Dawson AI Accelerator Program

Fire Guard

AI Project Blueprint

Ethan Tran

Gabriel Saberian

Parsa Jafarian

Anaïs Perron

AI Launch Lab

Dawson AI

1 PROBLEM IDENTIFICATION

1.1 Problem Definition

The primary problem we aim to solve is the prediction of future wildfires using historical data and determining optimal locations for fire shelters based on those predictions. The goals of this project include:

- Develop a machine learning model to analyze historical wildfire data and predict the likelihood, location, and severity of future wildfires
- Utilize various data sources, including weather patterns, vegetation types, historical fire records, and human activity, to improve prediction accuracy
- Use the predictions from the wildfire model to identify safe and strategic locations for fire shelters
- Optimize shelter locations to ensure maximum coverage and accessibility for at-risk populations

Significance

Wildfires pose a significant threat to both human life and property. By accurately predicting wildfires and strategically placing fire shelters, we can:

- Enhance public safety and preparedness
- Reduce the loss of life and property
- Improve resource allocation for firefighting and emergency response teams

1.2 Target Audience

Primary Beneficiaries

- Residents in wildfire-prone areas
- Local and state government agencies responsible for public safety and emergency management
- Firefighting and emergency response teams

Secondary Stakeholders

- Environmental researchers and scientists
- Insurance companies assessing wildfire risks
- Urban planners and developers in wildfire-prone regions

1.3 Scope & Constraints

In-Scope

- Collection and analysis of historical wildfire data
- Development of ML model for regression to be used for predicting at-risk locations for wildfires
- Identify optimal placements of fire shelter locations based on predictions
- Display of historical data wildfires on an interactive map
- Display of future predictions on the same map with a heatmap overlay
- Ability to add/update wildfire data using
- Evaluation and validation of model accuracy

Out-of-Scope

- Alerting systems
- Real-time monitoring of wildfires

Constraints

- Data availability and quality
- Computational resources - Handling large datasets
- Model interpretability: Ensuring stakeholders can understand and trust the model's predictions

1.4 Success Criteria

- **Model Accuracy:** Achieve an accuracy rate of at least 85% in predicting high-risk areas for fire incidents.
- **Precision:** Attain high precision, correctly identify true positive high-risk areas with minimal false positives
- **Heat Map Usability:** Obtain positive feedback from users regarding the clarity and usability of the heat map, ensuring users can easily interpret and use the heat map for decision-making.
- **Visual Display Quality:** Create a visually appealing and user-friendly interface for displaying the heat map and other data visualizations. The display should be interactive, easily navigable, and provide clear, actionable insights.
- **AI and Machine Learning Performance:** Ensure the AI and machine learning models used in the project are efficient. The models should be able to process large datasets and near-real-time predictions.

2 DATA COLLECTION & PREPARATION

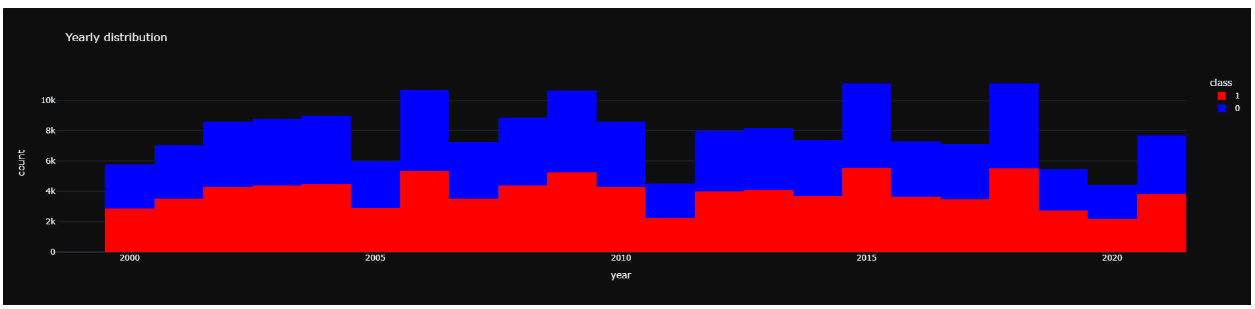
- Weather data: ClimateData.ca
- 1950-2021 Wildfire dataset
- Canada Historical Climate Data Search
- Canada Climate data extraction tool
- ArcGIS for live wildfire data

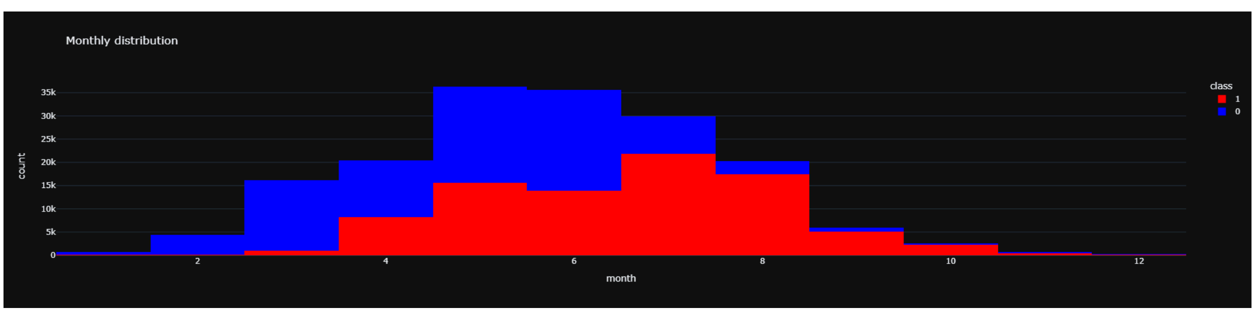
3 EXPLORATORY DATA ANALYSIS

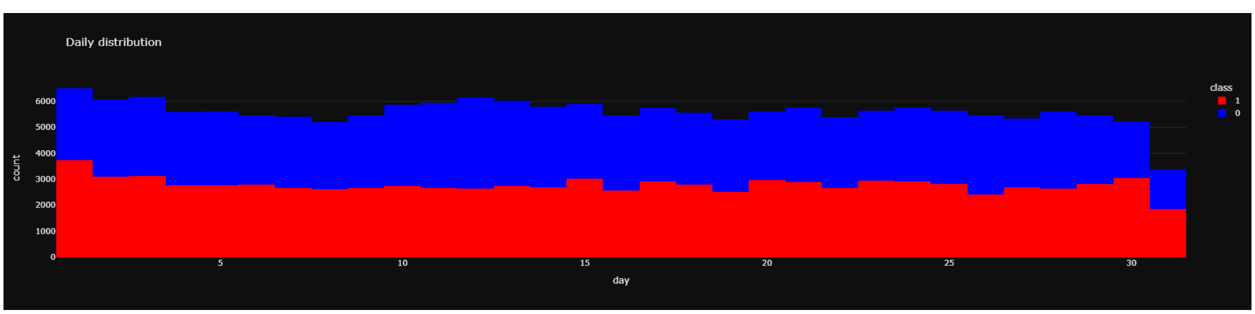
Further details about the data analysis and dataset format can be found in `exploratory-data-anlysis.ipnyb`. This section is a summary of what was found with the data visualizations.

3.1 Data Visualizations

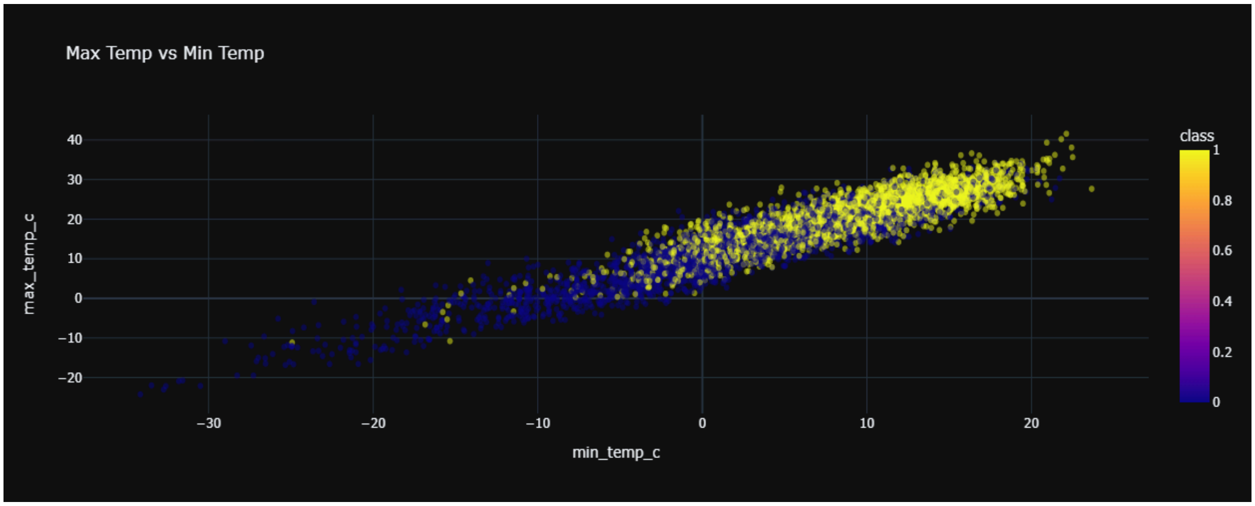
Distributions of data across time

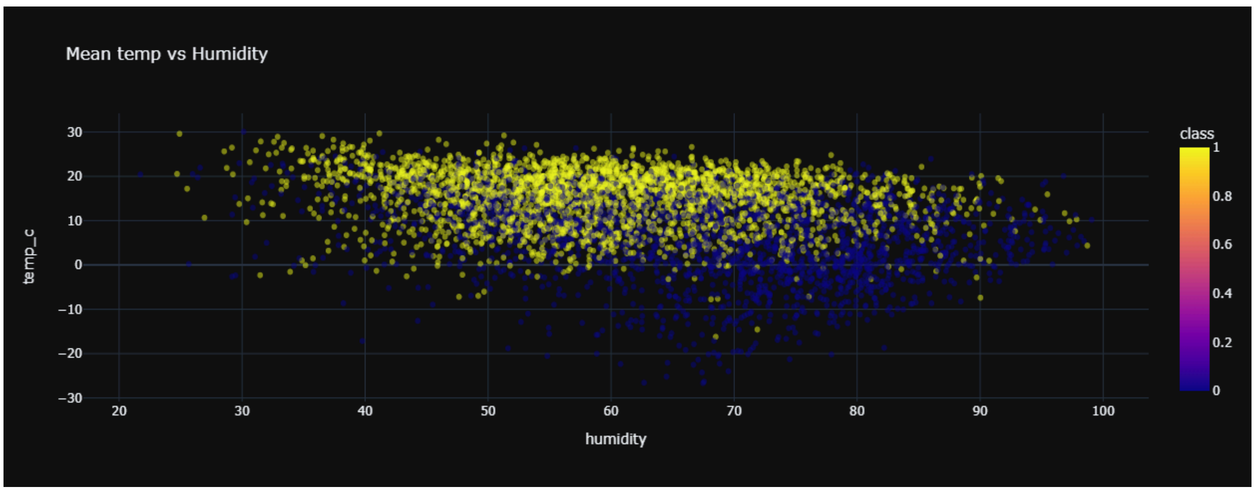


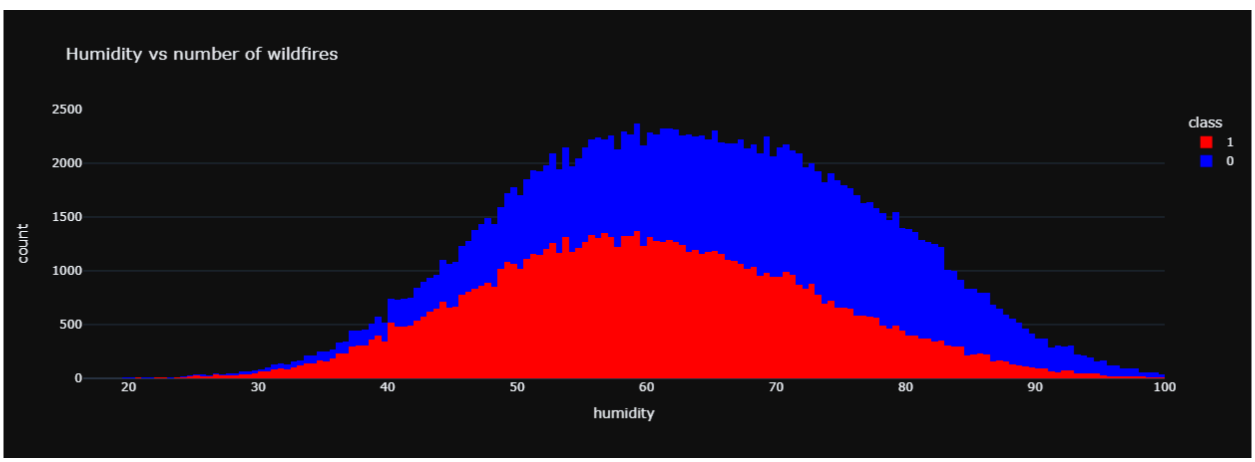


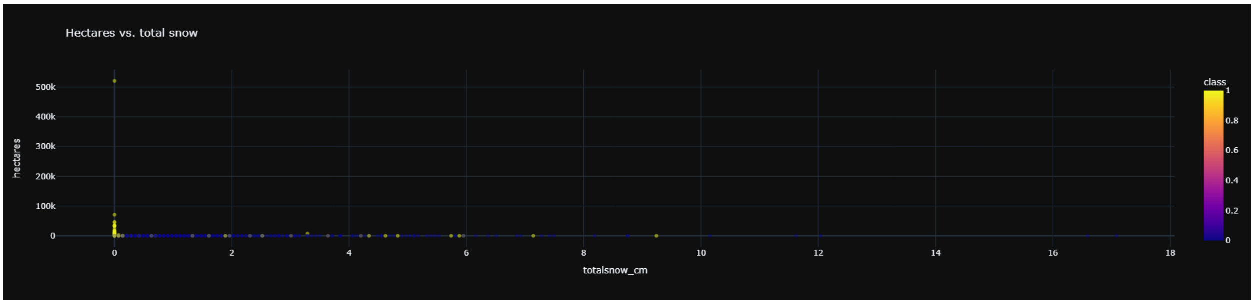


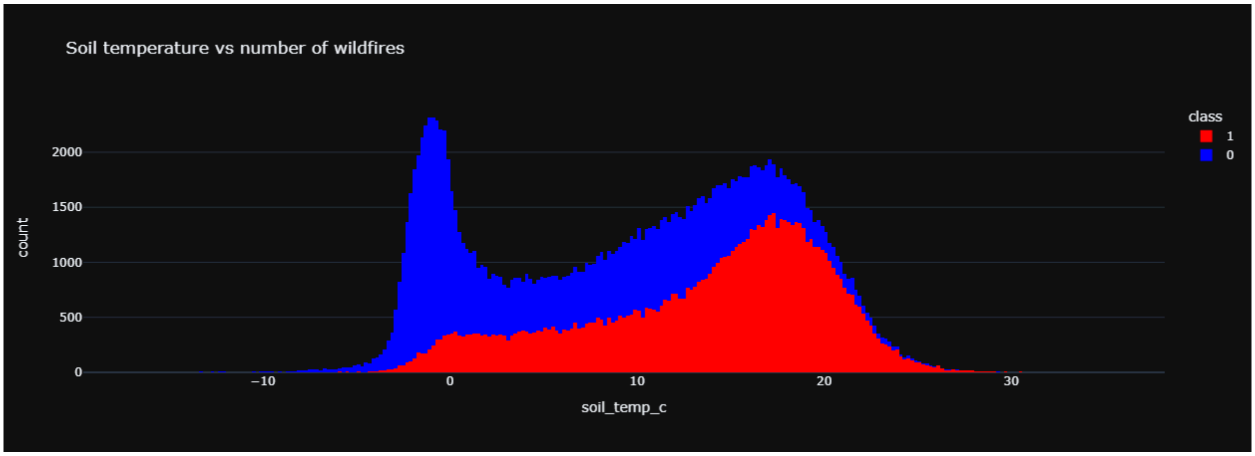
Weather data analysis

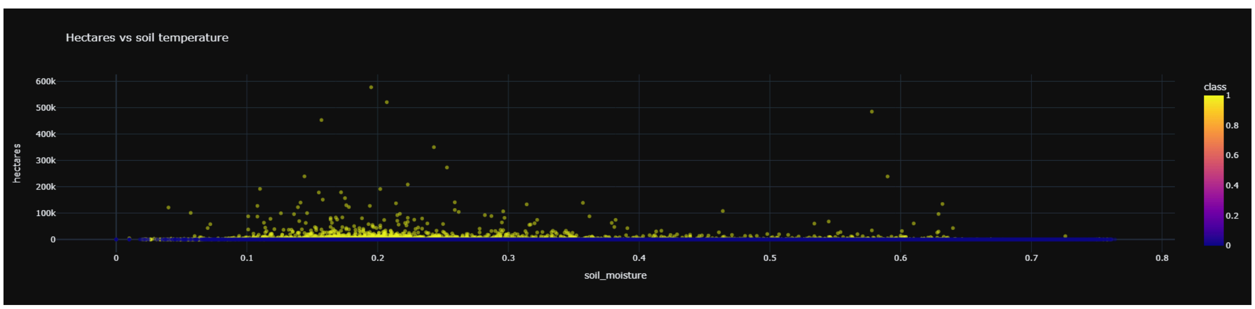


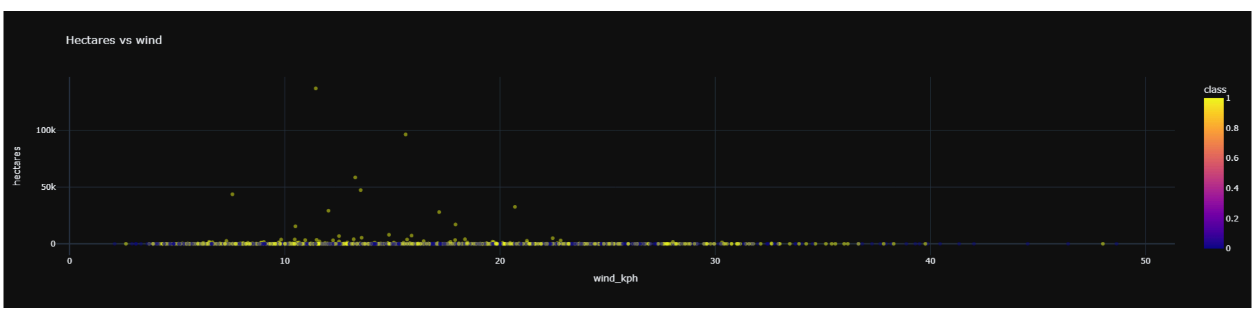




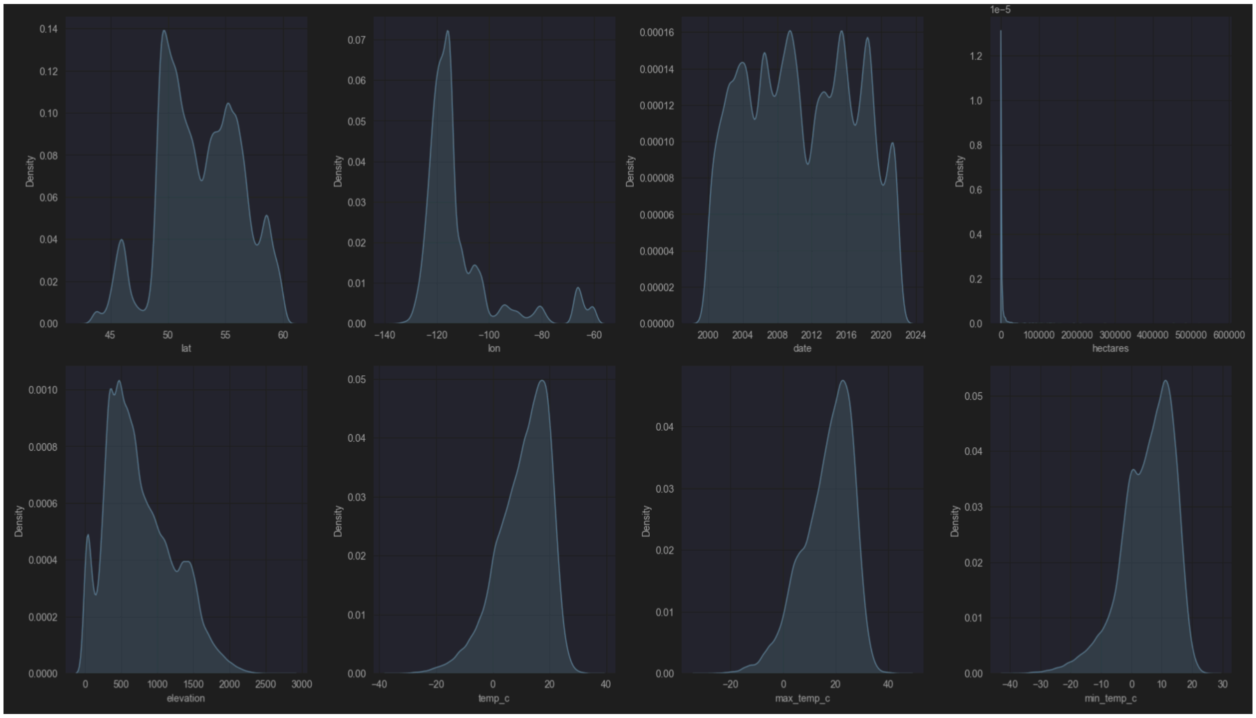

Hectares vs. total snow
hectares
500k
400k
300k
200k
100k
0
0
2
4
6
8
10
12
14
16
18
totalsnow_cm
class
1
0.8
0.6
0.4
0.2
0

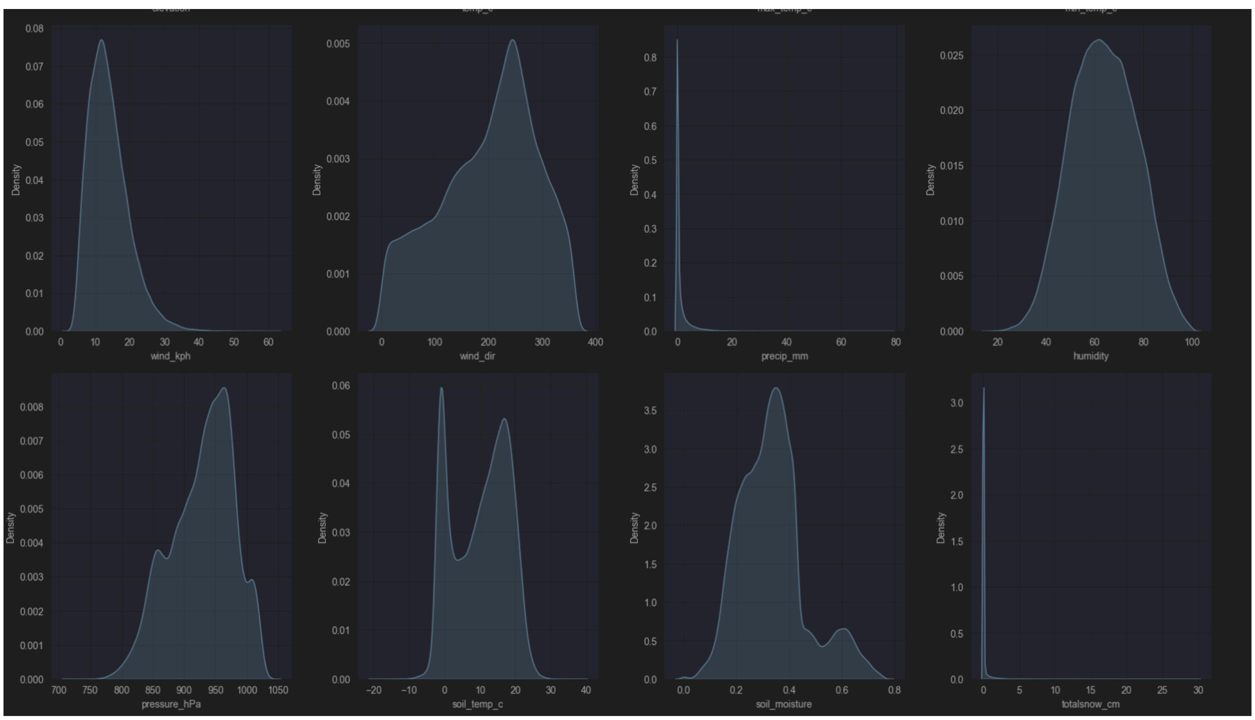

Soil temperature vs number of wildfires
count
2000
1500
1000
500
0
−10
0
10
20
30
soil_temp_c
class
1
0


Hectares vs soil temperature
hectares
600k
500k
400k
300k
200k
100k
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
soil_moisture
class
1
0.8
0.6
0.4
0.2
0

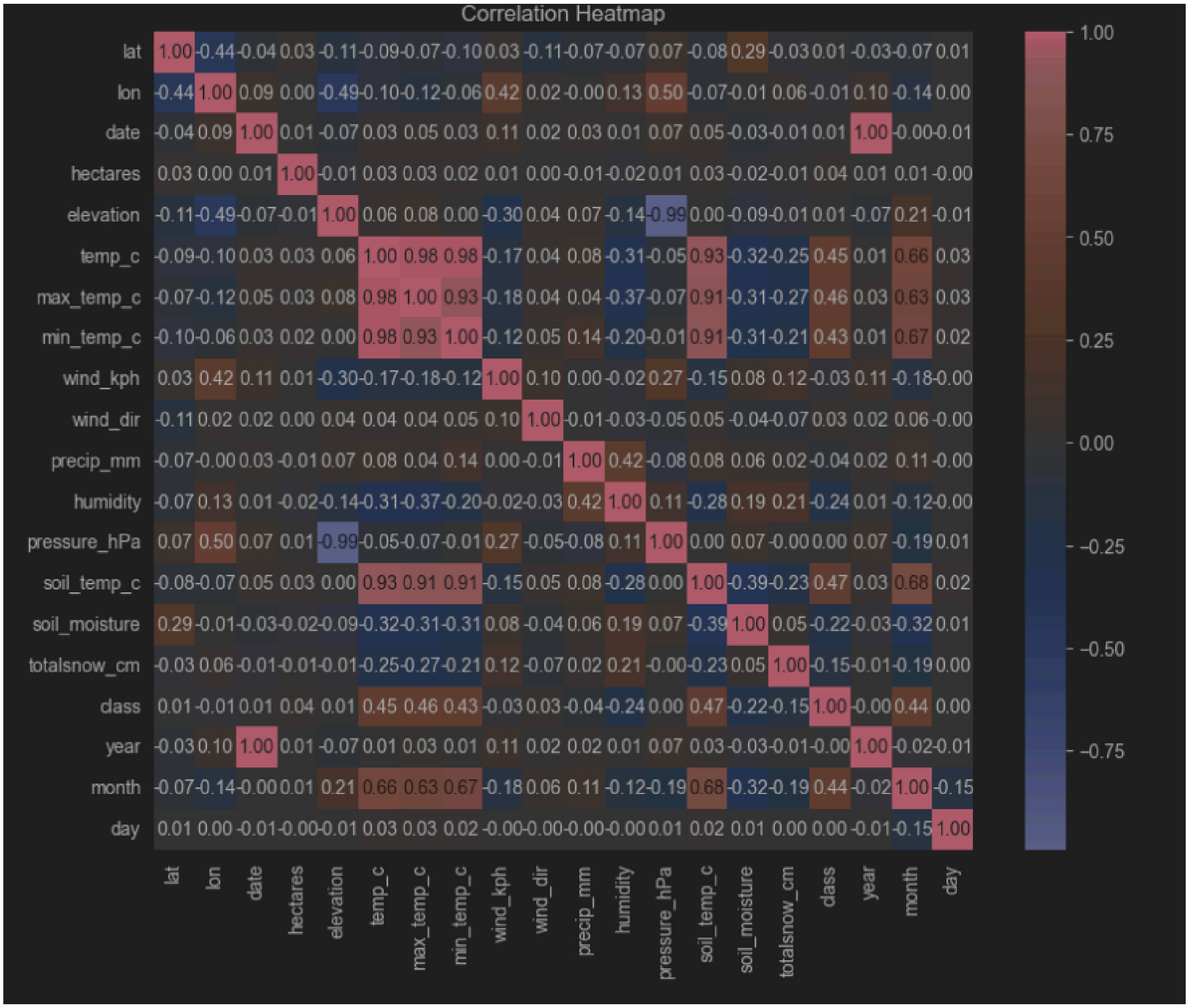

Hectares vs wind
hectares
100k
50k
0
0
10
20
30
40
50
wind_kph
class
1
0.8
0.6
0.4
0.2
0

Distributions of data





Correlation



3.2 Pattern Identification

We can draw many conclusions from the weather data available. The first thing we verified was the distribution of data across time. From this, it is clear that the distributions of data across time and weather data are spread as we would expect. There is roughly an even number of available data between years and days, and this is to be expected since there is no reason why there would be more or less wildfires for a given day/year. The most interesting, however, is the monthly distribution. We can also see that the monthly distribution follows a gaussian curve, but towards the middle (hotter months) it is a lot more likely to have wildfire than not. The peak for non-wildfire points are more aggregated towards the left, cooler months, suggesting that it is less likely to have wildfires. This makes sense, since in hotter months (june-july-august) the temperature is higher and therefore more likely to have wildfires. The distribution for most of the weather data also seems to follow a gaussian distribution, including temperature, soil temperature, humidity, wind, pressure.

We can clearly see from the plots the mean, max, and min temperature seem to have the greatest impact on the likelihood of a wildfire. The distribution shows that it is a lot more likely to have wildfires when it is hot, and when it is cool significantly less likely. We can also see that humidity seems to play a significant role. The scatter plot of Temperature vs Humidity shows that wildfires are much more likely when it is higher temperature and lower humidity. The non-wildfire points are all gathered in the bottom right of the plot (high humidity, low temperature) From the plot of soil temperature we can roughly see the same thing, but in greater effect. It does not follow a typical gaussian distribution, but rather has 2 peaks: the peak in the higher temperatures is when there are a lot more wildfires, whereas there is a peak at a soil temperature near 0, where there it is significantly more likely to not have a wildfire. A similar conclusion can be drawn for humidity: unlike soil temperature, this one follows a gaussian distribution, but the peak for wildfires is in the lower humidity area, whereas the peak for non-wildfire data seems to be at the higher humidity. This suggests that there would be a negative correlation between likelihood of wildfire and humidity. The more humid it is, the less likely it is to have a wildfire, and vice-versa. We can also see that there are certain weather features that do not seem to have a great impact, such as precipitation, wind, and snow. This might be due to the fact that precipitation data is too concentrated at 0 mm.

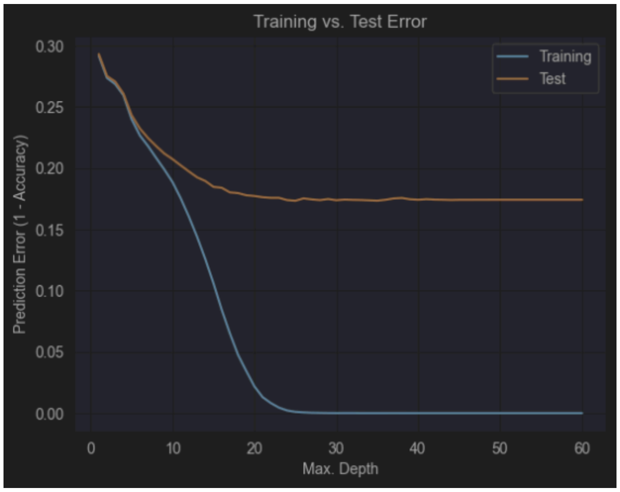
From the correlation plot heatmap, we can see that class (wildfire/no wildfire) is most closely correlated to temperature (mean, max, min), soil temperature, and month. Following this would be humidity, soil moisture, and snow. Pressure, precipitation, wind, as well as location (lat, lon, elevation) seem to have very little correlation with likelihood of wildfire.

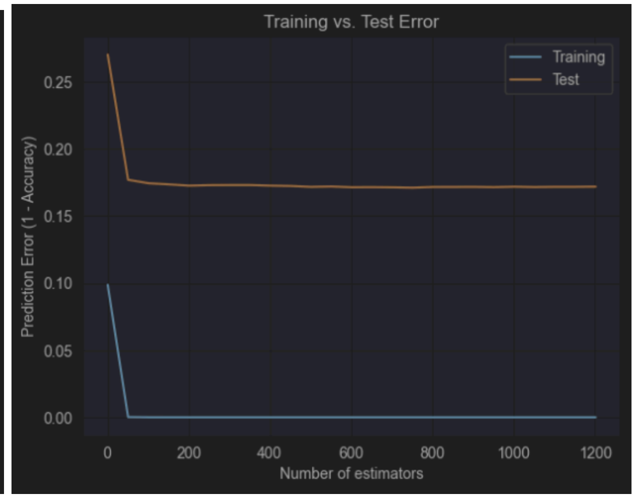
4 MODEL SELECTION AND DESIGN

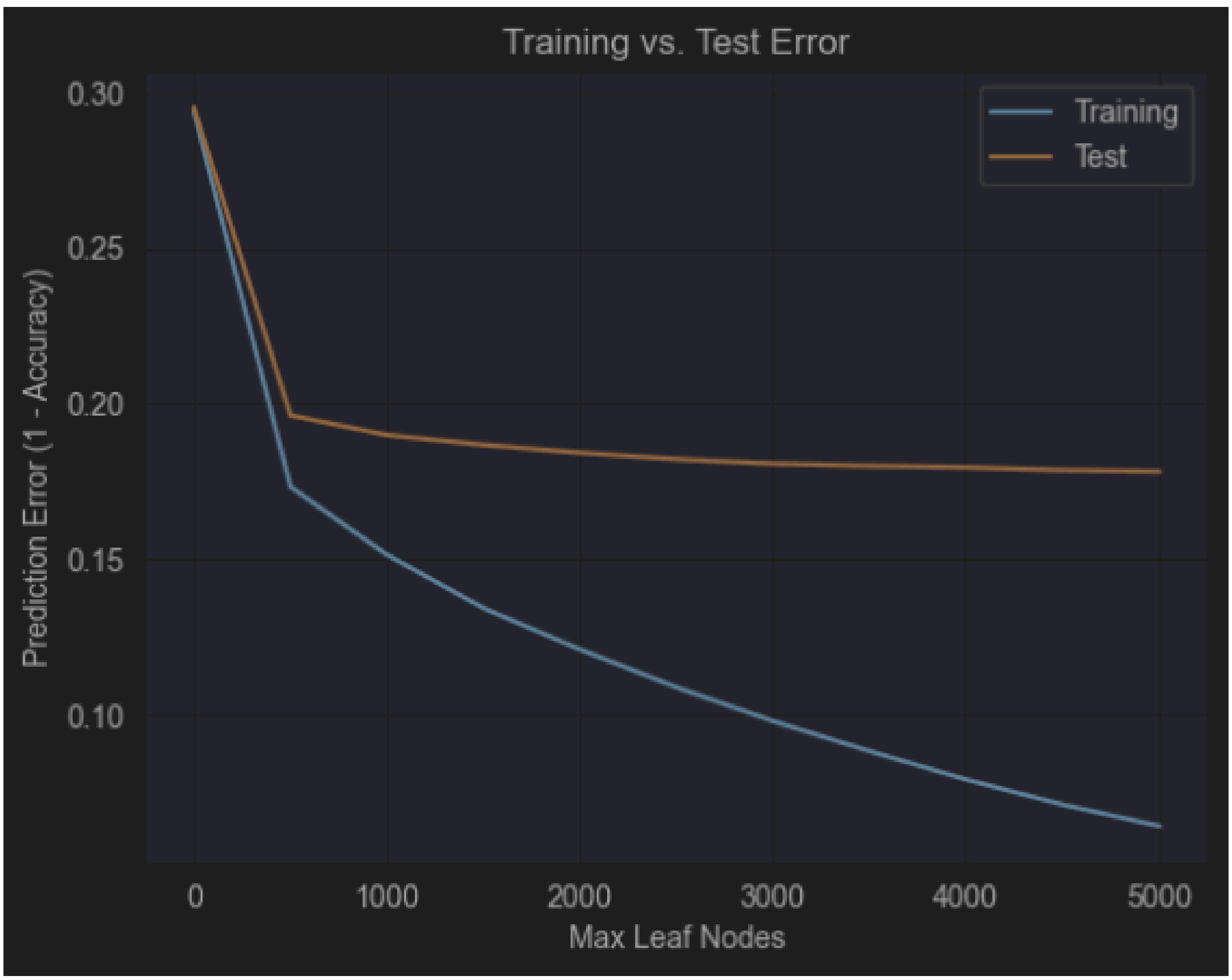
4.1 Process

We first split the dataset into training, test, and validation. Following this, we selected features that were important and applied a `MinMax Scaler` to the input columns. Following this, we performed simple tests to check the viability of the types of model (accuracy, precision, recall, f1). We found that the two best performing models were the XGB booster and the Random Forests model. We then proceeded to hyperparameter tuning. To do this, we first proceeded to test out the most important hyperparameters by plotting a graph of train/test error versus different values for the hyperparameters. After we selected the values for the most important ones, we proceeded with grid search CV to find other optimal parameters. Lastly, we evaluate the model by checking its most important features and plotting a confusion matrix. At least, with best hyperparameters we arrive at an accuracy of about 83%. This is a summary, further details can be found in `modeling.ipnyb.`

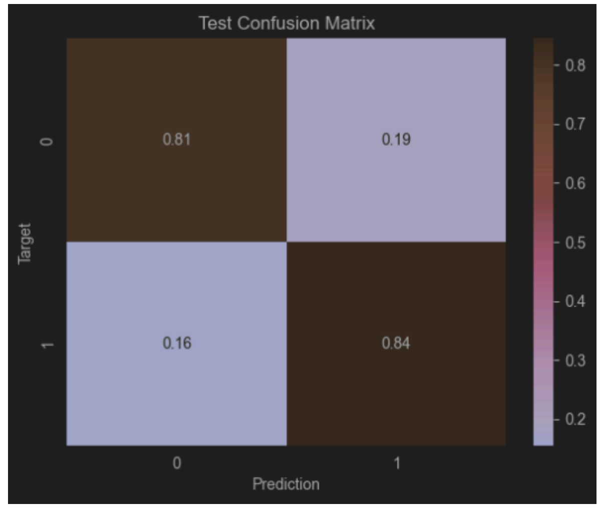
4.3 Hyperparameter tuning

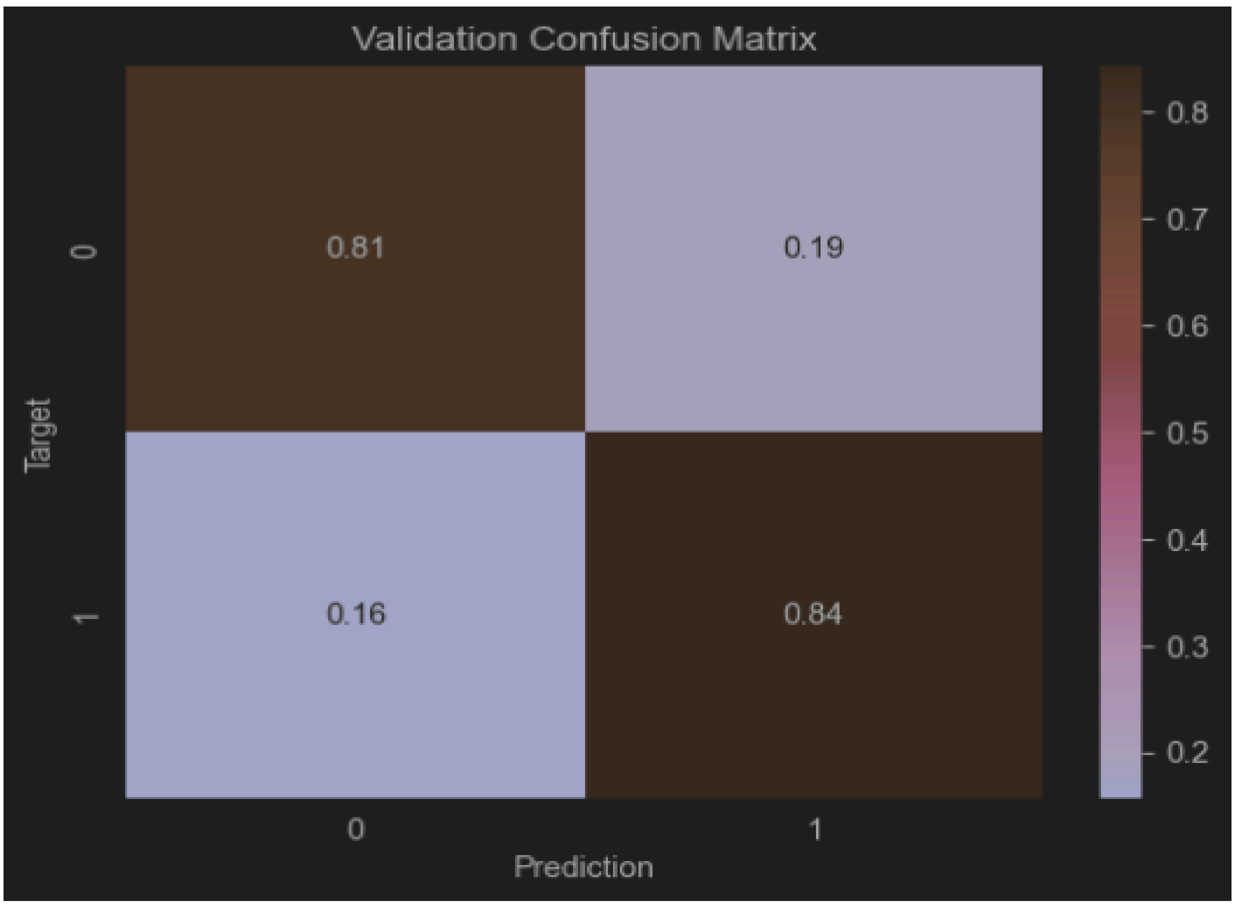


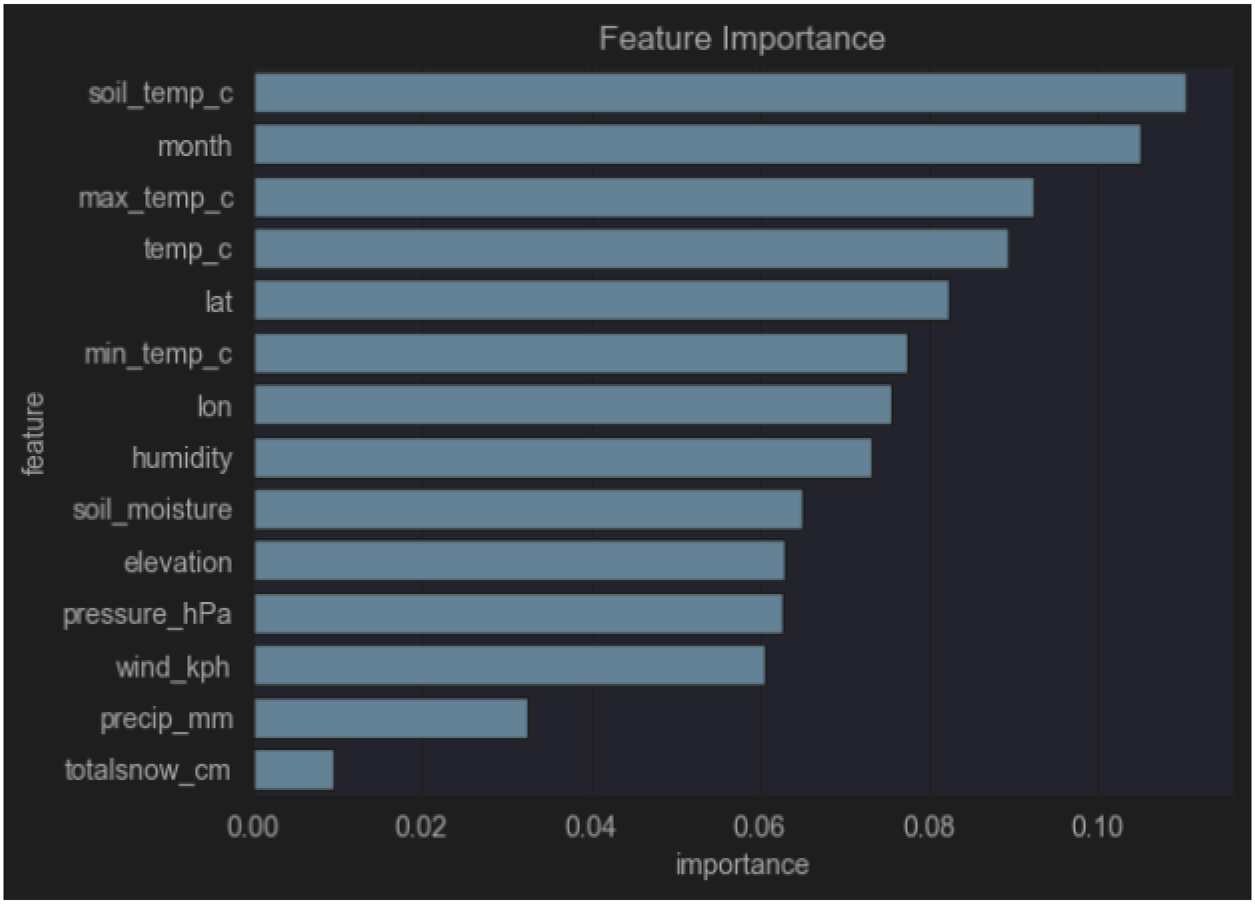




4.2 Evaluation






Feature Importance
soil_temp_c
month
max_temp_c
temp_c
lat
min_temp_c
lon
humidity
soil_moisture
elevation
pressure_hPa
wind_kph
precip_mm
totalsnow_cm
feature
0.00
0.02
0.04
0.06
0.08
0.10
importance