information

MDPI

Article

Information Mining from Heterogeneous Data Sources: A Case Study on Drought Predictions

Getachew B. Demisse [1,*], Tsegaye Tadesse [1], Solomon Atnafu [2], Shawndra Hill [3], Brian D. Wardlow [1], Yared Bayissa [1] and Andualem Shiferaw [1]

1 National Drought Mitigation Center, School of Natural Resources, University of Nebraska-Lincoln, Hardin Hall, 3310 Holdrege Street, P.O. Box 830988, Lincoln, NE 68583-0988, USA; ttadesse2@unl.edu (T.T.); bwardlow2@unl.edu (B.D.W.); ybayissa2@unl.edu (Y.B.); ashiferaw2@unl.edu (A.S.)

2 Department of Computer Science, Addis Ababa University, P.O. Box 1176, Addis Ababa, Ethiopia; solomon.atnafu@aau.edu.et

3 The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA; shawndra@wharton.upenn.edu

* Correspondence: gdemisse2@unl.edu; Tel.: +1-402-601-4260

Received: 2 May 2017; Accepted: 28 June 2017; Published: 3 July 2017

Abstract: The objective of this study was to develop information mining methodology for drought modeling and predictions using historical records of climate, satellite, environmental, and oceanic data. The classification and regression tree (CART) approach was used for extracting drought episodes at different time-lag prediction intervals. Using the CART approach, a number of successful model trees were constructed, which can easily be interpreted and used by decision makers in their drought management decisions. The regression rules produced by CART were found to have correlation coefficients from 0.71–0.95 in rules-alone modeling. The accuracies of the models were found to be higher in the instance and rules model (0.77–0.96) compared to the rules-alone model. From the experimental analysis, it was concluded that different combinations of the nearest neighbor and committee models significantly increase the performances of CART drought models. For more robust results from the developed methodology, it is recommended that future research focus on selecting relevant attributes for slow-onset drought episode identification and prediction.

Keywords: CART; drought; information mining; instances; regression tree; rules

1. Introduction

Information mining is a sub-discipline of the information systems field that supports business intelligence tools to transform information into knowledge [1,2], with a focus on searching for interesting patterns and important regularities in large bodies of information [3]. Here, knowledge is a justified true belief, and is created and organized by the flow of information. Knowledge for solving human challenges can be produced from the collection of useful information, through interactions among people sharing their experiences and other information sources [4]. In this process, information is a flow of messages, while knowledge is created and organized by the flow of information, anchored on the commitment and beliefs of its holder [5].

The present work is concerned with explicit knowledge [6], specifically patterns observed in data that can be easily understood by humans and validated by test data with some degree of certainty [7]. The practice of finding useful patterns in data known by a variety of names, including data mining [7], knowledge extraction [8], information discovery [9], information harvesting [8,10], data archaeology [9], and data pattern processing [8,10,11]. In this research, information mining is defined as the mining of meaningful patterns from heterogeneous data sources for drought modeling and prediction.

Drought is defined as "the naturally occurring phenomenon that exists when precipitation has been significantly below normal recorded levels, causing serious hydrological imbalances that adversely affect land resource production systems" [12,13]. Drought is also defined as a prolonged abnormally dry period when there is insufficient water for users' normal needs, resulting in extensive damage to crops and loss of yields [14].

Drought is one of the major natural hazard challenges to human beings. It is a recurrent climatic phenomenon across the world and affects humanity in a number of ways, such as causing loss of life, crop failures, food shortages that may lead to famine in many regions, malnutrition, health issues and mass migration [15]. Drought also causes considerable damage to the natural environment and is regarded as a major cause of land degradation, aridity, and desertification [16].

Different modeling methods are used for drought modeling and prediction, such as regression analysis [17,18], time series analysis [19–23], probability models [24–30], artificial neural network models [31,32], hybrid models [33–35], long-lead drought forecasting [36–38], and data mining [39–42]. These different drought modeling approaches have their own advantages and disadvantages for slow-onset drought modeling [43], which needs multiple variables for the modeling experiments. Specifically, in managing heterogeneous data sources for modeling and predicting drought from these existing methods, Mishira et al. [43] recommended data mining for its robustness in data integration and pattern extractions. Data mining is also considered to be a powerful technology that helps with extracting predictive information from large databases [44,45].

Despite different modeling drought efforts [17–42], there is no clear methodology that can be used for practically and reliably identifying, modeling, and predicting drought under climatic change scenarios with uncertainty [46–50], where the modeling products can give actionable information to the decision makers. The goal of this work was to develop an effective methodology (specifically, a practical data and information mining approach) for drought modeling and predictions using historical records of climate, satellite, environmental, and oceanic data. This requires (1) identifying the appropriate regression tree information mining parameters for retrieving actionable information for improved accuracy and ease of use of the model outputs; (2) developing interpretable regression tree information mining techniques for predicting drought at monthly temporal resolution; and (3) evaluating the performance of the developed regression tree drought models. Materials and methods are presented in detail in Section 2. Section 3 discusses the major findings, and Section 4 presents the conclusions.

2. Materials and Methods

2.1. Experimental Data Sources

The experimental data used for this study were obtained for Ethiopia, located in East Africa. Ethiopia was selected as an experimental study area because drought was frequently reported in the past [51], and also because the information mining concepts can easily be exercised in this real world problem-solving scenario. Ethiopia occupies the interior of the Greater Horn of Africa, stretching between 3° and 14° N latitude and 33° and 48° E longitude, with a total area of 1.13 million km^2 [52].

For the experimental analysis, 11 attributes were used for modeling drought (Table 1). In Table 1, SDNDVI is Standardized Deviation of Normalized Difference Vegetation Index (unitless) [53,54]; DEM is digital elevation model (meter) [55]; WHC is soil water-holding capacity (unitless) [55]; ER is ecological region (ecosystems of Ethiopia), (which is a categorical data type with a value representation of 1 = desert and semi-desert scrubland, 2 = Acacia-Commiphora woodland and bush land proper, 3 = acacia wooded grassland of the rift valley, 4 = wooded grassland of the Western Gambela region, 5 = Combretum/Terminalia woodland and wooded grassland, 6 = dry evergreen Afromontane forest and grassland complex, 7 = moist evergreen Afromontane forest, 8 = transitional rain forest, 9 = ericaceous belt, 10 = afro-alpine vegetation, 12 = freshwater lakes and open water vegetation, 13 = freshwater marshes and swamps, floodplains and lake shore vegetation, 14 = salt lake open water vegetation, 15 = salt pans, saline/brackish and intermittent wetlands and salt-lake shore vegetation [56];

LC is land use and land cover (which is a categorical data type with a value representation of 14 = rainfed croplands, 20 = mosaic cropland, 30 = mosaic vegetation (grassland/shrubland/forest), 40 = closed to open broadleaved evergreen or semi-deciduous forest, 60 = open broadleaved deciduous forest/woodland, 110 = mosaic forest or shrubland grassland, 120 = mosaic grassland/forest or shrubland, 130 = closed to open (broadleaved or needle-leaved, evergreen or deciduous) shrubland, 140 = closed to open herbaceous vegetation (grassland, savannas or lichens/mosses), 150 = sparse vegetation, 180 = closed to open grassland or woody vegetation on regularly flooded or waterlogged soil with fresh, brackish or saline water, 190 = artificial surfaces and associated areas, 200 = bare areas, and 210 = water bodies) [57,58]; SPI_3month is the three-month Standard Precipitation Index (unitless); PDO is Pacific Decadal Oscillation (unitless) [59]; AMO is Atlantic Multidecadal Oscillation Index (unitless) [59,60]; NAO is North Atlantic Oscillation (unitless) [59,61,62]; PNA is Pacific North American Index (unitless) [59]; and MEI is Multivariate ENSO Index (unitless) [59,63]. Detailed domain explanations for these attributes are available in the references listed. Twenty-four years of data (1983–2006) were extracted from each attribute and used for developing the time lag prediction models and performance evaluations.

Table 1. An excerpt from the list of attributes and data used for modeling drought.

DEM	SDNDVI	LC	AWC	ER	SPI	PDO	AMO	NAO	PNA	MEI
598	−0.40	200	97.5	2	0.21	0.68	0.243	−0.07	−0.53	0.068
700	−0.40	110	56.42857	2	0.72	−1.47	−0.11	−0.82	−0.92	1.005
2850	0.00	14	123.5	6	0.48	0.68	0.243	−0.07	−0.53	0.068
524	0.00	30	175	2	0.86	2.36	−0.027	0.99	2.10	1.700
258	−0.10	30	157.7778	1	1.28	1.10	−0.10	0.56	−1.18	−0.221
1697	0.30	14	159.5	6	1.07	1.10	−0.10	0.56	−1.18	−0.221
572	−0.40	110	127.7778	2	0.33	0.18	−0.299	−0.42	−0.36	−0.152
1876	−0.70	30	158.3334	6	−1.39	0.89	−0.226	1.22	0.39	0.394
1106	−0.20	110	33.33334	2	−1.32	−0.44	0.015	−0.03	−1.19	−0.219
1213	0.00	130	91.5	2	1.38	0.46	−0.194	1.52	−1.36	0.807
1567	0.20	110	197.5	5	0.17	1.10	−0.100	0.56	−1.18	−0.221
422	−0.20	30	157.7778	2	−2.54	−1.3	0.222	1.12	0.43	−0.500
1182	0.10	110	159.5	2	−0.59	0.74	0.197	0.88	1.31	−1.187
1612	0.20	20	175	6	−0.17	0.18	−0.299	−0.42	−0.36	−0.152
1589	−0.60	20	91.5	5	−0.03	0.46	−0.194	1.52	−1.36	0.807
3159	0.50	110	170	9	1.32	1.10	−0.100	0.56	−1.18	−0.221
2835	0.10	20	65.55556	6	−0.41	1.26	−0.125	0.2	1.36	0.952
1385	0.20	20	144.5	5	−0.57	1.27	0.396	0.13	0.90	0.177
908	0.20	200	88	2	1.82	1.10	−0.100	0.56	−1.18	−0.221

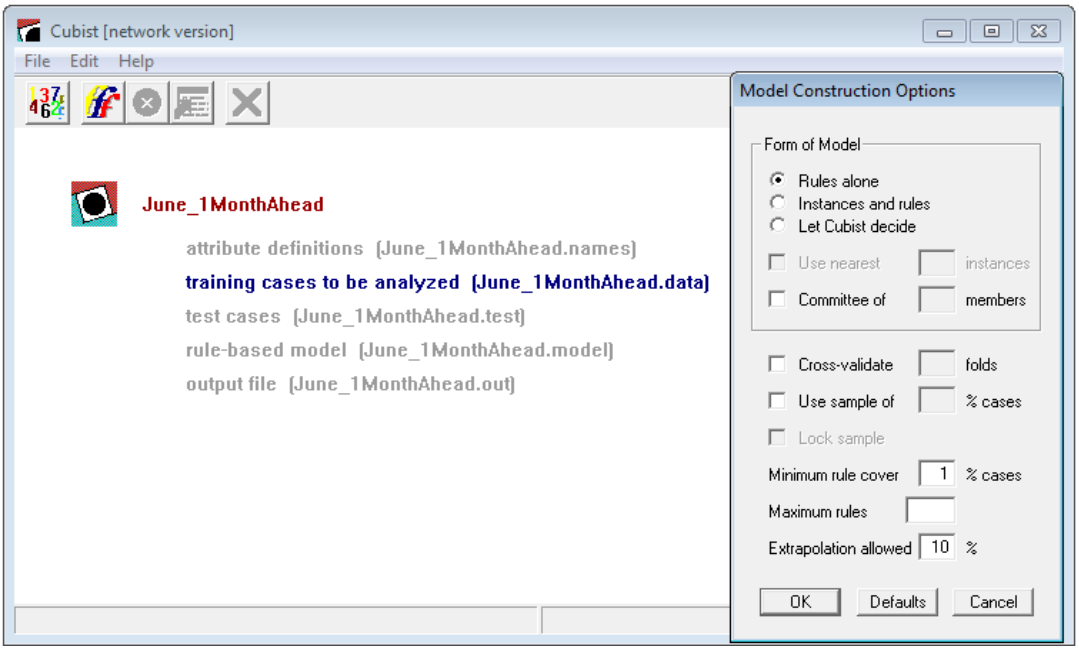
2.2. Classification and Regression Tree Modeling

In modeling drought, one of the major challenges is integrating the types of different parameters and subsequent data reduction. There are several approaches for data reduction and pattern extraction, such as principal component analysis and partial least squares [64], Bayesian method [65], neural networks and support vector machines [66], multiple adaptive regression splines [67], random forest [68], and classification and regression tree (CART) [69]. Past studies [70–73] used the CART model for identifying drought episodes at different time-lag prediction intervals. The CART model has also been found to be an effective method in studies in a variety of applications, such as soil properties analysis [74] and medical tablet formulation analysis [75] because of its high performance and ease of use of understanding by the domain experts.

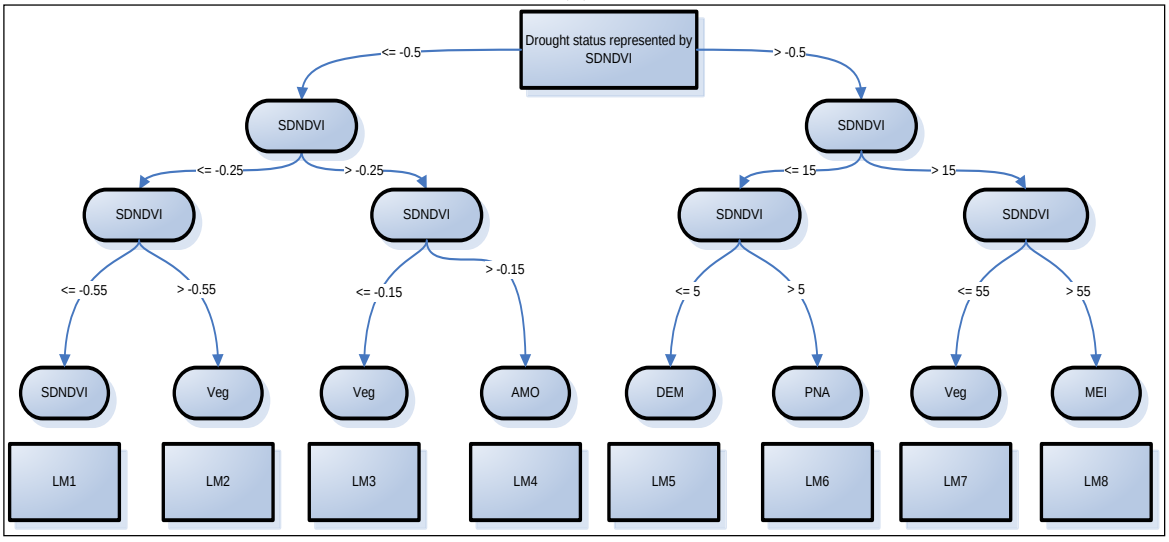
The CART modeling approach is implemented in RuleQuest [69]. Here, the term Rulequest is produced from two different terms: (1) *rule* is from the pieces of rules produced from the CART model, and (2) *quest* is an acronym for Quick, Unbiased, Efficient, Statistical Tree [76].

The CART in RuleQuest has the options to produce a model with rules alone, instance and rules (composite model), and *let cubist decide* selections (Figure 1a). The final model of RuleQuest [69]

in Cubist is a set of comprehensible rules, where each rule has an associated multivariate linear model. Figure 1b demonstrates the concept with a sample model tree for drought outlook predictions. Whenever a situation matches a rule's conditions (i.e., an individual rule set), the associated model is used to calculate the predicted value. This is in effect transforming regression into a classification problem. The model consists of a set of rules, and each rule consists of a linear model (see Figure 1b for linear model [LM] and Figure 2 for pieces of regression functions). Regression rules are analogous to piecewise linear functions [69]. Details on the algorithm of RuleQuest Cubist can be found in Quinlan [77]. In addition to rules, RuleQuest also has an instance-based modeling option (Figure 1a).



(**a**)



(**b**)

Figure 1. Graphical user interface for RuleQuest Cubist (**a**), and a structure of a model tree with a hypothetical tree and drought tree splits for the Standardized Deviation of Normalized Difference Vegetation Index (SDNDVI) (**b**).

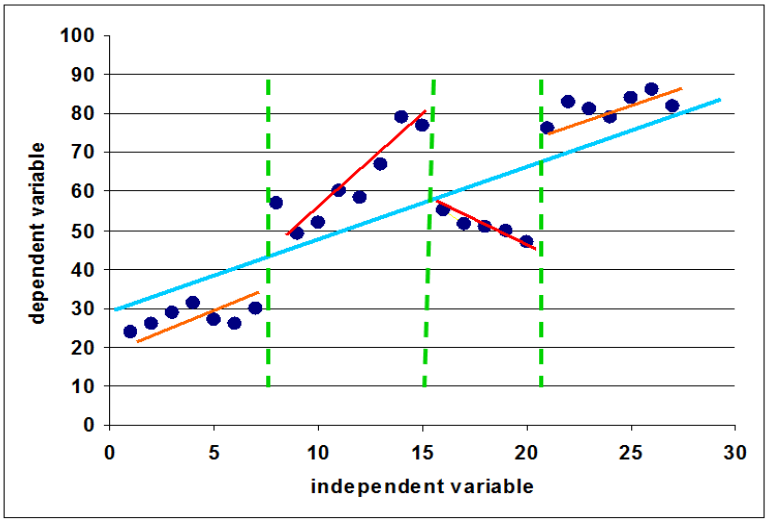


Figure 2. Conceptual demonstration of how a regression tree produces split rules fitting local models for final classification assignments. The red lines are produced for the clustered dots, and the green lines are classifying the different fits into different classes (the if conditions in RuleQuest [69] regression tree).

Instance modeling is based on the concept of an instance or a case as a basis for knowledge representation and reasoning [78]. Hullermeier [78] noted that a case or observation can be thought of as a single experience, such as a pattern (along with its classification) in a problem along with a solution in case-based reasoning. In instance-based modeling, an instance-based learning (IBL) algorithm learns by simply storing some of the observed examples [79]. As opposed to model-based machine learning methods, which induce a general model from the data and use that model for further reasoning, IBL algorithms store the data itself, and the algorithm processes the data until a prediction is actually requested [78–80]. In RuleQuest Cubist, it is possible to combine instances and rules (Figure 1a). In addition to these, there is also an option to use *let cubist decide* (Figure 1a).

In the *let cubist decide* option, Cubist derives from the training data a heuristic estimate of the accuracy of each type of model, and chooses the form that appears more accurate. In the experimental analysis, it was observed that the derivation of these estimates required considerable computation.

In addition to the instance model, Cubist CART has a committee modeling option (Figure 1a). The committee modeling option is based on the principle that each member of the committee should be as competent as possible (similar to establishing a committee of people), but the members should be complementary to one another. If the members are not complementary (i.e., if they always agree), then the committee is unnecessary and any one member is sufficient. If the members are complementary, then when one or a few members make an error, the probability is high that the remaining members can correct this error [69].

In the graphical user interface (GUI) (Figure 1a) implementations of the Cubist CART modeling tool, different modeling parameters can be selected for modeling drought: rules alone, instances and rules, *let cubist decide*, or a combination of these options with the use nearest instances; and use of committee members (Figure 1a). Since there are multiple options, there is a challenge to determine which combinations of the modeling parameters perform best. Therefore, in this research we used the original RuleQuest C code, which was produced under the GNU General Public License [69], to retrieve the pattern from the large dataset extracted from the 11 attributes. We also combined the different modeling parameters for selecting the best performing parameter combinations. This experimental analysis helped us compare the "rules alone", "instance and rules" (composite model), and *let cubist decide* modeling options from RuleQuest with the different parameter combinations to determine the best performances among the various drought models.

2.3. Accuracy Assessment

In data mining modeling, there are four approaches for estimating the accuracies of the models: holdout, random sub-sampling, k-fold cross validation, and bootstrap [7]. For our current drought data modeling experiment, we used the holdout approach, where the data split for training and testing were also experimented for 50/50%, 60/40%, 70/30%, 80/20%, 90/10%, and 99/1%. Mean average difference (MAD), relative error (RE), and correlation coefficient (CC) [69,81] were used for the comparisons of the performances of the models.

The MAD (Equation (1)) [69,82] and RE (Equation (2)) [69,82] were calculated as:

$$MAD = \frac{\sum_{i=1}^{d} |y_i - y'|}{d} \tag{1}$$

$$RE = \frac{\sum_{i=1}^{d} |y_i - y'|}{\sum_{i=1}^{d} \left| y_i - \bar{y} \right|} \tag{2}$$

where d is number of observation, y_i is model predicted values, y' is observed values, and $\bar{y}$ is mean value.

Lower MAD indicates that the predicted values are closer to the observed value. In the case of relative error (RE), if there is little improvement on the mean, the environmental variables have little predictive capacity and the relative error is close to 1. Generally, the smaller the relative error, the more useful the model [83].

CC (Equation (3)) [69,81] measures the agreement between measured and predicted samples, or how close the model predictions fall along a 45-degree line from the origin with the measured data (or a slope of exactly 1).

$$CC = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2} \tag{3}$$

where $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$, $s_y^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$, $s_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$, $\bar{x}$ and $\bar{y}$ are sample means for populations X (measured) and Y (predicted), x_i and y_i are paired with values from populations X and Y. The value of the index is scaled between -1 and 1, with a value of 1 representing complete agreement between all paired values.

In our current analysis, the performances of our models were also assessed in terms of a number of key indicators, such as root mean square error (RMSE) and coefficient of determination (r^2). The RMSE gives an estimate of the standard deviation of the errors. A lower RMSE is associated with greater predictive ability. Henderson et al. [83] indicated that RMSE values cannot be compared between different properties because they depend critically on the scale used. Therefore, care was taken in our experimental drought model assessment to prevent such mismatching.

The RMSE was calculated using Equation (4) [84].

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2} \tag{4}$$

where N = number of prediction/observation pairs, f is forecast, and o is observation.

In order to validate the predictability of trained models, the coefficient of determination (r^2) [85] was computed against the validation dataset. Higher r^2 values represent a higher predictive accuracy of the trained model.

3. Results and Discussions

3.1. CART Model Trees for Drought Predictions

Using the CART approach, a number of successful model trees were constructed, which can be easily interpretable and used by decision makers in their drought management decisions. Conceptually, the model trees were produced as upside down trees (Figure 1b) and have branches that lead to the actual leaves of the tree. When there are no more splits for the branches and/or leaves, there are linear models representing the node in the form of rules, which have if-conditions and linear-regression equations.

At each node, the model trees were expressed as collections of rules, where each rule has an associated multivariate linear equation. Whenever a situation matches a rule's conditions, the associated equation is used to calculate the predicted value of drought outlooks. A sample rule (*Rule 1 of June model*) for predicting July drought conditions in June (June one-month outlook) is presented in Figure 3. In this case, July drought values range from −3.90 to 1.40 with an average value of −0.96. Here, the drought values are in the units of standard deviation values. The regression tree model finds that the target value of these or other cases satisfying the conditions can be modeled by the following formula:

$$SDNDVI_{July} = -123 + 4.14\ AMO - 0.85\ PDO + 1.16\ SDNDVI_{June} + 0.45\ NAO + 0.3\ PNA - 0.09\ SPI\_3month + 0.009\ DEM$$

```
Rule 1: [748 cases, mean -0.96, range -3.90 to 1.40, est err 0.414]

    if
      SDNDVIJune <= -0.20
      LC in {20, 130, 140, 180, 200}
      ER in {1, 2, 12, 15}
      AMO > 0.23
    then
      JulyDrought = -123 + 4.14 AMO - 0.85 PDO + 1.16 SDNDVIJune + 0.45 NAO + 0.3 PNA - 0.09 SPI_3Month + 0.009 DEM
```

Figure 3. An excerpt rule produced in a June one-month outlook model for predicting July drought. In rule 1, the land cover and ecological region (ER) are categorical values. For the land cover, 20 = Mosaic cropland (50%–70%)/vegetation (grassland/shrubland/forest) (20%–50%); 130 = Closed to open (>15%) (broad-leaved or needle-leaved, evergreen or deciduous) shrubland (<5 m); 140 = Closed to open (>15%) herbaceous vegetation (grassland, savannas, or lichens/mosses); 180 = Closed to open (>15%) grassland or woody vegetation on regularly flooded or waterlogged (fresh, brackish or saline water); 200 = Bare areas. For the ER, 1 = Desert and semi-desert scrubland (DSS); 2 = Acacia-Commiphora woodland and bush land proper (ACB); 12 = Freshwater marshes and swamps, floodplains, and lake shore vegetation (FLV/MFS); 15 = Salt pans, saline/brackish and intermittent wetlands, and salt-lake shore vegetation (SLV/SSS).

During this modeling, the estimated error is 0.414. For a case covered by this rule, AMO has the greatest effect on the drought estimate and DEM has the least effect on determining this drought outlook value. Each rule generated during the modeling was interpreted this way, and whenever a case satisfied all the conditions, the linear model was used for predicting the value of the target attribute. If two or more rules applied to a case, then the calculated values from the respective multiple linear regression models were averaged to arrive at the final prediction of drought conditions.

There are two components of the models produced that include: the if-condition and the actual multiple linear regression model. The 11 attributes used were assessed in the model construction (Table 2). The last column indicated the rank of each attribute, which was determined by summing the attribute usage in the if-conditions and the regression model and dividing it by 2. In our drought modeling experiment, the PDO was the most important attribute for modeling July drought, followed by AMO and SDNDVI, respectively. This means that these three attributes have strong relationships with drought status in the study area. The model performances and selection of the parameters are explored in subsequent sections.

Table 2. Attribute usage assessment for drought prediction experiment.

Attribute	Attribute Usage in the If-Conditions	Attribute Usage in the Regression Model	Rank
AMO	83%	54%	2
MEI	81%	23%	7
ER	72%	36%	6
DEM	61%	52%	4
PDO	51%	87%	1
SPI	50%	61%	5
AWC	40%	39%	8
LC	35%	35%	9
SDNDVI	32%	100%	3
PNA	13%	12%	10
NAO	0%	24%	11

3.2. Model Performance Evaluations

A 24-year historical record of data (1983–2006) was used for developing the time lag prediction models and performance evaluations metrics. In cross-validating the drought models in the regression tree, the data were randomly split into training and testing sets. The learned parameters from the data in the training set were subjected to the parameters of the test dataset, and the quality of the predictions on the test set was evaluated. This approach usually gives an idea of how well the models generalize the unseen data points, with better generalized and more robust models having a higher predictive accuracy [69] of future drought conditions.

Table 3 presents the performance of the "rules alone" model on test data. The MAD, which measures the deviation of the predicted value from the actual known value, ranged from 0.22 to 1.9 for the 10 models assessed. These error values are in terms of standard deviations. The highest error values were observed for the July three-month prediction followed by the August two-month prediction. All of the October month predictions (June four-month outlook, July three-month outlook, August two-month outlook, and September one-month outlook) were found to have about two standard deviation values. The possible explanation for this high value is that October is dry and at the end of the growing season [45,86–89] compared to the predictor months, and this high variability is expected for the study area.

The ratio of the average error magnitude to the error magnitude that would result from always predicting the mean value (RE) ranged from 0.29 to 0.67. The lowest RE was recorded for the August one-month outlook and the highest was for the June-three month outlook. In all 10 models, the RE is <1, which indicates that the average error magnitude is lower in the overall observations. RuleQuest [69] also indicated that for useful models the relative errors should be less than one.

The CC ranged from 0.71 to 0.95. The highest CC was found for the August one-month prediction, which is predicting September drought conditions using August data. This is in agreement with our expectation in that both of these months are vigorous plant-growing months [45,86–89] and there is a similarity in their vegetation conditions over this period of time. The lowest correlation value was for the June three-month prediction. This is in line with our expectation that the predictive accuracy would likely be lower during June, which is early in the growing season when vegetation conditions (e.g., amount and vigor) are more variable because of several early season factors (e.g., moisture and air temperature affecting the initial plant growth rates, resulting in vegetation condition variations produced by several environmental factors beyond drought). In comparison, the August period with higher predictive accuracy is a vigorous growing month [45,86–89] with considerable accumulated plant biomass, leading to more consistent conditions from year to year with major deviations more likely be related to drought stress during the early to mid-growing season.

Table 3. Performances of the drought models using the "rules alone" modeling option on the test dataset.

Model	MAD	RE	CC
June one-month outlook	0.3750	0.49	0.85
June two-month outlook	0.46488	0.61	0.77
June three-month outlook	0.51748	0.67	0.71
June four-month outlook	1.79753	0.57	0.77
July one-month outlook	0.27255	0.36	0.92
July two-month outlook	0.39691	0.51	0.84
July three-month outlook	1.89925	0.59	0.75
August one-month outlook	0.22184	0.29	0.95
August two-month outlook	1.86404	0.58	0.75
September one-month outlook	1.80224	0.57	0.77

Figure 4 presents the performance of the 10 "rules alone" models, and "instances and rules" CART Cubist models. In the 10 assessed models, the MAD in the "instance and rules" model was found to have significantly lower value than the rules-alone model (Figure 4i). The reason for these performance improvements in all of the models for the "instance and rules" model approach is that the latter modeling approach uses bagging (bootstrap aggregating, which is designed to improve the accuracy of the regression tree model) [90,91]. Witten et al. [82] supported this result that the predictive accuracy of a rule-based model can be improved by combining it with nearest neighbor or an instance-based model, which is achieved through the use of bagging in machine learning.

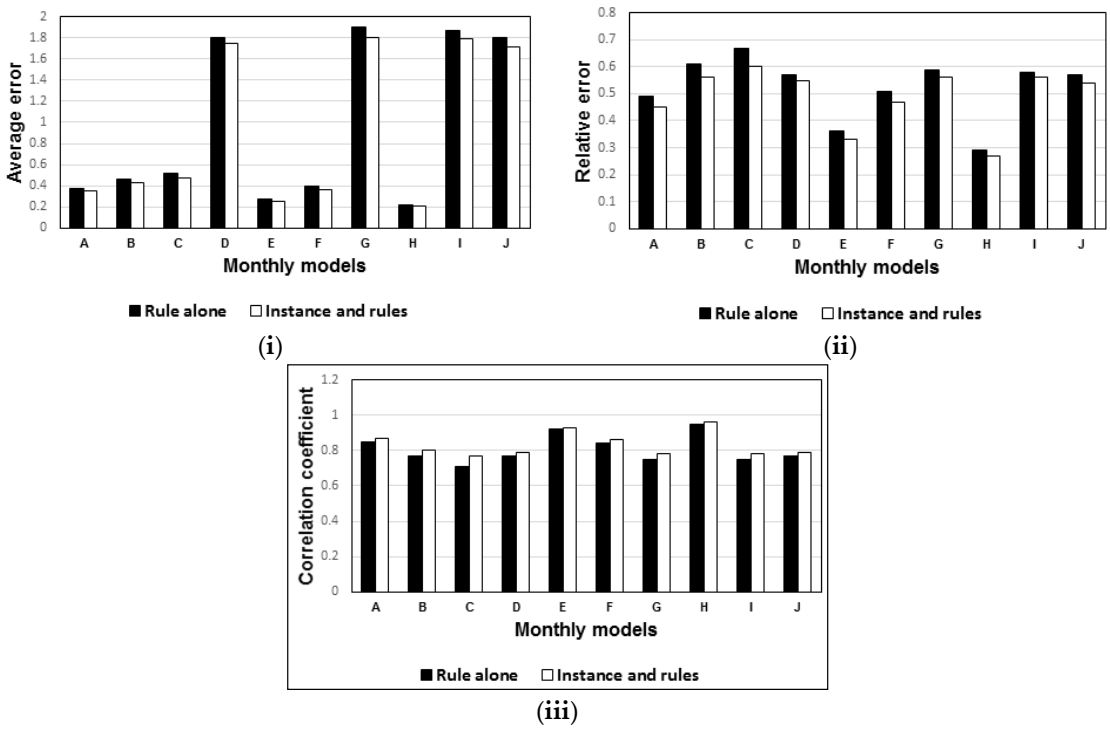


Figure 4. Model performance evaluation: (**i**) MAD; (**ii**) relative error; (**iii**) correlation coefficient for the 10 monthly models; (A) June one-month outlook, (B) June two-month outlook, (C) June three-month outlook, (D) June four-month outlook, (E) July one-month outlook, (F) July two-month outlook, (G) July three-month outlook, (H) August one-month outlook, (I) August two-month outlook, and (J) September one-month outlook.

Fortmann-Roe [92] indicated that bagging and other resampling techniques can be used to reduce the variance in model predictions, where numerous replicates of the original dataset are created using

a random selection with replacement method. Each derivative dataset is then used to construct a new model and the models are gathered together into an ensemble. To make a prediction, all of the models in the ensemble are polled and their results are averaged. Using this intelligence, the MAD can be reduced and the overall modeling performance can be improved.

Moreover, the MAD consistently increased as the prediction period lengths increased (Figure 4i). It was also observed that the June four-month outlook, July three-month outlook, August two-month outlook, and September one-month outlook had the highest MAD both for the instance and rules, and rules-alone models (Figure 4i). The highest MAD values were found during the October month (the to-be predicted month), which is out of the growing season [45,86–89] and these much large errors are expected for this period of the growing season, as reported earlier. The relative error comparison between the "instance and rules" model and the rules-alone model was found to have a pattern similar to the MAD pattern (Figure 4ii).

The CC for the 10 monthly outlook models showed that the "instance and rules" model had consistently higher correlations than the "rules alone" models (Figure 4iii). Our assessment also showed that for all 10 models, the "instance and rules" model version were found to have higher accuracy than the "rules alone" model (Figure 4iii). The comparison of average CC between the "instance and rules" model and the "rules alone" model showed the "instance and rules" model CC to be significantly higher than the "rules alone" model ($p < 0.05$).

In addition to the instance and rules options, there is *let Cubist decide* in the Cubist regression tree modeling option [69]. For the 10 models assessed, the *let Cubist decide* option was found to have the same accuracy (in terms of MAD, RE, and CC values) as composite models (which combine instances and rules), showing equivalent performances of the two modeling options in the Cubist modeling tool. Similar results were also found by Ruefenacht et al. [93] in that the *let cubist decide* model has the same accuracy as the composite models. The challenge using the *let cubist decide* option was the increased time required to build the model, as the optimal decision was determined by Cubist.

In addition to the composite rule-based nearest-neighbor models, RuleQuest [69] can also generate committee models made up of several rule-based models. Each member of the committee predicts the target value for a case and the members' predictions are averaged to give a final prediction.

The first member of a committee model is always exactly the same as the model generated without the committee option. The second member is a rule-based model designed to correct the predictions of the first member; if the first member's prediction is too low for a case, the second member will attempt to compensate by predicting a higher value. The third member tries to correct the predictions of the second member, and so on. The default number of members is five, a value that balances the benefits of the committee approach against the cost of generating extra models [69].

Before combining the committee and neighbor models, the performances of these approaches were separately analyzed (Figure 5). For deciding the actual threshold values for the number of committees and neighbors to be used in Cubist drought modeling, the *r*-squared values and RMSE were assessed.

Figure 5 presents the *r*-squared values and RMSE for the different committees and number of neighbors to be used. In the committee models, it can be observed that the highest average *r*-squared value was obtained for 30 committees, and there was no performance improvement gained with the addition of more committees. The RMSE decreased as the number of committees increased from 1 to 30, but remained the same as further committees were added. Therefore, in using the committee models, it is imperative to use 30 committee models in future drought modeling experiments.

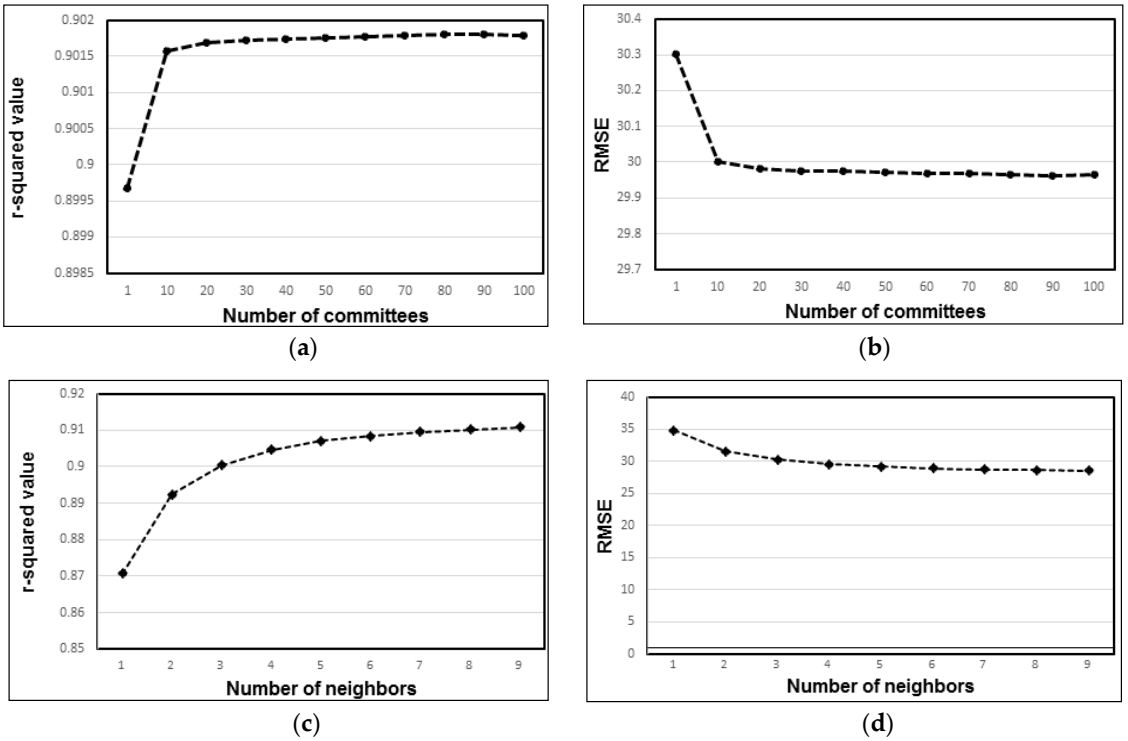


Figure 5. Performance evaluations using (**a**) *r*-squared values for 1–100 committees; (**b**) Root mean square error (RMSE) for 1–100 committees; (**c**) *r*-squared values for 1–9 neighbors; (**d**) RMSE for 1–9 neighbors.

For the number of neighbors, the *r*-squared value was found to increase up to seven neighbors with the *r*-squared values remaining unchanged, as additional neighbors were included. The RMSE was found to decrease up to seven neighbors and remain the same thereafter. Therefore, the optimum neighbor threshold was found to be seven to achieve the highest predictive accuracy.

The comprehensive set of prediction performances on models for three months (June, July and August) using 0–9 neighbors and with committee ranges 1–100 are presented in Figure 6. This analysis was done in line with the GUI-based implementations of RuleQuest [69] for modeling. The GUI has different checkbox options and also has options for specifying the values for the number of nearest instances and committees. Therefore, Figure 6 gives us the benefits that we may achieve by specifying the parameters in the GUI options, as well as parameter specification in the command line, batch-processing options.

For the 0-neighbor and 1-committee options, the highest *r*-squared value and lowest RMSE values were achieved for the August one-month outlook (see the * symbol in Figure 6) with the lowest *r*-squared and the highest RMSE observed for the June three-month outlook (see the black square symbol in Figure 6). These model performances are in line with our previous explanations. From the assessed possible combinations with 0–9-neighbors and 1–100 committee June–August outlooks, the same pattern was observed for all with the exception of 1-neighbor with 1-committee, 10-committee, 50-committee, and 100-committee options (Figure 6). In all four combinations with 1-neighbor, the RMSE was the highest and the *r*-squared was the lowest. The possible explanation for this is that having only one instance and creating the model based on this instance only generated a biased model, and the model performance becomes highly affected. The zero-neighbors option did not use the instance modeling option, which is better than the 1-neighbor instance modeling option (Figure 6).

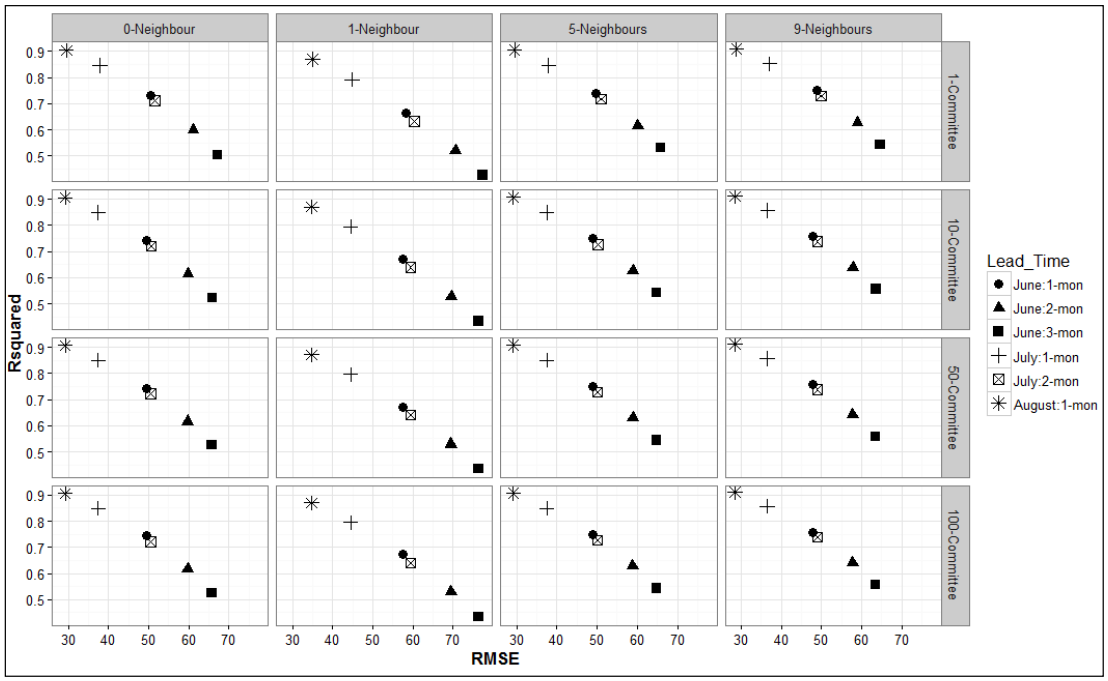


Figure 6. Performance of June, July, August outlook models on the test dataset. This figure presents the possible combinations of the instance and committee models.

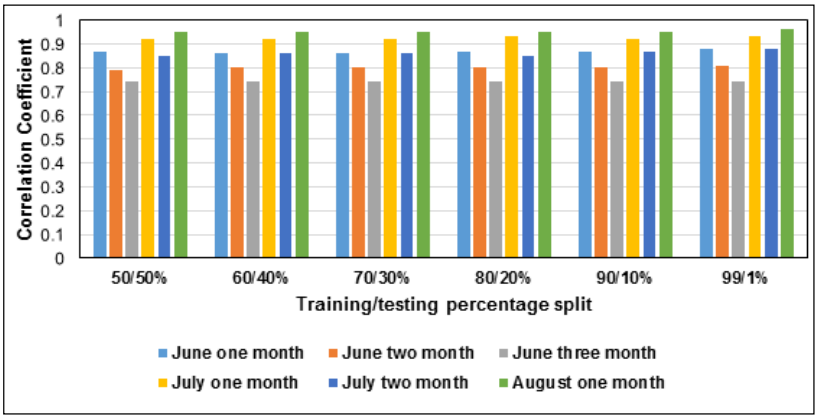
3.3. Percentage Splits for Training and Testing

The percentage splits for training–testing were done for 50/50%, 60/40%, 70/30%, 80/20%, 90/10%, and 99/1%. The relative performances of the models were compared for six models during the core of the growing season (June one-month outlook, June two-month outlook, June three-month outlook, July one-month outlook, July two-month outlook, and August one-month outlook) (Figure 7). The MAD ranged from 0.2 to 0.5. In all of the assessed models, the minimum MAD was found for the August one-month outlook, and the maximum MAD was observed for the June three-month outlook (Figure 7a). This is in agreement with our expectation that as prediction length increases from one month to three months, the error increases and the performance of the models decreases.

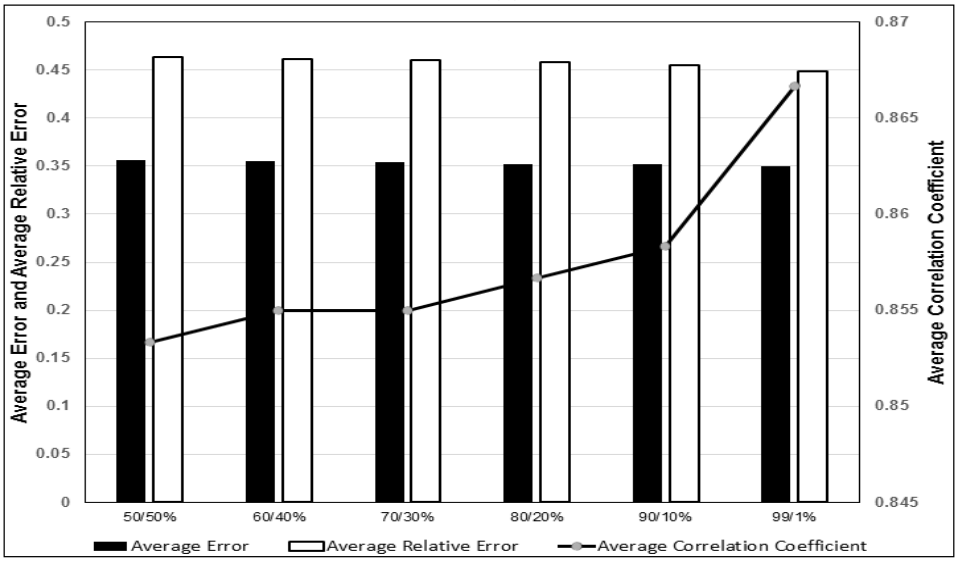
The descriptive statistics for the MAD of the six models are presented in Table 4. The average of the MAD for the six models ranged from 0.35 to 0.36. The minimum error was recorded for the 99/1% split and the maximum error was observed for the 50/50% split. The same pattern was observed for the RE (not presented here) as the MAD. The possible explanation for this is that as more data are used for training the model and less data are assigned to the test set, the MAD was found to be decreasing. Therefore, in terms of the MAD parameter, 99/1% is performing the best.

Table 4. Performance of the monthly outlook models for the MAD on the percentage splits.

Monthly Outlooks	Percentage Splits					
	50/50%	**60/40%**	**70/30%**	**80/20%**	**90/10%**	**99/1%**
June one-month	0.365	0.367	0.365	0.363	0.363	0.361
June two-month	0.4408	0.4386	0.4354	0.4336	0.4398	0.4392
June three-month	0.4886	0.487	0.4901	0.4851	0.4912	0.4999
July one-month	0.2622	0.2623	0.2625	0.258	0.2592	0.2497
July two-month	0.3672	0.3655	0.3658	0.3625	0.3501	0.3424
August one-month	0.2118	0.2091	0.2082	0.2095	0.2077	0.2044
Average	0.355933	0.354917	0.3545	0.35195	0.351833	0.349433
Maximum	0.4886	0.487	0.4901	0.4851	0.4912	0.4999
Minimum	0.2118	0.2091	0.2082	0.2095	0.2077	0.2044
Standard deviation	0.09537	0.095348	0.095776	0.094646	0.097163	0.101541



(**a**)



(**b**)

Figure 7. Percentage split comparisons for the training and testing; (**a**) monthly models comparisons; (**b**) average performances evaluations for MAD, RE, and CC.

The relative performances of the six models described above were also compared in terms of accuracy CC (Figure 7b). The CC ranged from 0.74 to 0.96. In all models assessed, the maximum CC was found for the August one-month outlook, and the minimum CC was observed for the June three-month outlook. This is in agreement with our expectation in that the August one-month outlook predicted September conditions, where these two months have similar vegetation conditions for the study area near the end of the core growing season [45]. In general, as the prediction length increases, the CC was found to be decreasing (Figure 7a). There are not significant differences between the percentage splits in terms of both MAD and RE (Figure 7b). The CC was found to be increasing consistently as split percentage changes from 50/50% to 99/1% (Figure 7b). For the practical implementation of drought modeling using CART, it seems imperative to use the 90/10% splits (compared to both 80/20% and 99/1%) for two reasons: (1) the accuracy is higher compared to 80/20% and reasonably lower compared to 99/1% splits, and (2) evaluation of the model's performance on unseen data will be more representative and robust using 10% of the data for testing rather than 1% of the total dataset.

4. Conclusions

In this paper, CART modeling approaches were explored to develop a clear methodology, specifically a practical data and information mining approach, for drought modeling and predictions.

The developed approach is targeted to help decision makers convert large volumes of data available from different sources (e.g., satellite observations and in situ measurements) into actionable information that can be used for drought prediction, management, and mitigation activities.

The CART approach was used to construct a number of successful model trees, which can be interpreted by decision makers for use in drought management decisions. The model trees produced are easily understandable and can be interpreted for actionable information. After assessing the modeling options, it was concluded that ensemble models had the best performance.

Among the 10 models assessed, the "instance and rules" models had significantly lower MAD values than the "rules alone" models. The models' accuracies were also found to be significantly higher in the "instance and rules" models (in terms of correlation coefficients) (0.77–0.96) compared to the "rules alone" models. The better performance of models generated using the "instance and rules" approach can be attributed to the bagging (i.e., bootstrap aggregating) method that was added to this type of model, which has been shown to improve model performance compared to approaches such as the "rules alone" option tested here.

In the experimental drought prediction modeling, there were two modeling options, the nearest neighbor and the committee options, which were found to significantly improve model performance. Assessment of the different combinations of nearest-neighbor and committee models revealed that these two options significantly increased performance of the CART drought models when proper parameters were defined compared to the default parameter specifications.

For the practical implementation of this drought modeling experiment, the following conclusions can be drawn:

1. The "instance and rules" models significantly increased the performance of the drought model compared to the "rules alone" model. For further performance enhancement of the models, the instance-and-rules model can be combined with ensemble models.
2. When using ensemble model options, it was confirmed that the highest average *r*-squared value was obtained for the model with 30 committees. There was no gain in performance when additional committees were added to the models. RMSE consistently decreased for models with one to 30 committees but remained the same for all models with more than 30 committees. Therefore, it is imperative to use 30 committee models for future experiments employing committee models.
3. For ensemble models, the *r*-squared value increased consistently with up to seven neighbors, with *r*-squared values remaining constant when any additional neighbors were considered. RMSE also decreased with up to seven neighbors but remained constant when additional neighbors were considered. Therefore, the optimum threshold for the highest model accuracy is the use of seven neighbors.
4. In the iterative experiment to determine the best training/testing data split (i.e., 50/50%, 60/40%, 70/30%, 80/20%, 90/10%, or 99/1%), the CC results revealed a consistent increase as split percentage changed from 50/50% to 99/1%. For the practical implementation of regression tree modeling of drought, accuracy was higher when the original data set was split 90/10% into training and testing data sets.

In the experimental analysis, the "nearest neighbor" models or the "instance and rules" modeling option had the best performance. One of the challenges in "nearest neighbor" models was that the models were adversely affected by the presence of irrelevant attributes. All attributes were taken into account when evaluating the similarity of the two cases ("rules alone", and "instance and rules") and irrelevant attributes introduced a random factor into this measurement. Composite models are most effective when the numbers of attributes are relatively small and all attributes are relevant to the prediction task. However, there was uncertainty regarding the decision about which attributes are the most relevant and which are irrelevant. Future research is recommended for developing an approach for selecting the most relevant attributes for drought modeling experiments.

This research demonstrated the CART modeling approach for practically specifying the modeling parameters, which is again helpful for integrating the different data available from climate, satellite, environmental, and oceanic sources in drought modeling. In an overall assessment, a regression tree modeling approach, which is well known in other domains, was found to be an excellent approach in capturing the drought pattern.

Limitation of the current study is that the results are applicable to only one study area with a vegetation growing season from June to October and were not replicated for other study areas to observe spatio-temporal differences. Future research could determine the influence of spatio-temporal differences on model performance.

Acknowledgments: This work was supported by NASA Project NNX14AD30G. The authors are grateful to Deborah Wood of the National Drought Mitigation Center for her editorial comments.

Author Contributions: Getachew B. Demisse, Tsegaye Tadesse, Solomon Atnafu, and Shawndra Hill conceived and designed the experiments; Getachew B. Demisse performed the experiments; Yared Bayissa and Andualem Shiferaw analyzed the data; Shawndra Hill, Getachew B. Demisse, Brian D. Wardlow and Solomon Atnafu wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Negash, S.; Gray, P. Business Intelligence. In *Handbook on Decision Support Systems*; Burstein, F., Holsapple, C., Eds.; Springer: Heidelberg, Germany, 2008; Volume 2, pp. 175–193.
2. Langseth, J.; Vivatrat, N. Why Proactive Business Intelligence is a Hallmark of the Real-Time Enterprise: Outward Bound. *Intell. Enterp.* **2003**, *5*, 34–41.
3. Ali, M.; Bosse, T.; Hindriks, K.V.; Hoogendoorn, M.; Jonker, C.M.; Treur, J. Recent Trends in Applied Artificial Intelligence. In Proceedings of the 26th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE 2013), Amsterdam, The Netherlands, 17–21 June 2013.
4. Hor, C. Extracting Knowledge From Substations for Decision Support. *IEEE Trans. Power Deliv.* **2005**, *20*, 595–602. [CrossRef]
5. Nonaka, I. A Dynamic Theory of Organizational Knowledge Creation. *Organ. Sci.* **1994**, *5*, 14–37. [CrossRef]
6. Dienes, Z.; Perner, J. A theory of implicit and explicit knowledge. *Behav. Brain Sci.* **1999**, *22*, 735–808. [CrossRef] [PubMed]
7. Han, H.; Kamber, M. *Data Mining: Concepts and Techniques*, 2nd ed.; Morgan Kaufmann Publishers: San Francisco, CA, USA, 2006.
8. Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. The KDD process for extracting useful knowledge from volumes of data. *Commun. ACM* **1996**, *39*, 27–34. [CrossRef]
9. Jackson, J. Data Mining: A Conceptual Overview. *Commun. Assoc. Inf. Syst.* **2002**, *8*, 267–296.
10. Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96 AAAI), Portland, OR, USA, 2–4 August 1996; Volume 96, pp. 82–88.
11. Miller, H.J.; Han, J. *Geographic Data Mining and Knowledge Discovery*; Taylor & Francis: London, UK, 2001.
12. UNCCD. *United Nations Convention to Combat Desertification, Article 1*; United Nations: Bonn, Germany, 1999.
13. Dai, A. Drought under global warming: A review. *Adv. Rev. Natl. Center Atmos. Res.* **2011**, *2*, 45–65. [CrossRef]
14. Wilhite, D. *Drought and Water Crisis: Science, Technology and Management Issues*; Taylor & Francis: Boca Raton, FL, USA, 2005.
15. Masih, I.; Maskey, S.; Mussá, F.; Trambauer, P. A review of droughts on the African continent: A geospatial and long-term perspective. *Hydrol. Earth Syst. Sci.* **2014**, *18*, 3635–3649. [CrossRef]
16. EM-DAT. EM-DAT: The International Disaster Database. Available online: http://www.emdat.be/ (accessed on 22 August 2015).
17. Kumar, V.; Panu, U. Predictive assessment of severity of agricultural droughts based on agro-climatic factors. *J. Am. Water Resour. Assoc.* **1997**, *33*, 1255–1264. [CrossRef]
18. Leilah, A.A.; Al-Khate, S.A. Statistical analysis of wheat yield under drought conditions. *J. Arid Environ.* **2005**, *61*, 483–496. [CrossRef]

19. Mishra, A.K.; Desai, V.R. Drought forecasting using stochastic models. *Stoch. Environ. Res. Risk Assess.* **2005**, *19*, 326–339. [CrossRef]
20. Durdu, O.F. Application of linear stochastic models for drought forecasting in the Buyuk Menderes river basin, western Turkey. *Stoch. Environ. Res. Risk Assess.* **2010**, *24*, 1145–1162. [CrossRef]
21. Modarres, R. Streamflow drought time series forecasting. *Stoch. Environ. Res. Risk Assess.* **2007**, *21*, 223–233. [CrossRef]
22. Han, P.; Wang, P.X.; Zhang, S.Y.; Zhu, D.H. Drought forecasting based on the remote sensing data using. ARIMA Models. *ARIMA Model.* **2010**, *51*, 1398–1403. [CrossRef]
23. Fernandez, C.; Vega, J.A.; Fonturbel, T.; Jimenez, E. Streamflow drought time series forecasting: A case study in a small watershed in North West Spain. *Stoch. Environ. Res. Risk Assess.* **2009**, *23*, 1063–1070. [CrossRef]
24. Lohani, V.K.; Loganathan, G.V. An early warning system for drought management using the palmer drought index. *J. Am. Water Resour. Assoc.* **1997**, *33*, 1375–1386. [CrossRef]
25. Paulo, A.A.; Ferreira, E.; Coelho, C.; Pereira, L.S. Drought class transition analysis through Markov and Loglinear models, an approach to early warning. *Agric. Water Manag.* **2005**, *77*, 59–81. [CrossRef]
26. Cancelliere, A.; Mauro, G.D.; Bonaccorso, B.; Rossi, G. Drought forecasting using the Standardized Precipitation Index. *Water Resour. Manag.* **2007**, *21*, 801–819. [CrossRef]
27. Steinemann, A. Drought indicators and triggers: A stochastic approach to evaluation. *J. Am. Water Resour. Assoc.* **2003**, *39*, 1217–1233. [CrossRef]
28. Banik, P.; Mandal, A.; Rahman, M.S. Markov chain analysis of weekly rainfall data in determining drought-proneness. *Discret. Dyn. Nat. Soc.* **2002**, *7*, 231–239. [CrossRef]
29. Ochola, W.O.; Kerkides, P. A Markov chain simulation model for predicting critical wet and dry spells in Kenya: Analysing rainfall events in the Kano plains. *Irrig. Drain.* **2003**, *52*, 327–342. [CrossRef]
30. Moreira, E.E.; Paulo, A.A.; Pereira, L.S.; Mexia, J.T. Analysis of SPI drought class transitions using loglinear models. *J. Hydrol.* **2006**, *331*, 349–359. [CrossRef]
31. Morid, S.; Smakhtinb, V.; Bagherzadehc, K. Drought forecasting using artificial neural networks and time series of drought indices. *Int. J. Climatol.* **2007**, *27*, 2103–2111. [CrossRef]
32. Mishra, A.K.; Desai, V.R. Drought forecasting using feed-forward recursive neural network. *Ecol. Model.* **2006**, *198*, 127–138. [CrossRef]
33. Kim, T.; Valdes, J. A nonlinear model for drought forecasting based on conjunction of wavelet transforms and neural networks. *J. Hydrol. Eng.* **2003**, *8*, 319–328. [CrossRef]
34. Mishra, A.K.; Desai, V.R.; Singh, V.P. Drought forecasting using a hybrid stochastic and neural network model. *J. Hydrol. Eng.* **2007**, *12*, 626–638. [CrossRef]
35. Bacanli, U.G.; Firat, M.; Dikbas, E.F. Adaptive Neuro-Fuzzy Inference System for drought forecasting. *Stoch. Environ. Res. Risk Assess.* **2009**, *23*, 1143–1154. [CrossRef]
36. Pongracz, R.; Bogardi, I.; Duckstein, L. Application of fuzzy rule-based modeling technique to regional drought. *J. Hydrol.* **1999**, *224*, 100–114. [CrossRef]
37. Balling, J.; Goodrich, G.B. Analysis of drought determinants for the Colorado River Basin. *Clim. Chang.* **2007**, *82*, 179–194. [CrossRef]
38. Steinemann, A. Using climate forecasts for drought management. *J. Appl. Meteorol. Climatol.* **2006**, *75*, 1353–1361. [CrossRef]
39. Farokhnia, A.; Morid, S.; Byun, H.R. Application of global SST and SLP data for drought forecasting on Tehran plain using data mining and ANFIS techniques. *Theor. Appl. Climatol.* **2011**, *104*, 71–81. [CrossRef]
40. Dhanya, C.T.; Nagesh, K.D. Data mining for evolution of association rules for droughts and floods in India using climate inputs. *J. Geophys. Res.* **2009**, *114*, 1–15. [CrossRef]
41. Vasiliades, L.; Loukas, A. Spatiotemporal drought forecasting using nonlinear models. In Proceedings of the EGU General Assembly 2010, Vienna, Austria, 2–7 May 2010.
42. Tadesse, T.; Wilhite, D.; Harms, S.; Hayes, M.; Goddard, S. Drought Monitoring Using Data Mining Techniques: A Case Study for Nebraska, USA. *Nat. Hazards* **2004**, *33*, 137–159. [CrossRef]
43. Mishra, A.K.; Singh, V.P. Drought modeling—A review. *J. Hydrol.* **2011**, *403*, 157–175. [CrossRef]
44. Demisse, G.B.; Tadesse, T.; Solomon, A. Drought Spatial Object Prediction Approach using Artificial Neural Network. *Geoinform. Geostat. Overv.* **2015**, *3*, 1–7.
45. Demisse, G.B. Knowledge Discovery From Satellite Images for Drought Monitoring. Ph.D. Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2013.

46. Andreadis, K.M.; Clark, E.A.; Wood, A.W.; Hamlet, A.F.; Lettenmaier, D.P. Twentieth-century drought in the conterminous United States. *J. Hydrometeorol.* **2005**, *6*, 985–1001. [CrossRef]
47. Dubrovsky, M.; Svoboda, M.D.; Trnka, M.; Hayes, M.J.; Wilhite, D.A.; Zalud, Z.; Hlavinka, P. Application of relative drought indices in assessing climate-change impacts on drought conditions in Czechia. *Theor. Appl. Climatol.* **2008**, *96*, 155–171. [CrossRef]
48. NOAA. DROUGHT: Monitoring Economic, Environmental, and Social Impacts. Available online: http://www.ncdc.noaa.gov/news/drought-monitoring-economic-environmental-and-social-impacts (accessed on 4 January 2016).
49. Sheffield, J.; Wood, E.F. Projected changes in drought occurrence under future global warming from multi-model, multi-scenario, IPCC AR4 simulations. *Clim. Dyn.* **2008**, *31*, 79–105. [CrossRef]
50. UCS. Causes of Drought: What's the Climate Connection? Union of Concerned Scientists (UCS). Available online: http://www.ucsusa.org/global_warming/science_and_impacts/impacts/causes-of-drought-climate-change-connection.html#.VprO5k98wRI (accessed on 10 December 2016).
51. National Meteorological Services Agency (NMSA). *Assessment of Drought in Ethiopia*; Meteorological Research Reports Series, No. 2; NMSA: Addis Ababa, Ethiopia, 1996.
52. EMA. Ethiopian Mapping Agency (EMA). Available online: http://www.ema.gov.et/ (accessed on 22 December 2016).
53. FEWSNET. Normalized Difference Vegetation Index, Product Documentation. Available online: http://earlywarning.usgs.gov/fews/africa/web/readme.php?symbol=nd (accessed on 20 June 2011).
54. Holben, B.N. Characteristics of maximum-value composite images from temporal data. *Int. J. Remote Sens.* **1986**, *7*, 1417–1434. [CrossRef]
55. USGS. USGS—Earth Resources Observation and Science (EROS) Center-Elevation Data. Available online: http://eros.usgs.gov/#/Find_Data/Products_and_Data_Available/gtopo30/hydro/africa (accessed on 1 September 2011).
56. Ecodiv.org. Atlas of the Potential Vegetation of Ethiopia. Available online: http://ecodiv.org/atlas_ethiopia/index.html (accessed on 1 September 2011).
57. ESA. European Space Agency, Global Land Cover Map. Available online: http://ionia1.esrin.esa.int/index.asp (accessed on 10 November 2011).
58. GLCF. Global Land Cover Facility. Available online: http://www.landcover.org/aboutUs/ (accessed on 20 December 2010).
59. NOAA. National Oceanic and Atmospheric Administration, Climate Indices: Monthly Atmospheric and Ocean Time Series. Available online: http://www.esrl.noaa.gov/psd/data/climateindices/list/ (accessed on 1 September 2011).
60. Enfield, D.B.; Mestas-Nunez, A.M.; Trimble, P.J. The Atlantic multidecadal oscillation and it's relation to rainfall and river flows in the continental U.S. *Geophys. Res. Lett.* **2001**, *28*, 2077–2080. [CrossRef]
61. Hurrell, J.W. Decadal trends in the North Atlantic Oscillation and relationships to regional temperature and precipitation. *Science* **1995**, *269*, 676–679. [CrossRef] [PubMed]
62. Jones, P.D.; Jonsson, T.; Wheeler, D. Extension to the North Atlantic Oscillation using early instrumental pressure observations from Gibraltar and South-West Iceland. *Int. J. Climatol.* **1997**, *17*, 1433–1450. [CrossRef]
63. Wolter, K.; Timlin, M.S. Measuring the strength of ENSO—How does 1997/98 rank? *Weather Forecast.* **1998**, *53*, 315–324. [CrossRef]
64. Frank, I.; Kalivas, J.; Kowalski, B. Partial Lease Square Solutions for Multicomponent Analysis. *Lab. Chemom.* **1983**, *55*, 1800–1804.
65. Tadesse, M.; Sha, N.; Vannucci, M. Bayesian Variable Selection in Clustering HighDimensional Data. *J. Am. Stat. Assoc.* **2005**, *100*, 602–617. [CrossRef]
66. Pierna, J.; Dardenne, P. Soil parameter quantification by NIRS as a Chemometric challenge at 'Chimiométrie 2006'. *Chemom. Intell. Lab. Syst.* **2008**, *91*, 94–98. [CrossRef]
67. Xu, Q.; Daszykowski, M.; Walczak, B.; Daeyaert, F.; Jonge, M.R.; Koymans, H.; Lewi, P.J.; Vinkers, H.M.; Janssen, P.A.; Heeres, J.; et al. Multivariate adaptive regression splines—Studies of HIV reverse transcriptase inhibitors. *Chemom. Intell. Lab. Syst.* **2004**, *72*, 27–34. [CrossRef]
68. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
69. Rulequest. An Overview of Cubist. Available online: http://www.rulequest.com/cubistwinhtml (accessed on 20 September 2015).

70. Tadesse, T.; Wardlow, B.D.; Hayes, M.J.; Svoboda, M.D. The Vegetation Outlook (VegOut): A New Method for Predicting Vegetation Seasonal Greenness. *GIScience Remote Sens.* **2010**, *47*, 25–52. [CrossRef]
71. Brown, J.F.; Wardlow, B.D.; Tadesse, T.; Hayes, M.J.; Reed, B.C. The Vegetation Drought Response Index (VegDRI): A New Integrated Approach for Monitoring Drought Stress in Vegetation. *GIScience Remote Sens.* **2008**, *45*, 16–46. [CrossRef]
72. Tadesse, T.; Demisse, G.; Zaitchik, B.; Dinku, T. Satellite-based hybrid drought monitoring tool for prediction of vegetation condition in Eastern Africa: A case study for Ethiopia. *Water Resour. Res.* **2014**, *50*. [CrossRef]
73. Berhan, G.; Shawndra, H.; Tadesse, T.; Solomon, A. Drought Prediction System for Improved Climate Change Mitigation. *IEEE Transs Geosci. Remote Sens.* **2014**, *52*, 4032–4037. [CrossRef]
74. Minasny, B.; McBratney, A. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. *Chemom. Intell. Lab. Syst.* **2008**, *94*, 72–79. [CrossRef]
75. Shao, Q.; Rowe, R.C.; York, P. Investigation of an artificial intelligence technology-Model trees Novel applications for an immediate release tablet formulation database. *Eur. J. Pharm. Sci.* **2007**, *31*, 137–144. [CrossRef] [PubMed]
76. Loh, W.Y.; Shih, Y.S. Split selection methods for classification trees. *Stat. Sin.* **1997**, *7*, 815–840.
77. Quinlan, J.R. Learning with Continuous Classes. In Proceedings of the AI 92 (Adams & Sterling, Eds.), Hobart, Australia, 16–18 November 1992; World Scientific: Singapore, 1992; pp. 343–348.
78. Hullermeier, E. Possibilistic instance-based learning. *Artif. Intell.* **2003**, *148*, 335–383. [CrossRef]
79. Aha, D.W. Lazy Learning. *Artif. Intell. Rev.* **1997**, *11*, 7–10. [CrossRef]
80. Aha, D.W.; Kibler, D.; Albert, M.K. Instance-based learning algorithms. *Mach. Learn.* **1991**, *6*, 37–66. [CrossRef]
81. Taylor, R. Interpretation of the correlation coefficient: A basic review. *J. Diagn. Med. Sonogr.* **1990**, *6*, 35–39. [CrossRef]
82. Witten, I.; Frank, E.; Hall, M.A. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed.; Elsevier: San Francisco, CA, USA, 2011.
83. Henderson, B.L.; Bui, E.N.; Moran, C.J.; Simon, D.A.P. Australia-wide predictions of soil properties using decision trees. *Geoderma* **2005**, *124*, 383–398. [CrossRef]
84. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. [CrossRef]
85. Kvalseth, T.O. Cautionary note about R 2. *Am. Stat.* **1985**, *39*, 279–285. [CrossRef]
86. Cheung, W.H.; Senay, G.; Singh, A. Trends and spatial distribution of annual and seasonal rainfall in Ethiopia. *Int. J. Climatol.* **2008**, *28*, 1723–1734. [CrossRef]
87. Korecha, D.; Barnston, A.G. Predictability of June–September Rainfall in Ethiopia. *Mon. Weather Rev.* **2006**, *135*, 628–650. [CrossRef]
88. Segele, Z.T.; Lamb, P.J. Characterization and variability of Kiremt rainy season over Ethiopia. *Meteorol. Atmos. Phys.* **2005**, *89*, 153–180. [CrossRef]
89. Seleshi, Y.; Zanke, U. Recent Changes In Rainfall and Rainy Days In Ethiopia. *Int. J. Climatol.* **2004**, *24*, 973–983. [CrossRef]
90. Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
91. Oza, N.C. *Ensemble Data Mining Methods*; NASA Ames Research Center: Moffett Field, CA, USA, 2004.
92. Fortmann-Roe, S. Understanding the Bias-Variance Tradeoff. Available online: http://scott.fortmann-roe.com/docs/BiasVariance.html (accessed on 14 October 2016).
93. Ruefenacht, B.; Hoppus, A.; Caylor, J.; Nowak, D.; Walton, J.; Yang, L.; Koeln, G. *Analysis of Canopy Cover and Impervious Surface Cover of Zone 41*; San Dimas Technology & Development Center: San Dimas, CA, USA, 2002.