The impact of droughts across the United States is far more than just a lack of drinking water for a certain part of the country. According to the CDC, they also can negatively impact aquatic ecosystems, drinking water quality, air quality, food and nutrition, and hygiene. An improvement in our ability to predict droughts and their severity could lead to better preparation and quicker response, potentially saving both lives and money. The National Centers for Environmental Information attribute over 4,500 deaths to just 31 significant drought effects, and billions of dollars of damages.

The United States' Drought Monitor is a joint project between the National Drought Mitigation Center at the University of Nebraska-Lincoln, the National Oceanic and Atmospheric Administration and the U.S. Department of Agriculture. Data from this shared group of meteorologists and climatologists formed the focal point of our project. The drought level data that formed our target variable is exclusively from the USDM, while additional pieces of data are from:

- NASA Langley Research Center POWER Project
- Harmonized World Soil Database
- U.S. Drought Monitor

As a result of this amalgamation of sources, the data set that we had at our disposal was extensive with a large number of variables to consider. First were descriptive variables that would play no role in our modeling but were important for indexing - FIPS county codes, a unique indicator for each county in the US, and date - self-explanatory.

We then had meteorological data for each county:

- Precipitation in millimeters
- Surface Pressure in kPA

- Specific Humidity in k/kg measured at 2 meters
- Temperature in degrees Celsius measured at 2 meters
- Dew/Frost point in degrees Celsius measured at 2 meters
- Wet Bulb Temperature in degrees Celsius measured at 2 meters
- Temperature Maximum, Minimum, and Range in degrees Celsius measured at 2 meters
- Earth Skin Temperature in degrees Celsius
- Wind Speed Maximum, Minimum, and Range in m/s measured at 10 and 50 meters

The most important variable is drought score, measured 0-5, further detailed in the image below from the USDM.

Category	Description	Possible Impacts
D0	Abnormally Dry	Going into drought: ▪ short-term dryness slowing planting, growth of crops or pastures Coming out of drought: ▪ some lingering water deficits ▪ pastures or crops not fully recovered
D1	Moderate Drought	▪ Some damage to crops, pastures ▪ Streams, reservoirs, or wells low, some water shortages developing or imminent ▪ Voluntary water-use restrictions requested
D2	Severe Drought	▪ Crop or pasture losses likely ▪ Water shortages common ▪ Water restrictions imposed
D3	Extreme Drought	▪ Major crop/pasture losses ▪ Widespread water shortages or restrictions
D4	Exceptional Drought	▪ Exceptional and widespread crop/pasture losses ▪ Shortages of water in reservoirs, streams, and wells creating water emergencies

In addition to the meteorological data, we were also equipped with soil data for each FIPS coded county. This data included the following:

- Latitude, Longitude, and Elevation
- Slope
- Mapped water bodies, sparsely vegetated land, built-up land, grass/scrub/woodland, forest land, cultivated land, irrigated cultivated land
- Nutrient availability, nutrient retention capacity, rooting conditions, oxygen available, excess salts, toxicity, workability

Between these two sets of data we were able to create a comprehensive report on the factors influencing drought risk for each listed county.

Cleaning/preprocessing steps

The first step in our preprocessing was to drop the variable “fips”, this was a numerical assignment to a location that could have clouded our model. We also converted “date” into 3 separate variables Any instance that had an outlier of any value was also removed from the data we were working with. To get a general understanding of how the variables in our dataset were related to each other, a correlation plot was created and is shown below. While no individual variables have an obvious correlation with the “score” variable, we can see that the temperature range at 2 meters(T2M_RANGE), the surface pressure(PS), and the maximum temperature at 2 meters(T2M_MAX) have the three strongest correlations with “score”. After running a random forest model to select features, we were able to identify the 15 most important features, shown to the left. In addition to the three mentioned above, other features determined to be important were humidity at 2 meters, dew point at 2 meters, minimum temperature at 2 meters, surface temperature, wind speed at 10 and 50 meters, the range of wind speed at 10 and 50 meters, the maximum wind speed at 50 meters, the year, and the day.

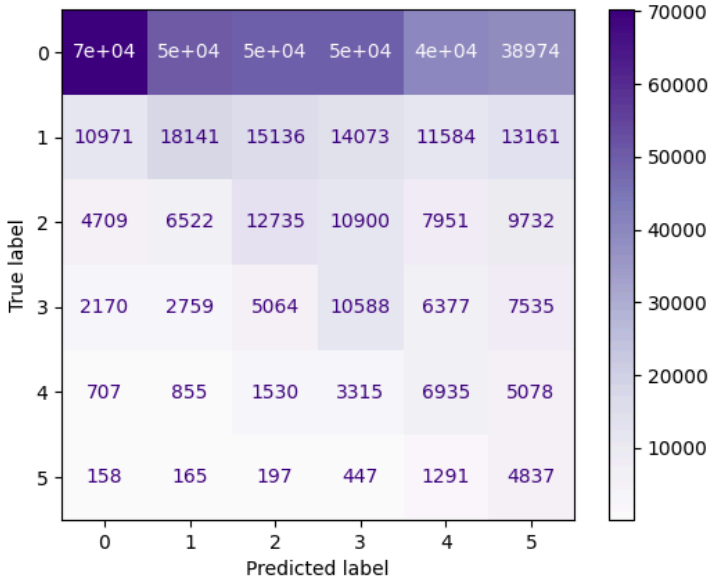
	PRECTOT	PS	QV2M	T2M	T2MDEW	T2MWET	T2M_MAX	T2M_MIN	T2M_RANGE	TS
PRECTOT	1.000000	0.051470	0.284634	0.117007	0.262642	0.262567	0.043930	0.177954	-0.343956	0.113807
PS	0.051470	1.000000	0.258551	0.121002	0.301899	0.301761	0.074383	0.161513	-0.211641	0.116805
QV2M	0.284634	0.258551	1.000000	0.875095	0.962860	0.963820	0.810068	0.911854	-0.058398	0.867656
T2M	0.117007	0.121002	0.875095	1.000000	0.915036	0.915705	0.983863	0.982230	0.261250	0.997745
T2MDEW	0.262642	0.301899	0.962860	0.915036	1.000000	0.999975	0.858324	0.940878	0.005209	0.906540
T2MWET	0.262567	0.301761	0.963820	0.915705	0.999975	1.000000	0.858991	0.941555	0.005357	0.907248
T2M_MAX	0.043930	0.074383	0.810068	0.983863	0.858324	0.858991	1.000000	0.939692	0.421068	0.981023
T2M_MIN	0.177954	0.161513	0.911854	0.982230	0.940878	0.941555	0.939692	1.000000	0.085450	0.979753
T2M_RANGE	-0.343956	-0.211641	-0.058398	0.261250	0.005209	0.005357	0.421068	0.085450	1.000000	0.259544
TS	0.113807	0.116805	0.867656	0.997745	0.906540	0.907248	0.981023	0.979753	0.259544	1.000000
WS10M	0.001778	-0.106831	-0.217262	-0.193542	-0.230004	-0.229845	-0.199604	-0.193177	-0.069167	-0.174931
WS10M_MAX	0.010417	-0.162464	-0.246679	-0.204741	-0.259305	-0.259076	-0.204828	-0.209936	-0.039938	-0.186172
WS10M_MIN	-0.009614	0.012949	-0.101331	-0.113367	-0.106937	-0.106992	-0.125578	-0.104081	-0.089801	-0.098454
WS10M_RANGE	0.020658	-0.227981	-0.262929	-0.198097	-0.276098	-0.275752	-0.189813	-0.211491	0.007924	-0.183316
WS50M	0.014459	-0.060541	-0.201012	-0.185187	-0.198160	-0.198248	-0.183040	-0.192209	-0.023481	-0.172521
WS50M_MAX	0.006348	-0.119919	-0.250720	-0.198463	-0.245111	-0.245136	-0.182207	-0.221578	0.056832	-0.184477
WS50M_MIN	0.033493	0.030308	-0.076152	-0.106887	-0.076230	-0.076387	-0.124655	-0.092348	-0.118228	-0.096901
WS50M_RANGE	-0.025964	-0.184376	-0.243475	-0.145288	-0.236215	-0.236087	-0.106390	-0.189681	0.193101	-0.137547
score	-0.060739	-0.182149	-0.040668	0.098238	-0.045372	-0.044671	0.134755	0.068713	0.210332	0.106373

```
Index(['PS', 'QV2M', 'T2M', 'T2MDEW', 'T2M_MAX', 'T2M_MIN', 'T2M_RANGE', 'TS',
       'WS10M', 'WS10M_RANGE', 'WS50M', 'WS50M_MAX', 'WS50M_RANGE', 'year',
       'day'],
```

Our biggest challenge towards our modeling was the imbalance of the data set. To combat this, we attempted both upsampling and downsampling of the data. With each data set (upsampled and downsampled) we chose to run both KNN and Decision Tree algorithms to attempt to model the data. These four models were then evaluated by accuracy, precision, recall, F1 score, and Cohen Kappa Score. We used the Scikit-Learn package to create these models, and then evaluate them.

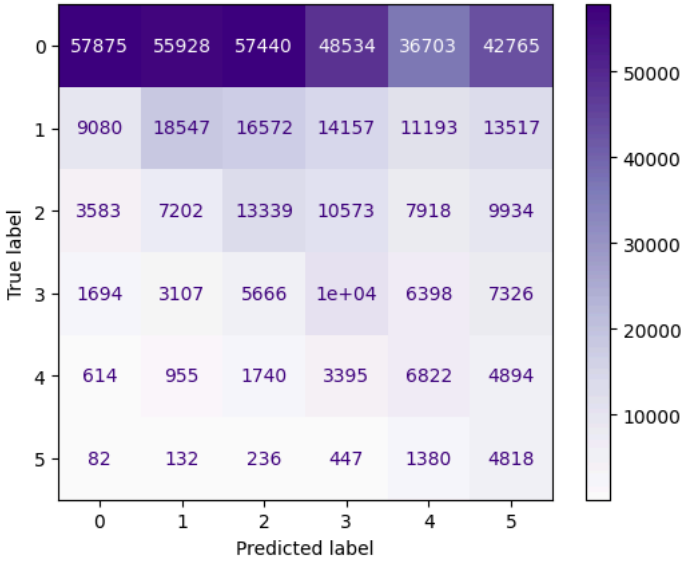
Downsampling involves eliminating a significant portion of the skewed portion of the dataset in order to have an equal number of observations for each value in our target variable. This brought us from a significantly skewed data set to a uniformly balanced dataset, we had 28,807 instances of each drought score. This did not produce a good model at all. The confusion matrices and accuracy scores can be seen below for both the KNN and Decision tree models.

KNN:




```
Accuracy: 0.24948471107446835
Precision: 0.5447049424476852
Recall: 0.24948471107446835
F1 Score: 0.29526079289304746
Cohen Kappa Score: 0.09140799383984988
```

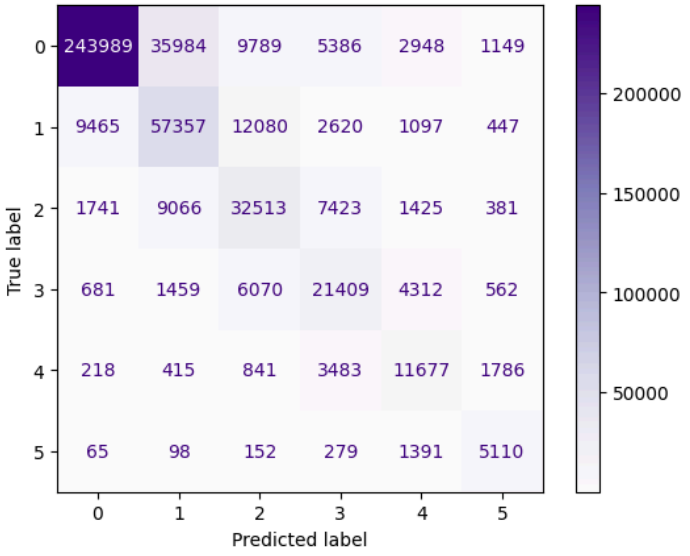
Decision Tree:




```
Accuracy: 0.22572281901436342
Precision: 0.5436981925460196
Recall: 0.22572281901436342
F1 Score: 0.26315149308048763
Cohen Kappa Score: 0.07972816519597348
```

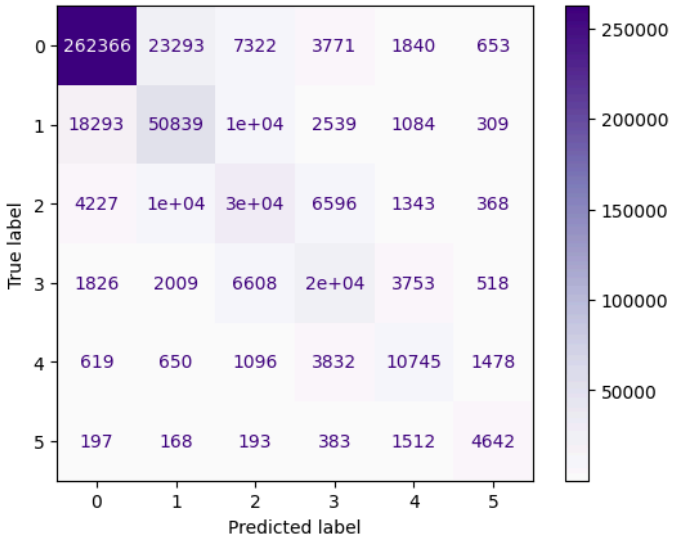
Upsampling balances the data by randomly generating instances for each value of the target until each category has the same number of instances as the most common category. After upsampling we had 1,195,531 instances of each drought score. This vastly improved our model. The confusion matrices and accuracy scores can be seen below for both the KNN and Decision tree models.

KNN:




```
Accuracy: 0.7518267497595318
Precision: 0.7879345908004509
Recall: 0.7518267497595318
F1 Score: 0.7642559970107387
Cohen Kappa Score: 0.6065055697349082
```

Decision Tree:




```
Accuracy: 0.7643957580607353
Precision: 0.7728293195381504
Recall: 0.7643957580607353
F1 Score: 0.7681860155127902
Cohen Kappa Score: 0.6075513413487583
```

The table below shows how much more effective upsampling was than downsampling was with this dataset when creating predictive models.

	Accuracy	Precision	Recall	F1	Cohen-Kappa
KNN Up	0.75	0.79	0.75	0.76	0.61
KNN Down	0.24	0.54	0.25	0.3	0.09
DT Up	0.76	0.77	0.76	0.77	0.61
DT Down	0.23	0.54	0.23	0.26	0.08

An increase in the ability to predict droughts could significantly decrease the impact of droughts in the United States. Droughts are the third most destructive natural phenomenon in terms of financial impact, behind hurricanes and severe storms, and drought events - and the United States National Integrated Drought Information System cites a figure of over $9bn yearly in damages due to droughts. Droughts in 2012 cost $14.5bn, and droughts in California in 2015 caused $1.84bn in damages, a loss of 8.7m acre-feet of water, and over 10,000 jobs. Every significant, and "insignificant" drought has a large impact on both a local and national scale. Preparedness and mitigation of damage are crucial. With the ability to predict droughts based on past events and conditions, local and national government agencies can more efficiently allocate resources with maximum efficiency. While this project is just a start, with accuracy that is not high enough for significant consideration, improvements on the models created in this project could have significant impacts down the line.

Sources

https://www.cdc.gov/nceh/drought/implications.htm#:~:text=Drought%20can%20limit%20the%20growing,of%20livestock%20raised%20for%20food.

https://www.ncei.noaa.gov/access/monitoring/drought-recovery/#:~:text=From%201980%E2%80%932024%20(as%20of,effects%20on%20the%20areas%20impacted.

https://www.drought.gov/sectors/public-health

https://www.drought.gov/sectors/agriculture#:~:text=The%20primary%20direct%20economic%20impact,through%20government%20disaster%20assistance%20programs.