**Iteration 1** – Data Specifics

File Name – ‘TCGA\_data.csv’

* After preprocessing the data, we have total 8704 rows and 152 columns.
* There are total 32 different cancer types and the most occurring ones is “BRCA” with a frequency of 970. Therefore, this has been chosen for the analysis.
* To have consistency among the results, we would be using XGBOOST for classification.
* 20%-80% dataset is divided into test-train dataset.

After applying the model, the accuracy is calculated from confusion matrix, i.e.

The accuracy came out to be 89.83% over test dataset of size of 1741.

![A close up of a map

Description automatically generated]()

From the confusion matrix, False Negative was 171, which means the BRCA cancer was not predicted accurately. The aim is to decrease FN/FP as much as possible.

|  |  |  |
| --- | --- | --- |
| **Setting nan values to** | **Accuracy** | **AUC** |
| 0 | ﻿0.896598060290192 | ﻿0.8940048573878775 |
| Mean of column | ﻿0.8964525298408624 | ﻿0.9024087275869518 |
| Median of column | ﻿0.8958786399770806 | ﻿0.8979293911244657 |

So far the data we gathered in this iteration would act as the baseline for future experiments.