An interactive visualization platform for California Wildfire

CSE6242 Team 6: Stephanie Cho, Dhyan Gandhi, Wenjia Hu, Huanci Yang, Elisa Yuen Ting Ching

I. Introduction

Wildfire causes significant casualties and long-term economic and social consequences. With climate change intensifying fire risks, predictive modeling is crucial for mitigation and preparedness.

This project aims to develop an advanced visualization dashboard that displays the probability of wildfires and assesses potential damage across California, enhancing both public understanding and emergency response strategies. By leveraging machine learning, we will identify key environmental and meteorological factors influencing wildfire occurrences and estimate their impact.

II. Problem Definition

We will visualize the predictive model, wildfire trends, and confidence levels across the U.S. to enhance understanding of their spatial and temporal distribution. These insights will support emergency response planning, enabling proactive resource allocation and early warning systems. By demonstrating wildfire risk prediction, this project aims to enhance preparedness, minimize damage, and contribute to more effective wildfire management strategies in California and beyond.

We will use three datasets: 1) California Environmental Conditions (OpenML/Originally sourced from CIMIS, 128,126 observations) – includes meteorological factors like temperature, humidity, and wind speed; 2) California Fire Incident Data (Department of Forestry and Fire Protection, 1,647 observations) – provides historical fire incidents and damage records, and 3) NASA's Satellite Data on U.S. Wildfire Incidents (1,291,385 observations) – offers real-time fire detection and trends.

III. Literature Survey

Data and visual analytics are increasingly vital in wildfire research. While earlier studies focused on statistical models to forecast wildfire events and sizes [1], [2], recent advancements in machine learning (ML) have enabled its wide application in wildfire analysis, including fire detection, behavior prediction, and effects [3]. Researchers have trained classifiers using traditional ML algorithms like Decision Tree, Random Forest, Naive Bayes, and advanced Neural Networks to identify fire occurrences [4], [5]. Tree-based models like Random Forest have shown strong performance in predicting wildfires with high accuracy [6], [7], while some studies explore ML for predicting wildfire damages [8].

The research benefits this project in several ways: 1) it provides useful wildfire datasets, including California Environmental Conditions to be used in this project; 2) it guides feature selection by incorporating both fire and environmental data; 3) it helps establish a baseline for model training; and 4) it offers insights that can be integrated with visualization. However, existing fire prediction models are limited by binary classification, narrow feature sets, and ineffective visualizations, reducing their predictive depth and social impact. Our project addresses these gaps with multi-class classification, diverse datasets, and interactive visualizations to enhance accuracy and decision-making.

Another research focus is using visualization to improve fire decision-making. Chew et al. [9] developed user-friendly platforms for dynamic fire trend visualization, while forestry research used 3D models to simulate fire occurrence, spread, and extinguishing [10]. Researchers have analyzed existing studies on wildfire visualization to compare their effects, such as immersive vs. non-immersive and map-based vs. text-based approaches [11], [12]. Brewer et al. [13] explored the method of classification in choropleth maps, and Cromley et al. [14] proposed cumulative frequency legends to enhance map communication. Guldåker [15] examined the effects of combining point data, kernel density, and choropleth mapping.

These papers are valuable as they demonstrate the dynamic presentation of wildfires using various visualization techniques, comparing their advantages, limitations, and common use cases. They also highlight key considerations such as user-friendliness and the ease of fire information interpretation.

However, a major drawback is the inadequacy of more advanced data analysis in these approaches. Our improvement involves integrating these visualization techniques with ML model outputs to enhance data-driven insights and improve the effectiveness of information delivery.

IV. Proposed Method

1. Data Cleansing & Preprocessing: We performed data preprocessing and cleaning on the CA_Fire_Incidents_Data and CA_Env_Data. The cleansing pipeline included importing data, checking basic information, examining missing values, unifying data types and formats, selecting features relevant to modeling, and conducting basic feature engineering. For the CA_Fire_Incidents_Data, we derived the feature "Note" by splitting the feature "CanonicalUrl" which is the unique identifier for each wildfire incident, and extracting the last part as a description of the fire. We also created the feature "Duration_Hours" by subtracting the fire start time from the extinguished time.

We applied different missing value strategies for the two datasets. For the CA_Fire_Incidents_Data, we noticed that the percentage of missing values is significantly higher for certain fields, such as 98.65% for "Fatalities, " 95.74% for "StructuresDamaged", and 94.72% for "Helicopters". However, we didn't simply drop these features as they are important measures of the impact of wildfire incidents. Such extensively high proportions of missing values are likely due to the fact that most wildfires extinguished naturally or occurred far from human residential areas, and thus were not properly recorded. In this case, it would cause bias to use mean/mode imputation for these missing values, as these incidents are likely to have no casualties or structural damage; these fields were typically filled in for "MajorIncidents", which we found highly correlated with "Fatalities" and "StructuresDamaged". Therefore, we created binary indicators for missing values and filled missing fields with 0, preserving information about absence versus true zero. In contrast, CA_Env_Data had very few missing entries (less than 0.02%). As a result, we simply dropped the rows with missing values to maintain a clean dataset with minimal effect on its size. This tailored cleaning strategy ensured data integrity and modeled relevance across both datasets.

2. Machine Learning: Our goal was to predict the probability of wildfire occurrences in California using environmental conditions. To achieve this, we used data containing environmental factors during the wildfire such as temperature, humidity, and wind speed.

- **The dataset was highly imbalanced (4,000 wildfire days vs. 116,000 non-wildfire days):** To address this, we applied SMOTE (Synthetic Minority Over-sampling Technique) to oversample the minority class in the training data. Additionally, we used scale_pos_weight in XGBoost to penalize misclassification of minority class during training.
- **Adding geographical and time data:** We further preprocessed the data to encode location and temporal information. From timestamps, we extracted the *month* as a feature. We also encoded *station_name* to preserve geographical context.
- **Model Selection & Training:** We selected XGBoost, a gradient boosting decision tree algorithm. We chose to use this since it can natively handle missing values, and is robust against overfitting. After this, our next step was to train an initial model with default parameters, then improve it.
- **Bayesian Hyperparameter Optimization:** We used BayesSearchCV from scikit-optimize to automatically tune: *max_depth, learning_rate, n_estimators, scale_pos_weight, subsample, colsample_bytree.* The optimization was guided by F1 score to balance precision and recall. Our optimal parameters were *max_depth = 6, learning_rate = 0.3, n_estimators = 300, scale_pos_weight = 20, subsample = 0.8,* and *colsample_bytree = 1*

- **Threshold Optimization:** The default threshold (0.5) was suboptimal due to class imbalance. To address this, we used precision-recall (PR) curve to determine an optimal threshold dynamically (between 0.3-0.7). We also avoided extreme thresholds which could lead to imbalanced predictions.

3. Visualization: We created three interactive Tableau dashboards to visualize our analytical insights and machine learning results.

(1) U.S. Wildfire Dashboard (Based on NASA Data) - This interactive dashboard visualizes wildfire trends across the U.S. using NASA satellite data from 2014 to 2023. It includes four interconnected parts: 1) Density Map: Plots each fire occurrences and groups overlapping points. Larger and darker areas show concentrated fire occurrences. Users can hover to view detailed information (state, coordinates, brightness, and Fire Radiative Power) and use drag-and-zoom to explore more. 2) Choropleth Map: Shows state-level wildfire intensity over 10 years, with darker shades representing more fire occurrences. 3) Heatmap: Displays average monthly Fire Radiative Power, where darker and larger rectangles indicate higher fire severity. 4) Area Chart: Shows monthly trends in total wildfire occurrences over the 10 years. Selecting a state of the choropleth map will update the density map, heatmap, and area chart to reveal wildfire patterns specific to the selected state.

(2) California Wildfire Incidents Impact Dashboard – This dashboard explores the spatial distribution of wildfire incidents in California and analyzes the associated damage and resource involvement. We first created a density map and a bubble map to visualize wildfire incidents between 2013 and 2019. The density map highlights areas with frequent incidents. The bubble map illustrates fire severity from three perspectives: 1) acres burned – shown by bubble size; 2) major incidents (with fatalities/injuries) – shown by shape (cross vs. circle); and 3) population density – shown by color intensity. These two maps help identify not only where wildfires were most frequent and destructive, but also whether they occurred in densely populated regions. To further investigate fire impact over time, we built a county-level bubble chart. Bubble size represents the number of injuries and fatalities, while color indicates structures damaged or destroyed. This chart also allows us to assess whether fire severity aligns with acres burned. Lastly, we conducted a trend analysis comparing standardized metrics for fire damage (acres burned, structures damaged) and emergency response (personnel involved, resources involved), providing insight into how resources involved have changed relative to wildfire impact.

(3) California Wildfire Prediction by XGBoost Dashboard – The predicted probabilities of wildfires in California, as generated by the XGBoost model, are visualized on a map. Larger and darker bubbles indicate regions with a higher probability of a wildfire occurring. To provide additional insights, a bar plot shows the top 5 regions with the highest wildfire probability, alongside a distribution of these probabilities. A page filter is applied to these visualizations, enabling users to examine how the probabilities change across different months, thus supporting better wildfire preparation and management. In addition, to highlight key features influencing wildfire probabilities, a bar plot of the top 5 features from the XGBoost model is shown. Filters for environmental factors such as air temperature, humidity, and solar radiation are included, enabling users to explore how they influence wildfire probability. Model performance is assessed with visualizations of the ROC curve and confusion matrix.

(4) Time Series Model Visualization – To gain a deeper understanding of temporal wildfire patterns in California, we conducted a time series decomposition of the average Fire Radiative Power (FRP) from NASA's satellite data (2014–2023). FRP is a critical metric reflecting the intensity and energy released by wildfires.

We first filtered for incidents within California and aggregated the FRP values by week. After ensuring proper datetime formatting, we applied additive seasonal decomposition using the statsmodels library to isolate trend, seasonality, and residuals.

- Trend: We can see an increase in average FRP during late summer to early fall, aligning with California's known wildfire season (July to September).
- Seasonal Component: We can see recurring peaks every year during summer indicate a strong seasonal cycle in fire intensity.
- Residuals: This helps us capture local irregularities such as sudden spikes or dips, which could potentially indicate major events or anomalous weather patterns.

Innovations Over Previous Research:

- While existing research has focused on binary classification – predicting whether a wildfire will occur or not – we introduce a novel approach by predicting the probability of a wildfire occurrence based on specific latitude and longitude, and weather conditions. This approach improves wildfire risk assessments by incorporating environmental factors, enabling more effective strategies for wildfire preparedness

- We also use XGBoost, an improvement from Logistic Regression and Random Forests seen in multiple previous papers.

- Technically, rather than using a static probability cutoff, we dynamically adjust the threshold based on the model's performance as evaluated by the PR curve.

- Our interactive visualization presents state-level wildfire data using four interconnected charts. Users can click on any state in the choropleth heatmap to dynamically update the other charts, revealing both seasonal patterns of wildfire occurrences and intensity. Unlike traditional dashboards that focus solely on fire occurrences, our approach integrates additional layers, such as population density, resource involvement, and damage caused, offering a more comprehensive view that supports both wildfire prediction and management. This makes the dashboard a powerful tool for detecting hidden trends of the dataset and supporting informed decision-making.

V. Evaluation

1. Machine Learning:

To assess the performance of the wildfire prediction model, we used metrics aligning with a binary classification model. Since we were predicting probabilities, this required us to go beyond simple accuracy and prioritize metrics that accounted for class imbalance.

Our evaluation metrics of choice are *Precision, F1 Score, and AUC-ROC Score*.

In our pipeline after tuning the hyperparameters, the model was evaluated on a stratified 20% hold-out test set. All metrics (Precision, F1, ROC-AUC) are computed on this set.

Our model achieved:

- Precision: 0.6930
- F1 Score: 0.6365
- ROC-AUC Score: 0.9771

These results indicate that the model performs well in distinguishing wildfire risk, with a high ROC-AUC reflecting strong overall discrimination capability, while the precision and F1 scores suggest reasonable balance between false positives and false negatives, making it effective for targeted wildfire risk prediction in practice.

2. Visualization:

(1) User Study Evaluation: To evaluate the effectiveness of our wildfire visualization, we conducted a user study using a custom-designed dashboard survey (survey link attached in Appendix). The survey assessed user experience across four key dimensions: clarity (do you agree that the visualization clearly helped me understand the distribution and severity of wildfires?), interactivity (do you agree that the interactive features of the visualization (e.g., zoom, click, hover) intuitive and helpful?), aesthetic design (do you agree that the visualization is well-designed and aesthetically pleasing?), and meaningfulness (do you agree that the visualization is useful for learning about wildfires, such as prevention, response, or forecasting?). Based on responses from 17 participants, the overall feedback was highly positive, with an average rating of 4.88/5. Most users strongly agreed that the dashboard was clear, interactive, and informative. However, a few responses indicated that there is still room for improvement in terms of visual design, clarity, and intuitiveness of interactive features. These insights will help guide future iterations of our dashboard to enhance user experience.

(2) Results & Observations:

Our interactive visualizations including our predictive model aim to address key questions: What are the yearly and seasonal trends of U.S. wildfires? Which regions are most wildfire-prone? How are a cres burned, damage caused, and resource allocation related? What factors are most predictive of wil dfires in California? These insights that users can gain from our dashboards demonstrate their effecti veness in uncovering patterns, supporting decision-making, and enhancing wildfire prediction and ma nagement.

- **Two annual wildfire peaks:** NASA data reveals two peak wildfire periods in the U.S. each year: February–April and July–September. The latter shows significantly more fire occurrences. These peaks vary by region. Central and southern states (e.g., Kansas, Texas, Florida) tend to experience more wildfires in the spring, while western states (e.g., California, Oregon, Washington) see more wildfires in the summer. Despite these seasonal patterns, no consistent yearly trend is observed over the 10-year period. Understanding these regional and seasonal differences is critical for formulating effective fire prevention strategies.

- **Severe wildfires concentrated in summer and western states:** Based on Fire Radiative Power (FRP), the most intense wildfires occurred in western states such as California, Oregon, and Washington. While the number of fire occurrences varies regionally each year, severe fires tend to happen in summer nationwide. For instance, Kansas saw more fires in April than August, but fire intensity was higher in August. Vegetation fires accounted for 94% of all wildfires over the 10-year period, highlighting the strong impact of environmental factors. In contrast, volcanic fires are a major concern in Hawaii.

- **Most major wildfire incidents occurred in less populated regions in California.** It suggests that firefighting efforts in urban areas have been reasonably effective. However, this pattern also indicates potential gaps in fire prevention and response in rural areas. Given that these regions have experienced more frequent and destructive wildfires over the years, it may be necessary to reallocate or increase resources to better support fire management efforts in less populated areas.

- **There is not always a direct correlation between acres burned and impact severity in California.** Some large-scale fires caused relatively minor damage, while others with fewer acres burned led to high fatalities or infrastructure loss. This highlights the importance of considering multiple indicators in fire impact analysis. In addition, our research suggests that certain wildfires in forested areas can be ecologically beneficial, aiding in pest control, eliminating diseased trees,

and maintaining healthy tree density. Therefore, fire prevention and treatments should account for such ecological trade-offs, rather than pursuing a one-fits-all approach.

- **Resource involvement does not always scale with fire severity based on California data.** Our trend analysis shows that while damage indicators peaked in 2018, resource deployment followed a decreasing trend, indicating possible inefficiencies in response or resource allocation. This decline in resources may also partially explain the increasing severity of wildfires over time, highlighting the need for better alignment between fire risk and emergency preparedness.

- **August and September show the highest predicted probabilities for wildfires in California.** In August, the regions with the highest probabilities are: Northeast Plateau (56%), Bishop (38%), Eastern Sierra Basin Range (31%), Los Angeles Basin (28%), and Sacramento Valley (21%). These insights suggest that wildfire preparedness efforts and resource allocation should be concentrated in the summer, particularly in the identified regions. In contrast, during other months, the distribution of predicted wildfire probability is right-skewed, indicating low likelihood of wildfires. The XGBoost model's feature importance analysis reveals that the key environmental factors that influence the predicted probability are soil temperature (10.7%), solar radiation (10.6%), air temperature (7.6%), location (7.4%), and wind run (6.4%). Understanding these environmental factors enables targeted risk assessments and more effective wildfire preparedness strategies.

VI. Conclusion and Discussion

Our project leverages machine learning models and interactive visualizations to uncover key patterns in U.S. and California wildfires, focusing on temporal trends, regional risks, and predictive factors. Key findings include two annual wildfire peaks, regional variations in fire intensity, and disparities in resource allocation versus fire severity. We show that fire impact is not always proportional to burned acres and that ecologically beneficial fires should be considered in prevention strategies. While our model predicts high wildfire risk in California's rural regions during summer, resource deployment does not always align with these risks. Limitations include potential data quality issues and model generalizability across different terrains. For future work we could expand predictive models to other high-risk regions in the U.S.

Contribution:

All team members contributed a similar amount of effort.

VII. Appendix:

User Study: https://forms.gle/Z5awHEhZoTgG4Nhm7

VIII. References:

[1] F. P. Schoenberg, R. Peng, and J. Woods, “On the distribution of wildfire sizes,” *Environmetrics*, vol. 14, no. 6, pp. 583–592, 2003, doi: https://doi.org/10.1002/env.605.

[2] H. K. Preisler and A. L. Westerling, “Statistical Model for Forecasting Monthly Large Wildfire Events in Western United States,” *Journal of Applied Meteorology and Climatology*, vol. 46, no. 7, pp. 1020–1030, Jul. 2007, doi: https://doi.org/10.1175/jam2513.1.

[3] P. Jain, S. C. P. Coogan, S. G. Subramanian, M. Crowley, S. Taylor, and M. D. Flannigan, “A review of machine learning applications in wildfire science and management,” *Environmental Reviews*, vol. 28, no. 4, pp. 478–505, Dec. 2020, doi: https://doi.org/10.1139/er-2020-0019.

[4] K. Hernandez and A. B. Hoskins, “Machine Learning Algorithms Applied to Wildfire Data in California’s Central Valley,” *Trees, forests and people*, vol. 15, Art. no. 100516, Mar. 2024, doi: https://doi.org/10.1016/j.tfp.2024.100516.

[5] K. Pham *et al*., "California Wildfire Prediction using Machine Learning," *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, Nassau, Bahamas, 2022, pp. 525-530, doi: 10.1109/ICMLA55696.2022.00086.

[6] B. Hahs, K. Sood and D. Gomez, "A Data-Driven Model for Wildfire Prediction in California," *2024 International Conference on Smart Applications, Communications and Networking (SmartNets)*, Harrisonburg, VA, USA, 2024, pp. 1-6, doi: 10.1109/SmartNets61466.2024.10577731.

[7] M. Rodrigues and J. de la Riva, “An insight into machine-learning algorithms to model human-caused wildfire occurrence,” *Environmental Modelling & Software*, vol. 57, pp. 192–201, Jul. 2014, doi: https://doi.org/10.1016/j.envsoft.2014.03.003.

[8] N. K. Pahuja and M. H. Rivero, "Predicting the Impact of Wildfire Using Machine Learning Techniques to Assist Effective Deployment of Resources," *2022 International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV, USA, 2022, pp. 201-205, doi: 10.1109/CSCI58124.2022.00039.

[9] Y. J. Chew, S. Y. Ooi, Y. H. Pang and J. K. Hoi, "Enhancing Forest Fire Management in Malaysia with a Power BI Dashboard: Leveraging VIIRS Hotspots and Environmental Data," *2024 IEEE 14th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, Penang, Malaysia, 2024, pp. 482-486, doi: 10.1109/ISCAIE61308.2024.10576420.

[10] Q. Meng, H. Lu, Y. Huai, H. Xu, and S. Yang, “Forest Fire Spread Simulation and Fire Extinguishing Visualization Research,” *Forests*, vol. 14, no. 7, Art. no. 1371, Jul. 2023, doi: https://doi.org/10.3390/f14071371.

[11] C. A. T. Cortes *et al*., "Analysis of Wildfire Visualization Systems for Research and Training: Are They Up for the Challenge of the Current State of Wildfires?," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 7, pp. 4285-4303, July 2024, doi: 10.1109/TVCG.2023.3258440.

[12] L. Cheong, S. Bleisch, A. Kealy, K. Tolhurst, T. Wilkening, and M. Duckham, "Evaluating the impact of visualization of wildfire hazard upon decision-making under uncertainty," *International Journal of Geographical Information Science*, vol. 30, no. 7, pp. 1377–1404, Jan. 2016, doi: https://doi.org/10.1080/13658816.2015.1131829.

[13] C. A. Brewer and L. Pickle, "Evaluation of Methods for Classifying Epidemiological Data on Choropleth Maps in Series," *Annals of the Association of American Geographers*, vol. 92, no. 4, pp. 662–681, 2002, doi: https://doi.org/10.1111/1467-8306.00310.

[14] R. G. Cromley and E. K. Cromley, "Choropleth map legend design for visualizing community health disparities," *International Journal of Health Geographics*, vol. 8, no. 1, Art. no. 52, 2009, doi: https://doi.org/10.1186/1476-072x-8-52.

[15] N. Guldåker, "Geovisualization and Geographical Analysis for Fire Prevention," *ISPRS International Journal of Geo-Information*, vol. 9, no. 6, Art. no. 355, May 2020, doi: https://doi.org/10.3390/ijgi9060355.