ISSN 1870-4069

Supervised Machine Learning Application for Developing a Predictive Model of the Monthly Phase of the Pacific Decadal Oscillation

Indalecio Mendoza Uribe

Instituto Mexicano de Tecnología del Agua,
Mexico

indalecio_mendoza@tlaloc.imta.mx

Abstract. In this work, supervised machine learning was applied, using regression trees, to develop a predictive model of the monthly phase of the Pacific Decadal Oscillation. This oscillation is associated with the alteration of weather patterns, mainly in the North Pacific and southwestern North America. As characteristics, the records of the PDO phase of the 24 months prior to the forecast target month were used. The predictive model developed presented an acceptable capacity to estimate the monthly phase of the PDO. This according to the performance evaluation statistics corresponding to the Mean Absolute Error, Maximum Error, Mean Quadratic Error and Pearson's Correlation, which obtained ranges of [0.55,1.07], [1.58,3.29], [0.55,1.82] and [0.30,0.74] respectively for 20% of test data for the period 1854-2020.

Keywords: Artificial intelligence, climate, regression trees.

1 Introduction

In climatology, machine learning has great potential, especially in phenomena of long temporal development, as is the case of the Pacific Decadal Oscillation (PDO). PDO is mainly characterized by changes in sea surface temperature (SST) in the Pacific Ocean over 20° north latitude, as well as variation in sea level pressure and wind patterns. The study of the PDO has gained relevance in recent years due to its association with the alteration of weather patterns, mainly in the North Pacific and southwestern North America [1, 2, 3, 4]. Alterations in the climate have significant socioeconomic impacts, especially in countries that base their development on the management of their natural resources [5].

In the area of artificial intelligence, various machine learning techniques have been applied to understand, describe and predict the behavior of natural phenomena. Ovando et al. [6] developed a model based on neural networks to predict the occurrence of frost in Argentina, based on meteorological data of temperature, relative humidity, cloud cover, wind direction and speed. On the other hand, Téllez-Valero et al. [7] developed a system based on machine learning methods that improves the acquisition of data from natural disasters, the system automatically populates a database of natural disasters with information extracted from online newspaper news. In addition, Haro-Rivera [8] applied a decision tree to identify predominant meteorological variables in the province of Chimborazo, Ecuador. Finally, in this list of examples, Suárez et al. [9] analyzed the meteorological phenomenon called DANA, which caused serious floods, human losses, economic and infrastructure damage in the southeast of Spain during the month of September 2019, studying the phenomenon from the perspective of data analysis.

Machine learning is a data analysis technique that gives computers the ability to learn from experience without relying on a given equation as a model. These algorithms look for natural patterns in the data that generate knowledge. Algorithms adaptively improve their performance as the number of samples available for learning increases. In a general way, we can classify machine learning techniques as supervised and unsupervised.

A supervised learning algorithm takes a set of known data (inputs) and known responses for this data (outputs) to train a model that can generate reasonable predictions in response to new data. Supervised learning uses classification and regression techniques to develop predictive models. In comparison, unsupervised learning looks for hidden patterns or intrinsic structures in the data. Used to infer information from data sets consisting of input data with no labeled responses. Among the most common unsupervised learning techniques are neural networks [10], k-means [11], among other.

The objective of this work was to apply supervised machine learning through regression trees to develop a predictive model of the monthly phase of the Pacific Decadal Oscillation. As characteristics, the records of the PDO phase of the 24 months prior to the target month of prognosis were used.

2 Method

The development of the predictive model was carried out by applying three procedures. First, the historical data set of the monthly value of the PDO was obtained for the period 1854-2020. These data were organized by month and grouped into training and test data. Second, for each month of the year the regression tree corresponding to the predictive model was generated with the training data. Third, the monthly predictive models were applied on the test data sets. The results were evaluated using three continuous error measurement metrics and one of correlation.

2.1 Dataset

The PDO is a pattern of anomalies of the SST, this fluctuation oscillates between -4 and 4 degrees centigrade, corresponding to the cold and warm phase respectively. The PDO values indicate the variation of the SST with respect to the historical average. The data was obtained from National Oceanic and Atmospheric Administration through the URL https://www.ncdc.noaa.gov/teleconnections/pdo/data.csv. The data set corresponds to the monthly deviation of the SST for the period 1854-2020 (see Fig. 1). For each forecast month (label) the values of the previous 24 months were assigned as characteristics. The characteristics and labels for each month of the year were grouped in separate files to facilitate their processing.

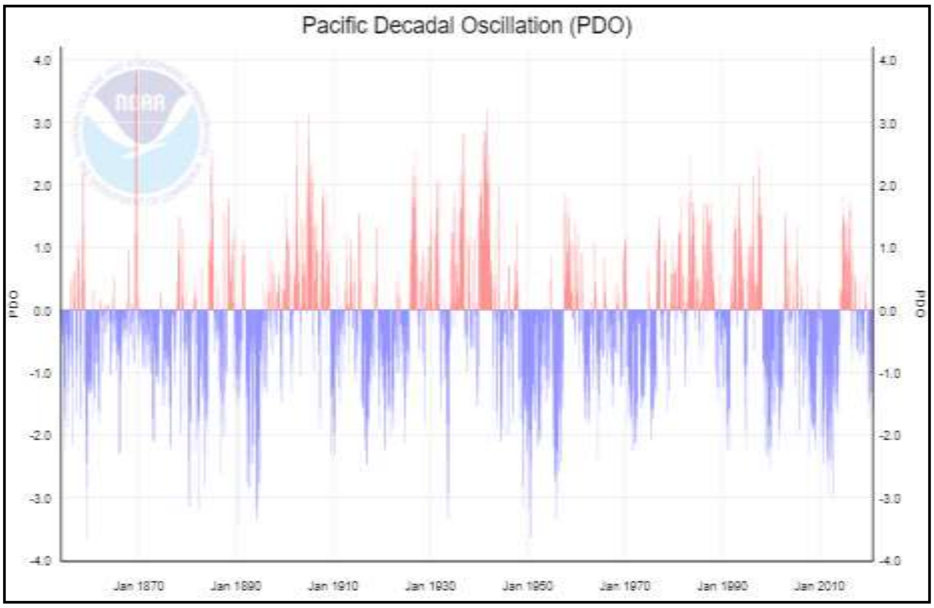

Fig. 1. Monthly anomaly of the Pacific Decadal Oscillation for the period 1854-2020 [12].

Machine learning consists of learning some properties of a data set and then verifying those properties with another data set. A common practice in machine learning is to evaluate an algorithm by dividing the data into two subsets. The majority set is dominated by training data, from which the algorithm learns some properties. While the second set of data is called test data, with which the ability of the model to predict through the learned properties is verified. For this study, the training and test data set were divided into a proportion of 80 and 20% respectively.

2.2 Generation of the Predictive Model

For each month of the year, the regression tree corresponding to the predictive model was generated. Each predictive model was trained with 80% corresponding training data.

Classification and regression trees (CART) were developed by Breiman et al. [13]. Tree models where the target variable can take a finite set of values are called classification trees. On the other hand, trees where the target variable can take continuous values are called regression trees.

Let Y be the response variable and x be the vector with the set of predictor variables, the problem corresponds to establishing a relationship between Y and x in such a way that it is possible to predict *Y* based on the values of *x*. Mathematically looking for probability $P(Y \mid x_1, x_2, \ldots, x_k)$.

The construction of the tree is done following a recursive binary division approach, let *N* be the number of data and N_j the number of cases in class *j*.

The probability that a case is in class *j* given that it was located in the terminal node *t*, is given by the Eq. 1.

$$P(j \mid t) = \frac{P(j,t)}{P(t)} = \frac{N_j(t)}{N} \tag{1}$$

and comply with:

$$\sum P(j \mid t) = 1. \tag{2}$$

Thus, the set of $P(j|t)$ are the relative proportions of the cases in class *j* at node *t* [8].

To obtain the optimal tree, evaluate each subdivision among all possible trees, get the root node and the subsequent ones, the algorithm must measure the predictions achieved and evaluate them to select the best one. Fig. 2 shows a simplified form of a regression tree.

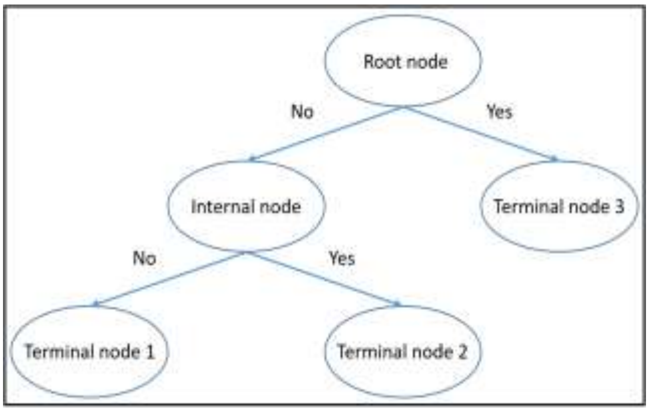

Fig. 2. Simplified form of a regression tree.

In this study, machine learning was applied through the *Scikit-Learn* library of the Python programming language, which integrates a wide range of machine learning algorithms for supervised and unsupervised problems [14].

Specifically, the *tree.DecisionTreeRegressor* method was used to create the instance of the predictive model; *train_test_split* to divide the training/test data set; *mean_absolute_error, mean_squared_error y max_error* to measure mean absolute error, mean square error and maximum error respectively; finally, the function *plot_tree* was used to graph the regression trees.

2.3 Statistical Validation of the Predictive Model

Monthly predictive models were applied on the corresponding test data sets. For the evaluation of the monthly predictive model of the PDO phase, three continuous error measurement metrics and Pearson's correlation were used. These metrics are recommended for evaluating forecasts of a deterministic nature. These metrics are described below.

The Mean Absolute Error (MAE) measures the magnitude of the errors in a set of predictions, regardless of their direction [15, 16]. It corresponds to the average of the absolute differences between the prediction and the observation where all the individual differences have the same weight (Eq. 3):

$$\mathrm{MAE} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{|\mathrm{P_i} - \mathrm{O_i}|}{n}, \tag{3}$$

where P_i is the prediction value at position i, O_i is the value observed at position i and n is the sample size.

The Maximum Error (ME) allows to identify the largest absolute value of the observed error between the prediction and the observation (Eq. 4). It belongs to the set of objective functions used for the calibration of models [17]:

$$\mathrm{ME} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \max\{|\mathrm{P_i} - \mathrm{O_i}|\}. \tag{4}$$

The Root Mean Square Root (RMSE) measures the mean magnitude of the error. Corresponds to the square root of the average of the squared differences between the prediction and the observation, therefore this measure has been used in the evaluation of forecasting models [18, 19]. Amplifies and penalizes with greater force those errors of greater magnitude (Eq. 5):

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{P_i} - \mathrm{O_i})^2}. \tag{5}$$

Pearson's Correlation, denoted as *r* (Eq. 6), is a normalized measure widely used to establish relationships between two continuous quantitative variables [20, 21]. It allows to show the joint variability and therefore to typify what happens with the data. The coefficient can score values ranging from -1.0 to 1.0 and is interpreted as follows: values close to 1.0 indicate that there is a strong association between the variables, that is, they increase or decrease in the same direction.

On the other hand, values close to -1.0 indicate that there is a strong negative association between the variables, that is, as one variable increases, the other decreases. A value of 0.0 indicates that there is no correlation or it is a null correlation [22].

$$r \; = \; \frac{\sum_{i=n}^{n}(P_i - \bar{P})(O_i - \bar{O})}{\sqrt{\sum_{i=n}^{n}(P_i - \bar{P})^2}\;\sqrt{\sum_{i=n}^{n}(O_i - \bar{O})^2}}, \tag{6}$$

where $\bar{P}$ is the mean value of the predictions and $\bar{O}$ is the mean value of the observations.

3 Results

For the creation of the monthly predictive models based on regressive trees, the constructor of the *DecisionTreeRegressor* class was used. Table 1 lists the parameters used during the creation of the predictive model with which the best results were obtained.

Table 1. Predictive model creation parameters.

Parameter	Value	Description
criterion	mse	Function to measure the quality of the division.
splitter	best	Strategy used to choose the division at each node.
max_depth	None	Maximum depth of the tree. None indicates that nodes are expanded until all sheets are pure or until all sheets contain less than *min_samples_split* samples.
min_samples_split	2	The minimum number of samples required to divide an internal node.
min_samples_leaf	1	The minimum number of samples required to be in a leaf node.
max_features	12	The number of features to consider when looking for the best division.
random_state	5	Controls the randomness of the estimator. To obtain a deterministic behavior during the setting *random_state* must be set to an integer.

As part of the training, the algorithm identifies the impact on the prognosis of each of the characteristics. As can be seen in Table 2, in general, with 12 characteristics, more than 90% importance is obtained in the forecast.

These 12 characteristics are not the same for all months of the year, therefore, in the training stage, the 24 characteristics are initially considered, but the algorithm is instructed to only select the 12 most relevant characteristics. This reduction in dimensions allows the algorithm to be optimized by eliminating characteristics that do not contribute to the forecast.

Table 2. Percentage of importance by characteristics for monthly predictive models.

		Month											
		Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
Characteristics	**1**	1	0	1	0	3	8	4	2	0	1	3	1
	2	0	1	1	0	0	3	1	0	0	2	4	7
	3	1	0	3	1	1	2	1	1	0	3	3	1
	4	0	1	2	0	2	0	0	3	5	4	0	0
	5	5	2	1	3	2	0	2	5	1	2	0	0
	6	0	1	1	0	1	3	0	2	0	2	1	0
	7	1	0	0	4	3	2	1	2	0	0	1	0
	8	0	2	0	0	2	0	1	1	3	0	1	0
	9	2	1	1	1	0	2	2	1	5	0	1	6
	10	4	3	2	0	3	1	0	3	2	2	1	3
	11	3	0	2	0	0	0	1	2	0	0	2	4
	12	0	1	4	0	0	0	1	0	3	2	4	2
	13	8	6	0	0	0	0	1	4	0	2	0	1
	14	0	1	1	2	1	0	2	0	1	0	1	0
	15	6	0	3	1	5	0	1	1	1	2	1	0
	16	3	1	2	1	1	0	0	2	3	1	0	0
	17	0	3	1	0	1	2	0	0	1	7	5	7
	18	1	2	1	0	0	3	0	5	0	1	1	11
	19	2	1	1	5	1	0	2	2	3	4	4	2
	20	0	3	1	1	2	6	3	3	0	2	1	2
	21	5	1	0	5	1	0	3	2	2	6	0	2
	22	0	1	1	3	2	2	1	37	7	1	1	0
	23	39	49	56	50	51	35	58	14	36	38	41	40
	24	19	20	15	23	18	31	15	8	27	18	24	11

Algorithm 1 presents in a simplified way the sequence of steps to divide the data into the training/test subsets, feed the classifier (predictive model) with the training data, apply the classifier on the test data, calculate model performance evaluation metrics, graphing and data storage. Clarification is made that the algorithm does not detail the modules of *dataReadingMonth()* and *graphingStorage()*.

Algorithm 1: Simplified sequence to generate, train, apply and validate the monthly predictive model

```
for month in range(0,12):
    totalCharacteristics, totalLabels = dataReadingMonth(month)
    trainingCharacteristics, testCharacteristics, trainingLabels, testLabels =                \
              train_test_split(totalCharacteristics, totalLabels,train_size=0.80,             \
              test_size=0.20, random_state= 5)
    # Creation of the instance (object) of type DecisionTreeRegressor (predictive model)
    predictiveModel = tree.DecisionTreeRegressor(criterion = 'mse', splitter = 'best',        \
               max_depth = None, min_samples_split = 2, min_samples_leaf = 1,                 \
               max_features = 12, random_state=5)
    # Feed the classifier with the training data (train the predictive model)
    predictiveModel.fit(trainingCharacteristics,trainingLabels)
    # Apply the predictive model to the test data set
    predictions = predictiveModel.predict([testCharacteristics])
    predictedLabels = predictions[0]
    # Calculate MAE, ME, RMSE and r performance metrics
    mae      = round(mean_absolute_error(testLabels,predictedLabels),2)
    me       = round(max_error(testLabels,predictedLabels),2)
    rmse     = round(mean_squared_error(testLabels,predictedLabels),2)
    pearson = sc.pearsonr(testLabels,predictedLabels)
    r = round(pearson[0],2)
    storageGraph(month,predictiveModel,mae,me,rmse,r)
```

Figures 3 and 4 show arbitrarily the trees corresponding to the predictive models for the months of June and December, respectively.

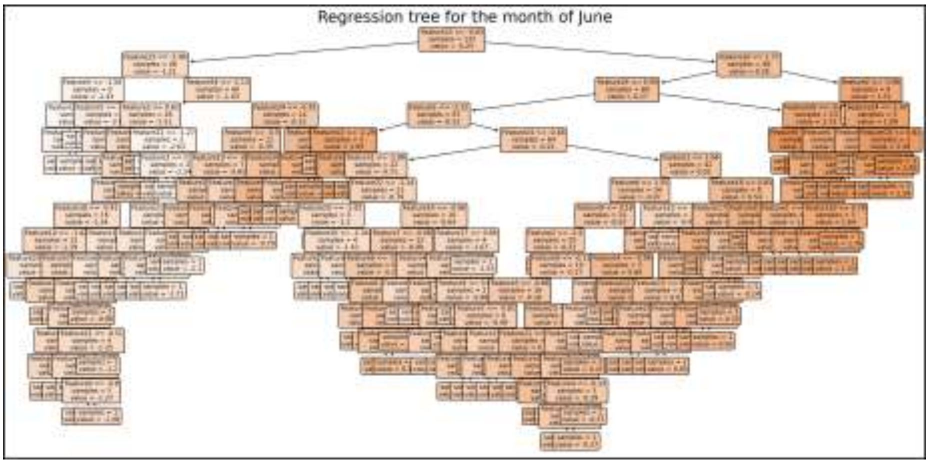

Fig. 3. Regression tree for the month of June. The predictive model was trained with 80% of data from the period 1854-2020. The strongest fill color indicates the majority class for classification.

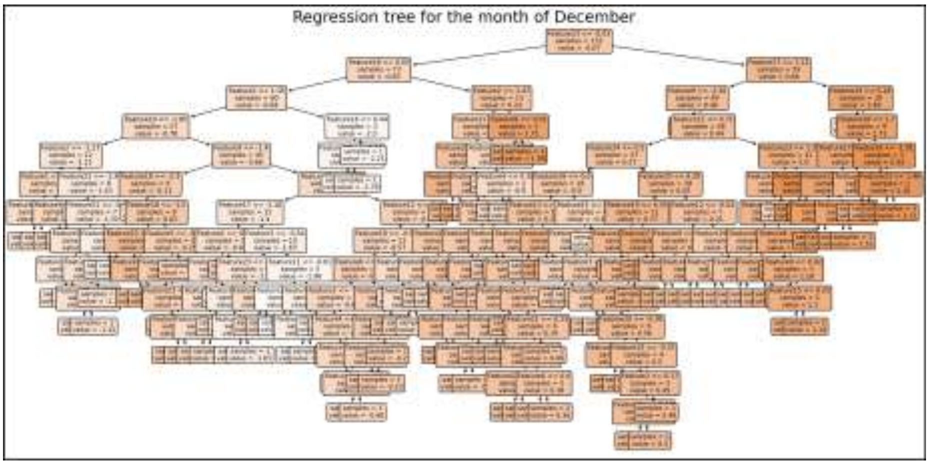

Fig. 4. Regression tree for the month of December. The predictive model was trained with 80% of data from the period 1854-2020. The strongest fill color indicates the majority class for classification.

Monthly predictive models were applied for 20% of test data. Table 3 shows the results of the four statistical metrics applied by the monthly predictive model. Besides that, Fig. 5 shows the dispersion diagrams with the comparison between the observed and predicted data.

Table 3. Result of the statistical metrics of the monthly predictive models.

Target Month	MAE	ME	RMSE	r
January	0.79	2.49	1.03	0.59
February	0.66	2.09	0.70	0.64
March	0.61	2.44	0.66	0.72
April	0.74	1.58	0.74	0.74
May	0.86	2.52	1.13	0.62
June	0.64	1.79	0.62	0.77
July	1.01	2.47	1.44	0.55
August	1.07	3.29	1.82	0.30
September	1.01	2.83	1.58	0.38
October	0.69	1.60	0.68	0.74
November	0.55	1.75	0.55	0.77
December	0.89	3.53	1.31	0.47

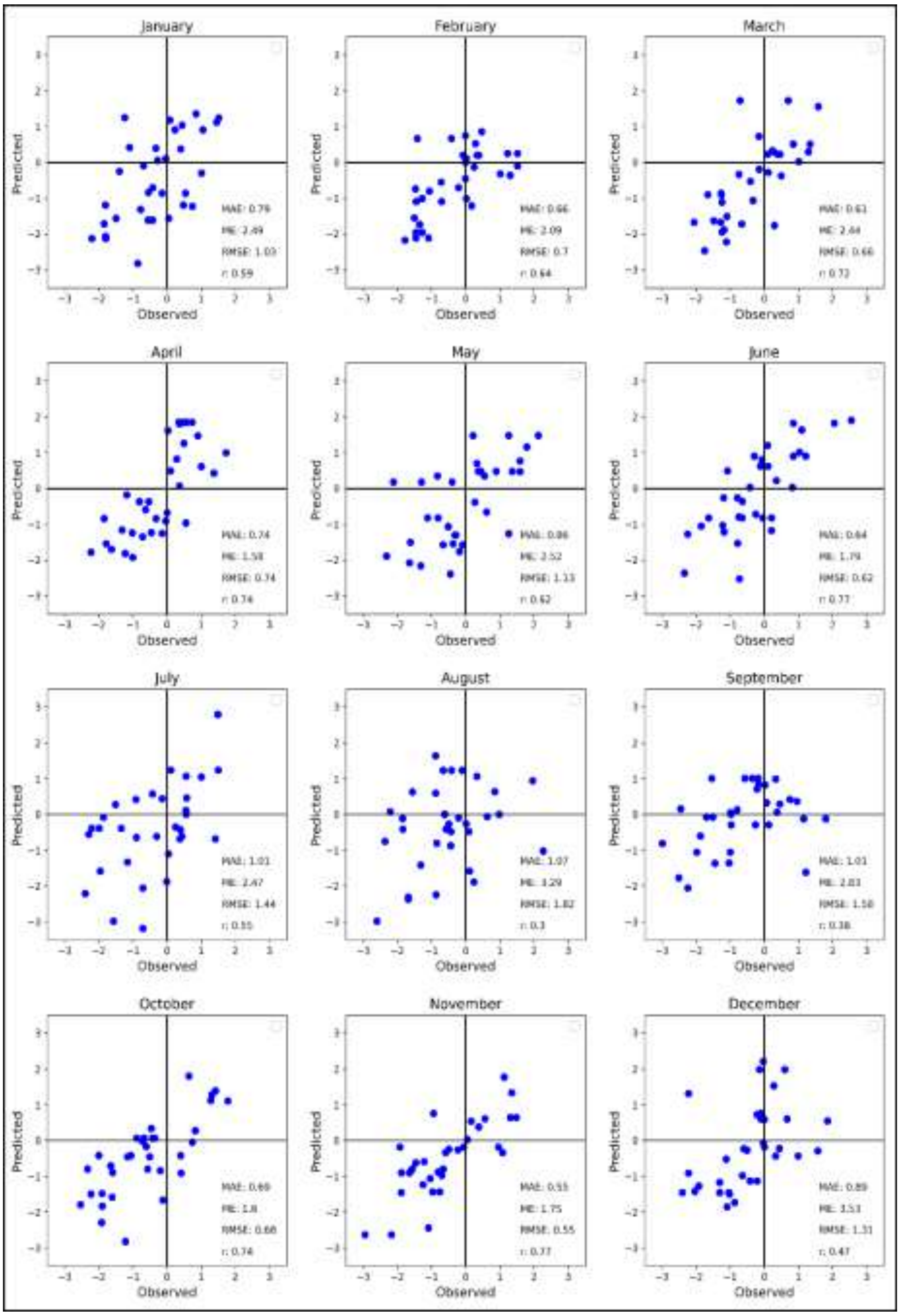

Fig. 5. Monthly dispersion diagrams between observed and predicted values for 20% of test data for the 1854-2020 period.

4 Conclusions

Of the 24 characteristics considered, it was identified that characteristic 23 in eleven months and characteristic 22 in the month of July, predominated as root node in the trees of the predictive models, that is, these characteristics have a greater impact on forecasts. In addition, it was distinguished that in 12 characteristics more than 90% of importance is obtained in the prognosis.

The predictive model developed using machine learning presented an acceptable capacity to estimate the monthly phase of the PDO. This according to the results of the performance evaluation statistics MAE, ME, RMSE and r obtained for 20% of test data, with ranges of [0.55, 1.07], [1.58, 3.29], [0.55, 1.82] y [0.30, 0.74] respectively. Therefore, it is considered that the predictive model developed can constitute a reference forecasting tool, but not an exact one.

As future work, it is proposed to continue with the validation and adjustment of the predictive model for its application in larger time windows, such as for seasonal forecast (3 months), or even annual forecast.

Regarding the functionality of the Scikit-Learn library, this turned out to be docile to implement and very efficient in its performance. The computational cost required for the training and testing of the predictive model was of the order of seconds on a personal computer.

References

1. Mantua, N.J., Hare, S.R.: The Pacific Decadal Oscillation. Journal of Oceanography, 58, 35–44 (2002). Doi: https://doi.org/10.1023/A:1015820616384
2. Cayan, D.R., Dettinger, M.D., Diaz, H.F., Graham, N.E.: Decadal variability of precipitation over western North America. Journal of Climate, 11, 3148-3166 (1998). Doi: https://doi.org/10.1175/1520-0442(1998)011<3148:DVOPOW>2.0.CO;2
3. Higgins, R.W., Leetmaa, A., Xue, Y., Barnston, A.: Dominant factors influencing the seasonal predictability of U.S. precipitation and surface air temperature. Journal of Climate, 13(22), 3994–4017 (2000). Doi: https://doi.org/10.1175/1520-0442(2000)013<3994:DFITSP>2.0.CO;2
4. Gutzler, D.S., Kann, D.M., Thornbrugh, C.: Modulation of ENSO-based long-lead outlooks of Southwest U.S. Winter precipitation by the Pacific Decadal Oscillation. Weather and Forecasting, 17, 1163–1172 (2002).
5. Méndez-González, J., Ramírez-Leyva, A., Zárate-Lupercio, A., Cavazos-Pérez, T.: Teleconexiones de la Oscilación Decadal del Pacífico (PDO) a la precipitación y temperatura en México. Investigaciones Geográficas, 73, 57–70 (2010).
6. Ovando, G., Bocco, M., Sayago, S.: Redes neuronales para modelar predicción de heladas. Agricultura Técnica, 65(1), 65–73 (2005). Doi: http://dx.doi.org/10.4067/S0365-28072005000100007
7. Téllez-Valero, A., Montes, M., Villaseñor-Pineda, L.: Using Machine Learning for Extracting Information from Natural Disasters News Reports. Computación y Sistemas, 13(1), 33–44 (2009).
8. Haro-Rivera, S.M.: Árbol de decisión, aplicación con datos meteorológicos. KnE Engineering, 5(2), 37–46 (2020). Doi: https://doi.org/10.18502/keg.v5i2.6217

9. Suárez, L., Alarcon, P.A.: Inteligencia artificial para la comprensión de desastres naturales. Telefónica Digital, España (2020).
10. Mercado-Polo, D., Pedraza-Caballero, L., Martínez-Gómez, E.: Comparación de Redes Neuronales aplicadas a la predicción de Series de Tiempo. Prospectiva, 13(2), 88–95 (2015). Doi: http://dx.doi.org/10.15665/rp.v13i2.491
11. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An efficient k-means clustering algorithm: analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 881–892 (2002). Doi: https://doi.org/10.1109/TPAMI.2002.1017616
12. NOAA (National Oceanic and Atmospheric Administration) Pacific Decadal Oscillation (PDO), https://www.ncdc.noaa.gov/teleconnections/pdo, last accessed 2021/02/22.
13. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Chapman & Hall/CRC, New York (1984). Doi: https://doi.org/10.1201/9781315139470
14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12(85), 2825–2830 (2011).
15. Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30, 79–82 (2005).
16. Karamirad, M., Omid, M., Alimardani, R., Mousazadeh, H., Heidari, S.N.: ANN based simulation and experimental verification of analytical four and five parameters models of PV modules. Simulation Modelling Practice and Theory, 34, 86–98 (2013).
17. Gupta, H.V., Sorooshian, S., Yapo, P.O.: Toward improved calibration of hydrologic models: Multiple and noncommesurable measure of information. Water Resources Research, 34(4), 751–763 (1998).
18. González-Leyva, F., Ibáñez-Castillo, L.A., Valdés, J.B., Vázquez-Peña, M.A., Ruiz-García, A.: Pronóstico de caudales con Filtro de Kalman Discreto en el río Turbio. Tecnología y Ciencias del Agua, 6(4), 5-24 (2015).
19. Vázquez, M.: Predicción de series de tiempo usando un modelo híbrido basado en la descomposición wavelet. Comunicaciones estadísticas, 11(2), 257–283 (2018).
20. Restrepo, L.F., González, J.: De Pearson a Spearman. Revista Colombiana de Ciencias Pecuarias, 20, 183–192 (2007).
21. Martínez-Curbelo, G., Cortés-Cortés, M.E., Pérez-Fernández, A.C.: Metodología para el análisis de correlación y concordancia en equipos de mediciones similares. Universidad y Sociedad, 8(4), 65–70 (2016).
22. Anderson, R.B., Doherty, M.E., Friedrich, J.C.: Sample size and correlational inference. Journal of Experimental Psychology: Learning, Memory and Cognition, 34(4), 929–944 (2008). Doi: https://doi.org/10.1037/0278-7393.34.4.929