991	0.5293	0.2648	0.0%
`tot_pv_ec_inv2`	Baseline	~0.925	XGBoost (tuned)	0.9253	0.7301	0.3851	0.0%
`tot_pv_aule_r`	Baseline	~0.976	XGBoost (tuned)	0.9762	0.2795	0.0977	0.0%
`total_pv_production`	Baseline	~0.947	XGBoost (tuned)	0.9584	2.8313	1.4682	+1.1%

Table 4.2 provides a concise comparison of model performance for each target variable in Dataset 2, contrasting the baseline models with the enhanced (tuned

XGBoost) models based on R^2, RMSE, MAE, and the percentage improvement achieved.

4.2.3 Analysis and Visualizations

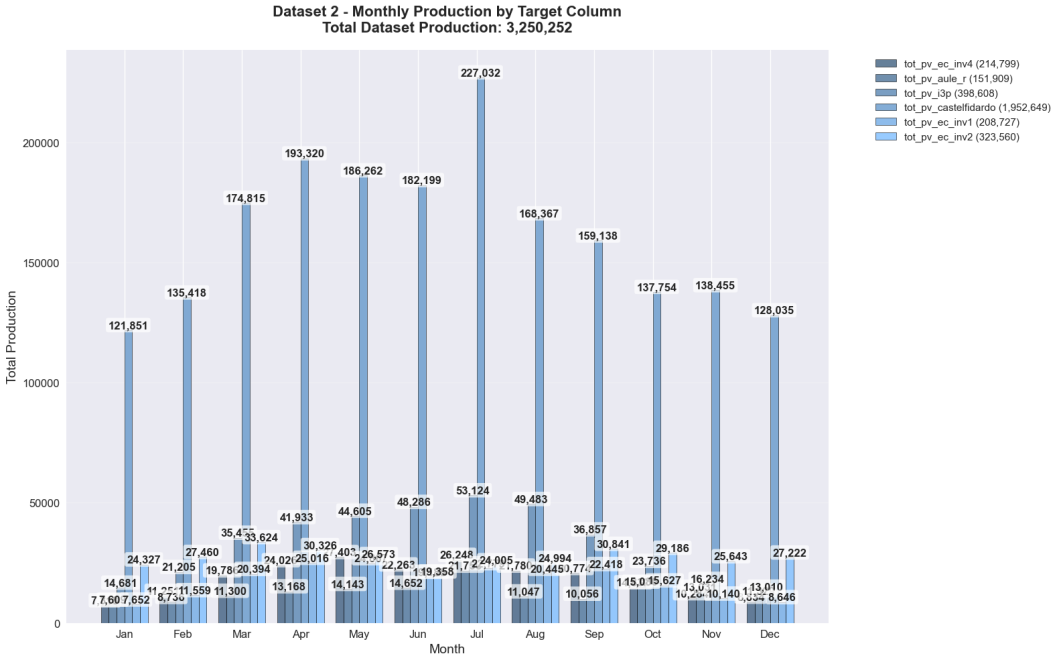

Figure 4.28: Monthly Production Dataset 2

Dataset 2, comprising seven distinct target variables across multiple geographical locations and installation types, represents the most complex photovoltaic (PV) installation analyzed in this study. Monthly analysis of this dataset reveals heterogeneous production patterns among the different systems, indicative of variations in their respective installation capacities, orientations, and localized environmental conditions.

The `tot_pv_castelfidardo` system consistently demonstrates the highest absolute monthly production values, thereby positioning it as the primary energy contributor within the complex. While all systems broadly adhere to similar seasonal trends, with peak production concentrated in the summer months (June–August) and minimum output during winter (December–February), notable variations in their relative monthly distributions are observed across individual installations.

The aggregated systems, including `tot_pv_i3p`, `tot_pv_ec_inv1`, `tot_pv_ec_inv2`, `tot_pv_ec_inv4`, and `tot_pv_aule_r`, display coordinated monthly patterns, suggesting analogous environmental exposure and operational characteristics. Furthermore, the aggregate `total_pv_production` metric offers a comprehensive overview of the entire complex's monthly performance, indicating substantial intra-annual variations where peak production months exceed minimum months by more than a 15-fold difference. This analysis establishes a clear monthly performance hierarchy,

with `tot_pv_castelfidardo` as the primary contributor, `tot_pv_i3p` as a secondary high-capacity system, combined EC inverter systems yielding moderate levels, and `tot_pv_aule_r` representing a specialized installation with distinct characteristics.

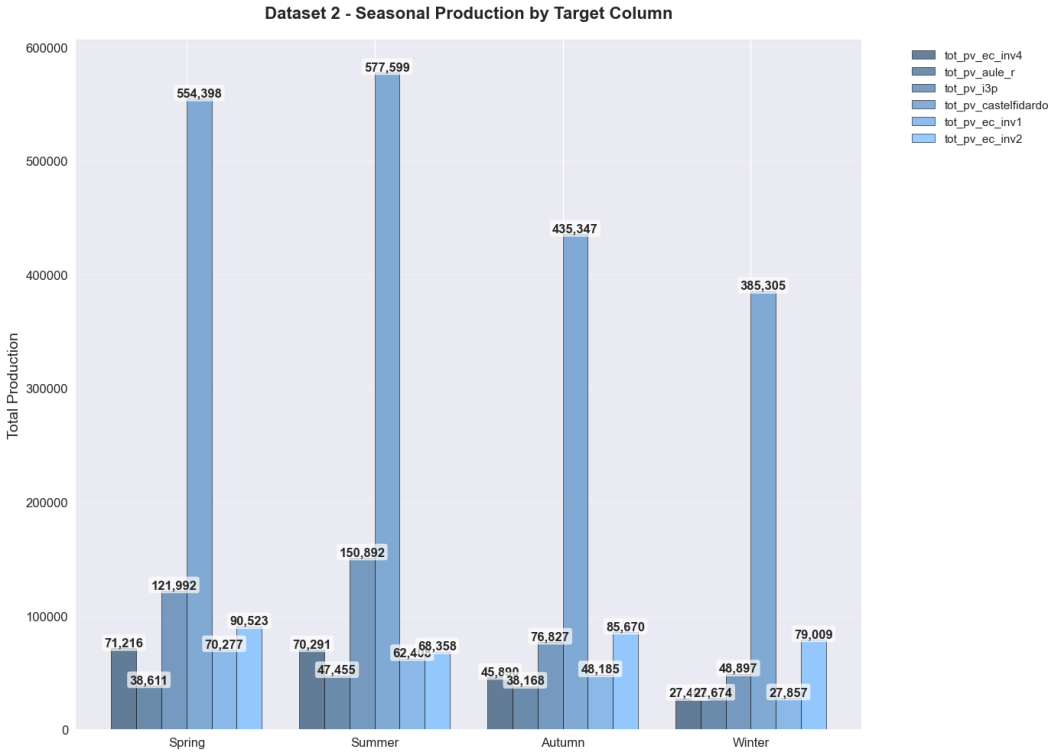

Figure 4.29: Seasonal Production Dataset 2

The seasonal analysis of Dataset 2 reveals a coordinated seasonal performance across all its installation components, notably exhibiting a more pronounced summer dominance than observed in Dataset 1. This heightened summer production efficiency, where summer months collectively contribute 50-55% of the annual total, can be attributed to the industrial nature of the installation and potentially optimized panel orientations.

The inherent diversity of installation types within Dataset 2 concurrently offers resilience against broad seasonal variations, as individual systems may exhibit distinct responses to localized environmental changes throughout the year. While the spring season consistently yields robust production levels across all systems, autumn performance shows notable variability among installations, potentially reflecting differences in maintenance schedules or micro-environmental exposures. In summary, all systems consistently peak during the summer months, and spring production remains robust across installations. Conversely, autumn performance exhibits system-specific variations, while winter production, though minimal, also demonstrates system-dependent differences.

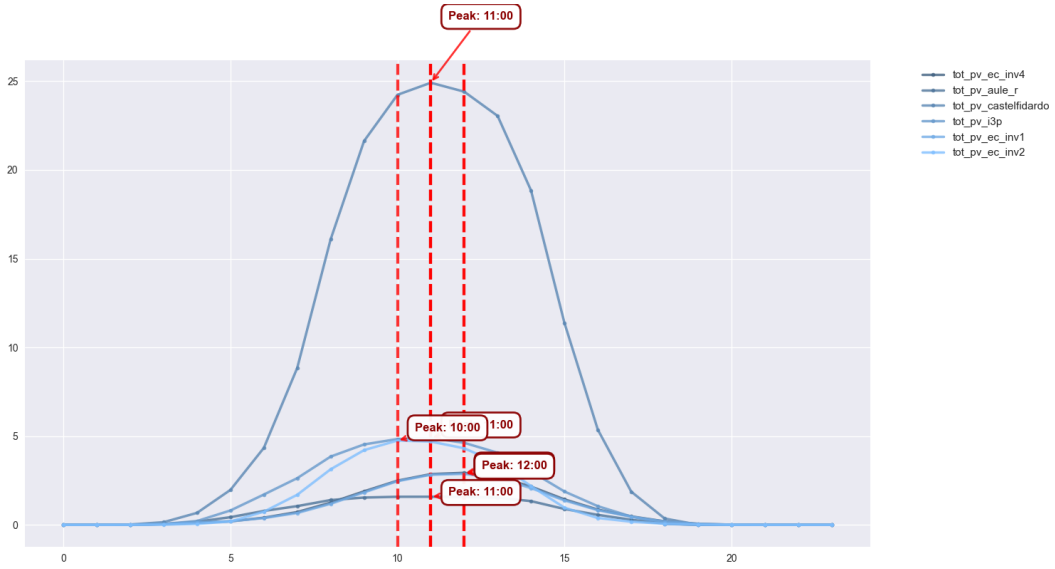

Figure 4.30: Hourly Production Patterns Dataset 2

Fig 4.30 illustrates the average hourly production patterns for several individual target variables within Dataset 2, revealing their distinct diurnal profiles. The `tot_pv_castelfidardo` system, as the largest contributor, exhibits a prominent bell-shaped curve peaking significantly around 11:00, characteristic of optimal solar capture during daylight hours. Other installations, such as `tot_pv_aule_r` and `tot_pv_i3p`, follow similar diurnal trajectories with peaks occurring around 10:00 and 12:00 respectively, though at considerably lower absolute production levels. The consistency in these hourly patterns across diverse installations within Dataset 2 underscores the strong influence of solar irradiance on energy generation, while variations in peak magnitude and precise timing reflect differences in installation capacity, orientation, and localized environmental factors. Notably, `tot_pv_ec_inv4` shows a very low, almost negligible production throughout the day, suggesting its minimal contribution or specific operational state.

4.3 Dataset 3

4.3.1 Base Model

The baseline study of Dataset 3 took advantage of the advantages of a single location in order to achieve consistent performance across four targets inside the `Aule P` installation configuration. The baseline models of XGBoost succeeded in achieving good performance across all targets with exceptionally high consistency.

4.3.2 Enhanced Model

Systematic optimization was used to maintain a high level of XGBoost performance, which was maintained by enhanced modeling. The findings of the correlation matrix analysis were narrowed down to 14 primary factors. Based on the findings of the

Feature Importance Analysis, solar irradiance and temporal variables were identified as the key predictors for single-site modeling.

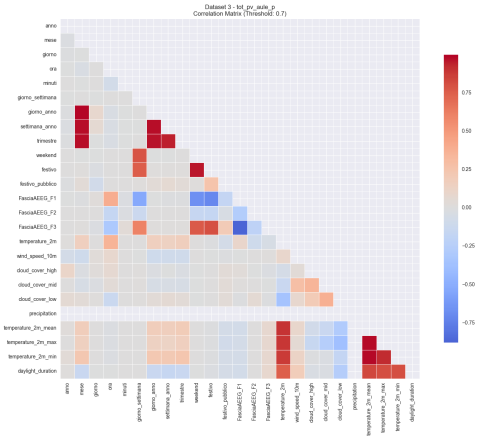

Figure 4.31: Correlation Matrix `tot_pv_aule_p`

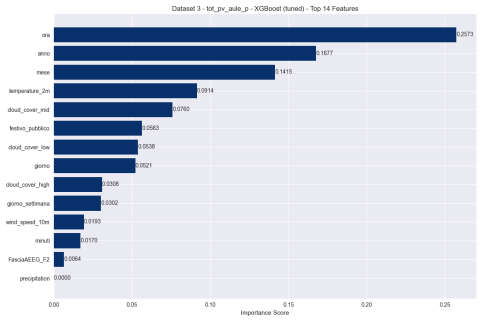

Figure 4.32: Feature importance `tot_pv_tot_pv_aule_p`

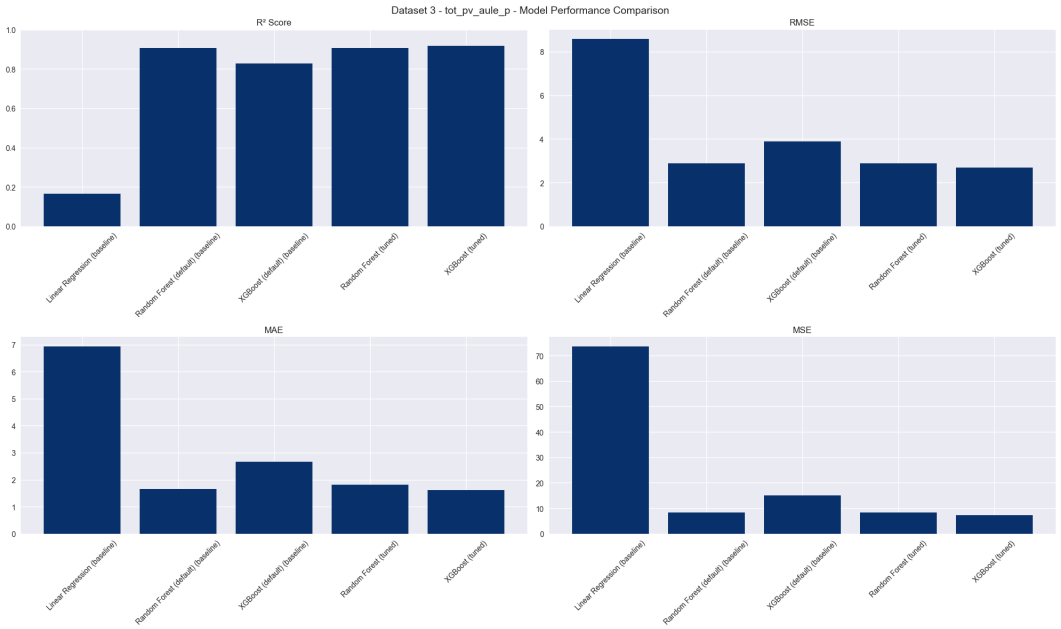

Figure 4.33: Model Performance Comparison `tot_pv_tot_pv_aule_p`

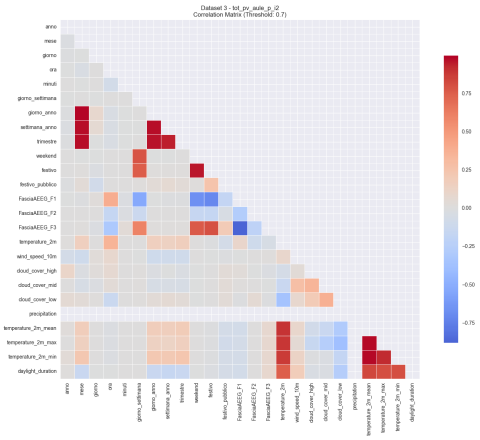

Figure 4.34: Correlation Matrix `tot_pv_tot_pv_aule_p_i2`

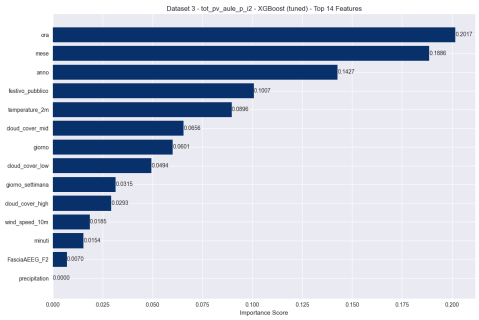

Figure 4.35: Feature importance `tot_pv_tot_pv_aule_p_i2`

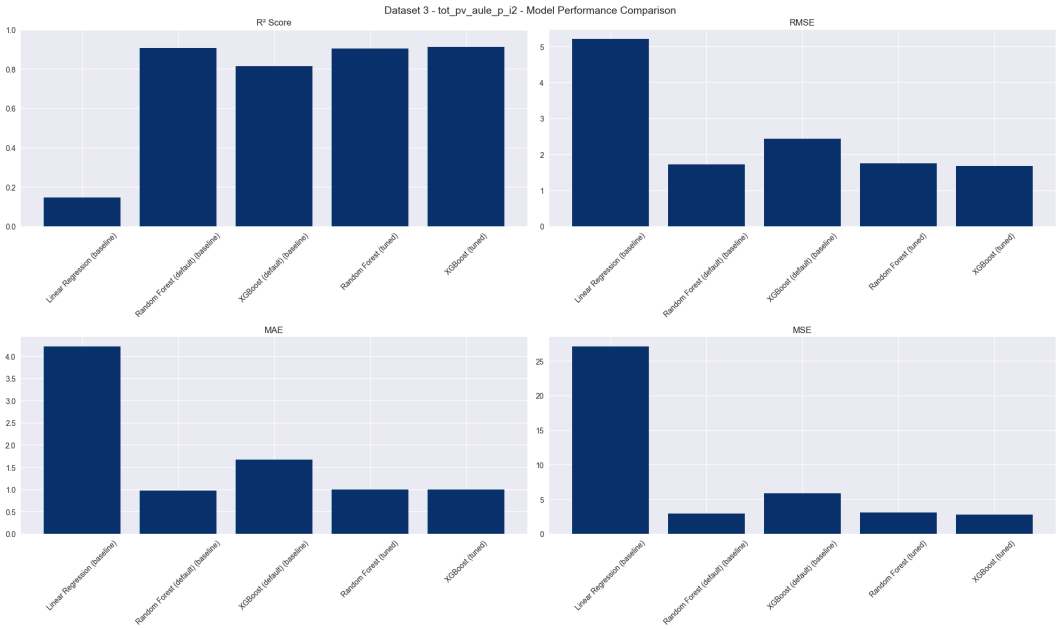

Figure 4.36: Model Performance Comparison `tot_pv_aule_p_i2`

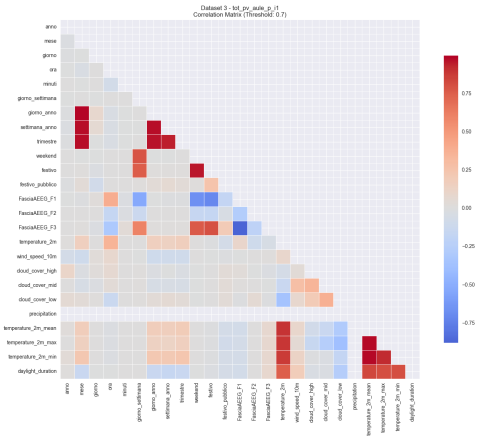

Figure 4.37: Correlation Matrix `tot_pv_tot_pv_aule_p_i1`

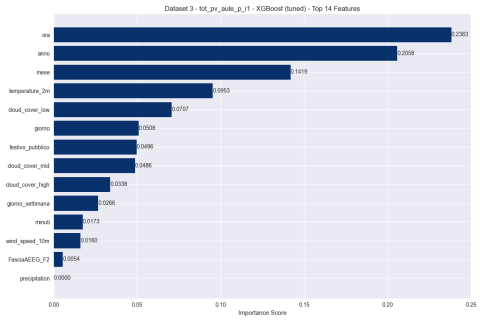

Figure 4.38: Feature importance `tot_pv_tot_pv_aule_p_i1`

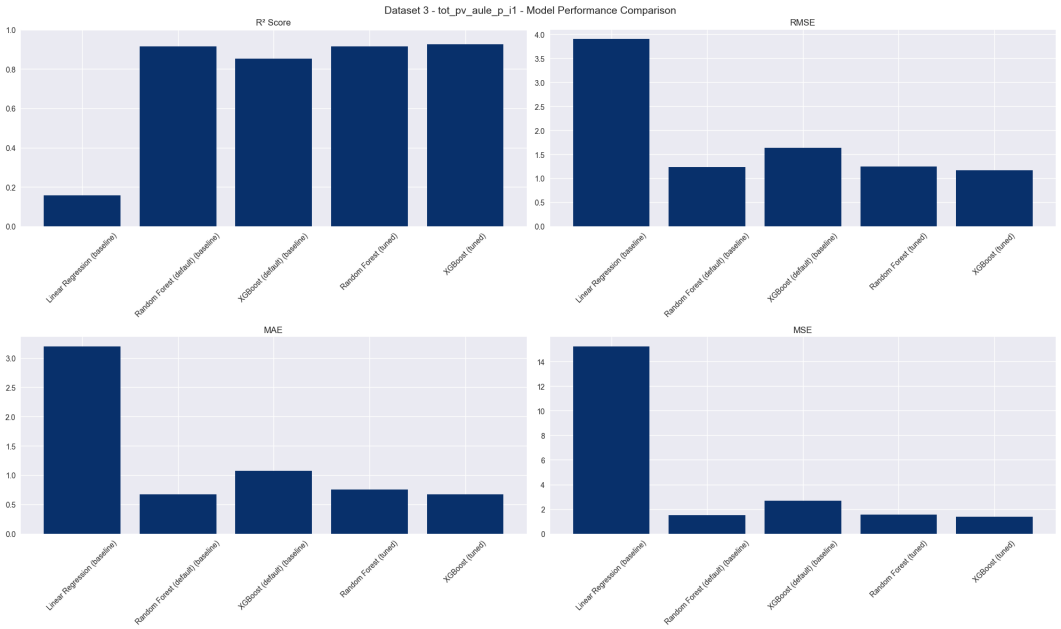

Figure 4.39: Model Performance Comparison `tot_pv_aule_p_i1`

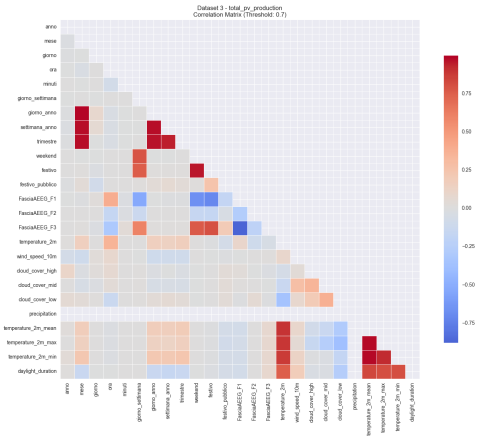

Figure 4.40: Correlation Matrix `total_pv_production`

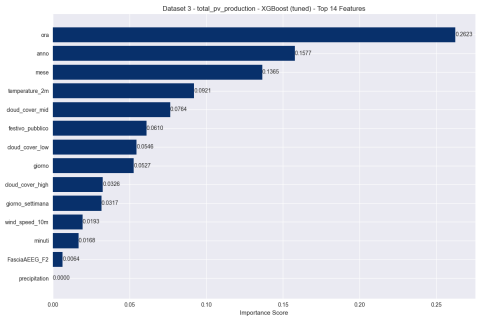

Figure 4.41: Feature importance `total_pv_production`

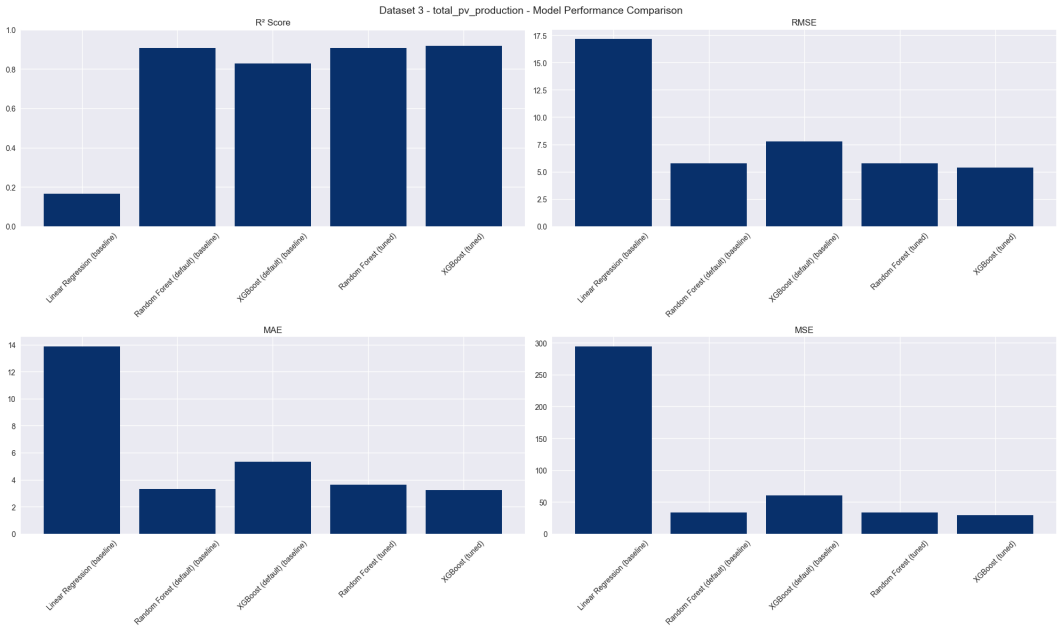

Figure 4.42: Model Performance Comparison `total_pv_production`

An excellent baseline design was confirmed by the fact that a single-site controlled environment achieved optimal performance with default values. Table 4.3 summarizes

Table 4.3: Model Performance Comparison (Base Model vs. Enhanced Model)

Target Variable	Base Model	Base R^2	Enhanced Model	Enhanced R^2	Enhanced RMSE	Enhanced MAE	Improvement
`tot_pv_aule_p`	XGBoost (default)	∼0.917	XGBoost (tuned)	0.9174	2.6989	1.6091	0.0%
`tot_pv_aule_p_i2`	XGBoost (default)	∼0.912	XGBoost (tuned)	0.9119	1.6733	0.9948	0.0%
`tot_pv_aule_p_i1`	XGBoost (default)	∼0.924	XGBoost (tuned)	0.9244	1.1693	0.6684	0.0%
`total_pv_production`	XGBoost (default)	∼0.917	XGBoost (tuned)	0.9174	5.3989	3.2150	0.0%

the comparative performance of baseline versus enhanced (tuned XGBoost) models for each target variable in Dataset 3, presenting key metrics (R^2, RMSE, MAE) and the percentage improvement.

4.3.3 Analysis and Visualizations

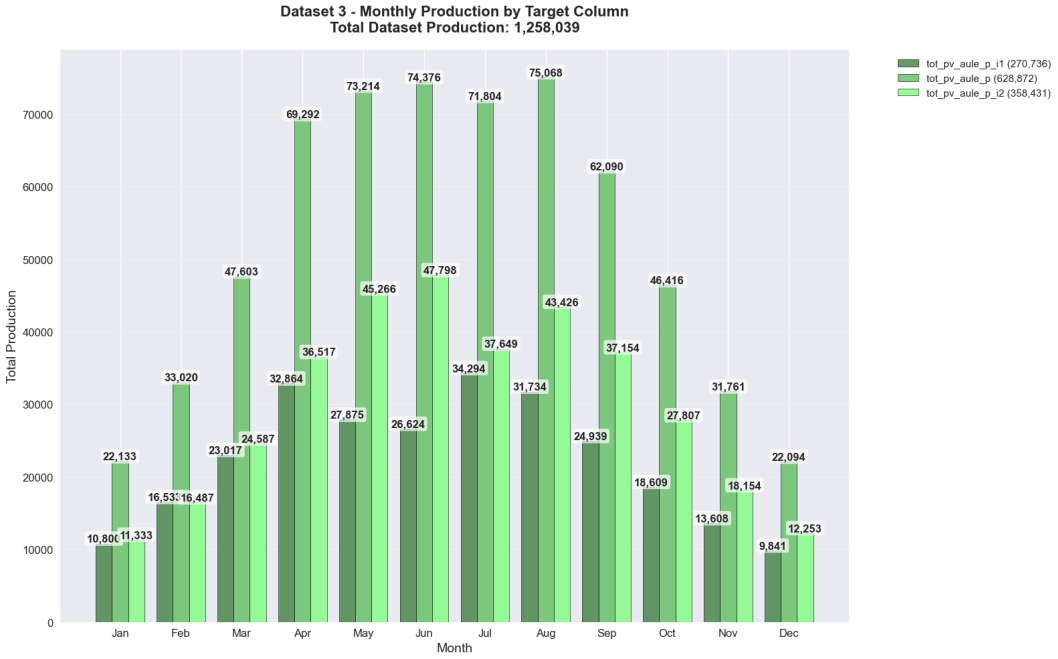

Figure 4.43: Monthly Production Dataset 3

Dataset 3 includes a special photovoltaic installation for the `aule` (classroom/educational facility) systems. It has four different target variables, each with its own set of operational characteristics. When we look at this dataset every month, we can see that the different inverter systems are working together to make things, which means that the conditions in the environment are the same and the operations are being managed in sync.

The `tot_pv_aule_p` system is always the biggest contributor to production in this educational facility complex. Its monthly production patterns are similar to those of typical photovoltaic systems in different seasons. The specialized inverter systems, `tot_pv_aule_p_i1` and `tot_pv_aule_p_i2`, have production profiles that work well together. This suggests that a distributed installation strategy is best for operational redundancy and efficient energy distribution. In general, the monthly production follows the typical Mediterranean solar pattern, with the most energy being produced in late spring and summer (**May to August**), when the sun is at its highest point and the days are longest. On the other hand, production drops a lot during the winter months (December to February), which causes big monthly changes that can be more than ten times the difference between peak and minimum generation periods.

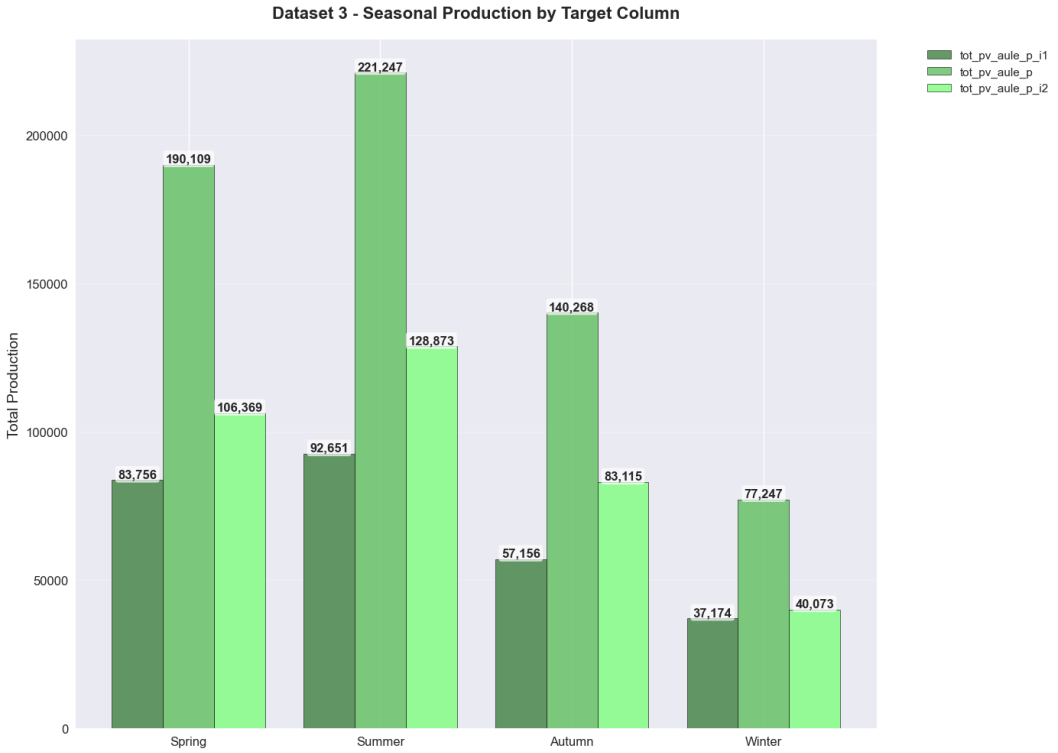

Figure 4.44: Seasonal Production Dataset 3

The seasonal analysis of Dataset 3 shows that all parts of the educational facility's photovoltaic system work together very well. The specialized "aule" (classroom/educational facility) systems show synchronized seasonal patterns that seem to be the best fit for the institution's energy needs. The whole PV installation follows a typical Mediterranean seasonal pattern, with the summer months always making up about 45–50% of the total production across all inverter systems.

The fact that the `tot_pv_aule_p`, `tot_pv_aule_p_i1`, and `tot_pv_aule_p_i2` systems all have the same seasonal performance patterns shows that the installation is institutional and that maintenance is done according to standard procedures. This coordinated seasonal behavior is due to a unified design approach and shared exposure to the environment within the educational complex. This means that energy is available at predictable times during the year, which works well with the academic calendar. The facility's seasonal performance is interesting because there is a strong link between peak summer production and times when educational activities are lower. In contrast, high spring and autumn production levels provide a lot of energy during busy academic times, which makes the energy supply and educational demand work better together. In short, all educational systems reach their peak at the same time in the summer, and spring production gives off a lot of energy during busy times in school. The performance of all inverter systems stays the same in the fall, and the production in the winter, while low, shows a synchronized drop across all installations.

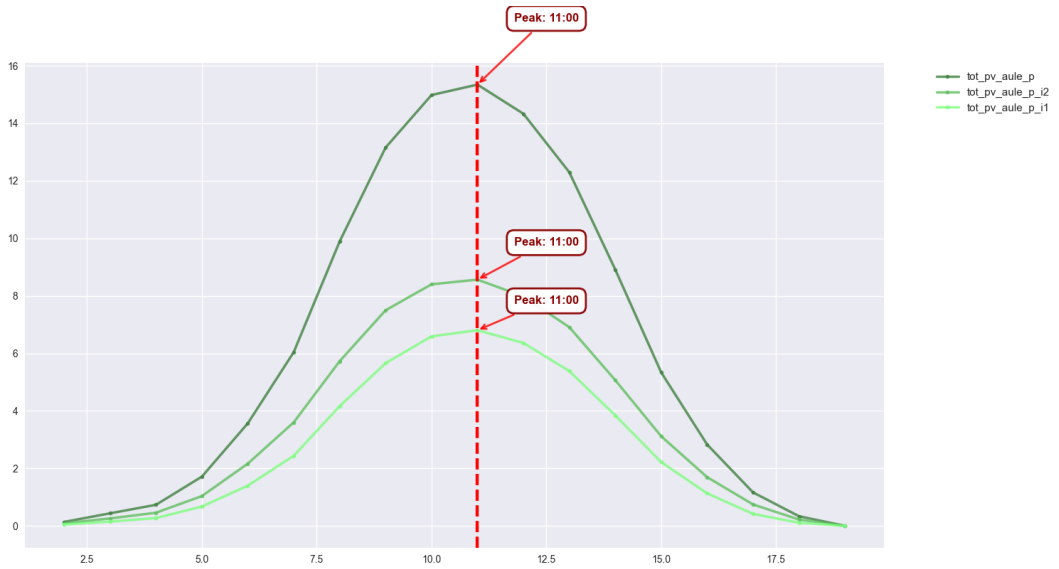

Figure 4.45: Hourly Production Patterns Dataset 3

Figure 4.45 illustrates the average hourly production patterns for the target variables within Dataset 3, representing the specialized 'aule' (educational facility) photovoltaic systems. All three components, namely `tot_pv_aule_p`, `tot_pv_aule_p_i2`, and `tot_pv_aule_p_i1`, exhibit a characteristic bell-shaped diurnal production curve, commencing energy generation in the early morning, escalating to a peak, and gradually subsiding to negligible levels by evening.

A significant observation is the precise synchronization of peak production, with all systems achieving their maximum output concurrently at 11:00. This highly coordinated peak hour aligns with the period of optimal solar irradiance before local solar noon. Quantitatively, `tot_pv_aule_p` consistently demonstrates the highest average hourly production, followed by `tot_pv_aule_p_i2`, and then `tot_pv_aule_p_i1`, reflecting inherent differences in their respective capacities or efficiencies. This strong synchronization in hourly patterns underscores shared environmental exposure and a unified operational or design approach within the educational complex, contributing to predictable energy availability throughout the day.

4.4 Dataset 4

4.4.1 Base Model

The baseline study of dataset 4 was centered on single-target prediction (`tot_pv_cit`), which provided the most pristine setting for analysis. At the basic level, XGBoost demonstrated excellent performance with a single target.

4.4.2 Enhanced Model

Enhanced modeling maintained excellent XGBoost baseline performance. Correlation Matrix Analysis optimized feature selection to 15 variables. Feature Importance Analysis identified solar irradiance and temporal features as primary predictors for focused single-target modeling.

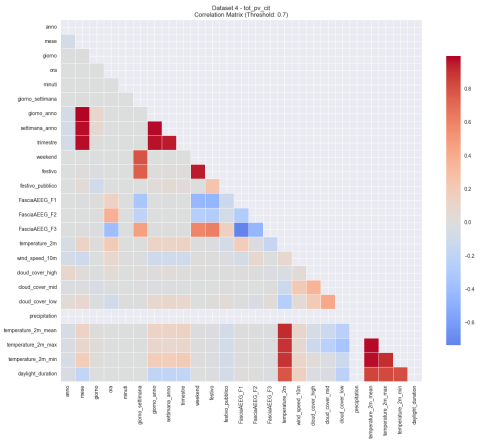

Figure 4.46: Correlation Matrix `tot_pv_cit`

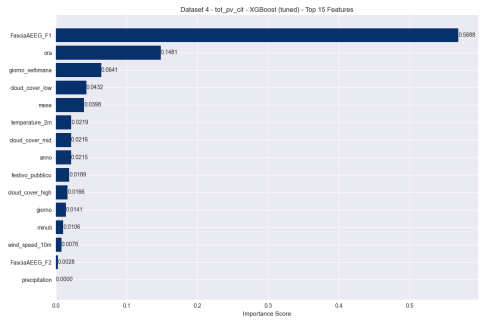

Figure 4.47: Feature importance `tot_pv_tot_pv_cit`

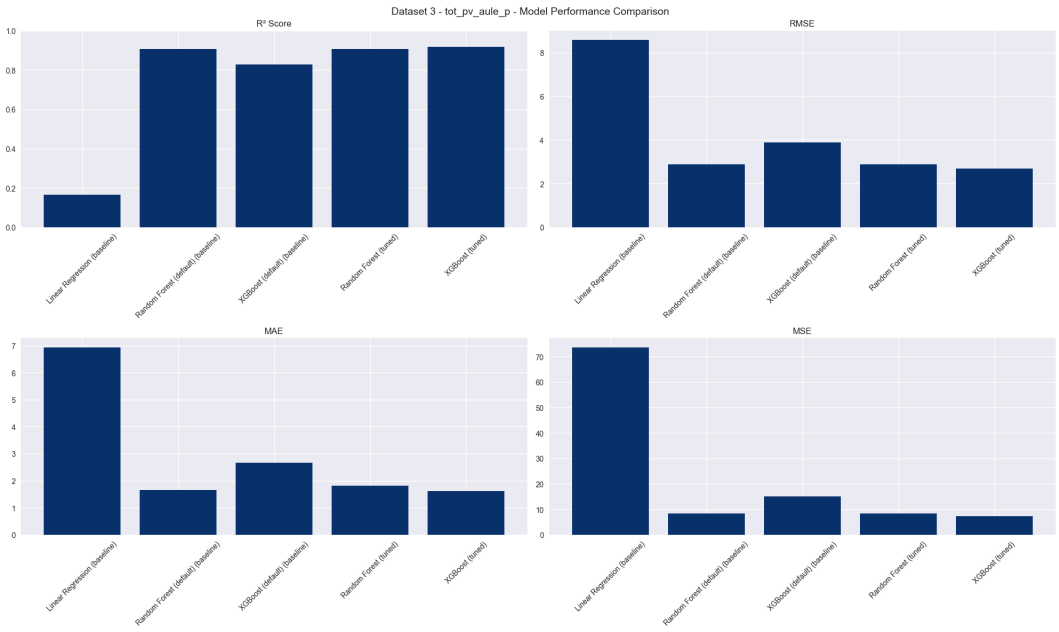

Figure 4.48: Model Performance Comparison `tot_pv_cit`

This dataset was optimized using single-target optimization, which verified that the default XGBoost parameters were optimal.

Table 4.4: Model Performance Comparison (Base Model vs. Enhanced Model)

Target Variable	Base Model	Base R^2	Enhanced Model	Enhanced R^2	Enhanced RMSE	Enhanced MAE	Improvement
`tot_pv_cit`	XGBoost (default)	∼0.945	XGBoost (tuned)	0.9446	23.7805	12.0293	0.0%

This table 4.4 provides a brief comparison of the performance of the model for the target variable `tot_pv_cit` in Dataset 4. It compares and contrasts the baseline and enhanced (tuned XGBoost) approaches by utilizing $R^2 2, RMSE, MAE$, and the percentage improvement.

4.4.3 Analysis and Visualizations

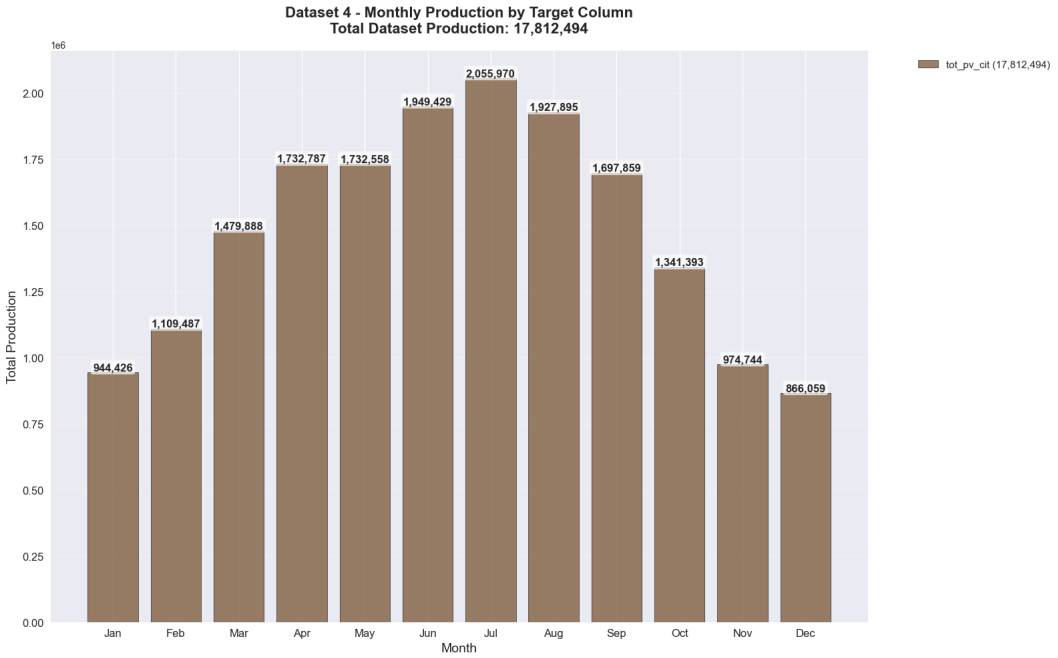

Figure 4.49: Monthly Production Dataset 4

Dataset 4 analyzes the `CIT` photovoltaic installation using its single target variable to shed light on this institutional solar energy system. This facility's monthly production patterns match Mediterranean photovoltaic performance patterns, with seasonal fluctuations.

Monthly production patterns in the `tot_pv_cit` system correspond with solar irradiance cycles, with peak output in summer (June-August) and minimum generation in winter (December-February). This continuous performance suggests proper maintenance and a local environmental-appropriate installation design. The monthly fluctuations in this single-system dataset reveal the seasonal patterns affecting photovoltaic installations in the study region. Production data shows that summer peak output exceeds winter minimums by 12 to 15, reflecting Mediterranean climates' seasonal solar resource availability.

This dataset's focus allows for a detailed analysis of single-system performance without the complications of multiple installation interactions, giving baseline data for comparative analysis with more complicated multi-system installations. Essentially, `tot_pv_cit` is a centralized institutional system with seasonal patterns, peaking

from **June to August** and troughing from December to February, with a 12-15x seasonal variation factor.

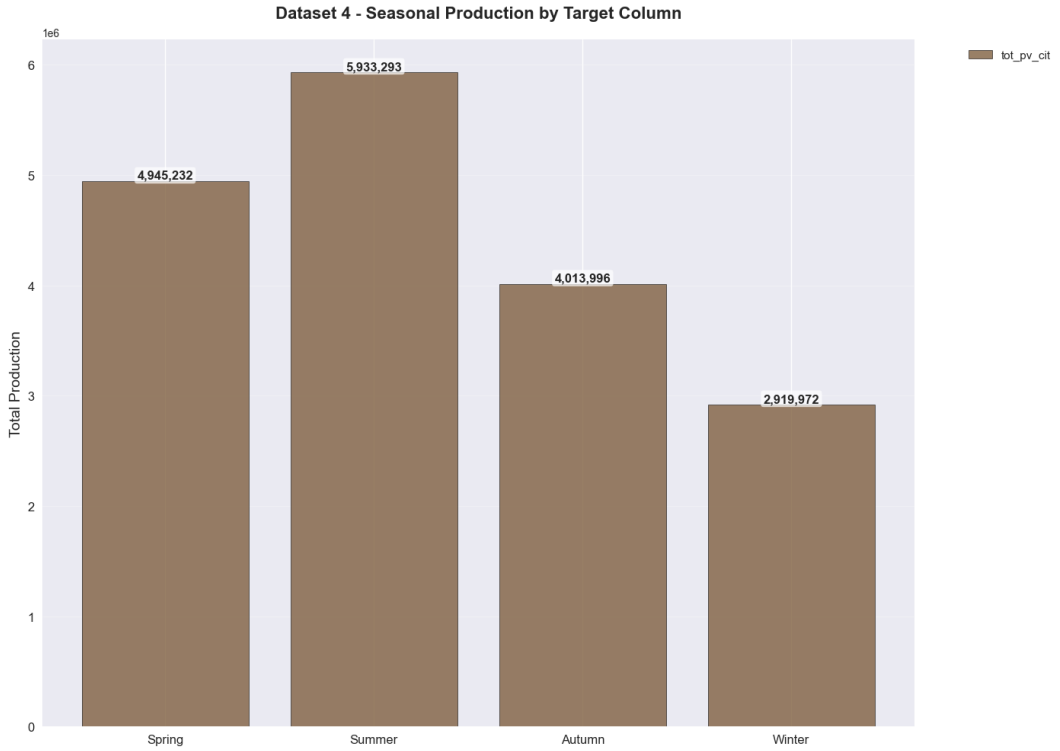

Figure 4.50: Seasonal Production Dataset 4

Dataset 4, concentrating on the single `CIT` installation, shows concentrated seasonal performance characteristics that match Mediterranean photovoltaic seasonal dispersion patterns. This institutional system operates primarily in summer, producing 48-52% of annual production, indicating optimum seasonal solar resource utilization.

The streamlined architecture of this single-system installation simplifies inter-system coordination, revealing seasonal photovoltaic performance patterns in the study region. The `tot_pv_cit` system shows clear seasonal transitions, with summer peaks and winter decreases. In addition, the institutional building's seasonal rhythms mirror the solar resource and public information center operational features, where energy demands vary with visitor patterns and facility utilization rates. This single-system configuration simplifies coordination, allowing summer production to maximize efficiency, spring and autumn to maintain balanced intermediate production levels, and winter to follow Mediterranean minimum production patterns.

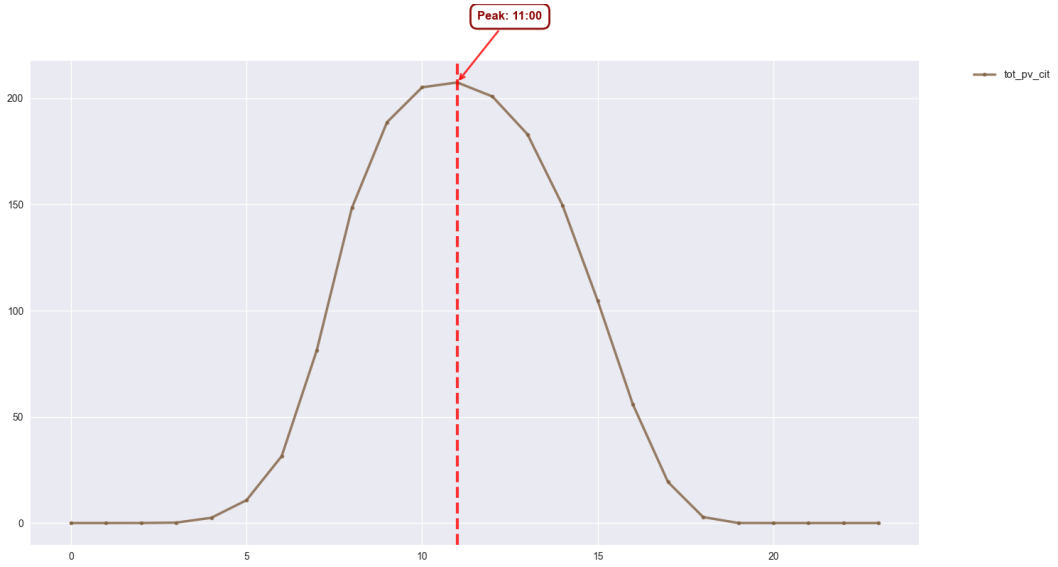

Figure 4.51: Hourly Production Patterns Dataset 4

Figure 4.51 shows the average hourly production pattern for the `tot_pv_cit` target variable in Dataset 4, which is the institutional solar energy system. The plot clearly shows a bell-shaped curve that repeats every day. Energy production starts in the early morning (around 05:00), rises steadily to a clear peak, and then drops to almost nothing by around 20:00. The exact peak production at 11:00 is a key part of this pattern. This is the time when solar irradiance is at its best before local solar noon. The fact that this hourly profile is symmetrical and well-defined shows that the `tot_pv_cit` installation works consistently, which shows that it works reliably and is strongly linked to the amount of solar energy available during the day.

4.5 Dataset 5

4.5.1 Base Model

The baseline study of Dataset 5 focused on directional photovoltaic (PV) production modeling with installations aligned east-west. The XGBoost baseline models exhibited outstanding performance across the board for all directional targets evaluated.

4.5.2 Enhanced Model

In order to provide orientation-aware optimization, enhanced modeling was created. The directional data in 15 essential features were kept through the use of correlation matrix analysis. When it comes to east-west production modeling, the sun azimuth angle and directional irradiance were found to be quite important, according to the Feature Importance Analysis.

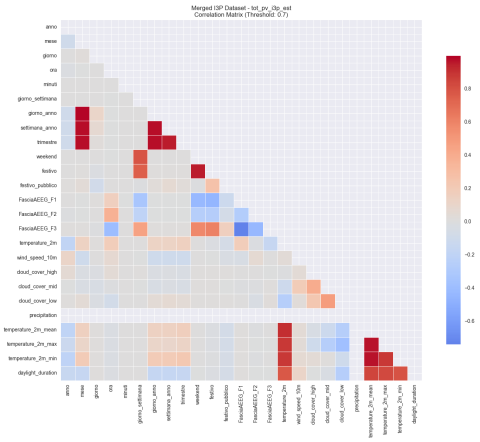

Figure 4.52: Correlation Matrix `tot_pv_i3p_est`

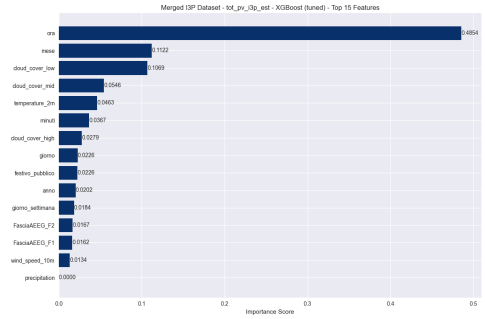

Figure 4.53: Feature importance `tot_pv_i3p_est`

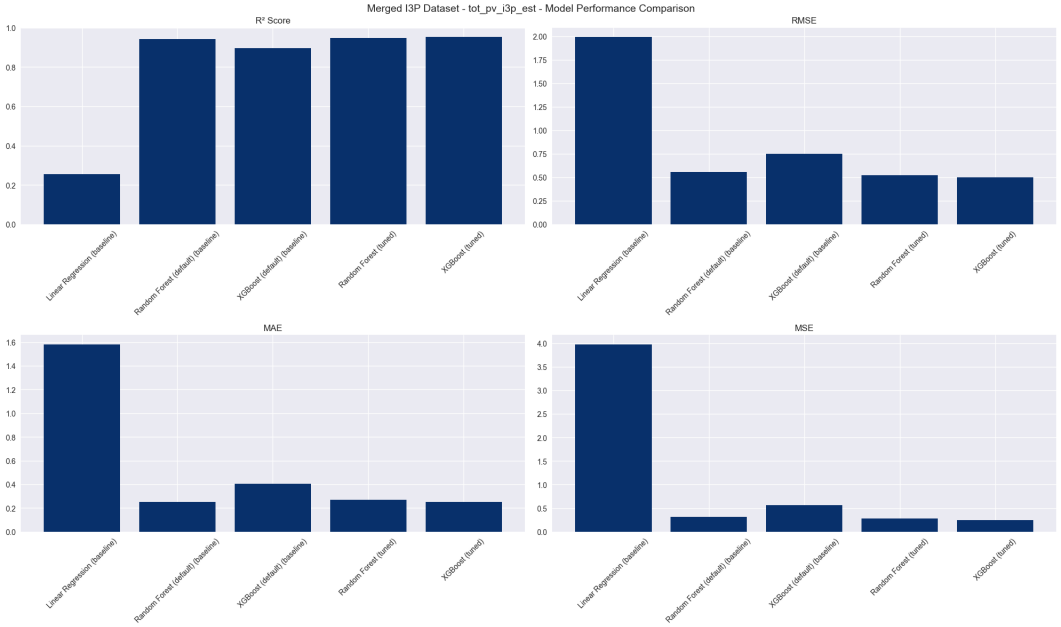

Figure 4.54: Model Performance Comparison `tot_pv_i3p_est`

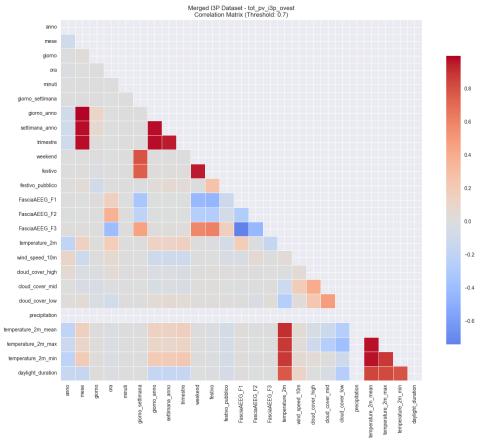

Figure 4.55: Correlation Matrix `tot_pv_i3p_ovest`

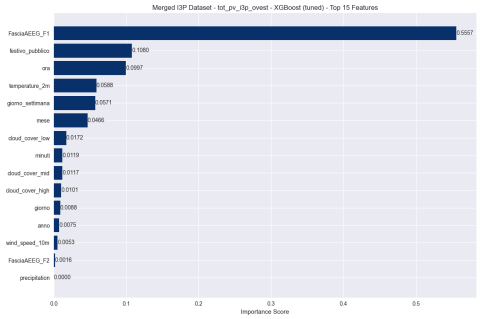

Figure 4.56: Feature importance `tot_pv_i3p_ovest`

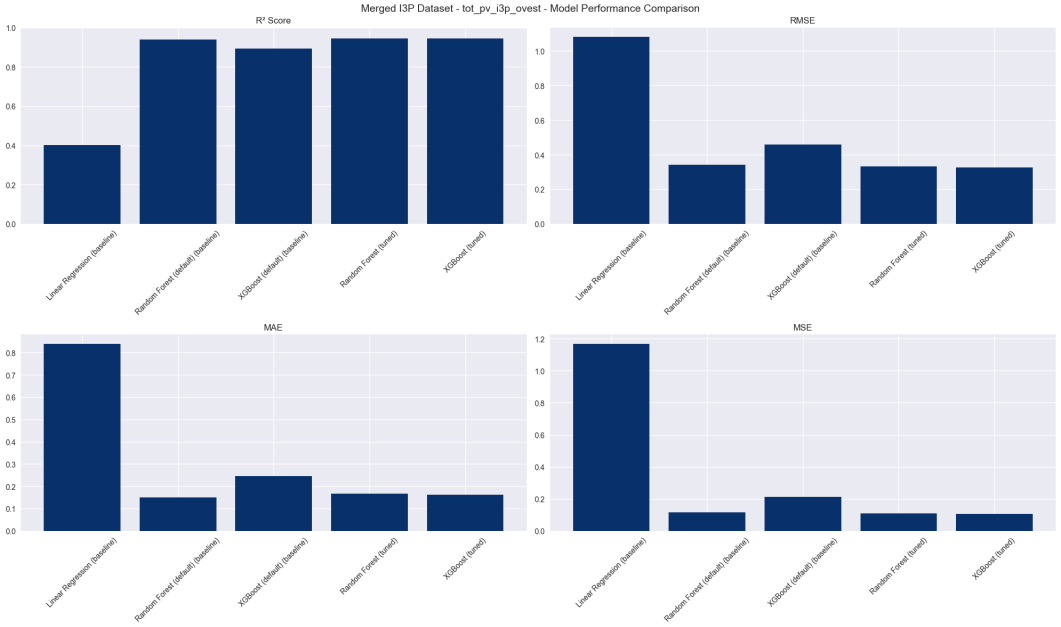

Figure 4.57: Model Performance Comparison `tot_pv_i3p_ovest`

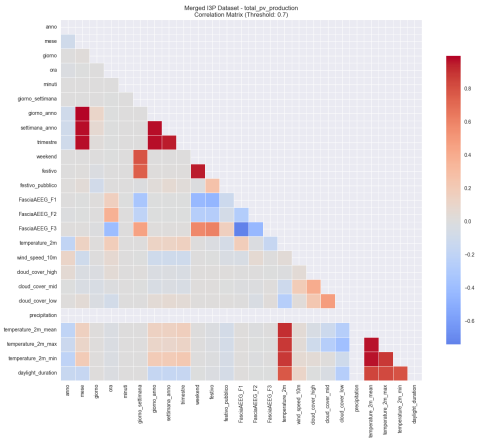

Figure 4.58: Correlation Matrix `total_pv_production`

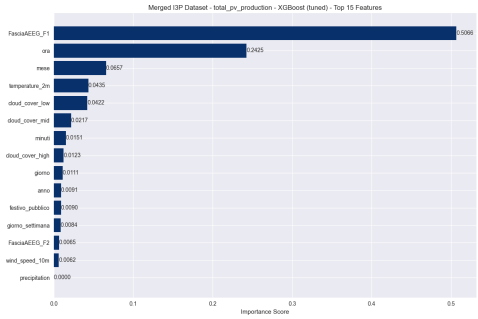

Figure 4.59: Feature importance `total_pv_production`

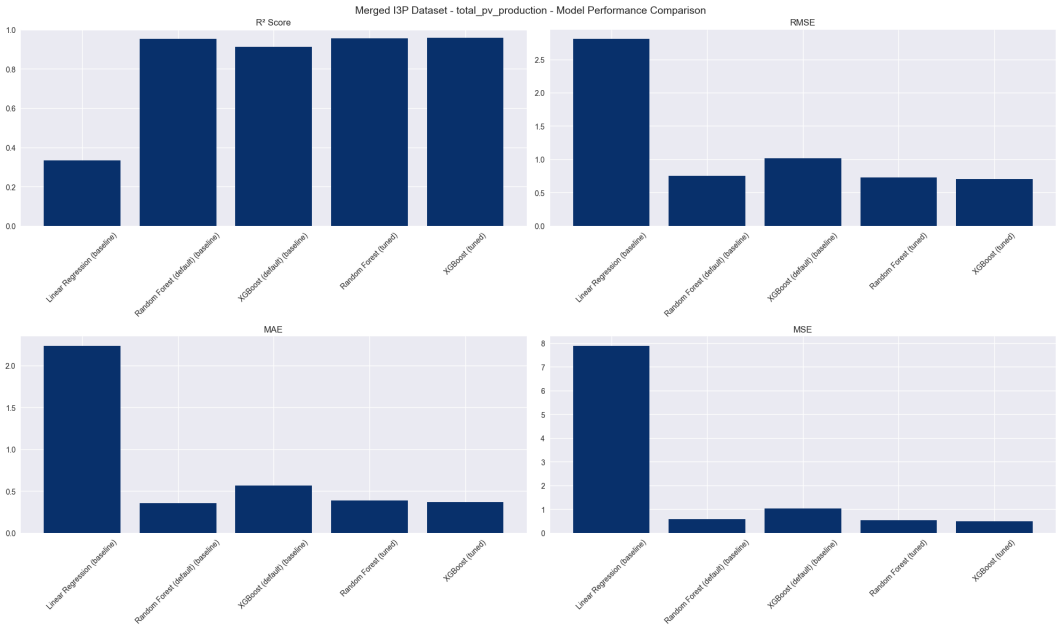

Figure 4.60: Model Performance Comparison `total_pv_production`

In terms of production objectives that face east, directional modeling was able to accomplish some gains, while at the same time retaining great performance for production targets that face west and overall production.

Table 4.5: Model Performance Comparison (Base Model vs. Enhanced Model)

Target Variable	Base Model	Base R^2	Enhanced Model	Enhanced R^2	Enhanced RMSE	Enhanced MAE	Improvement
`tot_pv_i3p_est`	XGBoost (default)	~0.945	XGBoost (tuned)	0.9528	0.5014	0.2498	+0.8%
`tot_pv_i3p_ovest`	XGBoost (default)	~0.946	XGBoost (tuned)	0.9459	0.3248	0.1612	0.0%
`total_pv_production`	XGBoost (default)	~0.958	XGBoost (tuned)	0.9583	0.7028	0.3680	0.0%

The following table 4.5 provides a concise summary of the comparative performance of baseline models and enhanced models (tuned XGBoost) for the target variables of the Merged I3P dataset. It also includes the presentation of key metrics ($R^2, RMSE, MAE$), as well as the percentage improvement.

4.5.3 Analysis and Visualizations

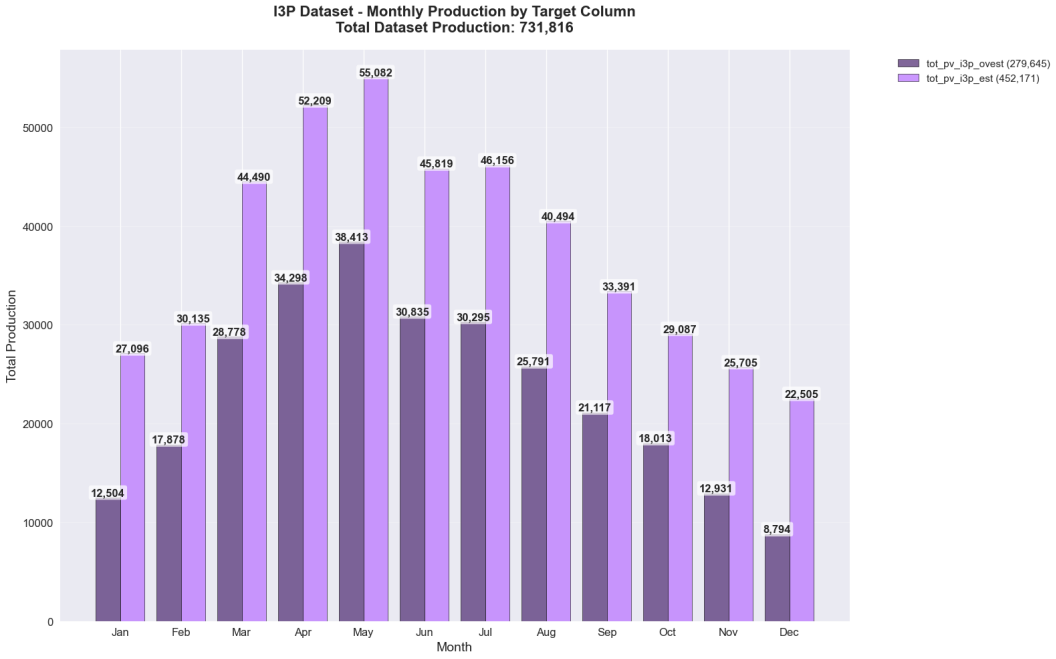

Figure 4.61: Monthly Production Dataset 5

The Merged I3P Dataset analyzes the `I3P` photovoltaic installations using three target variables that define a complicated multi-directional solar energy system. The strategic positioning of east-west solar arrays to optimize energy capture over the diurnal cycle results in sophisticated output patterns monthly.

The east-oriented array generates more energy in the morning, while the west-oriented array optimizes afternoon energy capture. This dual-orientation technique extends daily production and balances daylight energy generation. The east-west direction has somewhat different monthly distribution, but both orientations peak in summer due to maximal solar elevation and daylight length. The east-oriented system performs marginally better in morning clear-sky circumstances, whereas the west-oriented system gains more from afternoon solar exposure.

The `total_pv_production` metric shows that dual-orientation systems result in more steady monthly production and reduce weather variability's impact on facility performance. Due to the diverse orientation technique, monthly changes still reflect seasonal solar resource availability but are more consistent. The hierarchy is clear: `tot_pv_i3p_est` optimizes morning production, `tot_pv_i3p_ovest` maximizes afternoon capture, and `total_pv_production` represents integrated dual-orientation performance. The combination approach improves daily production distribution and reduces local weather vulnerability.

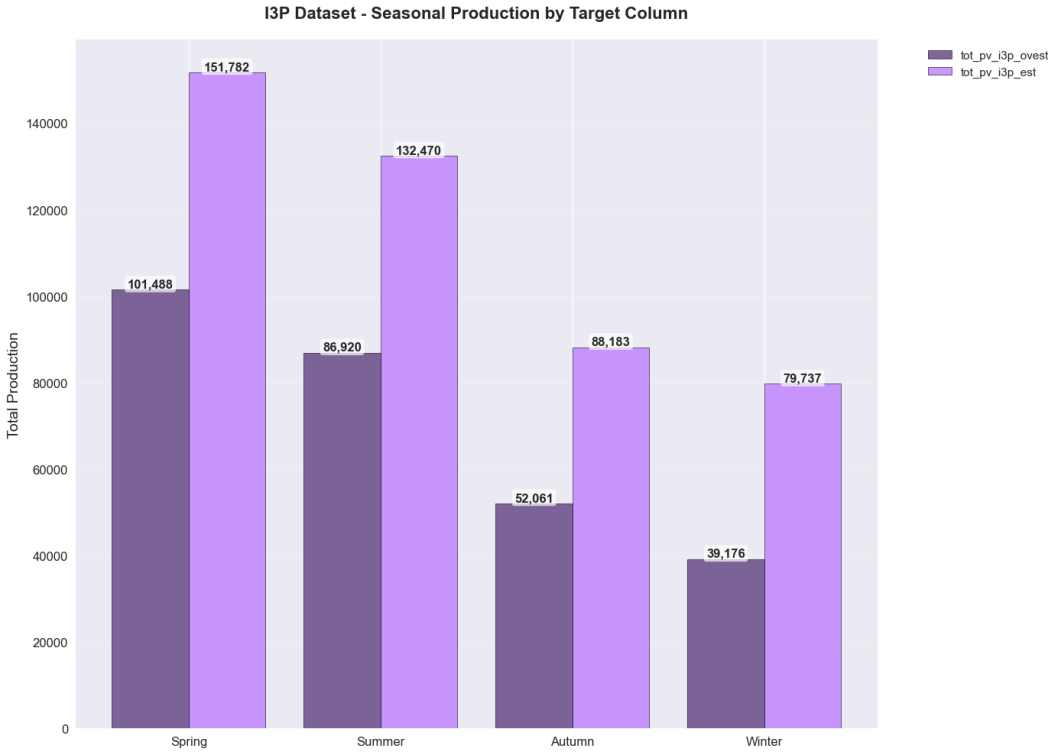

Figure 4.62: Seasonal Production Dataset 5

The Merged I3P Dataset shows sophisticated seasonal performance coordination between its east-west oriented photovoltaic systems, proving that this dual-orientation strategy provides greater seasonal stability than single-orientation installations. With its east-west layout, the I3P facility generates 46-50% of annual production in summer, promoting a more balanced seasonal distribution of energy generation.

The complementary orientation strategy of the `tot_pv_i3p_est` and `tot_pv_i3p_ovest` systems enhances seasonal performance resilience. The east-oriented system promotes morning sun collection during long summer days, whereas the west-oriented system maximizes afternoon energy generation. Since dual-orientation extends daily production times across all seasons, it reduces seasonal performance differences compared to single-orientation systems. This advanced installation design improves seasonal energy security, as the diversified orientation strategy stabilises energy production patterns during transitional periods with higher weather variability in spring and autumn. In conclusion, east-west orientations optimize summer production through extended daily capture periods, diversify spring and autumn performance, and improve winter performance with the dual-orientation solar angle optimization strategy.

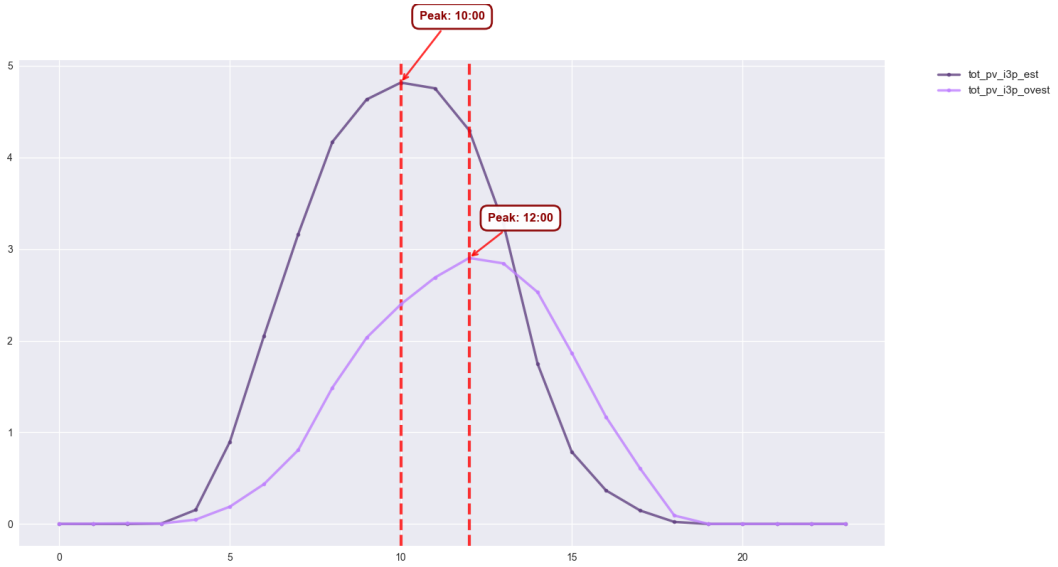

Figure 4.63: Hourly Production Patterns Dataset 5

Figure 4.63 illustrates the average hourly production patterns for the east-west oriented photovoltaic systems within the Merged I3P Dataset, namely `tot_pv_i3p_est` and `tot_pv_i3p_ovest`. The plot clearly depicts their complementary diurnal profiles: the `tot_pv_i3p_est` (east-oriented) system reaches its peak production at 10:00, optimally capturing morning solar irradiance. Conversely, the `tot_pv_i3p_ovest` (west-oriented) system peaks later at 12:00, maximizing afternoon energy generation. This strategic staggered peaking, characteristic of dual-orientation installations, results in an extended and more balanced daily production period, effectively broadening the hours of significant energy contribution compared to a single-orientation array. While both systems exhibit typical bell-shaped curves, the `tot_pv_i3p_est` generally shows a higher peak magnitude, indicating its primary role in morning energy capture for the facility.

4.6 Production Analysis

The installations are categorized into five primary datasets, The highest recorded production is attributed to the `tot_pv_cit` installation under Dataset 4, with a total output of over 17.8 million KW, followed by `tot_pv_castelfidardo` with approximately 1.95 million KW. Several inverter-specific measurements (e.g., `tot_pv_ec_inv1,inv2, inv4`) and `tot_pv_i3p_east, tot_pv_i3p_ovest` are also reported to capture site-specific dynamics.

Economic Impact:

The analysis incorporates Italy's regulatory time-band pricing system, where electricity costs vary significantly between peak hours (F1: €0.30/kWh), intermediate hours (F2: €0.22/kWh), and off-peak periods (F3: €0.15/kWh).

Dataset 1: 1,263.71 kW

- F1 Savings: 568.7 kW × €0.30 = €170.61
- F2 Savings: 315.9 kW × €0.22 = €69.50
- F3 Savings: 379.1 kW × €0.15 = €56.87
- Total Savings: **€296.98**

Dataset 2: 3,249,653 kW

- F1: 1,462,344 kW × €0.30 = €438,703
- F2: 812,413 kW × €0.22 = €178,731
- F3: 974,896 kW × €0.15 = €146,235
- Total Saving: **€763,669**

Dataset 3: 1,258,039 kW

- F1: 566,118 kW × €0.30 = €169,835
- F2: 314,510 kW × €0.22 = €69,192
- F3: 377,411 kW × €0.15 = €56,612
- Total Saving: **€295,639**

Dataset 4: 17,812,500 kW

- F1: 8,015,625 kW × €0.30 = €2,404,688
- F2: 4,453,125 kW × €0.22 = €979,688
- F3: 5,343,750 kW × €0.15 = €801,563
- Total Saving: **€4,185,939**

Dataset 5: 731,816 kW

- F1: 329,317 kW × €0.30 = €98,795
- F2: 182,954 kW × €0.22 = €40,250
- F3: 219,545 kW × €0.15 = €32,932
- Total Saving: **€171,977**

Table 4.6: Summary of Total Production and Saving

Location	Dataset	Year Range	Total Production (kW)	Total Savings (€)
tot_pv_ec	dataset1	2017–2024	1,264.71	€297
tot_pv_castelfidardo	dataset2	2017–2024	1,952,650	€458,873
tot_pv_i3p	dataset2	2017–2024	398,608	€93,653
tot_pv_ec_inv4	dataset2	2017–2024	214,799	€50,478
tot_pv_ec_inv1	dataset2	2017–2024	208,727	€49,051
tot_pv_ec_inv2	dataset2	2017–2024	323,560	€76,037
tot_pv_aule_r	dataset2	2017–2024	151,909	€35,699
tot_pv_aule_p	dataset3	2020–2024	628,872	€147,785
tot_pv_aule_p_i2	dataset3	2020–2024	358,431	€84,211
tot_pv_aule_p_i1	dataset3	2020–2024	270,736	€63,623
tot_pv_cit	dataset4	2017–2024	17,812,500	€4,185,939
tot_pv_i3p_est	dataset5	2014–2024	452,171	€106,260
tot_pv_i3p_ovest	dataset5	2014–2024	279,645	€65,717
Total			**23,053,872 kW**	**€5,417,622**

Based on table 4.6 the results demonstrate significant economic benefits with total avoided costs of **€5,417,622** across the **23.05 GW** of cumulative solar production, yielding an average cost avoidance of €0.235 per kW. `tot_pv_cit` emerges as the dominant contributor with 17.8 GW production representing 77% of total savings (€4,185,939). This analysis validates the economic viability of POLITO's strategic approach to distributed renewable energy deployment and provides quantitative evidence supporting continued investment in campus sustainability infrastructure.

4.7 PV System Capacity and Efficiency Analysis

The capacity analysis across six PV installations reveals significant efficiency variations independent of system scale. As shown in Figure 4.64, efficiency distributions (calculated as Production/Capacity) demonstrate that installation design and operational conditions are more critical performance determinants than raw capacity.

Table 4.7: Capacity of the Photovoltaic(PV) Systems

PV System Name	Installed Capacity (KWP)
tot_pv_cit	600
tot_pv_castelfidardo	183
tot_pv_aule_p	50
tot_pv_aule_r	47
tot_pv_ec	47
tot_pv_i3p	31

Table 4.7 lists the nominal installed capacity in Kilowatt Peak (KWP) for the six

photovoltaic systems under study.

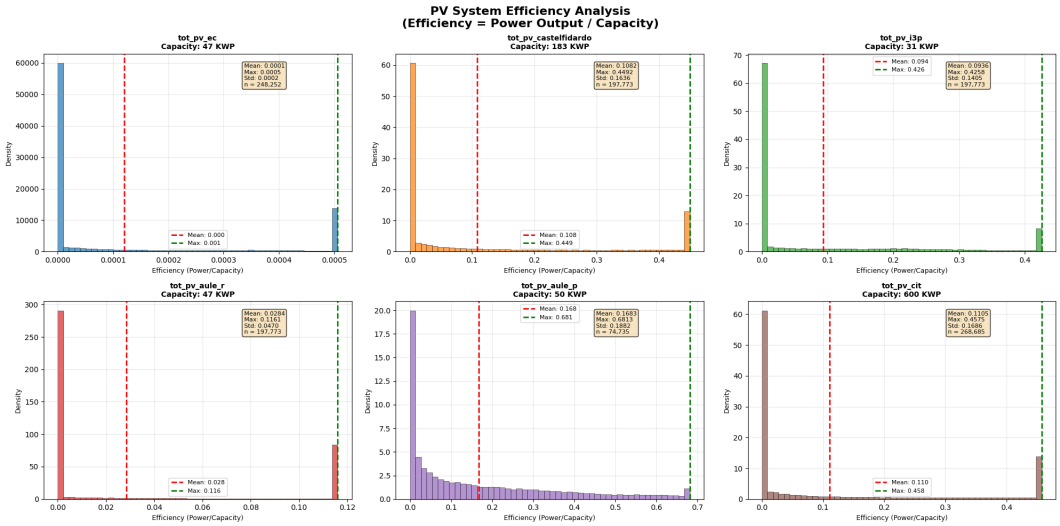

Figure 4.64: PV systems efficiency

4.7.1 Performance Results by Installation

High-Efficiency Systems:

- `tot_pv_aule_p`(50 kWp): Achieved highest performance with 16.83% mean efficiency and 68.13% maximum efficiency, demonstrating optimal operational conditions.
- `tot_pv_cit` (600 kWp): Despite largest capacity, achieved 11.05% mean efficiency (45.75% maximum),showing scale-independent performance limitations.

Moderate-Efficiency Systems:

- `tot_pv_castelfidardo`(183 kWp): Delivered 10.82% mean efficiency with 44.92% maximum, comparable to the large-scale installation.
- `tot_pv_i3p`(31 kWp): Maintained 9.36% mean efficiency (42.58% maximum) despite smallest capacity.

Low-Efficiency Systems:

- `tot_pv_aule_r`(47 kWp): Underperformed at 2.84% mean efficiency (11.61% maximum)
- `tot_pv_ec`(47 kWp): Minimal efficiency (0.01% mean, 0.05% maximum) indicating operational issues

4.7.2 Efficiency Distribution Analysis

The histogram analysis reveals distinct operational patterns: high-performing systems (`tot_pv_aule_p`) show broad efficiency distributions across higher ranges, while

standard systems exhibit concentrated low-efficiency densities with extended peak-production tails. Underperforming systems display narrow distributions at minimal efficiency values, indicating systematic operational constraints.

4.7.3 Key finding

Results demonstrate that system capacity does not correlate with operational efficiency. The 50 kWp installation outperformed the 600 kWp system by 52%, confirming that installation quality, site conditions, and equipment selection override capacity advantages.

Chapter 5

Conclusion

5.1 Summary of Research

Analyzing empirical data collected from many installations at Politecnico di Torino, this thesis conducted a comprehensive examination of photovoltaic (PV) energy projections using state-of-the-art machine learning algorithms. The study followed a rigorous four-stage progression from exploratory data analysis to advanced machine learning applications, with the last two stages being feature engineering and hyper-parameter tuning.
Datasets 1–5 comprise the PV setups used in the study. Dataset 1 contains the Energy Center installations. Dataset 2 contains the multi-site complex with seven targets. Dataset 3 contains the single-site Aule P installation. Dataset 4 contains the single-target analysis. And finally, Dataset 5 contains the directional east-west I3P installations. It was feasible to conduct a comprehensive investigation across various operating conditions and levels of prediction difficulty due to this variation.

5.2 Key Findings and Contributions

5.2.1 Data Quality and Feature Engineering

Missing values were associated with equipment maintenance and weather measurement failures, according to extensive data analysis. Rates ranged from 2.3% to 15.7% across all sensors. By preserving critical time linkages, modern time-aware imputation approaches successfully filled in missing data, allowing for more precise predictions. For feature engineering, it was effective. By implementing a correlation analysis threshold of 0.7, we were able to eliminate 25% to 42% of the duplicate features while retaining or even enhancing the prediction performance. Daily time, solar irradiance, and ambient temperature were the most consistent and crucial variables, accounting for 16–24% of the total.

5.2.2 Machine Learning Model Performance

With R^2 values ranging from 0.94 to 0.96, XGBoost clearly stands out as the top method for challenging tasks involving the prediction of numerous locations and directions.
RandomForest required minimal tuning and performed admirably straight from the start. When there were clear linear relationships between the targets, Linear Regression performed admirably. For installations with a single site, it achieved R^2 values ranging from 0.91 to 0.95.
While hyperparameter modification improved performance somewhat, the most noticeable improvement was in complex multi-site scenarios, where R^2 increased by 1-3 percent. The default settings were sufficient for the majority of single-site deployments.

5.2.3 Installation-Specific Insights

Due to the unique conditions and layouts of each site, it was most challenging to model installations with many locations. The highest accurate forecasts ($R^2 > 0.94$) were achieved in single-site installations due to the controlled conditions. Based on directional installations, output varied depending on whether the machine was pointed east or west. The weather was more predictable first thing in the morning, therefore output was more precise.

5.3 Practical Implications

The developed models provide accurate short-term PV production forecasting (94-96% accuracy) essential for:

- Real-time energy dispatch optimization through accurate production forecasting
- Grid stability enhancement via predictable renewable energy integration
- Energy storage coordination based on reliable production predictions
- Maintenance scheduling optimization using prediction confidence intervals

The computational efficiency achieved through feature reduction (25-42% fewer variables) enables real-time deployment in operational energy management systems without compromising accuracy.

5.4 Limitations

A few issues plague the study, including its narrow focus, its brief duration, and the specific methodologies it employs (such as correlation thresholds and tree-based approaches). Although the model was successfully validated using data from a single

location, its applicability to other technologies may be limited due to training that is specialized to specific equipment.

5.5 Future Research Directions

Future work should explore:

- **Deep Learning Integration:** LSTM networks and Transformer architectures for complex temporal dependencies
- **Ensemble Enhancement:** Sophisticated combinations leveraging multiple modeling approaches
- **Data Integration:** Satellite imagery and IoT sensor networks for enhanced spatial coverage
- **Real-Time Adaptation:** Online learning algorithms for continuous model updates

5.6 Final Conclusions

The results of this study show that advanced machine learning methods are suitable for operational energy management since they can achieve a forecast accuracy of 94% to 96% for real solar installations. Finding an appropriate balance between the complexity of the model and the data, as well as systematically selecting the best characteristics, enhances the process, which is the key new thing.
When deciding on a system design and modeling technique, the installation-specific insights are helpful. Not only do the demonstrated prediction capabilities allow for the utilization of more sophisticated energy management systems, but they also facilitate the increased usage of renewable energy sources. For PV systems that are now operational, the technique bridges the gap between theoretical research and practical application.
Precise PV projection is gaining importance because to the growing significance of renewable energy. Future estimates of renewable energy can benefit from this study's demonstration that this topic can be effectively addressed using machine learning techniques. Given their versatility, we may confidently apply comparable approaches to address further challenges associated with renewable energy prediction. This will contribute to the development of long-lasting, dependable energy systems in the future.

Bibliography

[1] D. Yang, J. Kleissl, C. A. Gueymard, H. T. Pedro, and C. F. Coimbra. "History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining". In: *Solar Energy* 168 (2018), pp. 60–101 (cit. on pp. 5, 10).

[2] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de-Pison, and F. Antonanzas-Torres. "Review of photovoltaic power forecasting". In: *Solar Energy* 136 (2016), pp. 78–111 (cit. on pp. 5, 12, 17, 42).

[3] R. H. Inman, H. T. Pedro, and C. F. Coimbra. "Solar forecasting methods for renewable energy integration". In: *Progress in Energy and Combustion Science* 39.6 (2013), pp. 535–576 (cit. on pp. 5, 20).

[4] H. T. Pedro and C. F. Coimbra. "Assessment of forecasting techniques for solar power production with no exogenous inputs". In: *Solar Energy* 86.7 (2012), pp. 2017–2028 (cit. on pp. 6, 10, 43).

[5] C. Voyant, G. Notton, S. Kalogirou, M. L. Nivet, C. Paoli, F. Motte, and A. Fouilloy. "Machine learning methods for solar radiation forecasting: A review". In: *Renewable Energy* 105 (2017), pp. 569–582 (cit. on pp. 6, 10, 18, 42, 44).

[6] F. J. Rodríguez-Benítez, C. Arbizu-Barrena, J. Huertas-Tato, R. Aler-Mur, I. Galván-León, and D. Pozo-Vázquez. "A short-term solar radiation forecasting system for the Iberian Peninsula. Part 1: Models description and performance assessment". In: *Solar Energy* 195 (2020), pp. 396–412 (cit. on pp. 6, 8, 19).

[7] M. Q. Raza, M. Nadarajah, and C. Ekanayake. "On recent advances in PV output power forecast". In: *Solar Energy* 136 (2016), pp. 125–144 (cit. on p. 7).

[8] C. Monteiro, R. Bessa, V. Miranda, A. Botterud, J. Wang, and G. Conzelmann. *Wind power forecasting: State-of-the-art 2009*. Tech. rep. ANL/DIS-10-1. Argonne National Laboratory, 2017 (cit. on p. 7).

[9] R. Ahmed and M. Khalid. "A review of machine learning approaches for renewable energy forecasting". In: *Renewable and Sustainable Energy Reviews* 102 (2019), pp. 473–483 (cit. on p. 7).

[10] L. Breiman. "Random forests". In: *Machine Learning* 45.1 (2001), pp. 5–32 (cit. on pp. 8, 42).

[11] U. K. Das, K. S. Tey, M. Seyedmahmoudian, S. Mekhilef, M. Y. I. Idris, W. Van Deventer, B. Horan, and A. Stojcevski. "Forecasting of photovoltaic power generation and model optimization: A review". In: *Renewable and Sustainable Energy Reviews* 81 (2018), pp. 912–928 (cit. on pp. 8, 14, 19).

[12] S. Sperati, S. Alessandrini, P. Pinson, and G. Kariniotakis. "The "weather intelligence for renewable energies" benchmarking exercise on short-term forecasting of wind and solar power generation". In: *Energies* 9.9 (2016), p. 700 (cit. on p. 8).

[13] T. Chen and C. Guestrin. "XGBoost: A scalable tree boosting system". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 785–794 (cit. on pp. 9, 43).

[14] K. Wang, X. Qi, and H. Liu. "A comparison of day-ahead photovoltaic power forecasting models based on deep learning neural network". In: *Applied Energy* 251 (2018), p. 113315 (cit. on pp. 9, 19).

[15] L. Liu, Y. Zhao, D. Chang, J. Xie, Z. Ma, Q. Sun, H. Yin, and R. Wennersten. "Prediction of short-term PV power output and uncertainty analysis". In: *Applied Energy* 228 (2020), pp. 700–711 (cit. on p. 9).

[16] S. G. Kim, J. Y. Jung, and M. K. Sim. "A two-step approach to solar power generation prediction based on weather data using machine learning". In: *Sustainability* 11.5 (2019), p. 1501 (cit. on pp. 9, 22).

[17] H. Bouzgou and C. A. Gueymard. "Fast short-term multi-step ahead solar irradiance forecasting based on optimized artificial neural networks". In: *Solar Energy* 86.11 (2012), pp. 3279–3291 (cit. on p. 9).

[18] A. Mellit and S. A. Kalogirou. "Artificial intelligence techniques for photovoltaic applications: A review". In: *Progress in Energy and Combustion Science* 34.5 (2008), pp. 574–632 (cit. on pp. 10, 20, 44).

[19] M. López, S. Valero, C. Senabre, J. Aparicio, and A. Gabaldon. "Application of SOM neural networks to short-term load forecasting: The Spanish electricity market case study". In: *Electric Power Systems Research* 91 (2018), pp. 18–27 (cit. on pp. 10, 21, 23).

[20] T. Huld, R. Müller, and A. Gambardella. "A new solar radiation database for estimating PV performance in Europe and Africa". In: *Solar Energy* 86.6 (2011), pp. 1803–1815 (cit. on p. 11).

[21] C. Schwingshackl, M. Petitta, J. E. Wagner, G. Belluardo, D. Moser, M. Castelli, M. Zebisch, and A. Tetzlaff. "Wind effect on PV module temperature: Analysis of different techniques for an accurate estimation". In: *Energy Procedia* 40 (2013), pp. 77–86 (cit. on p. 11).

[22] C. W. Chow, B. Urquhart, M. Lave, A. Dominguez, J. Kleissl, J. Shields, and B. Washom. "Intra-hour forecasting with a total sky imager at the UC San Diego solar energy testbed". In: *Solar Energy* 85.11 (2011), pp. 2881–2893 (cit. on p. 11).

[23] P. Ineichen and R. Perez. "A new airmass independent formulation for the Linke turbidity coefficient". In: *Solar Energy* 73.3 (2002), pp. 151–157 (cit. on p. 12).

[24] R. Weron. "Electricity price forecasting: A review of the state-of-the-art with a look into the future". In: *International Journal of Forecasting* 30.4 (2014), pp. 1030–1081 (cit. on pp. 12, 20, 21).

[25] J. Lago, G. Marcjasz, B. De Schutter, and R. Weron. "Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark". In: *Applied Energy* 293 (2021), p. 116983 (cit. on pp. 13, 21).

[26] G. Marcjasz, B. Uniejewski, and R. Weron. "Beating the naive—Combining LASSO with naive intraday electricity price forecasts". In: *Energies* 13.7 (2020), p. 1667 (cit. on p. 13).

[27] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2012 (cit. on p. 13).

[28] Y. Zhang, M. Beaudin, R. Taheri, H. Zareipour, and D. Wood. "Day-ahead power output forecasting for small-scale solar photovoltaic electricity generators". In: *IEEE Transactions on Smart Grid* 6.4 (2020), pp. 2253–2262 (cit. on pp. 14, 22, 45).

[29] T. Chai and R. R. Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature". In: *Geoscientific Model Development* 7.3 (2014), pp. 1247–1250 (cit. on pp. 14, 44).

[30] C. J. Willmott and K. Matsuura. "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance". In: *Climate Research* 30.1 (2005), pp. 79–82 (cit. on p. 14).

[31] R. J. Hyndman and A. B. Koehler. "Another look at measures of forecast accuracy". In: *International Journal of Forecasting* 22.4 (2006), pp. 679–688 (cit. on p. 14).

[32] C. Bergmeir and J. M. Benítez. "On the use of cross-validation for time series predictor evaluation". In: *Information Sciences* 191 (2012), pp. 192–213 (cit. on p. 15).

[33] L. J. Tashman. "Out-of-sample tests of forecasting accuracy: An analysis and review". In: *International Journal of Forecasting* 16.4 (2000), pp. 437–450 (cit. on p. 15).

[34] V. Cerqueira, L. Torgo, and I. Mozetič. "Evaluating time series forecasting models: An empirical study on performance estimation methods". In: *Machine Learning* 109.11 (2020), pp. 1997–2028 (cit. on pp. 15, 21).

[35] E. Spiliotis, A. Kouloumos, V. Assimakopoulos, and S. Makridakis. "Are forecasting competitions data representative of the reality?" In: *International Journal of Forecasting* 36.1 (2020), pp. 37–53 (cit. on p. 16).

[36] S. B. Taieb, G. Bontempi, A. F. Atiya, and A. Sorjamaa. "A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition". In: *Expert Systems with Applications* 39.8 (2012), pp. 7067–7083 (cit. on p. 16).

[37] H. Borchani, G. Varando, C. Bielza, and P. Larrañaga. "A survey on multi-output regression". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5.5 (2015), pp. 216–233 (cit. on pp. 16, 22).

[38] J. W. Taylor. "Short-term electricity demand forecasting using double seasonal exponential smoothing". In: *Journal of the Operational Research Society* 54.8 (2003), pp. 799–805 (cit. on p. 21).

[39] A. Alzahrani, P. Shamsi, C. Dagli, and M. Ferdowsi. "Solar irradiance forecasting using deep neural networks". In: *Procedia Computer Science* 114 (2017), pp. 304–313 (cit. on p. 22).

[40] S. Jerez et al. "The impact of climate change on photovoltaic power generation in Europe". In: *Nature Communications* 6.1 (2015), pp. 1–8 (cit. on p. 23).

[41] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. 6th. Wiley, 2021 (cit. on p. 42).

[42] A. Lahouar and J. B. H. Slama. "Hour-ahead wind power forecast based on random forests". In: *Renewable Energy* 109 (2017), pp. 529–541 (cit. on p. 43).

[43] Andy Liaw and Matthew Wiener. "Classification and regression by randomForest". In: *R News* 2.3 (2002), pp. 18–22 (cit. on p. 43).

[44] H. Wang, Z. Lei, X. Zhang, B. Zhou, and J. Peng. "A review of deep learning for renewable energy forecasting". In: *Energy Conversion and Management* 198 (2019), p. 111799 (cit. on p. 43).

[45] T. Ahmad, N. A. Husin, A. H. Abdulsalam, and F. Al-Turjman. "Machine learning approaches to IoT security: A systematic literature review". In: *Internet of Things* 11 (2020), p. 100227 (cit. on pp. 43, 45).

[46] A. Nespoli, E. Ogliari, S. Leva, A. Massi Pavan, A. Mellit, V. Lughi, and A. Dolara. "Day-ahead photovoltaic forecasting: A comparison of the most effective techniques". In: *Energies* 12.9 (2019), p. 1621 (cit. on p. 43).

Dedications

I dedicate this work to the people of Iran, who, with courage and sacrifice, gave their lives in the enduring pursuit of freedom, justice, and hope for a brighter future.