Forest Fire In Brazil

Open Programme : ADS A

Daviano Almeyda Kustaryanto – 3212610

Contents

Introduction

Nowadays, Data has become the crucial part in professional world. Many company used data in order to make decision in the future therefore they won’t be making any wrong decision. Data also can be used to predict something is likely to happen in the future. Based on this statement, we believe that data can also predict what accident will be happen in the future, since the global warming threat is becoming more dangerous and we need to do everything to prevent their impact growing on Earth.   This Prediction is intended to prevent any forest fire in Brazil since, Brazil is one of many countries that have large number of rainforest and Brazil also part of the Earth’s Lungs country because of their rainforest.

This Document is intended to explain about how we cleaned the data, method that we are using, The analysis of the data and the result of the data.

Assumption

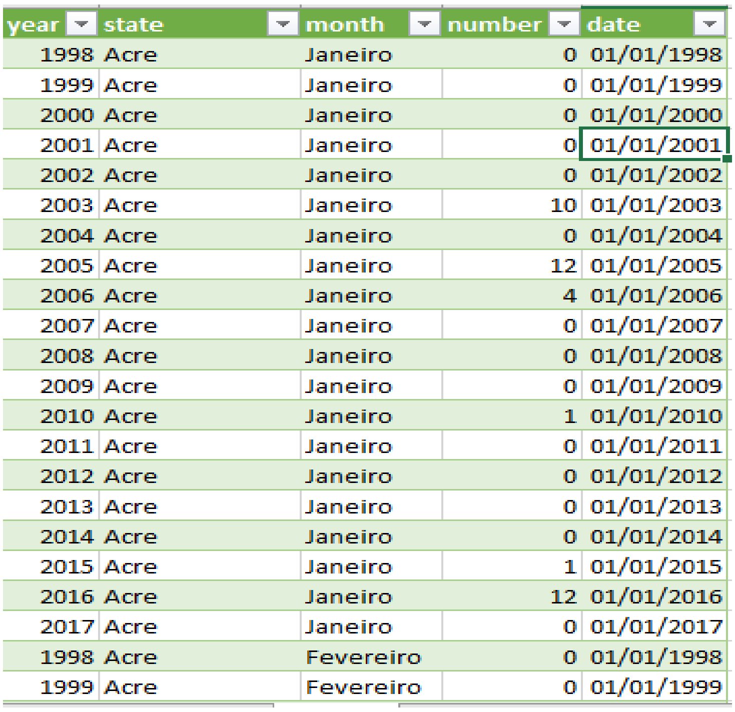
First, we are going to assume that the data that we get from Kaggle is from an official data and we get the data from Brazil’s Government itself. Otherwise, this entire project will be illegal since we’re using data from a third-party that the source of the data is not very credible or possible a stolen data, hacked data and etc. We make an assumption that data that we get from Kaggle is legal and free to use since for us Data Ethics and Laws are very important.

Second, we are going to assume that the data that we get from Kaggle is based on real life situation from Brazil’s Government. Therefore, if there any mis value from the data that we get from Kaggle and the official data from Brazil’s Government, it’s not our fault.

Third, We are assuming that any result that we get from machine learning that we made, our clients will consider to prevent it .

Data

We assumed that the dataset is originally from Brazil's Government that we retrieved from Kaggle. The dataset contains the Year, State Name, Month is Portuguese, number of forest that caught in fire and the date of the accident. The Rough data will look like the picture Below :

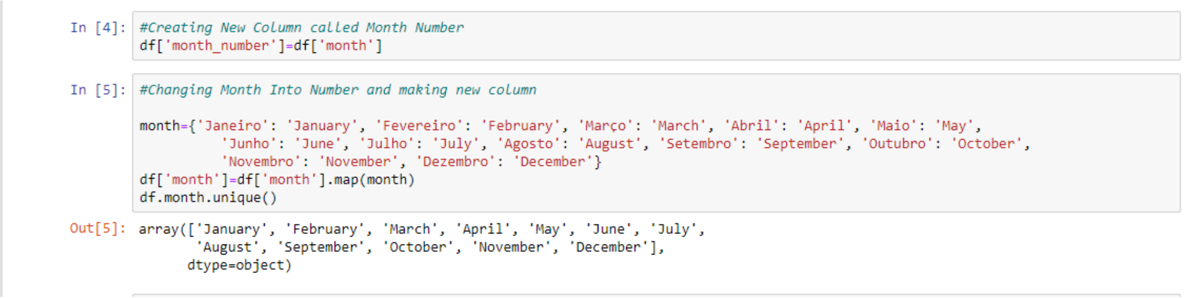


year	state	month	number	date
1998	Acre	Janeiro	0	01/01/1998
1999	Acre	Janeiro	0	01/01/1999
2000	Acre	Janeiro	0	01/01/2000
2001	Acre	Janeiro	0	01/01/2001
2002	Acre	Janeiro	0	01/01/2002
2003	Acre	Janeiro	10	01/01/2003
2004	Acre	Janeiro	0	01/01/2004
2005	Acre	Janeiro	12	01/01/2005
2006	Acre	Janeiro	4	01/01/2006
2007	Acre	Janeiro	0	01/01/2007
2008	Acre	Janeiro	0	01/01/2008
2009	Acre	Janeiro	0	01/01/2009
2010	Acre	Janeiro	1	01/01/2010
2011	Acre	Janeiro	0	01/01/2011
2012	Acre	Janeiro	0	01/01/2012
2013	Acre	Janeiro	0	01/01/2013
2014	Acre	Janeiro	0	01/01/2014
2015	Acre	Janeiro	1	01/01/2015
2016	Acre	Janeiro	12	01/01/2016
2017	Acre	Janeiro	0	01/01/2017
1998	Acre	Fevereiro	0	01/01/1998
1999	Acre	Fevereiro	0	01/01/1999

Cleaning Data

In this section we're going to explain how we cleaned the dataset.

Creating New Column called Month Number and Change the language



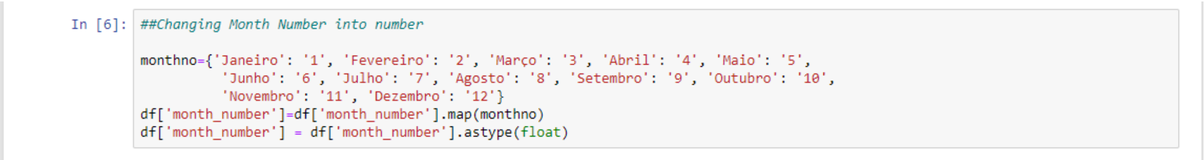
```
In [4]: #Creating New Column called Month Number
        df['month_number']=df['month']

In [5]: #Changing Month Into Number and making new column

        month={'Janeiro': 'January', 'Fevereiro': 'February', 'Março': 'March', 'Abril': 'April', 'Maio': 'May',
                 'Junho': 'June', 'Julho': 'July', 'Agosto': 'August', 'Setembro': 'September', 'Outubro': 'October',
                 'Novembro': 'November', 'Dezembro': 'December'}
        df['month']=df['month'].map(month)
        df.month.unique()

Out[5]: array(['January', 'February', 'March', 'April', 'May', 'June', 'July',
               'August', 'September', 'October', 'November', 'December'],
              dtype=object)
```

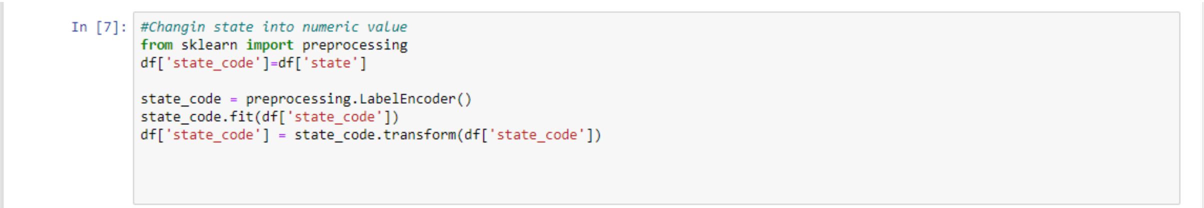
This step, i make a new column called Month number by copying the month column and for the month name i changed to English since the month name from the dataset used Portuguese language



In [6]:
```
##Changing Month Number into number

monthno={'Janeiro': '1', 'Fevereiro': '2', 'Março': '3', 'Abril': '4', 'Maio': '5',
          'Junho': '6', 'Julho': '7', 'Agosto': '8', 'Setembro': '9', 'Outubro': '10',
          'Novembro': '11', 'Dezembro': '12'}
df['month_number']=df['month_number'].map(monthno)
df['month_number'] = df['month_number'].astype(float)
```

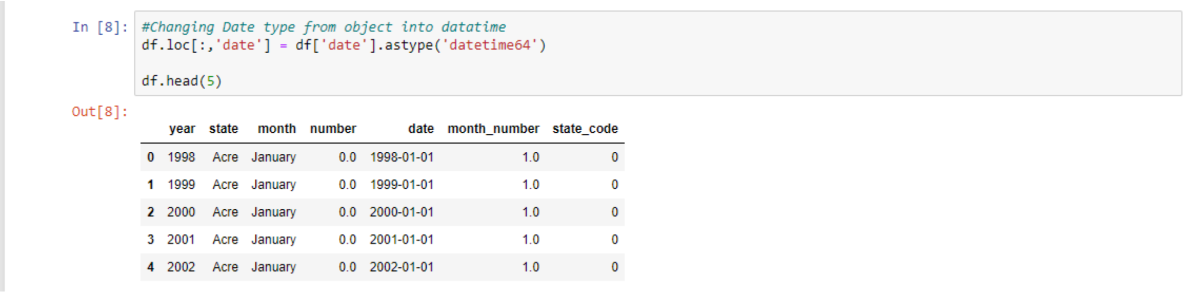
After that, the month number column i need to change the value in numeric and since the datatype is object i need to change into numeric value such as float



In [7]:
```
#Changin state into numeric value
from sklearn import preprocessing
df['state_code']=df['state']

state_code = preprocessing.LabelEncoder()
state_code.fit(df['state_code'])
df['state_code'] = state_code.transform(df['state_code'])
```

Since the state code is string and in order to get more accuracy in predictive performance i need to change it into numeric value. Therefore i make a new column called state code that contain a numeric value for a state.



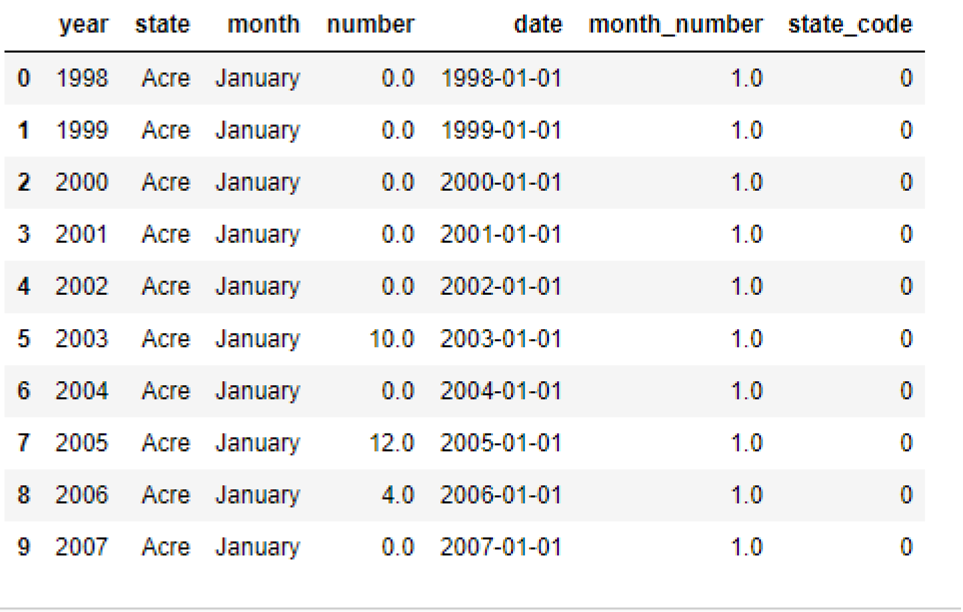
In [8]:
```
#Changing Date type from object into datatime
df.loc[:,'date'] = df['date'].astype('datetime64')

df.head(5)
```

Out[8]:

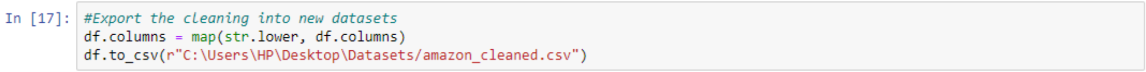
	year	state	month	number	date	month_number	state_code
0	1998	Acre	January	0.0	1998-01-01	1.0	0
1	1999	Acre	January	0.0	1999-01-01	1.0	0
2	2000	Acre	January	0.0	2000-01-01	1.0	0
3	2001	Acre	January	0.0	2001-01-01	1.0	0
4	2002	Acre	January	0.0	2002-01-01	1.0	0

Finally i need to change the date column type into Datetime since when i got it from Kaggle the data type is object.



	year	state	month	number	date	month_number	state_code
0	1998	Acre	January	0.0	1998-01-01	1.0	0
1	1999	Acre	January	0.0	1999-01-01	1.0	0
2	2000	Acre	January	0.0	2000-01-01	1.0	0
3	2001	Acre	January	0.0	2001-01-01	1.0	0
4	2002	Acre	January	0.0	2002-01-01	1.0	0
5	2003	Acre	January	10.0	2003-01-01	1.0	0
6	2004	Acre	January	0.0	2004-01-01	1.0	0
7	2005	Acre	January	12.0	2005-01-01	1.0	0
8	2006	Acre	January	4.0	2006-01-01	1.0	0
9	2007	Acre	January	0.0	2007-01-01	1.0	0

After i finished with cleaning, my dataset will look like this. Compare to the original it have 2 more column that i have explained above.



```
In [17]: #Export the cleaning into new datasets
         df.columns = map(str.lower, df.columns)
         df.to_csv(r"C:\Users\HP\Desktop\Datasets/amazon_cleaned.csv")
```

Finally, i export the new dataset into a new csv file called amazon_cleaned.

Analysis Report

`

In order to complete this project, we are going to used some machine learning modules. We used many Machine Learning model to compare the result between each other. Below is the type of machine learning that we used :

1. KNN Algorithm(K Nearest Neighbours)

The K-Nearest Algorithm is a supervised classification algorithm, It takes a lot marked points and use them to learn how to label another point. The reason we used this technique is this technique is quite common among data scientist, It is very simple to implemented and many major company used this technique to predict Something.

2. Minkowski Model

The minkowski model with the Knn. Minkowski is a method for machine learning where they spot a common divisor of a number, where the largest number can be divided without leaving a remainder. The minkowski is also known as the generalized model of both Euclidean and Manhattan model.

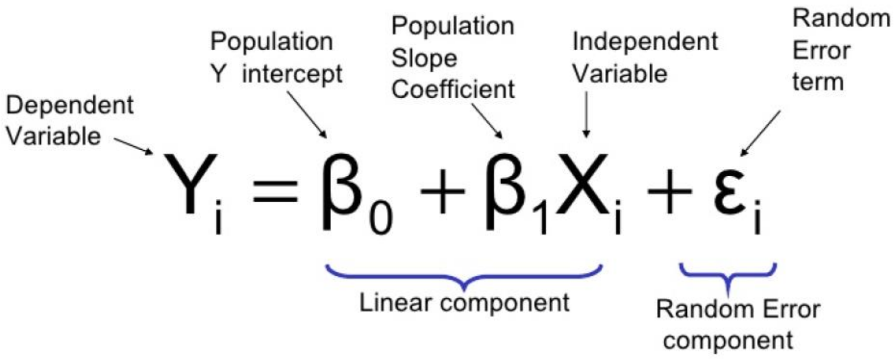
$$D(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{l=1}^{d} |x_{il} - x_{jl}|^{1/p} \right)^{p}.$$

3. Decision Tree Regressor

$$S(T, X) = \sum_{c \in X} P(c) S(c)$$

Decision Tree is a machine learning method where they break down a data set into smaller subset. Decision tree also part of the supervised machine learning category and It also work for continuous (Regression) or Categorical output (Classification). Decision Tree Regressor observe an object and trains the model in the structure of a tree to produce a continuous output.

4. Linear Regression

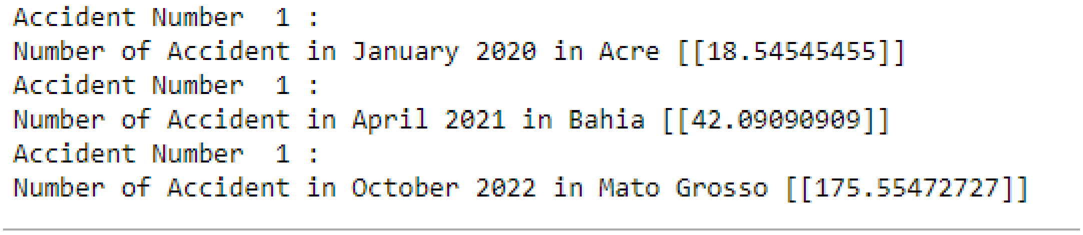

Population Y intercept
Population Slope Coefficient
Independent Variable
Random Error term
Dependent Variable
$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$
Linear component
Random Error component

This model in algebra refers to linear relation ship between two or more variables. Linear Regression Predict a dependent Variable Value (y) based on independent value (x). Therefore this model finds out the linear relationship between x (input) and y (output).

Result

K Nearest Algorithm

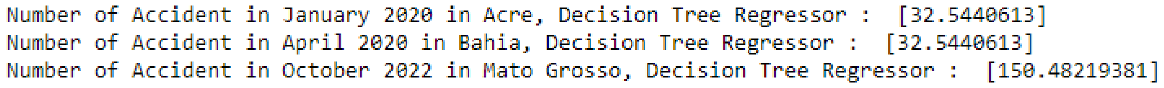
The KNN algorithm was successfully applied in our machine learning. The algorithm show the Prediction of how many Forest will burn in certain location and time. Picture of proof are shown below :



```
Accident Number  1 :
Number of Accident in January 2020 in Acre [[18.54545455]]
Accident Number  1 :
Number of Accident in April 2021 in Bahia [[42.09090909]]
Accident Number  1 :
Number of Accident in October 2022 in Mato Grosso [[175.55472727]]
```

Decision Tree

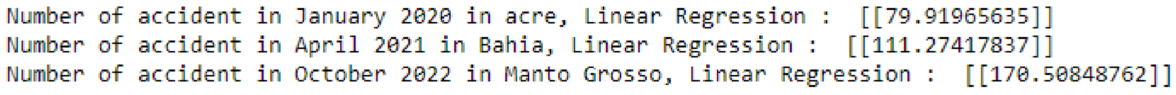
The Decision Tree algorithm was successfully applied in our machine learning. The algorithm show the Prediction of how many Forest will burn in certain location and time. Picture of proof are shown below :



```
Number of Accident in January 2020 in Acre, Decision Tree Regressor :  [32.5440613]
Number of Accident in April 2020 in Bahia, Decision Tree Regressor :  [32.5440613]
Number of Accident in October 2022 in Mato Grosso, Decision Tree Regressor :  [150.48219381]
```

Linear Regression

The Linear Regression algorithm was successfully applied in our machine learning. The algorithm show the Prediction of how many Forest will burn in certain location and time. Picture of proof are shown below :

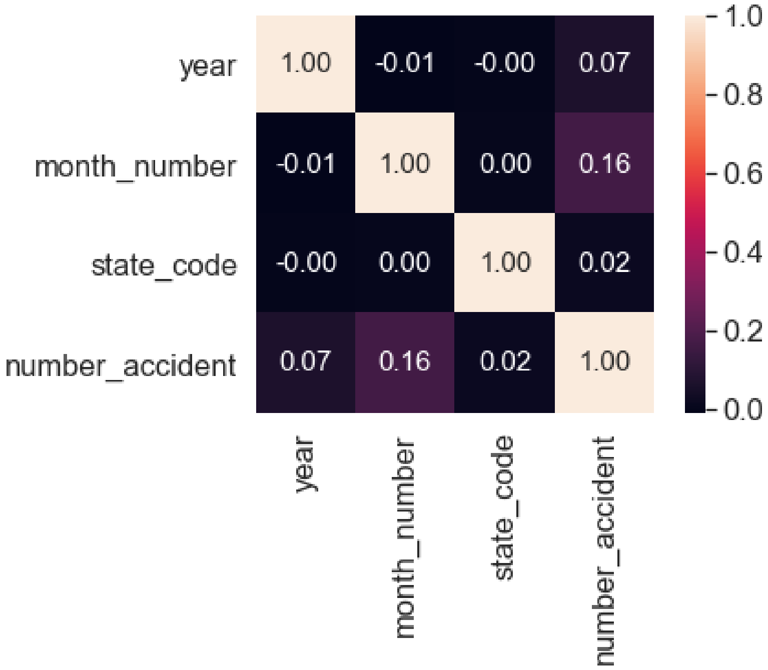


```
Number of accident in January 2020 in acre, Linear Regression :  [[79.91965635]]
Number of accident in April 2021 in Bahia, Linear Regression :  [[111.27417837]]
Number of accident in October 2022 in Manto Grosso, Linear Regression :  [[170.50848762]]
```

To get this result, we predict number of forest burn based on the Year, Month Number and State Code. For example : We are going to predict the result of number of forest burn in Acre in January 2020 there when we applied Linear Regression Method the result is there will be

79 forest burn in Acre in January 2020. This Step also applied to the other Machine Learning Method.

Conclusion



Based on this diagram we can see the correlation between the column in the datasets. I can make a conclusion that our machine learning is working perfectly fine since all of the method produce a result even though the result is not quite the same but it was the formula since every algorithm used different formula to calculated. The accuracy of the algorithm also not that good however it’s not our fault since we only used the dataset that has been given to us and the correlation between the data is quite low.

For the recommendation, i would like to suggest the government of brazil to add more column that relate with the data such as cause of fire or total damaged of the accident. Therefore we can predict more accurate score

What Have I Learned

I have learned many things in this open program, from data science topic and personal soft skills. For the data science stuff i learned how to use Regression algorithm such as Linear Regression, Decision Tree and KNN. I learn how to used regression because on the individual challenge i used many of classifier method and i want to push my self to learn more about machine learning and predictive performance. I also learned how to clean the data because not all the dataset that i will get in the future will be clean and there always be messy data frame that i need to clean before i use it.

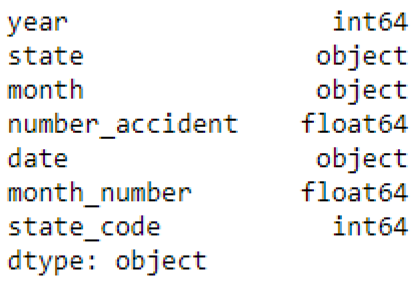
On my soft skills, i also improved my time management because to do this assignment i need to make a planning because i have a lot of courses this semester and if i can’t use time management that well my assignment will be chaos and there will be assignment that i will late to submit it.

Topics That i learned :

1. Predictive Performance
2. Machine Learning
3. Data Quality

Appendix

Here is our complete dataset with their type look like :



```
year                 int64
state               object
month               object
number_accident    float64
date                object
month_number       float64
state_code           int64
dtype: object
```