

Which predictive models most accurately forecast forest fires along trail routes in  
Table Mountains, South Africa.

Brian Goodluck Mmari

30th April, 2024

Class Instructor: Dr. Jonathan Hanke

I hereby declare that this paper represents my own work in accordance with University  
regulations.

Your Signature

Brian Goodluck Mmari

Which predictive models most accurately forecast forest fires along trail routes in  
Table Mountains, South Africa.

Brian Goodluck Mmari

### **Abstract**

Forest fires have recently gained traction over what seems to be a result of global warming, as the earth's surface gets hotter and hotter. Arguably, multiple studies have shown that climate change has led to an increase in forest fires frequency, length, and extent. According to Western Fire Chiefs Association, wildfires across the globe are most frequent in the United States, Canada, Australia, Western Cape of South Africa, and Southern Europe. According to a research paper most frequent forest fires in South Africa have occurred along the mountain regions of the Western province, aiding the necessity for my research[4]. In this paper, I use 4 trailing routes along the Table Mountains: MCSA Hut via Kasteelspoort Trail, Cape Point Overnight Hike, Hoerikwaggo trail: Day 4, and Hoerikwaggo trail: Day 2 to sample my data. Using these trail routes, I aim to analyze different predictive models to better predict forest fires and analyze their performance overall.

## **ACKNOWLEDGEMENTS**

I want to thank Dr. Jonathan Hanke for first helping me throughout this course. His input has been invaluable, which has really helped make this process possible. I am thankful for his time and effort to make this class more interactive and fun, and also being very supportive of assignment deadlines and overall students' workload. Furthermore, I have had the privilege of learning a lot through this class which would definitely have been hard without his presence. Thank you for providing feedback on this report and the overall project.

I also want to thank our preceptor Visheshs for his unwavering teaching and exceptional tutoring during precept times and office hours. Under his guidance, I have developed formidable problem-solving skills, techniques and work ethic. Thank you for everything!

## Table of Contents

1.	Background	5 - 7
	1.1 History of Wildfires in South Africa	5 - 6
	1.2 Previous Research on forest fires	6 - 7
2.	About Dataset	7 - 12
	2.1 Sources of Datasets	7 - 10
	2.2 Aggregation of Datasets	10 - 12
3.	Exploratory Data Analysis	12 - 15
	3.1 Explore & Analyze Feature column	12 - 15
	3.2 Explore & Analyze Target column	15
4.	Choose and Train Models with Hyperparameter tuning	16 - 19
	4.1 Logistic Regression	16 - 17
	4.2 Random Forest Classifiers	17 - 18
	4.3 Neural Networks	18 - 19
5.	Gain Insights and Conclusions	19 - 20
6.	References	20 - 21

## 1. Background

### 1.1 History of Wildfires in South Africa

Wildfires have had tremendous consequences in many parts of the world. Just recently, 2023, a total of 324,917 acres of land in California were engulfed with fires. According to the California Department of Forestry and Fire Protection, extended drought and increased temperature is one of the leading causes of forest fires around the area[2]. However, the United States is not the only affected country. According to the International Fire & Safety Journal, wildfires happen most frequently in the United States, Canada, Australia, the Western Cape of South Africa, and Southern Europe[1]. Particularly in South Africa, most fires actually happen in the mountainous regions along the Western provinces. The Table Mountains in South Africa, which is my area of interest, has experienced forest fires in 2000, 2006, 2009 and 2021. This frequency is quite astonishing and one that prompted me to do this research.

Using Machine Learning(ML) to understand forest fires is particularly important because having the best predictive model means we can safely avoid future forest fires. This effectiveness in different ML models thereby helps humans better predict and safely avoid the repercussion sooner rather than later. Predicting forest fires will also help in preserving biodiversity as well as recreational and aesthetic values. Because of the aforementioned reasons I was prompted to answer the question, If most of the fires happen in the mountainous regions of the Western province, how can we better predict them?

### 1.2 Previous Research on forest fires

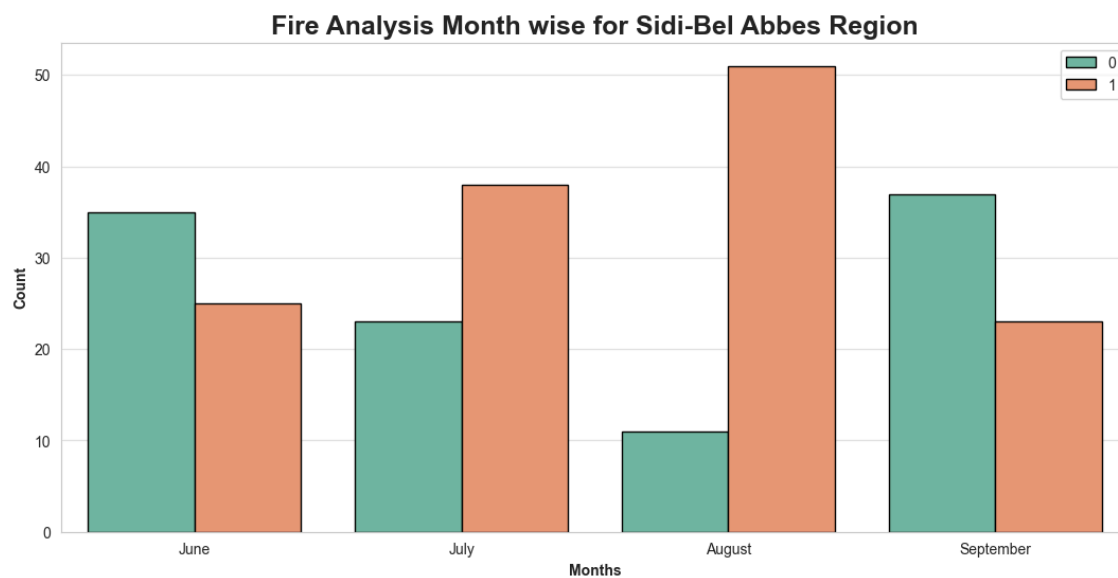
There has been various researches done on this topic, ones that are even outside South Africa.

One research project was conducted in **Montesinho Park, Portugal** for the year 2007 with the dataset obtained from Kaggle. This project relied on using only Neural Networks(NN's) with linear kernels, which are undoubtedly very powerful, to predict forest fires. The diagram below shows exactly how powerful this technique was.

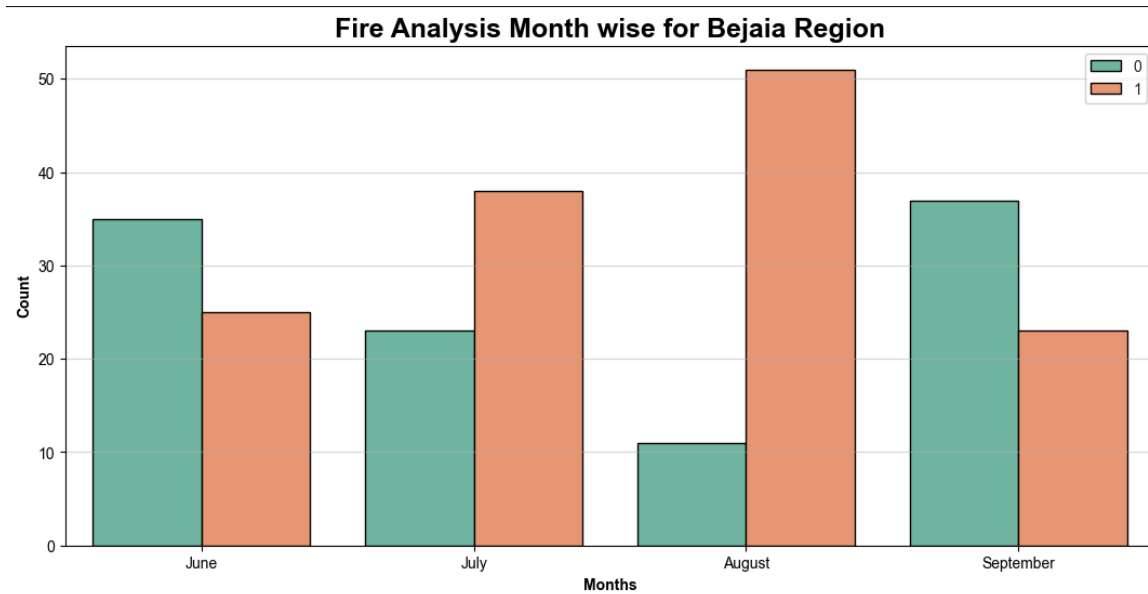


**Figure 1:** Increase in model accuracy over the number of epochs

Another research project was conducted in **Bejaia region** located in the **northeast of Algeria** as well as the **Sidi Bel-abbes region** located in the **northwest of Algeria** for the year 2012. Comparatively, this research project used a dataset collected from the UCI repository. In addition, the researcher wanted to compare performance accuracies of different models (“Logistic Regression”, “Decision Trees”, “Random Forest”, and “K-Nearest Neighbors”) in predicting forest fires. The researcher also performed regression analyses using “Linear regression”, “Lasso Regression”, “Ridge Regression” etc.



**Figure 2:** A visualization of the model prediction in Sidi-Bel Abbas Region



**Figure 3:** A visualization of the model prediction in Bejala Region

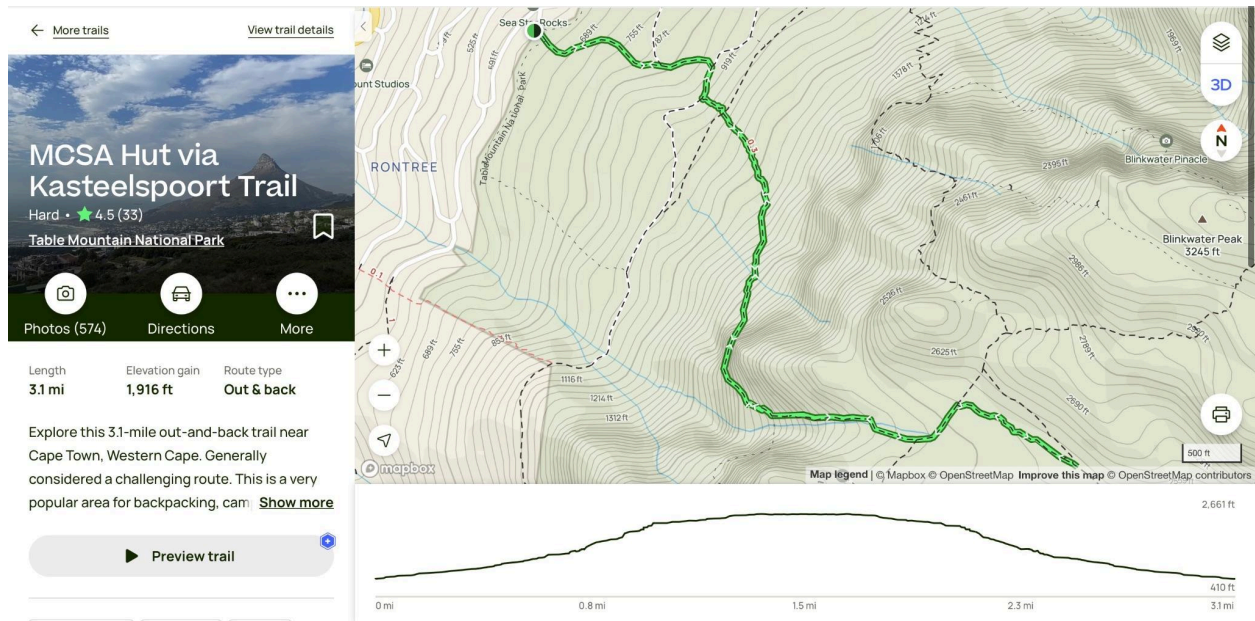
Compared to the two mentioned researches, my research will pivot in two major ways. First and foremost my research is intended to be on a more niche area, trail routes. The intended trail routes are particularly ones where most visitors/hikers tend to camp a lot on the Table Mountains. The chosen trail routes are: Cape Point Overnight Hike, MCSA Hut via Kasteelspoort Trail, Hoerikwaggo Trail: Day 4, and Hoerikwaggo Trail: Day 2.

In addition my research will also be conducted for an extended period of time, 2012 - 2021 contrary to the previous research. This should provide my model with a much better predictive power. I also intend to deploy different models to see and test my predictive accuracy.

## 2 About Dataset

### 2.1 Sources of Datasets

#### 1) DATASET 1: Trail data



**Figure 4:** An example of visualized dataset of the MCSA Hut via Kastelspoort Trail

#### Data Available at:

<https://www.alltrails.com/explore/trail/south-africa/western-cape/mcsa-hut-via-kasteelspoort-trail?mobileMap=false&ref=sidebar-static-map&u=i>

#### Data Description

The first dataset as discussed previously was obtained from the popular website: **AllTrails**, where all the four trails were downloaded as csv files with information about their latitudes, longitudes and elevation. Also, as previously mentioned because latitudes and longitudes are unique identifiers of any geographical location especially for this context, elevation data was never used in this analysis.

#### 2) DATASET 2: Climate dataset

Data Available at: <https://www.ncei.noaa.gov/cdo-web/search>

#### Data Description

The second dataset was obtained from National Centers for Environmental Information, National Oceanic and Atmospheric Administration (NOAA), where the data was daily weather condition summaries from the year, 2012 to 2021 for the whole country, South Africa collected in steps due to the large file size.

Here are the descriptions of the variables in the dataset

- **STATION:** station identification code
- **NAME:** name of the station (usually city/airport name)



- **LATITUDE:** latitude (decimated degrees w/northern hemisphere values > 0, southern hemisphere values < 0)
- **LONGITUDE:** decimated degrees w/western hemisphere values < 0, eastern hemisphere values > 0)
- **ELEVATION:** above mean sea level (tenths of meters)
- **DATE:** is the year of the record (4 digits) followed by month (2 digits) and day (2 digits).
- **PRCP:** Precipitation (mm or inches as per user preference, inches to hundredths on Daily Form pdf file
- **TAVG:** Average temperature (Fahrenheit or Celsius as per user preference, Fahrenheit to tenths on Daily Form pdf file
- **TMAX:** Maximum temperature (Fahrenheit or Celsius as per user preference, Fahrenheit to tenths on Daily Form pdf file
- **TMIN:** Minimum temperature (Fahrenheit or Celsius as per user preference, Fahrenheit to tenths on Daily Form pdf file

### 3) **DATASET 3:** Fire dataset

**Data Available at:** <https://firms.modaps.eosdis.nasa.gov/country/>

#### **Data Description**

The third dataset was obtained from NASA Fire Information for Resource Management System(FIRMS) which provides near real-time active fire data from satellite observations. The satellites used are called Visible Infrared Imaging Radiometer Suite(VIIRS) due to a satellite project called Suomi National Polar-orbiting Partnership (Suomi NPP) as a joint mission between NASA and NOAA. The target column of my aggregate dataset is extracted from the confidence factor as a categorical variable.

Here are the descriptions of the variables in the dataset

- **LATITUDE:** The latitude coordinate where the observation was made, indicating the north-south position.
- **LONGITUDE:** The longitude coordinate where the observation was made, indicating the east-west position
- **BRIGHT TI4:** Brightness temperature of the fire in the I4 band, measured in Kelvin. This is a measure of the intensity of radiation emitted by the fire in the infrared spectrum.
- **SCAN:** The size of the pixel in kilometers from which the observation was made, representing the east-west dimension of the ground area covered.

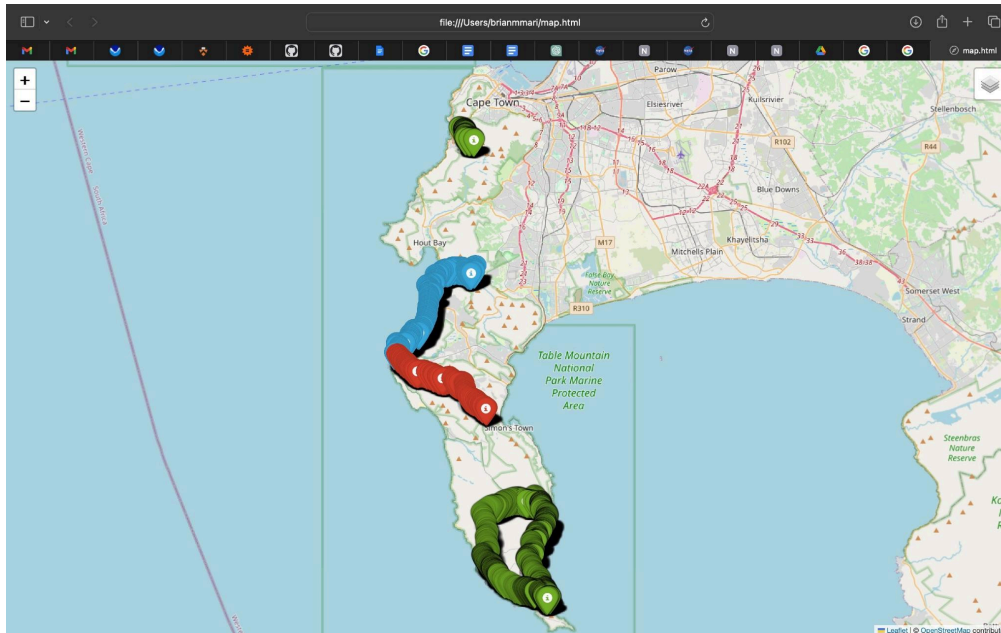
- **TRACK:** Similar to scan, but represents the north-south dimension of the ground area covered.
- **ACQ\_DATE:** The acquisition date of the observation.
- **ACQ\_TIME:** The acquisition time of the observation.
- **SATELLITE:** The name of the satellite that provided the data, "N" likely stands for Suomi National Polar-orbiting Partnership (Suomi NPP).
- **INSTRUMENT:** The instrument used for the observation, which in this case is VIIRS (Visible Infrared Imaging Radiometer Suite).=
- **CONFIDENCE:** A qualitative measure of the confidence in the fire detection: 'n' stand for 'nominal', 'h' for 'high', and 'l' for 'low'.
- **VERSION:** The version of the algorithm or processing version used to generate the data.
- **BRIGHT\_T15:** Brightness temperature of the fire in the I5 band, also measured in Kelvin. This can be used to provide a more refined measure of the fire's characteristics.
- **FRP:** Fire Radiative Power, measured in Megawatts (MW), which quantifies the heat energy released by the fire.
- **DAYNIGHT:** Indicator of whether the observation was made during the day ('D') or night ('N').
- **TYPE:** Designates the type of fire or thermal anomaly detected. The meaning of each value (e.g., '0', '2') depends on specific definitions used in the data collection, often indicating whether the detected heat source is a confirmed fire, a hot spot, or another type of thermal anomaly

## 2.2 Aggregation of Datasets

A common attribute to all three types of the previously mentioned datasets is their geographical location. Because of this, aggregation of the three datasets involved reading geographical data((latitude and longitude) from these files and creating a geometric polygon for each trail. This technique was implemented using the **Geospatial clustering algorithm**. This algorithm and its implementations is described in the steps below

### Step 1)

This polygon represents the shape of the trail as a continuous line formed by connecting points from start to finish and then back to the start to close the loop.



**Figure 5:** A visual representation of the polygons created for each trail when run from Visual Studio Code and displayed as an html file

After creating individual geometric representations for each trail, my algorithm proceeds to combine them into a single geographic data frame(a type of table that understands geography).

I then proceed to use the following important methods:

- **Spatial join:** It uses a spatial join to combine this trail data with another dataset called `climate_dataset`. This operation links data based on their spatial location, essentially matching climate data points with the trails they are closest to or overlap.
- **Calculating nearest trails:** For each point in the climate data, the script calculates which trail polygon is the closest and updates the climate data with the name of the nearest trail.

## Step 2)

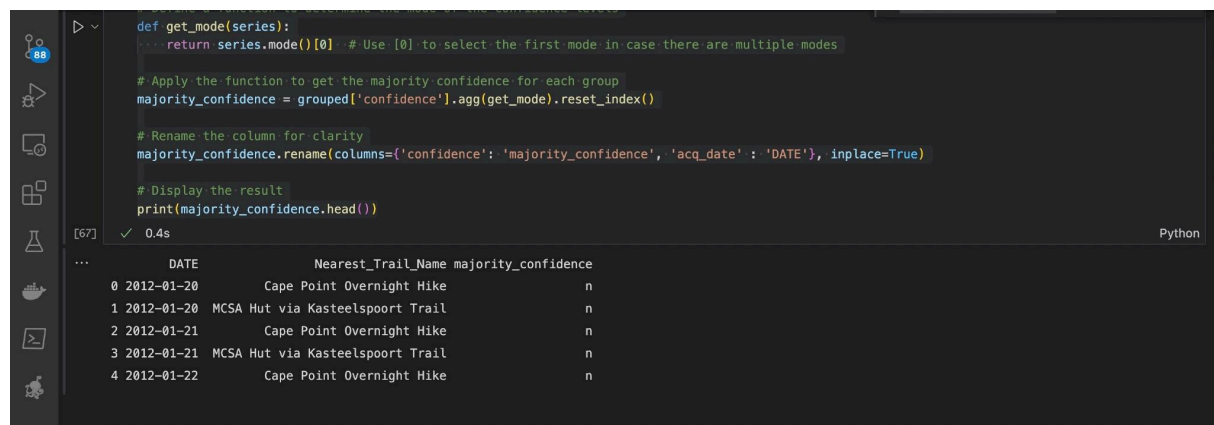
The same is done for the fire dataset, where for each fire incident point, my algorithm calculates how far it is from the nearest trail, which helps in understanding how close the fires are occurring relative to the trails.

### Step 3)

The final procedure is combining these two newly created datasets into one. Here I group the `fire_data` data frame by both `'acq_date'` and `'Nearest_Trail_Name'`. I

I also proceed to use a function called `get_mode()` to calculate the most frequently occurring value of the `'confidence'` levels within each group. This is useful for understanding the most common confidence assessment for fire detections on specific dates at specific trails.

Because the `majority_confidence` dataset has confidence factors for different trail names for a single day I choose to aggregate this dataset with the `"climate_dataset"` using the `"Nearest_Trail_Name"` column, which is common in both datasets.



```
def get_mode(series):  
    return series.mode()[0] # Use [0] to select the first mode in case there are multiple modes  
  
# Apply the function to get the majority confidence for each group  
majority_confidence = grouped['confidence'].agg(get_mode).reset_index()  
  
# Rename the column for clarity  
majority_confidence.rename(columns={'confidence': 'majority_confidence', 'acq_date': 'DATE'}, inplace=True)  
  
# Display the result  
print(majority_confidence.head())
```

[67] ✓ 0.4s Python

	DATE	Nearest_Trail_Name	majority_confidence
0	2012-01-20	Cape Point Overnight Hike	n
1	2012-01-20	MCSA Hut via Kasteelspoort Trail	n
2	2012-01-21	Cape Point Overnight Hike	n
3	2012-01-21	MCSA Hut via Kasteelspoort Trail	n
4	2012-01-22	Cape Point Overnight Hike	n

**Figure 6.** A visual representation of the `majority_confidence` dataframe

I then proceed to use the `"outer"` join method which ensures that all entries from both datasets are included in the merged dataframe, filling in missing values with `Nan` where data from one dataset has no corresponding entry in the other.

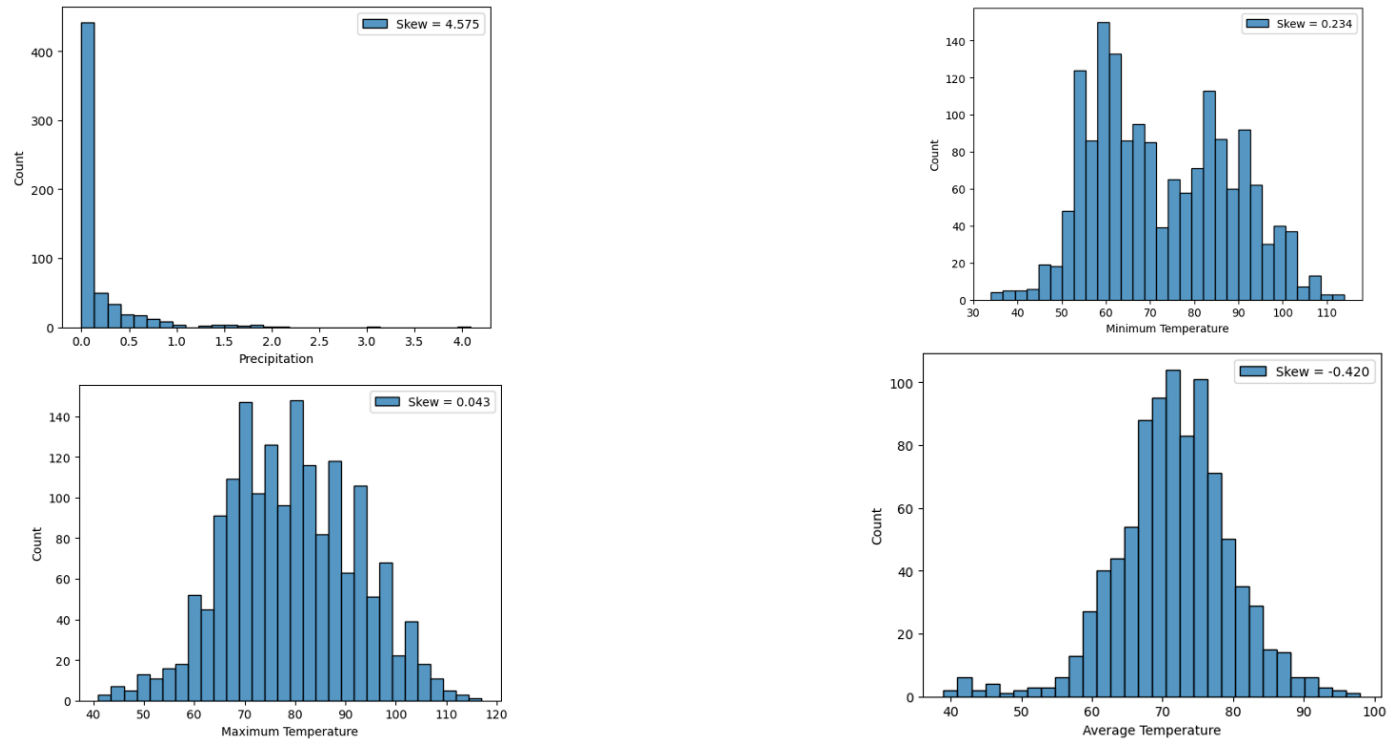
## 2 Exploratory Data Analysis

It was important to perform Exploratory Data Analysis(EDA) before choosing and training a particular model. I performed EDA analysis on all of my columns to understand what kind of data i was dealing with, and what results to expect when I modeled the data in the future.

### 3.1 Explore & Analyze Feature column

I pursued this procedure by asking myself the following questions:

1. What feature columns do I have?
  - 'Latitude', 'Longitude', 'ELEVATION', 'Date', 'Precipitation', 'Average Temperature', 'Maximum Temperature', 'Minimum Temperature', 'Nearest\_Trail\_Name\_x', 'Nearest\_Trail\_Name\_y'
  - Choices made:
    - Pick features relevant to my prediction:
    - 'Date', 'Precipitation', 'Average Temperature', 'Maximum Temperature', 'Minimum Temperature', 'Nearest\_Trail\_Name\_y', 'majority\_confidence'
2. Do I have any missing entries?
  - Percentage of missing values for Precipitation: 65.18%
  - Percentage of missing values for Average Temperature: 47.37%
  - Percentage of missing values for Maximum Temperature: 2.14%
  - Percentage of missing values for Minimum Temperature: 4.92%
  - Choices made:
    - Filter out Precipitation data
3. What is my distribution of values in my features column?



**Figure 7.** Histograms indicating distribution of data in all my relevant features columns.

From the histograms made, I deduced the following:

1. Precipitation is right-skewed indicating that most of the data is near the lower end of the range, and higher values are infrequent. Also we have very few bars owing to missing entries
2. Average Temperature is left-skewed indicating that most of the data is near the higher end of the range. This means that the median of this data is higher than its mean.
3. Maximum Temperature is close to benign symmetrical indicating that most of the data is evenly split between higher and lower ranges
4. Minimum Temperature is right-skewed which is also almost like a non-symmetric bimodal distribution. Thai means that the mean of this data is higher than its median.

It's also important to note that none of the histograms show any signs of outliers for any particular feature columns. This is a good sanity check before modeling our data.

### 3.1 Explore & Analyze Target column

I pursued this procedure by asking myself the following questions:

1. What type of variable is my target column?
  - Categorical
    - “l” = low
    - “n” = nominal
    - “h” = high
  - Choices made:
    - Convert categories to numbers: ‘0’, ‘1’, ‘2’ respectively.
2. What is the distribution of my labels?

majority confidence	
0.0 = ‘low’	1312
1.0 = ‘nominal	191793
2.0 = ‘high’	417

Highly imbalanced classifications. More 1’s than 0’s and 2’s

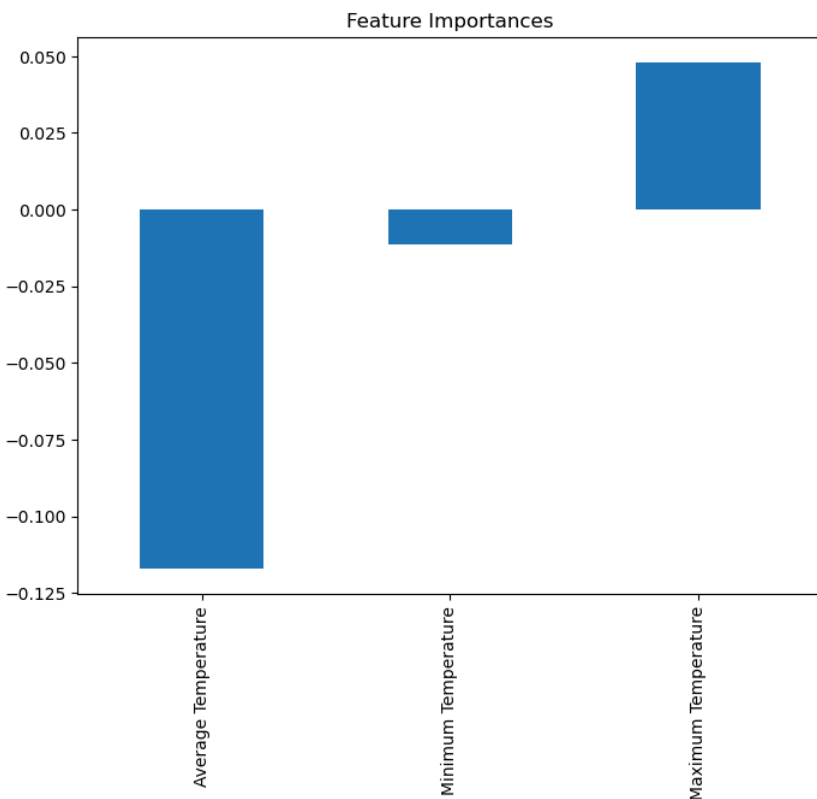
- Choices made:
  - Filter the dataset to only include 0’s and 2’s
    - Pros:
      - More balanced dataset
      - Using high confidence factors for my model
    - Cons
      - Smaller dataset.
      - (267441, 11)-> (1729, 11)

#### 4. Choose and Train Models Hyperparameter tuning

After performing a `train_test_split`, which was considered on the basis of the size of my dataset, I went to use a couple of models for training. Because the target column is not continuous, my modeling technique was classification.

#### 4.1 Logistic Regression

Logistic Regression was first chosen because of its ability to capture linear relationships among feature columns. This was particularly important because aspects such as average temperature, maximum and minimum temperature are all linearly related. From my logistic regression model I then plotted a coefficient plot see the underlying importance of each feature with the predictions.

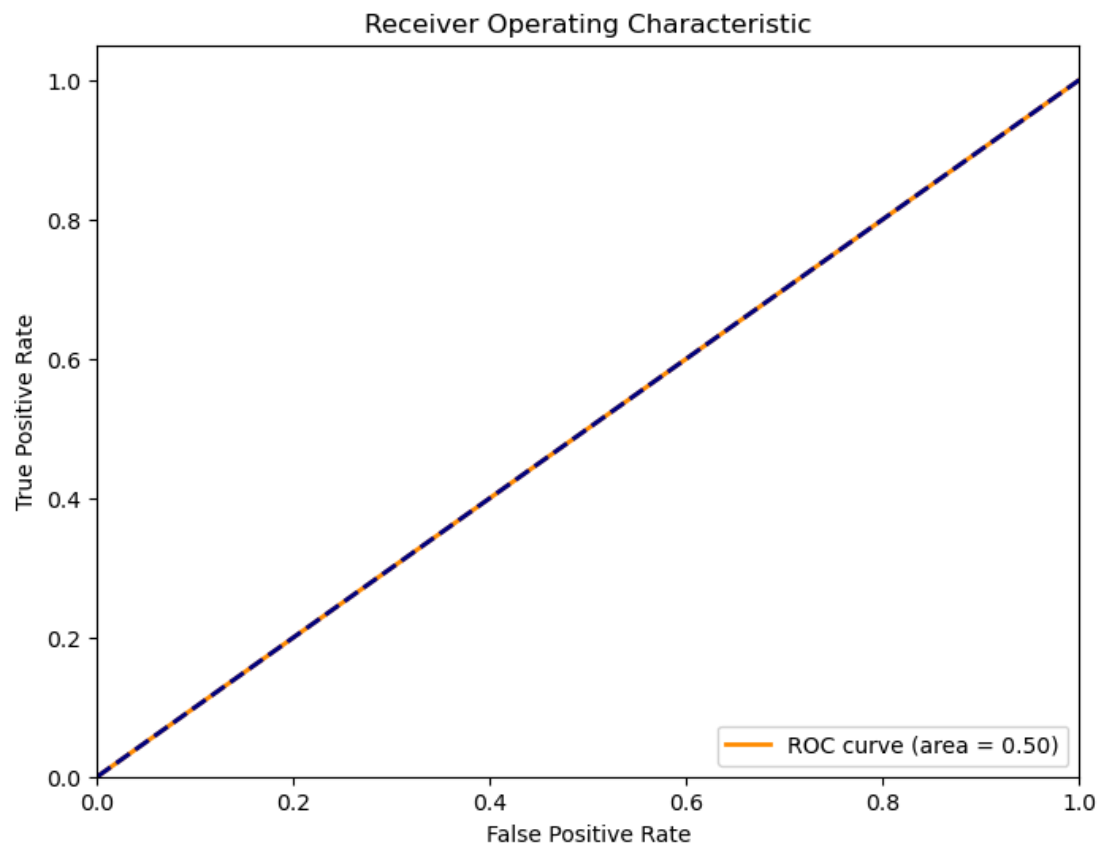


**Figure 8.** Coefficient plot showing importance of each feature on prediction

Key insights:



It's clear that the order of importance of the features goes from Average to Maximum to Minimum Temperature.



**Figure 9.** ROC Curve for logistic regression model

The ROC curve has an area of 0.5 indicating that the model is no better than a random guesser.

As a result this means that we can do better

#### **4.2 Random Forest Classifiers**

The switch to random forest classifiers was prompted by the following reasons:

- 1) Can handle missing NaN values
  - This means that I can use 'Average Temperature' column and hence capture more understanding of my data
- 2) Can capture non-linearity between features

- This is extremely important especially when we don't know how our features are related to each other

3) Can handle both categorical and continuous data hence effective for both regression and classification tasks

This therefore means I get the flexibility of including all my features: Precipitation, Average Temperature, Maximum Temperature and Minimum Temperature.

After training my model with Random Forests, predictably my accuracy increases compared to Logistic Regression. My accuracy remains almost the same as in Logistic regression, even with an increasing number of estimators(trees). However after conducting hyperparameter tuning, my model's accuracy increases to 78%. This compared to the original model which used 100,000 trees suggesting that perhaps too much complexity can lead to overfitting.

```

# Setup the grid search
grid_search = GridSearchCV(RandomForestClassifier(random_state=42), param_grid, cv=5, scoring='accuracy')
grid_search.fit(X_train, y_train)

# Best parameters and best score
print("Best parameters:", grid_search.best_params_)
print("Best cross-validation score: {:.2f}".format(grid_search.best_score_))

[593] ✓ 2m 58.8s Python

... Best parameters: {'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 100}
Best cross-validation score: 0.76

# Re-train with the best parameters
best_params_rf = RandomForestClassifier(**grid_search.best_params_, random_state=42)
best_params_rf.fit(X_train, y_train)

# Re-evaluate
new_y_pred = best_params_rf.predict(X_test)
new_accuracy = accuracy_score(y_test, new_y_pred)
new_roc_auc = roc_auc_score(y_test, best_params_rf.predict_proba(X_test)[:, 1])

print(f"New Accuracy: {new_accuracy}")
print(f"New ROC AUC: {new_roc_auc}")

[595] ✓ 0.2s Python

... New Accuracy: 0.7832369942196532
New ROC AUC: 0.5916737198476514

```

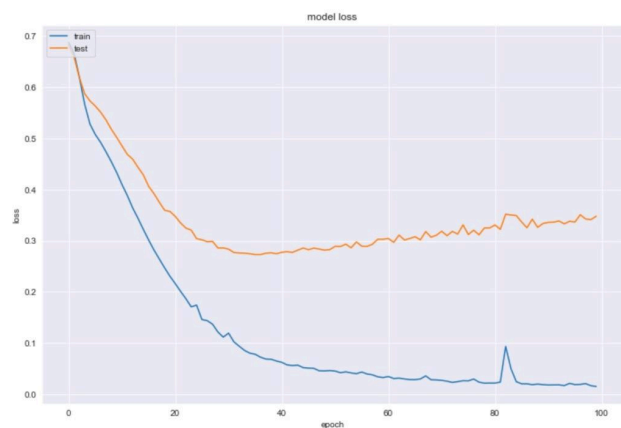
**Figure 10.** Increased accuracy from hyperparameter tuning using random forests

## 4.3 Neural Networks

After some analysis, neural networks offer the best of all worlds. Not only can neural networks understand the complexity involved, can perform regression, classification and even unsupervised learning but also neural networks can be tuned for time-series data, like in our case.

The only caveat to using neural networks is its really high processing time.

By using neural networks, I was able to jump the accuracy score to about 83% with a few adjustments.



**Figure 11.** Decreasing in model loss curve as the number of epochs increases

## 5. Gain Insights and Conclusions

From the different models deployed, there's a lot of key insights that were discovered. The saying that, "the simplest model is always the best"(Occam's razor) was my standard for starting the model evaluation process

However, as a data scientist it's important to note that sometimes the simplest model doesn't really do much. This can be argued in my case because the better question is, "How good is my data?"

Another key lesson learnt from this is although hyperparameter tuning is always better, larger parameter size is not always better. When using Random Forests I had initially set the number

of estimators = 10000 thinking that the larger number of trees would help capture complexity of my data. However after doing GridSearchCV where I found the best model by adjusting different parameters. This resulted in a model that gave me a higher accuracy but used less number of trees.

Therefore this research not only taught me the importance of EDA before doing modelling but also the advantages and disadvantages of every model. Choosing and training a model even though it depends on the data scientist, ultimately the decisions made have to be reasonably justifiable.

## **6. References**

[1] USGCRP (U.S. Global Change Research Program). 2018. Impacts, risks, and adaptation in the United States: Fourth National Climate Assessment, volume II. Reidmiller, D.R., C.W. Avery, D.R. Easterling, K.E. Kunkel, K.L.M. Lewis, T.K. Maycock, and B.C. Stewart (eds.).

<https://nca2018.globalchange.gov/downloads>.

[2] Westerling, A.L. 2016. Increasing western U.S. forest wildfire activity: Sensitivity to changes in the timing of spring. Phil. Trans. R. Soc. B. 371:20150178.

<https://royalsocietypublishing.org/doi/10.1098/rstb.2015.0178>

[3] Defending People, Wildland, and California's Way of Life.

[Our Impact | CAL FIRE California Department of Forestry and Fire Protection \(.gov\)https://www.fire.ca.gov › our-impact](https://www.fire.ca.gov/our-impact)

[4] Strydom Sheldon, Savage J. Michael. A spatio-temporal analysis of fires in South Africa. Nov./Dec. 2016

[https://scielo.org.za/scielo.php?script=sci\\_arttext&pid=S0038-23532016000600019#:~:text=The%20study%20included%20the%20mapping,and%20in%20the%20Western%20Cape](https://scielo.org.za/scielo.php?script=sci_arttext&pid=S0038-23532016000600019#:~:text=The%20study%20included%20the%20mapping,and%20in%20the%20Western%20Cape).

[5] Burge Simon. What Causes Wildfires - 11 Common Reasons. August 7th, 2023.

[What Causes Wildfires - 11 Common ReasonsInternational Fire & Safety  
Journalhttps://internationalfireandsafetyjournal.com › what-cau...](https://internationalfireandsafetyjournal.com/what-causes-wildfires-11-common-reasons/)