****

**MCTA 4362: Machine Learning**

**(Sec. 1)**

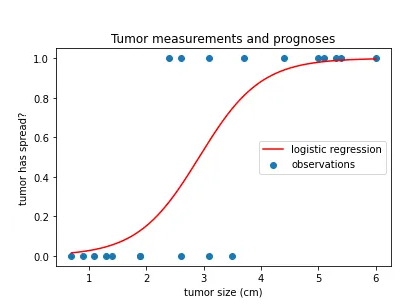
**Quiz/Test #1**

*(LO 1)*

Name: Muhamad Nurhakimie Thaqif Bin Abdullah Matric No: 2213217

**Answer ALL Questions.**

1. Calculate the likelihood for the following logistic regression problem. (you may estimate the values of some of the projections to the sigmoid curve since the graph here does not have a grid)



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tumor Size | Tumor Has Spread | Estimate Likelihood | Likelihood | Log(likelihood) |
| 0.5 | 0 | 0.005 | = 1- 0.005= 0.995 | -0.0050 |
| 0.8 | 0 | 0.05 | = 1- 0.05 = 0.95 | -0.0513 |
| 1.2 | 0 | 0.07 | = 1- 0.07 = 0.93 | -0.0726 |
| 1.4 | 0 | 0.09 | = 1- 0.09 = 0.91 | -0.0943 |
| 1.5 | 0 | 0.12 | = 1- 0.12 = 0.88 | -0.1278 |
| 1.9 | 0 | 0.16 | = 1- 0.16 = 0.84 | -0.1744 |
| 2.4 | 1 | 0.30 | = 0.30 | |  | | --- | |  |   -1.204 |
| 2.5 | 0 | 0.33 | = 1- 0.33 = 0.67 | -0.4005 |
| 2.6 | 1 | 0.35 | = 0.35 | -1.0498 |
| 3.0 | 1 | 0.58 | = 0.58 | -0.5447 |
| 3.2 | 0 | 0.59 | = 1- 0.59 = 0.41 | -0.8916 |
| 3.5 | 0 | 0.67 | = 1- 0.67 = 0.33 | -1.1087 |
| 3.7 | 1 | 0.80 | = 0.80 | -0.2231 |
| 4.2 | 1 | 0.90 | = 0.90 | |  | | --- | |  |  |  | | --- | | -0.1054 | |
| 4.9 | 1 | 0.94 | = 0.94 | |  | | --- | |  |  |  | | --- | | -0.0619 | |
| 5.0 | 1 | 0.95 | = 0.95 | |  | | --- | |  |  |  | | --- | | -0.0513 | |
| 5.2 | 1 | 0.97 | = 0.97 | |  | | --- | |  |  |  | | --- | | -0.0305 | |
| 5.4 | 1 | 0.98 | = 0.98 | |  | | --- | |  |  |  | | --- | | -0.0202 | |
| 6.0 | 1 | 1.0 | = 1.0 | |  | | --- | |  |  |  | | --- | | 0.0000 | |
| Total |  |  | 1.995 10-3 | -6.217 |

Therefore, The likelihood is around -6.217

1. Please refer to the following distribution of data which shows two classes of data. In a KNN algorithm what does this circle mean and what could be the impact of the diameter of this circle on your classification accuracy

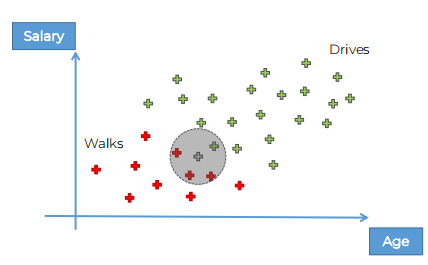


Figure 1: Walking and Driving Clusters of data.

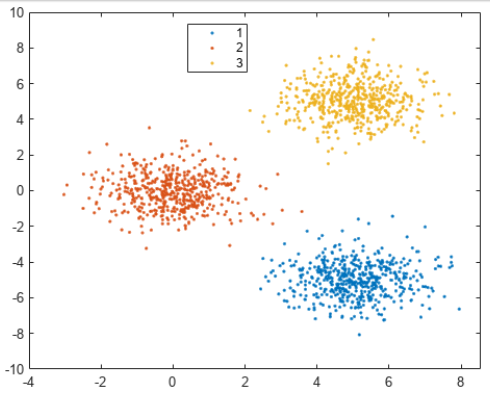
The Circle means neighborhood around a query point, which signifies area KNN select Its nearest neighbor, the center of the circle is class that has not been classified.

The larger the area means larger K, which means more points can be considered therefore more stable as it selects more data points, and the smaller area means smaller K which means fewer points would be considered which allow to capture local points easier.

If the area is too large it would cause underfit as it would be influenced by other neighboring easier. However, if it is too small then it would be overfit as it would be easier to be influenced by noise.

1. Please refer to Figure 3 which shows three classes of data. Draw a Decision Tree which will distinguish the three classes.

X2



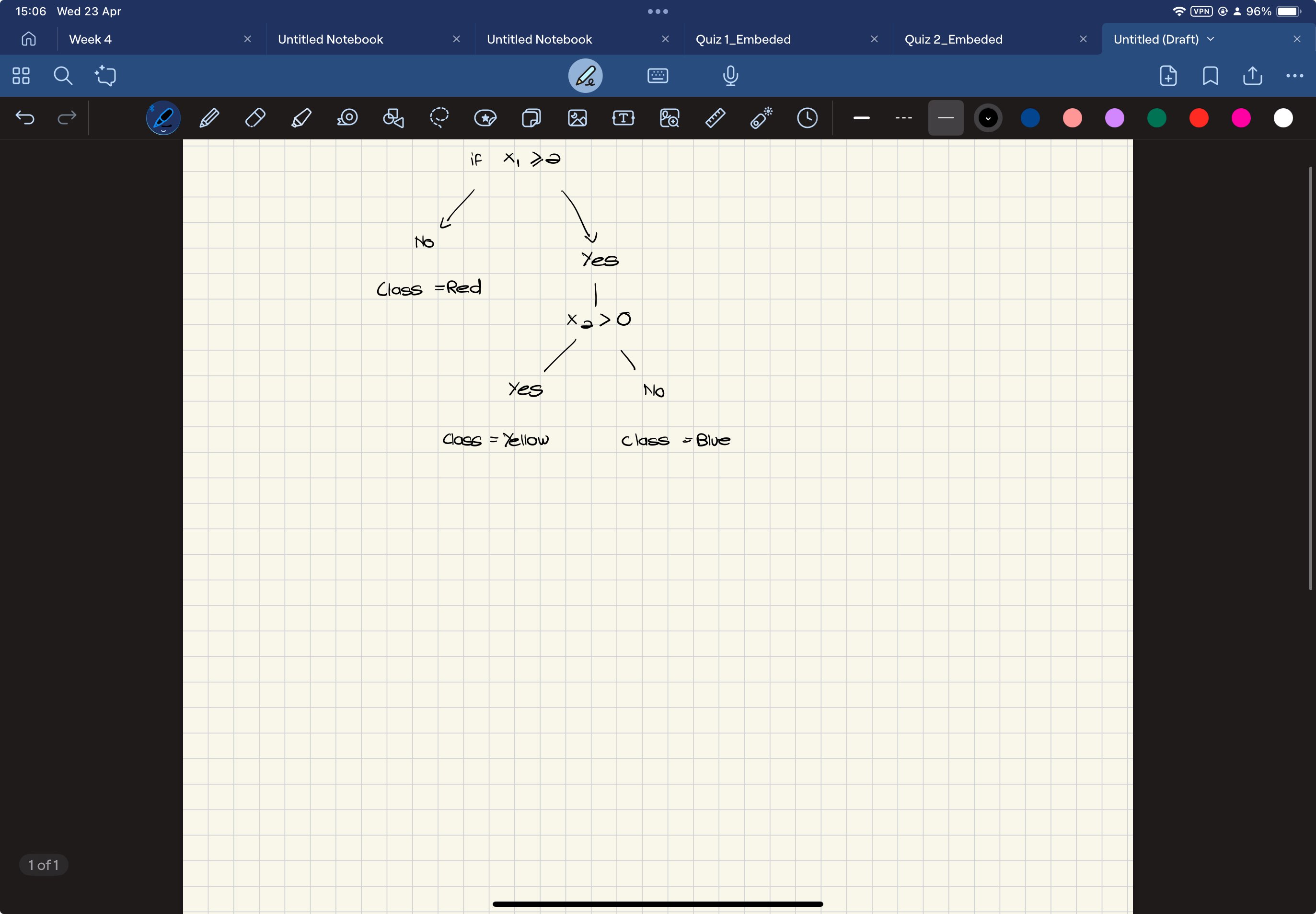
Yellow

Blue

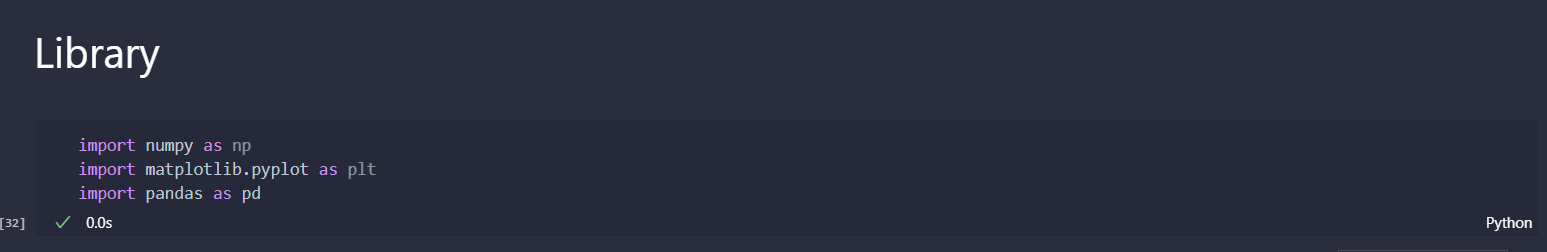
Red

X1

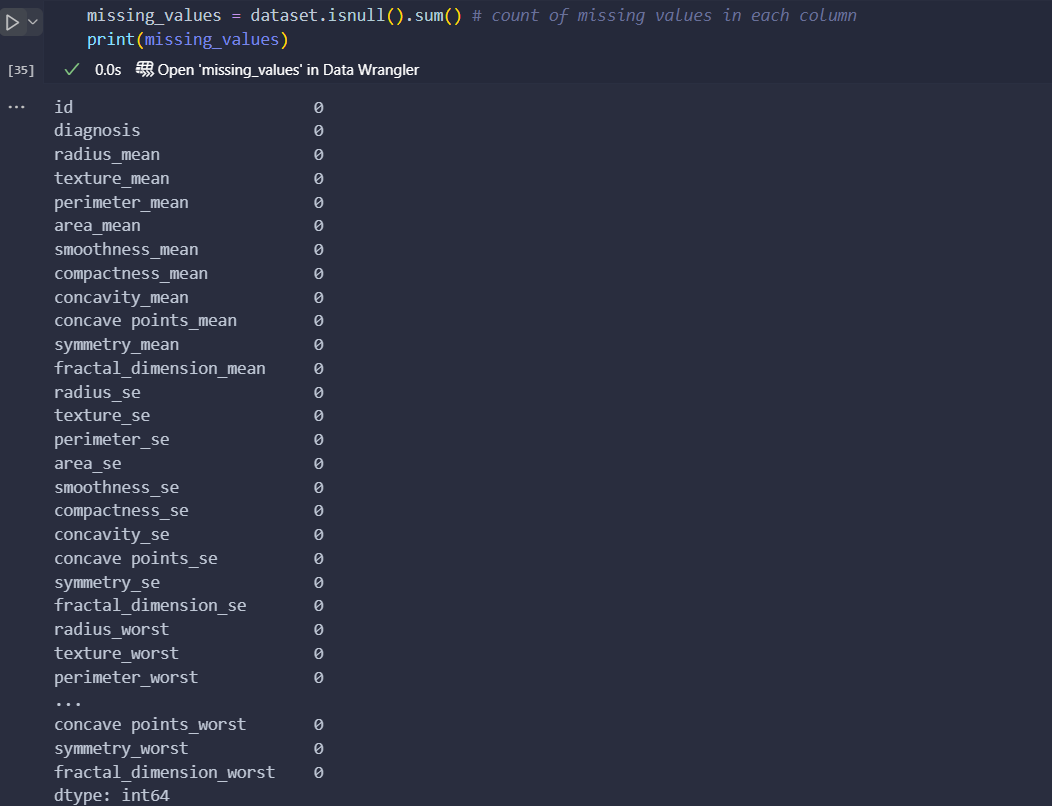
Figure 3: Three classes of data



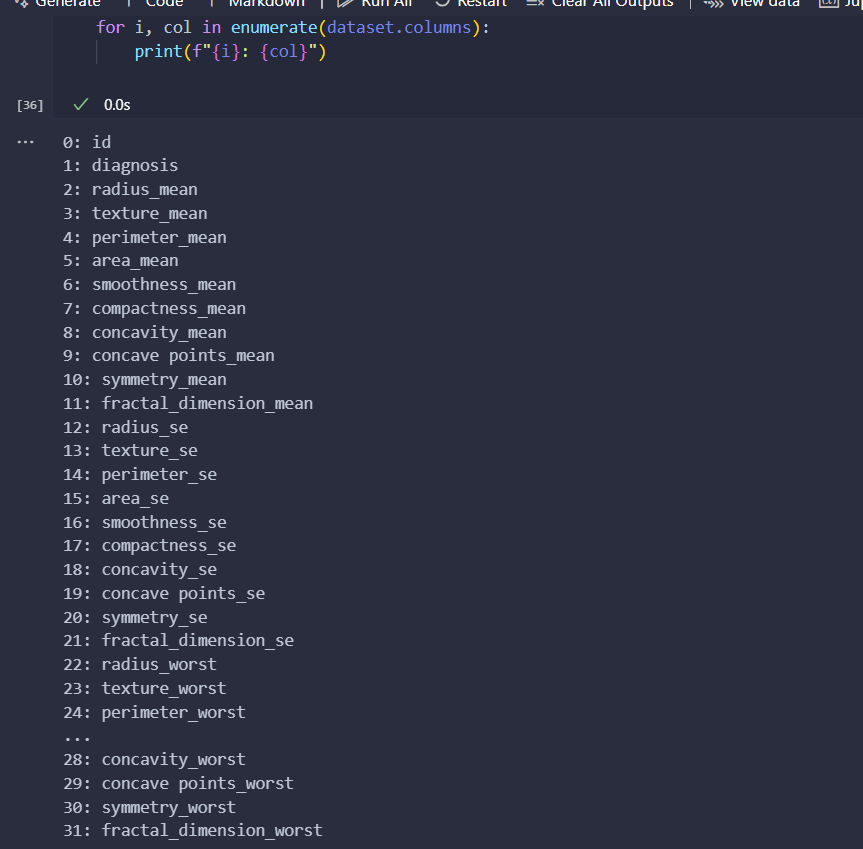
1. The following dataset contains the characteristics of patients diagnosed with cancer. The dataset contains a unique ID for each patient, the type of cancer (diagnosis), the visual characteristics of the cancer and the average values of these characteristics. By using KNN, train a model to classify the two different classes.



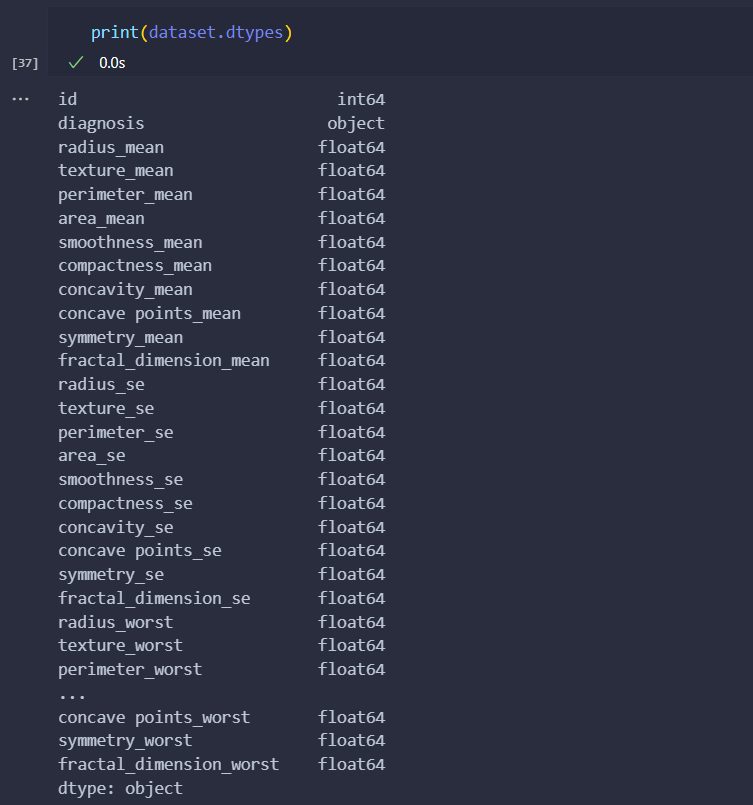
Import Main Library



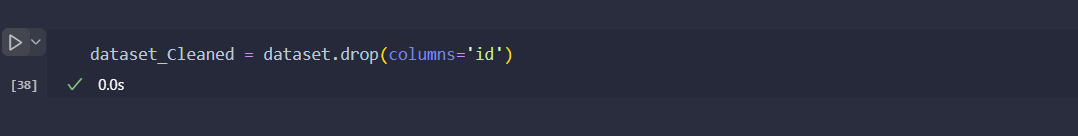
Check for missing Value it looks like none



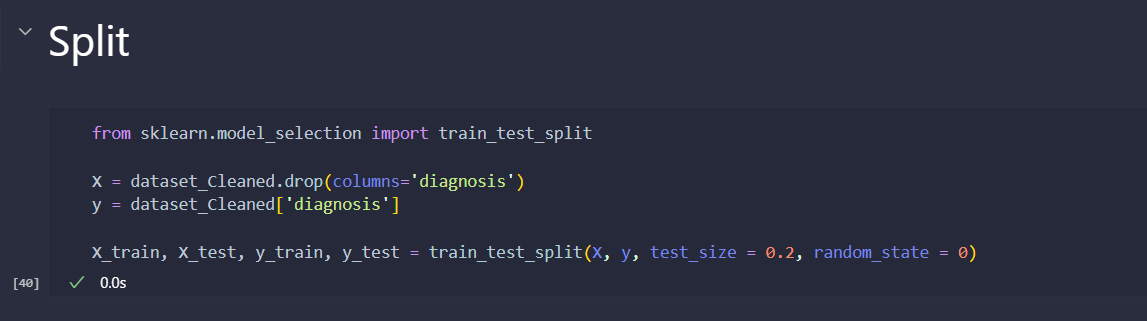
Get the Index Id in for checking if we want to use ID instead of Name Of header



Checking for data type to know if we want to use one hot encoding or not



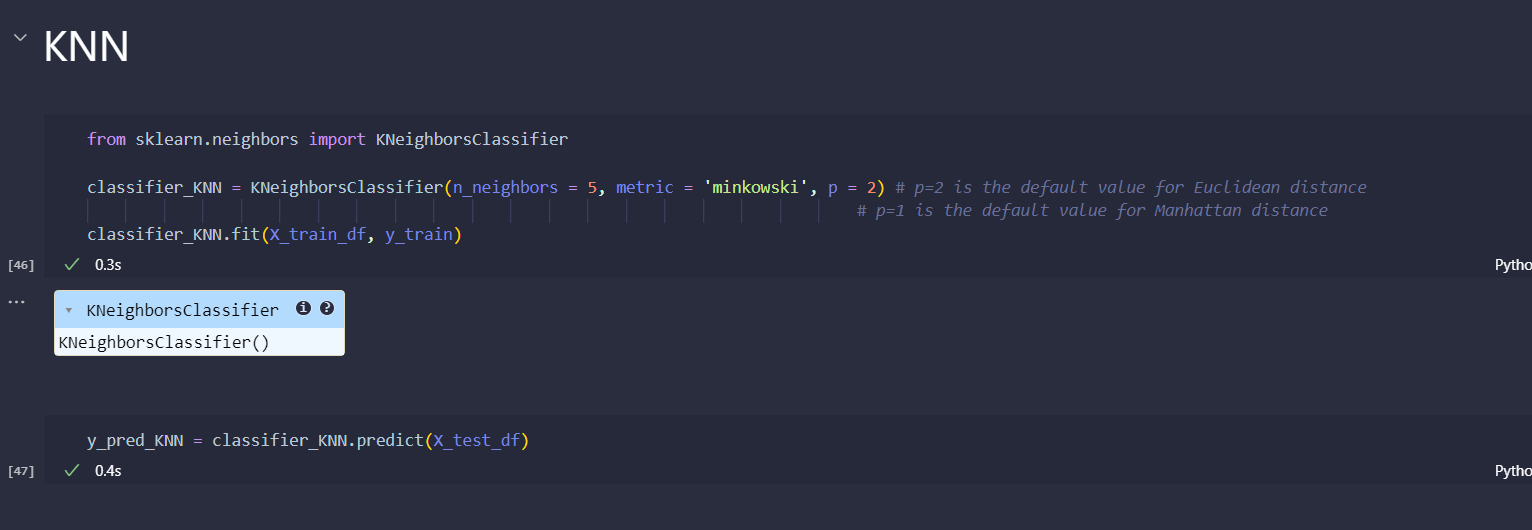
Remove ID as it is useless in training



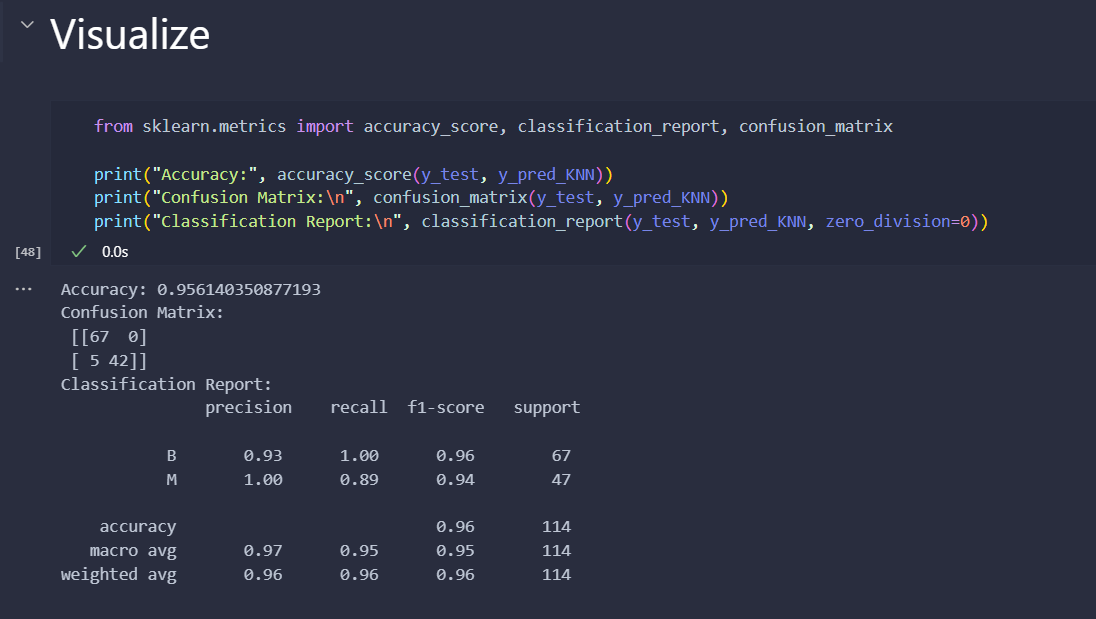
Split Data set to training 80 % Test 20 %



Apply Feature Scaling To make sure no bias of bigger data value happening and kept everything in range [3,-3]



Doing KNN with 5 Neighbor and Metric Minkowski with p=2 for Euclidean



Visualize to check performance, here we got **95.61 %**