Wildcard Project

Mohammad Irfan Uddin

Description of the Problem:

The Climate-Leaf Analysis Multivariate Program (CLAMP) is a widely used method for reconstructing climatic variables up to 90 million years ago. However, the statistical methods that are at the core of CLAMP have not changed for a long time. Using new machine learning techniques, we can attempt to find better fits between the CLAMP dataset and the climate, and thus improve our understanding of the development of the Earth's climate in the past.

CLAMP (Climate-Leaf Analysis Multivariate Program) is a program that utilizes a form of multivariate statistical analysis called Canonical Correspondence Analysis (CCA). The analysis estimates a mathematical equation that aims to relate climatic features to particular floras using 31 different leaf characteristics. CLAMP's statistical analysis is calibrated from a dataset of modern flora's leaf physiognomy and their respective climates; each of the two types of data are held in two different matrices; these two matrices then derive the mathematical equations which relate the dataset to their corresponding climate features.

The CLAMP dataset comprises 144 samples, each with 31 features representing distinct characteristics of plant leaves from various regions in the USA. These features encompass attributes like lobe shape, tooth type, length-to-width ratio (L:W), and other morphological traits. Each entry in the dataset corresponds to a specific plant specimen, providing values for particular characteristics. Leveraging these leaf features, our model can effectively predict diverse climate parameters, including Relative Humidity (RH), Specific Humidity (SH), Mean Annual Temperature, Precipitation (GSP), and Enthalpy.

Experimental Setup:

Tuned Hyperparameter and ranges:

SVR	
Hyperparameter	Grid Search
C	10
Kernel	rbf

Random Forest Regression	
Hyperparameter	Grid Search
max_depth	None
min_samples_leaf	1
min_samples_split	5
n_estimators	50

Decision Tree Regression	
Hyperparameter	Grid Search
max_depth	20
min_samples_leaf	2
min_samples_split	2

Hyperparameter ranges:

SVR:

C: [0.1, 0.8, 2, 10]

kernel: linear', 'rbf', 'poly', 'sigmoid

Random Forest:

n_estimators: [50, 150, 100, 200]

max_depth: [None, 10, 18, 30]

min_samples_split: [2, 5, 10]

min_samples_leaf: [1, 2, 5]

Decision Tree:

max_depth: [None, 10, 20, 30, 50, 100]

min_samples_split: [2, 5, 10, 20]

min_samples_leaf: [1, 2, 4, 7]

HPO technique: GridSearch

Run Time: 18 minutes

Total number of evaluations during HPO: **1290**

Analysis of the results:

Model	GSP		SH		MAT	
	B_HPO	A_HPO	B_HPO	A_HPO	B_HPO	A_HPO
CLAMP(CCA)	0.45		0.68		0.72	
Random Forest	0.73	0.75	0.73	0.68	0.91	0.92
SVR	0.08	0.84	0.73	0.84	0.85	0.94
Linear Regression	0.81	0.81	0.8	0.8	0.9	0.93
Decision Tree	0.59	0.55	0.65	0.58	0.67	0.7

B_HPO= Before HPO, A_HPO= After HPO
MAT=mean Annual Temperature
GSP= Growing Season Precipitation
SH= Specific Humidity
Can also predict: RH, Enthalpy

The obtained results clearly indicate that our model's prediction performance surpasses that of the original CLAMP. While the original CLAMP solely employed Support Vector Regression (SVR), I've enhanced our model's versatility. It dynamically explores various models based on the input features, allowing it to select the model that yields the optimal performance for predictions.

Some location Climate prediction with our model:

Sandstone Creek OR:

Climate Parameters	Actual Measurement	Predictions
Enthalpy	30.12	29.76
Mean Annual Temperature	11.2	9.6
Growing Season Precipitation	60	47
Specific Humidity	5.9	4.6
Relative Humidity	70	70

Kissinger Lake, Wyoming climate Parameters Prediction:

Enthalpy : 30.61
Mean Annual Temperature : 13.6
Growing Season Precipitation : 100
Specific Humidity : 7.3
Relative Humidity : 76

For Kissinger Lake, we don’t have the actual measurements to compare the predictions, however, got the leaf features from CLAMP.