Prediction of Electricity Consumption for Optimization of Shared Credit Systems Using Machine Learning.

Rhuan Lucas Alves Soares*, Luiz Felipe Pugliese*, Elcio Franklin de Arruda*, Giovani Bernardes Vitor*,
Rodrigo Aparecido da Silva Braga†
*Institute of Science and Technology, Federal University of Itajubá (Unifei), Itabira - MG, Brazil
Emails: {rhuanlucas123, pugliese, elcio.arruda, giovanibernardes}@unifei.edu.br
†Institute of Science, Technology and Innovation, Federal University of Lavras (UFLA), São Sebastião do Paraíso - MG, Brazil
Email: rodrigobraga@ufla.br

***Abstract*—Brazil has witnessed remarkable growth in distributed energy generation, driven by government incentives and the credit compensation market. This scenario, particularly within the solar energy sector, has led to the emergence of companies investing in large-scale areas for the installation of solar panels, enabling high-volume energy production and the sale of surplus. However, this exponential growth has introduced significant challenges in the photovoltaic sector, including the need to forecast plant generation, anticipate individual customer consumption, and optimize energy credit allocation. In this context, this work proposes the application of Machine Learning algorithms using Python and cloud resources with SQL queries—specifically through linear regression—to predict each customer's electricity consumption for the following month. The objective is to achieve the lowest possible error rate and thereby enable optimization of the credit allocation system.**

***Index Terms*—Distributed generation, credit clearing market, Machine Learning, cloud resource, consumption forecast.**

I. Introduction

According to [1], in 2012, Brazil had only six distributed generation units. However, by 2021, this number had increased dramatically, marking a 2000% growth. This exponential expansion is largely attributed to government incentives that significantly boosted the energy credit compensation market in the country.

Distributed generation refers to electricity produced at or near the point of consumption, as defined by [2]. When solar panel energy production exceeds local consumption, there are two options for selling the surplus. According to [3], the first option is to participate in auctions regulated by ANEEL or enter the free contracting market through solar power plants. The second alternative is to sell the generated solar energy back to the distribution grid, receiving energy credits in return. These credits are valid for up to 60 months (5 years) for future use.

With the continuous growth of the solar energy market, the creation of companies focused on selling solar energy in the free market has been envisioned. These companies invest in extensive land areas for the installation of solar panels, enabling large-scale energy production and making the sale of surplus energy feasible.

In this scenario, with the aim of improving operational efficiency and maximizing financial returns, several challenges have emerged in the photovoltaic energy sector. These challenges include forecasting the plant's energy generation, predicting each customer's consumption, and optimizing the allocation of energy credits. Thus, the objective of this study is to develop an on-premise [4] algorithm capable of predicting each customer's electricity consumption for the following month, while achieving the lowest possible error rate. Additionally, a model is proposed to be implemented on a cloud service provider platform for the same purpose, leveraging the robustness and efficiency offered by such platforms.

To achieve the project's objectives, Machine Learning (ML) techniques related to linear regression will be applied. These algorithms can identify patterns within the data and, based on the study's scope, make predictions using past consumption periods. Moreover, the linear regression model available on BigQuery (BQ)—a data analytics platform within Google Cloud Platform (GCP) [5]—will be implemented. Therefore, by employing ML techniques and tuning their parameters, the goal was to create a solution capable of predicting customer energy consumption with the lowest possible error rate. This led to the development of a program capable of generating plausible predictions given the constraints of the dataset used.

To further improve prediction accuracy and maximize efficiency, a new approach was adopted to handle missing data, replacing the previous method with interpolation techniques. Subsequently, GridSearchCV (GSCV) was employed—an algorithm that performs an exhaustive search over a grid of specified parameters using cross-validation to assess the performance of each parameter combination. This allowed for the selection of the most effective ML model configurations [6]. Finally, the two most effective techniques identified during the initial research were selected, along with an alternative solution provided by the GCP cloud service.

II. Related Works

A. Photovoltaic energy credit system

According to [7], the Electric Energy Compensation System (SCEE) is a framework in which a consumer unit generates more electricity than it consumes.

This dynamic also applies to photovoltaic energy, where electricity is produced through solar panels. As mentioned by [1], any surplus generated in this process and sold to the power distribution grid is classified as energy credits.

The SCEE received tax incentives to promote the adoption of sustainable practices, such as renewable energy sources. PIS and Cofins are taxes established by the Brazilian Federal Constitution, but under this system, customers are exempt from paying these fees [8].

Therefore, the credit allocation system consists of sharing and distributing the electricity generated in order to profit from the surplus energy produced by the consumer unit's solar panels. The system operates by collecting electricity generated by the panels at a central unit and then distributing it to individual consumer units according to specific rules and agreements.

Forecasting customer consumption plays a crucial role in the credit allocation system, as it enables efficient management of energy distribution per customer before actual consumption is recorded by the utility grid. In this way, companies can operate more efficiently, reduce unnecessary costs, and optimize the use of financial resources.

Despite its benefits, energy consumption forecasting presents challenges that affect its effectiveness in the context of the credit allocation system. These challenges include the lack or insufficiency of historical energy consumption data from customers, which undermines the ability to make accurate predictions that truly reflect each consumer's profile. Moreover, due to unexpected events over the years, such as the pandemic, the data often fails to follow predictable patterns, making it more difficult for algorithms to accurately forecast future consumption values.

B. Machine Learning

According to [9], with the increasing adoption of renewable energy in the global power grid, improving the accuracy of clean energy forecasts becomes essential for effective planning, management, and operation of the energy system. Thus, the application of advanced ML techniques, particularly linear regression, plays a crucial role in accurately predicting energy consumption, enabling the efficient Aintegration of renewable energy sources and the optimization of grid operations.

It can therefore be stated that the ever-evolving field of ML plays a fundamental role in a wide range of applications, and its impact on linear regression is particularly remarkable. The research and application of linear regression ML algorithms have experienced notable growth, generating valuable insights and substantial advancements across various disciplines such as economics, medicine, environmental sciences, telecommunications, and many others.

In this context, [10] employed an Artificial Neural Network (ANN) known as a Multi-Layer Perceptron (MLP). The author highlights that this technique was chosen for being relatively more efficient to build and train compared to more complex architectures, while also being capable of capturing complex and nonlinear relationships between numerical data. The objective of the study was to forecast the energy consumption of a building one day ahead, using historical energy demand data collected at 15-minute intervals from July 2014 to May 2016, in addition to temperature data from four locations within the building.

The Gradient Boosting Regressor (GBR) is a technique widely used in regression problems, including linear regression scenarios. It stands out for its high performance and stability in predictions [11]. However, to maximize its performance, it is essential to apply proper data preprocessing, ensuring the quality of input variables and the consistency of the values used during training.

In [12], ML techniques were also used for the same purpose: predicting electricity consumption. The algorithms applied included: K-Nearest Neighbors, XGBOOST, Random Forest (RF), and ANN. This study used one year of hourly energy consumption data. Additionally, [12] emphasizes that data preprocessing was necessary to handle missing values or outliers.

To test whether an ML model would be capable of yielding good results in the context of energy forecasting, [13] applied the Long Short-Term Memory model for training the dataset. According to [13], six years of electricity consumption data in Finland were used, forming a univariate time series with seasonal variations in the data.

Due to the infrastructure and the number of residents in buildings, [14] states that such structures must be energy-efficient and sustainable, as buildings significantly contribute to global energy consumption and greenhouse gas emissions. Therefore, to predict electricity consumption in buildings, [14] proposed a forecasting model using RF during both training and testing phases. The datasets were collected from five buildings as part of the Building Data Genome Project [15].

The following subsections explore the linear regression techniques used, that have significantly contributed to advancements in this field.

1) Gradient Boosting Regressor: The Gradient Boosting Regressor (GBR) is a tool also applied to problems involving linear regression, with the objective of predicting values over time. Its effectiveness becomes especially evident when temporal data exhibit complex behaviors, nonlinear trends, and seasonal patterns that are difficult to capture using conventional methods. However, to optimize its performance, it is essential to subject the data to appropriate preprocessing, which may include steps such as differencing and seasonal adjustments.

As demonstrated in [16], GBR shows robust performance in regression tasks, particularly after hyperparameter tuning. The model's superiority was established by outperforming a variety of other models when evaluated with standard metrics

such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2.

2) Multi-Layer Perceptrons: Feedforward Artificial Neural Networks, commonly referred to as MLPs (Multi-Layer Perceptrons), have been successfully applied to linear regression problems, offering a versatile approach to data analysis and forecasting.

MLPs consist of layers of interconnected neurons, including an input layer, one or more hidden layers, and an output layer. These networks are trained to learn patterns and dependencies within the data, making them adaptable to time series with various characteristics such as seasonality, trends, and irregular fluctuations. As stated in [17], this algorithm, due to the multilayer perceptron theory, is capable of more easily capturing the correlation between features in the dataset through linear transformation of the feature space.

In [18], a study was conducted to estimate energy consumption in Canadian manufacturing industries, and the MLP achieved the highest ranking. It was followed by the Radial Basis Function Network and the Support Vector Machine.

However, it is important to note that the performance of MLPs in linear regression tasks can depend on several factors, such as the quantity and quality of available data, the network architecture, model hyperparameters, and proper preprocessing techniques such as differencing or normalization.

C. Cloud Computing

Cloud computing refers to the use of computing resources such as servers, storage, and services provided over the Internet. Instead of maintaining physical servers locally, companies can utilize these resources from cloud providers such as Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP), Oracle, among others. This allows for scalable resources in a more cost-effective manner, flexibility to access servers, machines, or other tools from anywhere, and cost reduction since resources can be scaled as needed [19].

The most commonly used cloud computing service models are IaaS (Infrastructure as a Service), PaaS (Platform as a Service), SaaS (Software as a Service) [20], CaaS (Container as a Service), and FaaS (Function as a Service). Each of these services differs in terms of responsibility distribution between the consumer and the provider. In the IaaS model, the user is responsible for most of the operational infrastructure, while in SaaS, the provider takes over most responsibilities by offering ready-to-use software solutions.

The Pay as you go (PAYG) model is widely used in cloud computing. This system involves billing the consumer based on their usage of the platform's available resources. Therefore, if a consumer keeps a resource reserved in the cloud—even without actively using it—charges will still apply, as the resource is being held for them.

BigQuery is a tool for data analysis and SQL queries [21]. While it is widely used for querying large datasets, it can also be integrated as part of a machine learning pipeline. BigQuery is categorized as a PaaS, which means it provides a managed environment that allows users to develop, run, and manage applications without the complexity of building and maintaining the infrastructure typically associated with developing and launching an app.

One of the main advantages of this tool is its scalability. It can handle large volumes of data and automatically distributes queries across multiple machines, speeding up processing. In addition, BQ ML [22] is an extension that allows users to create, train, and evaluate machine learning models directly within the query environment.

According to [5], there is no need to be proficient in commonly used programming languages for data analysis or data science. In this case, the steps of model creation, training, validation, and prediction can be performed through SQL queries.

With this resource extension, algorithms such as linear regression, logistic regression, decision trees, and k-means can be used to train models using the data stored in BQ. This eliminates the need to transfer data to another platform and simplifies the machine learning workflow.

D. Hyperparameter Tuning

Hyperparameter tuning is the process of adjusting the values of a model's hyperparameters with the aim of achieving the best possible performance and consequently reducing the error rate during training. During the development of ML models, this process is crucial. When a model is trained, it is often necessary to fine-tune these hyperparameters to obtain more accurate results. These are configurable values that are not learned directly by the model.

GridSearchCV (Grid Search Cross-Validation or GSCV) is a technique used to find the best hyperparameter values in ML models [6]. It explores all possible combinations of specified values for the hyperparameters, generating various options. It then evaluates the model's performance for each combination using cross-validation. At the end of the search, the algorithm returns the best hyperparameter values based on some evaluation metric, such as R^2 score, maximum residual error, mean absolute error (MAE) loss, among others [23], commonly used for linear regression models.

In this study, the R^2 score, also known as the coefficient of determination, was selected. This metric ranges from 0 to 1, where 1 indicates that the model is capable of explaining all the variability of the response data around its mean.

In this process, the objective is to find the model that best fits the used data. Therefore, by using R^2, the goal is to find the hyperparameter combination that produces the linear regression model capable of explaining the highest proportion of variance in the dependent variable. This enhances the model's accuracy and effectiveness.

As a demonstration, Table I presents, using alphabetical examples and hyperparameters represented as Hp_x where $x = A, B, C$, a representation of the value grid used in the tuning process. Once the process is completed, Table II displays the best hyperparameter set returned for the selected model, based on the maximization of the R^2 score evaluation metric.

TABLE I: List of hyperparameter values for tuning.

	Hp_A	Hp_B	Hp_C
Value 1	X_1	Y_1	Z_1
Value 2	X_2	Y_2	Z_2
Value 3	X_3	Y_3	Z_3

TABLE II: Best hyperparameter set.

	Hp_A	Hp_B	Hp_C
Final values	X_2	Y_1	Z_3

III. Methodology

The research methodology used was Design Science Research (DSR). DSR is a research methodology focused on the creation and evaluation of artifacts to solve practical and relevant problems. According to [24], the evaluation of design artifacts and design theories plays a fundamental role in research, as it not only informs future development but also, when properly conducted, ensures the quality and soundness of the research. Accordingly, the steps of this methodology can be visualized in Figure 1.

Considering the proposed context, each stage of the DSR and its implementation in this work is described below:

Problem identification: In the given context, the central challenge consists of forecasting customer energy consumption in order to optimize the distribution of energy credits.

Problem awareness: Given the problem, the process begins by recognizing the challenge posed by the available data, where many customers lack standardized information or extensive histories, making forecasting inaccurate. Furthermore, the dynamic nature of energy consumption, influenced by seasonal factors and unpredictable events, adds another layer of complexity to the problem.

Systematic literature review: To address this issue, a systematic literature review was conducted. This review analyzed several studies related to energy consumption forecasting in compensation scenarios, with a particular focus on those that faced similar challenges. The analysis of these studies enabled the identification of the most promising techniques and best practices used to deal with the lack of standardized data and the inherent volatility in energy consumption, as detailed in Section II-B.

Identification of artifacts and problem class configuration: With the problem and its context defined, the identified artifacts were: (i) linear regression techniques implemented both locally and using a cloud service provider, (ii) training datasets, and (iii) a hyperparameter tuning algorithm for the models. The problem classes included: handling missing data, short- or long-term forecasting, unstandardized data, and seasonality.

Artifact proposal to solve the specific problem: The main proposed artifacts include preprocessing of missing and unstandardized data, implementation of an algorithm incorporating various linear regression techniques using Python as the programming language to explore different forecasting approaches, fine-tuning of hyperparameters using an optimization algorithm, the application of a data analysis platform with embedded linear regression models, and SQL [25] for querying the platform. These steps aim to improve forecast quality, ensure data consistency, and maximize model accuracy.

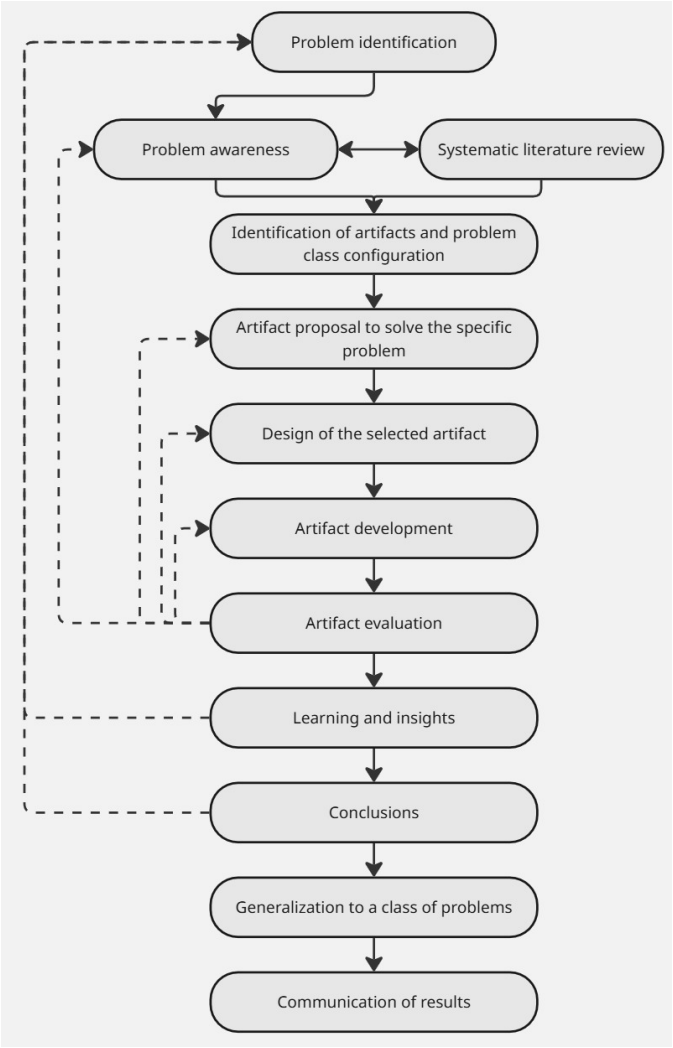

Fig. 1: Design Science Research methodology.

Design of the selected artifact: The artifact design consists of an algorithm using Python and SQL queries, along with libraries such as Pandas and the BigQuery client library [21], for consumption prediction using various linear regression techniques, a hyperparameter optimization algorithm for the mentioned techniques, and model creation on a cloud service provider as an alternative prediction method.

Artifact development: The development process of the algorithms began with data acquisition, including details about how the data were obtained, their characteristics, and the types of customers (new and existing). As for the tools used, a specific programming language, platform, and libraries were employed. Each ML technique was detailed, including the best hyperparameters found and the improvements made. The main objective was to forecast the next month's energy

consumption.

Artifact evaluation: The evaluation of the developed artifact demonstrated the algorithm's efficiency in the proposed context. The primary goal of predicting customer consumption was successfully achieved, and the obtained results are promising. Overall, the developed artifact proved to be a solid and effective solution for forecasting energy consumption in the energy compensation market.

Learning and insights: The artifact development process highlighted the importance of incremental evolution throughout the project. From the initial phase to final implementation, there was clear progress in the accuracy and effectiveness of the consumption forecasts. This demonstrates that the iterative development approach adopted in the project was appropriate. Each iteration allowed for the identification and resolution of specific issues, understanding how hyperparameter changes affected results, and refinement of applied techniques. At each step, the results were analyzed and used as a basis for subsequent improvements. These incremental evolutions testify to the flexibility and adaptability of the development methodology, enabling adjustments as new insights were gained. As a result, the artifact reached a solid performance level, and this continuous improvement approach proves valuable for future projects in the field of energy consumption forecasting in time series.

Conclusions: The main conclusions include more accurate and reliable consumption forecasts for each customer following the application of linear regression techniques with hyperparameter tuning and the alternative model in the cloud provider. Moreover, the algorithm performed acceptably even when customers lacked sufficient data or when the data were not standardized. However, results could be further improved if all customers had adequate data.

Generalization to a class of problems: There are numerous problems where the developed algorithm could be applied, such as sales forecasting, exchange rate prediction, and water consumption forecasting, among others. In each of these cases, the linear regression algorithm can be adapted and configured to find the most suitable technique that produces accurate and reliable forecasts, contributing to informed and effective decision-making in various contexts.

Communication of results: The research materials used during the project development will be made available to disseminate the problem, encouraging further research, and, if necessary, the pseudocode of the applied algorithms will also be shared.

IV. System Implementation

In Section III, the problem was conceptualized, providing a broader view of what will be developed in the current implementation section.

The data used for this research is provided by the distribution network in an XML document. Thus, the customer's consumption value for the corresponding month is extracted and added to the dataset.

This dataset is also updated when new customers are added, as they will also be included in the prediction process, even if they do not yet have sufficient historical data.

Figure 2 illustrates all the steps of the process flow used throughout the project. This structure covers the entire process, from raw data to the monthly prediction for each customer.

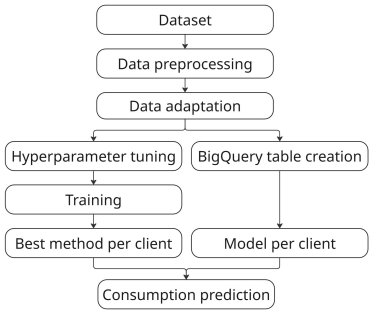

Fig. 2: Process flow.

A. Dataset

The structure begins with the definition of the project's main dataset. It contains crucial information about the monthly energy consumption of the involved customers, data that is necessary for the interpolation technique discussed in Section IV-B.

B. Data Preprocessing

Data preprocessing was necessary because many customers had months with missing information, and the column names were complex. There are several ways to perform data preprocessing, depending on the data format (text, images, etc.).

Therefore, the preprocessing of the dataset was divided into the following steps:

1) Remove rows with duplicate customers;
2) If the customer has a missing first consumption value, generate a fictitious first consumption, being a random number between the customer's minimum and maximum recorded consumption;
3) Reset the dataset index;
4) Remove customers with no valid consumption data;
5) Fill in missing data using the interpolation technique;
6) Remove the fictitious consumption value created in step 2;
7) Normalize the consumption values;
8) Format the columns by renaming them.

C. Data Adaptation

After filling and performing the essential data manipulations, the dataset was adjusted to suit the techniques selected prior to Section IV-D. Table III shows, with numerical examples, how the data was adapted to be used in the selected techniques. This process involves predicting consumption for the subsequent month within a predefined limit.

Initially, the model is trained with data from month 1 to month N and is used to predict consumption for month N+1.

In the next iteration, the training interval shifts to start from month 2 to month N+1, and the model is then used to predict month N+2. This process continues, adjusting for each new iteration, allowing the model to be trained with the most recent data and predict the upcoming month.

TABLE III: Training model (consumption in kWh).

Month 1	Month 2	Month ...	Month N	Month to Predict
2.263	1.989	...	1.762	1.730
1.989	2.277	...	1.730	1.587

To perform the prediction on the GCP data analysis platform, the aforementioned adjustment was not necessary. The dataset was structured as shown in Table IV. In this format, the first row labeled "Months" represents the independent variable, corresponding to sequential time indices. The second row labeled "Consumption" contains the dependent variable, representing the actual electricity consumption (in kWh) for each respective month. The model is trained using the pairs (Month, Consumption), and the goal is to predict the consumption for month N+1.

TABLE IV: Training style 2 (consumption in kWh).

Months	1	2	...	N+1
Consumption	2.263	1.989	...	1.730

D. Hyperparameter Tuning

After adapting the data, the GridSearchCV algorithm was applied to tune the hyperparameters of each technique in order to achieve predictions with lower error rates. The techniques used for tuning are those described in subsections II-B1 and II-B2. As a result, a dataset was generated that presented error rates (absolute and percentage) along with the actual and predicted values for each locally used linear regression model. Furthermore, the MAE was calculated to evaluate the model's performance [26].

E. Training

With Sections IV-C and IV-D completed, the training stage was initiated. Training was performed using the adjusted input data and the best-found hyperparameters. The goal in this phase was to minimize the error for each technique.

F. Best Technique per Client

Using the dataset generated in Section IV-D, it was possible to verify the error rates per customer for each technique. Since the prediction values varied depending on each customer's consumption, it was decided to select the technique with the lowest error rate for each customer.

G. BigQuery table creation

Seeking an alternative using a cloud service provider's resource, a table was created in BQ and filled with the dataset resulting from Section IV-A.

H. Model per client

As energy consumption varied from customer to customer, a specific prediction model was created for each one. After the dataset was prepared and loaded into BigQuery, a linear regression model was trained individually for each customer using SQL queries. These queries filtered the dataset by customer ID and applied the CREATE MODEL statement from BigQuery ML to generate a separate regression model for each case. This approach allowed the prediction to be tailored to the specific consumption patterns of each client.

I. Consumption Prediction

After Section IV-F, it became possible to identify the most effective technique within the current dataset, specifically tailored to each customer. Consequently, the energy consumption prediction for all customers was carried out using the most appropriate method for each one. This way, it is expected to provide energy consumption forecasts that accurately reflect each consumer's individual reality.

V. Results and Discussion

A. Metric for Evaluating Results

Section IV-D mentioned values that needed to be calculated. Therefore, this section presents the formulas used.

ϵ_{ab}: is the absolute error obtained by the technique during prediction. It is calculated by subtracting the predicted consumption value from the actual consumption value.

$$\epsilon_{ab} = Actual_{value} - Predicted_{value} \tag{1}$$

$\epsilon_{\%}$: is the percentage error obtained by the technique during prediction. The absolute error is divided by the actual consumption value and multiplied by 100.

$$\epsilon_{\%} = \frac{\epsilon_{ab}}{Actual_{value} * 100} \tag{2}$$

MAE: is the Mean Absolute Error obtained by the technique during prediction. It is calculated by summing all absolute errors and dividing by the number of samples (n) used [26].

$$\text{MAE} = \frac{\sum_{i=1}^{n} \epsilon_{ab_i}}{n} \tag{3}$$

$Classif_{\epsilon}$: Table V shows the classification of error levels, which is based on the absolute error.

TABLE V: Error level classification.

$Classif_{\epsilon}$	Values (in kWh)
Low	$\epsilon_{absolute} < 1000$
Medium	$1000 < \epsilon_{absolute} < 2000$
High	$2000 \leq \epsilon_{absolute}$

These thresholds were defined based on the suggestion of a specialist in energy-sharing systems using solar credit compensation.

TABLE VI: Calculation of MAE by technique.

Techniques	MAE (in kWh)
BigQuery	552.02
MLP with GSCV	592.60
MLP without GSCV	723.90
GBR with GSCV	651.90
GBR without GSCV	859.50

B. Results and Analysis

With the chosen techniques and the developed algorithm, it was possible to visualize the results. Table VI shows the MAE for each technique after consumption prediction.

From Table VI, it is evident that the technique employed on the GCP data analytics platform yielded superior results in terms of client consumption prediction. Additionally, using the hyperparameter tuner allowed the techniques to achieve lower MAE values, justifying its implementation to reduce model error rates.

After the hyperparameter tuning process, different values were observed compared to those used previously. In the MLP technique, the hyperparameters `hidden_layer_sizes`, `activation`, and `alpha` showed the most significant changes after the process in Section IV-D.

The `hidden_layer_sizes` parameter determines the number of neurons in each hidden layer of the neural network. Changes in size and architecture affect the model's ability to learn complex data patterns.

The `activation` value defines the activation function used in the hidden layers. Each function has unique properties and influences the training process differently.

The `alpha` parameter controls the regularization applied to the neural network weights to prevent overfitting. A higher `alpha` increases the penalty, balancing data fitting and model complexity.

A similar tuning process occurred for the GBR technique. Noticeable changes were observed in the parameters `n_estimators`, `learning_rate`, and `min_samples_split`.

The `n_estimators` parameter specifies the number of boosting stages. More stages can improve learning but may also increase the risk of overfitting if excessive.

The `learning_rate` adjusts the contribution of each tree in the model. A smaller value reduces the influence of individual trees, requiring more trees to learn the data.

The `min_samples_split` defines the minimum number of samples required to split an internal node. A higher value can prevent overfitting but may also limit the model's ability to capture complex patterns.

Figures 3 and 4 illustrate predictions for different clients. Figure 3 shows that although BQ reported a lower MAE value compared to other techniques, it exhibited the worst prediction performance.

In contrast, Figure 4 presents a different scenario. In this case, most techniques adapted efficiently to the data, resulting in predicted values that were closely aligned.

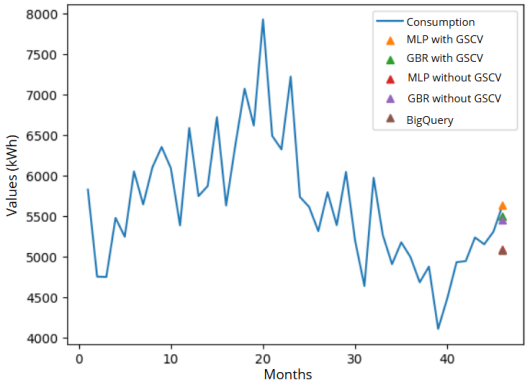

Fig. 3: Prediction for Client 1.

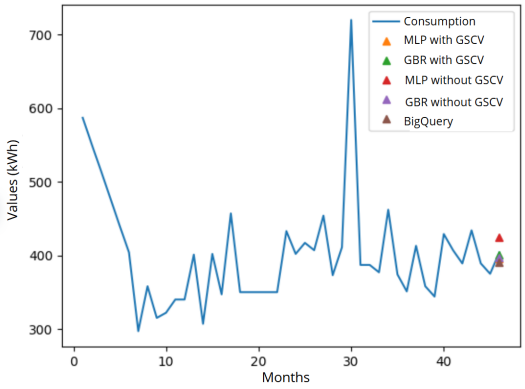

Fig. 4: Prediction for Client 2.

Regarding processing time, in local environments, exhaustive hyperparameter tuning can be extremely slow and resource-intensive due to hardware limitations and the need for multiple training and validation iterations. As seen in Section II-C, BQ leverages Google's cloud computing infrastructure, enabling large-scale parallel processing and access to virtually unlimited resources. This results in faster model optimization and training, significantly reducing the overall time required to develop and deploy machine learning models.

To define which technique should be used for each client, several steps were followed, from data loading and preprocessing to the final energy consumption prediction using the technique with the lowest error rate for each specific client.

VI. Conclusion

The growth of distributed solar energy generation in Brazil, driven by government incentives and the energy credit compensation market, has introduced significant challenges, such as the need to forecast energy generation, anticipate customer consumption, and optimize the allocation of credits.

This work proposed the application of Machine Learning algorithms—specifically linear regression—using Python and cloud-based resources with SQL queries to predict the electricity consumption of each customer for the following month. The results demonstrated that the use of machine learning

techniques can be more effective in forecasting electricity consumption when either hyperparameter tuning algorithms are applied or cloud computing resources are leveraged. This provides a valuable tool for energy credit management and contributes to the operational efficiency of companies in the photovoltaic sector.

Therefore, the implementation of ML-based solutions for consumption prediction proves to be a viable approach to address the challenges posed by the rapid growth of distributed energy generation in Brazil. This approach not only improves the accuracy of consumption forecasts but also facilitates the efficient management of energy resources, contributing to a more sustainable and energy-balanced future.

REFERENCES

[1] C. H. Flesch, C. A. Cambani, P. G. Dallepiane, L. N. Canha, D. C. Heman, E. D. Garcia, J. B. Parizzi, L. Losinkas, and M. d. S. Carvalho, "Análise financeira da energia fotovoltaica no mercado livre de energia," *Congresso Brasileiro de Energia Solar*, 2022, accessed: May 8, 2025. [Online]. Available: https://doi.org/10.59627/cbens.2022.1166

[2] P. Solar. Geração distribuída de energia (gd): o que é, regras, benefícios e como fazer parte. Accessed: May 8, 2025. [Online]. Available: https://www.portalsolar.com.br/geracao-distribuida-de-energia.html

[3] ——. Como vender energia solar. Accessed: May 8, 2025. [Online]. Available: https://www.portalsolar.com.br/como-vender-energia-solar

[4] S. Younus, K. Kumar, I. Ali, A. Laghari, and A. Ali, "Systematic analysis of on premise and cloud services," *International Journal of Cloud Computing*, vol. 13, pp. 214–242, 06 2024.

[5] "Machine learning in google cloud big query using sql," *International Journal of Computer Science and Engineering*, 2023, accessed: May 8, 2025. [Online]. Available: https://doi.org/10.14445/23488387/IJCSE-V10I5P103

[6] K. Alemerien, S. Alsarayreh, and E. Altarawneh, "Diagnosing cardiovascular diseases using optimized machine learning algorithms with gridsearchcv," *Journal of Applied Data Sciences*, vol. 5, no. 4, pp. 1539–1552, 2024. [Online]. Available: https://bright-journal.org/Journal/index.php/JADS/article/view/280

[7] L. S. Monteiro, "Estudo para a utilização de energia solar no âmbito do sistema de compensação de energia elétrica no brasil," 2014, accessed: May 8, 2025. [Online]. Available: https://sistemabu.udesc.br/pergamumweb/vinculos/000000/000000f0.pdf

[8] H. C. Camargo, "Efetividade dos incentivos fiscais concedidos ao sistema de compensação de energia solar como forma de estímulo ao desenvolvimento sustentável," 2018, accessed: May 8, 2025. [Online]. Available: https://repositorio.jesuita.org.br/handle/UNISINOS/7328

[9] H. Wang, Z. Lei, X. Zhang, B. Zhou, and J. Peng, "A review of deep learning for renewable energy forecasting," *Energy Conversion and Management*, vol. 198, p. 111799, 2019, accessed: May 8, 2025. [Online]. Available: https://doi.org/10.1016/j.enconman.2019.111799

[10] M. H. Lin. (2025) Predicting energy demand with neural networks. Accessed: May 8, 2025. [Online]. Available: https://towardsdatascience.com/forecasting-energy-consumption-using-neural-networks-xgboost-2032b6e6f7e2

[11] H. Wang and G. Gu, "Wavelet gradient boosting regression method study in short-term load forecasting," *Smart Grid*, vol. 5, no. 4, pp. 189–196, 2015, accessed: July 19, 2025.

[12] C. P. N. S. K. R. S. M. G. Gurram Vijendar Reddy, Lakshmi Jaswitha Aitha, "Electricity consumption prediction using machine learning," 2023.

[13] M. Nachawati. Energy consumption prediction using machine learning. Accessed: May 8, 2025. [Online]. Available: https://github.com/MohamadNach/Machine-Learning-to-Predict-Energy-Consumptionenergy-consumption-prediction-using-machine-learning

[14] A.-D. Pham, N.-T. Ngo, T. T. Ha Truong, N.-T. Huynh, and N.-S. Truong, "Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability," *Journal of Cleaner Production*, vol. 260, p. 121082, 2020, accessed: May 8, 2025. [Online]. Available: https://doi.org/10.1016/j.jclepro.2020.121082

[15] C. Miller and F. Meggers, "The building data genome project: An open, public data set from non-residential building electrical meters," *Energy Procedia*, vol. 122, pp. 439–444, 2017, accessed: May 8, 2025. [Online]. Available: https://doi.org/10.1016/j.egypro.2017.07.400

[16] M. Samuel, O. Simon, I. Ndibatya, N. Kimbugwe, and G. Marvin, "Transparent multi-strategy learning for bike distribution forecasting in localised operational environments," in *Data Science and Applications*, S. J. Nanda, R. P. Yadav, A. H. Gandomi, and M. Saraswat, Eds. Singapore: Springer Nature Singapore, 2025, pp. 413–426.

[17] Y. Liu, C. Zhao, and Y. Huang, "A combined model for multivariate time series forecasting based on mlp-feedforward attention-lstm," *IEEE Access*, vol. 10, pp. 88 644–88 654, 2022, accessed: May 8, 2025. [Online]. Available: https://doi.org/10.1109/access.2022.3192430

[18] O. A. Olanrewaju and C. Mbohwa, "Comparison of artificial intelligence techniques for energy consumption estimation," in *2016 IEEE Electrical Power and Energy Conference (EPEC)*, 2016, pp. 1–5, accessed: May 8, 2025. [Online]. Available: https://doi.org/10.1109/epec.2016.7771702

[19] A. C. M. Paz and M. J. Loos, "A importância da computação em nuvem para a indústria 4.0." [Online]. Available: http://dx.doi.org/10.3895/gi.v16n2.9317

[20] I. Ashraf, "An overview of service models of cloud computing," *International Journal of Multidisciplinary and Current Research*, vol. 2, no. 1, pp. 779–783, 2014.

[21] G. Cloud. Bigquery api client libraries. Accessed: May 8, 2025. [Online]. Available: https://cloud.google.com/bigquery/docs/reference/libraries?hl=pt-brclient-libraries-install-python

[22] ——. Introduction to ai and ml in bigquery. Accessed: May 16, 2025. [Online]. Available: https://cloud.google.com/bigquery/docs/bqml-introduction

[23] S. Learn, "Metrics and scoring: quantifying the quality of predictions," accessed: May 8, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/model$_evaluation.htmlscoring - parameter$

[24] J. P.-H. John Venable and R. Baskerville, "Feds: a framework for evaluation in design science research," *European Journal of Information Systems*, vol. 25, no. 1, pp. 77–89, 2016, accessed: May 8, 2025. [Online]. Available: https://doi.org/10.1057/ejis.2014.36

[25] G. Cloud. Introduction to sql in bigquery. Accessed: May 8, 2025. [Online]. Available: https://cloud.google.com/bigquery/docs/introduction-sql

[26] T. O. Hodson, "Root-mean-square error (rmse) or mean absolute error (mae): when to use them or not," *Geoscientific Model Development*, vol. 15, no. 14, pp. 5481–5487, 2022, accessed: May 8, 2025. [Online]. Available: https://gmd.copernicus.org/articles/15/5481/2022/