Acadlore Transactions on AI and Machine Learning
https://www.acadlore.com/journals/ATAIML

Impact of Data Preprocessing Techniques on the Performance of Machine Learning Models for Drought Prediction

Serap Erçel*, Sinem Akyol

Software Engineering Department, Engineering Faculty, Firat University, 23279 Elazig, Turkey

* Correspondence: Serap Erçel (ercelserap@gmail.com)

Received: 12-22-2024 **Revised:** 02-10-2025 **Accepted:** 02-15-2025

Citation: S. Erçel and S. Akyol, "Impact of data preprocessing techniques on the performance of machine learning models for drought prediction," *Acadlore Trans. Mach. Learn.*, vol. 4, no. 1, pp. 14–24, 2025. https://doi.org/10.56578/ataiml040102.

Abstract: Drought, a complex natural phenomenon with profound global impacts, including the depletion of water resources, reduced agricultural productivity, and ecological disruption, has become a critical challenge in the context of climate change. Effective drought prediction models are essential for mitigating these adverse effects. This study investigates the contribution of various data preprocessing steps—specifically class imbalance handling and dimensionality reduction techniques—to the performance of machine learning models for drought prediction. Synthetic Minority Over-sampling Technique (SMOTE) and near miss sampling methods were employed to address class imbalances within the dataset. Additionally, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) were applied for dimensionality reduction, aiming to improve computational efficiency while retaining essential features. Decision tree algorithms were trained on the preprocessed data to assess the impact of these preprocessing techniques on model accuracy, precision, recall, and F1-score. The results indicate that the SMOTE-based sampling approach significantly enhances the overall performance of the drought prediction model, particularly in terms of accuracy and robustness. Furthermore, the combination of SMOTE, PCA, and LDA demonstrates a substantial improvement in model reliability and generalizability. These findings underscore the critical importance of carefully selecting and applying appropriate data preprocessing techniques to address class imbalances and reduce feature space, thus optimizing the performance of machine learning models in drought prediction. This study highlights the potential of preprocessing strategies in improving the predictive capabilities of models, providing valuable insights for future research in climate-related prediction tasks.

Keywords: Machine learning; Drought prediction; SMOTE; Near miss; PCA; LDA; Decision trees

1 Introduction

Climate change is increasingly affecting water resources, agriculture, and ecosystems worldwide. Therefore, drought prediction holds a strategic position in critical areas such as combating adverse effects and resource management. Due to the significant impacts of drought on ecology, environment, hydrology, and agriculture, drought monitoring and early warning system development have become important research areas [1]. Risk assessments of agricultural drought disasters are necessary to quantitatively understand drought and guide prevention and relief efforts [2]. Additionally, numerous studies have been conducted in the field of drought risk assessment, reflecting increasing interest in this critical area [3]. Drought mitigation is an integral part of quantitative drought risk assessment, and its effects serve as an intensified expression of drought mitigation [4].

Drought risk assessments are crucial for informing proactive and appropriate drought risk management decisions, as they contribute to improving risk knowledge and identifying priority management areas [5]. The impact of drought on agriculture is significant, with climatic factors dominating intra-annual variations, emphasizing the importance of integrating drought indices with agricultural survey data for accurate assessments [6].

The ability to predict the impacts of droughts in advance is critical for minimizing crop losses in agriculture, alleviating pressure on water resources, and preserving the sustainability of ecosystems. Droughts have far-reaching consequences on water ecosystems and human consumption, underscoring the need to understand their ecological impacts and human responses [7]. Agriculture, including the livestock sector, is particularly vulnerable to the effects of drought, leading to significant losses in agricultural production [8].

This study aims to achieve more precise and accurate results in drought predictions using weather and soil data. Models created by combining meteorological data, soil moisture content, climate variables, and other factors were used to predict future drought risks. The development of models contributes to creating more effective strategies in combating drought. Increasing model accuracy and ensuring prediction continuity support communities in being prepared for drought risks and taking measures for a sustainable future. Traditional drought prediction methods are carried out through systems built on meteorological data and climate models. However, these methods have various limitations. The first striking limitation is that the lack of data and low-resolution data sources significantly affect the accuracy of the predictions. The situation that reduces the accuracy of drought prediction in large geographical areas is that traditional methods generally work with data obtained from a limited number of meteorological stations.

In addition, traditional approaches are mostly based on univariate analyses. For example, they focus on a single meteorological parameter such as precipitation, temperature or evaporation. Despite this, drought is a very complex natural event that occurs as a result of the interaction of more than one variable. Therefore, traditional models cannot fully reflect the multidimensional nature of drought. Another limitation of traditional prediction methods is that they are based on static models. These models make future predictions based on past data, but their ability to adapt to changing climate and environmental conditions over time is limited. Dynamic processes such as climate change further reduce the accuracy and reliability of traditional methods.

In addition, these methods require a high degree of human intervention. Forecasting processes largely rely on the knowledge and experience of experts. This increases the risk that human errors and biases may affect forecast results. Machine learning methods offer significant potential in overcoming these limitations. Thanks to its capacity to process large data sets, machine learning can use a wide variety of data sources, including satellite images, meteorological station data, and soil moisture sensors. In addition, machine learning systems have an automatic and scalable structure. In addition, by recognizing patterns and anomalies in past data, it can better predict future drought events.

As a result, machine learning methods have the potential to overcome the limitations of traditional drought forecasting methods and provide more dynamic, accurate, and comprehensive forecasts. It is envisaged that these technologies can be used as an important tool in drought management and planning. The obtained results can assist in making strategic decisions in reducing crop losses in the agricultural sector, effectively managing water resources, and protecting ecosystems. This study aims to prepare communities for climate change-related drought risks on a scientific basis.

The introduction defines the research's purpose and scope, explaining the significance of drought and its context in Turkey. The second section shows the existing literature on the topic. The methodology, including data preprocessing, exploratory analysis, and model training, is detailed in the third section. The fourth section presents findings from modeling with various balancing techniques and offers recommendations. Finally, the conclusion summarizes the results, compares them with the literature, and highlights the study's contributions.

2 Literature Review

Drought is a climate event where water resources are insufficient or precipitation amounts decrease from normal for an extended period. Three types of drought are commonly defined in the literature: meteorological, agricultural, and hydrological drought [9]. There is also a concept that emphasizes the socio-economic effects of drought, defined in some sources as socio-economic drought, which occurs when certain economic products cannot be supplied and demanded due to drought [10]. Meteorological drought occurs when the amount of precipitation falling to the earth's surface drops below average during a certain period. When meteorological drought duration extends, agricultural drought occurs, defined as water scarcity to the degree that plants cannot meet their physiological water needs during the growing period due to decreased soil moisture. With further increases in precipitation deficit, after soil moisture deficiency, decreases observed in surface and groundwater resources availability indicate hydrological drought. Socio-economic drought is related to the above three types of drought. It represents the water resources system's inadequacy in meeting water demand or its effects on human health [9]. All drought conditions begin and develop with meteorological drought.

When examining the frequency, duration, and severity of drought on a global scale between 1951 and 2010, Spinoni et al. [11] found that North America and Australia were affected by drought between 1951 and 1970, the equatorial region between 1971 and 1990, and the Mediterranean basin between 1991 and 2010. When evaluating studies for the Eastern Mediterranean Basin and Turkey, especially after the 1970s, Ceylan et al. [12] revealed a decrease in total precipitation amounts and an increase in the effect of dry conditions and even the presence of areas prone to desertification. In addition, it was stated that areas prone to desertification in Turkey progress from the east of the Konya Plain towards the Eastern Mediterranean section.

Turkey is located in a semi-arid region in climate classification. Therefore, monitoring drought and taking precautions in advance are important. Drought, which occurs in large areas in different parts of the world, is a meteorological natural disaster with lasting effects on society as it brings hunger, famine, and unemployment. If drought can be monitored, its effects can be minimized. The spatial and temporal variation of drought can be

tracked [13].

To understand the impact of drought in Turkey, it is crucial to consider the country's sensitivity to global warming and the increasing frequency and severity of drought events. Turkey is particularly vulnerable to drought as it is located in a semi-arid region [14]. Research has shown that the effects of drought in Turkey can vary both spatially and temporally in terms of economic, environmental, and social aspects [15]. Akbaş [16] suggested using the Palmer Drought Severity Index (PDSI) as a useful tool for monitoring and mitigating drought effects in Turkey. Furthermore, An et al. [17] proposed that drought in Turkey may lead to migrations as people are forced to seek alternative living conditions.

The Eastern Mediterranean basin, including Turkey, faces various problems due to drought events caused by low precipitation [18]. The impact of drought extends to water resources, and studies show increasing trends in hydrological drought in certain regions [19]. As a result of global warming, it has been emphasized that drought in Turkey has the potential to lead to ecological deterioration, desertification, and weakening of water resources [20]. It is clear that drought in Turkey is a multifaceted problem affecting various sectors such as agriculture, forestry, and water management [21]. The need for effective water management strategies, particularly in the face of declining water resources and potential ecological deterioration, further emphasizes the country's sensitivity to drought [22].

3 Methodology

3.1 Data and Preprocessing

Meteorological indicators include measurements of weather and soil data. Wind speed, temperature, humidity, pressure, precipitation, and other meteorological values are included in the dataset. Data from the US Drought Monitor provides an analysis of drought manually created by experts using a wide range of data (droughts using weather & soil data). The purpose of this dataset is to help investigate whether droughts can be predicted using only meteorological data and potentially lead to the generalization of US predictions to other parts of the world. The dataset was created for a classification model containing non-drought conditions and six different drought levels. The classes in the dataset and the number of samples they contain are shown in Table 1.

Table 1. Classes in the dataset and their numbers of examples

Class	Size
0	1652230
1	466944
2	295331
3	196802
4	106265
5	39224

The dataset underwent comprehensive cleaning and preprocessing steps. Inconsistencies, duplicates, and anomalies were detected and corrected. The drought score was the target variable and was based on weekly measurements. Missing values were considered meaningful but were processed with an appropriate strategy to not affect the analysis's accuracy. Date and score information was reformatted. These steps both increase the reliability of the analysis and enable more accurate interpretation of results.

3.2 Exploratory Data Analysis (EDA)

EDA is a crucial step in understanding the characteristics of a dataset and revealing patterns that can guide further analysis. The primary purpose of EDA is to generate hypotheses and gain insights about the data rather than reaching final conclusions [23].

3.2.1 Distribution of continuous variables

By plotting histograms of continuous variables, distributions in the dataset were visually examined. This step provides insight into the central tendencies and distribution characteristics of variables.

Bivariate analysis aims to obtain deeper insights by examining relationships and possible correlations between two variables. At this stage, methods such as scatter plots, correlation coefficients, and regression analysis were used to explain relationships between variables. Bivariate analysis is an important step in understanding interactions between variables in the dataset. It is shown in Figure 1.

The distribution of continuous variables is shown in Figure 2.

3.2.2 Correlation between independent variables

Dependent and independent variables were determined for analysis. While the dependent variable usually forms the focus of the analysis, understanding relationships between independent variables helps with the modeling process. It is shown in Figure 3.

3.3 Data Preparation Before Modeling

Data standardization was performed to bring the scales of different features to the same level. In this process, features were transformed into standard normal distribution using Z-score. Recursive Feature Elimination (RFE) and random forest algorithms were applied for feature selection, and 15 out of 21 features that contributed most to model performance were selected.

To address the class imbalance problem, minority class samples were increased using SMOTE, while majority class samples were reduced using the near miss method. These operations enabled the model to better learn the minority class. Additionally, dimensionality reduction methods such as PCA and LDA were applied to the dataset, reducing the number of features to five for both methods. These processes increased the model's generalization ability while reducing computational costs.

3.3.1 RFE

This method was initially used to build a model using all features and iteratively remove the least contributing features while updating the model. The goal is to achieve a more efficient model by eliminating features that do not significantly impact performance.

3.3.2 Random forest algorithm

Random forest is one of the ensemble learning methods and creates a more robust and stable model by combining multiple decision trees. This algorithm creates random trees from different subsamples in the dataset, allowing each tree to make a separate prediction. Then, the final decision is made by combining these predictions. Random Forest can be used effectively for both classification and regression problems. The robustness and accuracy of the model are especially prominent in high-dimensional and complex datasets.

(a) Identifying key meteorological factors of the model

The random forest algorithm uses feature importance scores to measure the contribution of each feature to the model performance. These scores are calculated based on how much each feature increases the prediction accuracy of the model. The model ranks the meteorological factors according to these importance scores and determines the most important ones. For example:

• Soil moisture can be a determining factor in the occurrence of drought, since moisture levels directly affect plant growth and agricultural production.

• Rainfall is a primary indicator of drought duration and severity.

• Air temperature affects the evaporation rate, changing the soil and plant water balance.

• Wind speed can increase drought severity by accelerating moisture loss from the surface.

Random forest's feature importance ranking shows which of these factors play a critical role in drought occurrence.

(b) Effects on drought occurrence and severity

The model provides a clear visualization of how key factors selected to predict drought classes affect drought classes. For example:

• Low values for attributes such as soil moisture and precipitation can be a factor that increases drought risk.

• If air temperatures are high and precipitation is low, drought severity is likely to increase.

• Wind speed can increase evaporation, which can cause already limited water resources to deplete more quickly.

3.3.3 Addressing class imbalance

In the field of drought prediction, data imbalance stands out as a critical problem that negatively affects the performance of machine learning models. In imbalanced datasets, the number of examples belonging to the minority class is numerically different from the majority class, which causes models to fail to learn this minority class well enough. This problem is a challenge to solve, especially in critical real-world applications such as drought prediction.

SMOTE and near miss are preferred to be used in drought prediction models. Their effects on prediction accuracy, error rate and model generalization capacity have been observed in depth. While SMOTE increases the importance of the minority class and keeps the prediction precision for this class high, near miss enables the model to generalize in a more balanced way.

Comparing these two methods with traditional data balancing methods can make the scientific contributions of the study more obvious. Traditional methods, such as simple resampling, usually mechanically remove the imbalance between minority and majority classes. However, these methods may limit the learning capacity of the model and lead to the overfitting problem. These processes aim to tackle the class imbalance problem by enabling the model to better learn from minority class examples.

(a) Oversampling with SMOTE

SMOTE is a method specifically designed to increase the number of minority class instances. By generating synthetic examples for the minority class, it reduces the imbalance between classes, ensuring a more balanced class distribution. This method creates new examples by interpolating between existing minority class examples, thus enabling the model to learn the minority class better.

SMOTE can be used to:

• Identify the k-nearest neighbors for each minority class sample.

• Choose one of these neighbors randomly.

• Generate a new sample along the line segment joining the minority class sample and the chosen neighbor. This technique ensures that the minority class is represented with diverse samples, reducing the overfitting risk.

(b) Downsampling with near miss

Near miss aims to reduce class imbalance by decreasing the number of majority class instances. It selects majority class samples that are closest to the minority class samples, thereby balancing the class distribution. Near miss provides an effective strategy to increase the generalization ability of the model, especially due to the dominance of the majority class.

3.3.4 Dimensionality reduction processes

These processes aim to make the dataset more effective by highlighting the features of the minority class and minimizing unnecessary information loss. Additionally, they enhance the model's generalization capability and reduce the risk of overfitting.

(a) PCA

PCA is a dimensionality reduction technique that transforms the feature matrix into principal components. It helps reduce the number of features while preserving the variance within the dataset. PCA is an unsupervised technique used to reduce dimensionality while retaining maximum variance in the data.

• PCA computes the covariance matrix of the features.

• Eigenvalues and eigenvectors are extracted from this matrix.

• The eigenvectors corresponding to the largest eigenvalues form the principal components.

• The number of components to retain is chosen based on the cumulative explained variance (e.g., $>90\%$).

(b) LDA

LDA is a dimensionality reduction technique that transforms the feature matrix to enhance the separation between classes. Its primary goal is to highlight differences between classes more effectively. It computes:

• Within-class scatter (variation within each class).

• Between-class scatter (variation between the class means).

• The linear discriminants are the directions that maximize the ratio of between-class scatter to within-class scatter.

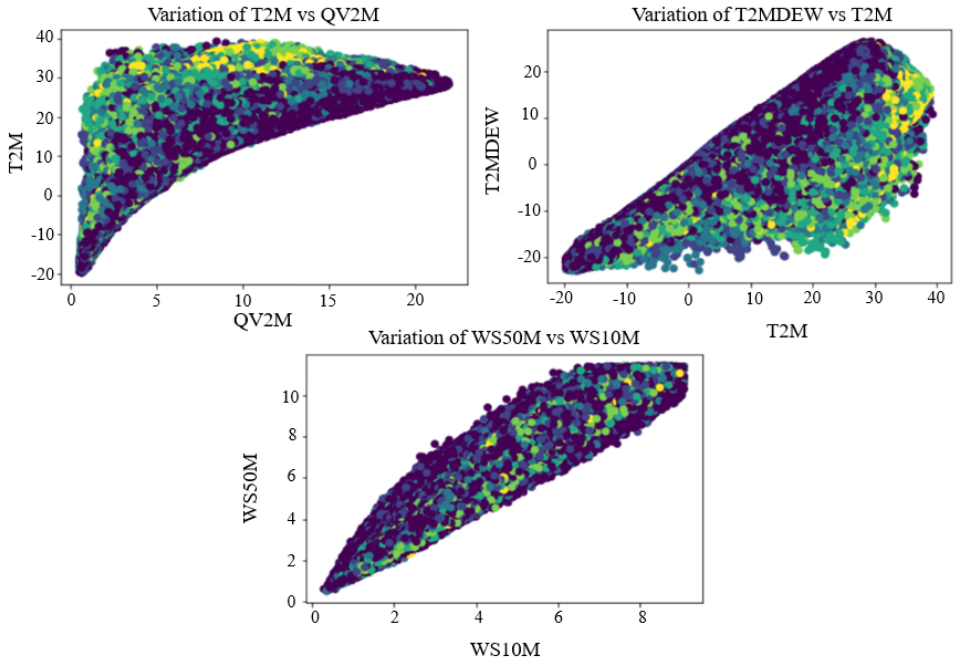
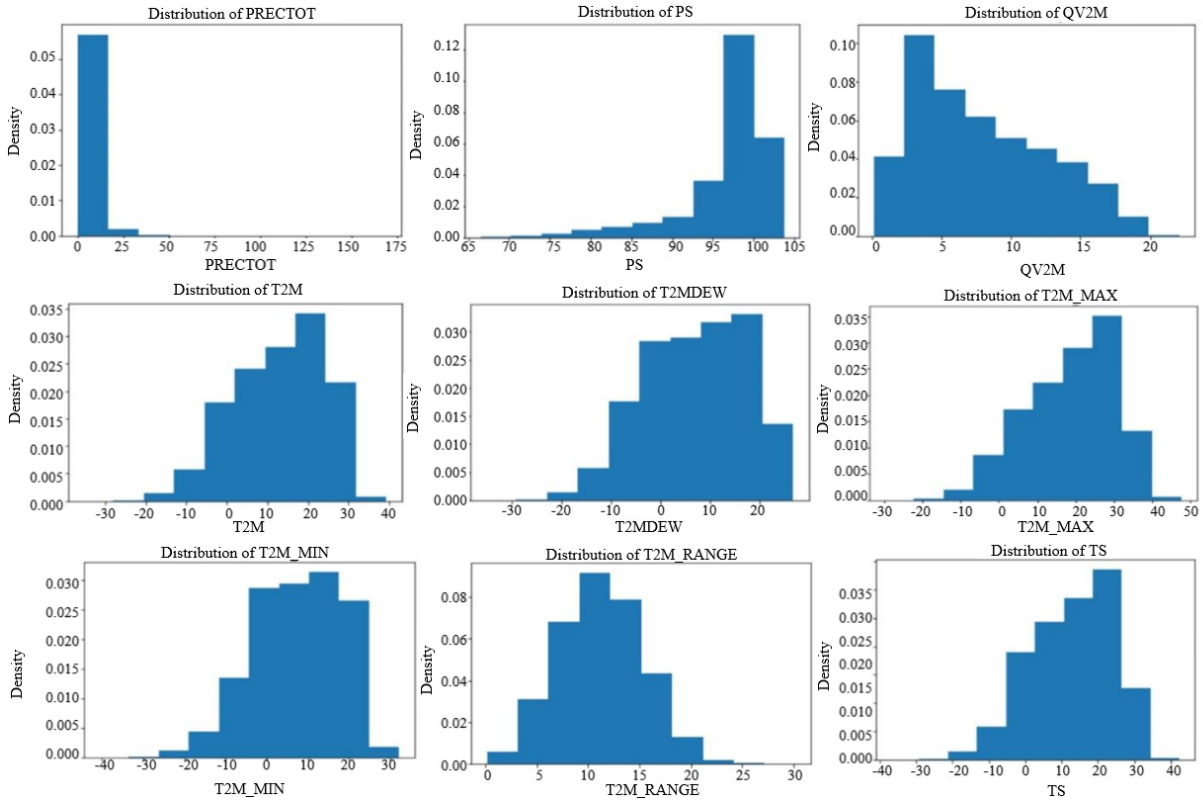
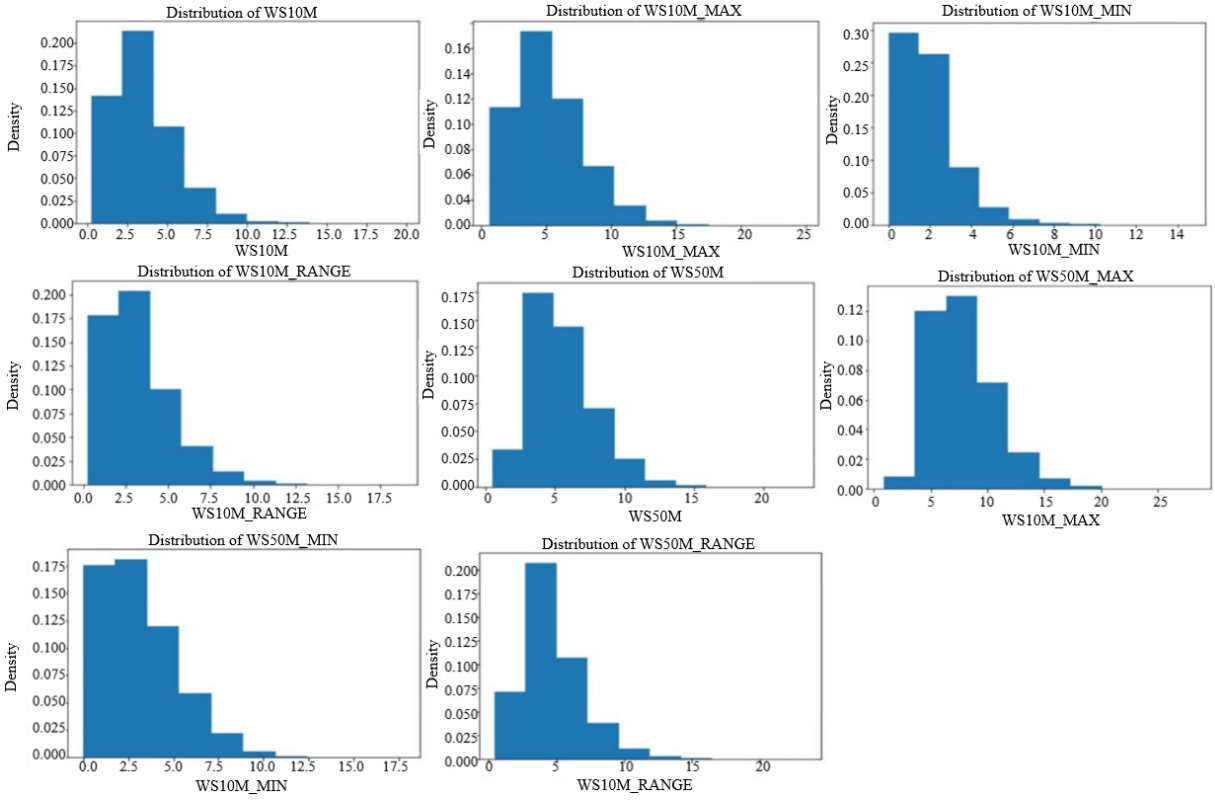


Figure 1. Bivariate analysis



(a)



(b)

Figure 2. Distribution of continuous variables

	PRECTOT	PS	QV2M	T2M	T2MDEW	T2MWET	T2M_MAX	T2M_MIN	T2M_RANGE	TS	WS10M	WS10M_MAX	WS10M_MIN	WS10M_RANGE	WS50M	WS50M_MAX	WS50M_MIN	WS50M_RANGE
PRECTOT	1.000000	0.068775	0.245081	0.093258	0.231035	0.230975	0.026773	0.144929	-0.304171	0.089598	0.049730	0.060981	0.023346	0.065755	0.069057	0.079508	0.057816	0.047477
PS	0.068775	1.000000	0.282412	0.164160	0.341234	0.341252	0.111979	0.208285	-0.225935	0.163830	-0.080747	-0.135905	0.022932	-0.198332	-0.043315	-0.091821	0.036238	-0.154479
QV2M	0.245081	0.282412	1.000000	0.870242	0.959385	0.960434	0.804338	0.906144	-0.071547	0.862559	-0.225449	-0.256452	-0.108789	-0.269203	-0.205971	-0.249961	-0.081554	-0.246203
T2M	0.093258	0.164160	0.870242	1.000000	0.913530	0.914218	0.983356	0.981629	0.244357	0.997515	-0.207874	-0.220192	-0.125407	-0.209030	-0.193196	-0.206444	-0.112579	-0.159589
T2MDEW	0.231035	0.341234	0.959385	0.913530	1.000000	0.999970	0.854716	0.939934	-0.015643	0.905184	-0.238299	-0.268686	-0.115920	-0.280702	-0.204238	-0.245323	-0.082416	-0.239335
T2MWET	0.230975	0.341252	0.960434	0.914218	0.999970	1.000000	0.855401	0.940629	-0.015500	0.905911	-0.237971	-0.268292	-0.115882	-0.280199	-0.204143	-0.245147	-0.082497	-0.239029
T2M_MAX	0.026773	0.111979	0.804338	0.983356	0.854716	0.855401	1.000000	0.937762	0.407534	0.980101	-0.216764	-0.221671	-0.141911	-0.199614	-0.195727	-0.196236	-0.133234	-0.126331
T2M_MIN	0.144929	0.208285	0.906144	0.981629	0.939934	0.940629	0.937762	1.000000	0.065037	0.979134	-0.206382	-0.225829	-0.112878	-0.225256	-0.197991	-0.225744	-0.096593	-0.200157
T2M_RANGE	-0.304171	-0.225935	-0.071547	0.244357	-0.015643	-0.015500	0.407534	0.065037	1.000000	0.241564	-0.080163	-0.043127	-0.110952	0.018746	-0.041778	0.029737	-0.128844	0.163320
TS	0.089598	0.163830	0.862559	0.997515	0.905184	0.905911	0.980101	0.979134	0.241564	1.000000	-0.189823	-0.202713	-0.110273	-0.196015	-0.180665	-0.193347	-0.102367	-0.152434
WS10M	0.049730	-0.080747	-0.225449	-0.207874	-0.238299	-0.237971	-0.216764	-0.206382	-0.080163	-0.189823	1.000000	0.952217	0.833340	0.702896	0.966275	0.908750	0.795424	0.412412
WS10M_MAX	0.060981	-0.135905	-0.256452	-0.220192	-0.268686	-0.268292	-0.221671	-0.225829	-0.043127	-0.202713	0.952217	1.000000	0.690087	0.866026	0.910717	0.946710	0.660428	0.592380
WS10M_MIN	0.023346	0.022932	-0.108789	-0.125407	-0.115920	-0.115882	-0.141911	-0.112878	-0.110952	-0.110273	0.833340	0.690087	1.000000	0.235775	0.839187	0.666629	0.943983	-0.046209
WS10M_RANGE	0.065755	-0.198332	-0.269203	-0.209030	-0.280702	-0.280199	-0.199614	-0.225256	0.018746	-0.196015	0.702896	0.866026	0.235775	1.000000	0.643131	0.810677	0.234645	0.827364
WS50M	0.069057	-0.043315	-0.205971	-0.193196	-0.204238	-0.204143	-0.195727	-0.197991	-0.041778	-0.180665	0.966275	0.910717	0.839187	0.643131	1.000000	0.917883	0.847885	0.373539
WS50M_MAX	0.079508	-0.091821	-0.249961	-0.206444	-0.245323	-0.245147	-0.196236	-0.225744	0.029737	-0.193347	0.908750	0.946710	0.666629	0.810677	0.917883	1.000000	0.646726	0.674944
WS50M_MIN	0.057816	0.036238	-0.081554	-0.112579	-0.082416	-0.082497	-0.133234	-0.096593	-0.128844	-0.102367	0.795424	0.660428	0.943983	0.234645	0.847885	0.646726	1.000000	-0.126283
WS50M_RANGE	0.047477	-0.154479	-0.246203	-0.159589	-0.239335	-0.239029	-0.126331	-0.200157	0.163320	-0.152434	0.412412	0.592380	-0.046209	0.827364	0.373539	0.674944	-0.126283	1.000000

Figure 3. Correlation between independent variables for feature selection

3.4 Model Training and Evaluation

The decision tree algorithm was used for drought prediction. After splitting the dataset into training and test sets, data were standardized during the training process, and the selected 15 features were used. The decision tree algorithm was trained on class-balanced datasets, and hyperparameter optimization was performed.

To evaluate the robustness of the model, cross-validation was employed. Cross-validation ensures consistent performance across different subsets of the data, reducing the risk of overfitting. Additionally, the model's scalability and applicability to different regions were explored by considering variations in meteorological data. These steps, combined with cross-validation and scalability evaluations, demonstrate a methodical approach to preparing the data for training and improving the classifier's performance in diverse scenarios.

Model performance was evaluated using metrics such as accuracy, precision, recall, F1-score, and Cohen Kappa. The results were compared to analyze the effects of addressing class imbalance and dimensionality reduction operations on model performance.

4 Results

The main findings of this study are below. Oversampling with SMOTE significantly improved the model's performance. Additionally, a positive effect on the model's overall performance was observed with the use of dimensionality reduction methods of LDA and PCA. Hyperparameter optimization also contributed to the model's potential to exhibit higher performance. However, it was recommended to test other alternative algorithms without being limited to classification algorithms.

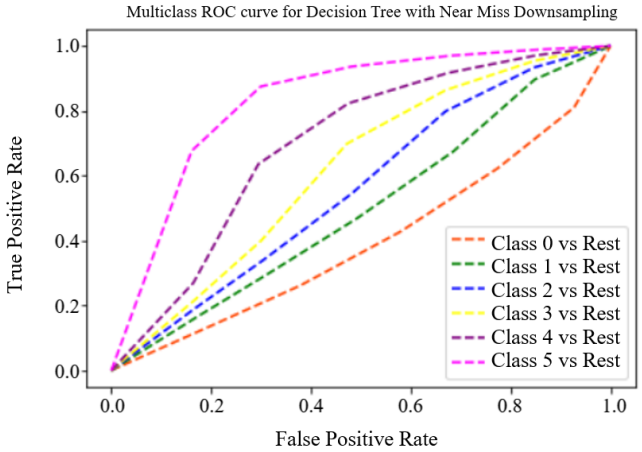


Figure 4. Multiclass ROC curve for decision tree with near miss downsampling

Figures 4-9 compare the performance of the decision tree algorithm with various preprocessing techniques. The dataset was balanced using the near miss and SMOTE methods, and then dimensionality reduction was applied using PCA and LDA. Each figure visualizes the multiclass Receiver Operating Characteristic (ROC) curve of the model. For example, Figure 4 and Figure 5 show the cases where near miss and SMOTE are used alone, while Figure 6

and Figure 7 consider the cases where dimensionality is reduced using PCA. Figure 8 and Figure 9 show the class separation enhancing effect of LDA. These figures provide an opportunity to evaluate the strengths and weaknesses of the proposed approaches by visualizing the class separation success of the model and the effect of each method on the ROC curve.

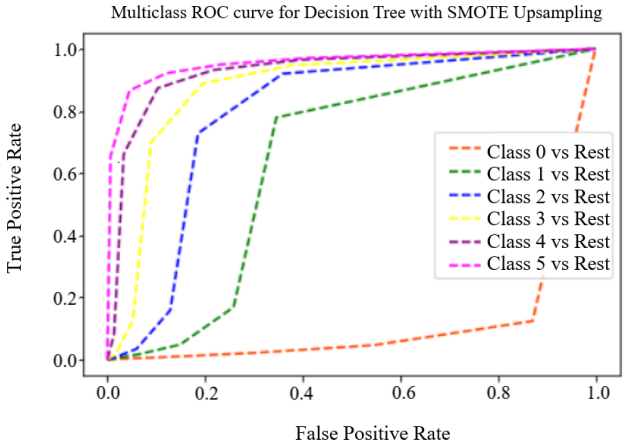


Figure 5. Multiclass ROC curve for decision tree with SMOTE upsampling

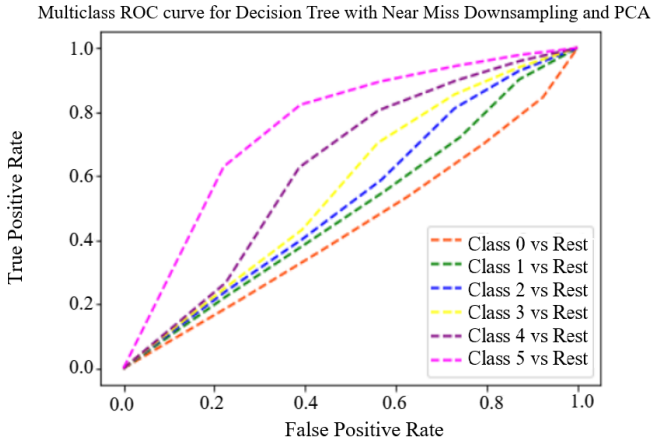


Figure 6. Multiclass ROC curve for decision tree with near miss downsampling and PCA

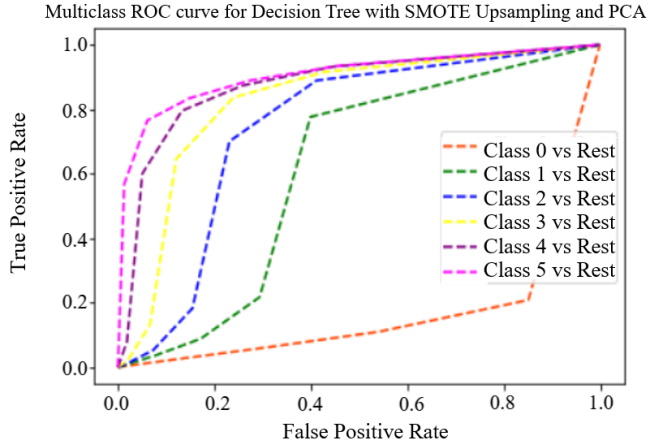


Figure 7. Multiclass ROC curve for decision tree with SMOTE upsampling and PCA

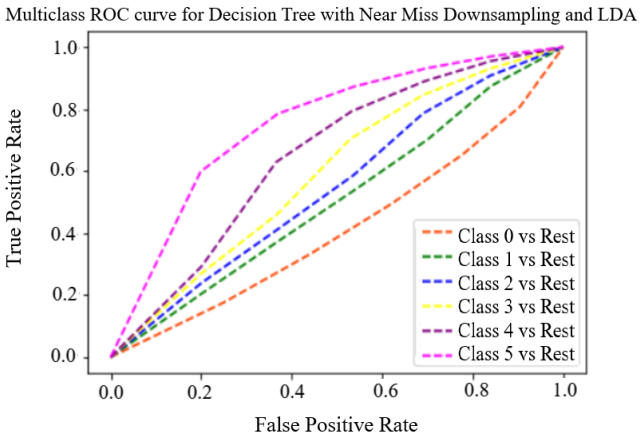


Figure 8. Multiclass ROC curve for decision tree with near miss downsampling and LDA

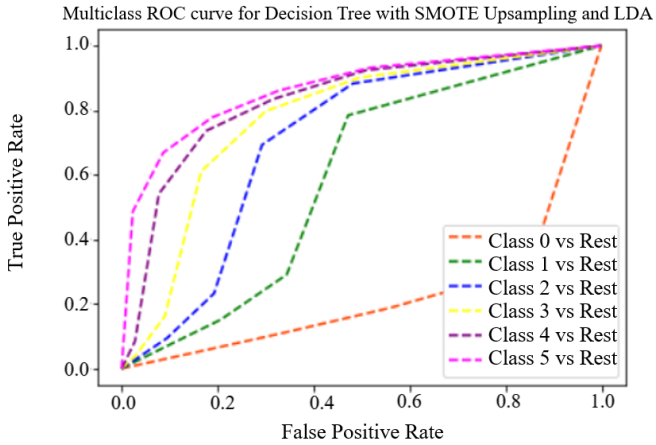


Figure 9. Multiclass ROC curve for decision tree with SMOTE upsampling and LDA

Table 2. Classes in the dataset and their numbers of examples

Balancing Method	Accuracy	Precision	Recall	Fl-Score	Cohen Kappa Score
Near miss downsampling	0.2263	0.5431	0.2263	0.2646	0.0794
SMOTE upsampling	0.7642	0.7728	0.7642	0.7680	0.6073
Near miss downsampling and PCA	0.1884	0.5195	0.1884	0.2234	0.0589
SMOTE upsampling and PCA	0.6912	0.7219	0.6912	0.7033	0.5046
Near miss downsampling and LDA	0.2057	0.5153	0.2057	0.2511	0.0586
SMOTE upsampling and LDA	0.6031	0.6747	0.6031	0.6286	0.3951
Decision tree (without applying any imbalance process)	0.7628	0.7617	0.7628	0.7622	0.5957

The obtained results provide a foundation for future studies and developments while strengthening the model's current state. In this context, potential improvements that can be made to increase the model's complexity and bring it to a broader perspective should be emphasized. The findings of the models according to the performance metrics with the applied methods are shown in Table 2.

5 Conclusions

This research aims to examine the application of various machine learning techniques for drought detection and strategies for dealing with class imbalance. This study focuses on the decision tree algorithm using meteorological data containing a broad feature set. The insights from this study provide a solid scientific basis for improving decision-making in water resources management and agricultural strategies. By leveraging machine learning models that accurately identify and predict drought conditions, policymakers and stakeholders can implement more targeted and effective measures to reduce drought impacts.

In water resources management, models can enable actions such as predicting water shortages, optimizing reservoir operations, prioritizing water needs for key sectors, and implementing conservation strategies. Integrating real-time monitoring systems with historical and predicted data can also support dynamic adjustments in water allocation. In agriculture, machine learning insights can guide the selection of drought-tolerant crops and inform irrigation planning. Early warnings can help farmers prepare for droughts, allowing them to adjust planting plans or adopt alternative agricultural practices. Analysis results reveal that the decision tree model performs well when there is no class imbalance. However, class imbalance is a frequently encountered problem in real-world applications. In this context, methods correcting class imbalance need to be applied.

Upsampling performed using SMOTE was successful in addressing the imbalance between classes and improved the model's overall performance. Particularly notable improvements were seen in accuracy, precision, recall, and F1-score values. On the other hand, downsampling performed using the near miss method negatively affected the model's performance. This situation indicates that class representation was insufficient due to excessive data loss. Additionally, the effect of the feature subset obtained using dimensionality reduction techniques of PCA and LDA on model performance was evaluated. However, results obtained using these methods were generally more effective when used after class imbalance operations.

Guided by these findings, the recommendations for future studies are as follows:

• Advanced machine learning algorithms such as ensemble methods (e.g., random forest, gradient boosting) could be used to improve model accuracy and robustness.

• Alternative class imbalance techniques beyond SMOTE, such as Adaptive Synthetic Sampling (ADASYN) or cost-sensitive learning, could be investigated to improve minority class detection.

• Climate projection data could be integrated to model long-term drought trends to map various climate change scenarios.

• Real-time forecast systems combining meteorological and satellite data could be implemented for dynamic drought monitoring and management.

In future research, different machine learning algorithms could be explored to enhance model diversity and compare their performance. In addition, alternative balancing methods beyond SMOTE as well as additional dimensionality reduction techniques could be investigated to further optimize model performance. Finally, the findings could be integrated into real-time prediction systems to develop practical field applications and test them across different climatic regions to achieve generalizable results.

In conclusion, this study incorporated class imbalance operations and dimensionality reduction techniques to improve the performance of machine learning models for drought detection. The obtained results demonstrate that using SMOTE and dimensionality reduction techniques on the class-balanced dataset enables the decision tree model to perform reliable and effective drought detection. These findings provide an important foundation for future drought prediction studies and applications.

Data Availability

The data used to support the research findings are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] B. R. Nikam, S. P. Aggarwal, P. K. Thakur, V. Garg, S. Roy, A. Chouksey, P. R. Dhote, and P. Chauhan, "Assessment of early season agricultural drought using remote sensing," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, pp. 1691–1695, 2020. https://doi.org/10.5194/isprs-archives-XLIII-B3-2020-1691-2020

[2] H. F. Xu, K. X. Xu, and Y. J. Yang, "Risk assessment model of agricultural drought disaster based on grey matter-element analysis theory," *Nat. Hazards*, vol. 107, pp. 2693–2707, 2021. https://doi.org/10.1007/s11069-021-04681-1

[3] L. Hao, X. Y. Zhang, and S. D. Liu, "Risk assessment to China's agricultural drought disaster in county unit," *Nat. Hazards*, vol. 61, pp. 785–801, 2012. https://doi.org/10.1007/s11069-011-0066-4

[4] Y. L. Zhang, J. L. Jin, S. M. Jiang, S. W. Ning, Y. L. Zhou, and Z. Y. Wu, "Quantitative assessment model for the effects of drought mitigation on regional agriculture based on an expectation index of drought mitigation effects," *Water*, vol. 11, no. 3, p. 464, 2019. https://doi.org/10.3390/w11030464

[5] I. Aitkenhead, Y. Kuleshov, J. Bhardwaj, Z. W. Chua, C. Sun, and S. Choy, "Validating a tailored drought risk assessment methodology: Drought risk assessment in local Papua New Guinea regions," *Nat. Hazards Earth Syst. Sci.*, vol. 23, pp. 553–586, 2023. https://doi.org/10.5194/nhess-23-553-2023

[6] H. Yan, S. Q. Wang, J. B. Wang, H. Q. Lu, A. H. Guo, Z. C. Zhu, R. B. Myneni, and H. H. Shugart, "Assessing spatiotemporal variation of drought in China and its impact on agriculture during 1982–2011 by using PDSI indices and agriculture drought survey data," *J. Geophys. Res. Atmos.*, vol. 121, no. 5, pp. 2283–2298, 2016. https://doi.org/10.1002/2015JD024285

[7] N. R. Bond, P. S. Lake, and A. H. Arthington, "The impacts of drought on freshwater ecosystems: An Australian perspective," *Hydrobiologia*, vol. 600, pp. 3–16, 2008. https://doi.org/10.1007/s10750-008-9326-z

[8] V. A. Myeki and Y. T. Bahta, "Determinants of smallholder livestock farmers' household resilience to food insecurity in South Africa," *Climate*, vol. 9, no. 7, p. 117, 2021. https://doi.org/10.3390/cli9070117

[9] A. F. Van Loon, "Hydrological drought explained," *WIREs Water*, vol. 2, no. 4, pp. 359–392, 2015. https://doi.org/10.1002/wat2.1085

[10] G. P. Mengü, S. Anaç, and E. Özçakal, "Kuraklık yönetim stratejileri," *J. Agric. Fac. Ege Univ.*, vol. 48, no. 2, pp. 175–181, 2011.

[11] J. Spinoni, G. Naumann, H. Carra, P. Barbarosa, and J. Vogt, "World drought frequency, duration and severity for 1951-2010," *Int. J. Climatol.*, vol. 34, no. 8, pp. 2792–2804, 2014. https://doi.org/10.1002/joc.3875

[12] A. Ceylan, S. Akgündüz, Z. Dermirörs, A. Erkan, S. Çınar, and E. Özveren, "Determination of changes in desertification-prone areas in Turkey using the Aridity Index," in *the First National Drought and Desertification Symposium*, Konya, Turkey, 2009.

[13] S. Sırdaş and Z. Şen, "Meteorological drought modeling and application in Turkey," Ph.D. dissertation, Istanbul Technical University, Graduate School of Science, Engineering and Technology, Istanbul, Turkey, 2002.

[14] G. Aktürk and O. Yıldız, "Effects of rainfall deficiency on various hydrological systems in Çatalan Dam Basin," *Int. J. Eng. Res. Dev.*, vol. 10, no. 2, pp. 10–28, 2018. https://doi.org/10.29137/umagd.441389

[15] C. Ayva, A. Atalay Dutucu, and B. Ustaoğlu, "The impact of climate change on water resources and adaptation recommendations: The case of Kirazdere basin," *Firat Univ. J. Soc. Sci.*, vol. 33, no. 1, pp. 47–64, 2023. https://doi.org/10.18069/firatsbed.1131015

[16] A. Akbaş, "Important drought years over Turkey," *J. Geogr. Sci.*, vol. 12, no. 2, pp. 101–118, 2014. https://doi.org/10.1501/Cogbil_0000000155

[17] N. An, M. T. Turp, and L. Kurnaz, "The effect of environmental degradation due to climate change on migration decision: An overview," *Aegean Geogr. J.*, vol. 30, no. 2, pp. 383–403, 2021. https://doi.org/10.51800/ecd.932879

[18] K. Oğuz, M. A. Pekin, H. Gürkan, E. Oğuz, and M. Coşkun, "Determination of drought in the Eastern Mediterranean basin using ERA-Interim data," *Anadolu J. Agric. Sci.*, vol. 32, no. 2, pp. 229–236, 2017. https://doi.org/10.7161/omuanajas.321080

[19] O. M. Katipoğlu, S. N. Yeşilyurt, and H. Y. Dalkılıç, "Trend analysis of hydrological droughts in Yesilırmak Basin by Mann Kendall and Sen innovative trend analysis," *Gümüşhane Univ. J. Sci. Technol.*, vol. 12, no. 2, pp. 422–442, 2022. https://doi.org/10.17714/gumusfenbil.1026893

[20] M. Ekinci, S. Örs, M. Turan, and E. Yıldırım, "Effects of nitric oxide applications on tolerance of plants in abiotic stress conditions," *Yuzuncu Yıl Univ. J. Agric. Sci.*, vol. 28, no. 2, pp. 254–265, 2018. https://doi.org/10.29133/yyutbd.427960

[21] I. Koç, "The effect of global climate change on some climate parameters and climate types in Bolu," *J. Bartin Fac. For.*, vol. 23, no. 2, pp. 706–719, 2021. https://doi.org/10.24011/barofd.947981

[22] A. Özdemir, "Assessment of climate change effects on flow and sediment amount at basin scale: Yuvacık Dam Lake Basin," *J. Geol. Eng.*, vol. 45, no. 1, pp. 129–154, 2021. https://doi.org/10.24232/jmd.941528

[23] K. Abt, "Descriptive data analysis: A concept between confirmatory and exploratory data analysis," *Methods Inf. Med.*, vol. 26, no. 2, pp. 77–88, 1987. https://doi.org/10.1055/s-0038-1635488