**PROJECT REPORT**

**Introduction**

Wildfires are a growing concern in Alberta, Canada, with devastating impacts on the environment, economy, and human life. The unpredictable nature of wildfires makes it challenging to prepare for and respond to these events.

One of the most notable effects on the economy is that the region around Fort McMurray is rich in natural resources, including oil and gas, and the wildfires can disrupt the extraction and production of these resources. The wildfires also have impact on the timber industry, as large of fires are burened affecting timber values and livelihood of people dependent on it. Also wildfires fires in Alberta also affect human health. Massive fires force thousands of people to displace and the air quality can become hazardous.

So, such kind of projects become important for the following reasons:

1. Forest management : These type of project hels to predict the wildfires in advance which allow the forest department to develop effective strategies to tackle it.

2. Minimizing the health risks : Analyzing the aftereffects of wildfires on air-quality, respiratory health can help in controlling the health of the communities living there.

Methodology

We are provided with the real life data of wildfires in the Alberta region of Canada from the year 2006 to the year 2021. Since the data is huge we need to preprocess it before we draw any conclusions from it. Firstly we drop the columns which have nothing to do with the severity or frequency of the wildfires, like ‘fire\_number’ , ‘fire\_name’ , ‘dicovered\_size’ .

Then after dealing with some null values, we will merge some columns and store only the necessary value in a new column. For example using the ‘pd.to\_datetime’ function we will take the time difference between the ‘fire\_start\_date’ and ‘reported\_date’ and create a new column named ‘reporting\_time’. Similarly with the ‘uc\_fs\_date’ and ‘ex\_fs\_date’.

For deling with null values in a column we replace the null values with the mean of the column.

Now we have to deal with the latitude and longitude, first we drop the invalid values of lat and long, then calculate the distance from Alberta using the Haversine formula.

Then using LabelEncoder, we encoded the ‘size\_class’ column. We have to reduce the dimensionality of the columns for that we use Z-scoring. And then encode all the object columns and apply K-means clustering in numbers so that they can be used in the machine learning model.

After this applied DecisionTreeRegressor model to estimate the current size

Question 1:

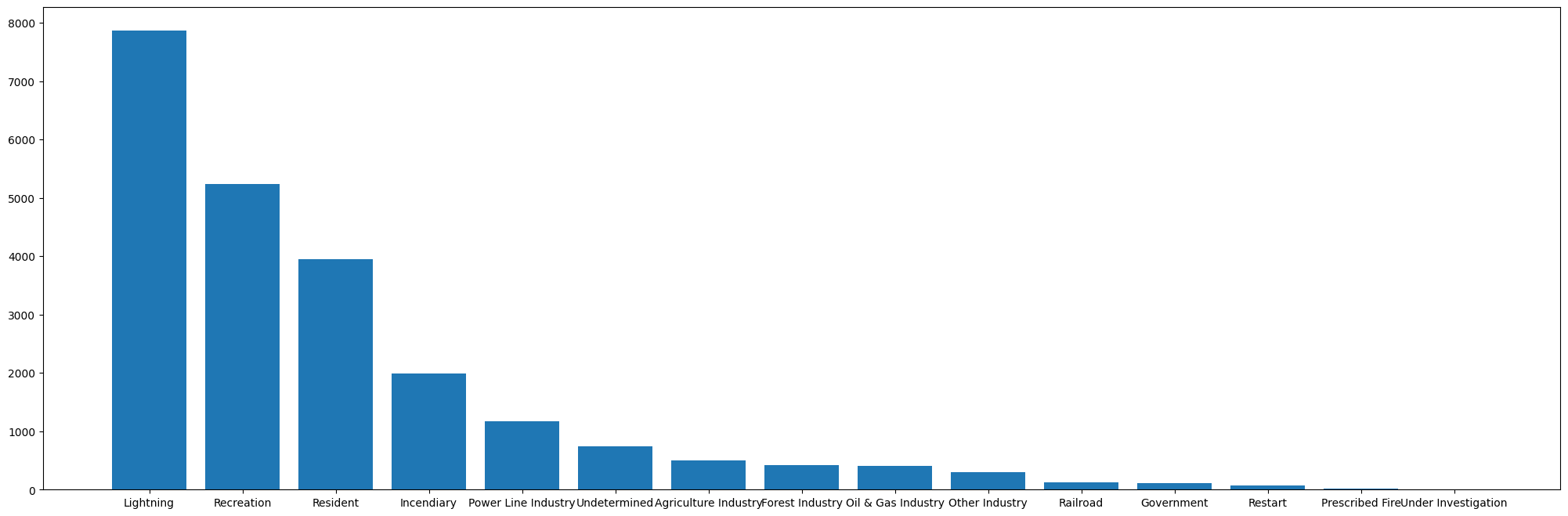
For the first problem, current size is responsible for vulnerabilities of FSA region i.e, if current size of a FSA region is more then that region is more vulnerable. To obtain the FSA region, I searched it in the postal Canada file from statistics, Canada and map these regions using Geopandas. I found out **TOH** and **TOP** are most vulnerable FSA region.

Question 2

For the second question, **Lightning** is main reason which usually cause wildfire near each vulnerable FSA region. To determine this I looked on the ‘general\_cause\_desc’. I obtained the value count of ‘general\_cause\_desc’ which result in the maximum value of lightning. Also **Recreation , Resident** are probable reasons which cause maximum burn sizes. Also after studying the **correlation matrix** we can see that features like **‘fire\_speed\_rate’,‘temperature ’,’relative humidity’ ‘wind speed’** cause fires with large burn areas.

Wildfires have a huge impact on the environment as it destroys vegetation and affects the air quality. Also wildfires can impact the economy as well as the residents nearby discussed in the second paragraph of the report

Wildfires have impact on the environment as it destroys vegetation and affects air quality. Also, Wildfire in these regions impacts on resident nearby because it spreads fastly causing threats to communities living there.



Question 4

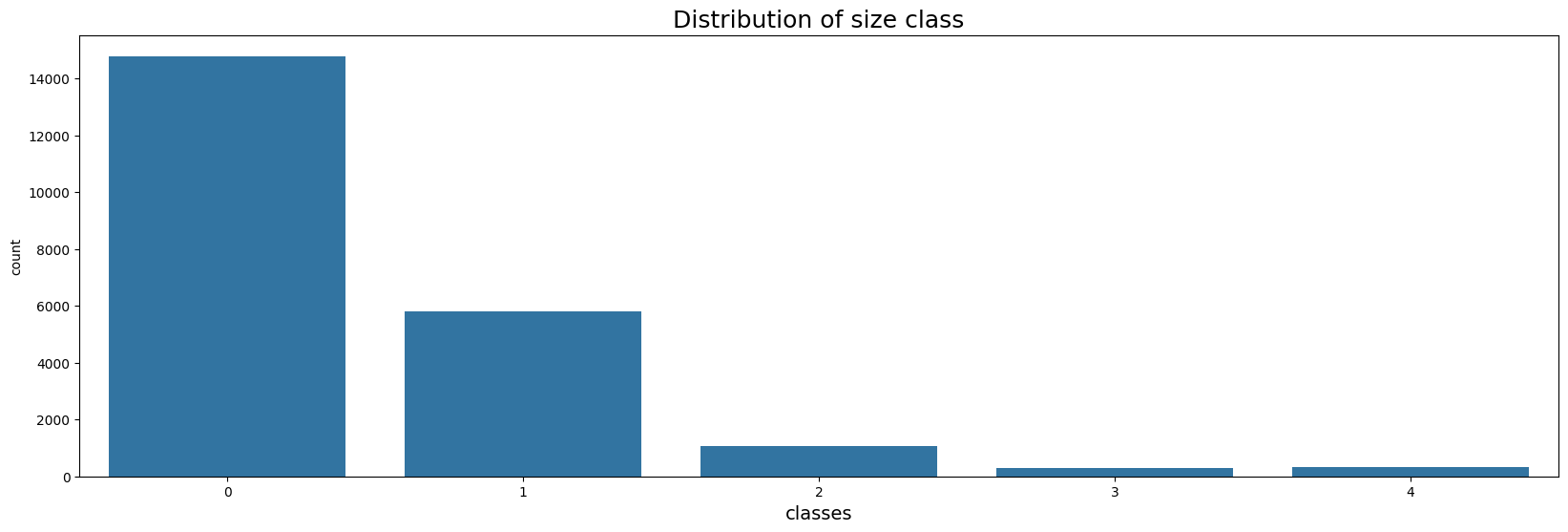
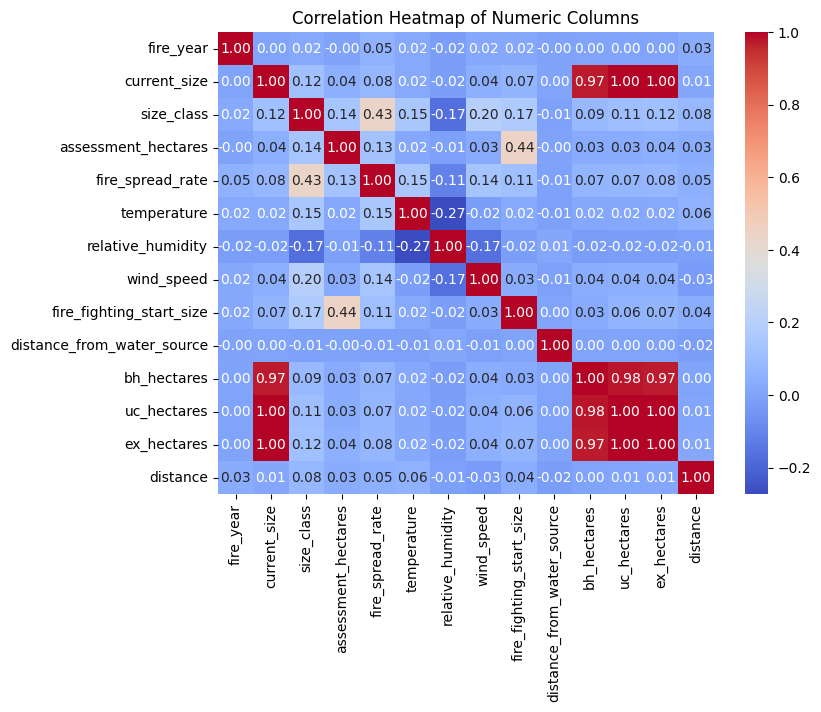
For question 4, I prepared and trained a model to predict the final size (current size) of fires. To do this I trained a DecisionTreeRegressor model the code for which is attached in jupyter notebbok.

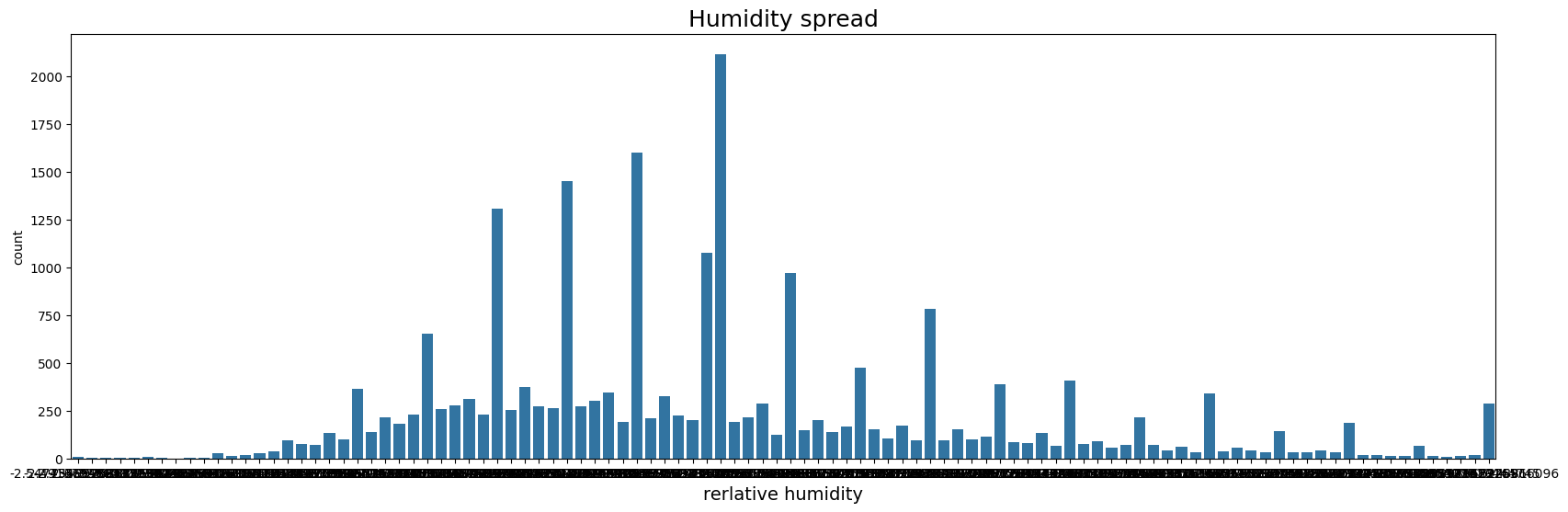
I got r2\_score = 0.9367310603498255

and RMSE = 1.115326147271815 both on the train data.

With the help of correlation matrix I was able to analyse the features which affected

the burn size of the wildfire.





Reference:

[**https://www12.statcan.gc.ca/census-recensement/2021/geo/sip-pis/boundary-limites/index2021-eng.cfm?year=21**](https://www12.statcan.gc.ca/census-recensement/2021/geo/sip-pis/boundary-limites/index2021-eng.cfm?year=21)

[**https://geopandas.org/en/stable/**](https://geopandas.org/en/stable/)

[**https://snyk.io/advisor/python/shapely/functions/shapely.geometry.Point**](https://snyk.io/advisor/python/shapely/functions/shapely.geometry.Point)

[**https://pypi.org/project/shapely/**](https://pypi.org/project/shapely/)