SFU SIMON FRASER UNIVERSITY

CMPT 353

Computational Data Science

Group Project

Bhakti Bhanushali
301448639

Ketan Dhingra
301439331

Harveer Singh Virk
301451685

Introduction

Forest fires are an issue, in areas with dense forests like Canada, where they pose serious threats to nature, communities, and economies. These natural disasters can devastate plant and animal life, cause property damage, and create health hazards due to smoke and air pollution. Over the years the frequency and severity of forest fires have risen, influenced by factors such as climate change, human actions, and natural cycles. This highlights the need for prediction systems to manage and reduce the dangers linked to forest fires.

Our project utilizes data from GHCN weather records, NASA fire data, and Canadian government humidity data to explore the dependencies between weather attributes—specifically precipitation, temperature, and humidity—and forest fire occurrences. The primary focus is on identifying these dependencies and subsequently developing a model capable of predicting the probability of forest fires in different regions across Canada.

Problem Statement

Forest fires present risks to the environment, economy and public health. Canada has seen an increase in the frequency and intensity of these fires in times due to factors like climate change, human actions and natural elements. The vast Canadian forests and varied climate conditions make it difficult to predict and manage these fires effectively. Having forecasts of forest fire occurrences is crucial for minimizing their impact and implementing safety measures.

Our project initiative, Forest Fire Prediction, aims to analyze fire and weather trends to find connections between them. Additionally, we concentrate on predicting the probability of forest fires in locations to improve preparedness efforts and resource allocation.

Data Collection and Preprocessing

The success of our forest fire prediction system lies very heavily on the quality and extent of data used. So, in this regard, we are combining datasets of fire incidents with current weather reports, from GHCN weather data, and humidity levels, from the Canadian Government to give a robust framework for our forecasting model. Gathering and organizing the data was a very important procedure that ensures the ingesting of trustworthy data by our model.

Data Sources

1. **Historical Fire Data:** We obtained detailed records of occurrences of forest fires in the past years across Canada, including their location, date, time, size, intensity, etc. This data was sourced from NASA's FIRM System (https://firms.modaps.eosdis.nasa.gov/country/).
2. **Weather Data:** Temperature, precipitation, wind speed, snow, and other meteorological variables were extracted from all weather stations that had operated in Canada. We used data supplied by GHCN.
3. **Humidity Data:** Relative humidity data at different meteorological stations was also retrieved, as humidity is one of the leading factors of fire behavior. This dataset will allow for an analysis of the moisture content in the atmosphere (https://climate-change.canada.ca/climate-data/#/daily-climate-data).

Data Extracting

1. GHCN Weather Data: We accessed GHCN weather data using the compute cluster and extracted data from years after 2015. We ensured that the data was filtered to include only relevant attributes, such as average temperature, precipitation, snow, and maximum temperature.
2. Humidity Data: We collected the humidity data manually by downloading information for all provinces in Canada for years after 2015.
3. Fire Data: We retrieved the fire data by manually downloading it from NASA's website for years after 2015. This data provided essential information about past fire occurrences, which was crucial for training our predictive model and understanding the relationship between weather conditions and fire incidents.

Data Storage and Cleaning

1. GHCN Weather Data: The data for the specified time frame was organized and stored in a dedicated folder named 'ghcn-subset.' To ensure data integrity and optimize storage efficiency, the files were saved in JSON Gzip format ('*.json.gz'). This choice of format not only minimized storage requirements but also facilitated seamless data handling and access for subsequent analysis.
2. Humidity Data: Since this data was downloaded individually for each province, the data was initially in separate CSV files. We wrote a script using pandas to merge all these CSV files into a single, unified dataset. This merged dataset was then converted into a spark DataFrame, allowing us to store the data as partitioned files. This process streamlined the data management and ensured consistency across all provincial data, facilitating more efficient analysis.
3. NASA Fire Data: A similar approach was applied to the NASA dataset. The data, available as different CSV files for various years, contained numerous attributes. We read all the files into a Spark DataFrame, from which we extracted the useful fields. The refined data was then stored as partitioned CSV files.

Data Integration

After cleaning, the next step was to integrate the various datasets. This involved:

Merging Datasets

1. Merging Weather and Humidity Dataset:
    a. We calculated the distance between all distinct locations in the weather data and the distinct locations in the humidity data.
    b. For each weather record, we identified the closest location in the humidity dataset by finding the smallest difference in distance.
    c. The humidity value from the nearest corresponding humidity data point was then added to the weather dataset.
    d. This integration resulted in a comprehensive dataset, referred to as the weather_humidity_dataset, containing both weather and humidity information for accurate analysis and model training.
2. Merging Weather and Fire Dataset:
    a. We applied a similar distance calculation method, this time between locations in the fire data and the weather_humidity_dataset.

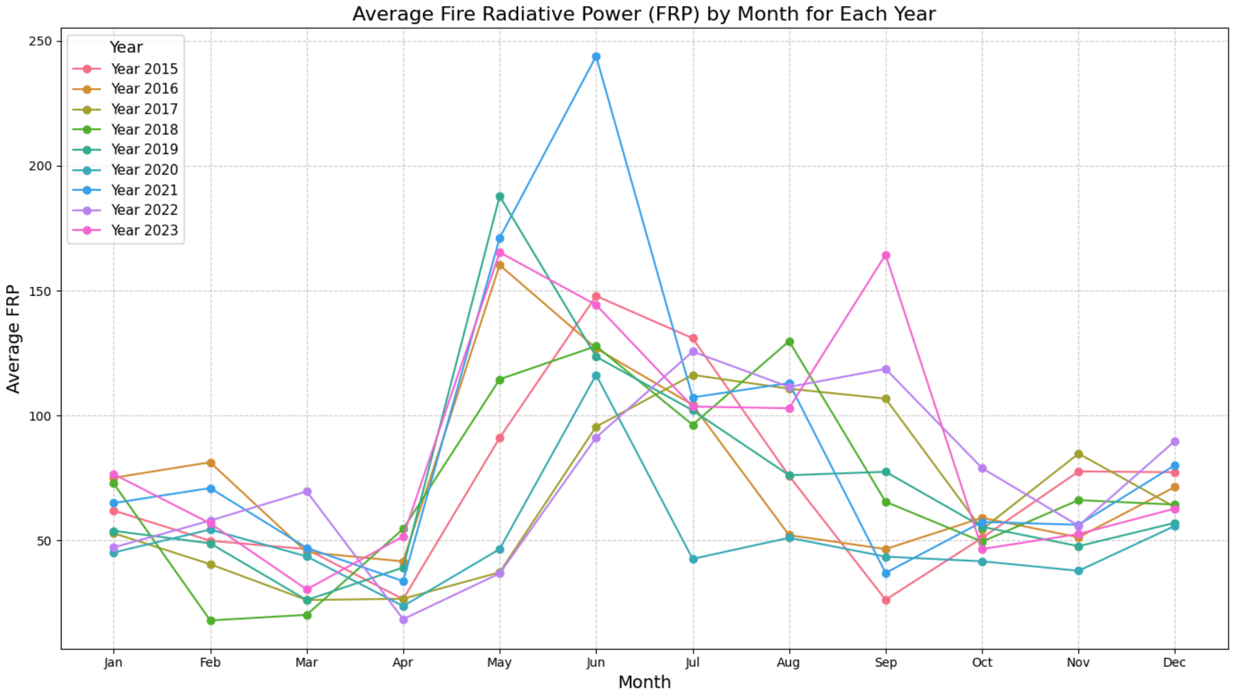
    b. By identifying the smallest difference in distance, we matched the relevant weather and humidity data to each fire record.
    c. This provided detailed information about the temperature, humidity, and other weather conditions on the days of the fires, allowing for a comprehensive analysis of the factors influencing fire occurrences.

3. Classifying No-Fire Points:
    a. To classify data points as "no fire," we needed to identify weather conditions not associated with fire incidents.
    b. We performed a left anti-join between the weather_humidity_dataset and the fire dataset.
    c. Since we had more ‘no-fire’ data points than ‘yes-fire’ we only used 1,000,000 of those points to make it more balanced but still accurately representative of the real world.
    d. This operation helped us find weather data points that were not matched with any fire occurrences, effectively creating a set of conditions under which no fires were reported.

All the data merging and integration processes were carried out using PySpark, which enabled streamlined handling and processing of substantial datasets.

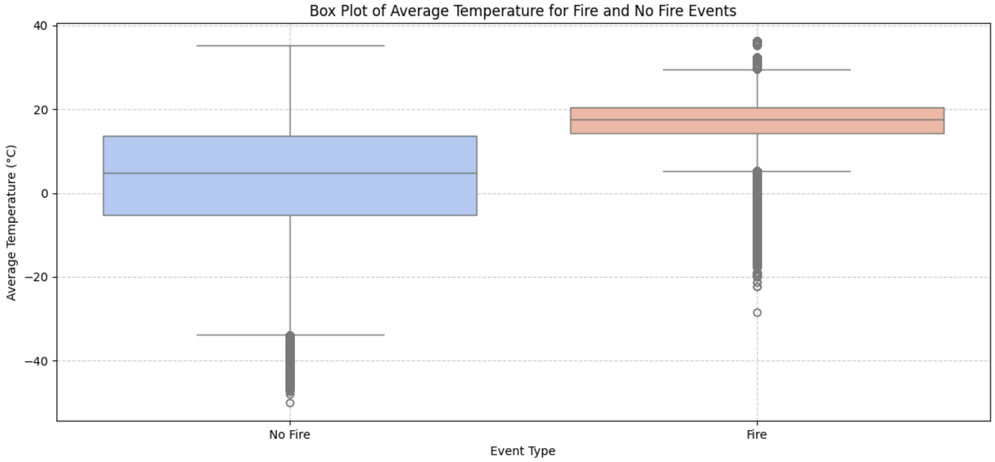
Graphical Analysis

The line graph provided shows the average Fire Radiative Power (FRP) by month for each year from 2015 to 2023. Each line represents a different year, with the x-axis denoting the months from January to December and the y-axis representing the average FRP.



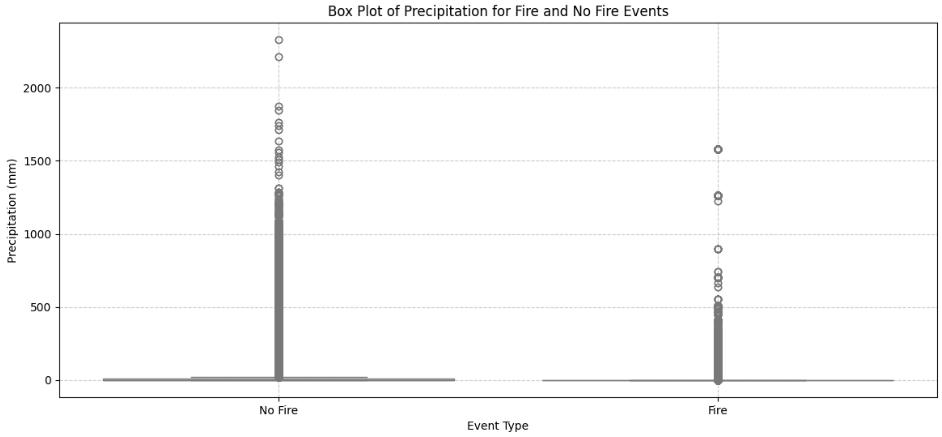
It can be observed that fire activity peaks during late spring to early summer, particularly in May and June, with a noticeable decline by September. From October to December, FRP values generally drop, suggesting lower fire activity in cooler, potentially wetter conditions.

The box plot illustrates the distribution of average temperature (TAVG) for two categories: "No Fire" and "Fire" events.

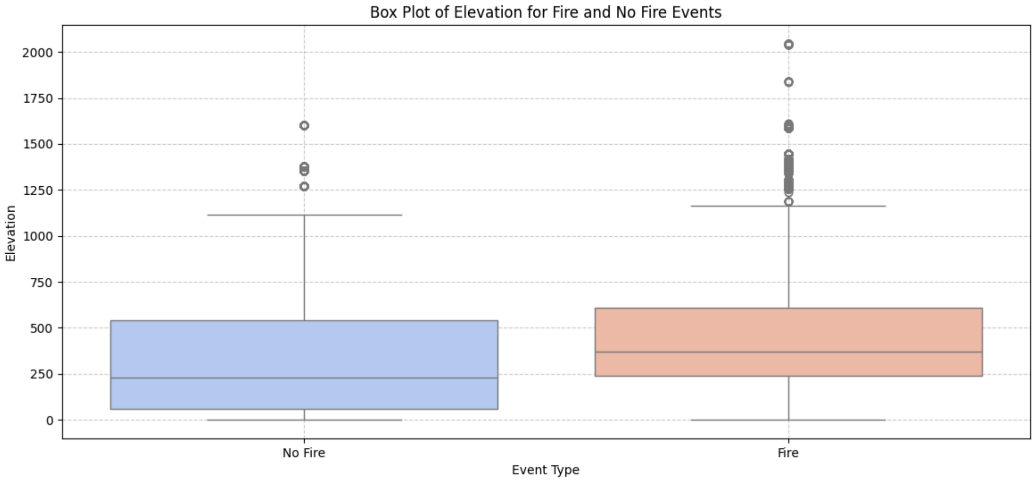


It reveals a strong association between higher temperatures and fire occurrences. Fires are more likely at around 20°C, with a tighter temperature range compared to no-fire events, which have a broader range and median around 5°C. This emphasizes the importance of temperature in fire risk assessment, with warmer conditions favoring fire occurrences.

The box plot illustrates the distribution of precipitation (PRCP) for "No Fire" and "Fire" events.



Even though both categories have a median precipitation close to 0 mm, indicating most events occur under dry conditions. High precipitation levels are more common in no-fire events, highlighting the protective effect of rainfall against fires.



From the boxplot, fires tend to occur at slightly higher median elevations compared to no-fire events. However, the elevation range for both events is quite similar, with most occurrences below 500 meters so the relationship, if any, between elevation and fire is not obvious.

Data Analysis and Techniques

In our forest fire prediction project, comprehensive data analysis was crucial to understanding the underlying factors contributing to fire occurrences. We employed a variety of statistical and computational techniques to explore the relationships between different weather attributes and the likelihood of fires.

Chi-Square Analysis

One of the primary methods used was the **Chi-Square Analysis**, a statistical test designed to determine the association between categorical variables.

The primary objective of conducting the chi-square analysis was to examine whether there was a significant association between the occurrence of fires (categorized as fire or no-fire) and specific weather attributes. We focused on three key attributes: average temperature, precipitation, and elevation. These factors were chosen based on their known or suspected influence on fire behavior and the environment.

1. Categorization of Weather Attributes:
    a. To facilitate the chi-square analysis, we categorized continuous weather attributes into discrete levels. For both temperature and precipitation, we defined three levels: **high, medium, and low**. These categories were determined based on natural breaks in the data distribution.
    b. For elevation, we divided points into : **Plains and Plateaus, Hills and Mountains**.
    c. For each category, we counted the occurrences in both fire and no-fire situations. This counting was essential for creating the contingency tables required for the chi-square test.
2. Data Preparation and Cross-Tabulation:
    a. We organized the data into a cross-tab format, which displayed the counts of observations for each combination of weather attribute level and fire category. This table provided a clear view of the distribution of data across different categories.

TAVG category	No Fire	Fire
Low	914842	59692
Medium	115812	373728
High	792490	858996

PRCP category	No Fire	Fire
Low	1725370	1277206
Medium	1596	840
High	96178	14370

Elevation category	No Fire	Fire
Plains-Plateaus	807710	125190
Hills	992724	1135190
Mountains	22710	32036

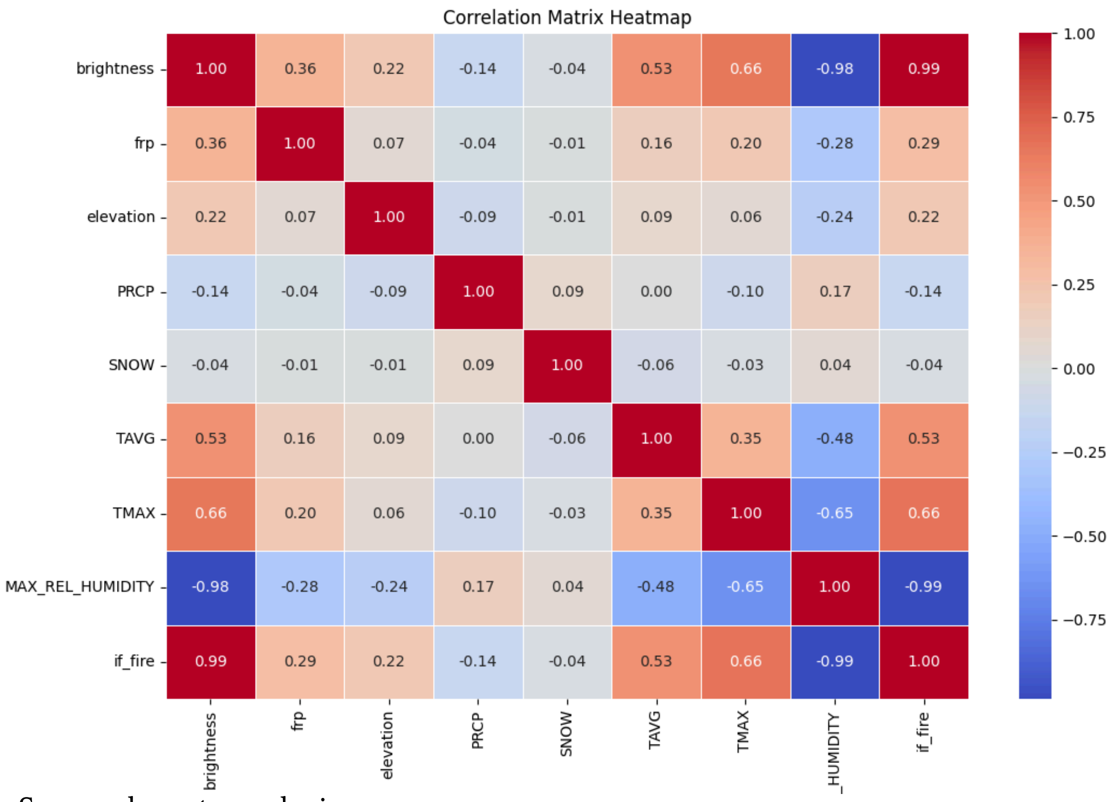
3. Chi-Square:
    a. We then applied the chi-square test. The test compared the observed frequencies (counts from the data) with the expected frequencies (calculated assuming no association between variables).

Category	P-Value
Fire and TAVG	0.0000000000
Precipitation and TAVG	0.0000000000
Elevation and TAVG	0.0000000000

    b. As evident from the extremely low p-value (< 0.05) obtained from our chi-square analysis, we rejected the null hypothesis, which stated that fire occurrences and weather attributes are independent. The low p-value indicates a statistically significant association between fire incidents and the weather attributes we analyzed, such as temperature, precipitation, and elevation.

Correlation Coefficient

We calculated the correlation coefficients between different weather attributes and the possibility of a fire.



	brightness	frp	elevation	PRCP	SNOW	TAVG	TMAX	_HUMIDITY	if_fire
brightness	1.00	0.36	0.22	-0.14	-0.04	0.53	0.66	-0.98	0.99
frp	0.36	1.00	0.07	-0.04	-0.01	0.16	0.20	-0.28	0.29
elevation	0.22	0.07	1.00	-0.09	-0.01	0.09	0.06	-0.24	0.22
PRCP	-0.14	-0.04	-0.09	1.00	0.09	0.00	-0.10	0.17	-0.14
SNOW	-0.04	-0.01	-0.01	0.09	1.00	-0.06	-0.03	0.04	-0.04
TAVG	0.53	0.16	0.09	0.00	-0.06	1.00	0.35	-0.48	0.53
TMAX	0.66	0.20	0.06	-0.10	-0.03	0.35	1.00	-0.65	0.66
MAX_REL_HUMIDITY	-0.98	-0.28	-0.24	0.17	0.04	-0.48	-0.65	1.00	-0.99
if_fire	0.99	0.29	0.22	-0.14	-0.04	0.53	0.66	-0.99	1.00

Some relevant conclusions were:

- **TMAX and if_fire**: A positive correlation (0.66) suggests that higher maximum temperatures are associated with fire events.
- **Brightness and TAVG**: A moderate positive correlation (0.53) indicates that higher average temperatures are associated with higher brightness values, which may relate to fire events.
- **MAX_REL_HUMIDITY and if_fire**: A strong negative correlation (-0.99) suggests that higher relative humidity significantly decreases the likelihood of fire events. This aligns with the understanding that drier conditions (lower humidity) are more conducive to fires.

Random Forest Classifier

The second analytical tool employed in our forest fire prediction project was the **Random Forest Classifier**, a powerful ensemble learning method used for classification tasks. We configured the classifier with a maximum depth of 10 and set the number of trees (numTrees) to 10. The model was designed to predict fire occurrences based on key weather attributes.

1. **Data Preparation:**
    a. We split our dataset, containing both fire and no-fire points, into two subsets: 75% for training and 25% for testing.
    b. The input features used for the model were:
    ○ **TAVG** (Average Temperature)
    ○ **PRCP** (Precipitation)
    ○ **Elevation**
    c. The output label was a binary value (0/1) representing the presence or absence of fire, respectively.
2. **Model Training and Prediction:**
    a. The Random Forest model was trained on the training dataset, learning the patterns and relationships between the input features and the target labels.
    b. After training, we used the model to make predictions on the testing dataset, allowing us to evaluate its performance on unseen data.
3. **Evaluation Metrics:**
    a. We employed the **MulticlassClassificationEvaluator** to assess the model's performance. The key metrics obtained were:
    ○ **Accuracy:** 0.8478858059580847
    ○ **Weighted Precision:** 0.8519792155722647
    ○ **Recall:** 0.8478858059580847
    ○ **F1 Score:** 0.8474320859976132
    b. These metrics indicated that the model performed well in distinguishing between fire and no-fire scenarios, with a high level of accuracy and balanced precision, recall, and F1 scores.
4. **Feature Importances:**
    a. The feature importances for our model were as follows:
    ○ **Elevation:** 0.3089914520582509
    ○ **PRCP:** 0.07579311644192799
    ○ **TAVG:** 0.615215431499821
    b. These values indicate that average temperature was the most influential factor in predicting fire occurrences, followed by elevation and precipitation.

c. Overall, the Random Forest Classifier proved to be an effective tool for our project, providing valuable insights into the factors influencing fire occurrences and enabling accurate predictions. This model's results will aid in better understanding and managing forest fire risks in various regions.

Conclusion

This project combines weather, fire and humidity data and aims to find relationships between weather attributes and the occurrence and intensity of fires. We also trained a model that can successfully predict the occurrence of fire based on those attributes.
Some key insights:

1. With some graphical analysis we found that intense fire activity peaks in early summer and diminishes in colder and wetter conditions later in the year. Furthermore, a strong association between higher temperatures and fire occurrences, with fires more likely at median temperatures around 20°C. This underscores the critical role of temperature in fire risk assessment, as warmer conditions favor fire events.
2. We found very low p-value in our Chi-Square Test analysis for occurrence of fire and temperature, precipitation and elevation, indicating dependence.
3. On further analysis, we found strong positive correlation (0.66) between fire and TAVG.
4. Additionally, our random forest classifier achieved an accuracy of approximately 84% in predicting fire and no-fire events, demonstrating the model's effectiveness in identifying fire risk based on environmental factors.

This comprehensive analysis provides valuable insights for fire management and prevention strategies, emphasizing the importance of monitoring weather conditions, particularly temperature and humidity.

Even with these positive results, some limitations were found regarding the quality of data and inconsistencies in mapping weather data onto cases of fires. These pose some challenges to the improvement of data collection methods and refinement of the model.

The Forest Fire Prediction project gave very valuable insight and a robust framework for understanding the risks associated with forest fires in Canada. It also gave critical insight into data science's role in resolving some of the most complex challenges in the environment, thereby setting a future way forward toward more effective management and response strategies for forest fires.

Limitations and Challenges

Our project had a lot of limitations and challenges to overcome.

1. **Quality of Humidity Data:** A challenge was the quality of the humidity data extracted for the project. We found that a significant portion of the dataset contained null values, making it unreliable for use in our model. The lack of complete and accurate humidity data limited our ability to analyze its impact on fire occurrences fully.
   - In future work, sourcing a more comprehensive and accurate humidity dataset would be crucial to improving model accuracy and reliability.
2. **Accuracy of Weather Data Mapping:** During the process of finding the nearest weather point to a fire location, we noticed that some distances were not sufficiently small. The potential discrepancy between the actual location of fire incidents and the corresponding weather data points could compromise the model's integrity.
   - An improvement would involve sourcing localized weather data to enhance the accuracy of the input features and the model's predictions.
3. **Data Integration Issues:** One significant challenge arose during the data integration phase. When performing a left anti-join to identify weather points not associated with fire incidents (to classify them as no-fire points), we unexpectedly received the entire dataset instead of a subset. This indicated that none of the fire data points were correctly mapped to weather data, despite prior merging efforts to associate weather values with each fire occurrence. This anomaly suggested a potential issue with data manipulation, possibly involving an exchange of locations or a schema mismatch. To resolve this, we had to recompute all our datasets, ensuring accurate data merging and mapping.

Project Experience Summary

Bhakti Bhanushali:

- Led data collection and preprocessing efforts to integrate multiple data sources—fire, weather, and humidity—into a single cohesive framework.
- Developed and implemented a random forest model by tuning the hyperparameters to optimize the performance of the model.
- Conducted exploratory data analysis to identify the key patterns and correlations within the data.
- He then made sure that the findings were communicated most effectively by means of designed visualizations, such as risk maps and feature importance plots.
- Assisted team members with model validation and report writing.

Ketan Dhingra:

- Focused on feature engineering through the creation of derived metrics, like the Fire Weather Index and lag features, to improve model accuracy.
- Assisted in cleaning the data in particular handling missing values and detecting outliers.
- Built the project's visualization dashboard, enabling dynamic exploration of the results of prediction.
- Contributed to the Project Report, detailing methodology and technical issues.

Harveer Virk:

- Managed project timelines and assigned tasks to facilitate smooth collaboration and an effective progress tracking process.
- Managed the predictive model's deployment by setting up the running infrastructure of the model and generating predictions.
- Literature reviewing and research were done in order to help choose algorithms and methodologies.
- Provided quality assurance on code and documentation through reviews, ensuring its accuracy and clarity.