Chicago Climate and Drought Analysis

DTSA 5742 Predicting Extreme Climate Behavior with Machine Learning

February 2025

Abstract

In this study, I explore how other climate and weather variables can help characterize and predict droughts in the Chicago area. The data include air and soil temperature, precipitation amounts, and groundwater levels across a 13-year period from 2011 to 2024. To my final dataset, I apply both supervised and unsupervised machine learning techniques including principal component analysis, K-Means clustering, and logistic regression. While the K-Means clustering only suggests a few variables that could increase the likelihood of a drought being present, my logistic regression model can predict if a drought is declared based on other factors with an accuracy of 85.55%. Overall, this study provides increased insight into the climate trends of the Chicagoland area and highlights climate and weather factors that could be monitored to predict future droughts.

Contents

1 Introduction

Climate change is already negatively impacting the world in a wide range of ways from disturbing our food and water supply to increasing extreme weather events.[1] The Intergovernmental Panel on Climate Change (IPCC) developed a set of scenarios known as Representative Concentration Pathways (RCPs) to quantify the impact of global climate change depending on the actions we take (or do not take) now to reduce carbon emissions and transition to cleaner energy and transportation sources.[2] RCP8.5, which is the closest path to current emissions levels, would result in a massive temperature increase, sea level rise, and increase in extreme weather.[3] This scenario is not a given, and the impact could be reduced if we reduce emissions quickly enough to conform to a less severe RCP pathway, but the potential dangers highlight the need for increased climate research across the globe, especially in localized contexts where impacts could be different than global trends.

While Chicago is not currently a drought-prone area, without implementing policies to reduce carbon emissions and limit global temperature increases, droughts and other extreme weather events could become more frequent or severe in the future. The city has already been facing new patterns of "flash droughts" over the past few years. Chicago had its worse drought in a decade in 2023, having to enforce water restrictions on residents.[4] All of Illinois faced one of the wettest Julys in history followed quickly by another drought in September of last year, which shows how quickly regional weather events can change and negatively impact communities as climate change develops.[5] This analysis explores the climate of the greater Chicago area and how these factors – including precipitation and temperature – contribute to drought conditions, which could help monitor and predict extreme weather events as local and global climates continue to change.

2 Methodology

Note: To make this report more readable, I have excluded the code segments that comprise my data preparation and analysis. To see my notebooks with the actual code, please visit my project's GitHub repository here: https://github.com/zoeh66/5742FinalProject.

2.1 Data

For this analysis, I utilized timeseries data from the National Oceanic and Atmospheric Administration (NOAA) and the United States Geological Survey (USGS). This data covered a range of climate-related variables explored in more detail in the subsections below from early 2011 through mid-2024. The target variable for my analysis indicates if there was a drought happening on a given date using NOAA data from the drought.gov portal. Because Chicago is located on Lake Michigan instead of a desert environment, severe droughts are less frequent, so I instead decided to have this variable indicate the presence of any drought from the D1 "moderate" category through the D4 "exceptional" category.

2.1.1 Air temperature and precipitation

The air temperature and precipitation for this analysis comes from Midway Airport in Chicago, one of the stations that has full coverage for my chosen date range. As shown in the figure below, Chicago has a pretty strong seasonality in temperatures, fluctuating between well below freezing in the winter to consistently hitting 90 degrees in the summer.

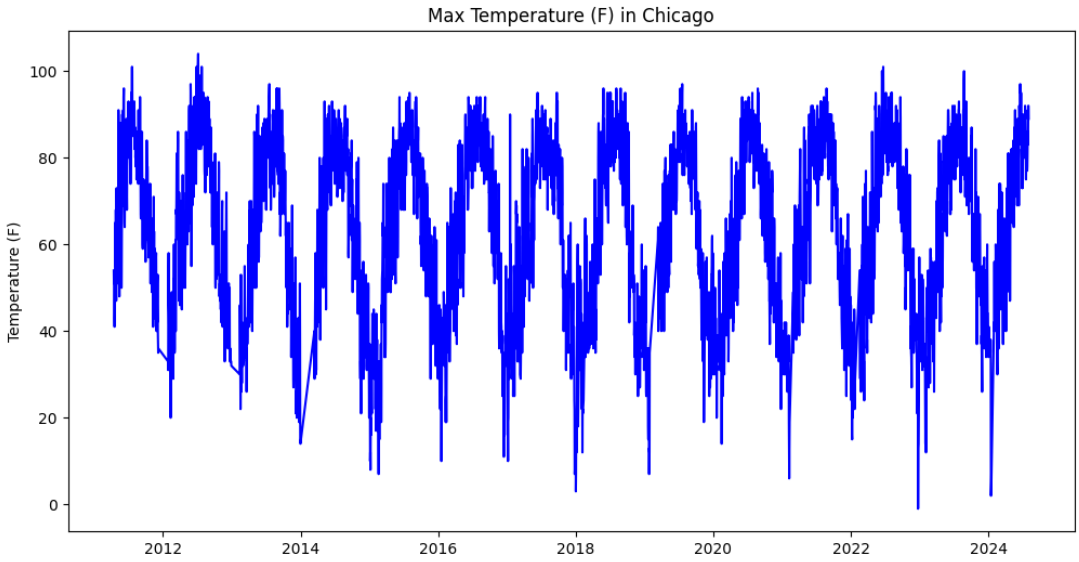

Figure 1: Max Temperature (F) in Chicago

2.1.2 Groundwater levels

The USGS groundwater level data comes from a station in Lake Barrington, a suburb to the northwest of downtown Chicago.

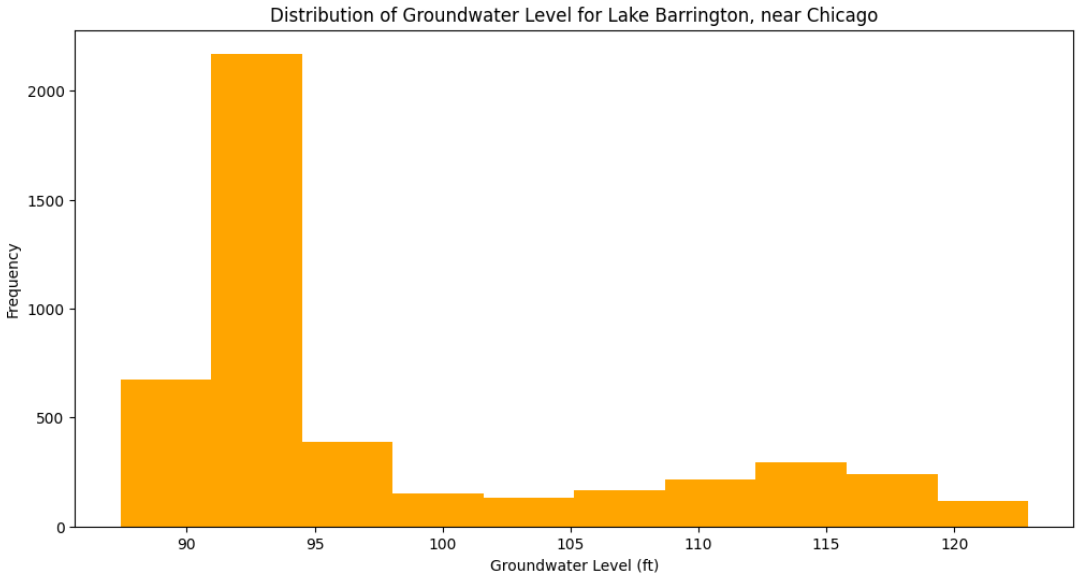

Figure 2: Distribution of Groundwater Level for Lake Barrington, near Chicago

2.1.3 Stream flow

The USGS stream flow data comes from a monitoring location along the northbound Chicago River near Niles, IL. As you can see in the figure below, the Chicago River is a relatively fast flowing river that has a large fluctuation in daily discharge. This dataset is the most complete data available in the region, but there are still spans in the wintertime with missing flow data when the river freezes and it is not possible to record measurements.

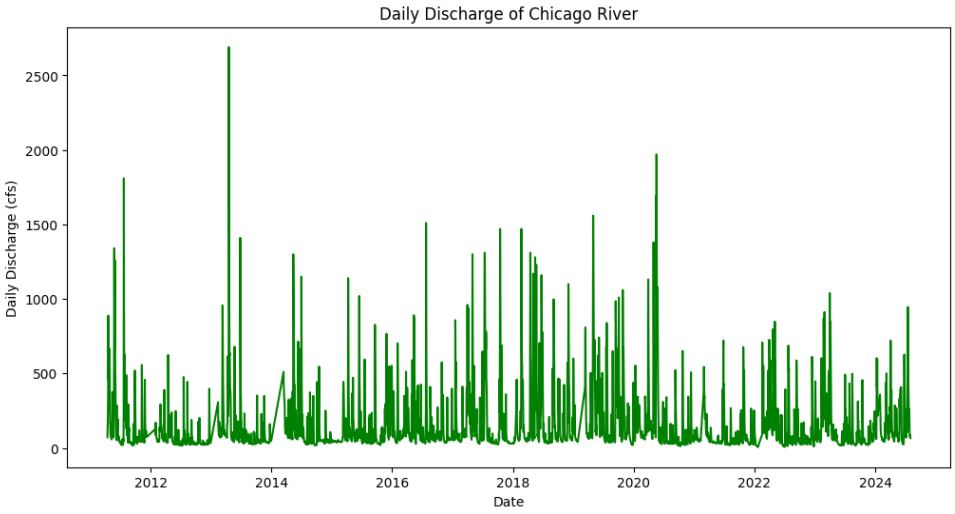

Figure 3: Daily Discharge of Chicago River

2.1.4 Soil moisture and temperature

Unfortunately, soil moisture and temperature data are hard to find near Chicago. The closest site in Shabbona, IL, has long periods of missing data across all variables, so I instead had to use the Champaign, IL site around 130 miles away from Chicago. This station has complete data for soil temperature at 20, 50, and 100 cm, but is missing soil moisture measurements for all depths except 20 cm.

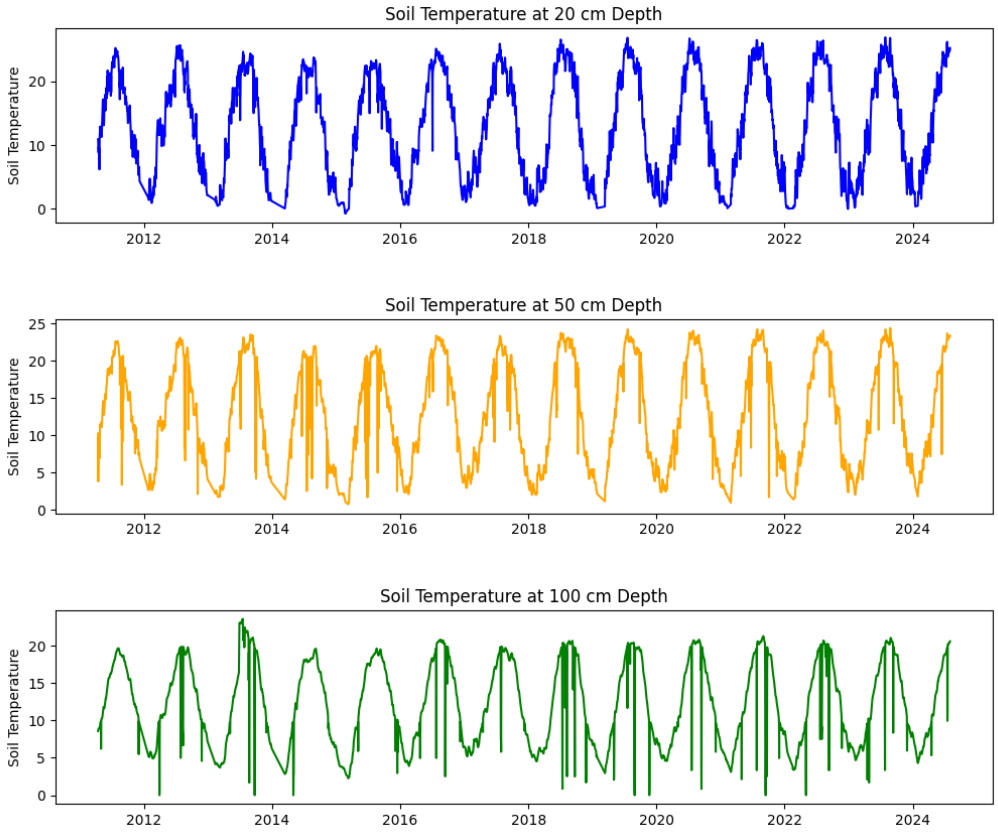

Figure 4: Soil Temperature at 20 cm Depth

2.2 Modeling

2.2.1 Data preparation

To prepare my final dataset for analysis, I imported all the above data from their respective U.S. government sources, cleaned the data, and joined them together. I then saved this dataset in a single Parquet data file for easy access and loaded the complete file into my analysis notebook. All the data is time series data, so I restricted the date range to where all variables have values and dropped any rows with missing observations.

Because there were three variables that all measured soil temperature, which were all highly correlated, I used the unsupervised learning technique Principal Component Analysis (PCA) to reduce the three variables into one principal component that explained over 90% of their variance. I also opted to remove the minimum temperature variable from any analysis since it is highly correlated with the maximum temperature variable. With these changes made, there appears to be no strong correlation between any explanatory variables, so the dataset is ready for modeling.

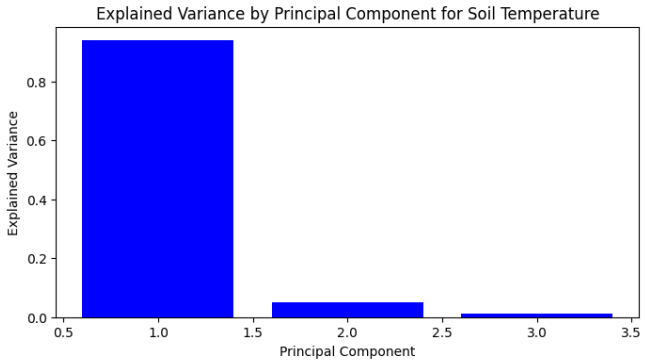

Figure 5: Explained Variance by Principal Component for Soil Temperature

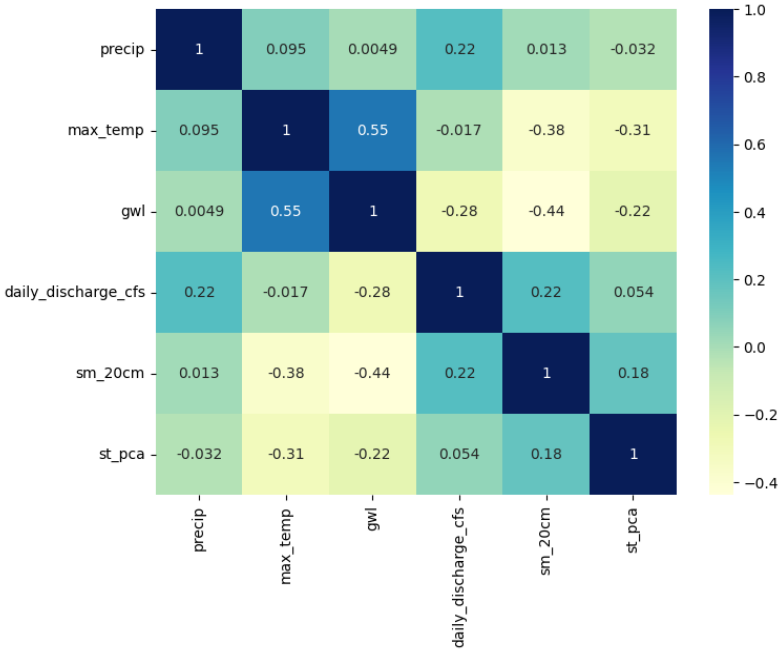

Figure 6: Correlation Matrix of Explanatory Variables

2.2.2 Unsupervised learning

For my unsupervised learning technique, I decided to use K-Means clustering to identify any trends across the explanatory variables. I knew it would be unlikely that there would be a single cluster for drought conditions and one for non-drought conditions, but I thought it could be interesting to see if there are certain clusters of characteristics that make drought conditions more likely. Based on the elbow plot and silhouette score, it appears as though 5 clusters was a decent choice for this dataset.

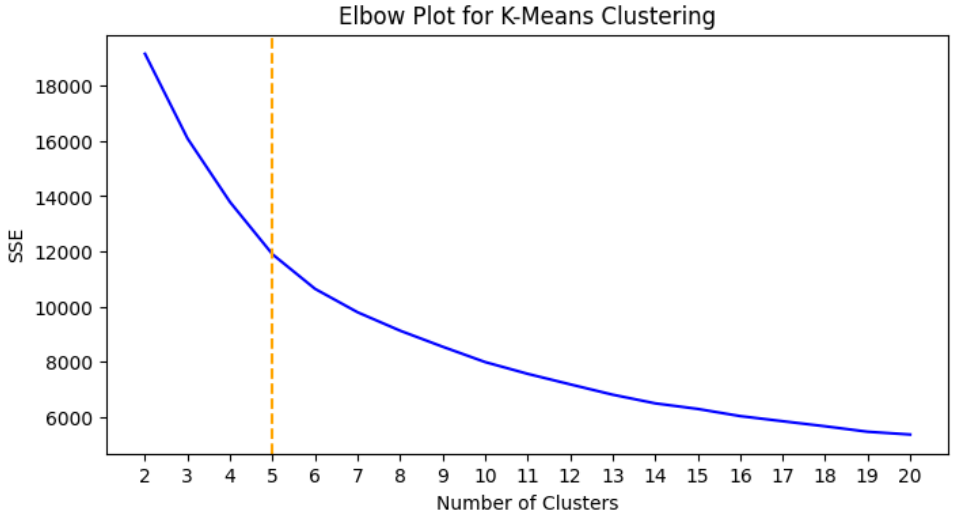

Figure 7: Elbow Plot for K-Means Clustering

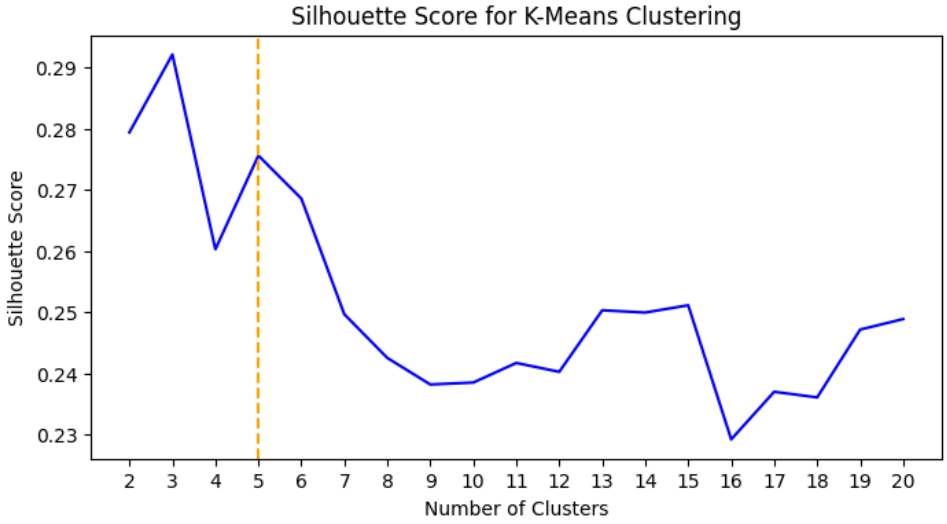

Figure 8: Silhouette Score for K-Means Clustering

2.2.3 Supervised learning

For my supervised learning technique, I chose to do logistic regression. Given that my dataset has multiple feature variables and a binary target (0 for no drought, 1 for drought), I thought this method would be the best choice. After playing around with parameters, I did have to standardize the feature variables to create the model, and I also ended up doing a 80-20 split to create a training and test dataset to validate my model.

3 Results

3.1 K-Means Clustering

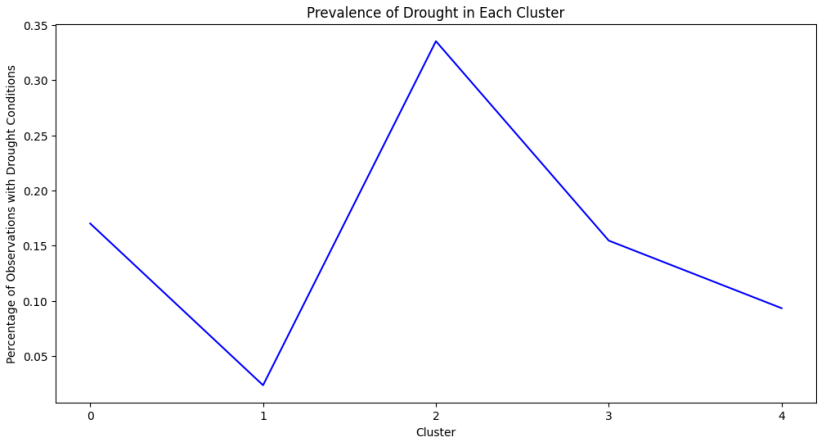

Figure 9: Prevalence of Drought in Each Cluster

After creating my final K-Means clustering model with 5 clusters, there does appear to be a cluster where drought conditions are more likely to be declared. In this instance, about one-third of dates in Cluster 2 have a drought, compared to less than 20% for any other cluster. Looking at the parallel coordinates chart below, this cluster has a higher maximum temperature and higher depth to groundwater level, which makes sense as both could indicate a hot, dry spell usually associated with drought.

Cluster 1, which has a very low rate of drought amongst its observations compared to the other clusters, is associated with a very high daily discharge measurement. This also makes sense intuitively as there is not likely to be a drought when there is a lot of water moving quickly through local waterways. However, Cluster 0 has a high percentage of observations with a declared drought despite having much higher precipitation amounts than all other clusters, which indicates that this model might not be completely accurate despite having some interesting observations.

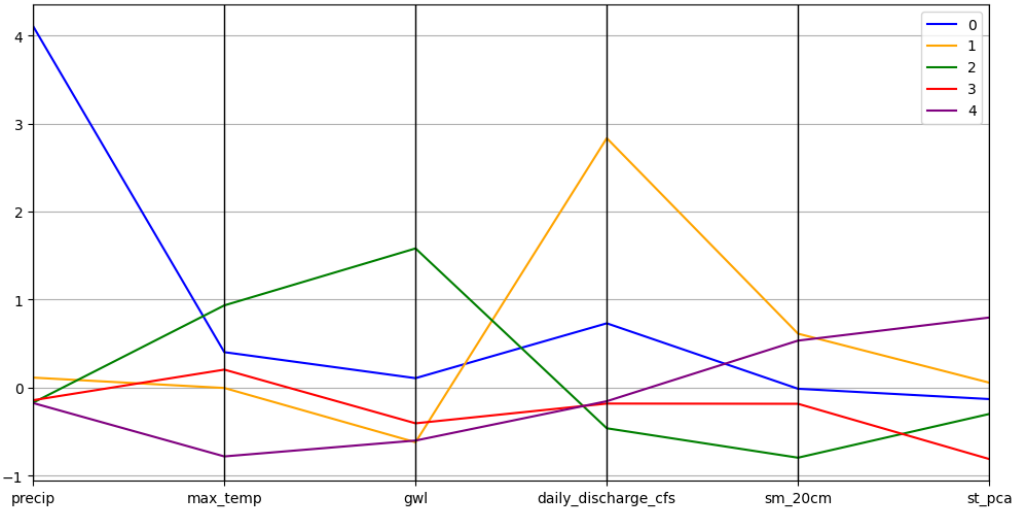

Figure 10: Parallel Coordinates of Cluster Centroids Across Features

3.2 Logistic Regression

In comparison with K-Means clustering, which simply provided some interesting insight into the variables when compared to the prevalence of drought, my logistic regression model was trained with the intention of predicting whether there would be drought conditions based on the other variables. Despite not performing much feature engineering or parameter tuning, this model actually performed quite well for this data. The R^2 value was 84.9%, and the groundwater level having the largest impact on the probability of drought with a coefficient of 0.89. The model also had an 85.55% accuracy predicting drought conditions on the test dataset. However, there were a lot of observations with drought conditions declared where my model falsely predicted that there was no drought, so there is definitely still room for improvement.

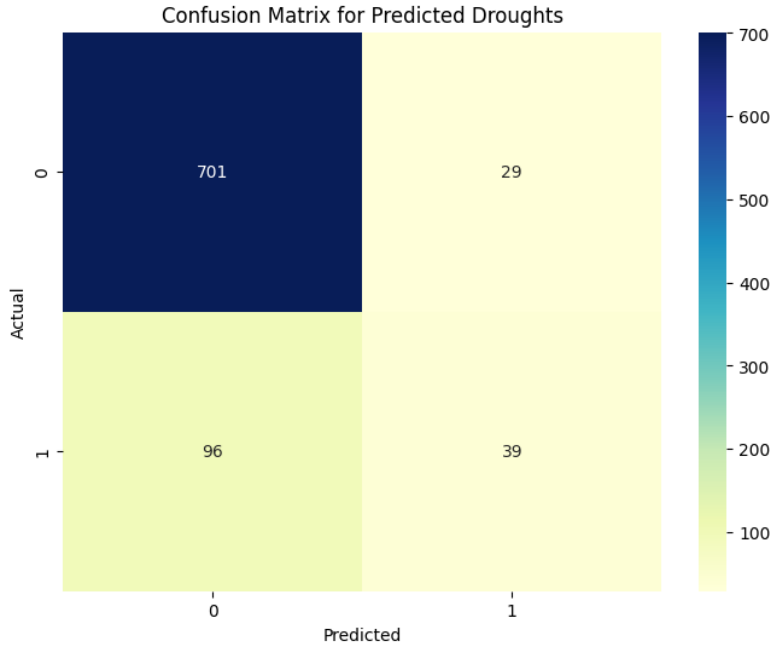

Figure 11: Confusion Matrix for Predicted Droughts

4 Conclusion

Overall, the logistic regression model provided relatively good accuracy for predicting the presence of a drought with the variables I chose, having a test accuracy of 85.55% with an R^2 value of 84.9%. While the K-Means clustering model was not able to identify a single drought cluster, it provided some interesting insight into which weather and climate variables could increase the chance of a drought occurring.

Although my analysis did provide some interesting results, there were limitations to this project and future improvements that could be made. The largest limitation for this project was data availability for the region. I wanted to focus closely on Chicago, but the lack of data available especially for a long-term date range meant that I had to use some sources that were a way away from the city itself, which could have negative implications for my results. Further research could try to identify other variables that could improve my models and extend the scope of this analysis to try to forecast how these variables and their relationships may be impacted by climate change in the future. Another future step could be to see how a regression model would do at predicting the severity of a drought, not just the presence of one.

5 Data and Code Availability

My data preparation notebook, final parquet data file, and data analysis notebook are all available at the GitHub repository here: https://github.com/zoeh66/5742FinalProject.

6 References

[1] National Oceanic and Atmospheric Administration. "Climate Change Impacts." Accessed February 24, 2025. https://www.noaa.gov/education/resource-collections/climate/climate-change-impacts.

[2] Department of the Environment and Energy. "What are the RCPs?" Australian Government. PDF Accessed February 24, 2025.

[3] Hausfather, Zeke. "Explainer: The High-emissions 'RCP8.5' Global Warming Scenario." CarbonBrief. August 21, 2019. Accessed February 24, 2025. https://www.carbonbrief.org/explainer-the-high-emissions-rcp8-5-global-warming-scenario/.

[4] White, Robyn. "Water Restricted as Chicago Suffers Worst Drought in a Decade." Newsweek. June 22, 2023. Accessed February 24, 2025. https://www.newsweek.com/water-restrictions-illinois-drought-1808492.

[5] Ford, Trent. "Drought is Back in Illinois." Illinois State Climatologist, University of Illinois. September 12, 2024. Accessed February 24, 2025. https://stateclimatologist.web.illinois.edu/2024/09/12/drought-is-back-in-illinois/.