Team Firewatch

Objective :

This project addresses the escalating global concern of wildfires, focusing on the Canadian context. The aim is to develop a machine learning model using data from NASA's VIIRS and MODIS satellites to predict wildfires. The primary goal is to enable early intervention and optimize resource allocation for effective wildfire management. Wildfires pose significant threats to ecosystems, communities, and contribute to environmental issues like increased CO2 emissions. The project narrows its focus to Canada, leveraging advanced machine learning techniques to enhance predictive capabilities.The team explores various machine learning models, including XGBoost, Random Forest, and Multilayer Perceptron Neural Network, to predict wildfires based on weather conditions. The emphasis is on harnessing the power of satellite data to strengthen the accuracy and reliability of the predictive model. With the increasing frequency and intensity of wildfires globally, the need for advanced predictive models is paramount. The proposed machine learning model aims to provide early warnings, empowering timely response and mitigation efforts. By leveraging NASA's satellite data, the project contributes to proactive wildfire management strategies, reducing the impact on both communities and the environment.

Dataset and inputs for the model :

The data that we are using is open source and is sourced from the NASA website. NASA's Fire Information for Resource Management Systems(FIRMS) recorded wildfire data was used as an **input** for training the models, along with this, weather data for the nearby regions was used as an **input** as well.

FIRMS data -The data has information about coordinates, date and time, and light levels available among other features measured through the MODIS(1000m resolution per pixel) and VIIRS(375m resolution per pixel) satellites which provides a confidence level and we used the highest confidence level out of the three provided levels in the dataset to be sure that an actual fire occurred at that time.

https://firms.modaps.eosdis.nasa.gov/download/

Weather data - The initial selection of weather data from Statistics Canada presented **challenges**, as it required downloading data per station from 1840 to the present. However, this approach limited the ability to obtain comprehensive data from all weather stations across Canada.

Subsequently, a more favorable dataset was identified from the National Oceanic and Atmospheric Administration (NOAA). This dataset covered all Canadian stations and included various weather parameters. Notably, the data retrieval process was **constrained** by the necessity to download **monthly data due to limitations on the NOAA website**.

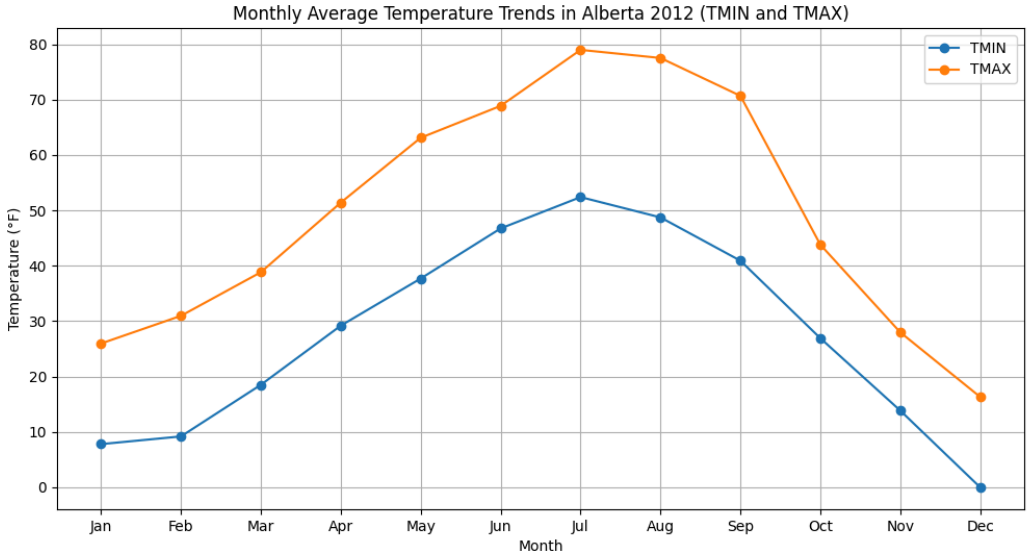
The dataset encompassed diverse weather attributes, including precipitation metrics (DAPR, MDPR, PRCP), snow-related information (SNOW, SNWD), wind characteristics (WDFG, WSFG), temperature parameters (TOBS, TMAX, TAVG, TMIN), and atmospheric conditions (WT03, WT06, WT01). The data was organized into separate folders for each year, underwent a cleaning process, and was consolidated into a unified CSV file.

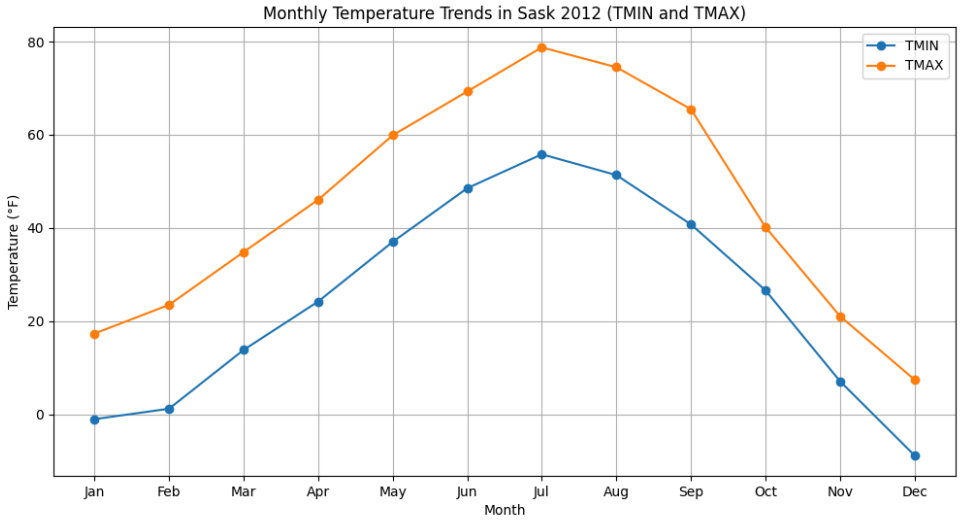
The five core values are:

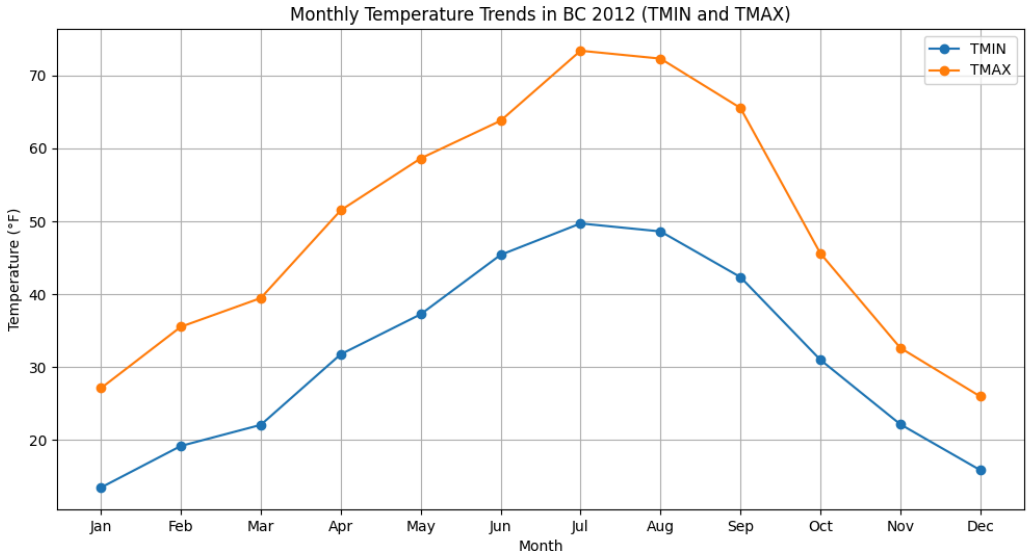
PRCP = Precipitation (mm or inches as per user preference, inches to hundredths on Daily Form pdf file)
SNOW = Snowfall (mm or inches as per user preference, inches to tenths on Daily Form pdf file)
SNWD = Snow depth (mm or inches as per user preference, inches on Daily Form pdf file)
TMAX = Maximum 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

https://www.ncdc.noaa.gov/cdo-web/search;jsessionid=1AB5E708B2355BA7AE94DF1998DCB7FA

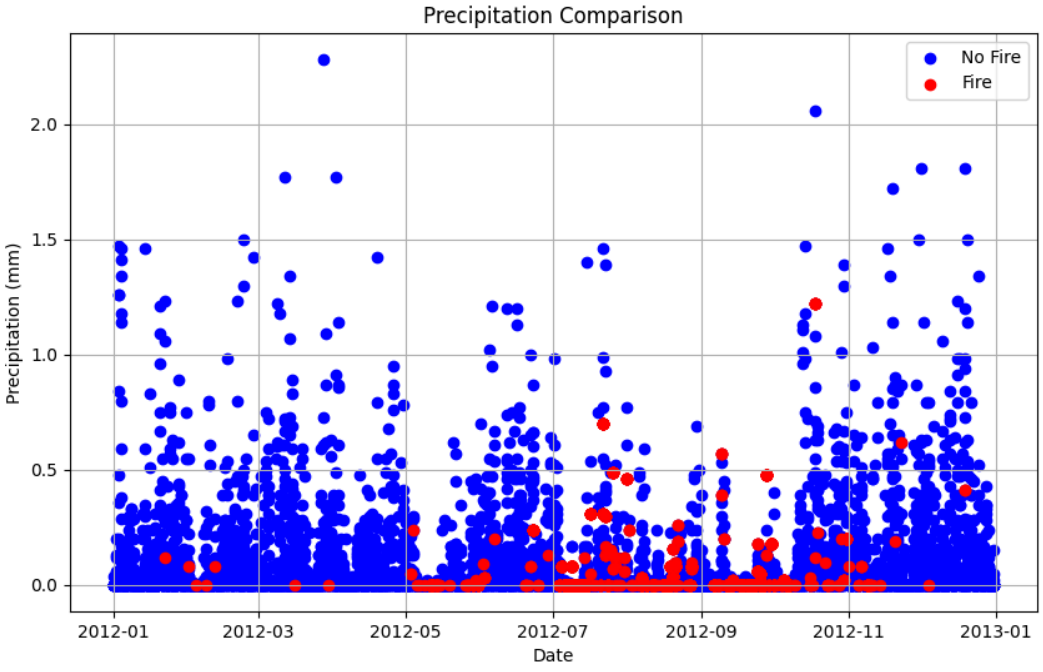
Graphs and charts





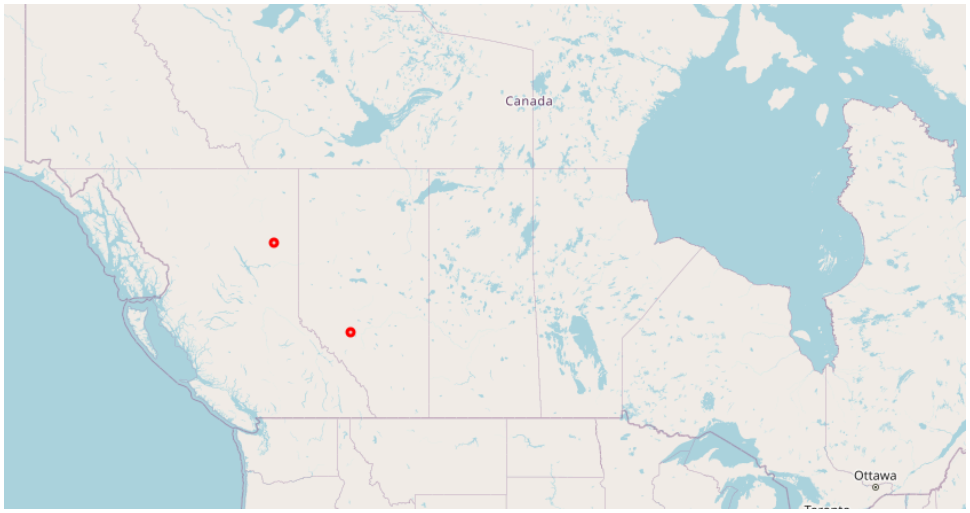


Precipitation in BC on days there was a fire vs on other days

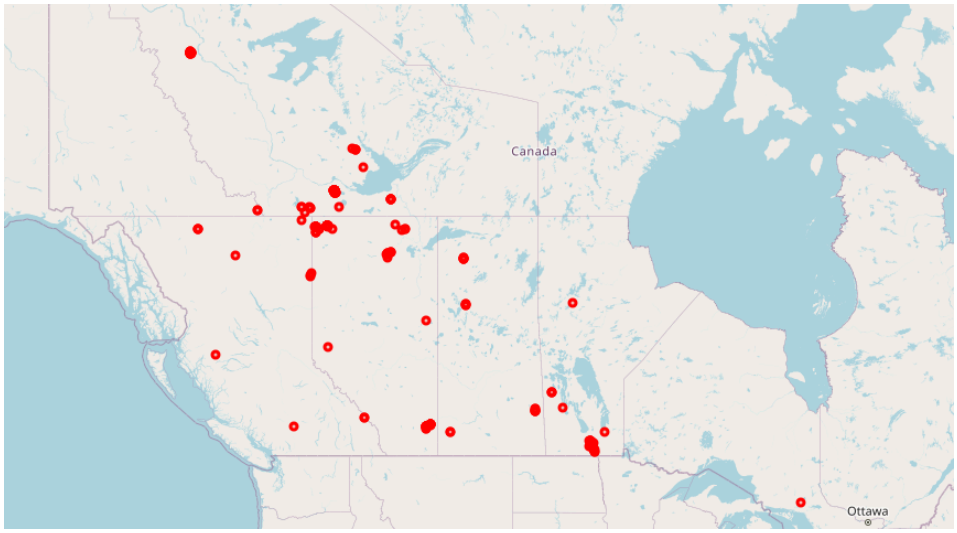


Wildfires with High Confidence mapped per Week - 2012

Week 05



Week 34



Week 51

Dataset used to train models

	Unnamed: 0	latitude	longitude	acq_date	confidence	geometry_fire	LATITUDE	LONGITUDE	DATE	PRCP	TMIN	TMAX	geometry_weather	distance	week_of_year	grid_label
0	0	58.728439	-117.236771	2012-01-22	h	POINT (-117.236771 58.728439)	53.0667	-121.5167	2012-01-22	0.12	3.0	27.0	POINT (-121.5167 53.0667)	7.097400	3	25
1	1	57.025085	-121.934845	2012-02-01	h	POINT (-121.934845 57.025085)	53.0333	-122.5167	2012-02-01	0.08	21.0	38.0	POINT (-122.5167 53.0333)	4.033969	5	19
2	2	53.131313	-115.990082	2012-02-04	h	POINT (-115.990082 53.131313)	51.5833	-119.7833	2012-02-04	0.00	16.0	32.0	POINT (-119.7833 51.5833)	4.096931	5	15
3	3	50.940090	-114.169312	2012-02-08	h	POINT (-114.169312 50.94009)	51.5833	-119.7833	2012-02-08	0.00	9.0	30.0	POINT (-119.7833 51.5833)	5.650715	6	10
4	4	58.049400	-114.157494	2012-02-12	h	POINT (-114.157494 58.0494)	51.5833	-119.7833	2012-02-12	0.08	32.0	45.0	POINT (-119.7833 51.5833)	8.570889	6	25

WEEKLY PROGRESS and CHALLENGES

Week 0 - Project Inception:

- Team formation and background research on potential project topics.
- Decision to focus on Canadian wildfires, given the severity of the issue.
- Identification of NASA's VIIRS and MODIS data as a valuable resource.
- Note : At this point we were not sure of how or where to get the weather data from and use it along with the fire data that wee were getting from NASA.

Week 1-2 - Project Kickoff:

- Initial meeting with Elahe to outline project goals and objectives.
- Literature review, discovering a relevant research repository using XGBoost for wildfire prediction.
- Exploration of VIIRS data and its organization by NASA.
- Efforts of finding the weather data were on, it was discovered that Historical weather data from
- Environment canada had to be downloaded for separate station and for each month separately, and there was no way to download the entire dataset in one go, it was decided to keep looking for more datasets

Week 3 - Dataset Exploration:

- Search for weather datasets to use as inputs for the predictive model was on and NOAA website was discovered, where the data could be downloaded for all the stations across Canada, but due to download restrictions, the entire data could not be downloaded in one go, the entire data file was 47.3 GB, we had to download the data for each month separately for the year 2012 to 2020.
- Identification of input and output variables for the machine learning models were discussed during the meeting with Elahe for the Fire Dataset, but the Weather dataset still had to be stitched.

Week 4 - Work Division:

- Division of tasks among team members.
- Dewang focused on visual representation of VIIRS data.
- Vrinda delved into XGBoost, Random Forest, and relevant research papers.
- Jasjeet collected weather data, faced challenges of stitching the Weather Data from NOAA website together, that was done using python and pandas.

Week 5 - Data Processing:

- It was discovered that data in the NOAA historical weather dataset had some important missing fields like the TMIN, TMAX, Latitude and Longitude.
- A data cleaning process was formed to delete the data, missing the fields mentioned above.

```python
import os
import pandas as pd

# Define the folder path containing the CSV files
folder_path = "/Users/jasjeetsingh/Desktop/Borealis/data/2020"

# List all CSV files in the folder
csv_files = [file for file in os.listdir(folder_path) if file.endswith(".csv")]

# Initialize an empty list to store DataFrames
dfs = []

# Loop through each CSV file, read, and clean it
for file in csv_files:
    file_path = os.path.join(folder_path, file)
    df = pd.read_csv(file_path)

    # Filter rows with missing values in specified columns
    columns_to_check = ["LATITUDE", "LONGITUDE", "TMAX", "TMIN"]
    df_cleaned = df.dropna(subset=columns_to_check)

    # Append the cleaned DataFrame to the list
    dfs.append(df_cleaned)

# Concatenate all DataFrames into one
combined_df = pd.concat(dfs, ignore_index=True)

# Save the combined data to a new CSV file
combined_df.to_csv("2020.csv", index=False)
```

- Also since the dataset was too big to be processed in one go, we made a smaller working model for the year 2012 and limited our datastations to 25-30 per province only for BC, Alberta and Saskatchewan to see if the model works and the results are fruitful.
- Initial implementation of XGBoost on a subset of 50 wildfires, achieving 80% accuracy.

Week 6 - Model Refinement:

- Increased the number of wildfires to 3500 for prediction.
- Employed Random Forest and Multilayer Perceptron Neural Network models.
- Implemented cross-validation for each model, with XGBoost achieving the highest accuracy.
- Requested permission to use SKYNET for processing large datasets.
- Progress documented and updated.

Models :

To get the best possible outcomes, tests were run on 3 different models, to figure out the best one out of them all, accuracy for all 3 of them was compared for the year 2012, and once we have the best out of them all, we will further choose one model and train the model on that.

Random Forest: An ensemble method that builds multiple decision trees and merges them together to get a more accurate and stable prediction. It's effective for both classification and regression tasks and known for handling overfitting well.

XGBoost: Stands for eXtreme Gradient Boosting, it's an efficient implementation of gradient boosting for classification and regression. It's known for its performance and speed and is widely used in machine learning competitions.

Multilayer Perceptron (MLP): A type of artificial neural network with multiple layers. It consists of an input layer, hidden layers, and an output layer, using non-linear activation functions. MLP is suitable for complex pattern recognition and classification problems.

We used the wildfires with only high confidence so that our model is predicting based on the most accurate data available. 25-30 weather stations were identified for each region. Then the data sets were merged, adding the closest weather station data to the wildfires.

Each region was divided into 25 Grid Labels that the models would use to classify the predicted wildfires into. This allows us to lessen the computing power required while still providing a prediction that is useful.

Findings :

After performing classification using all 3 models, it was observed that XGBoost is performing better than the other two.

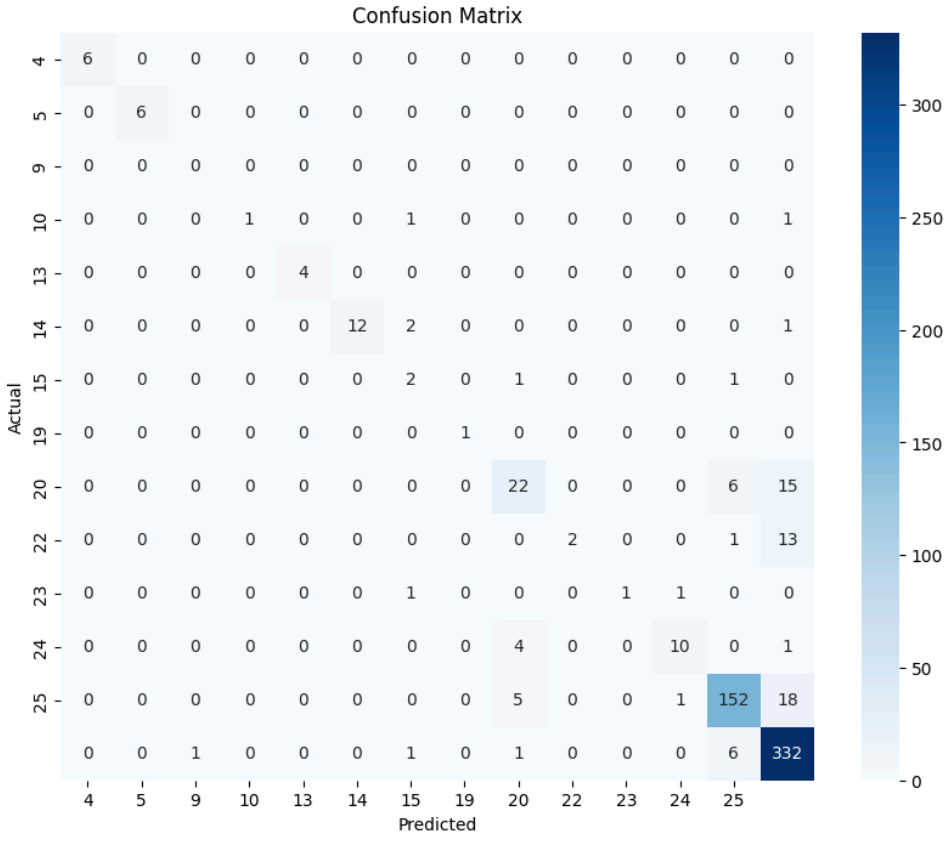
Models were trained using data from 2012 for British Columbia, Alberta and Saskatchewan for the prototype due to constraints with computing power. Here are the results -

British Columbia -

XG Boost -

```
Accuracy: 0.8704581358609794
Cross-validated scores: [0.85573123 0.84189723 0.85573123    nan    nan]
Average Score: nan
```

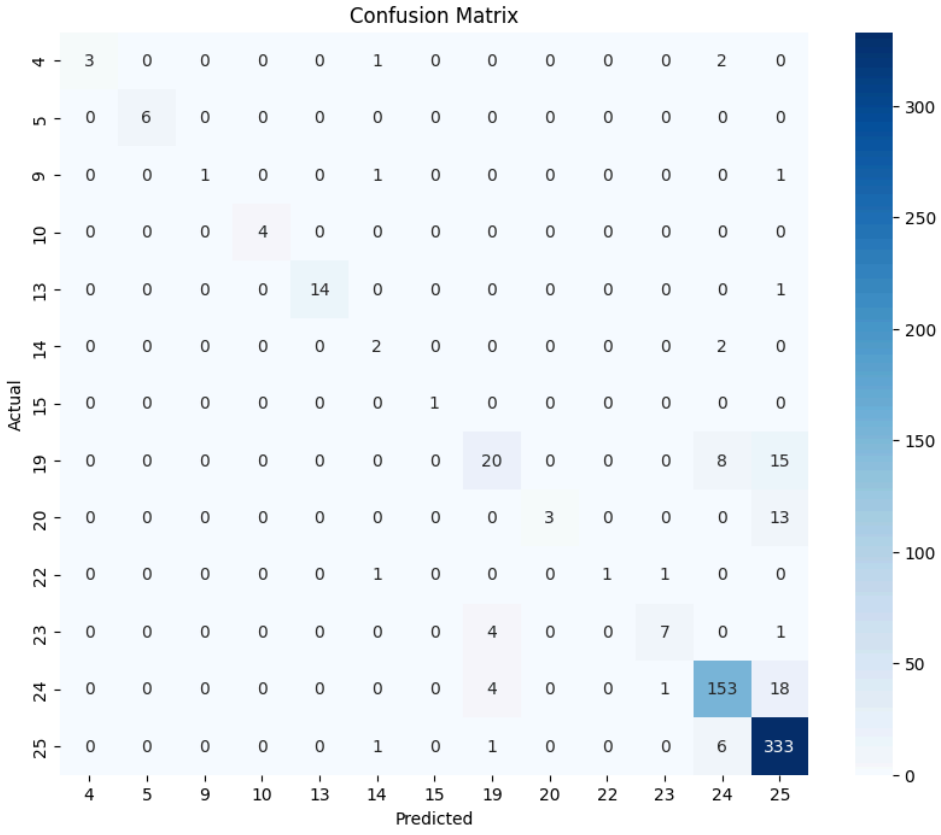
	precision	recall	f1-score	support
4	1.00	1.00	1.00	6
5	1.00	1.00	1.00	6
8	0.00	0.00	0.00	0
9	1.00	0.33	0.50	3
10	1.00	1.00	1.00	4
13	1.00	0.80	0.89	15
14	0.29	0.50	0.36	4
15	1.00	1.00	1.00	1
19	0.67	0.51	0.58	43
20	1.00	0.12	0.22	16
22	1.00	0.33	0.50	3
23	0.83	0.67	0.74	15
24	0.92	0.86	0.89	176
25	0.87	0.97	0.92	341
accuracy			0.87	633
macro avg	0.83	0.65	0.69	633
weighted avg	0.88	0.87	0.86	633



Random Forest-

```
Accuracy: 0.8698412698412699
Cross-validated scores: [0.85940594 0.83366337 0.85714286 0.84126984
0.86309524]
Average Score: 0.850915448687726
```

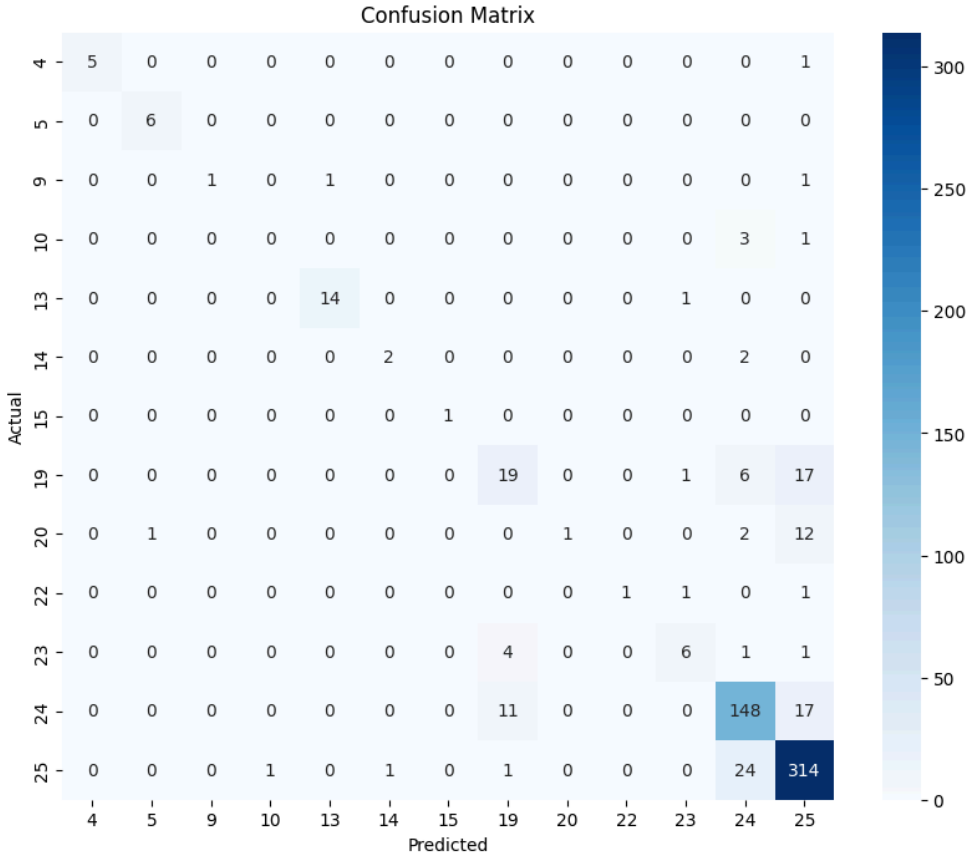
	precision	recall	f1-score	support
4	1.00	0.50	0.67	6
5	1.00	1.00	1.00	6
9	1.00	0.33	0.50	3
10	1.00	1.00	1.00	4
13	1.00	0.93	0.97	15
14	0.33	0.50	0.40	4
15	1.00	1.00	1.00	1
19	0.69	0.47	0.56	43
20	1.00	0.19	0.32	16
22	1.00	0.33	0.50	3
23	0.78	0.58	0.67	12
24	0.89	0.87	0.88	176
25	0.87	0.98	0.92	341
accuracy			0.87	630
macro avg	0.89	0.67	0.72	630
weighted avg	0.87	0.87	0.86	630



Multilayer Perceptron -

```
Accuracy: 0.8222222222222222
Cross-validated scores: [0.77425743 0.77425743 0.75   0.74801587 0.75 ]
Average Score: 0.7593061449002043
```

	precision	recall	f1-score	support
4	1.00	0.83	0.91	6
5	0.86	1.00	0.92	6
9	1.00	0.33	0.50	3
10	0.00	0.00	0.00	4
13	0.93	0.93	0.93	15
14	0.67	0.50	0.57	4
15	1.00	1.00	1.00	1
19	0.54	0.44	0.49	43
20	1.00	0.06	0.12	16
22	1.00	0.33	0.50	3
23	0.67	0.50	0.57	12
24	0.80	0.84	0.82	176
25	0.86	0.92	0.89	341
accuracy			0.82	630
macro avg	0.79	0.59	0.63	630
weighted avg	0.82	0.82	0.81	630

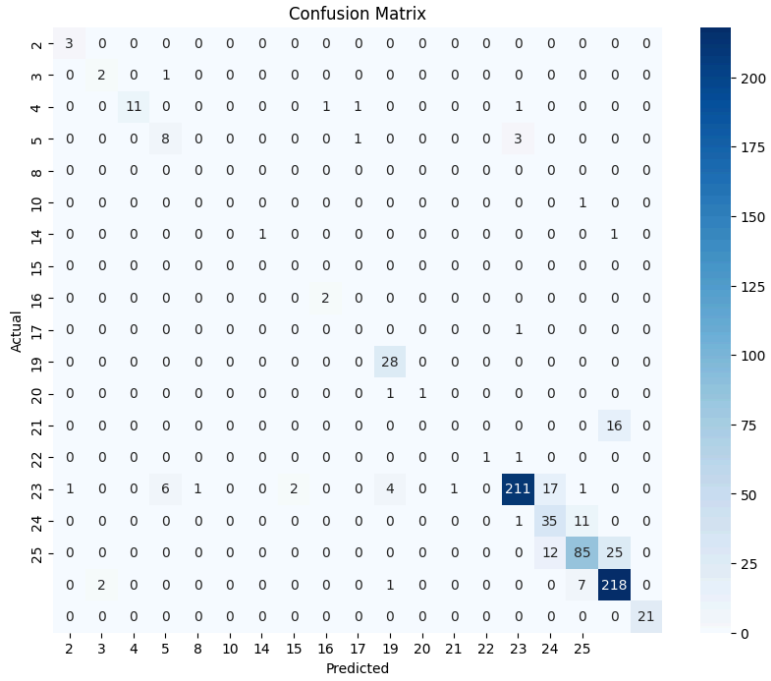


Alberta -

XG Boost -

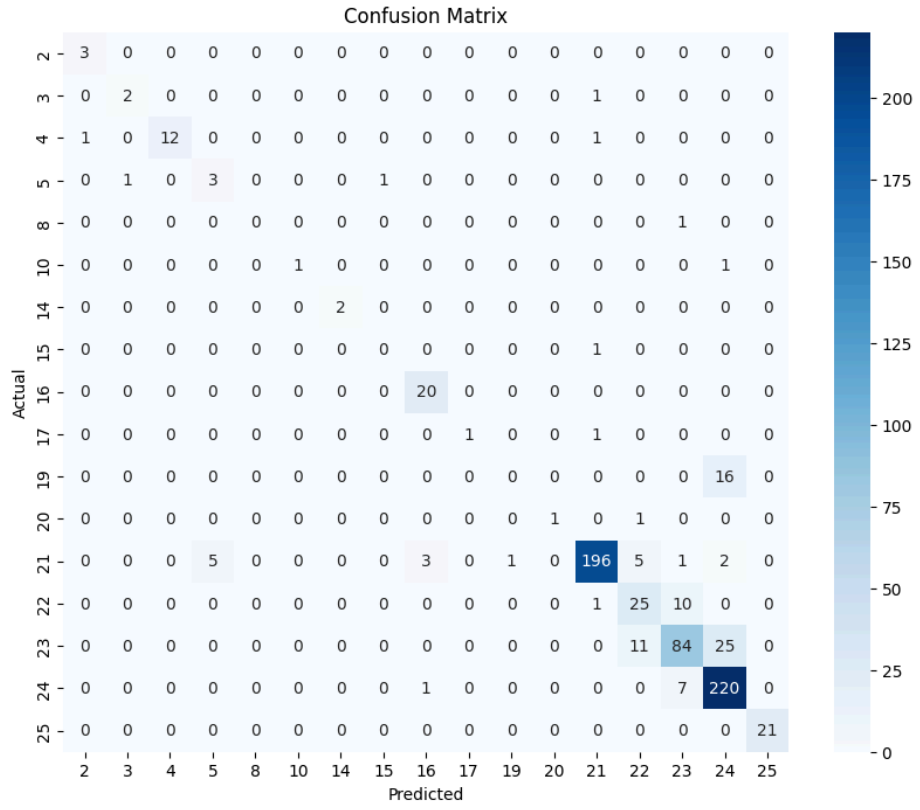
```
Accuracy: 0.8382352941176471
Cross-validated scores: [0.84307179 0.84448161 0.83277592 0.85451505
0.82608696]
Average Score: 0.8401862636166186
```

	precision	recall	f1-score	support
2	0.75	1.00	0.86	3
3	0.50	0.67	0.57	3
4	1.00	0.79	0.88	14
5	0.53	0.67	0.59	12
6	0.00	0.00	0.00	0
8	0.00	0.00	0.00	1
10	1.00	0.50	0.67	2
13	0.00	0.00	0.00	0
14	0.67	1.00	0.80	2
15	0.00	0.00	0.00	1
16	0.82	1.00	0.90	28
17	1.00	0.50	0.67	2
19	0.00	0.00	0.00	16
20	1.00	0.50	0.67	2
21	0.97	0.86	0.91	244
22	0.55	0.74	0.63	47
23	0.81	0.70	0.75	122
24	0.84	0.96	0.89	228
25	1.00	1.00	1.00	21
accuracy			0.84	748
macro avg	0.60	0.57	0.57	748
weighted avg	0.84	0.84	0.83	748



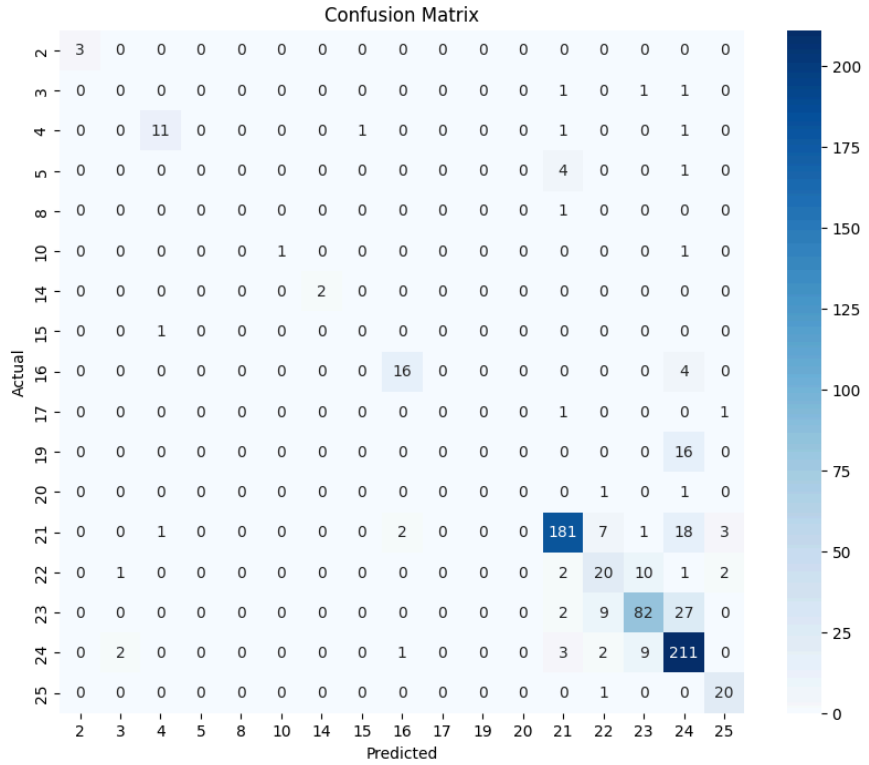
```
Random Forest -
Accuracy: 0.8577648766328012
Cross-validated scores: [0.85321101 0.85294118 0.85294118 0.87132353
0.84926471]
Average Score: 0.8559363194819213
```

	precision	recall	f1-score	support
2	0.75	1.00	0.86	3
3	0.67	0.67	0.67	3
4	1.00	0.86	0.92	14
5	0.38	0.60	0.46	5
8	0.00	0.00	0.00	1
10	1.00	0.50	0.67	2
14	1.00	1.00	1.00	2
15	0.00	0.00	0.00	1
16	0.83	1.00	0.91	20
17	1.00	0.50	0.67	2
19	0.00	0.00	0.00	16
20	1.00	0.50	0.67	2
21	0.98	0.92	0.95	213
22	0.60	0.69	0.64	36
23	0.82	0.70	0.75	120
24	0.83	0.96	0.89	228
25	1.00	1.00	1.00	21
accuracy			0.86	689
macro avg	0.70	0.64	0.65	689
weighted avg	0.85	0.86	0.85	689



Multilayer Perceptron -
Accuracy: 0.7939042089985486
Cross-validated scores: [0.7559633 0.76102941 0.79595588 0.78676471
0.72794118]
Average Score: 0.7655308958445763

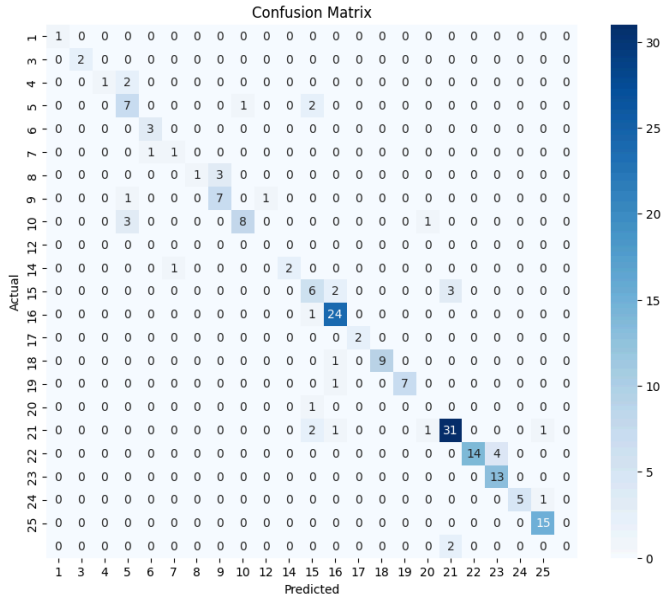
	precision	recall	f1-score	support
2	1.00	1.00	1.00	3
3	0.00	0.00	0.00	3
4	0.85	0.79	0.81	14
5	0.00	0.00	0.00	5
8	0.00	0.00	0.00	1
10	1.00	0.50	0.67	2
14	1.00	1.00	1.00	2
15	0.00	0.00	0.00	1
16	0.84	0.80	0.82	20
17	0.00	0.00	0.00	2
19	0.00	0.00	0.00	16
20	0.00	0.00	0.00	2
21	0.92	0.85	0.89	213
22	0.50	0.56	0.53	36
23	0.80	0.68	0.74	120
24	0.75	0.93	0.83	228
25	0.77	0.95	0.85	21
accuracy			0.79	689
macro avg	0.50	0.47	0.48	689
weighted avg	0.77	0.79	0.78	689



Saskatchewan - XGBoost -

```
Accuracy: 0.8112244897959183
Cross-validated scores: [0.7388535  0.75796178 0.80128205 0.80769231
0.75641026]
Average Score: 0.7724399804017639
```

	precision	recall	f1-score	support
1	1.00	1.00	1.00	1
3	1.00	1.00	1.00	2
4	1.00	0.33	0.50	3
5	0.54	0.70	0.61	10
6	0.75	1.00	0.86	3
7	0.50	0.50	0.50	2
8	1.00	0.25	0.40	4
9	0.70	0.78	0.74	9
10	0.89	0.67	0.76	12
11	0.00	0.00	0.00	0
12	1.00	0.67	0.80	3
14	0.50	0.55	0.52	11
15	0.83	0.96	0.89	25
16	1.00	1.00	1.00	2
17	1.00	0.90	0.95	10
18	1.00	0.88	0.93	8
19	0.00	0.00	0.00	1
20	0.86	0.86	0.86	36
21	1.00	0.78	0.88	18
22	0.76	1.00	0.87	13
23	1.00	0.83	0.91	6
24	0.88	1.00	0.94	15
25	0.00	0.00	0.00	2
...				
accuracy			0.81	196
macro avg	0.75	0.68	0.69	196
weighted avg	0.83	0.81	0.81	196



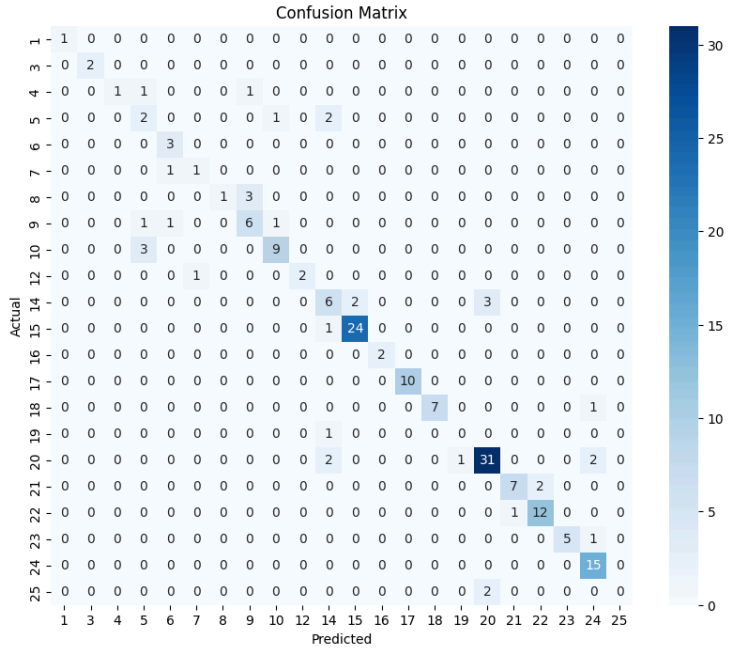
Random Forest -

Accuracy: 0.8076923076923077

Cross-validated scores: [0.74324324 0.74324324 0.79591837 0.79591837 0.75510204]

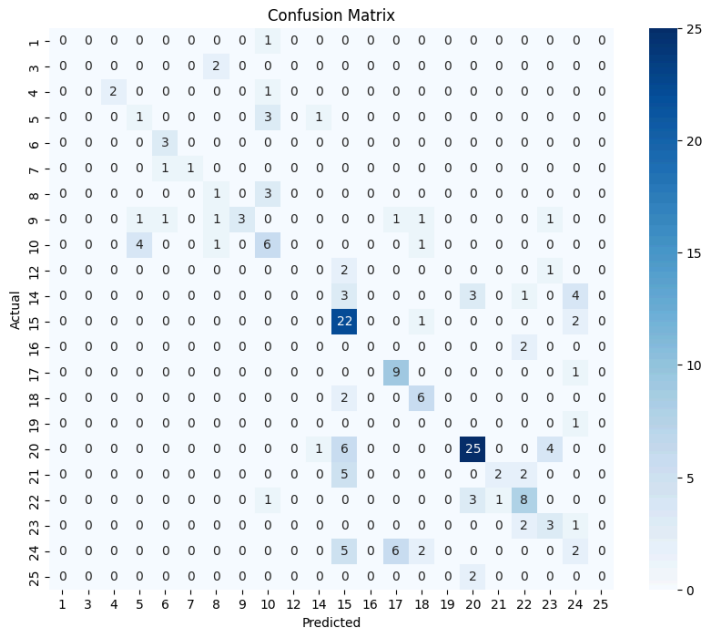
Average Score: 0.766685052399338

	precision	recall	f1-score	support
1	1.00	1.00	1.00	1
3	1.00	1.00	1.00	2
4	1.00	0.33	0.50	3
5	0.29	0.40	0.33	5
6	0.60	1.00	0.75	3
7	0.50	0.50	0.50	2
8	1.00	0.25	0.40	4
9	0.60	0.67	0.63	9
10	0.82	0.75	0.78	12
12	1.00	0.67	0.80	3
14	0.50	0.55	0.52	11
15	0.92	0.96	0.94	25
16	1.00	1.00	1.00	2
17	1.00	1.00	1.00	10
18	1.00	0.88	0.93	8
19	0.00	0.00	0.00	1
20	0.86	0.86	0.86	36
21	0.88	0.78	0.82	9
22	0.86	0.92	0.89	13
23	1.00	0.83	0.91	6
24	0.79	1.00	0.88	15
25	0.00	0.00	0.00	2
accuracy			0.81	182
macro avg	0.75	0.70	0.70	182
weighted avg	0.82	0.81	0.80	182



Multilayer Perceptron -
Accuracy: 0.5164835164835165
Cross-validated scores: [0.58108108 0.54054054 0.51020408 0.51020408
0.52380952]
Average Score: 0.5331678617392902

	precision	recall	f1-score	support
1	0.00	0.00	0.00	1
3	0.00	0.00	0.00	2
4	1.00	0.67	0.80	3
5	0.17	0.20	0.18	5
6	0.60	1.00	0.75	3
7	1.00	0.50	0.67	2
8	0.20	0.25	0.22	4
9	1.00	0.33	0.50	9
10	0.40	0.50	0.44	12
12	0.00	0.00	0.00	3
14	0.00	0.00	0.00	11
15	0.49	0.88	0.63	25
16	0.00	0.00	0.00	2
17	0.56	0.90	0.69	10
18	0.55	0.75	0.63	8
19	0.00	0.00	0.00	1
20	0.76	0.69	0.72	36
21	0.67	0.22	0.33	9
22	0.53	0.62	0.57	13
23	0.33	0.50	0.40	6
24	0.18	0.13	0.15	15
25	0.00	0.00	0.00	2
accuracy			0.52	182
macro avg	0.38	0.37	0.35	182
weighted avg	0.49	0.52	0.48	182



We also attempted to fine tune the XGboost model using Grid Search.

```
param_grid = {
    'n_estimators': [100, 200, 300],
    'max_depth': [3, 5, 7],
    'learning_rate': [0.01, 0.1, 0.2],
    'subsample': [0.7, 0.8, 0.9],
    'colsample_bytree': [0.7, 0.8, 0.9]
}

# Initialize GridSearchCV
grid_search = GridSearchCV(
    estimator=model,
    param_grid=param_grid,
    cv=5,  # 5-fold cross-validation
    scoring='accuracy',
    verbose=1
)
```

It was found that in 2 out of 3 cases, XGBoost beat Random Forest and Multilayer Perceptron. Random Forest did perform better for data based on Alberta. Multilayer Perceptron performed considerably worse in all 3 regions.

Although it should be noted that XGboost was trained and tested on data that still had some missing fields. The rows containing missing fields were dropped for the other 2 models.

It can be seen that since some of the classes only had small occurrences, while dividing the dataset into training and validation sets for cross validation some fits failed due to the small frequency of those classes such that they were either missing from the training set or the validation set. We tried StratifiedKFold i.e. proportional representation of classes in training and validation sets to tackle this situation but that did not solve the issue, the frequency was too less.

Conclusion :

XGboost was able to predict the grid label for wildfires based on weather data with around 80-85% accuracy. Random Forest was also able to predict with similar accuracy.

Both the models had some confusion as visible in the confusion matrices i.e. they predicted some classes wrong, biased towards a class with more frequency. We suspect this to be the case due the data being heavily leaning towards certain grid labels. This is because some regions are more susceptible to wildfires.

Future Direction:

We would like to continue working on these models, including more data, both regions and years. We have data for all of Canada from 2012 to 2021. Therefore, we would like to increase the training set to all

regions and 7 years and use 3 years as a testing set. We would like to increase the number of Grid labels as well to increase the precision of the location of the predictions. Currently the models are only able to predict where a wildfire is most probable to occur, given the week of the year and the weather data. The models are not able to predict how many and which grid labels could a wildfire occur in. This is due to the lack of weather data where wildfires do not occur. We would like to work on this. We would also like to see whether a regression model can be used to predict precise wildfire locations.

References :

D. Stojanova, P. Panov, A. Kobler, S. Džeroski, and K. Taškova, “Learning to predict wildfires,” Knowl. Creat. Diffus. Util., vol. 9, no. 14, pp. 255–258, 2006. https://www.researchgate.net/publication/228527438_Learning_to_predict_forest_fires_with_different_data_mining_techniques

Murali, Harsh & Yao. (2019). Predicting-Wild-Fire-with-Weather-Data-in-the-US-Geography https://github.com/muralits98/Predicting-Wild-Fire-with-Weather-Data-in-the-US-Geography

Pérez-Porras, F.-J., Triviño-Tarradas, P., Cima-Rodríguez, C., Meroño-de-Larriva, J.-E., García-Ferrer, A., & Mesas-Carrascosa, F.-J. (2021). Machine Learning Methods and Synthetic Data Generation to Predict Large Wildfires. Sensors, 21(11), 3694. MDPI AG. Retrieved from http://dx.doi.org/10.3390/s21113694