

Comparative Study of Machine Learning models in Drought Prediction

Raj Bharatbhai Choksi, San José State University
Piyush Gade, San José State University
Alekhya Pasupuleti, San José State University
Mohamed Shafeeq Usman, San José State University

November 2022

Abstract

Agriculture is an important part of the US economy. According to the US government agriculture contributed \$1.5 trillion to the economy in 2020 which is a 5% share. However global warming and changes in climate leads to significant drought in various parts of the country which adversely affects agriculture. Unlike other natural disasters, Drought develops slowly and has long term consequences. Hence by leveraging machine learning we can help farmers in taking preventive measures and minimize their loss.

Our aim is to provide a comparative study on the performance of different machine learning models in predicting five levels of drought ranging from moderate to extreme using meteorological data. Weather conditions and precipitation levels at different heights from the sea level will play an important indicators for predicting droughts. We aim to use supervised learning models such as Random Forest, Decision Tree, K Nearest Neighbors and Logistic Regression for the study and compare their results using performance metrics such as F1 score, accuracy, recall, precision and ROC curve.

Introduction

Data Preprocessing

Dimensionality Reduction

Upon observation of the above correlation matrix we observed that that is high correlation between certain features Eg. there is a 95% correlation between T2MDEW & QV2M and that can impact model performance and computing resources. Hence we implemented dimensionality reduction to reduce the number of features by keeping the as much variations as the original dataset. We implemented two dimensionality algorithms namely Principal Component analysis and Linear Discriminant Analysis.

Principal Component Analysis

Principal Component Analysis (PCA) is a linear dimensionality reduction algorithm which uses variance to reduce transform the data into low dimension. We implemented PCA on both downsampled data using near miss and upsampled data using SMOTE. We picked the features with more than 90% variance.

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) uses classes to separate training instances. Difference between LDA and PCA is that LDA uses linear combination of features which optimizes the class separability whereas

PCA uses variance to find set of uncorrelated components. We implemented LDA for both downsampled data with near miss and upsampled data using SMOTE. Using LDA we reduced the number of features to 5 out of 15 features.

Methods

Decision Tree Algorithm

The first method which we implemented is Decision Tree classifier where in we compared classification of each level of drought from class 0 to class 5 with the rest. Understanding the working of decision tree algorithm it breaks the dataset into small subsets while simultaneously expanding the tree nodes. This process continues until we can a pure leaf node without any impurities. The error rate is measured in a term called 'Entropy' wherein we try to reduce the entropy to minimum by increasing the information gain.

Comparisons

Example Analysis

Conclusions

References