Project Proposal: A Chained Model for Weather Forecasting and Wildfire Risk Prediction

2491015 방가은, 2491017 백규리, 2491025 이지원, 2491029 정아현

1. Business Value

1.1 Background

Recently, large-scale forest fires have occurred in succession in the United States and Korea. Global warming is creating an environment where wildfires can occur and spread easily. In particular, the forest fire that occurred in Gyeongbuk Province last March spread quickly and caused the largest amount of damage ever. Looking at these cases, we needed to predict the possibility of forest fires occurring in Korea in advance and prepare systematically.

1.2 Problem Definition & Value Proposition

We judged that an analysis focused on a specific area rather than the entire country would produce more effective prediction results. Accordingly, this project will focus on the Gyeongbuk region, where large-scale forest fire damage recently occurred. Our goal is to build a chained model that first predicts meteorological future conditions in the Gyeongbuk region, and calculates the probability of forest fire occurrence based on this. This model is expected to contribute to establishing a regionally tailored disaster management system and serve as a foundation for operating a national forest fire response system more efficiently in the future.

2. Available data

2.1 Data Acquisition

The first dataset was sourced from the KMA Weather Data Service and covers the Gyeongsangbuk-do region of South Korea from August 1938 to April 2025. It includes monthly climate variables such as average maximum temperature, average minimum temperature, rainfall, average wind speed, and wind direction, according to each location within the Gyeongsangbuk-do. The second dataset was sourced from Zenodo and provides a comprehensive compilation of weather observations and wildfire occurrences in California from 1984 to 2025, including maximum temperature, minimum temperature, temperature range, average wind speed, precipitation and wildfire occurrence status.

2.2 Data Review

Both datasets (8,029 and 14,988 rows, respectively) show low rates of missing value, recording 0% or under 0.2%, which will be removed in data preprocessing. Shared variables (maximum temperature, minimum temperature, average wind speed, and precipitation) were selected for modeling. Additionally, year and month variables are selected from first dataset, while wildfire occurrence (True/False) and temperature range are selected from second dataset. In first dataset, teperature range variable will be created using the maximum and minimum temperatures of itself, and it will be used as a shared wheather feature. Since the first dataset covers multiple locations within the Gyeongsangbuk-do, the climate variables will be averaged, so at each point in time (i.e., year and month), only one integrated set of climate data will be used.

3. Formulation

3.1 Choice of Algorithms and Rationale

In Model1, relatively irregular weather conditions must be carefully analyzed to accurately predict future climate. Therefore, we apply Gradient Boosting, which offers high prediction accuracy, to ca pture and model complex weather patterns in detail. For Model2, we use Random Forest Classifier. In this stage, the errors and uncertainties from Model 1's outputs can be effectively mitigated, thereby enhancing the stability and reliability of the wildfire prediction. In particular, it helps prevent overfitting and improves prediction accuracy through the voting mechanism of mult iple decision trees.

3.2 Input & Output Variables

This project adopts a dual-model architecture. The first model predicts future weather conditions in the Gyeongbuk region—specifically temperature, wind speed, and precipitation—based on user-provided inputs of year and month. The output from the first model is then used as input for the second model, which predicts the probability of wildfire occurrence at the given time. To utilize seasonal information as a derived feature, the month is first encoded as a categorical variable in the first model. In the second model, this encoded month is then converted into a season variable, which is included as a categorical input.

3.3 Expected Challenges and Alternatives

This project aims to predict wildfire occurrences in the Gyeongbuk region by integrating and analyzing two meteorological datasets with differing formats. To this end, discrepancies in precipitation units, differences in input variable formats, and data quality issues will be addressed through a comprehensive preprocessing stage. First, since Model 1 uses "month" as an input variable and Model 2 uses "season," an external function will be implemented to convert month values into their corresponding seasons, thereby standardizing the input format across models. Second, date values will be separated into year and month columns, categorical variables will be numerically encoded using one-hot encoding, and missing or anomalous values will be handled through techniques such as mean imputation or removal.