**Statistical Analysis of US Wildfire**

INFO 6105 Data Science Engineering Methods

**Team 13**

Shuai Shao: 001345385

Haonan Quan: 001305566

Lu Liu: 001087120

**Table of Contents**

[Introduction 2](#_Toc57927347)

[Obejctives 3](#_Toc57927348)

[Dataset 5](#_Toc57927349)

[**Description** 5](#_Toc57927350)

[**Data Source** 6](#_Toc57927351)

[Importing Data 6](#_Toc57927352)

[Visualization 8](#_Toc57927353)

[Correlation 13](#_Toc57927354)

[Algorithms & Implementation 14](#_Toc57927355)

[**Linear Regression** 15](#_Toc57927356)

[**Gaussian Naive Bayes** 17](#_Toc57927357)

[**Logistic Regression** 18](#_Toc57927358)

[**Decision tree** 19](#_Toc57927359)

[**Random Forest** 20](#_Toc57927360)

[Conclusion 24](#_Toc57927361)

[Reference 25](#_Toc57927362)

# Introduction

A wildfire is an unplanned, unwanted, uncontrolled fire in an area of combustible vegetation starting in rural areas and urban areas. Depending on the type of vegetation present, a wildfire can also be classified more specifically as a forest fire, brush fire, bushfire ([in Australia](https://en.wikipedia.org/wiki/Bushfires_in_Australia)), desert fire, grass fire, hill fire, peat fire, prairie fire, vegetation fire, or veld fire.

Just before the COVID-19 pandemic, wildfires in Australia destroyed more than 3,000 homes and burnt millions of hectares of vegetation.[1] 3 month ago, California suffered wildfires and it is still ongoing. As of November 12, 2020, over 9,177 fires[[2]](https://en.wikipedia.org/wiki/2020_California_wildfires#cite_note-CalFire_Stats-1)have burned 4,359,517 acres (1,764,234 ha),[[3]](https://en.wikipedia.org/wiki/2020_California_wildfires#cite_note-Large_Incident_Report-2) more than 4% of the state's roughly 100 million acres of land, making 2020 the largest wildfire season [recorded in California's modern history](https://en.wikipedia.org/wiki/List_of_California_wildfires) (according to the [California Department of Forestry and Fire Protection](https://en.wikipedia.org/wiki/California_Department_of_Forestry_and_Fire_Protection)), though roughly equivalent to the pre-1800 levels which averaged around 4.4 million acres yearly and up to 12 million in peak years [4][5].

Wildfires affect air quality, vegetation, human and animal habitats, and climate around the world. The intensity of the fires has been increased by drying and heating from [human-induced climate change](https://en.wikipedia.org/wiki/Climate_change),[[6]](https://en.wikipedia.org/wiki/2020_California_wildfires#cite_note-:0-5)[[7]](https://en.wikipedia.org/wiki/2020_California_wildfires#cite_note-techrev20200820-6) as well as decades of poor [forest management](https://en.wikipedia.org/wiki/Forest_management).

One of the most important ways to prevent wildfires is to predict the cause of wildfires and implement precise preventive measures. It is vital for the local fire department, state department to know how to take effective actions based on the causes of wildfires.

Statistical Analysis is one way to accomplish this important task.

Fortunately, it is a world where we could collect a huge amount of data. This has allowed us to use statistical machine learning to analyze and predict the causes of wildfires in the US.

In this project, we will show you how to analyze the dataset, visualize the data and implement various Machine Learning Algorithms to classify the causes of wildfires. The whole model will take the fire part of this dataset and provide percentage ratings for all possible reasons that could cause wildfires.

# Obejctives

* Listing out wildfire frequency from 1992 to 2015
* Exploratory Data Analysis (EDA) and Data visualization
* Figuring out the most and least fire-prone locations
* Implement Machine Learning Algorithms and predict the causes of wildfires

# Dataset

## **Description**

According to Kaggle, this data publication contains a spatial database of wildfires that occurred in the United States from 1992 to 2015. It is the third update of a publication originally generated to support the national Fire Program Analysis (FPA) system. The wildfire records were acquired from the reporting systems of federal, state, and local fire organizations.

表格

描述已自动生成Below are some of the main information:

Figure 1: Reading Data

Where each of the Columns stands for:

* FIRE\_CODE = Code used within the interagency wildland fire community to track and compile cost information for emergency fire suppression (<https://www.firecode.gov/)>.
* FIRE\_NAME = Name of the incident
* FIRE\_YEAR = Calendar year in which the fire was discovered
* DISCOVERY\_DATE = Date on which the fire was discovered
* DISCOVERY\_DOY = Day of year on which the fire was discovered
* STAT\_*CAUSE\_*CODE = Code for the (statistical) cause of the fire.
* STAT\_*CAUSE\_*DESCR = Description of the (statistical) cause of the fire.
* FIRE\_SIZE = Estimate of acres within the final perimeter of the fire.
* FIRE\_*SIZE\_*CLASS = Code for fire size based on the number of acres within the final fire perimeter expenditures (A= 0-0.25 acres, B=0.26-9.9 acres, C=10.0-99.9 acres, D=100-299 acres, E=300 to 999 acres, F=1000 to 4999 acres, and G=5000+ acres).
* LATITUDE = Latitude for point location of the fire (decimal degrees).
* LONGITUDE = Longitude for point location of the fire (decimal degrees).
* STATE = Two-letter alphabetic code for the state in which the fire burned (or originated)
* COUNTY = County, or equivalent, in which the fire burned (or originated)

**Data Source:** <https://www.kaggle.com/rtatman/188-million-us-wildfires>

# Methodology

This dataset has a huge amount of data, we should focus on the table named “fire” since it contains the causes and other information of wildfires.

In this project we should focus on the Fire Year, State, Fire\_Size, Longitude, Latitude, Day of the week, Month fields as our parameters.

Here are the algorithms we want to use.

1. Linear Regression

Linear regression is the next step up after correlation. It is used when we want to predict the value of a variable based on the value of another variable.

1. Gaussian Naive Bayes

Naive Bayes is a very handy, popular and important Machine Learning Algorithm especially for Text Analytics and General Classification.

1. Logistic Regression

Logistic regression (LR) is a statistical method similar to [linear regression](https://www.sciencedirect.com/topics/medicine-and-dentistry/linear-regression-analysis) since LR finds an equation that predicts an outcome for a binary variable

1. Decision Tree

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

1. Random Forest

Random forest algorithm can be used for both classifications and regression tasks. It provides higher accuracy through cross validation. Random forest classifier will handle the missing values and maintain the accuracy of a large proportion of data

With the algorithms above, we want to find a suitable algorithm to predict the causes of wildfires. We will show the implementation and explanation in the Algorithms & implementation part.

# Importing Data

We implemented sqlite3 to import the dataset and show the head.

表格

描述已自动生成

Figure 2: Importing Data

We then selected some of the data we thought useful (fire\_year, cause, etc.) and saved them.

表格

描述已自动生成

Figure 3: Select Useful Data

We also transferred the date format to Julian date for better understanding. 图片包含 文本

描述已自动生成

Figure 4: Transfer Date to Julian

# Visualization

We created a timeline and chart of fires per year. It showed that after the peak of 2006, total numbers of wildfires per year present a decreasing trend.

图表, 折线图

描述已自动生成

Figure 5: Timeline and Fires Per Year (5 Years Gap)

图表, 条形图

描述已自动生成

Figure 6: Timeline and Fires Per Year (1 Year Gap)

We then tried to analyze at which latitude and longitude did wildfire happen most, getting the result of around (34, -80).

图表

描述已自动生成

Figure 7: Latitude and Longitude Analyzing

图表, 散点图

描述已自动生成

Figure 8: Scatterplot of Latitude and Longitude

After creating the chart of the total fires for each state, we found CA, GA, and TX were the top three, which verified our analysis of latitude and longitude.

图表, 条形图

描述已自动生成

Figure 9: Top Three States in The Number of Wildfires

But what caused this high probability of wildfire? We decided to pick the top three states and analyze the causes.

图表, 条形图

描述已自动生成

Figure 10: Number of Wildfires with Causes in California

图表, 条形图, 瀑布图

描述已自动生成

Figure 11: Number of Wildfires with Causes in Georgia

图表, 条形图

描述已自动生成

Figure 12: Number of Wildfires with Causes in Texas

We also summed the total amounts of fire causes, but it seemed difficult to summarize the pattern of causes, so we picked fire cause as our next topic and did some algorithms & implementation to analyze and predict wildfires.

表格

描述已自动生成

Figure 13: Total Number of Wildfire with Different Causes

# Correlation

We transferred the states and causes to numeric numbers.

表格

描述已自动生成

Figure 14: States and Causes with Numeric Numbers

We then tried to calculate the correlations but found out the dataset didn’t show a strong correlation. The next step was to implement some algorithms to analyze the data.

图表

描述已自动生成

Figure 15: Correlations Among Parameters

# Algorithms & Implementation

**Algorithms:**

* **Linear Regression**
* **Gaussian Naive Bayes**
* **Logistic Regression**
* **Decision Tree**
* **Random Forest**

We have successfully implemented the algorithms above. A detailed explanation is provided in next pages

## **Linear Regression**

Our first try was Linear Regression. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.[8]

For our model, we employed the dataset used before for correlation, and added a column named month. We then dropped the null values and showed the header information.

表格

描述已自动生成

Figure 16: Data without Null and Date

We then picked the data and trained them, calculating the mean squared error, coefficient and score.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

Figure 17: Result of Linear Regression

We plotted the diagram and calculated the score, which didn’t show a satisfying result. We had to jump to the next algorithm, Gaussian Naive Bayes.

图片包含 图表

描述已自动生成

Figure 18: Diagram of The Linear Regression

## **Gaussian Naive Bayes**

As the name suggests, Gaussian Naïve Bayes classifier assumes that the data from each label is drawn from a simple Gaussian distribution. The Scikit-learn provides sklearn.naive\_bayes. GaussianNB to implement the Gaussian Naïve Bayes algorithm for classification.[9]

We followed the steps from lectures, imported the Gaussian Naive Bayes model, created a Gaussian Classifier, trained the model and predicted the response for the test dataset. We finally came to an accuracy of around 25%, which wasn’t what we wanted. Let’s go to the next one.

文本

描述已自动生成

Figure 19: Implementation of Gaussian Naive Bayes

表格

描述已自动生成

Figure 20: Result of Gaussian Naive Bayes

## **Logistic Regression**

Logistic regression (LR) is a statistical method similar to linear regression since LR finds an equation that predicts an outcome for a binary variable, *Y,* from one or more response variables, *X*. However, unlike linear regression the response variables can be categorical *or* continuous, as the model does not strictly require continuous data.[10]

We worked with the dataset while regularization strength C set to 10.0 instead of the default value. We then trained and evaluated the model, getting an accuracy of around 31%. We weren’t satisfied with that value either, so we went to the next model.

文本

描述已自动生成

Figure 21: Implementation of Logistic Regression

表格

描述已自动生成

Figure 22: Result of Logistic Regression

## **Decision tree**

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g., whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.[11]

Implementation:

Listed out some of the data required, transferred some data to numeric numbers.

表格

描述已自动生成

Figure 23: Data Preprocessing

We then trained the data and came out with a score of 33%. It still wasn’t what we want, so we had to turn to our last hope: random forest.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

Figure 24: Decision Tree Implementation

图形用户界面, 文本

描述已自动生成

Figure 25: Result of Decision Tree

## **Random Forest**

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction.[12]

Implementation:

文本, 信件

描述已自动生成

Figure 26: Implemenataion of Random Forest

The 58% score was much better than before. We wanted to focus a bit more on fire causes, so we split the causes into 4 classes: natural, accidental, malicious and other.

图片包含 表格

描述已自动生成

Figure 27: Four Classes of Wildfire Reasons

Then, we did a random forest test based on the new dataset, which gave a 70% score.

文本, 信件

描述已自动生成

Figure 28: Result of Random Forest

Seems that we are on the right track. Let's do some visualization and see what we have now:

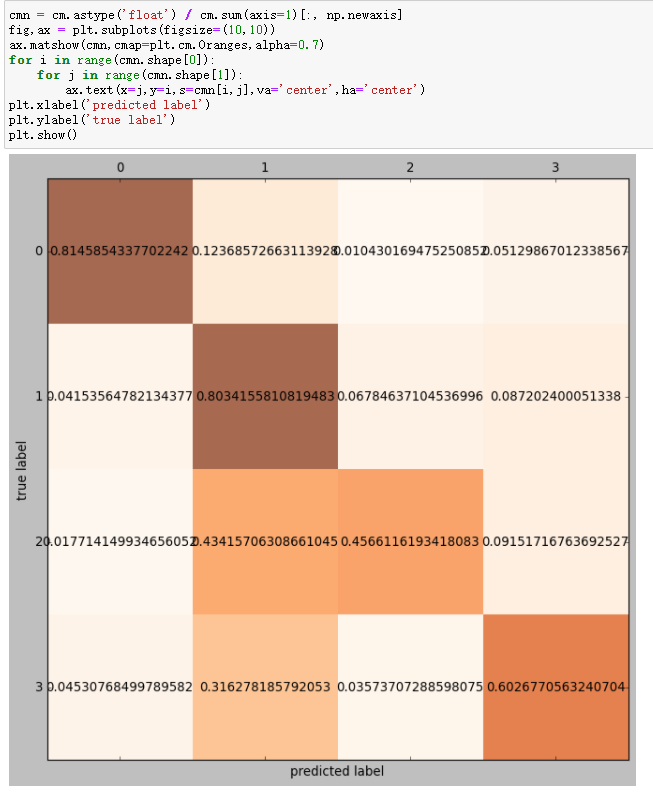


Figure 29: Random Forest Visualization

The first two labels, natural and accidental, worked well with around 80%. In a word, we could predict the cause of these wildfires, to an accuracy of 58% or better. Reducing the number of labels improved the prediction score to 70% while using random forest.

Our next step was to see if we could predict malicious fires in one state. We chose the top 3 states of wildfires: CA, GA and TX.

Implementaion:

表格

描述已自动生成

Figure 30: New Field Creation

CA:

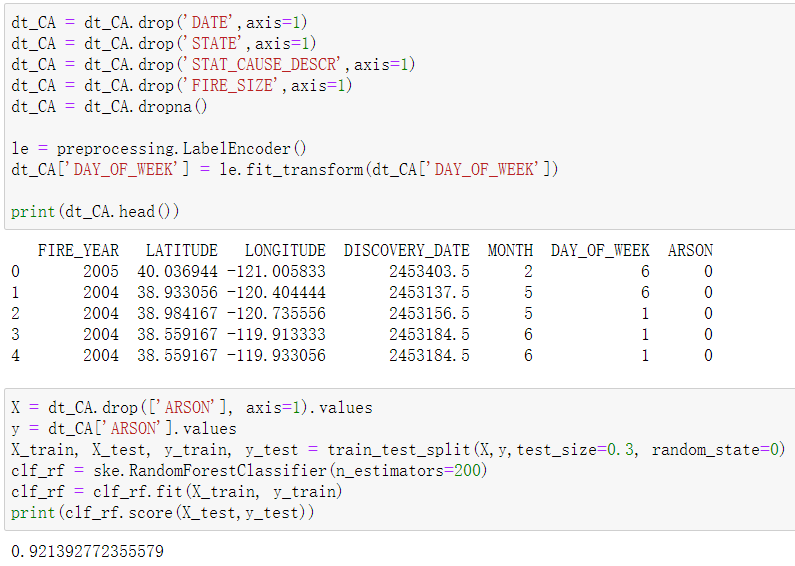


Figure 31: Prediction of California With Arson

GA:

文本, 信件

描述已自动生成



Figure 32: Prediction of Georgia With Arson

TX:

文本, 信件

描述已自动生成



Figure 33: Prediction of Texas With Arson

As we can see, we are able to predict malicious fires in these states, with an average accuracy of 90%.

# Conclusion

We have performed extract data by querying the partial columns of the fire table and we plotted the data to help us better understand the meaning behind the scenes. Post that, we implemented different Supervised Machine Learning Algorithms to check the ability of these models to learn from the training data set and make correct predictions on the test data set.

Among all the 6 models used in this project, Random Forest turns out to be the best model among them with a 70% predicted probability score. The model did badly at the malicious part, so in the next step, we did a more precise work by focusing on one state to determine whether the cause of wildfires in this state is malicious. Then we got a 90% predicted score.

Overall, Random Forest is a good model to predict the causes of US Wildfires.

# Reference

1. Bowman, D., Williamson, G., Yebra, M., Lizundia-Loiola, J., Pettinari, M., Shah, S., . . . Chuvieco, E. (2020, August 12). *Wildfires: Australia needs national monitoring agency*. Retrieved November 12, 2020, from <https://www.nature.com/articles/d41586-020-02306-4>
2. [*Fire Statistics*](https://www.fire.ca.gov/incidents/2020/). [CAL FIRE](https://en.wikipedia.org/wiki/CAL_FIRE). October 23, 2020. Retrieved October 27, 2020.
3. [2020 National Large Incident Year-to-Date Report](https://gacc.nifc.gov/sacc/predictive/intelligence/NationalLargeIncidentYTDReport.pdf) (PDF). Geographic Area Coordination Center (Report). [National Interagency Fire Center](https://en.wikipedia.org/wiki/National_Interagency_Fire_Center). October 27, 2020. Retrieved October 27, 2020.
4. Holly Yan; Cheri Mossberg; Artemis Moshtaghian; Paul Vercammen (September 6, 2020). [*California sets new record for land torched by wildfires as 224 people escape by air from a 'hellish' inferno*](https://www.cnn.com/2020/09/05/us/california-mammoth-pool-reservoir-camp-fire/index.html). [Cable News Network](https://en.wikipedia.org/wiki/Cable_News_Network). Retrieved September 6, 2020.
5. Silverman, Hollie (September 8, 2020). [*California wildfires have burned more than 2 million acres and prompted power outages for more than 170,000*](https://www.cnn.com/2020/09/08/us/california-fires-tuesday/index.html). CNN. Retrieved September 8, 2020.
6. Williams, A. Park; Abatzoglou, John T.; Gershunov, Alexander; Guzman‐Morales, Janin; Bishop, Daniel A.; Balch, Jennifer K.; Lettenmaier, Dennis P. (2019). [*Observed Impacts of Anthropogenic Climate Change on Wildfire in California*](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019EF001210). Earth's Future. 7 (8): 892–910. [doi](https://en.wikipedia.org/wiki/Doi_(identifier)):[10.1029/2019EF001210](https://doi.org/10.1029%2F2019EF001210). [ISSN](https://en.wikipedia.org/wiki/ISSN_(identifier)) [2328-4277](https://www.worldcat.org/issn/2328-4277).
7. Temple, James (August 20, 2020). [*Yes, climate change is almost certainly fueling California's massive fires*](https://www.technologyreview.com/2020/08/20/1007478/california-wildfires-climate-change-heatwaves/). MIT Technology Review.
8. (n.d.). Retrieved December 03, 2020, from <http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm>
9. *Scikit Learn - Gaussian Naïve Bayes. (n.d.).* Retrieved December 03, 2020, from <https://www.tutorialspoint.com/scikit_learn/scikit_learn_gaussian_naive_bayes.htm>
10. Hoffman, J. I. (2019). Logistic Regression Analysis. Retrieved December 03, 2020, from <https://www.sciencedirect.com/topics/medicine-and-dentistry/logistic-regression-analysis>
11. *Decision tree.* (2020, October 13). Retrieved December 03, 2020, from <https://en.wikipedia.org/wiki/Decision_tree>
12. Yiu, T. (2019, August 14). *Understanding Random Forest*. Retrieved December 03, 2020, from <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>