CS 760: Machine Learning Fall 2020

Investigating Causes of Wildfire in the United States

AUTHORS: Jingqi Duan, Huanran Li, Olivia Zhao

Wildfire is a major environmental concern that endangering human life, by affecting both physical and economical environments. Research has investigated the contributing factors that lead to wildfire occurrence. These factors include climate change, past forest fire behaviors, human activities, and weather information. In this project, we obtained data from each of the above modality from official sources, and explored machine learning models including K-nearest neighbor, decision tree, naive bayes and multinomial logistic regression, to predict wildfire's cause. We found that K-nearest neighbor was the best model overall. Our models had good accuracy with the fire caused by lightning, debris burning; but poor results for other causes, such as arson, fireworks, equipment use. Further investigation implied that we might not have enough data for those poorly performed causes.

1 Introduction

Wildfire has been one of the deadliest environmental concerns in the twenty-first century. As of December 22, 2019, 7860 fires have been recorded by California Department of Forestry and Fire Protection, destroying nearly 259,823 acres of land [1]. Nearly 74.1 thousand wildfire outbreaks were reported in Brazil just in 2020 [2]. Many countries including Australia, United States, and south American countries, have been suffered from wildfires. Massive fires put risk on our natural environments, generate greenhouse gas, and put human life and the world economy in jeopardy. The Verisk Wildfire Risk Analysis reported over 2 Million Californian estate properties at high risk from wildfire. Experts believed that the largest fire in CA in 2019 started with lightning, which was a natural phenomenon that is known impossible to foresee and avoid.

Wildfire may seem difficult to forecast due to its various causes. Research has shown that many factors including changeable weather, human activities and climate change could contribute the likelihood of wildfire [3]. Indeed, 90 percent of the wildfire on record were associated with human activities, including campfires, burning debris, power lines, and arson. Climate change, associated with industrial development, could also contribute to the cause. However, what we have known now is limited. Most of the data or evidence was collected on the post-fire scenes, which could be ambiguous. Knowing the exact cause of fire could benefit the weather or fire departments from successfully preventing catastrophes. Further, it could inform policy makers to produce more efficient actions.

In this project, we attempt to predict each alleged cause of wildfires with climate factors (temperature, humidity, precipitation), geographic location, and environmental factor (greenhouse gas emission). In section 3, we introduce the sources of our data, and data processing procedure. In section 4 and 5, we explain and present the results from the machine learning models that were implemented, including K-Nearest-Neighbors, Decision Trees, Logistic Regression and Naive Bayes. Finally, we conclude the best model for predicting the fire cause.

2 Related/Similar Work

Several research has utilized machine learning methods to predict wildfire with various input variables. Cortez and Morais [4] introduced a data mining approach to predict forest fires using meteorological data. Among several machine learning models, the best results derived was by Support Vector Machine (SVM) with 4 meteorological inputs as temperature, relative humidity, rain and wind.

Taylor et al. [5] obtained information from fire department, and implemented various predictive models to understand fire occurrences in regional and global scales. Fire information including fire size, fire ignition, fire growth, frequency and climate data were included in the model. They found that warmer climate years were associated with higher risk of wildfire. Greenhouse gas emission could be one of the potential factors.

Most research have utilized machine learning models to predict wildfire occurrence with meteorological data, categorical fire information, geographical data, historical data or environmental data separately. In our project, we propose to investigate the contributing factors of wildfire with a combination of inputs. We looked for data recording meteorological information, wildfire record and greenhouse gas emission from 1990 to 2017, and explored its significance in predicting each cause of wildfire occurrence in the United States.

3 Dataset

Our data include 28 features and 311,594 observations. Features can be divided into three parts: wildfire records (fire year, fire month, fire day of year, fire size, and geographic coordinates), meteorological data (precipitation, soil moisture, air temperature, specific humidity, and wind speed), and greenhouse gas emission. To observe the temporal effect of meteorological information, we included both data on the discovery day and data prior to the fire (1, 3, and 7 days prior to the fire for wind speed; 7, 15, and 30 days prior to the fire for the other elements). Data from different datasets were combined by longitude and latitude.

Wildfire records are from a database consisting of approximately 1.88 million cases of wildfires occurred in the U.S. from 1992 to 2015 [6]. Due to the limited computing power, we only analyzed the wildfires occurred from 2010 to 2015. Based on previous literature, we included variables that were related to fire cause including the discovery date of fire, geological information, and fire size from the database. As our interest was to predict the fire cause, we omitted the observations which fire cause was recorded as “missing” or “miscellaneous”, resulting in 11 class labels (lightning, equipment use, smoking, campfire, debris burning, railroad, arson, children, fireworks, powerline, structure).

Element	Precipitation	Soil moisture	Air temperature	Specific humidity	Wind speed
Level	At surface	From 0-10 cm	At 2 meters	At 2 meter	At 10 meters
Unit	mm	%	C	kg/kg	m/s
package	*rnoaa*	*RNCEP*	*RNCEP*	*RNCEP*	*RNCEP*

Table 1: Meteorological data: the level, unit, and source package of 5 elements

Meteorological data were obtained using *rnoaa* R package by Chamberlain [7] and *RNCEP* R package by Kemp [8]. Precipitation information includes the daily summary of average rainfall from the National Oceanic and Atmospheric Administration (NOAA) Climate Prediction Center. Weather data were retrieved from the National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) Reanalysis, containing soil moisture, air temperature, specific humidity, and wind speed. Each of these four feature is a daily summary, i.e., the average data of stations within the 100km of wildfire location. Table 1 shows the level and unit of meteorological data.

The greenhouse gas emission data was retrieved from the Greenhouse Gas Inventory in the United States

Environmental Protection Agency [9]. The data contains county- and state-level greenhouse gas (carbon dioxide) emission in million metric tons from 1990 to 2017. The original data was further detailed by release facilities, and cause of release (e.g. electricity, coal). To match with other data, we aggregated the data to county level and state level.

4 Approach

Previous research has proven that meteorological, fire information and greenhouse gas emission contribute to wildfire respectively. However, we did not have any a prior evidence that suggesting whether there is any feature useless for classification. So, the most promising method came in our sight was decision tree. When tree was produced trough the training data, non-relevant features can be ignored if the correct parameters were set up. Since we cannot came up with any interesting modification to the algorithm from our homework, we did not put much effort into this algorithm. It serves as a accuracy benchmark for other algorithms that we developed.

After we had the benchmark produced by decision trees, we moved on to the next method, which is a variant from K-Nearest-Neighbor. Then, multi-class logistic regression and Naive Bayes were implemented respectively. Below we described each of the method.

For the mathematical explanation in following subsections, we define several notations. $\{\mathbf{x}_i, y_i\}$ represents one observed pair, where $\mathbf{x}_i \in \mathbb{R}^D$ is a set of D features and $y_i \in \{1, \ldots, C\} = [C]$ is the class label of the ith sample. $\hat{y}_i$ is the predicted class label of the ith sample. In our project, we have 311,594 observations with 28 features and 11 class labels, i.e., $N = 311,594$, $D = 28$, and $C = 11$.

4.1 Naive Bayes

The main algorithm of Naive Bayes is to assign the label with the highest posterior probability under the independence assumption between features. Mathematically, it can be described as

$$\hat{y} = \arg\max_{y \in [C]} \mathrm{P}(y \mid \mathbf{x}) = \arg\max_{y \in [C]} \mathrm{P}(y) \prod_{j=1}^{D} \mathrm{P}(x_j \mid y).$$

To improve the performance of Naive Bayes, univariate feature selection is adopted. Generally, feature selection is aimed to reduce the number of inputs by selecting the features that are assumed to be most relevant to the outcome. In other wards, the significance of each feature is calculated and only the top significant features are used to predict the outcome. The significance of a feature is computed in two ways: F-value and mutual information. The test statistic of ANOVA under the null hypothesis that two variables are uncorrelated has a F-distribution. F-statistic is always non-negative, the larger the F-value, the less unlikely the test result happened by chance, the higher significance of that feature. One thing to notice is that F-statistic measures the linear relationship between variables. If two variables are dependent but uncorrelated, such as quadratic relationship, F-test result may be insignificant and mutual information should be considered. Mutual information, measuring the relationship between two variables, is defined as

$$I(x, y) = H(x) - H(x \mid y), \quad \text{where } H(x) = \mathrm{E}\Big[\log_2\Big(\frac{1}{\mathrm{P}(x)}\Big)\Big], \quad H(x, y) = \mathrm{E}\Big[\log_2\Big(\frac{1}{\mathrm{P}(x \mid y)}\Big)\Big].$$

Mutual information is always non-negative, the larger the mutual information, the greater the dependency between the variables, and the higher significance of that feature. Two methods for feature selection are performed and then the top significant features are used in Naive Bayes to predict the outcome.

4.2 Weighted-Distance K-Nearest-Neighbor Classification

The K-Nearest-Neighbors (KNN) classification algorithm is a relatively simple technique in machine learning. However, the major problem of this method was the process of finding the k-closest samples from the training data. A good method can increase the amount of correct labels within k samples, which increases the prediction accuracy. So, distance calculation becomes the key point for this algorithm. When x has 28 features, our default solution is:

$$d(x, x') = \sum_{i=1}^{28} ||x_i - x'_i||_2$$

As mentioned in the lecture notes, two of the common limitations of KNN are: 1) model is sensitive to different ranges of the feature; 2) model is not robust against irrelevant features. In order to mitigate those two limitations, we added weighted distance to the features. Since each feature will have different significance towards the label, distance from each feature can have different impact on the classification result. Therefore, we created a group of weights $\{w_1, w_2, ..., w_{28}\}$, based on the below distance formula:

$$d_w(x, x') = \sum_{i=1}^{28} w_i||x_i - x'_i||_2$$

The next step is to determine the weight of each feature. So far, since there is only 28 features, we used brute-force methods to try several values on each w_i and found values with the highest cross-validation accuracy.

4.3 Multinomial Logistic Regression

From what we learned this semester, given a parameter vector θ, we can predict if $\mathbf{x}$ belongs to class 0 or 1 with logistic regression. However, for our dataset, a 2-class logistic regression cannot predict 11 classes at the same time. Here in our project, we define class as the type of fire cause recorded (e.g. arson, lightening, railroad etc.). Therefore, we implemented a multinomial logistic regression as follow:

$$f(x) = \arg_k \max \frac{e^{\theta_k^T \mathbf{x}}}{\sum_{i=1}^{11} e^{\theta_i^T \mathbf{x}}} \qquad c \in [1, 11]$$

The equation above is also known as softmax [10]. For this formula to work as a classifier, we would have $\{\theta_k \in \mathbb{R}^{28} | k \in [1, 11]\}$ and one-hot encoded $Y = \{Y_i \in \{0, 1\}^{11} | Y_i[k] = 1 \text{ if } y_i = k, \text{otherwise } Y_i[k] = 0\}$. Noted that different features may have various ranges and mean value. In order to have a more effective learning process, all features got standardized. To find θ_k, we used gradient descent method as follow:

1. Initialize θ_k to 0s with 10^{-10} variance.
2. $\nabla\theta_k \leftarrow \sum_i \left(Y_i[k] - \frac{e^{\theta_k^T \mathbf{x}_i}}{\sum_{j=1}^{11} e^{\theta_j^T \mathbf{x}_i}} \right) x_i$
3. $\theta_i \leftarrow \theta_i + \eta\nabla\theta_i$
4. Go back to step 2 until it converges.

One advantage of this method is that, for a given x, the output represents the $P(\hat{y} = c|x)$ for all $c \in [1, 11]$. So, if the model does not have a satisfied result, we can always relax the constraints and have this model predicted 2 most possible labels at the same time.

5 Results

Figure 1 shows the distribution of the cause of fire with 11 classes. The most important problem of this dataset was the imbalance of data, which could be a serious harm for prediction accuracy. In another word, each class contains unequal amount of data. So, in addition to predicting the raw data, we also selected 80K entries uniformly from 4 largest classes (lightning(0), equipment use(1), debris burning(4), and arson(6)) and combined it to be a new dataset. The accuracy our model predicted on this new dataset would be labeled as ”4-class” prediction.

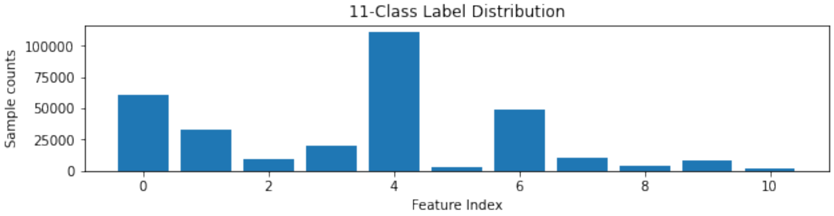


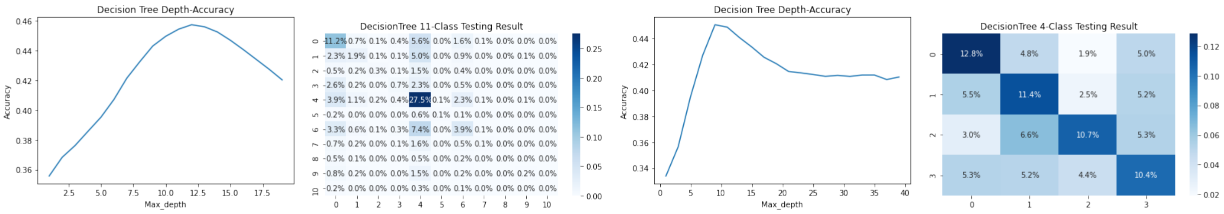
Figure 1: Distribution of the response (fire cause) with 11 class labels

5.1 Decision Tree Classification Benchmark

The model was initialized and trained by *sklearn* built-in method. Parameters including max depth and min split entropy/gini were tested by cross validations to prevent data overfitting. In the end, we found that by setting $MaxDepth = 13$ is the best choice for 11-class decision tree training. The tree was split based on gini scores. Table 2 presents all accuracy results of deicision tree. When predicting 11-class labels, the highest accuracy we got was 46.0%. For the 4-class predictions, the best accuracy from cross-validation was found at 45.1% with $MaxDepth = 9$.

—	Train Accuracy		Test Accuracy	
Model	11-Class	4-Class	11-Class	4-Class
Decision Tree(Best)	52.6%	49.1%	46.0%	45.4%

Table 2: Decision tree classification accuracy



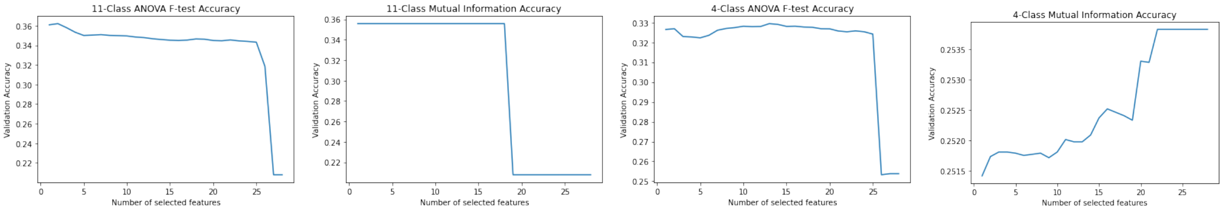
(a) 11-class accuracy vs. depth (b) Confusion matrix of 11-class (c) 4-class accuracy vs. depth (d) Confusion matrix of 4-class

Figure 2: Accuracy results and confusion matrix of decision tree

Figure 2a and 2c displays the relationship of decision tree accuracy and the max depth of the tree. The confusion matrices in Figure 2b and 2d indicate that although both models did not reach a high accuracy, 4-class decision tree showed a pattern of learning effort, which was our expectation from this dataset.

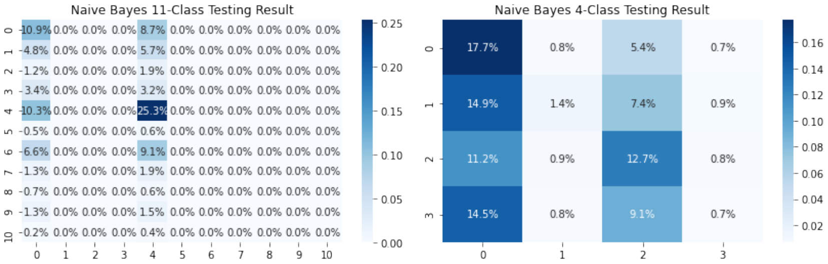
5.2 Naive Bayes

The feature selection and the model were both implemented by *sklearn* built-in method. Features were selected based on their significance, quantified by F-value and mutual information. Some features, such as greenhouse gas emission, were considered to be very significant using mutual information, but insignificant using ANOVA. The reason of this discrepancy is probably that ANOVA tests linear relationship while mutual information accounts dependency of one variable on the other. Figure 3 displays 11-class and 4-class Naive Bayes accuracy of different number of selected features, ranked by F-value from ANOVA and mutual information. While Figure 3a and 3c of feature selection using ANOVA show a clear pattern that excluding insignificant features led to an increase in validation accuracy and reached the maximum accuracy at 2 features and 13 features respectively, Figure 3b and 3d of feature selection using mutual information did not indicate much. Overall, using top significant features to predict the outcome led to an increase in validation accuracy by around 15% for 11-class and around 10% for 4-class. Figure 4 present the confusion matrices of the best results after feature selection using ANOVA F-value.



(a) ANOVA with 11-class (b) Mutual info with 11-class (c) ANOVA with 4-class (d) Mutual info with 4-class

Figure 3: 11-class and 4-class Naive Bayes accuracy of different number of selected features



(a) Confusion matrix of 11 classes (b) Confusion matrix of 4 classes

Figure 4: Confusion matrices of Naive Bayes after feature selection

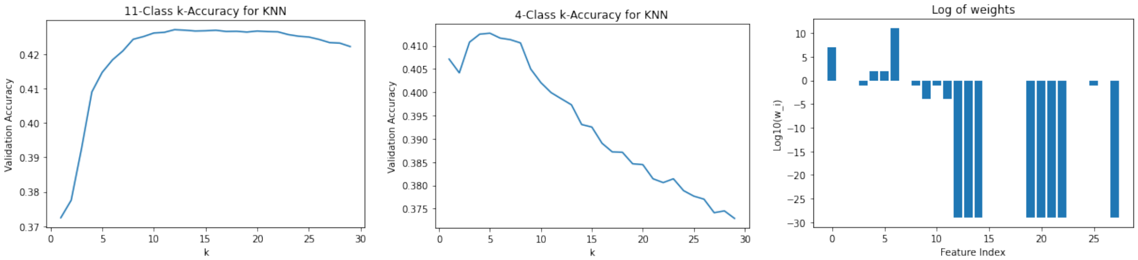
5.3 Weighted-Distance K-Nearest-Neighbor Classification

The model was initialized and trained by *sklearn* built-in method. We did not use the KNN from the homework due to the slow speed of classification. For a given test sample, our brute-force method needed to search through 200K training samples, which means it was impossible to finish the prediction on a 100K testing set.

The first step we took was determining the k for this algorithm. Cross validation was performed with various k values. The validation accuracy and corresponding k values can be referred at Figure 5. The optimum value we picked was $k = 20$ for 11-class and $k = 4$ for 4-class. The next step was finding the optimal weights for every feature with a similar method. Since weights were distributed with dramatic variance, we plotted $log_{10}w_i$ for $i \in [0, 27]$. Noted that this weights was calculated with non-standardized data, this is because *sklearn* KNN solver would take roughly 10 times longer to fit the standardized data than the raw data. The reason behind this should be further investigated.

There were eight features that have weights approximately equal to 10^{-29}, which can be considered as 0. That means those features should have no positive impact on accuracy of KNN. Besides, if we rank the features with their corresponding weight, the order was similar to the significance we calculated from Naive Bayes Method. Overall, latitude and fire year was the major factor for the model to predict the fire cause.

Shown in Table 3, with the parameters correctly set up, the test accuracy was slightly better than Decision Tree's results. With the weighted distance put into calculation, we had 1.2% gain on 11-class test and 3.1% gain on 4-class test from regular KNN algorithm. In fact, $k = 20$ worked better on 4-class as well, although $k = 4$ was calculated from cross-validation. Confusion matrices of two best results are shown in Figure 6.



(a) 11-class accuracy vs. k

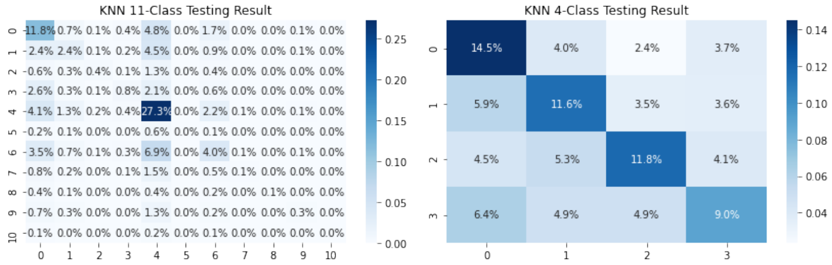
(b) 4-class accuracy vs. k

(c) Log weights of features

Figure 5: Accuracy results and weights of KNN

—	Train Accuracy		Test Accuracy	
Model	11-Class	4-Class	11-Class	4-Class
Decision Tree(Best)	52.6%	49.1%	46.0%	45.4%
KNN (k = 4)	60.1%	62.8%	41.5%	42.1%
KNN (k = 20)	48.0%	46.6%	43.3%	39.7%
Weighted-Distance KNN (k = 4)	62.1%	65.4%	44.4%	46.9%
Weighted-Distance KNN (k = 20)	51.5%	53.8%	47.2%	48.5%

Table 3: KNN and weighted-distance KNN accuracy compared to decision tree



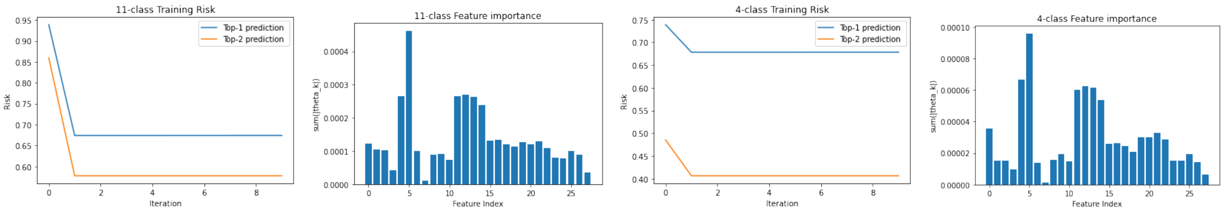
(a) Confusion matrix of 11 classes

(b) Confusion matrix of 4 classes

Figure 6: Confusion matrices of weighted-distance KNN

5.4 Multinomial Logistic Regression

This method was written only with the support from numpy. A regular gradient decent method was implemented. With numerous tests, learning rate was fixed at 10^{-9} and max iteration was limited to 10. The graphs below showed how fast the model converged and also the "importance" of each feature. Each feature's importance was calculated by $I_j = \sum_{k=1}^{C} |\theta_k|$, where j represents the feature index and C represents the total number of classes, ie. $C = 11$ for the left one and $C = 4$ for the right one. Unsurprisingly, we saw a similar pattern between 11-class's and 4-class's pattern. Besides, if we compare the importance with distance weights we got from KNN, they also shared similar patterns. See detail on Figure 7.

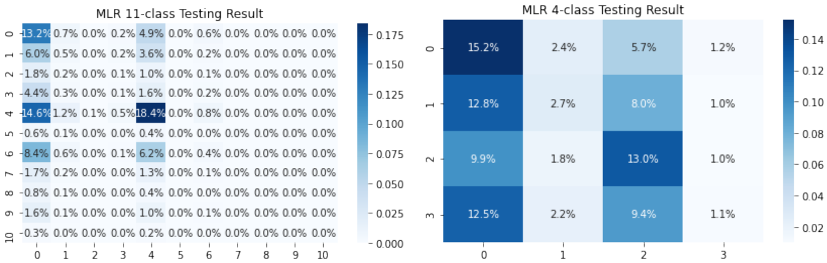


(a) 11-class training risk (b) 11-class feature importance (c) 4-class training risk (d) 4-class feature importance

Figure 7: Training risk and feature importance in multinomial logistics regression

—	Train Accuracy		Test Accuracy	
Model	11-Class	4-Class	11-Class	4-Class
Decision Tree(Best)	52.6%	49.1%	46.0%	45.4%
Weighted-Distance KNN (Best)	51.5%	53.8%	47.2%	48.5%
Top-1 Multinomial Logistic Regression	32.6%	32.2%	32.6%	32.0%
Top-2 Multinomial Logistic Regression	42.2%	59.4%	42.3%	59.2%

Table 4: Multinomial logistical regression accuracy compared to decision tree and weighted-distance KNN



(a) 11-class Confusion Matrix (b) 4-class Confusion Matrix

Figure 8: Logistic Regression Confusion Matrix

Table 4 shows the accuracy we had compared to Decision Tree and KNN. ”Top-1” represents the results when the model predicted the 1 label with the most probability, and ”Top-2” represents when the model predicted 2 labels for 1 entry with top 2 most probability. So, within expectation, Top-2 would have a better performance than Top-1, but predicting 2 labels at the same time would not be as meaningful as TOP-1. Moreover, the confusion matrices of the best results were plotted in Figure 8.

6 Conclusions and Future Work

To summarize, in this project, we attempted to look for the best model to predict each cause of wildfire from previous data. We combined data from historical fire record, meteorological record and environmental information. Weighted K-nearest neighbour, Naive Bayes, and Multinomial logistic regression were implemented. Weighted KNN was proved to be the best models for fire cause prediction with above 47% accuracy. If predicting two most possible cause were accepted, Multinomial Logistic Regression had a surprising 59.2% for 4-Class prediction.

Due to limited computing power and limited data, we were not able to produce a model that can predict all classes with high accuracies. Future work should further investigate with more data. In addition, we could also consider other methods like Maximum Expectation, Support Vector Machine or Neural Networks.

References

[1] Michelle Fay Shteyn. *Regional Extreme Weather Concern and its Relation to Support for Environmental Action*. PhD thesis, UC Santa Barbara, 2020.

[2] Karen Michael, Diane Davies, David S Green, Tian Yao, and Ryan Boller. Nasa's land, atmosphere near real-time capability for eos (lance): Delivering data and imagery to meet the needs of near real-time applications. 2020.

[3] Raffaella Lovreglio, Vittorio Leone, P Giaquinto, and Alessandra Notarnicola. Wildfire cause analysis: four case-studies in southern italy. *iForest-Biogeosciences and Forestry*, 3(1):8, 2010.

[4] Paulo Cortez and Aníbal de Jesus Raimundo Morais. A data mining approach to predict forest fires using meteorological data. 2007.

[5] Steve W Taylor, Douglas G Woolford, CB Dean, and David L Martell. Wildfire prediction to inform management: statistical science challenges. *Statistical Science*, pages 586–615, 2013.

[6] Karen C. Short. Spatial wildfire occurrence data for the united states, 1992-2015 [fpafod20170508]. *Fort Collins, CO: Forest Service Research Data Archive*, 2017.

[7] Scott Chamberlain. rnoaa: 'noaa' weather data from r. 2020. R package version 1.2.0.

[8] Michael U. Kemp, E. Emiel van Loon, Judy Shamoun-Baranes, and Willem Bouten. Rncep: global weather and climate data at your fingertips. methods in ecology and evolution. 3, 2011.

[9] Inventory of u.s. greenhouse gas emissions and sinks, Sep 2020.

[10] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.