Tree-based explainable clustering for drought severity predictions in United States

Stelios P. Neophytides*[a, b], Michalis Mavrovouniotis[a,b], Marinos Eliades[a,b], Felix Bachofer[c], Diofantos G. Hadjimitsis[a,b]

[a]ERATOSTHENES Centre of Excellence, Franklin Roosevelt 82, 3012, Limassol, Cyprus; [b]Department of Civil Engineering and Geomatics, Cyprus University of Technology, Archiepiskopou Kyprianou, 3036, Limassol, Cyprus; [c]German Aerospace Centre, D-82234 Weßling, Oberpfaffenhofen, Germany

*stelios.neophytides@eratosthenes.org.cy

ABSTRACT

Climate change drives the environment to more extreme weather events. Increased air, land surface and canopy surface temperatures affect the industry of agriculture in different ways. Significant crop damages and losses are emerging and spreading throughout different regions, accompanied by water scarcity and imposed restrictions on farmers' water usage. The Eastern Mediterranean, Middle East, and North Africa (EMMENA) region is one of the most affected areas globally. The United States (US) developed a system for monitoring droughts in different counties and classifying them into six categories (i.e., no drought, abnormally dry, moderate drought, severe drought, extreme drought, and exceptional drought) based on the assigned drought score. To predict drought scores, Artificial Intelligence (AI) methodologies are applied to a dataset that combines meteorological variables from the NASA Langley Research Center with drought scores from the US drought monitor system. The main objective of this work is to propose a novel explainable AI technique based on unsupervised learning for drought severity predictions and raise the awareness for drought events in the wider EMMENA region.

Keywords: Climate change, droughts, explainable artificial intelligence.

INTRODUCTION

Numerous extreme weather events are the outcome of the climate crisis. Among these events, droughts are a major threat to global water security, agricultural productivity, and ecosystem's health. Accurate prediction of drought events is essential for developing mitigation strategies[1], [2]. By leveraging the ability of Artificial Intelligence (AI) to analyze vast datasets of environment and climate variables, advanced methods are developed for drought predictions.

Different clustering algorithms are tested on segmentation of Southern It0aly to drought regions and then regression models are used for droughts' time-series forecasting. The hybrid M5P-SVR model achieved an R-squared (R^2 or coefficient of determination) of 0.91 [3]. Another work uses well-known drought indice0.s such as standardized precipitation index (SPI), standardized precipitation evapotranspiration Index (SPEI) and standardized runoff index (SRI) to assess the drought in semi-arid environments. Several machine learning (ML) models are utilized for the prediction of those indices using different meteorological variables (e.g., temperature, rainfall, etc.). The hybrid wavelet-GPR achieved the highest accuracy with R^2 of 0.809 [4]. A similar work showed that wavelet-enhanced multi-layer perceptron neural network (NN) achieved the highest accuracy in drought prediction [5]. Additional work uses a hybrid deep learning model which consists of a convolutional NN as the feature extractor and a long short-term memory NN as the temporal predictor for droughts[6].

However, all the aforementioned methodologies lack explainability and interpretation [7], [8]. Only a few works examine the capabilities of using explainable AI (XAI) to understand the features with positive contribution for accurate predictions related to drought monitoring. A study conducted in the area of New South Wales, Australia showed that it is important to interpret such models based on region and shorter time periods instead of decade-based explanations[9]. An extension of this study explains how climatic variables are important at a monthly scale, as well as their varying annual ranges based on SHAP-based (SHapley Additive exPlanations) explanations [10]. Extreme gradient boosting (XGB) model is used to explore the drought impacts in the United States (US). Specifically, the patterns between the SPI and drought impact reveal that negative values of the index are positively leading the model to complex drought impacts [11]. A study conducted for Canadian droughts using the interpretable XGB model achieved an overall accuracy of 71.3% in predicting drought maps.

Tenth International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2024), edited by A. Christofe, S. Michaelides, D. Hadjimitsis, C. Danezis, K. Themistocleous, N. Kyriakides, G. Schreier, Proc. of SPIE Vol. 13212, 132121B · © 2024 SPIE · 0277-786X · doi: 10.1117/12.3037320

The application of SHAP-based explanations identified the relation between the drought event that took place in 2015 in Prairie, Canada with the El-Niño event which reduced the water availability [12]. An extension of this work employed remote sensing data too, suggests that the satellite-based evaporative stress index, soil moisture and groundwater levels are effectual features for drought onset and intensification [13]. Countries and regions with advanced technologies and infrastructures like U.S. [14], Germany[15] and North America[16] have developed expert systems for drought monitoring in high temporal resolution. Similar systems are yet not developed for Cyprus and the rest of EMMENA region which is characterized as a climate change hotspot. Thus, there is an urgent need for systematic monitoring of such extreme events [17], [18]. However, there is a lack of data related to the drought severity in the EMMENA region.

In this work, a tree-based explainable clustering methodology is proposed using data from US drought monitor system. This methodology can be helpful in predicting drought severity by using meteorological data in the EMMENA region to raise awareness. Clustering ML models are characterized as black box [19]. Therefore, various methodologies have been developed to add explainability on clustering algorithms like k-means through decision trees [20]. Such methodologies are proposing iterative techniques to extract high distinct clusters[21]. Similarly, those techniques are also applied in kernel clustering[20] and k-medians clustering [21].The proposed methodology uses k-means algorithm and SHAP-based XAI techniques applied in drought monitoring, and thus, can be effective in cases where the ground truth labels are missing and/or interpretability is necessary. The rest of the paper is structured as follows. Section 2 describes the proposed methodology and the dataset used in this study, Section 3 presents the experimental setup and the evaluation strategy used. Section 4 gives an overview of the experimental results regarding model's performance and insights derived from the SHAP-based explanations. Finally, Section 5 concludes this work.

MATERIALS AND METHODS

2.1 Drought monitoring dataset

For this study a Kaggle dataset (https://www.kaggle.com/datasets/cdminix/us-drought-meteorological-data/data, visited on 05/03/2024) is collected. The dataset contains meteorological variables acquired from the NASA Langley Research Centre POWER Project and annotated based on the drought scores from the US drought monitor system. The measurements are acquired for the period of January 2000 to January 2020. Furthermore, each sample in the dataset is matched with the observation's date and the US county. Tables 1 and 2 describe all the different meteorological variables used and provide descriptive statistics for the dataset. Furthermore, Table 3 provides an overview of the various classes defined by the US drought monitor, which serve as the labels for this study.

Table 1. Description of each meteorological variable of the dataset.

Variable	Description	Unit
PRECTOT	Precipitation	mm/day
PS	Surface Pressure	kPA
Q2VM	Specific Humidity at 2 Meters	g/kg
T2MDEW	Dew/Frost Point at 2 Meters	oC
T2MWET	Wet Bulb Temperature at 2 Meters	oC
T2M	Temperature at 2 Meters	oC
T2M_MAX	Maximum Temperature at 2 Meters	oC
T2M_MIN	Minimum Temperature at 2 Meters	oC
T2M_RANGE	Range of Temperature at 2 Meters	oC
TS	Earth Skin Temperature	oC
WS10M	Wind Speed at 10 Meters	m/s
WS10M_MAX	Maximum Wind Speed at 10 Meters	m/s
WS10M_MIN	Minimum Wind Speed at 10 Meters	m/s

WS10M_RANGE	Range of Wind Speed at 10 Meters	m/s
WS50M	Wind Speed at 50 Meters	m/s
WS50M_MAX	Maximum Wind Speed at 50 Meters	m/s
WS50M_MIN	Minimum Wind Speed at 50 Meters	m/s
WS50M_RANGE	Range of Wind Speed at 50 Meters	m/s

Table 2. Basic descriptive statistics for all the meteorological data in which min, max, mean, and stdev represent the minimum, maximum, mean, and standard deviation values of the dataset.

Variable	**min**	**max**	**mean**	**stdev**
PRECTOT	0.000	21.440	1.764	3.702
PS	80.140	103.270	96.931	4.717
Q2VM	0.560	20.570	8.003	4.573
T2MDEW	-21.910	25.420	7.526	9.855
T2MWET	-20.920	25.420	7.567	9.783
T2M	-18.930	38.910	15.119	11.023
T2M_MAX	-15.03	46.15	21.590	11.735
T2M_MIN	-23.63	30.51	9.311	10.498
T2M_RANGE	0.47	23.03	12.279	4.014
TS	-19.51	40.75	15.331	11.370
WS10M	0.65	8.79	3.319	1.552
WS10M_MAX	0.96	12.36	4.869	2.214
WS10M_MIN	0.01	5.57	1.770	1.129
WS10M_RANGE	0.39	8.6	3.099	1.640
WS50M	1.09	11.37	5.203	1.971
WS50M_MAX	1.69	14.79	7.435	2.375
WS50M_MIN	0.02	8.63	2.856	1.863
WS50M_RANGE	0.74	10.26	4.578	1.850

Table 3. Drought severity classes as defined by US Drought Monitor

Drought Score	**Description**	**Label in dataset**
ND	No Drought	0
D0	Abnormally Dry	1
D1	Moderate Drought	2
D2	Severe Drought	3
D3	Extreme Drought	4
D4	Exceptional Drought	5

1.2 Explaining clustering through decision trees

The *k*-means [24] is an unsupervised machine learning algorithm able to discriminate samples (data points) into different clusters according to their similarities in the data space. The algorithm is always dependent on the value of k which is defined before the execution of the model. The *k*-means assigns each sample to the cluster with the nearest mean (cluster's centroid). At the end, the data space is split into Voronoi triangles.

For a set of observations $(x_1, x_2, \dots, x_n)$ where each observation represents a *d*-dimensional real vector, the algorithms aim to split the observations into *k* clusters ($\leq n$) denoted as $S = \{S_1, S_2, \dots, S_k\}$, in order to minimize the intra-cluster variance using sum of squares as defined in equation 1:

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x \in S_i} \| x - \mu_i \|^2 = \underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} |S_i| \qquad (1),$$

where μ_i is the mean (cluster's centroid) of points in S_i and calculated with the equation (2):

$$\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} x \quad (2),$$

where $|S_i|$ is the size of S_i.

After the execution of clustering and the evaluation of the agreement with the ground-truth classes, a decision tree is trained on all the samples. Decision trees are non-parametric supervised learning methods for classification and regression tasks. A tree is generated to predict the values of an output variable by exploring potential decision rules from the features. The trees are in general interpretable and explainable models. Thus, it eases the process of interpretation and explanation of clustering algorithms like *k*-means.

Following the training of a decision tree model, SHAP-based explanations are applied. SHAP is an XAI methodology based on the cooperative game theory and consequently uses the Shapley values. Each feature is considered as a "player" and the Shapley value for each player deputize its contribution to the output value. Shapley values are calculated by equation 3:

$$\varphi_i(v) = \sum_{S \subseteq N\{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} \big(v(S \cup \{i\}) - v(S)\big) \qquad (3),$$

where *N* is the number of players (features), and *v* is the function which subsets the players and represents a characteristic function. The *v* means that if *S* is a set of players, the $v(S)$ is the total worth of coalition *S* and describes the expected sum of payoffs the coalition can obtain by cooperation. The *n* is the total number of players and *i* is the current player.

EXPERIMENTAL SETUP

The *k*-means algorithm is employed for clustering. A trial-and-error strategy is used to determine the number of maximum iterations until the algorithm reaches convergence. During the tuning, the algorithm is tested with 200,300, 400 and 500 maximum iterations. According to the applied tuning, maximum iterations are set to 500. The number of clusters is set to 6, according to the number of drought scores defined by U.S. Drought Monitor. The decision tree which is used for the explainability assessment is tuned as follows: gini criterion is selected, the best split is selected at each node, and the nodes are expanded until all leaves are pure. In contrast to the typical ML methodologies, in this study a train/validation/test split is not necessary. The objective of training a decision tree is to use its explainability to calculate the SHAP values and proceed with clusters' exploration. Therefore, all the available data are used.

3.1 Evaluation Metrics

In this study, four distinct metrics are used to evaluate the accuracy of the proposed methodology in the subsection 2.2. The agreement between the ground-truth classes and the predicted clusters of the *k*-means algorithm is defined as accuracy. Furthermore, Silhouette score is calculated to understand the distinction between the different clusters. The first metric is the Rand Index (RI) (or Rand Score) as defined in equation 4 which quantifies the similarity between two data clusterings.

$$RI = \frac{TP + TN}{TP + TN + FP + FN} \quad (4),$$

where TP is the number of True Positive pairs (both points belong in the same cluster in predicted cluster and ground truth class), TN is the number of True Negative pairs (both points belong in different cluster in predicted cluster and ground truth class), FP is the number of False Positive pairs (both points belong in the same cluster in the predicted cluster and in different clusters in ground truth classification) and FN is the number of False Negative pairs (both points are in different predicted cluster but in the same ground truth class).

The second metric is the adjusted rand index (ARI), defined in equation 5, which calculates the similarity between predicted clusters and ground truth classes for all the pairs in a certain dataset and counts the number of pairs that are correctly clustered or not.

$$ARI = \frac{RI - \mathbb{E}[RI]}{\max(RI) - \mathbb{E}[RI]} \quad (5),$$

where RI is as defined in equation 4, $\mathbb{E}[RI]$ is the expected RI and max (RI) is the maximum RI.

Fowlkes-Mallows (FM) index computes the similarity between the two clusters by comparing the pairs of points and is defined in equation 6**.**

$$FM = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}} \quad (6),$$

where TP is the True Positive, FP is the False Positive, and FN is the False Negative.

The final metric is the homogeneity score which asses if a cluster contains only data points from the ground truth classes. Homogeneity score is defined in equation 7**.**

$$Homogeneity = 1 - \frac{H(C|K)}{H(C)} \quad (7),$$

where $H(C|K)$ is the conditional entropy of the class distribution in the given cluster and $H(C)$ is the entropy of the class distribution.

All the above metrics are taking a range from 0 to 1, where 0 indicates no similarity between clusters and ground truth classes and 1 indicates absolute similarity. Another metric used to determine the similarity of different data points in each cluster with the rest data points in the cluster, is the Silhouette which is defined by the equations 8 and 9**.**

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)} \quad (8),$$

where $s(i)$ is the silhouette score of a single data point i, $a(i)$ is the average distance from the point to the other points in the same cluster, $b(i)$ is the minimum average distance from the point to points of a different cluster.

$$Silhouette = \frac{1}{n}\sum_{i=1}^{n} s(i) \quad (9),$$

where Silhouette is the overall silhouette score of the clustering analysis and n is the total number of data points.

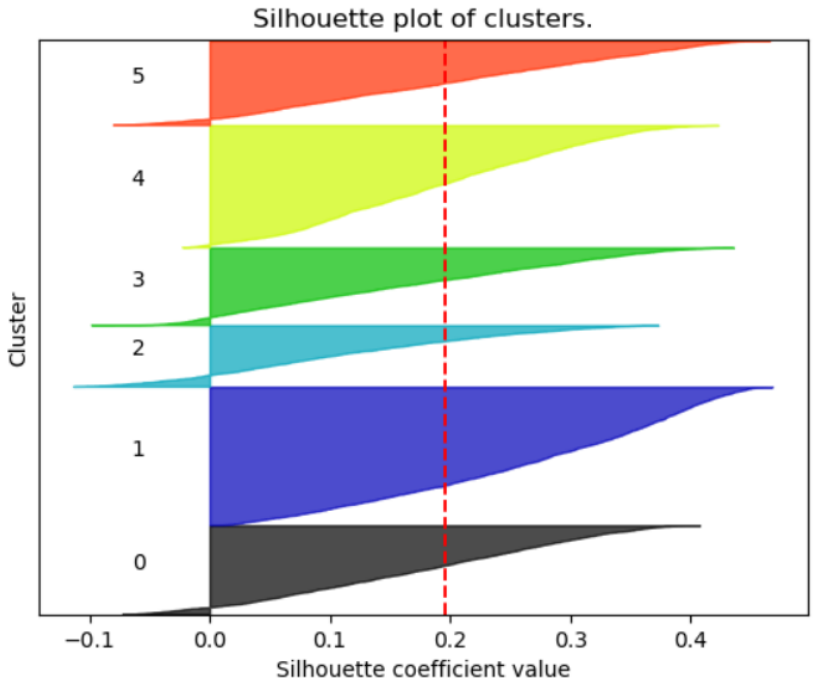

Figure 1. Silhouette of clusters. X-axis represents the silhouette coefficient value while the y-axis represents each cluster. The red dashed line shows the overall silhouette score.

EXPERIMENTAL RESULTS

4.1 Model's performance

The metrics defined in equations 4-7 and 9 are used to evaluate the agreement between predicted clusters and ground truth classes. Table 1 gives the experimental results of the proposed model. According to the evaluation metrics, it can be observed that the clustering achieves a full agreement of the predicted clusters with the ground truth classes, i.e., RI = 1.0, ARI = 1.0, FM = 1.0 and Homogeneity = 1.0. On the other hand, Silhouette analysis achieves an average score of 0.19, which means that there are samples assigned to a specific cluster that are similar with samples in a different cluster. Figure 1 shows the silhouette analysis' results of the predicted clusters. All the clusters exceed the average silhouette score (showed with red dashed line) while at the same time all, but cluster "1", there are data points who achieved a negative silhouette value which means that those data points are assigned to wrong cluster.

In fact, a typical strategy before clustering is to determine the optimal number of clusters. There are different techniques like the nbclust [25], which provides the optimal number of clusters based on an exhaustive analysis of 30 different evaluation metrics. However, this strategy is not applicable for this study since the number of different drought states for each county are defined by US drought monitor system.

Table 1. Experimental results of the k-means algorithm

RI	ARI	FM	Homogeneity	Silhouette
1.0	1.0	1.0	1.0	0.19

1.3 Explainability

At each step of building a decision tree, the algorithm chooses the most informative feature to split the data. In this specific case this separation is measured by gini impurity [26]. At the end, the decision tree can estimate the contribution of each feature to decrease impurity. Figure 2 shows the score of each feature in terms of importance in impurity reduce.

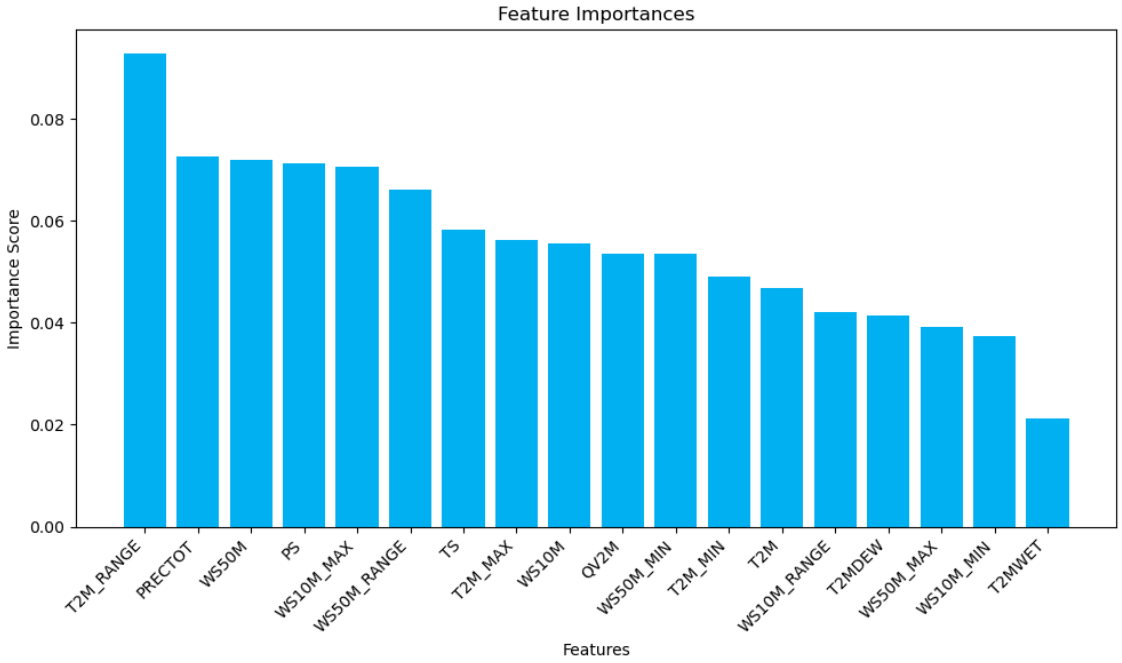

Figure 2. Feature importance according to the Decision Tree. From left to right, is the most important to the least important feature. X-axis define each feature and y-axis represents their importance score.

According to the feature importance, as estimated by the supervised algorithm, the range of air temperature (i.e., T2M_RANGE) at 2 meters and the precipitation (i.e., PRECTOT) are the most important features in discriminating the samples into the clusters. Conversely, the wind speed at 10 meters (i.e., WS10M_MIN) and the temperature of the wet bulb (i.e., T2MWET) are the least important features. From the importance scores, it can be observed that none of the features is highly important, but the difference of the most and least important in comparison with the rest of the features is noticeable.

A negative Shapley (or SHAP) value for a specific feature means that is pushing the model towards the examined class whereas a positive a Shapley value means the opposite. The SHAP summary plots for the classes "No Drought" and "Exceptional Drought" are presented in Figures 3 and 4, respectively. Both plots present the contribution of each feature (y-axis) in model's decisions according to the Shapley value (x-axis). The majority of high surface pressure (i.e., PS) values have a positive Shapley value, which means that aids the model to classify the samples as "No Drought". In contrast, high surface pressure values have a negative impact to the model for class "Exceptional Drought". Furthermore, higher maximum temperature (i.e., T2M_MAX) values tend to push the model away from "No Drought" class whereas lower values of the same feature push the model towards this class. From Figures 4 and 5 it can be that higher recorded T2M_MAX are not leading the model to the class "Exceptional Drought" class. Low precipitation (i.e., PRECTOT) values show that are not sufficiently helping the model, while the majority have a negative Shapley value. In contrast, most of the low PRECTOT values have a positive contribution towards the "Exceptional Drought" class. Lower surface temperature (TS) values are leading the model to classify data points as "No Drought", while they have a negative impact to the "Exceptional Drought" class. It is clear that higher temperature range (i.e., T2M_RANGE) drives the model towards the "Exceptional Drought" class, but it is questionable for the "No Drought" class.

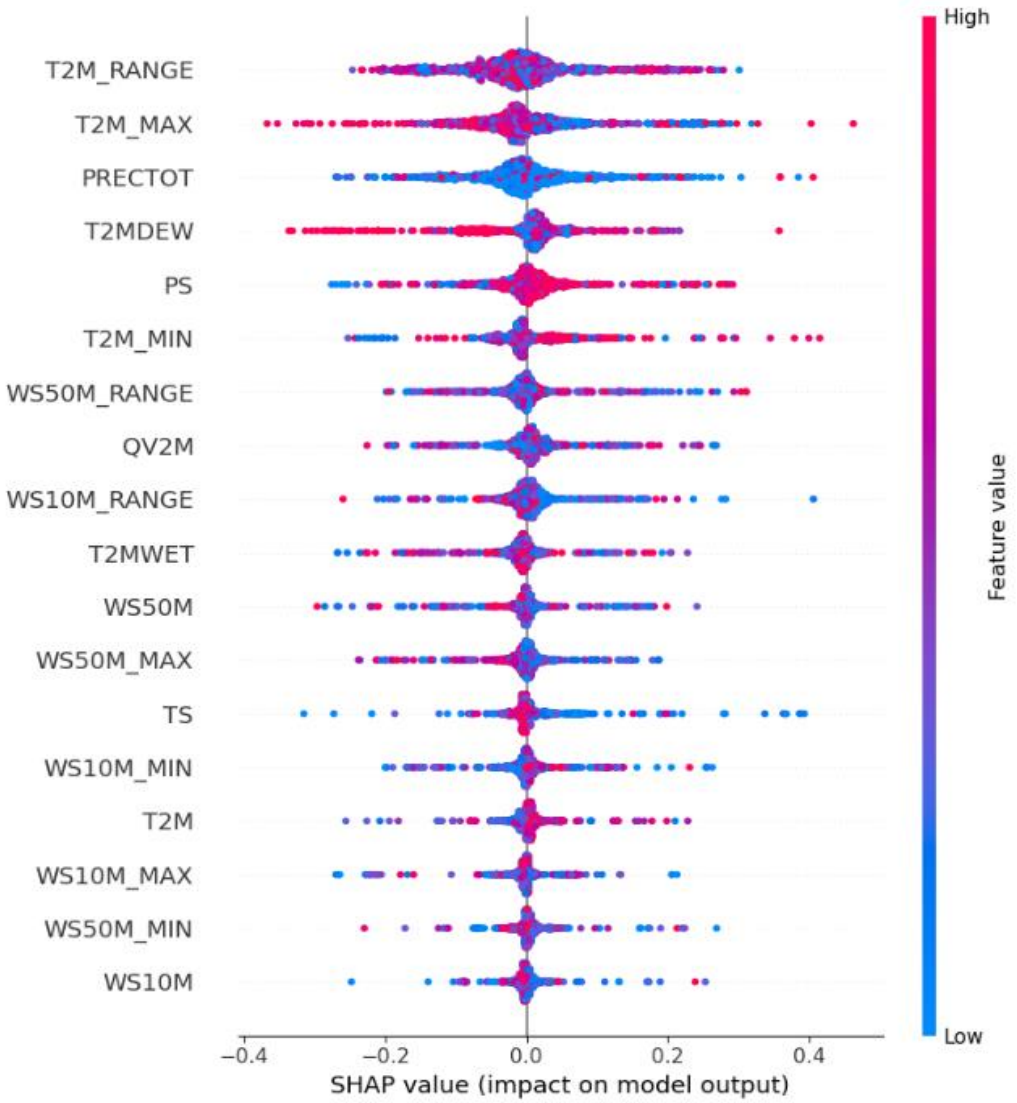

Figure 3. SHAP summary plot for samples classified as “No drought”. X-axis defines the calculated Shapley value for each data point and y-axis defines the different features involved in the dataset. The colour differentiation distinct the feature value for each point (low values are blue and high values are red).

Figure 4. SHAP summary plot for samples classified as "Exceptional drought". X-axis defines the calculated Shapley value for each data point and y-axis defines the different features involved in the dataset. The colour differentiation distinct the feature value for each point (low values are blue and high values are red).

CONCLUSION

In this work, a novel XAI strategy based on decision trees is used on the unsupervised learning algorithm *k*-means. It is observed that the *k*-means clustering algorithm with 500 iterations is in a full agreement with the ground truth classes of the dataset, while at the same time the silhouette evaluation metric suggests that the clusters are not distinct to each other. According to the SHAP-based XAI applied to the decision tree trained on clusters, the most effective predictive features are the maximum air temperatures at 2 meters, the range of air temperatures at 2 meters, the precipitation, the surface temperature, and the surface pressure. Based on XAI techniques, it is easier to understand the importance of these features in classifying extreme classes like "No Drought" and "Exceptional Drought". A further examination on this work is going to be conducted with the incorporation of data related to soil variables.

ACKNOWLEDGEMENTS

This work was partially supported by the European Union's HORIZON Research and Innovation Programme under grant agreement No 101120657, project ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI), the AI-OBSERVER project funded from the European Union's Horizon Europe Framework Programme HORIZON WIDERA-2021-ACCESS-03 (Twinning) under the Grant Agreement No 101079468, and the 'EXCELSIOR': ERATOSTHENES: Excellence Research Centre for Earth Surveillance and Space-Based Monitoring of the Environment H2020 Widespread Teaming project (www.excelsior2020.eu). The 'EXCELSIOR' project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 857510, from the Government of the Republic of Cyprus through the Directorate General for the European Programmes, Coordination and Development and the Cyprus University of Technology.

REFERENCES

[1] J. Qiu, Z. Shen, and H. Xie, 'Drought impacts on hydrology and water quality under climate change', *Science of The Total Environment*, vol. 858, p. 159854, Feb. 2023, doi: 10.1016/j.scitotenv.2022.159854.

[2] K. Furtak and A. Wolińska, 'The impact of extreme weather events as a consequence of climate change on the soil moisture and on the quality of the soil environment and agriculture – A review', *CATENA*, vol. 231, p. 107378, Oct. 2023, doi: 10.1016/j.catena.2023.107378.

[3] F. Di Nunno and F. Granata, 'Spatio-temporal analysis of drought in Southern Italy: a combined clustering-forecasting approach based on SPEI index and artificial intelligence algorithms', *Stoch Environ Res Risk Assess*, vol. 37, no. 6, pp. 2349–2375, Jun. 2023, doi: 10.1007/s00477-023-02390-8.

[4] M. Achite, O. M. Katipoglu, S. Şenocak, N. Elshaboury, O. Bazrafshan, and H. Y. Dalkılıç, 'Modeling of meteorological, agricultural, and hydrological droughts in semi-arid environments with various machine learning and discrete wavelet transform', *Theor Appl Climatol*, vol. 154, no. 1–2, pp. 413–451, Oct. 2023, doi: 10.1007/s00704-023-04564-4.

[5] S. M. E. Azimi, S. J. Sadatinejad, A. Malekian, and M. H. Jahangir, 'Application of artificial intelligence hybrid models for meteorological drought prediction', *Nat Hazards*, Dec. 2022, doi: 10.1007/s11069-022-05779-w.

[6] A. Danandeh Mehr, A. Rikhtehgar Ghiasi, Z. M. Yaseen, A. U. Sorman, and L. Abualigah, 'A novel intelligent deep learning predictive model for meteorological drought forecasting', *J Ambient Intell Human Comput*, vol. 14, no. 8, pp. 10441–10455, Aug. 2023, doi: 10.1007/s12652-022-03701-7.

[7] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo, 'Benchmarking and survey of explanation methods for black box models', *Data Min Knowl Disc*, vol. 37, no. 5, pp. 1719–1778, Sep. 2023, doi: 10.1007/s10618-023-00933-9.

[8] V. Hassija *et al.*, 'Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence', *Cogn Comput*, vol. 16, no. 1, pp. 45–74, Jan. 2024, doi: 10.1007/s12559-023-10179-8.

[9] A. Dikshit and B. Pradhan, 'Explainable AI in drought forecasting', *Machine Learning with Applications*, vol. 6, p. 100192, Dec. 2021, doi: 10.1016/j.mlwa.2021.100192.

[10] A. Dikshit and B. Pradhan, 'Interpretable and explainable AI (XAI) model for spatial drought prediction', *Science of The Total Environment*, vol. 801, p. 149797, Dec. 2021, doi: 10.1016/j.scitotenv.2021.149797.

[11] B. Zhang, F. K. Abu Salem, M. J. Hayes, K. H. Smith, T. Tadesse, and B. D. Wardlow, 'Explainable machine learning for the prediction and assessment of complex drought impacts', *Science of The Total Environment*, vol. 898, p. 165509, Nov. 2023, doi: 10.1016/j.scitotenv.2023.165509.

[12] J. Mardian, C. Champagne, B. Bonsal, and A. Berg, 'A Machine Learning Framework for Predicting and Understanding the Canadian Drought Monitor', *Water Resources Research*, vol. 59, no. 8, p. e2022WR033847, Aug. 2023, doi: 10.1029/2022WR033847.

[13] J. Mardian, C. Champagne, B. Bonsal, and A. Berg, 'Understanding the Drivers of Drought Onset and Intensification in the Canadian Prairies: Insights from Explainable Artificial Intelligence (XAI)', *Journal of Hydrometeorology*, vol. 24, no. 11, pp. 2035–2055, Nov. 2023, doi: 10.1175/JHM-D-23-0036.1.

[14] Y. Kuwayama, A. Thompson, R. Bernknopf, B. Zaitchik, and P. Vail, 'Estimating the Impact of Drought on Agriculture Using the U.S. Drought Monitor', *American J Agri Economics*, vol. 101, no. 1, pp. 193–210, Jan. 2019, doi: 10.1093/ajae/aay037.

[15] M. Zink *et al.*, ‘The German drought monitor’, *Environ. Res. Lett.*, vol. 11, no. 7, p. 074002, Jul. 2016, doi: 10.1088/1748-9326/11/7/074002.
[16] M. D. Svoboda, M. J. Hayes, D. A. Wilhite, and T. Tadesse, ‘Recent Advances in Drought Monitoring’.
[17] K. Themistocleous *et al.*, ‘Cyprus enters the space arena with "Excelsior " H2020 Teaming project and the Eratosthenes Centre of Excellence: Why Cyprus? Why Excelsior? What are the needs and opportunities?’ Accessed: Jun. 05, 2024. [Online]. Available: https://meetingorganizer.copernicus.org/EGU2020/EGU2020-21801.html
[18] M. Eliades *et al.*, ‘Earth Observation in the EMMENA Region: Scoping Review of Current Applications and Knowledge Gaps’, *Remote Sensing*, vol. 15, no. 17, p. 4202, Aug. 2023, doi: 10.3390/rs15174202.
[19] M. Louhichi, R. Nesmaoui, M. Mbarek, and M. Lazaar, ‘Shapley Values for Explaining the Black Box Nature of Machine Learning Model Clustering’, *Procedia Computer Science*, vol. 220, pp. 806–811, 2023, doi: 10.1016/j.procs.2023.03.107.
[20] C. Kingsford and S. L. Salzberg, ‘What are decision trees?’, *Nat Biotechnol*, vol. 26, no. 9, pp. 1011–1013, Sep. 2008, doi: 10.1038/nbt0908-1011.
[21] E. Laber, L. Murtinho, and F. Oliveira, ‘Shallow decision trees for explainable k -means clustering’, *Pattern Recognition*, vol. 137, p. 109239, May 2023, doi: 10.1016/j.patcog.2022.109239.
[22] M. Fleissner, L. C. Vankadara, and D. Ghoshdastidar, ‘Explaining Kernel Clustering via Decision Trees’. arXiv, Feb. 15, 2024. Accessed: Jun. 05, 2024. [Online]. Available: http://arxiv.org/abs/2402.09881
[23] S. Dasgupta, N. Frost, M. Moshkovitz, and C. Rashtchian, ‘Explainable k-Means and k-Medians Clustering’. arXiv, Sep. 21, 2020. Accessed: Jun. 05, 2024. [Online]. Available: http://arxiv.org/abs/2002.12538
[24] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, ‘K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data’, *Information Sciences*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.
[25] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs, ‘**NbClust** : An *R* Package for Determining the Relevant Number of Clusters in a Data Set’, *J. Stat. Soft.*, vol. 61, no. 6, 2014, doi: 10.18637/jss.v061.i06.
[26] Y. Yuan, L. Wu, and X. Zhang, ‘Gini-Impurity Index Analysis’, *IEEE Trans.Inform.Forensic Secur.*, vol. 16, pp. 3154–3169, 2021, doi: 10.1109/TIFS.2021.3076932.