



Universidad Andrés Bello®

Electric Coordinator Project

Authors:

Trureo Vicente

Mahaluf Rodrigo

Subject:

Machine Learning

Teacher:

Pablo Schwarzenberg

Date:

May 5th, 2024

Table of Contents

Introduction.....	2
Bibliographic References.....	3
Dataset Description.....	4
Data Quality Study.....	5
Descriptive Statistics.....	8
Relevant Graphs.....	9
Feature Selection Proposal.....	13
Dataset Generation Procedure.....	14
Model Evaluation Metrics.....	14
Machine Learning Technique.....	15
Description of the Neuronal Model.....	15
Description of the features comprising the model input.....	16
Description of the parameters used to train the model.....	16
Description of the architecture of the model used.....	17
The results obtained.....	18
Conclusions.....	23
Recommendations.....	24

Introduction

In this project, we face the challenge of predicting electricity consumption in three specific zones of the National Electric System (SEN) of Chile. The ability to accurately anticipate electricity consumption is fundamental for efficient management of the electric grid. The overall objective is to develop a disaggregated electricity consumption forecasting model based on Deep Learning, with a forecast horizon of 7 days and hourly resolution for the central (Santiago), north (Iquique and Antofagasta), and south (from Valdivia southward) zones of the SEN.

To achieve this objective, we have several specific goals. Firstly, we will construct a consumption data model for each of the three zones of the SEN, disaggregated by bars. Additionally, we will judiciously select a multivariable Machine Learning or Deep Learning algorithm, based on a comprehensive review of the state of the art. This algorithm should be capable of handling complex data and capturing non-linear relationships between variables, considering at least consumption per bar, temperatures, and type of day (working day, holiday, special, among others).

Once the appropriate algorithm is selected, we will develop and implement the forecasting model for each of the three zones of the SEN. Subsequently, we will evaluate the performance of the generated forecasts using error metrics and compare them with the forecasts currently used by the SEN Coordinator.

Bibliographic References

We have identified two articles that have some relevance to what we intend to accomplish:

Al Metrik, M. A., & Musleh, D. A. (2022). Machine learning empowered electricity consumption prediction. *Computers, Materials & Continua*, 72(1), 1427–1444.

This study focuses on predicting electricity consumption in Saudi Arabia using Artificial Neural Networks (ANN) and the Bagging Ensemble method. The authors utilize an extensive dataset collected from the Saudi Electric Company and demonstrate the effectiveness of the Bagging Ensemble approach in improving the accuracy of electricity consumption predictions.

Oh, S., Oh, S., Shin, H., Um, T., & Kim, J. (2023). Deep Learning Model Performance and Optimal Model Study for Hourly Fine Power Consumption Prediction. *Electronics*, 12(16), 3528. <https://doi.org/10.3390/electronics12163528>

In this article, the authors address electricity consumption prediction using deep neural network models. They conduct experiments to predict electrical demand at specific time intervals and compare the performance of various machine learning models. The results highlight the importance of accuracy in electricity consumption prediction for efficient energy management.

These two articles are related to the project as they are predictive models of electricity consumption for a specific area or using deep learning-based models, which is why they are mentioned, in addition to providing interesting insights that may help us progress along our path.

Dataset Description

First of all, it is necessary to clarify that the complete dataset ranged from the year 2017 to 2023. However, due to many missing months of data in these years, they were excluded from the experiment, and only the first week of available data in 2023 was considered for the final activity of this document.

a. Columns:

- The first column is "substation," which contains the name of the substation.
- The next four columns are "year," "month," "day," and "hour," representing the date and time of the measurement.
- The last column is "consumption," which contains energy consumption measurements in each record.

b. Number of records:

- The dataset has a total of 1,789,553 records.

c. Missing values:

- No missing values are observed in any of the columns.

d. Descriptive statistics:

- The most frequent substation is "CNAVIA."
- The year of the records ranges from 2018 to 2022.
- The month of the records ranges from 1 to 12.
- The day of the records ranges from 1 to 31.
- The hour of the records ranges from 0 to 23.
- Energy consumption has a minimum value of -377.6391 and a maximum value of 319.9903.

e. Data types:

- The "substation" column is of object type.
- The "year," "month," "day," and "hour" columns are of integer type.
- The "consumption" column is of float type.

Data Quality Study

Missing Values: No missing values were found in any of the columns of the dataset. This suggests that there is no incomplete data that need to be addressed before analysis.

```
[ ] missing_values = data.isnull().sum()
    print("Valores faltantes por columna:\n", missing_values)
```

```
Valores faltantes por columna:
substation      0
year            0
month           0
day             0
hour            0
consumption     0
dtype: int64
```

Image 1: Missing Values

Duplicated Values: The duplicated values found are due to repeated data given the way the data is arranged, for example, for each hour we have four different consumption samples but they are the four taken in hour "1". This can be observed in Image 3.

```
[ ] duplicate_values = data.duplicated().sum()
    print("Cantidad de valores duplicados:", duplicate_values)

Cantidad de valores duplicados: 16364
```

Image 2: Duplicated Values

	substation	year	month	day	hour	consumption
1	AJAHUEL	2018	1	1	0	-28.761417
2	AJAHUEL	2018	1	1	0	94.511155
3	AJAHUEL	2018	1	1	0	100.250373
4	AJAHUEL	2018	1	1	0	-28.669271
5	AJAHUEL	2018	1	1	1	-26.769412
6	AJAHUEL	2018	1	1	1	85.91531
7	AJAHUEL	2018	1	1	1	91.790629
8	AJAHUEL	2018	1	1	1	-26.890978
9	AJAHUEL	2018	1	1	2	-27.283222
10	AJAHUEL	2018	1	1	2	83.372304
11	AJAHUEL	2018	1	1	2	89.421208
12	AJAHUEL	2018	1	1	2	-27.372479
13	AJAHUEL	2018	1	1	3	-28.469701
14	AJAHUEL	2018	1	1	3	-28.393221
15	AJAHUEL	2018	1	1	3	80.109874
16	AJAHUEL	2018	1	1	3	84.652547
17	AJAHUEL	2018	1	1	4	-26.645466
18	AJAHUEL	2018	1	1	4	-26.544979
19	AJAHUEL	2018	1	1	4	74.917414
20	AJAHUEL	2018	1	1	4	78.6351
21	AJAHUEL	2018	1	1	5	-26.5462
22	AJAHUEL	2018	1	1	5	-26.443638
23	AJAHUEL	2018	1	1	5	71.5948
24	AJAHUEL	2018	1	1	5	74.912173
25	AJAHUEL	2018	1	1	6	-26.398043
26	AJAHUEL	2018	1	1	6	-26.32667
27	AJAHUEL	2018	1	1	6	67.754026
28	AJAHUEL	2018	1	1	6	71.49263
29	AJAHUEL	2018	1	1	7	-26.750084
30	AJAHUEL	2018	1	1	7	-26.682251

Image 3: Data

Data Types: Most of the columns have appropriate data types for their content. Numeric variables are represented as integers (int64) or floats (float64), while date and time are represented as objects. However, it would be recommended to convert the date and time column to a suitable date and time format to facilitate temporal analysis.

```
[ ] data_types = data.dtypes
    print("Tipos de datos por columna:\n", data_types)

Tipos de datos por columna:
substation      object
year            int64
month           int64
day             int64
hour            int64
consumption     float64
dtype: object
```

Image 4: Data Types

Unique Values: A variety of unique values is observed in each column. For example, the 'substation' column has 7 unique values, indicating the presence of data from 7 different substations. The 'year' column has 5 unique values, representing data from 5 different years. The 'month' column has 12 unique values, suggesting data from all 12 months of the year. The 'day' column has 31 unique values, indicating data covering all days of the month. The 'hour' column has 24 unique values, representing data from all 24 hours of the day. Lastly, the 'consumption' column has 1,583,210 unique values, indicating a wide range of consumption levels recorded in the dataset.

```
[ ] unique_values = data.nunique()
    print("Valores únicos por columna:\n", unique_values)

Valores únicos por columna:
substation      7
year            5
month           12
day             31
hour            24
consumption     1583210
dtype: int64
```

Image 5: Unique Values

Descriptive Statistics

The descriptive statistics provide insight into the distribution and variability of the data in each column. For example, considering the 'year' column, the mean year is approximately 2020, with a standard deviation of about 1.45, indicating the spread of data around the mean. The 'month' column has a mean of approximately 6.41, suggesting that the data spans across multiple months, with a standard deviation of around 3.41. The 'day' column has a mean of approximately 15.70, indicating that the data covers various days of the month, with a standard deviation of approximately 8.80. The 'hour' column has a mean of approximately 11.50, representing the distribution of data across different hours of the day, with a standard deviation of about 6.92. Lastly, the 'consumption' column has a mean of approximately 32.52, with a standard deviation of approximately 74.14, indicating the variability in energy consumption levels across the dataset. Additionally, the descriptive statistics provide information on the minimum, maximum, and quartile values, which further describe the distribution of data.

```
[ ] descriptive_stats = data.describe(include='all')
    print("Estadísticas descriptivas:\n", descriptive_stats)
```

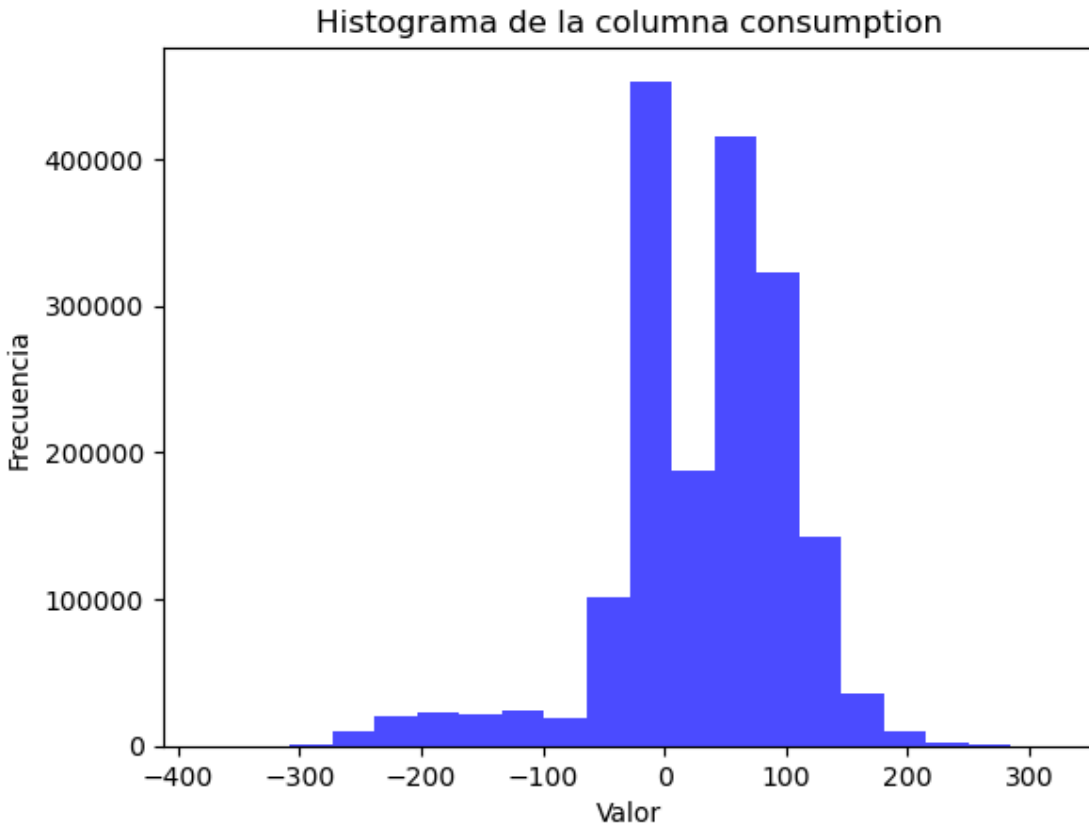
```
Estadísticas descriptivas:
      substation      year      month      day      hour \
count      1789553  1.789553e+06  1.789553e+06  1.789553e+06  1.789553e+06
unique         7         NaN         NaN         NaN         NaN
top      CNAVIA         NaN         NaN         NaN         NaN
freq      412974         NaN         NaN         NaN         NaN
mean         NaN  2.020094e+03  6.412154e+00  1.570166e+01  1.150212e+01
std         NaN  1.445269e+00  3.412864e+00  8.795105e+00  6.922741e+00
min         NaN  2.018000e+03  1.000000e+00  1.000000e+00  0.000000e+00
25%         NaN  2.019000e+03  4.000000e+00  8.000000e+00  6.000000e+00
50%         NaN  2.020000e+03  6.000000e+00  1.600000e+01  1.200000e+01
75%         NaN  2.021000e+03  9.000000e+00  2.300000e+01  1.800000e+01
max         NaN  2.022000e+03  1.200000e+01  3.100000e+01  2.300000e+01

      consumption
count      1.789553e+06
unique         NaN
top         NaN
freq         NaN
mean      3.252468e+01
std       7.413949e+01
min      -3.776391e+02
25%      -2.807174e+00
50%       4.452032e+01
75%       8.163271e+01
max       3.199903e+02
```

Image 6: Descriptive Stats

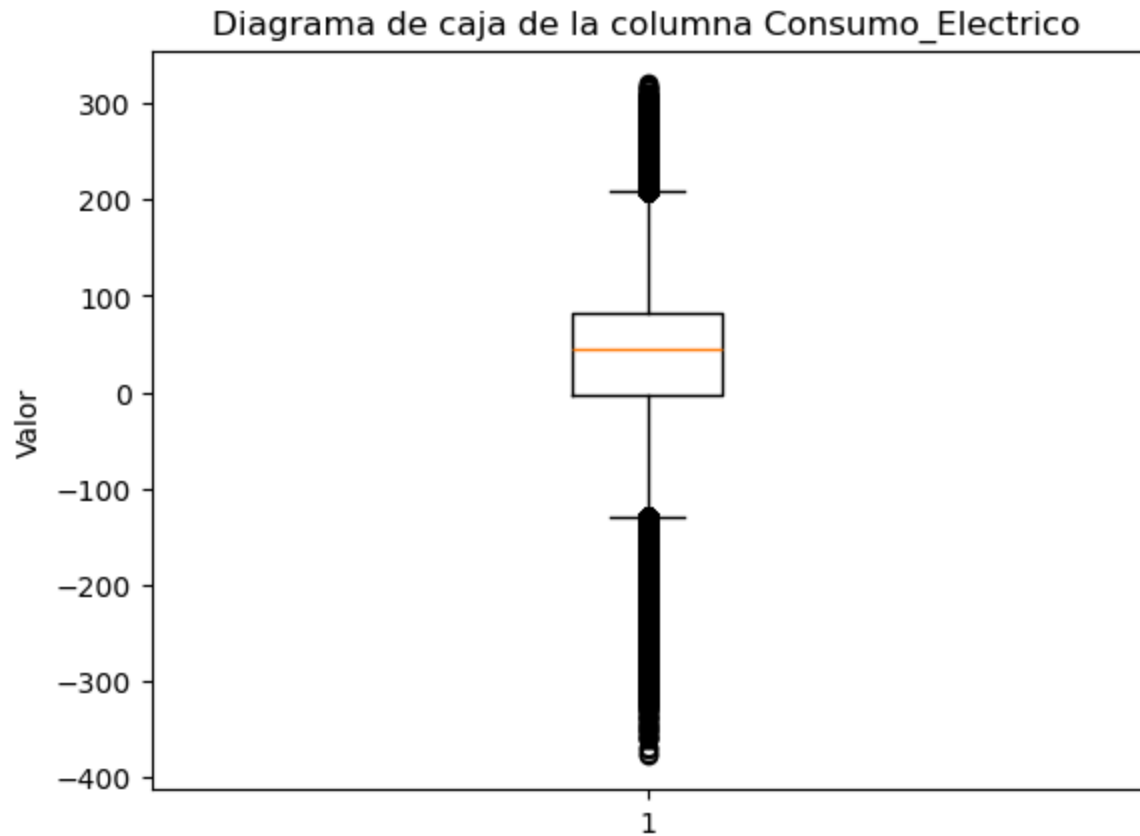
Relevant Graphs

The following graphs show us different aspects that provide us with an important initial analysis for the following stages.



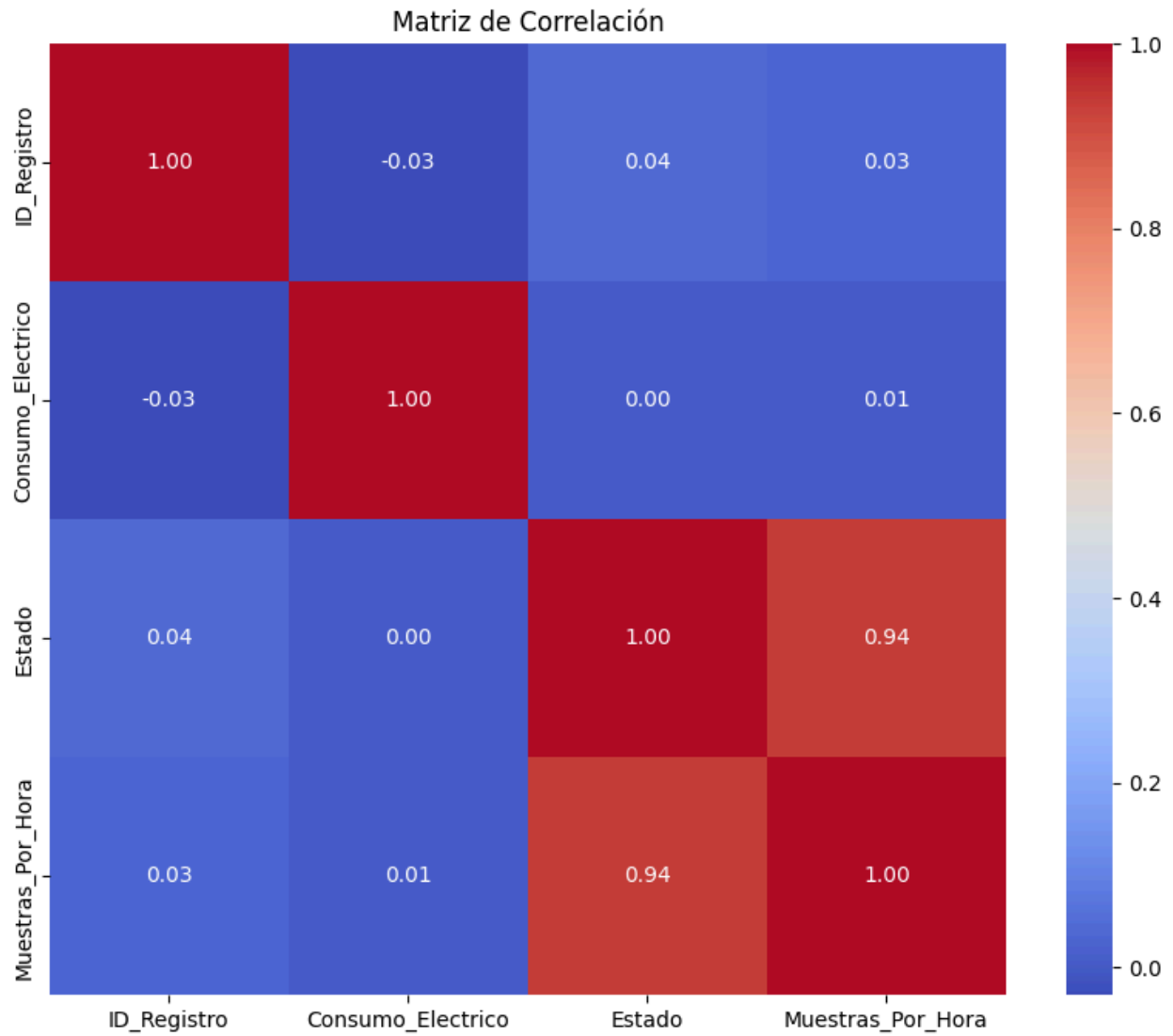
Graph 1

The histogram shows us that the data is mainly distributed between -50 and 150, indicating that the majority of data has electrical consumption similar to those units.



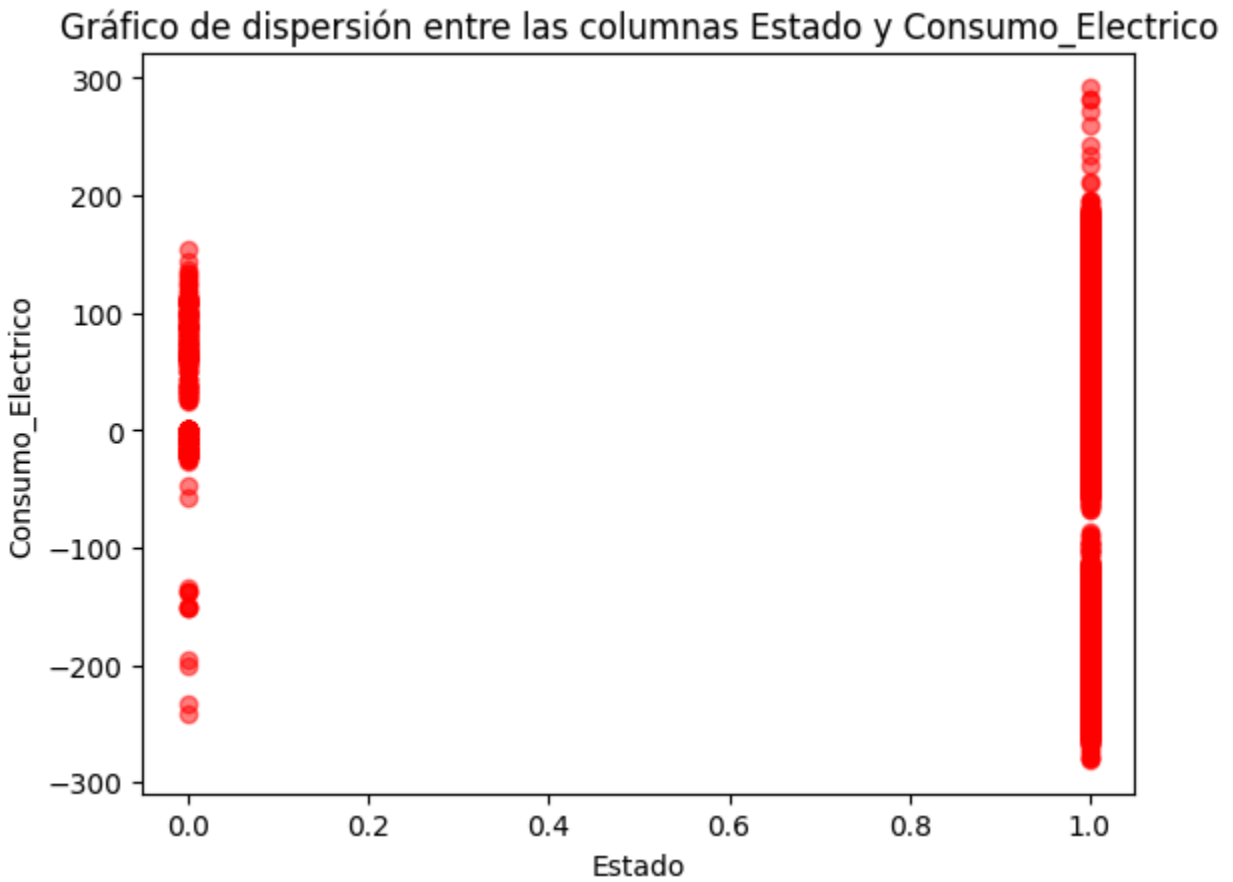
Graph 2

This box plot makes it clear that the consumption data is skewed towards positive values close to 100.



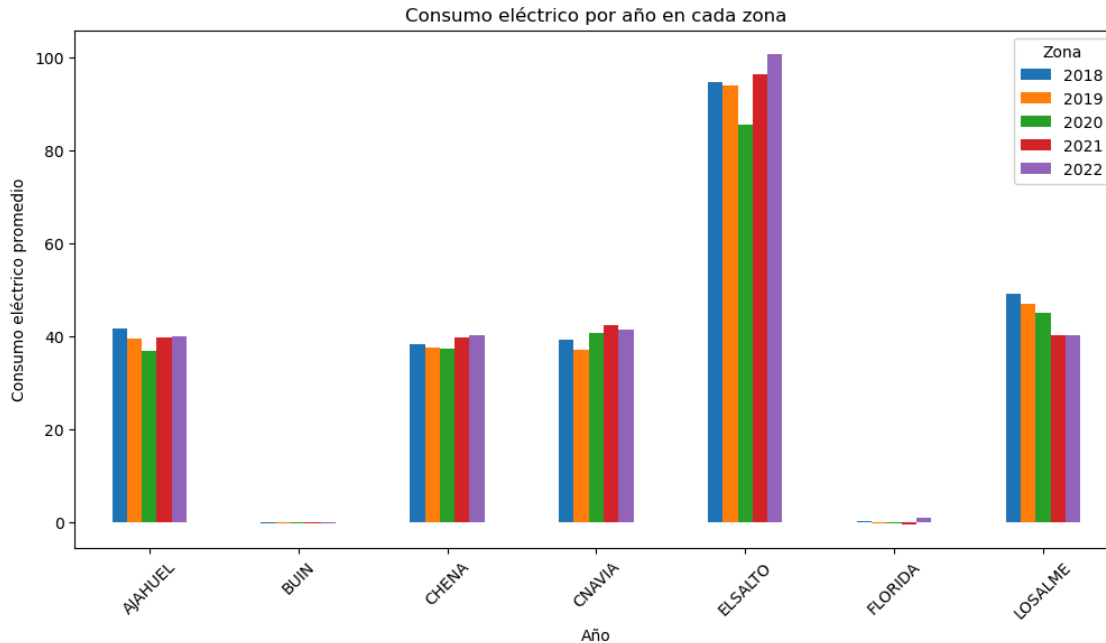
Graph 3

Among the numeric variables considered for the correlation matrix for Graph 3, only "state" and "samples per hour" show a strong correlation with each other.



Graph 4

In this scatter plot, we can see a strong correlation between electrical consumption and state, which is evident. However, it is important to mention that there are more instances in the enabled state than in the disabled state. It is also important to note that even the instances in the disabled state have consumption.



Graph 5

Graph 5 shows a comparison of consumption between substations. It can be observed that Florida and Buin have lower consumption over the years, whereas El Salto, on the contrary, has the highest electrical consumption over the years.

Feature Selection Proposal

Following an initial analysis, we have concluded that the samples per hour are simply a technical feature of the measurement process, indicating how many samples are taken on an hourly average. That is to say, for the electrical consumption prediction model, what really matters are the factors directly affecting consumption, such as the time of day, temperature, holidays, among others. These are the elements that provide significant information for accurately predicting electrical consumption. In addition, we will consider the ID records as a less important feature when predicting electrical consumption, so it will be one of the characteristics that may not be used in the final model, but initially, we will test it and compare results.

Dataset Generation Procedure

To generate different datasets, we will consider different years and substations. In this case, we will use a dataset with records from 2018 to 2021 to train the model and data from 2022 for testing (test). Additionally, we will create different datasets with the same year-based separation and, furthermore, by substation.

Finally, for a prediction of one week, we will use a dataset from the first week of 2023 to attempt the prediction. The years 2017 and 2023 will not be considered in the other datasets due to lack of data in these years.

Model Evaluation Metrics

To evaluate the performance of our model, we utilized several key metrics:

1. Mean Absolute Error (MAE): This metric provides the average absolute difference between the predicted values and the actual values. It gives us a measure of the average magnitude of errors in the predictions without considering their direction.
2. Mean Squared Error (MSE): MSE calculates the average of the squares of the errors between predicted and actual values. It amplifies larger errors due to the squaring operation, making it particularly useful for penalizing larger deviations from the actual values.
3. Mean Absolute Percentage Error (MAPE): MAPE measures the average percentage difference between predicted and actual values relative to the actual values. It provides insights into the accuracy of the model's predictions in terms of percentage errors.
4. R-squared (R^2) Score: R-squared represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with 1 indicating a perfect fit.

Among these metrics, Mean Squared Error (MSE) is a particularly useful measure for our evaluation due to its sensitivity to larger errors. This sensitivity is essential for our application because deviations from the actual values could result in significant consequences, such as increased operational expenses or infrastructure damage. By penalizing larger errors more heavily, MSE provides us with a more accurate assessment of our model's performance in handling such deviations.

Machine Learning Technique

This neural model is a Multilayer Perceptron (MLP) designed for regression tasks. It predicts electrical consumption based on historical data from 2018 to 2021. The model comprises dense layers with various activation functions. It's trained using the Adam optimizer and Mean Squared Error (MSE) loss function. Input data is normalized using Min-Max scaling. The aim is to accurately predict future consumption trends.

Description of the Neuronal Model

This neural model is a Multilayer Perceptron (MLP) designed for regression tasks, where it predicts a continuous value (in this case, electrical consumption) based on input features. It comprises several fully connected dense layers with different activation functions.

Proposed Inputs:

- Electrical consumption history: This represents the past measurements of electrical consumption, from 2018 to 2021, used to train the model. These historical data points serve as input features to the model, allowing it to learn temporal patterns and trends in consumption behavior.

Proposed Output:

- Electrical consumption prediction: The model predicts the future electrical consumption based on the provided historical data and the type of day. This prediction serves as an estimate for energy consumption in subsequent time periods, aiding in energy management and planning for efficient energy distribution.

Description of the features comprising the model input

The model utilizes electrical consumption data that includes several features such as substation, year, month, day, hour, and consumption. These data are structured in a tabular format, where each row represents an observation of electrical consumption at a specific substation, year, month, day, and hour, along with its corresponding consumption value.

Description of the parameters used to train the model

Optimizer: Adam optimizer has been chosen, known for its efficiency in optimizing neural network models. Adam automatically adjusts learning rates during training, which can lead to faster and more stable convergence, in this case we use a 0,001 as initial learning rate for the model.

Loss Function: The selected loss function is Mean Squared Error (MSE). This function is suitable for regression problems like electrical consumption forecasting, where the goal is to minimize the squared difference between the model predictions and the actual consumption values.

Number of Epochs and Batch Size: 30 epochs and a batch size of 150 have been chosen. These values may vary depending on the dataset's complexity and available computational capacity. The number of epochs controls how many times the training set passes through the neural network during training, while the batch size determines how many samples are used to compute the gradient in each weight update step of the network. These values have been selected as a starting point and can be adjusted during the model optimization process.

Description of the architecture of the model used

The model used is a feedforward neural network (also known as a densely connected neural network or multilayer perceptron). It consists of several densely connected layers, followed by a linear output layer.

Hidden Layers: The model has three hidden layers with 256, 128, and 64 neurons, respectively. These layers use sigmoid and hyperbolic tangent activation functions, which are commonly used in regression problems, to introduce nonlinearity into the network and allow for the approximation of more complex functions.

Output Layer: The output layer consists of a single neuron with a linear activation function. This configuration is suitable for regression problems where the goal is to predict a continuous numerical value, such as electrical consumption.

The model has been trained and validated using electrical consumption data from 2018 to 2021 as the training set and data from 2022 as the test set. Standard metrics such as mean squared error (MSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and the coefficient of determination (R^2) have been used to evaluate the model's performance on the test data. Experimental results show a good ability of the model to predict electrical consumption, supporting the selection of the architecture and model parameters.

The results obtained

In the main model with the architecture mentioned above, the following results were obtained, as shown in image 7 of key metrics:

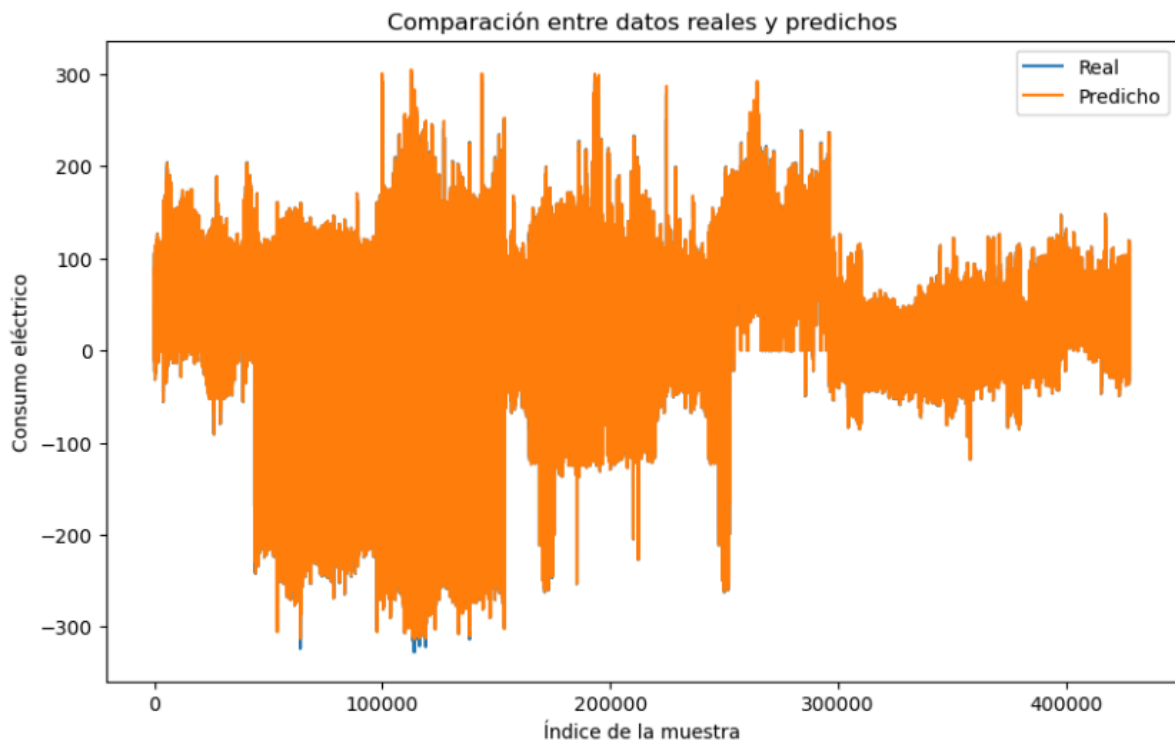
```
13369/13369 ————— 11s 848us/step
Mean Absolute Error (MAE): 0.043696733382828645
Mean Squared Error (MSE): 0.004133557154372792
Mean Absolute Percentage Error (MAPE): 0.40376296178001575
R^2 Score: 0.9999992988357339
```

Image 7: Metrics Result

1. Mean Absolute Error (MAE): The MAE measures the average of the absolute differences between the model predictions and the actual values. In this case, the MAE is approximately 0.0437, indicating that, on average, the model predictions have an absolute error of around 0.0437 units in relation to the actual electrical consumption values.
2. Mean Squared Error (MSE): The MSE calculates the average of the squares of the differences between the model predictions and the actual values. Here, the MSE is around 0.00413, suggesting that the model has good accuracy in predicting electrical consumption, as the mean squared error is quite low.
3. Mean Absolute Percentage Error (MAPE): The MAPE measures the average absolute percentage of the differences between the model predictions and the actual values. With a value of around 0.4038, it indicates that, on average, the model has an error of around 0.4038% in the predictions of electrical consumption relative to the actual values.
4. R² Score: The coefficient of determination (R²) is a measure of the proportion of the variance in the dependent variable that is predictable from the independent variables. An R² score close to 1 indicates a good fit of the model to the data. Here, the R² score is

approximately 0.999999, suggesting that the model explains almost all the variability in the electrical consumption data.

Based on these metrics, we can conclude that we are dealing with a good model for the datasets used. This is evident in graph 6.



Graph 6: Comparison between actual and predicted data

Despite the results in the main model, it had to be adapted for each substation by varying the epochs and batch sizes of each model due to the differences existing between datasets.

AJAHUEL Model(Epochs: 50 / Batch Size: 150):

1372/1372 ————— 1s 874us/step
Mean Absolute Error (MAE): 0.0991849356517602
Mean Squared Error (MSE): 0.02356147419310529
Mean Absolute Percentage Error (MAPE): 2.723898247159244
R^2 Score: 0.9999931680374908

Image 8: Metrics Result

Overall, the model performs similarly to that with the dataset containing all substations. However, we can see that the Mean Absolute Percentage Error is higher, indicating greater error. In fact, this is something that repeats in the other models of the substations as well.

BUIN Model(Epochs: 30 / Batch Size: 80):

1714/1714 ————— 2s 843us/step
Mean Absolute Error (MAE): 0.26280887472348335
Mean Squared Error (MSE): 0.09770194168356357
Mean Absolute Percentage Error (MAPE): 7.6442191530426795
R^2 Score: 0.9999897370731189

Image 9: Metrics Result

CHENA Model(Epochs: 50 / Batch Size: 150):

1714/1714 ————— 1s 835us/step
Mean Absolute Error (MAE): 0.29224891963657
Mean Squared Error (MSE): 0.0984543693235355
Mean Absolute Percentage Error (MAPE): 0.43344491810502866
R^2 Score: 0.9999938916730349

Image 10: Metrics Result

CNAVIA Model(Epochs: 20 / Batch Size: 20):


 3086/3086 ————— 3s 834us/step
Mean Absolute Error (MAE): 0.23309138686634756
Mean Squared Error (MSE): 2.6507527167127614
Mean Absolute Percentage Error (MAPE): 0.39941989549816226
R^2 Score: 0.999335203496377

Image 11: Metrics Result

ELSALTO Model(Epochs: 25 / Batch Size: 180):


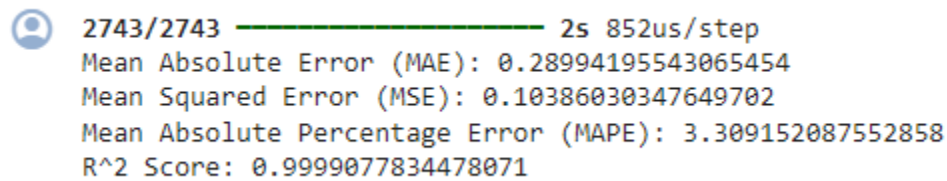
 1372/1372 ————— 1s 928us/step
Mean Absolute Error (MAE): 0.3690714323022265
Mean Squared Error (MSE): 0.16828011927797948
Mean Absolute Percentage Error (MAPE): 0.9676922887190755
R^2 Score: 0.9998945276032026

Image 12: Metrics Result

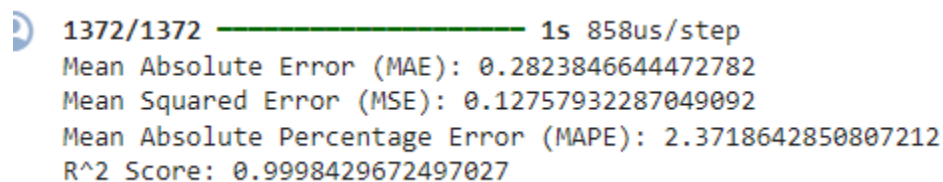
FLORIDA Model(Epochs: 15 / Batch Size: 150):



```
2743/2743 ————— 2s 852us/step
Mean Absolute Error (MAE): 0.28994195543065454
Mean Squared Error (MSE): 0.10386030347649702
Mean Absolute Percentage Error (MAPE): 3.309152087552858
R^2 Score: 0.9999077834478071
```

Image 13: Metrics Result

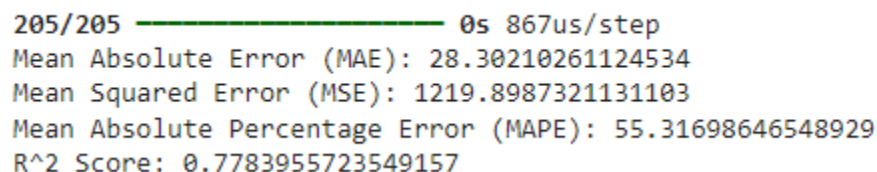
LOSALME Model(Epochs: 20 / Batch Size: 150):



```
1372/1372 ————— 1s 858us/step
Mean Absolute Error (MAE): 0.2823846644472782
Mean Squared Error (MSE): 0.12757932287049092
Mean Absolute Percentage Error (MAPE): 2.3718642850807212
R^2 Score: 0.9998429672497027
```

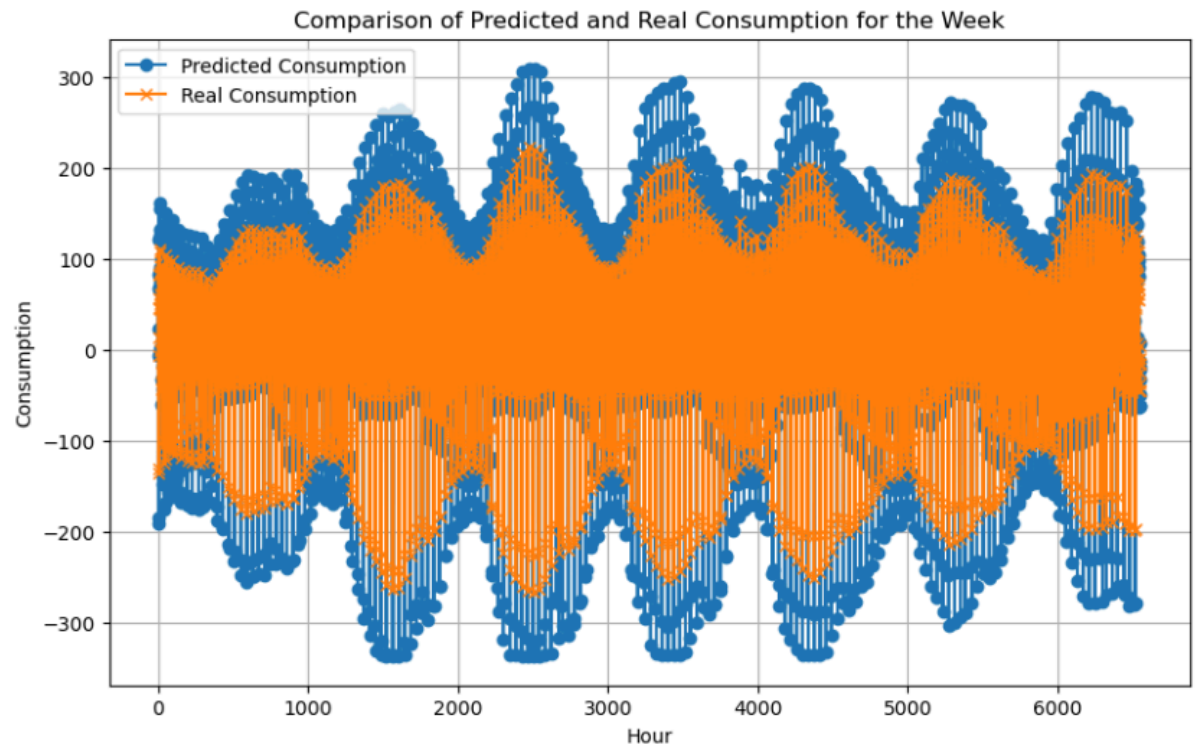
Image 14: Metrics Result

Finally, we attempted to predict one week with our model saved in Keras, using the data from the first week of 2023 available. For this purpose, a separate dataset was created with just one week of data. The results were considerably disastrous according to our metrics, as can be seen in image 15.



```
205/205 ————— 0s 867us/step
Mean Absolute Error (MAE): 28.30210261124534
Mean Squared Error (MSE): 1219.8987321131103
Mean Absolute Percentage Error (MAPE): 55.31698646548929
R^2 Score: 0.7783955723549157
```

Image 15: Metrics Result



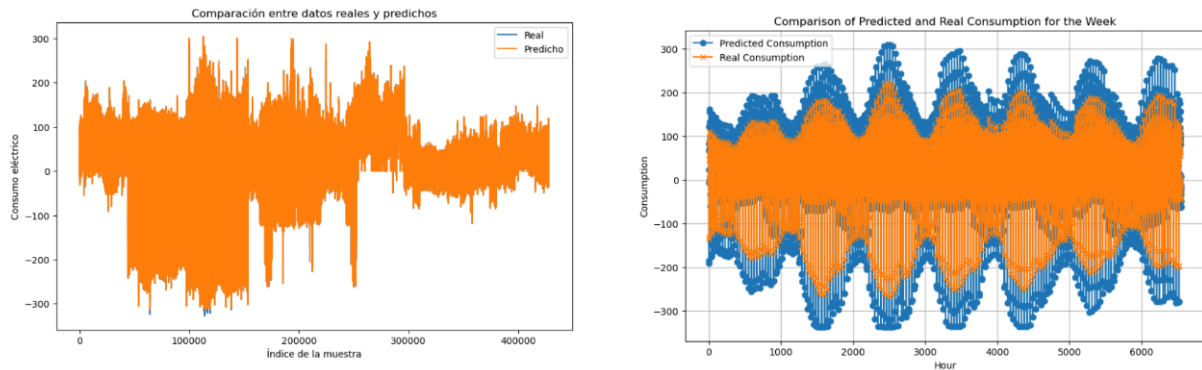
Graph 7: Comparison between actual and predicted data

As can be seen, the result of the model in predicting the first week of available data in 2023 is not satisfactory, as indicated by key metrics. For example, the mean absolute percentage error (MAPE) is approximately 55%, which is far from optimal. However, this is due to the limited adaptability of the model, leading us to believe that a Long Short-Term Memory (LSTM) network could be better suited for this type of project, given its properties as a model, unlike a Multilayer Perceptron, which has deficiencies in this area.

Conclusions

Certainly, we can say that we are facing a model that cannot adapt to significantly different datasets, or, in this case, samples of varying sizes, such as those from the first week of 2023. In fact, the latter confirms this, as there were suspicions when manually adapting it to the models of the datasets from each substation. However, the model that follows the formula $T_{2018} + T_{2019} + T_{2020} + T_{2021} = T_{2022}$, referring to using previous years to predict the year 2022, is not far from being a good result (Graph 6). Unfortunately, when applying it to predict the first week of available data in 2023, it falls short.

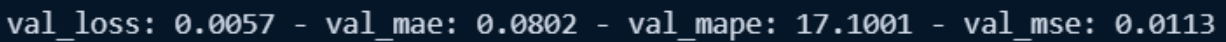
This is due to the model's limited adaptability to the differences in the data, such as variations in the quantity of data available. The models used for training had significantly more data than the data of a single week of records in 2023.



Graph 8: Comparison between actual and predicted data (Graph 6 and 7)

Recommendations

With that said, we've observed several things throughout the development of these models. As a recommendation, we propose using a different architecture for the model that has better adaptability, such as a Long Short-Term Memory (LSTM) network. We estimate that it could be much more accurate in capturing different data trends. During testing with an LSTM model, we achieved very low Mean Absolute Error (MAE) and Mean Squared Error (MSE) values, around 0.01 and 0.02 for both metrics. However, the main reason for choosing an MLP over this is the Mean Absolute Percentage Error (MAPE) metric. In this regard, the MLP achieves values ranging from 0.4 in the best cases to 7.6 in the worst cases, whereas the LSTM model obtains MAPE values of 17.0 in the best cases. This leads us to think that it is making a mistake in the sign of the consumption, meaning if the network is feeding another substation, it doesn't require power.



```
val_loss: 0.0057 - val_mae: 0.0802 - val_mape: 17.1001 - val_mse: 0.0113
```

Image 16: LSTM Metrics Result