COBEM 2021 Florianópolis - Brasil
ABCM Associação Brasileira de Engenharia e Ciências Mecânicas
26th ABCM International Congress of Mechanical Engineering
November 22-26, 2021. Florianópolis, SC, Brazil

COB-2021-0475
AUTOMATED MACHINE LEARNING APPROACH APPLIED TO NUCLEAR ENERGY GENERATION SHORT-TERM FORECASTING

Jorge Gustavo Sandoval Simão
Industrial and Systems Engineering Graduate Program (PPGEPS), Pontifical Catholic University of Parana (PUCPR). Curitiba, Parana, Brazil.
jorge.simao@pucpr.edu.br
Viviana Cocco Mariani
Mechanical Engineering Graduate Program (PPGEM), Pontifical Catholic University of Parana (PUCPR). Curitiba, Parana, Brazil.
viviana.mariani@pucpr.br
Leandro dos Santos Coelho
Department of Electrical Engineering, Federal University of Parana, Industrial and Systems Engineering Graduate Program (PPGEPS), Pontifical Catholic University of Parana, Curitiba, PR, Brazil.
leandro.coelho@pucpr.br

Abstract. *Machine learning models applied to the time series forecasting problem are gaining interest among researchers and the industry and energy systems. This work evaluates automated machine learning, and feature selection (RreliefF) applied to time series forecasting for the United States nuclear power plants case study in a clustered dataset. This research evaluated forecasting results in terms of different performance metrics, where the AutoML (Automated Machine Learning) approach pipelined into the RreliefF algorithm presents promising results based on training and test datasets. To achieve this goal, nuclear power generation patterns were divided into two groups using clustering. Subsequently, the RReliefF algorithm is applied in each cluster, aiming to find the ideal number of features. To increase the precision of the model, a normalization was performed on the energy generation data of each cluster individually, and their skewness and kurtosis were measured and compared. The AutoML models that generated the best results using RreliefF are analyzed, and the metrics of R^2 (Coefficient of Determination), MAE (Mean Absolute Error) and RMSLE (Root Mean Squared Logarithmic Error) are obtained. Despite a small difference, the number of features of each cluster shows that there is a difference in the generation of energy patterns, and that it is possible to generate accurate models of nuclear energy generation through the analysis of generation time series.*

Keywords: *Automated Machine Learning, Feature Selection, Nuclear Energy, Time Series, Forecasting*

1. INTRODUCTION

With the continuous development of society, the energy demand has also increasingly increased with the increasing economic development. It has led to an increasingly severe crisis of the exhaustion of the limited and nonrenewable fossil energy reserves. At the same time, the extensive use of fossil energy in the past hundred years has led to global warming, haze, and other environmental problems that have become increasingly prominent (Mi and Zhao, 2020). There has been a global effort to actively develop clean, efficient, and sustainable renewable energy sources to support future economic development to meet these challenges. The widely used new energy sources are solar energy, wind energy, tidal energy, nuclear energy, and geothermal energy (Wang *et al.*, 2016). Nuclear energy has been widely used worldwide because of its wide distribution, high energy frequency, and constant generation.

Electricity generation from commercial nuclear power plants in the United States of America began in 1958. At the end of December 2020, the United States had 94 operating commercial nuclear reactors at 56 nuclear power plants in 28 states. The average age of these nuclear reactors is about 39 years old. The oldest operating reactor, Nine Mile Point Unit 1 in New York, began commercial operation in December 1969. The newest reactor to enter service, Watts Bar Unit 2, came online in 2016—the first reactor to come online since 1996 when the Watts Bar Unit 1 came online. According to the U.S. Nuclear Regulatory Commission, as of November 2019, 23 shut down commercial nuclear power reactors at 19 sites in various stages of decommissioning. U.S. nuclear electricity generation capacity peaked in 2012 at about 102,000 MW when there were 104 operating nuclear reactors. At the end of 2020, 94 operating reactors with a combined generation capacity of about 96,555 MW. From 2014 through 2018, annual nuclear generation capacity and electricity generation increased each year even as the number of operating reactors declined. Power plant uprates—modifications to increase capacity—at nuclear power plants have made it possible for the entire operating nuclear reactor fleet to maintain

a relatively consistent total electricity generation capacity. These updates, combined with high-capacity utilization rates (or capacity factors), helped nuclear power plants maintain a consistent share of about 20% of total annual U.S. (United States) electricity generation from 1990 through 2019. Some reactors also increased annual electricity generation by shortening the time reactors are offline for refueling.

Researchers have developed methods for predicting power generation for NPPs (Nuclear PowerPlants) to assist in decision making. Papers as Moshkbar-Bakhshayesh (2020), Aizpurua *et al.* (2019) and Radaideh *et al.* (2020) help to see the need for methods to predict different types of events in critical structures like NPP's. However, traditional machine learning methods cannot extract feature information of time series from data in detail because they lack deep extraction capabilities (Duan *et al.*, 2021). According to their energy generation pattern, the 96 plants used in this study were divided into two clusters to achieve this goal. RreliefF was used in each of them to make the Feature Selection, aiming to extract the most significant variables for each cluster model. After that, there was an application of AutoML using the Coefficient of Determination as a measure to generate an efficient energy model forecasting.

The remainder of this paper is organized as follows. Section 2 presents the related works, Section 3 the material and methods used, and the data set. Section 4 discusses the results obtained, and Section 5 presents conclusions, remarks, and future research directions.

2. RELATED WORKS

For related works, we selected those that dealt with the prediction of nuclear energy (preferably) from 2016 onwards, except for Menyah and Wolde-Rufael (2010). In Radaideh *et al.* (2020) the researcher uses deep learning expert systems to model and predict the time series progression of a design-basis nuclear accident, featuring a loss of coolant accident. This work accomplishes two significant findings: first, train expert systems with high accuracy, which could help nuclear power plant operators figure out plant responses during the accident. Second, building fast, efficient, and accurate deep models to simulate nuclear phenomena, which could be valuable to computational nuclear science. In the work of Moshkbar-Bakhshayesh (2020) cross-correlation of measurable/unmeasurable parameters of nuclear power plants (NPPs) are detected. Correlation techniques including Pearson's, Spearman's, and Kendall-tau give appropriate input parameters for training/prediction of the target unmeasurable parameters. The case study target parameters used are fuel and clad maximum temperatures of uncontrolled withdrawal of control rods (UWCR) transient of Bushehr nuclear power plant (BNPP). Different model-free methods, including decision tree (DT), feed-forward back propagation neural network (FF-BPNN), accompany with different learning algorithms (i.e., gradient descent with momentum (GDM), scaled conjugate gradient (SCG), Levenberg-Marquardt (LM), and Bayesian regularization (BR)), and support vector machine (SVM) with different kernel functions (i.e., linear and Gaussian functions) are employed to predict the target parameters. In both works, despite the excellent performance of the metrics of the machine learning models, feature selection and automated machine learning is not used. Menyah and Wolde-Rufael (2010) consider a different approach: His study explores the causal relationship between carbon dioxide (CO^2) emissions, renewable and nuclear energy consumption, and real GDP for the U.S. for the period 1960-2007. Using a modified version of the Granger causality test, he founds a unidirectional causality running from nuclear energy consumption to CO^2 emissions without feedback but no causality running from renewable energy to CO^2 emissions, showing the use of nuclear energy time series for correlations. Wolde-Rufael (2010) uses a similar approach, attempting to examine the dynamic relationship between economic growth, nuclear energy consumption, labor, and capital for India for the period 1969-2006. Applying the bounds test approach to cointegration, he finds a short and long-run relationship between nuclear energy consumption and economic growth. Using four long-run estimators also found that nuclear energy consumption has a positive and a statistically significant impact on India's economic growth.

It Dian-Gang *et al.* (2018), establishes the simulation model of the transaction plan that meets the energy-saving power generation dispatching by using the time series simulation method, considering the factors such as the output characteristics, load characteristics, and peak shaving characteristics of various energy sources. In Jawad *et al.* (2020) 's research uses different supervised learning algorithms, including multiple linear regression, support vector regressors with different kernels, *k*-nearest neighbors, Random Forest and AdaBoost to forecast the time series data. However, the performance of these algorithms is data-dependent on the correlated meteorological parameters of the specific region. Similarly, in Aizpurua *et al.* (2019)] presented a novel transformer condition approach integrating uncertainty modeling, data-driven forecasting models, and model-based experimental models to increase prediction accuracy and handle uncertainty. These are done not for energy generation but to quantify measurement errors on transformer predictions and confirm that temperature and load measurement errors affect the estimation.

Developing an accurate and reliable multi-step ahead prediction model is a critical problem in many Prognostics and Health Management (PHM) applications. Inevitably, the further one attempts to predict the future, the harder it is to achieve an accurate and stable prediction due to increasing uncertainty and error accumulation. In Nguyen *et al.* (2020), this problem is addressing by proposing a prediction model based on Long Short-Term Memory (LSTM), a deep neural network developed for dealing with the long-term dependencies in time-series data. His proposed prediction model also tackles two additional issues: Firstly, a Bayesian optimization algorithm (called Tree-structured Parzen Estimator)

automatically tunes the hyperparameters (like this research, using AutoML - Automated Machine Learning). Secondly, the proposed model allows assessing the uncertainty on the prediction. A case study considering steam generator data acquired from different french nuclear power plants was carried out to validate the performance of the proposed model.

The last paper of this related works section proposes a novel hybrid ensemble learning paradigm integrating ensemble empirical mode decomposition (EEMD) and least squares support vector regression (LSSVR) for nuclear energy consumption forecasting. It is based on the principle of "decomposition and ensemble" This hybrid ensemble learning paradigm is explicitly formulated to address difficulties in modeling nuclear energy consumption, which has inherently high volatility, complexity, and irregularity (Tang *et al.*, 2012). Within the scope of the studies analyzed, despite the constant use of the nuclear energy datasets for forecasting, techniques such as AutoML and RreliefF are hardly used. Their combination can generate highly efficient models for predicting NPP's energy generation.

3. MATERIAL AND METHODS

3.1 Dataset

The raw dataset used for this research was obtained from the U.S. Energy Information Administration (2021) page, and the version used in this study is available on Kaggle [1], for future research. The whole dataset contains the power generation from nuclear power plants in the United States of America, separated by federation between January 2013 and November 2020. Its robust data comprises nine files containing monthly energy generation, separated by nuclear power plants in each US state. They also have information about the names of the states, the powerplants, their identifier, and utilization factor. Generation is measured in MHw (monthly) between 2013 to 2020.

The original dataset is composed of the following fields: "State" (the US state where the powerplant is located), "Plant ID" (the powerplant identification number), "Unit ID" (the powerplant unit identification, for each state), months from January to December describing the amount of energy generation in TW/h (Terawatts-hour) and the "Utilization Factor" field, that describes how effectively (how well are utilized) are thermal neutrons absorbed in the fuel.

For his forecasting study, data from one year of energy production (2020) were used, with this year subdivided into the total monthly amount, with a very short-term forecast (1 month), and the data divided into training (70%), testing (15%), and validation (15%).

3.2 Feature Engineering and Preprocessing

For the development of this work, not all fields in the data set were used. As the objective was to predict the generation of energy, it was necessary to apply feature engineering, aiming to separate only the most essential fields (at first) and increase the accuracy rate of the model.

It began with the creation of a sequential numeric field to identify the records ("Powerplant ID") since the fields "Plant ID" and "Unit ID" had repetitions between them, being unique only in the composition of both. After this, the fields "State" and "Unit ID" were removed, and "Plant ID" and "Utilization Factor" were separated for use in the preprocessing of the dataset. So for the forecast data set, there were only the months and the power generation in TW/h. Table 1presents the descriptive statistics of each month used (after normalization between 0 and 1 range).

Descriptive Statistics	0	1	2	3	4	5	6	7	8	9	10	11
Count	89	89	89	89	89	89	89	89	89	89	89	89
Mean	0.57	0.73	0.66	0.63	0.68	0.74	0.74	0.65	0.67	0.65	0.69	0.55
Standard Deviation	0.24	0.18	0.25	0.26	0.23	0.17	0.17	0.23	0.22	0.27	0.22	0.23
Minimum	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
First Quartile	0.40	0.64	0.57	0.48	0.62	0.66	0.65	0.51	0.55	0.45	0.63	0.41
Median	0.63	0.74	0.69	0.68	0.69	0.76	0.76	0.67	0.69	0.70	0.72	0.56
Third Quartile	0.76	0.87	0.87	0.86	0.86	0.87	0.88	0.84	0.84	0.87	0.87	0.73
Maximum	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 1. Normalized dataset statistics

In Table 1, the number 0 represents January, number 1 February, and so on until 11 (December). Once the data were already normalized and the features adjusted, for the preprocessing phase, the following techniques were applied: the white noise test to verify that the data set was suitable for pattern prediction, skewness to measure the symmetry of the distribution frequency, kurtosis to measure the shape of the flattening of the curve of the probability distribution function and the Yeh-Johnson power transformation to stabilize variance, make the data more normal distribution-like, improve the validity of measures of association such as the Pearson correlation between variables and for other data stabilization

[1] https://www.kaggle.com/jorgesandoval/us-nuclear-powerplant-energy-generation

procedures.

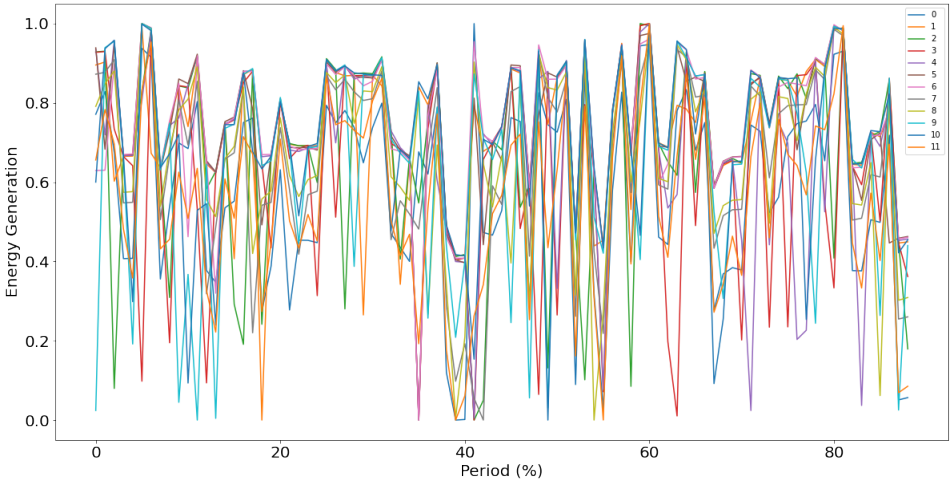

Figure 1. One year of montly normalized energy generation, from January to December to each reactor

In signal processing, white noise is a random signal having equal intensity at different frequencies, giving it a constant power spectral density that (Carter, 2013). In time series, white noise is defined by a zero mean, constant variance, and zero correlation (Box and Jenkins, 2015). Table 1 shows that the mean is not zero and the variance change over time, and Figure 1 shows that the series is random.

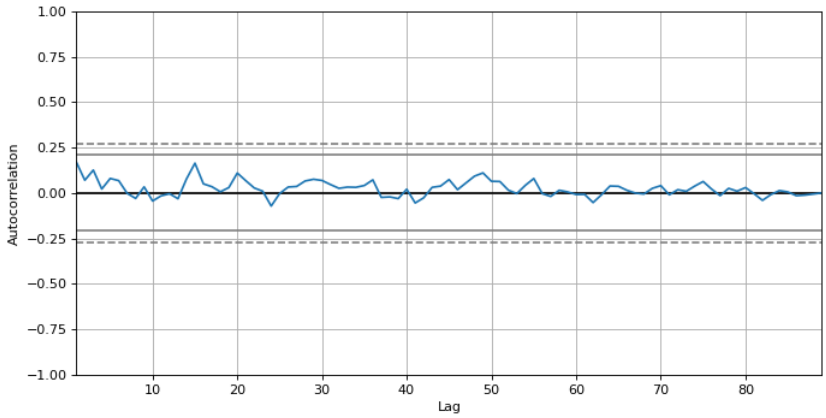

Figure 2. Correlation between lagged variables

To finish the white noise test, the correlogram (details in Figure 2) does not show any obvious autocorrelation pattern, which permanently rules out the possibility of white noise. Predictions are impossible in white noise time series, but it is possible to improve the model in this case.

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive, zero, negative, or undefined. The skewness of a random variable X is the third standardized moment $\tilde{\mu}_3$, and it is defined as (Brown, 2020):

$$\tilde{\mu}_3 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{\left(\mathrm{E}\left[(X-\mu)^2\right]\right)^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}} \tag{1}$$

where μ is the mean, σ is the standard deviation, E is the expectation operator, μ_3 is the third central moment, and k_t are the t-the cumulants. The dataset used in this study presented a skewness of -0.91, meaning that even if it is considered moderately asymmetric (between -0.5 and -1), it is still a very close asymmetry limit (-1), indicating the possibility of improvements through power transformation functions.

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. Data sets with high kurtosis tend to have heavy tails or outliers. Datasets with low kurtosis tend to have light tails or a lack of outliers. A uniform distribution would be the extreme case. The definition of the kurtosis formula for univariate data, as

(Pearson, 1905) presents is:

$$kurtosis = \frac{\sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)^4 / N}{s^4} \tag{2}$$

where $\bar{Y}$ is the mean, s is the standard deviation, and N is the number of data points (Sharma and Bhandari, 2015). A 0.64 kurtosis value was calculated for this research dataset, indicating a low tail (not too many outliers). However, it is possible to improve measurements again with a power transformation function.

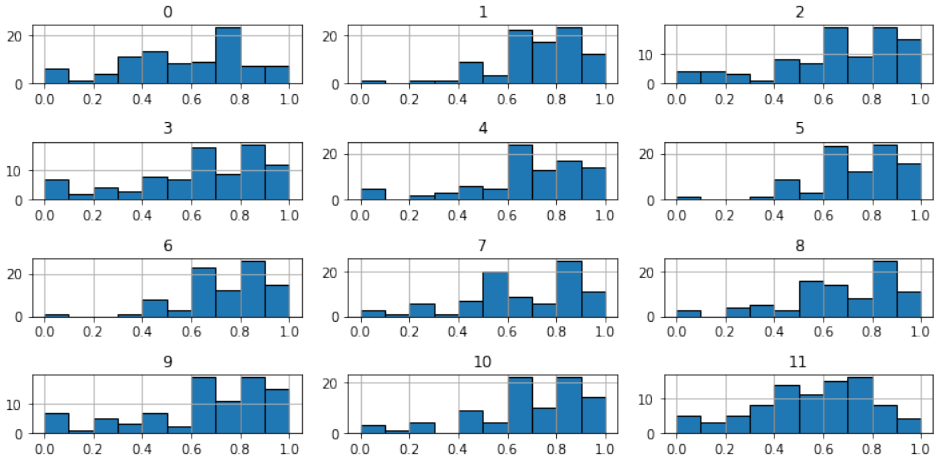

Figure 3. Data distributions from january to december to each reactor

It is possible to observe that except for month 12 (December), the rest of the distribution tends to a right tail, and it is not gaussian in its structure, in the Figure 3. That is why is used a power transformation function after the clustered dataset. It was helpful in stabilize variance, make the data more normal distribution-like, improve the validity of measures of association such as the Pearson correlation between variables, and for other data stabilization procedures.

Yeo and Johnson (2000) have proposed an new family of distributions that can be used without restrictions on that have many of the good properties of the Box-Cox power family. These transformations are defined by:

$$\psi(\lambda, y) = \begin{cases} \left((y+1)^{\lambda} - 1\right)/\lambda & if \lambda \neq 0, y \geq 0 \\ \log(y+1) & if \lambda = 0, y \geq 0 \\ -\left[(-y+1)^{2} - \lambda - 1\right)\right]/(2-\lambda) & if \lambda \neq 2, y < 0 \\ -\log(-y+1) & if \lambda = 2, y < 0 \end{cases} \tag{3}$$

If y is strictly positive, then the Yeo-Johnson transformation is the same as the BoxCox power transformation of $(y+1)$. If y is strictly negative, then the Yeo-Johnson transformation is the Box-Cox power transformation of $(-y+1)$, but with power $2-\lambda$. With both negative and positive values, the transformation is a mixture of these two; then different powers are used for positive and negative values (Weisberg, 2001). In this study, Yeh-Johnson power transformation was the best option because some negative values presented during the dataset preprocessing.

3.3 Clustering

The unsupervised classification process (clustering) is used to classify and place the data with similar characteristics in the same groups (Zahra *et al.*, 2015). K-means clustering is widely used for data sets classification in K cluster. Despite the simplicity and power of discovering the hidden patterns of the data set and the great convergence rate of K-means clustering in achieving the optimal focal point, the use of this method is limited. These limitations are due to introducing the number of clusters (k) before the clustering starts and the random selection of cluster centers (Afshoon *et al.*, 2021). The steps of data classification using the K-means method are as follows:

1. The user selects the number of clusters;
2. *k* points which are the centers of the clusters are selected randomly;
3. Comparing the distance of the available data with each of the cluster centers, the data are placed in the closest cluster;
4. The cluster centers are updated according to the mean of the cluster centers;

5. Based on the new center, the data will be classified again. In this case, one point may be placed in another new cluster;

6. Steps 3, 4, and 5 will be repeated until the centers are fixed, and the clusters converge.

For this study, the solution used a clustering algorithm to separate nuclear reactors based on its utilization factor and powerplant identification, with the help of silhouette and elbow functions.

3.4 Feature Selection

ReliefF and its counterpart for dealing with regression data, i.e., RReliefF (Kononenko *et al.*, 2000), are the most implemented Relief Based Algorithms (RBA) (Urbanowicz *et al.*, 2018). RReliefF works with continuous *y*, and similar to ReliefF, RReliefF also penalizes the predictors that give different values to neighbors with the same response values and rewards predictors that give different values to neighbors with different response values. However, RReliefF uses medium weights to compute the final predictor weights.

Given two nearest neighbors, assume the following:

- W_{dy} is the weight of having different values for the response *y*.
- W_{dj} is the weight of having different values for the predictor F_j.
- $W_{dy \wedge dj}$ is the weight of having different response values and different values for the predictor F_j.

RReliefF first sets the weights $W_{dy}, W_{dj}, W_{dy \wedge dj}$, and W_j equal to 0. Then, the algorithm iteratively selects a random observation x_r, finds the *k*-nearest observations to x_r, and updates, for each nearest neighbor x_q, all the intermediate weights as follows:

$W_{dy}{}^{i} = W_{dy}{}^{i-1} + \Delta_y\left(x_r, x_q\right) \cdot d_{rq}$
$W_{dj}^{i} = W_{dj}^{i-1} + \Delta_j\left(x_r, x_q\right) \cdot d_{rq}$
$W_{dy \wedge dj}^{i} = W_{dy \wedge dj}^{i-1} + \Delta_y\left(x_r, x_q\right) \cdot \Delta_j\left(x_r, x_q\right) \cdot d_{rq}$

The *i* and $i-1$ superscripts denote the iteration step number. *m* is the number of iterations specified by 'updates'. $\Delta_y\left(x_r, x_q\right)$ is the difference in the value of the continuous response *y* between observations x_r and x_q. Let y_r denote the value of the response for observation x_r, and let y_q denote the value of the response for observation x_q. (Robnik-Sikonja and Kononenko, 1997)

$$\Delta_y\left(x_r, x_q\right) = \frac{|y_r - y_q|}{\max(y) - \min(y)} \tag{4}$$

$\Delta_j\left(x_r, x_q\right)$ and d_{rq} functions are the same as for ReliefF, but RReliefF calculates the predictor weights W_j after fully updating all the intermediate weights.

$$W_j = \frac{W_{dy \wedge dj}}{W_{dy}} - \frac{W_{dj} - W_{dy \wedge dj}}{m - W_{dy}} \tag{5}$$

Relief algorithms are general and successful attribute estimators. They can detect conditional dependencies between attributes and provide a unified view on the attribute estimation in regression. (Robnik and Konenko, 2003)

3.5 Forecasting Model

In this work, there was the application of AutoML to short-term electric load forecasting tasks. According to Wang *et al.* (2019), from a data science perspective, the biggest challenge is to develop advanced pipelines of algorithms covering a numerical dynamic data application (real-time, optimized decision making). This pipeline would enable the development of a highly advanced level of decision-making, which is impossible to attain for human experts on their own. Thereby, AutoML contributes to the human capital skills in the field of short-term electric load forecasting. This study used the auto-sklearn AutoML system and its workflow presents in Figure. 4

Auto-sklearn uses Bayesian optimization on top of SciKit-learn for generating machine learning pipelines. Auto-sklearn uses SMAC (Sequential Model-Based Algorithm Configuration) (Hutter *et al.*, 2011) as the underlying Bayesian optimization method to optimize the machine learning pipeline preprocessing automatically and postprocessing method selection, model selection, and hyperparameter optimization. SMAC is a general-purpose automatic algorithm configurator and thus not limited to be used in the context of AutoML. It uses random forests as surrogate models combined with an expected improvement criterion for selecting candidate configurations. The update of the surrogate model occurs throughout the sequential optimization process, which tends to be slow initially, but usually shows good performance over time (Feurer *et al.*, 2015).

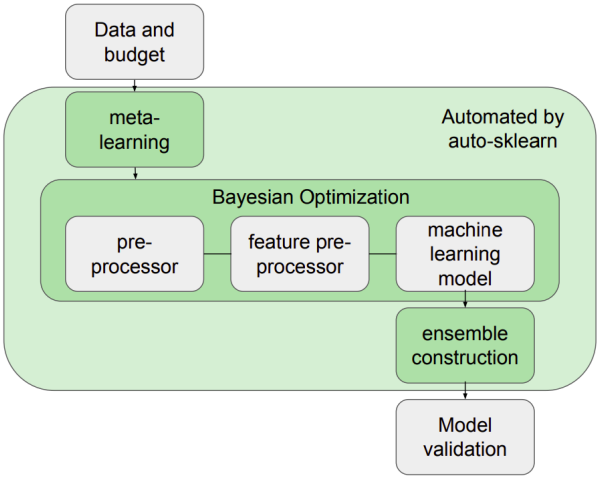

Figure 4. Auto-Sklearn workflow

3.6 Performance Criteria

To measure the model generated by AutoML, three performance criteria: the Coefficient of Determination (R^2), Root Mean Squared Logarithmic Error (RMLSE), and the Mean Absolute Error. The R^2, is used to analyze how differences in one variable can be explained by a difference in a second variable and was the base metric for the AutoML algorithm used in this study, as shown in the equation. 6

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}} \tag{6}$$

In this equation, *r* represents the correlation coefficient, *x* the values in the first set of data, *y* the values in the second set of data ,and *n* the total number of values. The value of R^2 increases after adding a new variable predictor, and it might not be associated with the result or outcome. The R^2 , which was adjusted, will include the same information as the original one, and the number of predictor variables in the model gets penalized. In short terms, the coefficient of determination (R^2) calculates the ratio of the variation of the predicted value described by the actual value (Mwangi *et al.*, 2021).

The second metric, the RMSLE, compares the activity concentrations predicted and the observed activity concentrations. The advantage of RMSLE is that it does not penalize significant differences in the modeled and the observed values when both modeled and observed values are large numbers (Dacre *et al.*, 2020).

$$\log_{10}(RMSLE) = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left[\log_{10}\left(x_t\right) - \log_{10}\left(y_t\right)\right]^2} \tag{7}$$

In the equation 7, x_t is the observed concentration ,and y_t is the simulated concentration at time *t*. This is calculated using *N* concentration values, when both x_t and y_t are non-zero. Subtracting two log-transformed values is equivalent to the log of the ratio of the value.

The last performance criteria measures the average magnitude of absolute differences between *N* predicted vectors $S = x_1, x_2, ..., x_N$ and $S = y_1, y_2, ..., y_N$, the corresponding loss function is defined as shown in equation 8:

$$\mathcal{L}_{MAE}\left(S, S^*\right) = \frac{1}{N}\sum_{i=1}^{N}\left\|x_i - y_i\right\|_1 \tag{8}$$

where $\|\cdot\|_1$ denotes L_1 normalization (Qi *et al.*, 2020). These three performance criteria ensure that the model generated in this study measured both the correlation rate and the errors between the predicted and current values.

4. RESULTS ANALYSIS

This section presents the application of Automated Machine Learning, pipelined into a RreliefF feature selection algorithm for nuclear power plants energy generation. The forecasting data frame consisted of a self-incrementing identifier, the powerplant identifier, and the months from January to December. These were all fields in the original data set. The

creation of the columns "Mean" and "Median" assists in observing the data based on each row's mean and median. The first action taken was removing columns that would not be used (State, Unit ID, and Plant Name). December was a null column in the 2020 data set, so its entries were composed of the average of the records from January to November.

This study used two dataframes from the original data: one for the forecast of values and another for clustering the nuclear powerplants. After these steps, clustering was performed, with the objective of dividing the data into groups that have a more similar pattern of energy generation. The dataset was tested using two functions, the elbow function and the silhouette function.

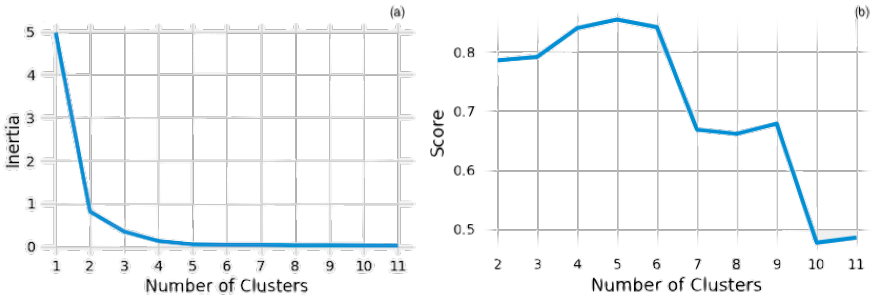

Figure 5. Elbow function (a) and Silhouette function (b) results for eleven clusters.

The figure 5 shows the results of the two functions applied to the dataset. Both functions have an ideal number of clusters out of 5. This number was tested, but it was found that the AutoML model maintains accuracy even with the division into 2 clusters, which shows that the sudden drop in inertia is the factor that really influences the accuracy of the model.

After dividing the dataset into groups, feature selection and AutoML were applied to each of them, being identified as C_0 (Cluster 0) and C_1 (Cluster 1) for the purposes of this study. Once they were clustered, the data were again submitted to statistical analysis; and the C_0 had values of 1.65 Skewness and 2.07 Kurtosis. The C_1 had Skewness of 1.8 Kurtosis of 2.44, and in both cases, a normalization was necessary, aiming to increase the precision of the AutoML model to be generated later. After normalization, the C_0 cluster had Skewness and Kurtosis at -0.57 and 0.01; and cluster C_1 -0.53 and -0.01, respectively.

Feature Selection tests were performed after normalization, and the objective was to find, for each of the clusters, the number of features needed to generate the model as accurately as possible. to achieve this goal, the AutoML was wrapped in the RReliefF technique, to find out which was the most suitable combination for each cluster. The Table 2 presents this results, the coefficient of determination (R^2) for each feature in each cluster, to a maximum total of 11.

	C_0			C_1		
	Train	Test	Valid	Train	Test	Valid
1	0.47	0.32	0.32	0.71	0.52	0.52
2	0.92	0.71	0.71	0.94	0.39	0.39
3	0.94	0.83	0.83	0.94	0.65	0.65
4	0.93	0.76	0.76	0.98	0.81	0.81
5	0.97	0.53	0.53	0.98	0.77	0.77
6	0.97	0.86	0.86	0.98	0.96	0.96
7	0.96	0.82	0.82	0.99	0.95	0.95
8	0.98	0.98	0.98	0.99	0.97	0.97
9	0.97	0.88	0.88	0.99	0.99	0.99
10	0.97	0.93	0.93	0.99	0.99	0.99
11	0.99	0.99	0.99	0.99	0.98	0.98

Table 2. R^2 train, test and valid for each feature in both clusters.

A behavior that can be observed is that in cluster C_0 (which has less defined patterns) the best results are achieved only using 11 features, with an improvement that starts with 6 features and grows alternately, maintaining a high difference between the number of initial and final features (13%). In cluster C_1 the values also start to improve after 6 features, but the growth is much more continuous and shorter, with a difference of only 3% between the best (9) and the worst (11) number of features.

Once the best number of features for each of the clusters was found, the model with the best results for each of them was run and its metrics were analyzed, as shown in Table 3. For cluster C_0, 11 features were used, while 9 were used for cluster C_1. As shown in Table 3, the division of energy generation patterns into clusters allowed a closer fit of the model,

	C_0 (11 Features)			C_1 (9 Features)		
	Train	Test	Valid	Train	Test	Valid
R2	0.99	0.99	0.99	0.99	0.99	0.99
MAE	0.005	0.02	0.02	0.007	0.01	0.01
RMSLE	0.02	0.03	0.03	0.009	0.02	0.02

Table 3. R^2, MAE and RMSLE train, test and valid for 11 and 9 features in both clusters.

allowing for greater precision.

5. CONCLUSIONS

The generation of nuclear energy follows patterns that vary from one plant to another, depending on its activity. Some generation patterns are more constant, others more erratic. However, with clustering methods it is possible to define these patterns, and with techniques such as Feature Selection and AutoML the power generation pattern can be identified and predicted.

AutoML proved to be able to predict trends in the two clusters used in this study, as well as Feature Selection methods were able to successfully identify the most important variables to be used. Thus, it is concluded that these techniques together can efficiently predict the energy generation of nuclear power plants, and that clustering increases the percentage of correctness of the models, by efficiently separating the generation patterns.

6. REFERENCES

Afshoon, I., Miri, M. and Mousavi, S.R., 2021. "Combining Kriging meta models with U-function and K-Means clustering for prediction of fracture energy of concrete". *Journal of Building Engineering*, Vol. 35. ISSN 23527102. doi:10.1016/j.jobe.2020.102050. URL https://doi.org/10.1016/j.jobe.2020.102050.

Aizpurua, J.I., McArthur, S.D., Stewart, B.G., Lambert, B., Cross, J.G. and Catterson, V.M., 2019. "Adaptive Power Transformer Lifetime Predictions Through Machine Learning and Uncertainty Modeling in Nuclear Power Plants". *IEEE Transactions on Industrial Electronics*, Vol. 66, No. 6, pp. 4726–4737. ISSN 02780046. doi: 10.1109/TIE.2018.2860532.

Box, G. and Jenkins, G.M., 2015. *Time Series Analysis: Forecasting and Control*. Holden-Day, New York.

Brown, S., 2020. "Measures of Shape: Skewness and Kurtosis". URL http://brownmath.com/stat/shape.htm.

Carter, B., 2013. *Op Amps for Everyone, 4th Edition*. Newnes, USA, 4th edition. ISBN 0123914957.

Dacre, H.F., Bedwell, P., Hertwig, D., Leadbetter, S.J., Loizou, P. and Webster, H.N., 2020. "Improved representation of particle size and solubility in model simulations of atmospheric dispersion and wet-deposition from Fukushima". *Journal of Environmental Radioactivity*, Vol. 217, No. August 2019. ISSN 18791700. doi:10.1016/j.jenvrad.2020.106193. URL https://doi.org/10.1016/j.jenvrad.2020.106193.

Dian-Gang, H., Jing-Jing, Z., Jing, P. and Yong, Y., 2018. "The study of monthly power generation plan based on energy saving dispatch considering wind, nuclear, water, thermal, and other energy sources". *China International Conference on Electricity Distribution, CICED*, , No. 201804270000022, pp. 1434–1438. ISSN 2161749X. doi: 10.1109/CICED.2018.8592210.

Duan, J., Zuo, H., Bai, Y., Duan, J., Chang, M. and Chen, B., 2021. "Short-term wind speed forecasting using recurrent neural networks with error correction". *Energy*, Vol. 217. ISSN 03605442. doi:10.1016/j.energy.2020.119397. URL https://doi.org/10.1016/j.energy.2020.119397.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J.T., Blum, M. and Hutter, F., 2015. "Efficient and robust automated machine learning". *Advances in Neural Information Processing Systems*, Vol. 2015-January, pp. 2962–2970. ISSN 10495258.

Hutter, F., Hoos, H. and Leyton-Brown, K., 2011. "Sequential Model-Based Optimization for General Algorithm Configuration Lecture Notes in Computer Science". *International Conference on Learning and Intelligent Optimization*, pp. 507–523. URL https://www.cs.ubc.ca/~hutter/papers/11-LION5-SMAC.pdf%0Ahttp://dl.acm.org/citation.cfm?id=2177360.2 //www.springerlink.com/index/pdf/10.1007/.

Jawad, M., Nadeem, M.S.A., Shim, S.O., Khan, I.R., Shaheen, A., Habib, N., Hussain, L. and Aziz, W., 2020. "machine learning based cost effective electricity load forecasting model using correlated meteorological parameters". *IEEE Access*, Vol. 8, pp. 146847–146864. ISSN 21693536. doi:10.1109/ACCESS.2020.3014086.

Kononenko, I., Robnik-Sikonja, M. and Pompe, S., 2000. "Relieff for estimation and discretization of attributes in classification, regression, and ilp problems".

Menyah, K. and Wolde-Rufael, Y., 2010. "CO2 emissions, nuclear energy, renewable energy and economic growth in

the US". *Energy Policy*, Vol. 38, No. 6, pp. 2911–2915. ISSN 03014215. doi:10.1016/j.enpol.2010.01.024. URL http://dx.doi.org/10.1016/j.enpol.2010.01.024.

Mi, X. and Zhao, S., 2020. "Wind speed prediction based on singular spectrum analysis and neural network structural learning". *Energy Conversion and Management*, Vol. 216, No. May. ISSN 01968904. doi:10.1016/j.enconman.2020.112956. URL https://doi.org/10.1016/j.enconman.2020.112956.

Moshkbar-Bakhshayesh, K., 2020. "Prediction of unmeasurable parameters of NPPs using different model-free methods based on cross-correlation detection of measurable/unmeasurable parameters: A comparative study". *Annals of Nuclear Energy*, Vol. 139. ISSN 18732100. doi:10.1016/j.anucene.2019.107232. URL https://doi.org/10.1016/j.anucene.2019.107232.

Mwangi, M.A., Yong-kuo, L. and Ochieng, A.S., 2021. "Small Break Loss of Coolant Accident (SB-LOCA) fault diagnosis using Adaptive Neuro-Fuzzy Inference System (ANFIS) Small Break Loss of Coolant Accident (SB-LOCA) fault diagnosis using Adaptive Neuro-Fuzzy Inference System (ANFIS)". doi:10.1088/1755-1315/675/1/012034.

Nguyen, H.P., Liu, J. and Zio, E., 2020. "A long-term prediction approach based on long short-term memory neural networks with automatic parameter optimization by Tree-structured Parzen Estimator and applied to time-series data of NPP steam generators". *Applied Soft Computing Journal*, Vol. 89. ISSN 15684946. doi:10.1016/j.asoc.2020.106116. URL https://doi.org/10.1016/j.asoc.2020.106116.

Pearson, K., 1905. "The Fault Law and its Generalisation by Fechner and Pearson". *Biometrics*, Vol. 4, No. 1/2, pp. 169–212.

Qi, J., Du, J., Siniscalchi, S.M., Ma, X. and Lee, C.H., 2020. "On Mean Absolute Error for Deep Neural Network Based Vector-to-Vector Regression". *IEEE Signal Processing Letters*, Vol. 27, pp. 1485–1489. ISSN 15582361. doi:10.1109/LSP.2020.3016837.

Radaideh, M.I., Pigg, C., Kozlowski, T., Deng, Y. and Qu, A., 2020. "Neural-based time series forecasting of loss of coolant accidents in nuclear power plants". *Expert Systems with Applications*, Vol. 160. ISSN 09574174. doi:10.1016/j.eswa.2020.113699. URL https://doi.org/10.1016/j.eswa.2020.113699.

Robnik, M. and Konenko, I., 2003. "Theoretical and empirical analysis of ReliefF and RReliefF". *Machine Learning*, Vol. 53, No. 1–2, pp. 23–69.

Robnik-Sikonja, M. and Kononenko, I., 1997. "An adaptation of relief for attribute estimation in regression".

Sharma, R. and Bhandari, R., 2015. "Skewness, kurtosis and Newton's inequality". *Rocky Mountain Journal of Mathematics*, Vol. 45, No. 5, pp. 1639–1643. ISSN 0035-7596. doi:10.1216/rmj-2015-45-5-1639.

Tang, L., Yu, L., Wang, S., Li, J. and Wang, S., 2012. "A novel hybrid ensemble learning paradigm for nuclear energy consumption forecasting". *Applied Energy*, Vol. 93, pp. 432–443. ISSN 03062619. doi:10.1016/j.apenergy.2011.12.030. URL http://dx.doi.org/10.1016/j.apenergy.2011.12.030.

Urbanowicz, R.J., Meeker, M., La Cava, W., Olson, R.S. and Moore, J.H., 2018. "Relief-based feature selection: Introduction and review". *Journal of Biomedical Informatics*, Vol. 85, pp. 189 – 203. ISSN 1532-0464. doi:https://doi.org/10.1016/j.jbi.2018.07.014. URL http://www.sciencedirect.com/science/article/pii/S1532046418301400.

U.S. Energy Information Administration, 2021. "Nuclear Uranium". URL https://www.eia.gov/nuclear/generation/index.html.

Wang, C., Back, T., Hoos, H.H., Baratchi, M., Limmer, S. and Olhofer, M., 2019. "Automated Machine Learning for Short-term Electric Load Forecasting". *2019 IEEE Symposium Series on Computational Intelligence, SSCI 2019*, pp. 314–321. doi:10.1109/SSCI44817.2019.9002839.

Wang, H.Z., Wang, G.B., Li, G.Q., Peng, J.C. and Liu, Y.T., 2016. "Deep belief network based deterministic and probabilistic wind speed forecasting approach". *Applied Energy*, Vol. 182, pp. 80–93. ISSN 03062619. doi:10.1016/j.apenergy.2016.08.108.

Weisberg, S., 2001. "Yeo-Johnson Power Transformations". *Department of Applied Statistics, University of Minnesota*, , No. 2, pp. 1–4. URL http://stat.umn.edu/arc/yjpower.pdf.

Wolde-Rufael, Y., 2010. "Bounds test approach to cointegration and causality between nuclear energy consumption and economic growth in India". *Energy Policy*, Vol. 38, No. 1, pp. 52–58. ISSN 03014215. doi:10.1016/j.enpol.2009.08.053. URL http://dx.doi.org/10.1016/j.enpol.2009.08.053.

Yeo, I. and Johnson, R.A., 2000. "A new family of power transformations to improve normality or symmetry". *Biometrika*, Vol. 87, No. 4, pp. 954–959. ISSN 0006-3444. doi:10.1093/biomet/87.4.954. URL https://doi.org/10.1093/biomet/87.4.954.

Zahra, S., Ghazanfar, M.A., Khalid, A., Azam, M.A., Naeem, U. and Prugel-Bennett, A., 2015. "Novel centroid selection approaches for KMeans-clustering based recommender systems". *Information Sciences*, Vol. 320, pp. 156–189. ISSN 00200255. doi:10.1016/j.ins.2015.03.062. URL http://dx.doi.org/10.1016/j.ins.2015.03.062.