**Wildfire Risk Prediction**

CS 6140 – Machine Learning - FALL 2023

by Elio Khouri and Ilana-Mahmea Siegel

# 1. Executive Summary

## 1.1 Brief overview of the project.

### 1.1.1 Problem Statement

California is an epicenter for relentless wildfires that bring devastation year after year. Aspects such as climate change, urban expansion, and mismanagement of forests compound the issue, resulting in an escalation of the frequency and magnitude of these wildfires. These events not only jeopardize the safety of California’s residents but also threaten the state's economic health and the integrity of its ecosystems. Despite the evolution in firefighting methodologies and increased resources, there is a pressing requirement for a proactive strategy to predict and prevent these wildfires. A well-crafted predictive model stands to offer invaluable insights into potential ignition points and the trajectory of active fires, thus affording authorities the advantage of time in resource allocation and strategic planning.

### 1.1.2 Significance of the Problem

The ramifications of California's wildfires are profound, affecting diverse facets of life in the state. Based on recent data from Cal Fire, the economic burden of these fires is staggering, with damages amounting to over 300 million dollars in 2022 (Tyler, 2022), and a somber tally of 133 lives lost in the past five years alone (*Statistics | CAL FIRE*, n.d.). Yet, these figures only scratch the surface. The broader economic implications span diminished property values, stifled tourism, and long-lasting health repercussions.

A 2016 study suggests that by mid-century, average PM2.5 emissions from wildfires in the Western U.S. could rise by over 60% compared to levels from 2004 to 2009. Additionally, the peak emissions of PM2.5 might see a surge of nearly 400% in certain regions (*Living Under Smoky Skies—Understanding the Challenges Posed by Wildfire Smoke in California*, 2022). Such increases in PM2.5 emissions have direct health implications.

Pratt et al. estimated that due to exposure to ozone generated by wildfire emissions, there's an associated increase of over 2,000 ER visits in children with asthma in the USA annually (Holm et al., 2020). Delfino et al. observed that wildfire PM2.5 exposure led to an 8% rise in asthma visits for children younger than five (Holm et al., 2020). In British Columbia, Canada, a significant association was found between wildfire smoke and adverse health outcomes. For instance, a 30 µg/m3 increase in total PM10 led to increased respiratory physician visits (odds ratio (OR) = 1.05) and specific asthma visits (OR = 1.16) (Youssouf et al., 2014). Furthermore, in Sydney, bushfire smoke events corresponded with a 6% increase in respiratory hospital admissions on the same day, with chronic obstructive pulmonary disease admissions rising by 13% and asthma admissions by 12% (Youssouf et al., 2014). Delfino et al. also studied the 2003 southern California wildfires and found asthma admissions across all ages increased by 4.8% during the wildfire period, with a more pronounced effect in the elderly (10.1% increase) and children aged 0-4 years (8.3%) (Holm et al., 2020). Vulnerable populations, notably children and the elderly, face the brunt of these health adversities. Beyond the immediate health concerns, the environmental ramifications are severe. Wildfires expedite soil erosion, induce water contamination, and disrupt vast habitats. Tackling the wildfire challenge with predictive strategies isn't just a scientific endeavor but an imperative humanitarian and ecological initiative.

Predicting wildfires is crucial for California. With accurate forecasts, we can evacuate areas at risk, reduce casualties and property damage. This foresight also allows us to strategically position firefighters and resources, making response efforts more effective. Furthermore, knowing when air quality will decline helps at-risk groups, like the elderly or children, to take precautions. Overall, anticipating wildfires isn't just about preventing destruction—it's a vital step toward protecting our communities and preserving California's environment.

## 1.2 Key findings or results.

Our comprehensive analysis led to several key findings:

1. Fire History's Significant Impact: Despite consisting of only a few features, fire history data proved to be more influential in predicting wildfires than 39 historical weather condition features. This underlines the critical importance of historical fire data in understanding wildfire incidents.
2. Superior Performance of Imputed Data Models: Models trained on datasets with imputed data (df2 and df3) consistently outperformed those trained on df1, where NaN values were dropped. This suggests that retaining and imputing data, rather than discarding it, enhances model accuracy.
3. Limited Effectiveness of Weather History Data: Contrary to expectations, weather history features did not significantly improve model performance compared to the baseline. This raises questions about the relative importance of these features in wildfire prediction models.
4. High Efficacy of Combined Feature Sets: Combining fire history with other immediate(day of incident) features generally yielded the best results, underscoring the value of a multifaceted approach to feature selection in model training.

## 1.3 Summary of recommendations or conclusions.

Based on our findings, we recommend the following:

1. Prioritizing Fire History in Model Training: Future models should emphasize fire history data, given its substantial impact on prediction accuracy.
2. Balanced Approach to Data Handling: Careful consideration should be given to handling missing data, with a preference for imputation over deletion to maximize the dataset's potential.
3. Further Research on Weather Data: Additional studies are needed to assess the role of weather history data in wildfire prediction, potentially exploring alternative data representation or longer-term weather trends.
4. Publication of Our Dataset: We aim to publish our dataset, which includes unique features not found in publicly available datasets, to benefit future research in this area.

# 2. Introduction

## 2.1 Background information about the project.

### 2.1.1 Previous Solutions and Approaches

Wildfire prediction has been the subject of myriad studies, spanning from elementary Kaggle challenges to advanced peer-reviewed publications. The studies largely differ in terms of data sets used, complexity of the algorithms, and the depth of the analysis.

Elementary studies, often sourced from platforms like Kaggle, typically leverage limited datasets and pre-built machine learning models. For instance, many Kaggle competitions focus on predicting wildfires based on basic meteorological data such as temperature, wind speed, and humidity (see Appendix A). Some have incorporated additional features like vegetation, topography, and even previous fires in the region. While these elementary studies provide a foundational understanding of wildfire prediction, they often rely on single datasets and might not capture the nuanced interplay of multiple factors influencing wildfire occurrence.

On the other hand, advanced peer-reviewed publications dive deeper, leveraging an array of sophisticated machine learning algorithms, particularly deep learning models. These studies often scrape data from multiple sources, ensuring a comprehensive dataset. For instance, the study on "Wildfire Danger Prediction and Understanding With Deep Learning" leverages daily weather data from ERA-5 Land, including metrics like maximum temperature, wind speed, and dewpoint temperature. They also incorporate satellite variables from MODIS, such as the Normalized Difference Vegetation Index and day/night Land Surface Temperature. Additional features include soil moisture index from the European Drought Observatory, details on roads and waterway distances, population density from WorldPop, as well as elevation, slope, and land cover fractions from Copernicus databases (Kondylatos et al., 2022).

Another notable study, "Wildfire Risk Prediction and Detection using Machine Learning in San Diego, California," utilized a range of machine learning and deep learning algorithms, including SVM, XGBoost, and Random Forest for weather-based fire predictions, and CNN, LTSM, and SVM for remote sensing-based models. These models were combined into a sophisticated ensemble model that leveraged multiple data sources, such as weather data, satellite imagery from Landsat 8, historical fire events, and vegetation indices like NDVI. Their ensemble approach achieved a remarkable 100% accuracy for fire risk prediction, underscoring the potency of integrating diverse algorithms and datasets (Malik et al., 2022). Such refined models, trained on a multitude of datasets, tend to have a more holistic understanding of the factors influencing wildfire danger.

The government's approach to determining fire danger provides another perspective. The U.S. National Park Service and the National Weather Service, for instance, consider an array of factors such as fuel moisture, drought indices, and historical fire data to determine fire danger levels (NPS, 2023). This highlights the multifaceted nature of wildfire prediction, which is not merely reliant on current weather conditions but also on historical and environmental data.

While the field of wildfire prediction research often benefits from the collaborative efforts of larger teams and the mentorship of seasoned experts, Our project is a testament to the dedication and initiative of two individuals taking their initial steps into the vast realm of machine learning. Our goal is clear: to bridge the gap between elementary and advanced studies. By recognizing the limitations of Kaggle-based models and recognizing the strengths of refined academic studies, we aspire to broaden our data scope, inching closer to the insights and performance offered by more seasoned methodologies.

## 2.1.2 Drawbacks of Existing Solutions

While the existing models provide valuable insights into wildfire predictions, they aren't without limitations. Elementary studies from platforms like Kaggle often face data limitation constraints. Relying on single datasets can lead to overfitting, where models perform exceptionally well on training data but falter when exposed to new, unseen data. Moreover, these models often utilize pre-built classifiers, which might not be specifically tailored for wildfire prediction, leading to potential inaccuracies.

Furthermore, even though advanced studies use sophisticated algorithms and vast datasets, they sometimes overlook crucial factors. For instance, the importance of weather parameters like air temperature, wind, and soil moisture has been underlined in several studies (see Appendix A). While these parameters play pivotal roles in influencing wildfires, not all refined models incorporate them, potentially leading to gaps in predictions.

Moreover, while the government's fire danger assessment tools, such as the Wildfire Danger Factors and the Wildfire Fire Index (WFI) codes (appendix), provide a comprehensive understanding of fire danger, they sometimes lack real-time adaptability. These systems are based on accumulated knowledge and historical data, and while they are robust, they might not be agile enough to predict sudden, anomalous wildfire outbreaks.

Recognizing these challenges, our approach seeks to strike a balance. We understand that as a team of two, in our first machine learning course, our logistic regression model might not rival the advanced solutions developed by larger teams with extensive resources and seasoned expertise in the field. However, by integrating a diverse set of data sources, we aim to leverage the strengths of both elementary and advanced models. Our goal is to improve upon the limitations of elementary studies while acknowledging and learning from the sophisticated methodologies of advanced research. In doing so, we hope to offer a fresh perspective that bridges the gap between these two ends of the spectrum.

## **2.2 Objectives and goals.**

Our primary objective was to investigate the effect of different feature sets on the performance of wildfire prediction models. This involved analyzing a variety of features, including meteorological data and fire history, to determine their relative importance in predicting wildfire occurrences in California.

## **2.3 Scope of the project.**

The scope of our project was focused on wildfire prediction in California, particularly from 2014 to the present. This timeframe allowed us to analyze recent data, taking into account the current climatic and environmental conditions influencing wildfire trends in the state. Our approach involved creating a comprehensive dataset by integrating and analyzing various data sources to build an effective predictive model.

# 3. Methodology

## **3.1 Description of the methods or approaches used.**

### 3.1.1 Data Collection

Data collection and dataset creation were pivotal stages of this project. We downloaded data from various sources, ensuring that each dataset was comprehensive and relevant to our study of California wildfires.

The primary datasets included:

* GHCN (Global Historical Climatology Network) Data: This dataset provided detailed climatological information, crucial for understanding the environmental conditions leading up to wildfires.
* LCD (Local Climatological Data): These data offered localized weather information, offering insights into microclimatic conditions in different parts of California.
* Cal Fire Fire Incidents Database: This dataset was central to our study, as it contained detailed records of past fire incidents across California.
* U.S. Forest Service Dataset (via Kaggle): This dataset complemented the Cal Fire data by providing historical fire records, thereby enriching our analysis.

Each of these datasets was unique in structure, requiring careful handling during integration. For instance, the weather datasets (GHCN and LCD) contained features such as temperature, humidity, wind speed, and other meteorological parameters. The fire datasets included details on fire locations, start and end times, fire names, and acres burned.

*A comprehensive list of the features used in our analysis will be provided in the index, detailing the specific attributes considered from each dataset.*

### 3.1.2 New Dataset Creation

#### 3.1.2.1 Creating Basic Data for Negative Examples

We pulled core features of location and time from the Cal Fire dataset for positive examples and generated them for negative examples. Fore the negative examples, locations were generated within approximate boundaries of California and then cross-checked with the actual shape of the state to get valid example locations. To partially counter the skew in the locations of the positive examples (Fig. 3.1.2.1-A), ¾ of the points we generated for the negative examples (Fig. 3.1.2.1-B) were completely random and ¼ were randomly chosen from locations where there were also positive fire examples. Times were completely randomly generated, but a future iteration of this dataset may consider biasing dates and times towards the skew in the positive examples dates and times in a similar way. All locations for negative examples were cross-checked with the locations and times of positive examples to ensure that there really were not fires at those location and time pairings.



#### 3.1.2.2 Reformatting Weather Data and Matching it to Basic Data

The process of reformulating weather data was a critical task in our data preparation stage.

This involved:

Converting Data Formats: Transforming date formats for consistency across datasets.

Aggregating Weather Data: We converted hourly weather data to daily metrics to align with the fire incident data. This step was essential to match weather conditions to the specific days of fire incidents.

Weather Data Integration: The weather data from GHCN and LCD was integrated with fire incident data, ensuring that each fire incident had corresponding weather information based on location and date. This step was key in creating a holistic dataset that captures both the incident details and the environmental conditions at the time.

Miscellaneous Conversions: The GHCN data was converted from metric to imperial units, and there were other little modifications or tweaks all along the process to bring it to fruition.

Through these methods, we were able to create a comprehensive dataset that not only documented fire incidents but also included relevant environmental and climatic conditions, setting the stage for an in-depth analysis of factors contributing to wildfires in California.

#### 3.1.2.1 Merging Fire History Data and Matching it to Basic Data

Our fire examples were drawn from the Cal Fire dataset, but that dataset only goes back to 2013. In order to obtain information to use to create fire\_history features, we combined the Cal Fire dataset with the U.S. Forest Service dataset from Kaggle, which covered fires from 1992-2015. To avoid duplicating incidents that were listed in both datasets during the overlap period, we matched incidents during that period based on similarities in their start times and their names, and the ratio of their listed locations to their sizes. The ratio of locations to sizes was key since two fires within, for example, .5 km on the same day are likely really the same fire if they each burned 100 km, but are probably different fires if they only burned .1 km. Kaggle fire incidents that matched Cal Fire incidents were removed before the datasets were joined together. The new “CombinedFires” dataset included a reduced feature set from the original datasets. The features included fire locations, start and end times, names, and acres burned.

The CombinedFires dataset we created was used to generate the fire\_history features for each positive or negative example: one near\_fire\_history capturing fires within 5 km and 3 yr, and one far\_fire\_history capturing fires within 15 km and 20 yr of the example location and time. The dataset was searched for locations and times near each example and the number of acres they burned was averaged across the number of years in the time window. We chose average as the aggregation method because the Kaggle data included many times more fires than the Cal Fire dataset, so counting the number of fires started per year would vary greatly for examples whose histories came from primarily from the Kaggle data (e.g. an example from 2014) and those whose histories drew mostly on Cal Fire data (e.g. an example from 2023). Most of the incidents included in the Kaggle dataset were very small, which meant that using the number of acres burned as the metric would essentially control for the issue and result in fairly similar measurements for examples whose histories relied almost entirely on data from one dataset or the other.

### 3.1.3 Data Cleaning and Exploration

Cleaning and exploring the data were critical steps to ensure the reliability of our model. This process involved:

**Data Cleaning:**

We meticulously cleaned the datasets, removing any physically unreasonable values and dealing with missing data. This included dropping or imputing NaN values, resulting in three distinct versions of our dataset:

df1: Dropped both NaN columns and rows.

df2: Dropped NaN columns but imputed NaN rows with mean, median, and mode.

df3: Dropped NaN columns and imputed NaN rows with mean and mode only.

**Data Exploration:**

We used various exploration techniques, including histograms to visualize data distributions across different features like time, month, and meteorological conditions. In fact, all features in all of the datasets were visualized in our 3 visualization notebooks. The exploration process was geared towards understanding nonlinear relationships, making machine learning techniques like SVM, Random Forest, or Neural Networks suitable for our analysis, with a particular emphasis on Neural Networks due to the number and complexity of features involved.

### 3.1.4 Model Building

### 3.1.4.1 Model from Scratch

We both separately explored building models from scratch using Numpy. Mea’s efforts were more successful so we tried training that model on our data. The code includes classes for Layer and Model as well as activation, loss, regularization and weight initialization functions that allow us to easily create models with different architectures. The code also includes a Metric class to calculate precision, recall and other metrics. This class was used to calculate the metrics for models we build using TensorFlow, too.

We built a model using our from-scratch code with one with comparable architecture built using TensorFlow. After training the models on the same data over 30 epochs, our from-scratch model was not able to successfully learn the data and predicted no positives. The TensorFlow model achieved an F1 Score of 0.6. Our model learned a little better with less regularization, but the balance between learning the features and not exploding the gradients was very difficult to manage. Our model only made limited use of vectorized functions and each of its training epochs took about 2 min while the TensorFlow model’s epochs were about 2 seconds.

Our experiment focused on the impact of different feature sets on model performance, rather than on optimizing model architecture. Given the much better performance of the TensorFlow model, we decided to use TensorFlow rather than our from-scratch code to build the model for our experiment. The much shorter training times for TensorFlow models also enabled us to perform more experiments.

### 3.1.4.2 Tensorflow Model

The models we ultimately used for our experiment were built with TensorFlow. We used the same architecture for each model, which we trained as described in 3.1.5. The layer dimensions can be seen in Fig 3.1.4.2-A. The final layer used a sigmoid activation function and the rest used relu. All layers included a bias (optional in TensorFlow). The models used binary cross-entropy loss and an Adam optimizer.



### 3.1.5 Model Training

We trained the same model architecture on various versions of the data. Each model was trained for 300 epochs with a learning rate of 0.001.

Although it was not part of our experiment proposal, we were curious about the effects of different methods of handling NaN values (dropping them or imputing the data), as mentioned above. We decided to perform our experiment comparing the effects of changing feature sets using each of the three datasets, df1-3, described in 3.1.3. For each dataset, we read in the data and split it 75-25 into training and validation samples. The same training-validation split was used for each model trained on a different subset of input features. For example, *df3\_baseline\_train\_x = df3\_train\_x[baseline\_features]* and *df3\_fire\_history\_train\_x = df3\_train\_x[baseline\_features + fire\_history\_features].* That way the only things that changed between models trained on data from a given dataset were the initial random weights and the features. A future experiment could use models with the same initial weights, too, but we don’t think the randomness of the weights had a meaningful impact on our experiment. Although the exact performance of the models changed slightly with different train-validate splits, the relative performance of the models trained on different feature sets was consistent across different runs of the code.

The different feature sets we used were:

* **minimal**: Just location and date/time features (included out of curiosity, not as an official baseline).
* **baseline**: Location and date/time features as well as immediate weather data for that day. This is the set of features most commonly found in our research on previous experiments with fire prediction.
* **weather\_history**: All the baseline features, plus 10-, 30-, and 60-day aggregated weather data. This additional weather data is used occasionally in other studies.
* **fire\_history**: All the baseline features, plus average acres burned within 5 km and 3 yr and 15 km and 20 yr of the example.
* **complete/all**: All the baseline, weather\_history and fire\_history features.

In total, each run of the notebook tested the 15 dataset-feature set combinations shown in table below.

| **feature set →**  **dataset ↓** | **minimal** | **baseline** | **weather\_history** | **fire\_history** | **complete** |
| --- | --- | --- | --- | --- | --- |
| **df1 (drop NaNs)** | df1- minimal | df1- baseline | df1- weather\_history | df1- fire\_history | df1- complete |
| **df2 (impute with mean and median)** | df2- minimal | df2- baseline | df2- weather\_history | df2- fire\_history | df2- complete |
| **df3 (impute with mean only)** | df3- minimal | df3- baseline | df3- weather\_history | df3- fire\_history | df3- complete |

## **3.2 Tools or technologies employed.**

Our project employed a variety of tools and technologies, including:

Python: Served as the primary programming language for data processing, analysis, and model building.

SQLite: Used for database management, particularly with the Kaggle fire dataset.

JavaScript (JS): Assisted in mapping and visualization tasks.

Google Colab: Provided a cloud-based environment for running our Jupyter Notebooks.

NumPy/Pandas: Crucial for data manipulation and numerical computations.

TensorFlow (TF): Used for building and training machine learning models, particularly neural networks.

## 

## **3.3 Data sources or materials.**

The data sources for our project were carefully selected to ensure a comprehensive and relevant dataset:

[GHCN (Global Historical Climatology Network):](https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C00861/html) Provided extensive climatological data.

[LCD (Local Climatological Data):](https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C00684/html) Offers localized weather information.

[Cal Fire Fire Incidents Database:](https://www.fire.ca.gov/incidents) Included detailed records of fire incidents in California.

[U.S. Forest Service Dataset (via Kaggle):](https://www.kaggle.com/datasets/rtatman/188-million-us-wildfires) Provided historical fire records.

# 4. Project Member Work Distribution

## 4.1 List of team members and their specific roles or responsibilities.

## Elio and Mea: Both Elio and Mea contributed significantly across various aspects of the project, sharing responsibilities evenly. Their collaborative efforts were pivotal in the project's success, encompassing everything from preliminary research to the final report.

## 

## 4.2 Breakdown of Tasks Assigned to Each Member

**Elio:**

* Managed the LCD data, including processing, converting, and extracting.
* Worked on the Kaggle fire dataset, handling the data via SQLite for manipulation and conversion.
* Contributed initial code for combining fire datasets.
* Developed weather-related features, weather aggregation data, additional station and elevation data.
* Led geohashing.
* Led the efforts in cleaning the enhanced dataset.
* Contributed to creating visualizations.
* Participated in the development of the from-scratch model and the TensorFlow implementations.

**Mea:**

* Focused on the GHCN data, handling its collection, processing, formatting, and conversion.
* Worked on the Cal Fire dataset in a similar capacity.
* Took over and finalized the combining fires script, merging data from the overlap period of 2013-2015.
* Created a California JSON polygon for bounding box definition and mapped data points using HTML, JS, and Python.
* Formulated the main functions in the combined dataset file, including the creation of fire history, positive and negative examples, and synthetic negatives.
* Initiated the clean data set, later taken over by Elio.
* Led creation of visualizations.
* Developed the from-scratch model and TensorFlow implementations.

**Collaborative Efforts:**

* Both Elio and Mea collaborated on the presentation, proposal, README files, preliminary research, and data source exploration.
* Worked together on the main file for merging datasets, contributing to various supporting functions.
* Shared administrative tasks, file and code maintenance, and engaged in extensive discussions and planning over numerous hours.

## **4.3 Timeline or schedule for each member’s contributions.**

October 2: Start of the project with dataset provision. Decision to create an original dataset.

October: Conducted preliminary research, sourced data, and developed the proposal document.

Late October to November: Focused on mining and extracting raw data.

November 20: Completed the project presentation and achieved the stage of having the enhanced dataset.

November 20 to December 11: Engaged in further data cleaning, model creation and training, visualization development, and final report compilation. Structured the entire repository and associated documentation.

# 5. Project Progress/Development

## **5.1 Detailed account of project activities.**

The project's course involved a detailed process of data collection, processing, and analysis. Initially, the team focused on an extensive phase of research and data sourcing. This foundational period was crucial for Elio and Mea, as they gathered preliminary research, explored potential data sources, and established the groundwork for their approach.

During the data extraction and processing phase, the team's efforts were divided based on the dataset characteristics. Mea took on the GHCN data, handling its collection, processing, formatting, and conversion, ensuring that this dataset was ready for integration. Simultaneously, Elio concentrated on the LCD data, undertaking tasks of processing, converting, and extracting valuable information. He also managed the Kaggle fire dataset, utilizing SQLite for data manipulation - a method distinct from the Python-based approach commonly used.

Both team members actively contributed to the creation of README files and the exploration of data sources. Their collaborative efforts were also evident in the development of a California JSON polygon, which was used to define a bounding box within the state. This tool was instrumental in accurately mapping data points.

The midpoint of the project saw a concerted effort in combining and integrating various datasets. Mea led the charge in combining the Cal Fire and Kaggle fire datasets, focusing on the period of 2013-2015 where data overlap occurred. Meanwhile, Elio's contributions included the initiation of code for this task, which Mea then took forward to completion.

As the project progressed, both Elio and Mea were involved in the development of the combined dataset file. This phase included the creation of fire history data, the formulation of positive and negative examples, and the generation of synthetic negatives to establish a comprehensive basic dataset. Additionally, Elio's work on weather-related features, data aggregation, and geohashing played a critical role in enhancing the dataset.

In the latter stages, the focus shifted to cleaning the enhanced dataset and creating visualizations. While Mea initiated this phase, Elio took over, bringing it to completion and ensuring the data was ready for the final analysis. Their joint efforts also extended to the implementation of the model, both from scratch and using TensorFlow, showcasing their versatility and adaptability in handling various aspects of machine learning.

Overall, the project's development was marked by a methodical and thorough approach, with both team members contributing significantly to each phase, ensuring the project's objectives were met within the planned timeline.

## 5.2 Stages or phases of the projec**t.**

## Initial Research and Data Sourcing (Early October): Focused on gathering preliminary data and establishing a foundational understanding of the project's scope.

## 

## Data Extraction and Processing (Late October - November): Involved in-depth work on mining and converting raw data from various sources.

## 

## Presentation and Enhanced Dataset Creation (By November 20): Culminated in a detailed project presentation and the creation of an enhanced dataset.

## 

## Data Cleaning and Model Development (November 20 - December 11): Involved refining the dataset, developing and training the model, and creating visualizations.

## **5.3 Progress against planned timelines.**

The project adhered closely to the planned timeline, with each phase executed within set timeframes. It's noteworthy that the workload was particularly intense leading up to the project's midway presentation and reached its peak in the final week before the submission of the final report. Despite these demanding periods, the team's efficient collaboration ensured the timely completion of a comprehensive final report and a fully structured repository.

# 6. Results and Findings

## **6.1 Presentation of the data or outcomes.**

Our study's findings are grounded in a detailed analysis of the data collected and processed from various sources. The datasets, after being meticulously cleaned and enhanced, offered a diverse range of features for analysis. Three primary datasets were created: df1, df2, and df3, each treated differently in terms of handling missing values.

df1: This dataset was the most stringent in terms of data cleanliness, where all columns and rows with NaN values were dropped. This led to a certain level of data reduction but ensured the highest data integrity.

df2: This dataset employed a more nuanced approach towards missing values. NaN values in rows were imputed using a combination of mean, median, and mode, depending on the nature of the data. Columns with NaN values were dropped. This approach allowed for a richer dataset while still maintaining a significant level of data reliability.

df3: Similar to df2, df3 involved dropping columns with NaN values but imputed missing values in rows using only mean and mode. This method struck a balance between data richness and integrity.

The analysis revealed a clear trend: the datasets with imputed data (df2 and df3) consistently outperformed df1. This improved performance can be attributed to the larger volume of incidents available for the model to learn from, adding to the robustness of the predictive analysis.

|  | Predicted Positive | Predicted Negative |
| --- | --- | --- |
| Positive | 381 | 106 |
| Negative | 86 | 882 |

| Precision | minimal | baseline | weather\_hist | fire\_hist | all |
| --- | --- | --- | --- | --- | --- |
| df1 | NaN | 0.585831 | 0.568528 | 0.744737 | 0.788512 |
| df2 | NaN | 0.655773 | 0.675676 | 0.801961 | 0.854839 |
| df3 | NaN | 0.660079 | 0.675565 | 0.798768 | 0.815846 |

| Recall | minimal | baseline | weather\_hist | fire\_hist | all |
| --- | --- | --- | --- | --- | --- |
| df1 | 0.0 | 0.581081 | 0.605405 | 0.764865 | 0.816216 |
| df2 | 0.0 | 0.606855 | 0.604839 | 0.824597 | 0.747984 |
| df3 | 0.0 | 0.685832 | 0.675565 | 0.798768 | 0.782341 |

| Accuracy | minimal | baseline | weather\_hist | fire\_hist | all |
| --- | --- | --- | --- | --- | --- |
| df1 | 0.697218 | 0.748773 | 0.741408 | 0.849427 | 0.878069 |
| df2 | 0.659107 | 0.757388 | 0.766323 | 0.870790 | 0.870790 |
| df3 | 0.665292 | 0.776632 | 0.782818 | 0.865292 | 0.868041 |

| F1 Score | minimal | baseline | weather\_hist | fire\_hist | all |
| --- | --- | --- | --- | --- | --- |
| df1 | 0.0 | 0.583446 | 0.586387 | 0.754667 | 0.802125 |
| df2 | 0.0 | 0.630366 | 0.638298 | 0.813121 | 0.797849 |
| df3 | 0.0 | 0.672709 | 0.675565 | 0.798768 | 0.798742 |

## 6.2 Analysis of the results.

A key discovery from our analysis was the significant impact of fire history data on the model's predictive accuracy, despite it comprising only a few features compared to the over 50 features related to immediate and historical weather conditions. This finding highlights the critical importance of historical fire data in understanding and predicting wildfire incidents.

The comparative analysis between different feature sets and datasets yielded insightful results. Models trained on df3 and df2, which included fire history or all features, consistently showed superior performance. This suggests that the combination of comprehensive historical data with a well-rounded approach to missing value imputation significantly enhances predictive accuracy.

In contrast, models relying solely on weather history or baseline data were less effective. Furthermore, df1, with its strict approach to data cleanliness, lagged in performance. These findings underscore the importance of a balanced approach in data preparation – one that carefully considers both the quantity and quality of data.

The effectiveness of fire history data, even with its limited feature set, in enhancing model performance is a notable finding. It indicates that certain key features can have a disproportionate impact on the success of predictive models, highlighting the importance of feature selection in machine learning.

## 

## 6.3 Graphics or charts to illustrate key points.

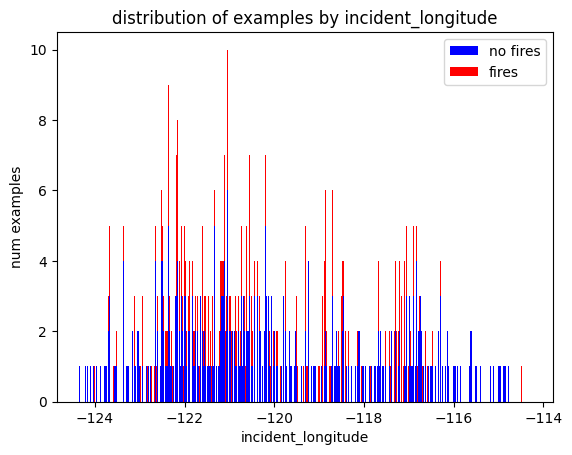
To visually represent our dataset, we have prepared a series of charts and graphics that provide a comprehensive depiction of the comparative performance distribution of features in our dataset. These visuals illustrate the trends and patterns observed in our study, aiding in a clearer understanding of the results.

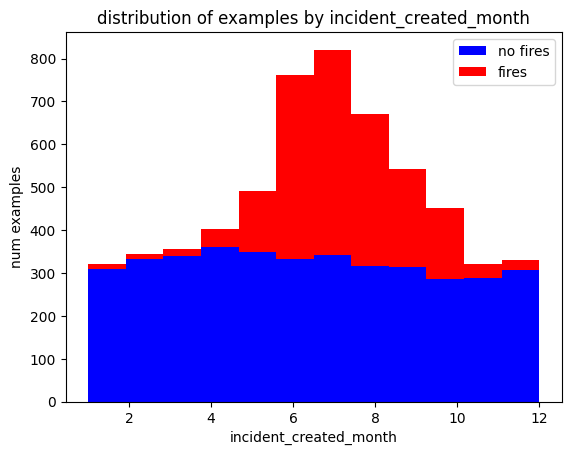
Below, we present a selection of histograms, each representing the class of fire and no fire, showcasing the spread over various parameters such as longitude, incident month, hour, minute, and dry bulb temperature. These histograms are a sample from over 60 similar visualizations, each providing unique insights into the data distribution and tendencies.

In addition to the histograms, we have also prepared two correlation matrices of the features. These matrices offer a detailed view of the relationships and interactions between different variables in our datasets, helping to understand the dynamics that influence fire occurrences.

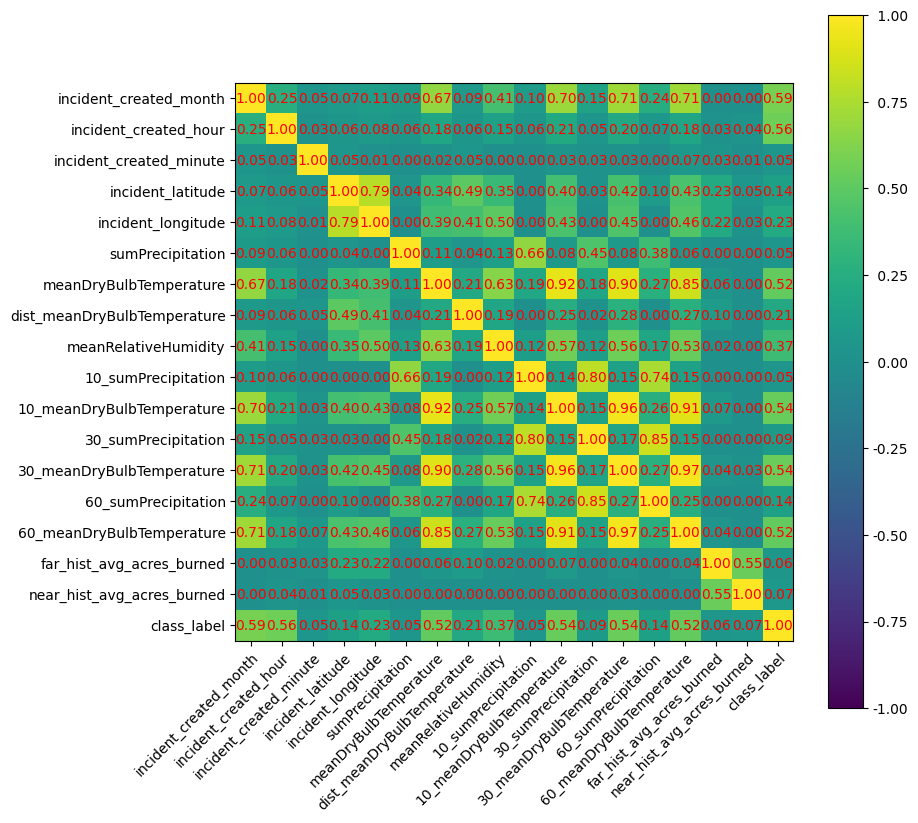
For a more extensive collection of all images and figures related to our analysis, we invite readers to visit our GitHub repository, specifically the 'visualization' section within the 'data' folder. This repository houses detailed charts and graphs that represent the outcomes of our study, providing a comprehensive view of our analytical process and its outcomes.

For graphics representing our results, see the charts in 6.1.

****

****

# 



# 7. Discussion

## **7.1 Interpretation of results.**

### 7.1.1 Comparison of Methods for Handling NaN Values

The datasets which imputed NaN values (df2 and df3) led to to models with significantly better results than the dataset which simply dropped NaN values (df1).Dropping NaN values meant that df1 was 89% the size of df2 and df3. This decrease in the number of training examples likely caused much of the decrease in model performance. With fewer training examples, models had less opportunity to learn useful features, especially the models like baseline, which had fewer features. Across multiple runs, there was not a clear trend to show that either imputation using just the mean or imputation that occasionally used the median was better than the other.

### 7.1.2 Comparison Across Feature Sets

Focusing on our project’s research question, we compared the performances of the models trained on different feature sets. Although there were 39 weather\_history features, including them generally only lead to slight improvements over the baseline performance. Conversely, the two fire\_history features cause a large improvement in model performance. Models trained with fire\_history features performed 9-18% across precision, recall, accuracy and F1 than models trained on weather\_history features, and 11-21% better than the baseline. Models using all features performed about the same as the models with the fire\_history features. Although not officially part of our experiment, we trained models on an additional minimal feature set that didn’t include any weather features. Those models never predicted fires.

Given the relatively frequent use of weather history data (as compared to fire history or vegetation data) in other fire prediction models and the impact of long-term weather on the flammability of forests, we had expected the weather\_history features to be more impactful. Their poor performance in our experiment does not rule out their possible importance in training fire prediction models generally.

We considered several reasons why the weather\_history may not have had a large positive impact on these models. One possibility is that the 10-, 30-, and 60-day weather histories did not add much new information since they varied so closely with the daily measurements. The covariation decreased with longer measurement windows, so it’s possible that using a 90-day measurement window instead of the 10-, 30-, and 60-day ones would be more effective. Medium-term weather trends like 30-day average temperature, are also correlated with the month so the incident\_created\_month feature may effectively capture much of the information weather\_history features would offer.

The distances between the weather measurements and the possible fire locations is likely another factor limiting the effectiveness of the weather\_history features. The mean distance between an incident location and the corresponding LCD and GHCN weather stations is ~32.7 km and ~16.8 km, respectively. At such a distance, weather measurements have limited use. Moreover, the distances vary from a minimum of ~0.13 km to a maximum of ~128.5 km, making measurement relevance unstable. A better weather measurement dataset may improve the usefulness of weather\_history features.

The size of the impact of the fire\_history features was as surprising as the weakness of the weather\_history features. Although we had hypothesized that fire history might be a good indicator of fire likelihood, during our exploratory data analysis, fire\_history features did not appear to clearly correlate with fire likelihood at all. They also did not correlate with the other features and it seems clear that they introduced a very different kind of information to the model.

We had hypothesized that fire history might be a good predictor of fire likelihood since a place that has burned a lot before may be prone to fires, and a place that has burned very recently may not have much fuel left to burn now. There are so many variables that can make an area fire prone (e.g. camping activity, proximity to high-voltage power lines, geographic features, etc.) that cannot all be accounted for explicitly, but fire history information may be a good approximation of them. It would be interesting to explore this category of features more to understand why the fire\_history features are as important as they are. Future experiments could try using fire\_history without any weather information, using different time and distance ranges for fire\_history aggregation windows, including the number of historic fires per year (we used the number of acres burned), or trying to capture some of the other variables like camping, topography, vegetation, etc. that fire\_history may be approximating.

Based on the results of our experiments, the best features on which to train a fire prediction model include fire\_history but not weather\_history.

## **7.2 Comparison with initial objectives.**

We set out to create our own fire/no-fire classification dataset that included more features than the available datasets we’d found. It took a lot of work, but we succeeded in that. We also intended to compare models trained on different feature sets to investigate the relative utility of different features. We succeeded in that, too. We also hoped to write a neural network from scratch which we did. We hit setbacks, and there is much more that can be explored, but overall we achieved our goals.

## **7.3 Challenges encountered and how they were addressed.**

The biggest challenge was with getting the data. We had wanted to include topographic and vegetation data, but were unable to access datasets for those. The weather data we found came from two different datasets, both of which were stored online. We had to figure out how to get the data (e.g. by querying S3), what data was there (many fields were blank and the documentation wasn’t always clear), and how to combine the data from different datasets. That took a lot of time and careful consideration. Figuring out how to combine all of the data was also challenging. We looked through the data and experimented with different approaches to match fires to each other, match fires to weather data, and aggregate weather data and fire history data.

Creating the neural network from scratch was also challenging. Figuring out the dimensions of the matrices and the gradients for backpropagation was difficult. At first, the network wouldn’t survive training past 15-20 epochs because the gradients exploded, but it took a lot of time to figure out what exactly was causing it to crash with NaN weights. Regularization was able to fix that. Even once the network started working, it was slow to train and was much less accurate than TensorFlow models. We realized trying to use our hand-built model for all of the experiments was impractical due to time constraints and may give unclear results so we used TensorFlow for our main experiments.

# 8. Conclusions and Recommendations

## **8.1 Summary of key findings.**

We found that the weather\_history features we used offered very little improvement over the baseline features. Fire history\_features lead to great improvements in model performance. Training a model with weather\_history and fire\_history gave approximately the same results as without the weather\_history.

## **8.2 Conclusions drawn from the project.**

Based on the performance of our model on the various datasets, it would be best to train a model on the feature set that includes fire\_history features and excludes weather\_history.

## **8.3 Recommendations for future work or improvements.**

### 8.3.1 Dataset

We created a new fire prediction dataset that includes features not found on other publicly available datasets. We would like to expand this data set (e.g. with the average number of fires in a region) or alter some of the generated data (e.g. skewing the dates and times of negative examples towards the skew in the positive examples like we did for the locations). We would like to publish the dataset we created so others can use it for their own experiments.

### 8.3.2 Model Performance and Feature Sets

There is much more to explore to learn why the weather\_history features were not useful, why the fire\_history features were, and what other features might improve model performance. Related to weather\_history, future work could try to compile weather data from nearer the fire locations, use satellite imagery to infer weather conditions, or use other features like drought indices in a region. Regarding fire history, as mentioned in 7.1.2, future work could experiment with aggregation windows and means of representing fire history (e.g. number of fires vs acres burned). They could also investigate why fire history is impactful by trying to find other features that correlate strongly with it and that it may encompass. Future work should also consider other ablation studies with other features like vegetation and topology. The success of the fire\_history features we created opens up exciting new possibilities for other features that could be used to predict fires.

## 

# 9. References/Bibliography

Cal Fire. (2023). *Incident Data* (No version listed)[Data set]. Cal Fire. <https://www.fire.ca.gov/incidents#:~:text=FOR%20SOFTWARE%20DEVELOPERS-,Incident%20Data,-Incident%20data%20is>

Gutierrez, A. A., Hantson, S., Langenbrunner, B., Chen, B., Jin, Y., Goulden, M. L., & Randerson, J. T. (2021). Wildfire response to changing daily temperature extremes in California's Sierra Nevada. *Science advances, 7*(47), eabe6417. <https://doi.org/10.1126/sciadv.abe6417>

Holm, S. M., Miller, M. D., & Balmes, J. R. (2020). Health effects of wildfire smoke in children and public health tools: a narrative review. *Journal of Exposure Science and Environmental Epidemiology*, *31*(1), 1–20. <https://doi.org/10.1038/s41370-020-00267-4>

Kondylatos, S., Prapas, I., Ronco, M., Papoutsis, I., Camps-Valls, G., Piles, M., et al. (2022). Wildfire danger prediction and understanding with Deep Learning. *Geophysical Research Letters* *49*(e2022GL099368). <https://doi.org/10.1029/2022GL099368>

*Living under Smoky Skies—Understanding the challenges posed by wildfire smoke in California*. (2022, November 14). Legislative Analyst’s Office. Retrieved October 22, 2023, from <https://lao.ca.gov/Publications/Report/4644>

*Malik, A., Jalin, N., Rani, S., Singhal, P., Jain, S., & Gao, J.* (2021, October 1). Wildfire Risk Prediction and Detection using Machine Learning in San Diego, California*. IEEE Conference Publication | IEEE Xplore.* Retrieved October 22, 2023, from <https://ieeexplore.ieee.org/document/9604370/authors#authors>Malik

Menne, M. J., Durre, I., Korzeniewski, B., McNeill, S., Thomas, K., Yin, X., Anthony, S., Ray, R., Vose, R. S., Gleason, B. E. & Houston, T. G. (2012). *Global Historical Climatology Network - Daily (GHCN-Daily)*(3.30)[Data set]. NOAA National Climatic Data Center. doi:10.7289/V5D21VHZ [Accessed 10/19/2023] <https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C00861/html>

National Centers for Environmental Information, NESDIS, NOAA, U.S. Department of Commerce. (2005). *U.S. Local Climatological Data (LCD)*(No version listed)[Data set]. [Accessed 10/19/2023] <https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C00684/html#>

National Park Service (NPS). *Understanding Fire Danger.* Retrieved October 20, 2023, from <https://www.nps.gov/articles/understanding-fire-danger.htm>

Shmuel, A. & Heifetz, E. (2023). Developing novel machine-learning-based fire weather indices. *Machine Learning: Sci. Technology* *4*(015029) <https://doi.org/10.1088/2632-2153/acc008>

Short, K. C. (2017). *Spatial wildfire occurrence data for the United States, 1992-2015* *(FPA\_FOD\_20170508)*.(4th Edition)[Data set]. Fort Collins, CO: Forest Service Research Data Archive. <https://doi.org/10.2737/RDS-2013-0009.4> Accessed via Kaggle: Tatman, R. *1.88 Million US Wildfires* [Data set]. Kaggle. <https://www.kaggle.com/datasets/rtatman/188-million-us-wildfires/data>

*Statistics | CAL FIRE*. (n.d.). California Department of Forestry and Fire Protection. Retrieved October 22, 2023, from <https://www.fire.ca.gov/our-impact/statistics>

Tyler, J. (Director). (2022). *2022 Wildfire Activity Statistics*. California Department of Forestry and Fire Protection. Retrieved October 22, 2023, from <https://34c031f8-c9fd-4018-8c5a-4159cdff6b0d-cdn-endpoint.azureedge.net/-/media/calfire-website/misc-doc/2022redbookfinalada.pdf?rev=19e8d1a007884c34a181e8138fc7017e&hash=7BA59394712297355511D58394E7B8E2>

USFSEnterpriseContent. (2022). *National USFS Fire Occurrence Point (Feature Layer)* (Modified 2023-09-28)[Data set]. U.S. Forest Service. <https://catalog.data.gov/dataset/national-usfs-fire-occurrence-point-feature-layer-d3233>

Youssouf, H., Liousse, C., Roblou, L., Assamoi, E., Salonen, R. O., Maesano, C. N., Banerjee, S., & Annesi‐Maesano, I. (2014). Non-Accidental health impacts of wildfire smoke. *International Journal of Environmental Research and Public Health*, *11*(11), 11772–11804. <https://doi.org/10.3390/ijerph111111772>

Zaidi A. (2023). Predicting wildfires in Algerian forests using machine learning models. *Heliyon, 9*(7), e18064. <https://doi.org/10.1016/j.heliyon.2023.e18064>

# 10. Appendices

## 10.1 Appendix A: Literature Review Notes

### 10.1.1 Peer-Reviewed (Advanced) Approaches

* 1. [Wildfire Danger Prediction and Understanding With Deep Learning](https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2022GL099368)
  2. Type of model
     1. Random forest [RF], XGBoost, Long-Short Term Memory [LSTM], and Convolutional Long-Short Term Memory [convLSTM]
  3. Features Used
     1. Daily weather data from ERA-5 Land (Muñoz-Sabater et al., 2021) of maximum 2 m temperature, maximum wind speed, minimum relative humidity, total precipitation, maximum 2 m dewpoint temperature, and maximum surface pressure. The type of daily aggregations chosen represents the most fire-aggravating conditions.
     2. Satellite variables from MODIS including Normalized Difference Vegetation Index (NDVI; Didan, 2015), day and night Land Surface Temperature (LST; Wan et al., 2015).
     3. Soil moisture index from the European Drought Observatory (Cammalleri et al., 2017).
     4. Roads distance, waterway distance, and yearly population density from WorldPop (Tatem, 2017).
     5. Elevation and Slope from Copernicus EU-DEM (Bashfield & Keim, 2011).
     6. Ten variables with the fraction of classes from Copernicus Corine Land Cover (Büttner, 2014).
  4. Keypoints
     1. The study aims to predict wildfire danger using deep learning models. The models were trained on large datasets of meteorological data and satellite images. The study also explored the interpretability of the models to understand which features are most influential in predicting wildfire danger. The models are used over 2 test sets (2020 and 2021) and metrics (F1, recall, etc) provide a gauge on the performance of each.
     2. Drawbacks
        1. "The results are better for 2021 than for 2020 for all the models. Precision is higher, probably because negatives are more distinguishable. We sample negatives on days when no fire occurred, but few days did not have fires in the summer of 2021. Thus, more negatives are sampled outside the summer, which might make their classification easier."
     3. Strengths
        1. The deep learning models, especially LSTM and ConvLSTM, demonstrated superior performance with F1-scores greater than 0.8, outperforming traditional models like RF and XGBoost.
        2. The models showed strong generalization capabilities, especially for the extreme fire season of 2021, indicating their ability to adapt to varying conditions and predict significant fire events.
  5. [Wildfire Risk Prediction and Detection using Machine Learning in San Diego, California](https://ieeexplore.ieee.org/document/9604370/citations#citations)
  6. Type of model
     1. We used Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost), and Random Forest (RF) algorithms for weather based fire prediction.
     2. CNN, LTSM and SVM for remote sensing based fire prediction models.
     3. Combine them into an ensemble model.
  7. Features Used
     1. Weather data (temperature, humidity, wind speed, etc.) (SECTION IV.A)
     2. Satellite data (Landsat 8 images) (SECTION IV.B)
     3. Historical fire data (SECTION IV.C)
     4. Vegetation data (Normalized Difference Vegetation Index - NDVI) (SECTION IV.D)
  8. Keypoints
     1. Aimed to develop a machine learning and big data-based fire risk prediction model. Achieved an accuracy of 100% using the ensemble model for fire risk prediction. Achieved 93% accuracy using Faster R-CNN for fire detection. Utilized various machine learning and deep learning techniques to achieve these results.
     2. Drawbacks
     3. Strengths
        1. Multiple data sources are used and they combine 2 ML models into one ensemble model. Obtained accuracy of 100% with our ensemble model for fire risk prediction and 93% for fire risk detection.
        2. Unlike other research that examines either fire risk detection or fire risk prediction with limited data and parameters, our work focuses on understanding these concepts using past fire events, weather, remote sensing, and satellite data.
  9. [Predictive modeling of wildfires: A new dataset and machine learning approach](https://www.sciencedirect.com/science/article/abs/pii/S0379711218303941?fr=RR-2&ref=pdf_download&rr=819eb5e4e95ef953)
  10. Neural nets MLP Classifier, SVM
  11. Features Used
      1. This is based on satellite imagery. They used fire zones to classify the images as fire or no fire.
      2. Go to the appendix at the bottom and you can download the data they used.
  12. Keypoints
      1. General
      2. Drawbacks
         1. Future works will mainly consist of strengthening the model by including weather data. Weather plays a major role in the occurrence, growth, spread and the Extinction of wildfires. It can impact on the strength and movement of fire, and thus burn more land, which makes its extinction even more difficult. There are three weather parameters that can affect wildfires: Air Temperature, Wind and Soil Moisture.
         2. Air Temperature influences the occurrence of wildfires, by heating trees and crops on the ground, which makes them sensitive toward catching fire.
         3. The Wind has the most prominent and strongest impact on wildfires behaviors. Wind speed and direction are unpredictable. Besides, winds supply the fire with additional oxygen, which pushes the fire to move faster across the land.
         4. Soil Moisture is directly affected by precipitation and air humidity. When the soil moisture is low, the risk of wildfires is high. Conversely, high soil moisture lowers the chances of a wildfire igniting,
      3. Strengths
  13. [Developing novel machine-learning-based fire weather indices](https://iopscience.iop.org/article/10.1088/2632-2153/acc008)
  14. Type of Model
      1. We develop four different classification models: (i) RF (Biau and Scornet 2016). (ii) Extreme Gradient Boosting (XGBoost) (Chen and Guestrin 2016). (iii) MLP, a form of Neural Network (Ramchoun et al 2016), and (iv) logistic regression (Lever et al 2016). We perform a simple train-test split where 25% of the observations are used for testing. We perform these analyses using Python's Scikit-learn package (Pedregosa et al 2011 ), apart from the XGBoost model which is based on the XGBoost package (Chen and Guestrin 2016).
  15. Features Used
      1. Daily ignition, 2 m temperature, Relative humidity, 10-m wind speed, Precipitation, Mean relative humidity in previous month, Mean precipitation in previous month, Mean relative humidity in previous year, Mean precipitation in previous year, Mean slope, Population density, Normalized difference vegetation index, Incoming short-wave solar radiation, Daily fire weather index, Daily build up index, Daily danger index, Daily drought code, Daily duff moisture code, Daily initial fire spread index, Daily fine fuel moisture code, Daily fire daily severity rating, Daily Keetch-Byram drought index, Daily fire danger index, Daily spread component, Daily energy release component, Daily burning index, Daily ignition component, Regional wildfire history.
  16. Keypoints
      1. General
         1. To enable accurate wildfire danger predictions by applying advanced ML models using various data. To analyze the most significant factors affecting wildfire risk, both individually and in interaction with additional factors. To demonstrate the potential of applying ML-based FWIs in actual fire warning systems. To examine the contribution of adding traditional FWIs (in addition to the raw data) as input data when training ML-based models. To examine the potential of predicting extreme wildfires using ML-based models.
         2. There is a figure showing the most significant features and their effects on the predictions and another one showing how logistic regression performs on this data.
         3. ML models significantly outperformed traditional Fire Weather Indices (FWIs) and Logistic Regression models.
         4. XGBoost achieved the highest accuracy with an ROC-AUC of 0.98.
         5. Temperature was the most influential factor in predicting wildfire risk.
         6. Complex interactions between features like wind and humidity were captured better by ML models.
      2. Drawbacks
      3. Strengths
         1. This seems like it has some applicable information to our work, they are using some similar data and they are even training a logistic regression model on it. Of course they use some more sophisticated models that perform better than the simpler models and the traditional fire weather indices.
  17. [A review of machine learning applications in wildfire science and management](https://cdnsciencepub.com/doi/full/10.1139/er-2020-0019)
  18. Keypoint
      1. This is actually a scoping review that provides a good background on the topic and I will actually use it for the background section.
  19. [Deep Learning Methods for Daily Wildfire Danger Forecasting](https://arxiv.org/abs/2111.02736)
  20. Type of Model
      1. Four different types of dataset samples (pixel, temporal, spatial, and spatio-temporal) are extracted from the datacube and then fed to the corresponding models (RF, LSTM, CNN, and convLSTM). For a dataset sample, nf is the number of input features, days is the number of days in the time series, h is the height and w is the width, wherever applicable. The pixel dataset has a different number of input features (n′f ), because it also includes aggregations of the dynamic features.
  21. Features Used
      1. Deep Learning Methods for Daily Wildfire Danger Forecasting" study encompass daily weather data from ERA-5 Land, which includes 2 m temperature, 10 m wind u-component, 10 m wind v-component, and total precipitation. Additionally, satellite variables from MODIS are incorporated, such as Leaf Area Index (LAI), Fraction of Photosynthetically Active Radiation (Fpar), Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), and Day/Night Land Surface Temperature (LST). The dataset also integrates roads density derived from OpenStreetMap, population density from WorldPop, land cover information from Copernicus Corine Land Cover (CLC), topography variables like elevation, aspect, and slope from Copernicus EU-DEM, and historical burned areas data from the European Forest Fire Information System (EFFIS) combined with the MODIS active fires product to determine the fire's start date
  22. Keypoints
      1. General
         1. We approach daily fire danger prediction as a machine learning task, using historical Earth observation data from the last decade to predict next-day’s fire danger. To that end, we collect, pre-process and harmonize an open- access datacube, featuring a set of covariates that jointly affect the fire occurrence and spread, such as weather conditions, satellite-derived products, topography features and variables related to human activity.
         2. We implement a variety of Deep Learning (DL) models to capture the spatial, temporal or spatio-temporal context and compare them against a Random Forest (RF) baseline. We find that either spatial or temporal context is enough to surpass the RF, while a ConvLSTM that exploits the spatio-temporal context performs best with a test Area Under the Receiver Operating Characteristic of 0.926.
      2. Drawbacks
      3. Strengths
  23. [Predicting wildfires in Algerian forests using machine learning models](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10372657/)
  24. Type of Model
      1. We developed an artificial neural network (ANN) with two hidden layers to predict wildfires in these cities (diagram in section 5.1). Next, we trained and compared the performance of our classifier with those provided by the Logistic Regression, K Nearest Neighbors, Support Vector Machine, and Random Forest classifiers, using a 10-fold stratified cross-validation. The experiment shows a slight superiority of the ANN classifier compared to the others, in terms of accuracy, precision, and recall.
  25. Features Used
      1. The attributes in the dataset include Temperature at 12 noon in degrees Celsius (Temp), Relative Humidity in % (RH), Wind speed in km/h (Ws), 24 hours of accumulated precipitation in mm from noon to noon (Rain), Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Buildup Index (BUI), Fire Weather Index (FWI), and the target variable representing fire occurrence (Target), which takes binary values {fire: 1 | no fire: 0}.
  26. Keypoints
      1. General
         1. There are many images and diagrams summarizing the process, explaining the models and comparing the scores.
      2. Drawbacks
      3. Strengths
  27. [interpretable machine learning fire model for burned-area predictions over tropics](https://gmd.copernicus.org/articles/16/869/2023/)
  28. Type of Model
      1. Note code is available in an online repo.
      2. Burned areas? Is this predicting the spread of wildfires or whether they will start? I dont think it's relevant but I will keep it in.
      3. The AttentionFire model is an interpretable machine learning model based on an attention-augmented Long Short-Term Memory (LSTM) framework, which uses attention mechanisms to dynamically select and assign weights to important drivers and time steps for wildfire predictions, addressing limitations of traditional LSTM models in terms of interpretability and feature importance. (Diagram in figure 1)
      4. Five other widely used machine learning (ML) models are used as baseline models to compare with AttentionFire model: ANN (Joshi and Sukumar, 2021; Zhu et al., 2022), decision tree (DT) (Amatulli et al., 2006; Coffield et al., 2019), random forest (RF) (Yu et al., 2020; Li et al., 2018; Gray et al., 2018), gradient-boosting decision tree (GBDT) (Coffield et al., 2019; Jain et al., 2020), and naive LSTM (Liang et al., 2019; Natekar et al., 2021; Gui et al., 2021; Mei and Li, 2019)
  29. Features Used
      1. Go to table 2 in section 2.3 to see the table of variables.
  30. Keypoints
      1. General
      2. Drawbacks
      3. Strengths
  31. [Fire danger forecasting using machine learning-based models and meteorological observation: a case study in Northeastern China](https://link.springer.com/article/10.1007/s11042-023-15881-1)
  32. Type of Model
      1. LTSM based time series prediction and Random Forest based fire danger occurrence probability.
  33. Features Used
      1. From 2004 to 2015, 3963 fire occurrences were recorded in Northeastern China, comprising 1088 natural wildfires (including forest, grassland, and shrubland fires) and 2875 human-prescribed burnings; data visualizations indicated peak occurrences in 2007 and more frequent fires during spring and summer.
      2. The study employed the Canadian Forest Fire Weather Index (FWI) system, which uses four key meteorological data types—temperature, relative humidity, wind speed, and 24-hour precipitation—sourced from daily observations between 1989 to 2019 at approximately 133 meteorological stations in the region, with fire weather data matched to each fire event based on its date and nearest meteorological station.
  34. Keypoints
      1. General
         1. Proposed a two-stage fire danger rating and forecasting model.
         2. First stage predicts future FWI system indexes using LSTM.
         3. Second stage uses Random Forest to determine fire danger occurrence probability and presents a classification scheme.
         4. Verified the scheme using forest fire danger data from Qipan Mountain.
         5. Fire danger forecasting is complex, and while traditional binary classification methods fall short, our two-stage prediction method uses the FWI system indexes from weather data to create a more nuanced classification, dividing fire danger into four categories.
      2. Drawbacks
         1. The study's model predicts fire danger ratings up to four days in advance, which is an improvement over typical short-term predictions but could benefit from further extension.
         2. The model's applicability is currently limited to Qipan Mountain and Northeastern China due to data collection constraints, but future work aims to expand and verify its effectiveness in other regions.
      3. Strengths
  35. [Predicting California Wildfire Risk with Deep Neural Networks](http://cs230.stanford.edu/projects_fall_2021/reports/103174984.pdf)
  36. Type of Model
      1. Conv Neural Networks (CNNs)
      2. Specifically, pretrained ImageNet models: MobileNetV2 and ResNetV2
  37. Features Used
      1. High-resolution satellite imagery of the California countryside.
      2. Historical fire map data from CalFire.
      3. NSGS Landsat 7 and 8 imagery for California over the last 10 years.
  38. Keypoints
      1. General
         1. The goal is to identify seasonal fire risk in California using satellite imagery.
         2. The models classify terrain based on historical images of areas that experienced wildfires.
         3. The study faced challenges due to high problem bias and applied techniques to reduce this bias.
         4. The input is a satellite image, and the output predicts if the area is at risk of wildfire for the season.
      2. Drawbacks
         1. Initial struggles with effective identification due to high problem bias.
         2. Data imbalance with the not-at-risk class being heavily overrepresented.
         3. Some satellite images were obscured by cloud cover.
         4. The model had difficulty identifying positive classifications.
         5. Fires caused by human incidents might not have identifiable features in satellite images.
      3. Strengths
         1. Utilized advanced CNNs like MobileNetV2 and ResNetV2.
         2. Incorporated learning rate optimization, channel feature addition, and cost weighting to improve model performance.
         3. Demonstrated the potential of AI-based techniques for estimating fire risk in California.
  39. [Identifification and Characterization of Forest Fire Risk Zones Leveraging Machine Learning Methods](https://scholar.smu.edu/cgi/viewcontent.cgi?article=1188&context=datasciencereview)
  40. Type of Model
      1. Binary Classification
  41. Features Used
      1. Historical weather data (temperature, precipitation, wind, etc.)
      2. Past fire occurrences
      3. Population density
      4. MGRS (Military Grid Reference System) coordinates
      5. Historical drought data
      6. Hours of sunlight per day
  42. Keypoints
      1. General
         1. They also have a feature importance figure.
         2. Literature review also included.
         3. Also they implemented a coordinate/grid system in this paper.
         4. The study focuses on the Southern California region.
         5. The model predicts the probability of a fire occurring in a specific area within the next one to five days.
         6. The model uses the MGRS grid system to define geographic areas and correlate multiple fires to a specific area.
         7. The study used a Kaggle dataset with 1.88 million georeferenced fires collected over 24 years.
         8. The model's predictions were visually inspected for accuracy by comparing them to known fire-prone areas.
      2. Drawbacks
         1. The model's predictions remain relatively constant throughout the year, not adjusting well for seasonal variations.
         2. The study is limited to the Southern California region.
         3. Ethical concerns include potential misuse by arsonists or for insurance fraud.
      3. Strengths
         1. The model can be a valuable tool for government and fire officials for proactive fire prevention measures.
         2. The study highlights the importance of feature engineering and the potential for enhancing current fire detection methods.
         3. The model's results align with historically known fire-prone areas, indicating its reliability.
  43. [Combining precipitation forecasts and vegetation health to predict fire risk at subseasonal timescale in the Amazon](https://iopscience.iop.org/article/10.1088/1748-9326/ac76d8/meta)
  44. Type of Model
      1. The model is a hybrid combining both dynamic and statistical methods, based on the NextGen prediction system.
      2. Its main goal is to predict fire risk in the Amazon.
      3. Fire risk is determined by the likelihood of active fires occurring due to precipitation and vegetation health.
      4. The model uses VIIRS active fires data and predicts based on SubX's week-2 precipitation forecasts from 2017-2021.
      5. A secondary experiment adds the NOAA Vegetation Health Index (VHI) as an extra predictor, which detects how ecosystems respond to weather conditions.
      6. The study then examines the model's accuracy in two specific Amazon regions with different fire seasons.
      7. It also investigates how different land covers in these regions affect the model's predictions.
  45. Features Used
      1. VIIRS Active Fires: Active fire detections.
      2. SubX MME Week-2 Precipitation: Rainfall predictions two weeks ahead (2017-2021).
      3. NOAA Vegetation Health Index (VHI): Satellite-based vegetation health measure.
      4. Mapbiomas Land Cover (2019): Brazil's Amazon land cover data.
      5. Fire Activity (FA): Most active fire trimester data from VIIRS.
      6. SubXPr and VHI Weekly Data: 13-week data within the peak fire trimester.
      7. Fire Probability Analysis: Logistic regression models using SubXPr and VHI.
      8. Land Cover Data (2019): Predominant land types in the Amazon from Mapbiomas.
  46. Keypoints
      1. General
         1. The study focuses on bridging the gap between atmospheric scientists and fire ecologists by combining observed states of vegetation with precipitation forecasts.
         2. The research identifies areas in the Amazon where fire risk can be skillfully predicted using SubXPr and VHI.
         3. In areas dominated by forests, fire risk is mainly driven by SubXPr. In contrast, in areas with prevalent savannas and grasslands, the status of VHI plays a significant role in determining fire risk.
         4. The study emphasizes the potential of using subseasonal forecasts in combination with vegetation health to forecast fire risk, providing crucial lead time for fire management resources allocation.
      2. Drawbacks
         1. The role of vegetation health as a predictor of fire risk is modest in areas dominated by forests.
         2. The study suggests that continuous improvement of statistical models' accuracy might be achieved using higher spatial resolution satellite vegetation conditions.
      3. Strengths
  47. [Improved Prediction of Forest Fire Risk in Central and Northern China by a Time-Decaying Precipitation Model](https://www.mdpi.com/1999-4907/13/3/480)
  48. Type of Model
      1. Support Vector Machine (SVM) regression model
      2. Time-decaying precipitation algorithm
  49. Features Used
      1. Relative humidity
      2. Daily maximum temperature
      3. Daily maximum wind speed
      4. Comprehensive precipitation index
      5. NDVI (Normalized Difference Vegetation Index)
      6. VSWI (Vegetation Supply Water Index)
      7. Altitude
      8. Fire density
      9. Slope
  50. Keypoints
      1. General
         1. Four forest areas in central and northern China were selected for training and testing.
         2. The model uses a time-decaying precipitation algorithm instead of traditional equal-weighting.
         3. Gaussian convolution was used to convert discrete fire spots into continuous forest fire density data.
         4. The model achieved accuracies of 0.98, 0.99, 0.95, and 0.93 in the four test areas.
      2. Drawbacks
         1. The study doesn't mention the use of deep learning, which might provide better accuracy with sufficient samples.
         2. The model might require re-optimization for different geographical and climatic conditions.
      3. Strengths
         1. The model improved fire risk prediction accuracy by approximately 10%.
         2. SVM regression model outperformed logistic regression and ANN models.
         3. The model is applicable to diverse climatic and geographical conditions, demonstrating its universality.

### 

### 10.1.2 Individual (Fundamental) Approaches

Kaggle Competitions

1. <https://www.kaggle.com/competitions/forest-fire-prediction/data>
   1. Data: day, month, year, temp, relative humidity, wind speed, rain, fine fuel moisture code, duff moisture code, drought code, initial spread index, buildup index, fire weather index
   2. approaches: I wasn’t able to see the competition code or discussion, but that may be because I wasn’t logged in. leaderboard scores are at 1.00000, but idk what that means
2. <https://www.kaggle.com/competitions/ausdm-student-competition-2015/data>
   1. Data: temp min/max, relative humidity min/max, feels like, rain since 9 am, wind speed min/max/avg, wind gust
   2. approaches: again, not able to see. best score on leaderboard was .58050, but idk how the score was calculated
   3. [Wildfires in California prediction (Kaggle)](https://www.kaggle.com/code/sandstorm0123/wildfires-in-california-prediction/notebook)
   4. Type of Model
      1. Machine Learning Model to predict wildfires in California based on weather events.
   5. Features Used
      1. US Weather Events (2016-2019)
         1. EventId, Type, Severity, StartTime(UTC), EndTime(UTC), TimeZone, AirportCode, LocationLat, LocationLng, City, County, State, ZipCode.
      2. California Wildfire Incidents (2013-2020)
         1. Same data we are using
   6. Keypoints
      1. General
         1. The goal is to combine weather events at different stations with wildfire incidents to uncover less obvious relations.
         2. The current version visualizes rain events and acres burned by wildfire.
         3. The model tries to predict if 2019 was a year with a lot of wildfires based on weather data.
         4. The model is trained on data from 2016 to 2018 and tested on 2019 data.
         5. Uses TensorFlow for creating and training the model.
      2. Drawbacks
         1. The approach is still in a rough state and not finished.
         2. The model is specialized in only three years and might have overfitting issues when applied to other years.
      3. Strengths
   7. [AI Geospatial Wildfire Risk Prediction (Kaggle)](https://www.kaggle.com/code/thjaquenoud/ai-geospatial-wildfire-risk-prediction)
   8. Type of Model
      1. Deep Learning Model: U-Net for multi-class semantic segmentation.
   9. Features Used
      1. MODIS: Daily U.S. coverage at 500m resolution.
      2. GRIDMET: Daily meteorological data at 4000m resolution, including temperature, humidity, and drought indices.
      3. LANDFIRE MFRI: 30m pixel categorization based on historical wildfire intervals. Useful for regional trends but static over time.
      4. USDA Cropland Data Layer: 30m resolution land cover map. Used to differentiate cultivated from non-cultivated areas.
      5. Output - Wildfire Hazard Potential (WHP) map: USDA's evaluation of wildfire hazard, available for 2014, 2018, and 2020.
      6. Processing: Data aligned using GDAL and segmented into smaller images for model processing.
      7. Final dataset: 13-layer geospatial images over the CONUS for three years.
   10. Keypoints
       1. General
          1. Objective: Generate wildfire risk maps weekly over the CONUS.
          2. U-Net architecture adapted for the specific input dimensions and multi-class output.
          3. Model trained with a weighted loss function to account for class imbalances.
          4. Achieved 65% accuracy on training and validation data.
          5. Model generalizes to months outside of its training domain, e.g., predicting wildfire hazard in winter months.
       2. Drawbacks
          1. Difficulty in obtaining labeled data for "at-risk areas."
          2. Domain gap due to the use of WHP map which is generated once every few years.
          3. Model sometimes underestimates the highest level of risk and overestimates in certain areas.
          4. Noisy outputs and grid-like artifacts in winter predictions.
       3. Strengths
          1. Demonstrates the potential of using geospatial data in deep learning models for large-scale tasks.
          2. Model can recreate the WHP map for years when it wasn't available.
          3. Evidence of model learning to generalize temporal differences from geographic ones.
   11. [Brazilian Wildfire Prediction (Kaggle)](https://www.kaggle.com/code/data13/brazilian-wildfire-prediction/input)
   12. Type of Model
       1. Linear Regression
   13. Features Used
       1. Year: When Forest Fires happened (Range: 1998-2017)
       2. State: Brazilian State (e.g., Rio, Mato Grosso)
       3. Month: Month when Forest Fires happened (e.g., Janeiro, Fevereiro)
       4. Number: Number of Forest Fires reported (Range: 0-998)
       5. Date: Date when Forest Fires were reported (Range: 01/01/1998 - 01/01/2017)
   14. Keypoints
       1. General
          1. The dataset reports the number of forest fires in Brazil divided by states over approximately 10 years (1998 to 2017).
          2. Data Exploration revealed that the dataset is structured and almost cleaned.
          3. Data Cleaning involved removing duplicates, changing month names to English, and handling special characters in state names.
          4. Exploratory Data Analysis (EDA) showed the distribution of fires per year, month, and state.
          5. Data Preprocessing included encoding categorical data, extracting new features like day of the week, and determining weekends.
          6. The Linear Regression model was used to predict the number of wildfires in the Amazon rainforest.
       2. Drawbacks
          1. The accuracy of the model is not satisfactory. The dataset's small size leads to overfitting, resulting in a low training error and a high test error.
       3. Strengths
          1. The dataset provides a comprehensive overview of forest fires in Brazil over a significant period.
          2. The notebook includes a thorough exploratory data analysis, visualizations, and preprocessing steps.
   15. [Australian Wildfires Prediction (Kaggle)](https://www.kaggle.com/code/ahmednobi/australian-wildfires-prediction/input)
   16. Type of Model
       1. RandomForestRegressor
   17. Features Used
       1. Latitude
       2. Longitude
       3. Brightness
       4. Scan and Track (pixel sizes; later dropped 'track')
       5. Acquisition Date and Time
       6. Satellite (Terra and Aqua)
       7. Confidence (target variable)
       8. Brightness temperature 31 (dropped during feature engineering)
       9. Fire Radiative Power
       10. Day/Night
       11. Type (encoded into type\_0, type\_2, and type\_3 during preprocessing)
   18. Keypoints
       1. General
          1. The dataset has over 36,000 entries and 15 initial features.
          2. Features such as 'track', 'instrument', and 'version' were dropped based on correlation.
          3. Features 'daynight' and 'satellite' were converted to binary values; 'type' was one-hot encoded.
          4. Derived year, month, and day from 'acq\_date' for better feature representation.
          5. The initial RandomForestRegressor had a testing accuracy of 64.66%.
          6. After hyperparameter tuning, the testing accuracy improved to 66.61%.
       2. Drawbacks
          1. Model might be overfitting given the high training accuracy compared to testing accuracy.
          2. Some features like 'bright\_t31' and 'type\_0' were dropped without clear justification.
       3. Strengths
          1. Extensive data preprocessing to enhance model performance.
          2. Hyperparameter tuning helped in improving the testing accuracy.
          3. The model captures a significant amount of variance in the training set.

## 10.2 Feature Sets

The feature sets on which we trained models are:

* minimal (not even the baseline in our proposal):
  + incident created year
  + incident created month
  + incident created day
  + incident created hour
  + incident created minute
  + incident latitude
  + incident longitude
* baseline (the baseline from our proposal):
  + minimal features plus
    - sum precipitation for the day
    - distance to the sum precipitation measurement station
    - max dry bulb temperature
    - distance to the max dry bulb temperature measurement station
    - min dry bulb temperature
    - distance to the min dry bulb temperature measurement station
    - mean dry bulb temperature
    - distance to the mean dry bulb temperature measurement station
    - mean dew point temperature
    - distance to the mean dew point temperature measurement station
    - mean wet bulb temperature
    - distance to the mean wet bulb temperature measurement station
    - mean wind speed
    - distance to the mean wind speed measurement station
    - mean relative humidity
    - distance to the mean relative humidity measurement station
    - min relative humidity
    - distance to the min relative humidity measurement station
    - max relative humidity
    - distance to the max relative humidity measurement station
    - max wind speed
    - distance to the max wind speed measurement station
    - circular mean wind direction
    - distance to the circular mean wind direction measurement station
    - mode wind direction
    - distance to the mode wind direction measurement station
* weather\_history:
  + baseline features plus
  + 10-, 30- and 60-day averages for
    - sum precipitation
    - max dry bulb temperature
    - min dry bulb temperature
    - mean dry bulb temperature
    - mean dew point temperature
    - mean wet bulb temperature
    - mean wind speed
    - mean relative humidity
    - min relative humidity
    - max relative humidity
    - max wind speed
    - circular mean wind direction
    - mode wind direction
* fire\_history:
  + baseline features plus
    - average acres burned per year in the last 3 years within 5 km
    - average acres burned per year in the last 20 years within 15 km
* complete/all:
  + baseline
  + weather\_history
  + fire\_history