Daviano Almeyda Kustaryanto – 3212610



Forest Fire In Brazil

Open Programme : ADS A

Contents

[Introduction 2](#_Toc29577521)

[Assumption 2](#_Toc29577522)

[Data 3](#_Toc29577523)

[Cleaning Data 3](#_Toc29577524)

[Creating New Column called Month Number and Change the language 3](#_Toc29577525)

[Analysis Report 6](#_Toc29577526)

[Result 8](#_Toc29577527)

[K Nearest Algorithm 8](#_Toc29577528)

[Decision Tree 8](#_Toc29577529)

[Linear Regression 8](#_Toc29577530)

[Conclusion 9](#_Toc29577531)

[What Have I Learned 9](#_Toc29577532)

[Appendix 10](#_Toc29577533)

# 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 :

A screenshot of a cell phone

Description automatically generated

# 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

A screenshot of a social media post

Description automatically generated

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

A screenshot of a social media post

Description automatically generated

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

A screenshot of a social media post

Description automatically generated

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.

A screenshot of a cell phone

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

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.

A close up of a logo

Description automatically generated

Finally, i export the new dataset into a new csv file called amazon\_cleaned.

# Analysis Report

`

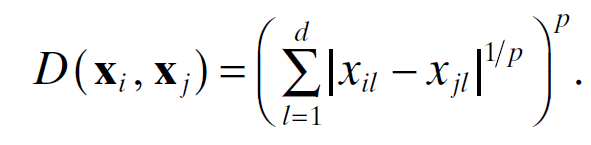
In order to complete this project, we are going to used some machine learning modules. We used many Machine Learning model to compare the result between each other. Below is the type of machine learning that we used :

1. KNN Algorithm( K Nearest Neighbours)

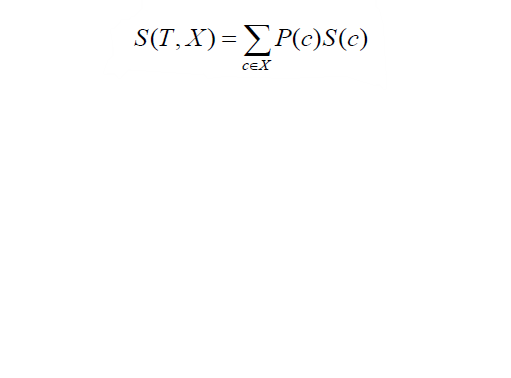
The K-Nearest Algorithm is a supervised classification algorithm, It takes a lot marked points and use them to learn how to label another point. The reason we used this technique is this technique is quite common among data scientist, It is very simple to implemented and many major company used this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

1. 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.

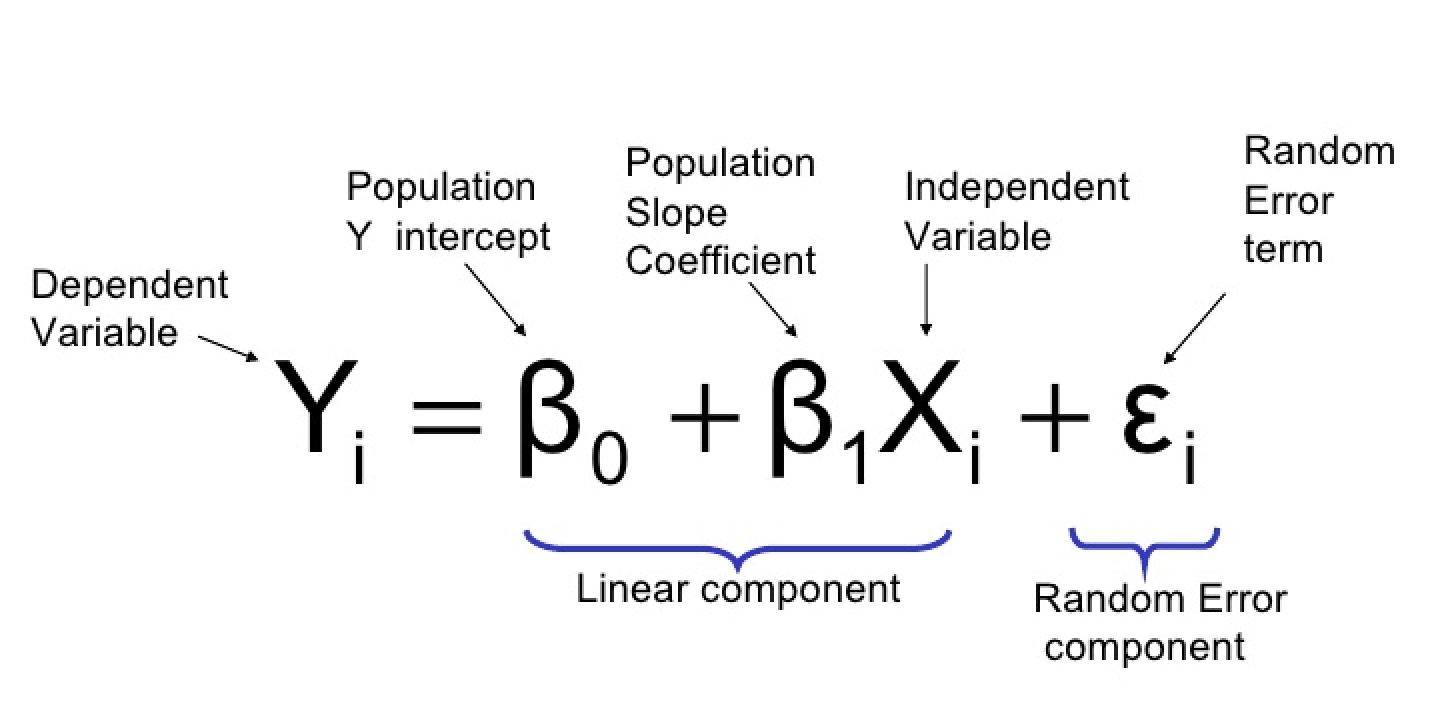


1. Decision Tree Regressor



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.

1. Linear Regression



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 :

A screenshot of a cell phone

Description automatically generated

### 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 :

A picture containing indoor, orange

Description automatically generated

### 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 :

A close up of a person

Description automatically generated

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

A screenshot of a cell phone

Description automatically generated

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 :

A screenshot of a cell phone

Description automatically generated