Lab 2 Summary of Discussion report

DMLL Group 3 Lab 2 Report
Date: 29-09-22

Members:-
Gowresh
Yehia
Sheetal
Gunjan
Nasir

Summary of Discussion:-

Attributes and its features

There are totally 15 attributes in the dataset.

The **latitude and longitude** attributes refer to the geographic location of the hotspots detected by the satellites.

The **brightness** attribute is the pixel measurement of the temperature in Kelvin.

Scan and Track attributes refers to the pixel measurement of the scanned area.

acq_date refers to the date of acquisition of data while acq_time refers to the time spent during the acquisition of data.

Satellite attribute refers to which satellite was used for acquiring the data while **instrument** refers to the instrument used by the satellites for measurment.

Confidence refers to the quality of the hostpot in the specified region of interest. This parameter can be used to predict whether an hostspot is a fired area or not.

Confidence will be our target feature (Values greater than 60% or 70% ususally mean the hotspot is actually a firespot).

Version attribute refers to the collection of source.

bright_t31 is the pixel measurement of temperature in the 3-1 Channel.

frp attribute depicts the pixel-integrated fire radiative power in MW (Mega Watts).

Daynight attribute refers to when the instance were obtained, whether day or night.

Finally the **type** attribute refers to the type of environment of the instance.

Null Values and Data Cleaning

There are no missing values in the dataset. This will probably help us in accurately predicting the target feature. However, in order to utilise the Data Cleaning technique, we inserted five null values in the acq_time attribute. Also, we are working on the Data cleaning process by Scikit-learn.

The **value_count()** displays the number of values in each category of an attribute across the dataset.

Here there are 28,203 values indicating that the respective measurements were taken during daytime while 7,808 values indicating that those measurements were taken during night time. The satellite dataset has two categories indicating the model of satellite that was used for measurment.

Analysis through Histogram

The purpose of a histogram (Chambers) is **to graphically summarize the distribution of a data set**. As the histogram was observed with the concerned dataset, **we analysed the presence of outliers in the attribute Confidence.** It is observed that the **instrument and type attribute have only one category across the entire dataset, which clearly indicates that the attribute can dropped from the dataset while processing as it is not going to affect the data.**The statistical summary and the histograms do not consider the categorical data or more precisely they do not consider the object type of data.

While the acq_date is not a categorical data, we have to consider the activity of changing the data type. The first plot does not help us in finding any particular pattern. The second plot is comparitively better in representing the density of areas where the measurement were taken. The third plot is more indicative as we can understand the different confidence levels of hotspot across the areas.

Correlation between attributes

The most **correlated attributes with the target attribute confidence are the brightness and the frp such that they have strong positive correlation.** According to the definitions of these two attributes, it makes sense that **the brightness, which is the temperature of hotspot region, and the frp, which is the radiative power of the hotspot, can affect the confidence of the hotspot more than the other features. The scan, track and acq_time attributes have negative correlation with the confidence attribute.** This may be due to the reason that the more amount of time the satellite spends on a particular region of interest, the less likely it can be that that region has low confidence of hotspot. Due to the upward trend, it can be concluded that the correlation is very strong. There are straight horizontal lines at around 100, 90, 85, 75 and even around 0. This needs to be considered in order to prevent data quirks.

We have represented the categorical values in numbers. Generally in ML, there may be a possibility that nearby values are more similar than distant values. However as far as this dataset is considered the categorical values have only binary classification.