Oct. 11th – 15th:

* Based on the conversation with the supervisor, we decided to have a more accurate data group with adding rows of data that also shows statistics regarding to pipes without broken till now. To accomplish this, the following changes to the “data processing” and “Naïve Bayes and Prediction” files are made:

1. Currently, we have already data groups with “Times\_Broken” as 1 and 2.

A picture containing diagram

Description automatically generated

By doing the above code, returns all the rows that never broken since they were constructed.

1. Then, by adding a column called “Times\_Broken” also to data group “datacopy”, the three tables of data can be merged as one now. And it was called as “final\_data\_with\_all\_times\_broken”.

Text

Description automatically generated with low confidence

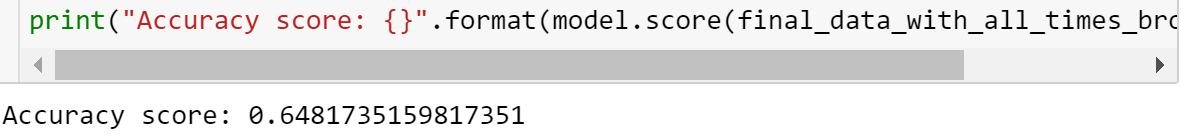
1. As for the file “Naïve Bayes and Prediction “, what to do was to modify the data group to the new one but the consider how many rows of data (“Times\_Broken” as 0) should be picked. We have a data set of 2190 rows that reports broken (“Times\_Broken” as 1: 2122, “Times\_Broken” as 2: 68), and to balance the data group, it should be a number around 2190.



Then, we randomly pick 2000 rows from it. The accuracy turned out to be 

We tried with 2100 rows and the accuracy was: 

With 2190 rows, the accuracy was shown as below:



With 2300 rows, the accuracy was shown as below: 

So, it came up with a question: how many rows should be chosen here as a better solution? And why?

* Another try was also based on the coding from last week, which is whether “fra\_kote” affect the broken of the pipes or not.

From last week, we had the result as below by using Ridge and Lasso Regression:

Text

Description automatically generated A picture containing table

Description automatically generated

So, this time, with the new data group and only input “fra\_kote”, the results of both show as below:

Graphical user interface, text, application, chat or text message

Description automatically generated

A picture containing graphical user interface

Description automatically generated

With a total data group of 4190 rows. Text

Description automatically generated

So, the conclusion is very clear that “fra\_kote” does not affect the broken of pipes using Ridge and Lasso Regression. And the value was large only because there is a “til\_kote”. By doing the addition of these two, it also shows that neither of them affects that since the result is very low to 0.