# Fire Prediction Modelling

Team Members - Jordi Bardia, Samuel Liechty, Daksh Mehta

**Description:**

Forest fires are a major issue that lead to huge losses every year in the US. Our project aims to determine the factors behind these forest fires and what causes them. We have used logistic regression to find the factors which are associated with these fires and identify the most influential ones which causes them.

**Datasets:**

We have combined three different datasets to feed to our model. These contained features such as latitude, longitude, name of the city (USA) They were obtained from different sources and some data pre processing steps were performed so that they could be used with the logistic regression model.

**Data Pre-Processing:**

The fire dataset contained the co-ordinates of fires which were used to calculate the nearest city distance. The nearest city co-ordinates were found out using the K-d tree algorithm. After that the distance between the occurrence of the fire and the closest city was found out using the Haversine Formula.

The dataset contained some co-ordinates outside the US as well. Our regression model is limited to co-ordinates inside the US. Thus, these co-ordinates were filtered.

The intersection of locations where fire occurred with the precipitation, temperature and elevation data was performed to combine the data. Raster data is data saved in pixels or grids. So, these were converted to co-ordinates as well.

After combining these datasets there were some entries where some of the values were missing. These were omitted from the main dataframe and records having fire power below 15W were also omitted.

We also created a correlation matrix to find out which predictors are correlated with each other to perform feature elimination so that our model performs better.

The data pre-processing steps were performed using Python and R.

**Hypothesis:**

Our hypothesis is that – “ Fire risk is associated with temperature, precipitation, elevation, distance to closest city, city population, and city density.”   
To confirm our hypothesis, we performed linear regression on the data using the predictors mentioned above and found out the p-values which could confirm or deny our hypothesis.

Our experiment plans to establish association and not causality because we are not performing a randomized control experiment.

**Model Used:**

We have used a logistic regression model to establish association between our predictors and the dependent variable. It predicts the probability of a fire occurring given the inputs. The model was fit using R programming language.

**Results:**

After fitting the logistic regression model with our dataset, we obtained the p values for the corresponding predictors. All the p-values were found out to be statistically significant as they were less than 0.05.

We also tested the model by splitting the dataset into train and test. The classification accuracy obtained was 64.8%. If our model predicted the probability of a fire occurring >0.5 it was classified as ‘fire will occur’ else ‘fire will not occur’.