Here we have text file in our project report on cancer.

So, the project report contains text data as we discussed earlier, which is nothing just contains the brief information about the type of cancer and genes associated with the cancer.

Now, form one perspective this is very important for the researcher’s point of view but here this report we cannot directly feed to the ML-based algorithm. So, here first we need to cleanse the data like we do in the NLP.

In cleansing now what needs to be removed/ replaced and what to keep so that we can smooth the data as per our requirement without losing any information.

* Remove the double spaces and tabs present in the dataset.
* There are so many unnecessary numbers present in the dataset which we do not need right now for our processing the data.
* There are so many words in the data set where same words use in different format like gene and Gene.
* There are so many stop words present in the text data file like an, a, the etc…

Now, after cleansing we are going to merge our dataset with the variants file, there is common component needed to merge and that common component here is “ID”.

After merging we create a new data frame, so the question arises that is there any missing value present in the dataset and if there is how to deal with it??

There are many methods to deal with the missing values but just because we do not have very large dataset so I opt to impute the missing values with the genes and variations.

Elsewise, if I had a very large data, I could have dropped out the missing values.

Now, this data is ready for the next step…... i.e., Training, Testing the dataset etc…...