* What is the main objective? What are we trying to predict?

Potential number of vacancies in the next quarter by NOC

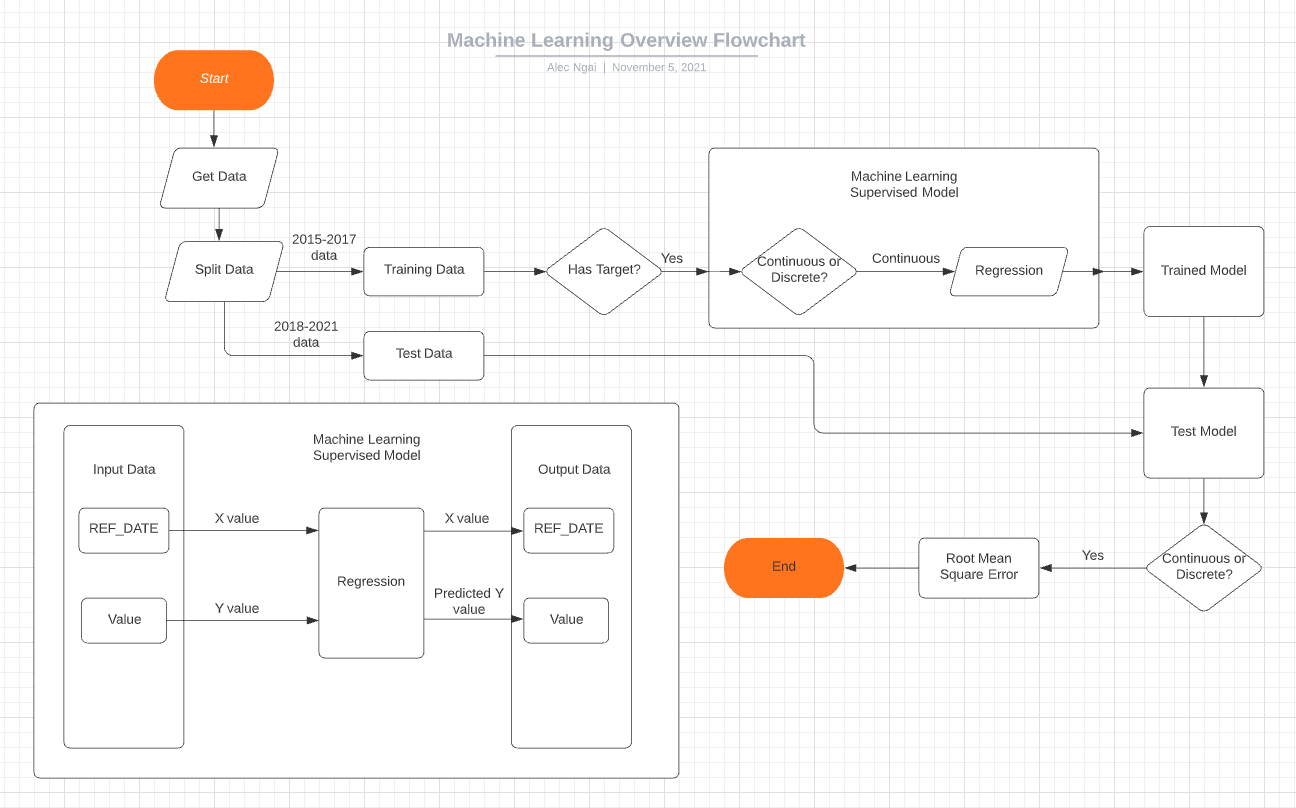
* What are the target features?

We want to target National Occupational Classification as a grouping, Job Vacancy characteristic as a grouping and predict the value in Value column in relation to REF\_DATE.

* What is the input data? Is it available?

| Column Name | Data Type | Process |
| --- | --- | --- |
| REF\_DATE | Object | Convert to date format, this will establish timeline as x-axis |
| GEO or DGUID | Object | Can convert DGUID into int64 to be used as location identifier |
| National Occupational Classification | Object | Can convert to int64, this will be grouping of different jobs, there are 692 different Classifications, these can be grouped into more broader industries |
| Job Vacancy Characteristics | Object | Can convert to int64, this will be grouping of different types of work, full time vs part time, i.e. there is 47 different characteristic |
| Statistics | Object | Only 3 statistics, Job Vacancies Proportion of job vacancies and average offered hourly wage |
| UOM or UOM\_ID | Object / int 64 | This column just defines if value is a percentage or a number or in dollars. |
| VALUE | float 64 | This value relies on UOM and Statistic to be understood. |

* What kind of problem are we facing? Binary classification? Clustering?
* What is the expected improvement?
* What is the current status of the target feature?
* How is going to be measured the target feature?



Ref

<https://towardsdatascience.com/machine-learning-general-process-8f1b510bd8af>

<https://www.youtube.com/watch?v=1wVgtINZIT4>