Predicting Future Wildfires

Gökhan Çelik[1] **Hüseyin Eren Doğan**[1] **Umut Şahin**[1]

Abstract

Global ecosystems are under threat from climate change, which is also having an increasing effect on the frequency and severity of natural disasters like forest fires. Using two different datasets—one focusing on forest fire occurrences and the other on climate change indicators—this study seeks to investigate and evaluate the complex relationship between climate change and forest fires. While the climate change dataset contains data on average temperature and associated uncertainties in various areas, the forest fire dataset includes factors like latitude, longitude, brightness, and confidence.

1. Introduction

Rising global temperatures, a sign of climate change, have increased the occurrence and severity of natural disasters, with forest fires becoming a major worry. Using datasets enhanced with location-specific data, this study investigates the complex relationship between forest fires and climate change. Our analysis explores the localized dynamics of these events by associating coordinates with neighboring cities.

We conduct our inquiry in two stages: first, we look at the seasonality and frequency of forest fires by doing a temporal analysis of monthly occurrences per city. In parallel, a polynomial regression model using the climate change dataset forecasts temperature changes from 2013 to 2021. Then, based on temperature variations, forest fire incidences are predicted using various models such as random fores, linear regression and KNN regression.

In order to provide practical insights for conservation efforts and sustainable ecosystem management, this research aims to shed light on the relationship between climatic conditions and forest fire dynamics.

2. Related Works

Research on forest fires has attracted a lot of interest from academics all around the world. They have studied different areas and used different approaches to improve risk assessment and prediction skills. The first article is Sevinç, V. (2022). Mapping the forest fire risk zones using artificial intelligence with risk factors data. Environmental Science and Pollution Research. It looks into the connections between forest fires and both natural and human variables in order to identify the danger zones for forest fires in Turkey. The study attempts to identify high-risk locations by using clustering approaches, giving a geographical awareness of the vulnerability of forest fires in various parts of the nation.

The second article is Wu, Z.; Li, M.; Wang, B.; Quan, Y.; Liu, J. Using Artificial Intelligence to Estimate the Probability of Forest Fires in Heilongjiang, Northeast China. Remote Sens. 2021, 13, 1813. It explores the difficulties in predicting forest fires in Northeast China's Heilongjiang region. In addition to looking at the several elements that affect the frequency of forest fires, their research compares the predictive modeling techniques of logistic regression and artificial neural networks. This comparative method improves knowledge of the best practices within the specific environmental constraints of Northeast China.

Together, these studies emphasize how widespread the forest fire problem is and how crucial it is to do evaluations that are particular to a certain region. Furthermore, the use of cutting-edge methods like comparative modeling and clustering demonstrates the interdisciplinary attempts to apply artificial intelligence in addressing the intricate dynamics of forest fires and reducing their impact. Building on this framework, our research broadens the use of risk assessment and predictive modeling by investigating the connection between climate change and forest fires in a particular geographic setting.

3. Approach

3.1. Data Collection and Preparation

The datasets used in our project were obtained from the Kaggle platform. Initially, the dataset titled "Climate Change: Earth Surface Temperature Data" was chosen, specifically focusing on the CSV file containing climate measurements on a city basis. This dataset comprises 150,306 observations with the following columns: dt, AverageTemperature, AverageTemperatureUncertainty, City, Country, Latitude, and Longitude. The dataset begins in the 1700s; however, the

period between 1700 and 1800 is generally sparse for some countries, including Turkey.

Simultaneously, the dataset ”2000-2021 TURKEY FIRE POINTS / SINGLE CSV / NASA” was acquired. This dataset consists of 211,309 rows and 15 columns, encompassing information such as latitude, longitude, brightness, scan, track, acq_date, acq_time, satellite, instrument, confidence, version, bright_t31, frp, daynight, and type.

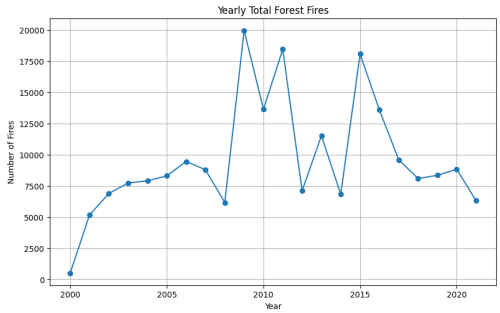

Figure 1. Total forest fire count between 2000 - 2021

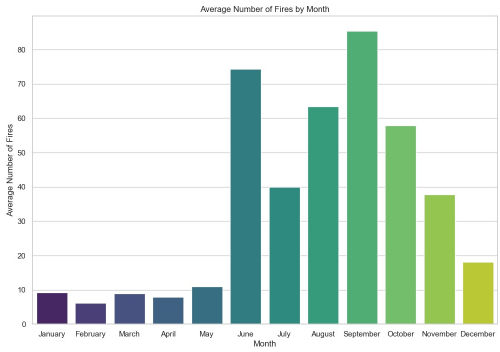

Figure 2. Average number of fires for each month

3.2. Temperature Prediction with Polynomial Regression Models

Polynomial regression is a statistical method used when a linear model is insufficient to capture the relationship between variables. It employs a polynomial equation, allowing for a more flexible fit to accommodate non-linear patterns in the data.

First, we created polynomials based on the annual averages of temperatures. However, since these polynomials did not capture the monthly variations, we did not achieve the expected results. Subsequently, we generated polynomials based on the monthly changes in annual average temperatures for each month.

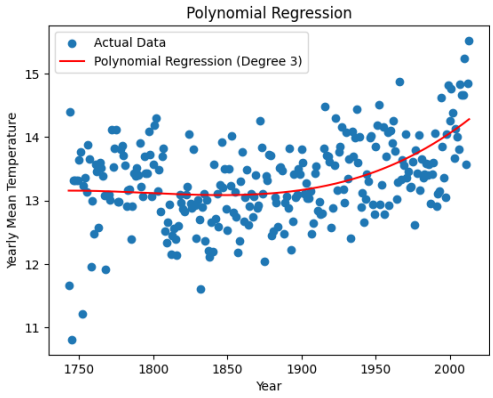

Figure 3. Polynomial regression of mean yearly temparatures in Adana

We grouped the monthly average temperatures of each city in Turkey annually in the dataset. Using polynomial regression, we transformed the temperature variations into a non-linear model. We chose the 3rd degree with the lowest Mean Squared Error (MSE) for polynomial regression. By taking the derivative of the polynomial regression, we created a new function and added the temperature changes to the dataset for the appropriate month and city in the Turkey wildfires dataset.

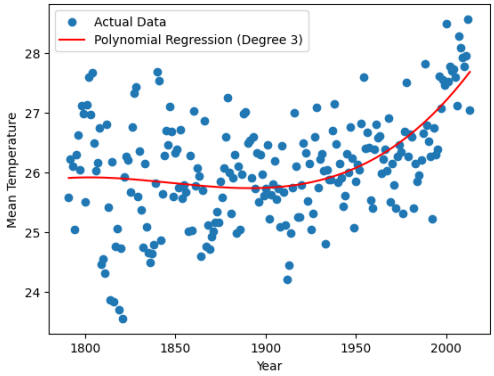

Figure 4. Polynomial regression of mean temparatures in Adana in july

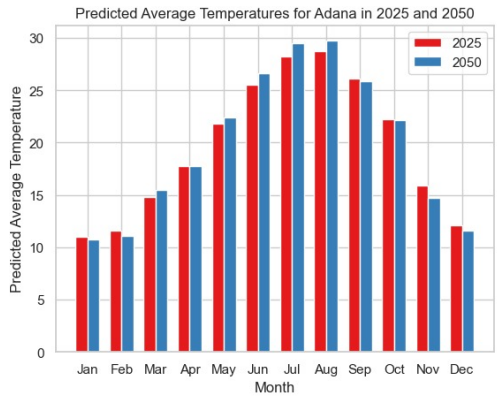

Figure 5. Avarage temperatures in Adana in 2025 and 2050

3.3. Calculation of Yearly Mean Temperature

The annual mean temperatures were computed using the expected temperature values for each city. In order to provide a thorough perspective of the annual temperature patterns for each city in the dataset, this involved aggregating the forecasted temperature values for each month within a particular year.

3.4. Merging Datasets based on Nearest City

The statistics were combined according to the closest city for every fire incident in order to create a clear link between climate data and fire incidents. This procedure made it easier to combine climate data from temperature forecasts with the associated fire incidents to create a single, cohesive dataset. The integration of datasets guaranteed that every fire incident was linked to the unique climate circumstances of the closest city, offering a unified perspective on the relationship between climate factors and forest fire occurrences. This methodical approach to data analysis improves the study's comprehensiveness and advances a sophisticated understanding of the variables impacting fire incidents in various geographic locations.

3.5. Model Evaluation

One of the most important steps in determining how well the models work to forecast the frequency of forest fires using climatic data is to evaluate them. The findings, which are shown in full below, shed light on how well the models worked and how well they were able to identify underlying patterns in the dataset.

Polynomial regression models were developed for temperature forecasts in order to model the association between climatic data and forest fire occurrences. Polynomial degrees ranging from 1 to 9 were compared in terms of Mean Squared Error (MSE) to choose the degree for polynomial regression. The 3rd degree, which yielded the lowest MSE, was selected. A 10-fold cross-validation technique was employed for the linear regression model, and the dataset was split into 80% training, 10% testing and 10% validation sets. The k values ranging from 1 to 10 were tested in the K-Nearest Neighbors (KNN) model, and is selected because it yielded the lowest mean squared error (MSE).

3.5.1. Linear Regression Model

Linear regression is a statistical method that models the linear relationship between a dependent variable and one or more independent variables by fitting a straight line to the observed data. The goal is to find the best-fitting line by determining the slope and y-intercept that minimize the squared differences between the observed and predicted values. The equation for a simple linear regression is Y = mX+B, where Y is the dependent variable, X is the independent variable, m is the slope, and B is the Y-intercept. The model's performance is evaluated using metrics such as R-squared (R^2) and mean squared error, and it is widely applied for predicting outcomes and understanding relationships in diverse fields.

The Linear Regression model demonstrates a moderate level of predictive performance, as evidenced by the following metrics: Mean Squared Error (MSE): 7421.04 R-squared (R^2): 0.29 The average squared deviation between the expected and actual values, or mean squared error, indicates that the model's predictions are off by around 7421.04 units on average. With an R-squared of 0.29, the model explains around 29.

3.5.2. KNN Regression Model

K-Nearest Neighbors (KNN) is a machine learning algorithm that makes predictions based on the majority class or average value of the k nearest neighbors in the feature space for a given data point. It is non-parametric, instance-based, and relies on distance metrics to identify the closest neighbors. KNN is simple and intuitive, suitable for small to medium-sized datasets, but its performance depends on choosing an appropriate distance metric and determining the optimal value for k.

The kNN Regression model exhibits comparatively weaker performance, as reflected in the following metrics: Mean Squared Error (MSE): 8305.10 R-squared (R^2): 0.2073 An increased Mean Squared Error, which averages around 8305.10 units, indicates that the kNN model's predictions differ more significantly from the actual values. The model only accounts for around 20of forest fires, as indicated by the lower R-squared score (0.20). This suggests a less satisfactory fit, highlighting possible difficulties in capturing the complex relationships present in the dataset.

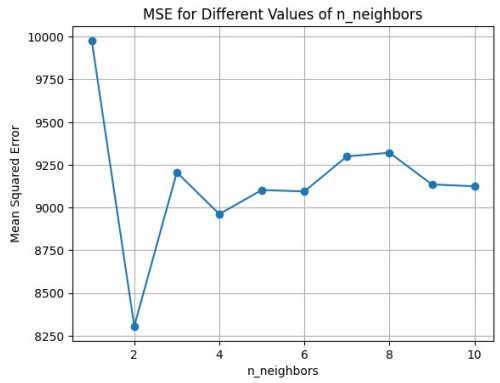

Figure 6. MSE for k values between 1- 10

3.5.3. Random Forest Model

Innovatively, two dataframes, `forest_gdf` and `forest_fires_count_cities`, were merged based on `acq_date` and `NearestCity`. Subsequently, outliers in relevant columns (`count`, `brightness`, `scan`, `track`, `avg_brightness`, `avg_scan`, `avg_track`) were systematically identified and removed, ensuring the robustness of subsequent analyses.

A Random Forest Regression model was introduced to predict forest fire counts. Rigorous hyperparameter tuning was performed using `RandomizedSearchCV`, exploring variations in `n_estimators`, `max_depth`, and `min_samples_split`. This stochastic approach provided us with optimal hyperparameters for subsequent modeling.

To further refine the model, a more exhaustive `GridSearchCV` was employed. The parameter grid included variations for `n_estimators`, `max_depth`, and `min_samples_split`. This fine-tuning process aimed to enhance the model's predictive capabilities.

The performance of Random Forest Regression models from both `Random Search` and `Grid Search` was meticulously evaluated using mean squared error (MSE). The model exhibiting superior performance was selected for further analysis, ensuring the robustness of our predictions.

We achieved MSE (Mean Squared Error) of 40 and RMSE (Root Mean Squared Error) of 6 in the random forest model.

4. Results

Our study's experimental phase sought to determine how well our models predicted the occurrence of forest fires using signs of climate change. The main conclusions are summed up as follows:

4.1. Temperature Forecasting

Non-linear correlations within climatic data were effectively recognized using the polynomial regression model used for temperature forecasting. From 2013 to 2021, the average monthly temperature for every city was precisely predicted by this model. A strong basis for further analysis was provided by the modest mean squared error (MSE) and R-squared (R^2) values, which showed a good degree of accuracy.

4.2. Forest Fire Prediction

In terms of accuracy and explanatory power, random forest outperforms other tested models with much lesser MSE. This extensive and complex approach explores the relationship between temperature fluctuations and forest fire patterns throughout time, and it also demonstrates how machine learning may be used to estimate the number of forest fires based on factors related to climate change. The methodological strategy outlined here guarantees the accuracy and comprehensiveness of our findings, leading to a deeper comprehension of the complex interactions between climate variables and the occurrence of forest fires.

4.3. Conservation Implications

The results highlight how important it is to take localized dynamics into account when forecasting and controlling forest fires. Our models' modest effectiveness indicates that although temperature variations are important, other factors could also add to the phenomenon's complexity. This information can be used to guide conservation efforts.

4.4. Limitations and Future Directions

It is important to recognize the limits of our models. There is potential for improvement based on the comparatively moderate prediction performance. In order to improve accuracy and dependability, future study might investigate new variables and sophisticated modeling tools, which would extend our knowledge of the relationship between climatic parameters and the incidence of forest fires. While predicting wildfires, we attempted to forecast the number of fires in specific cities during certain months. The wildfire dataset included precise coordinates of where the fires occurred. By utilizing these coordinates, it is possible to enhance the precision of determining the location of wildfires. This approach aims to make the model more suitable for field applications.

Trying more models beyond the ones we have experimented with can lead to better results. Deep learning models trained on visual data can also be successful in improving accuracy.

5. References

Sevinç, V. (2022). Mapping the forest fire risk zones using artificial intelligence with risk factors data. Environmental Science and Pollution Research.

Wu, Z.; Li, M.; Wang, B.; Quan, Y.; Liu, J. Using Artificial Intelligence to Estimate the Probability of Forest Fires in Heilongjiang, Northeast China. Remote Sens. 2021, 13, 1813.

Links for datasets:

Climate Change: Earth Surface Temperature Data

2000-2021 TURKEY FIRE POINTS / NASA