**NAME – PRABHAKAR KUMAR**

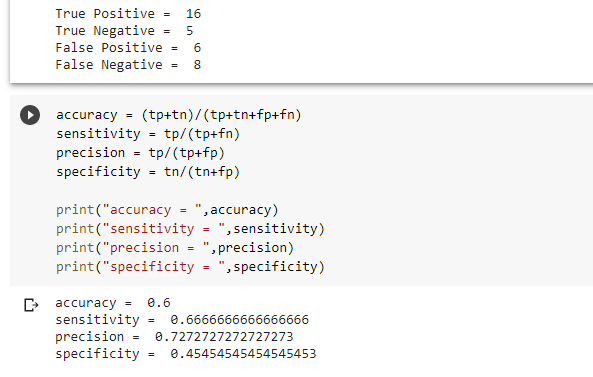
**Roll – IRM2017008**

**Assignment 4 Question 3 Analysis**

**Part 1**

In first part I implemented the Gaussian Discriminant Analysis algorithm on the Micorchip dataset, without performing any Feature Engineering. In the question it was specifically given that we had to create three more features from the only two existing features that were given in the question dataset, which could be done by feature engineering. But one thing to be noted was that the independent variables given to us were continuous numbers, rather than something like classes or labels, so there might have been possibility of under-performance of the model if trained on dataset after feature engineering.

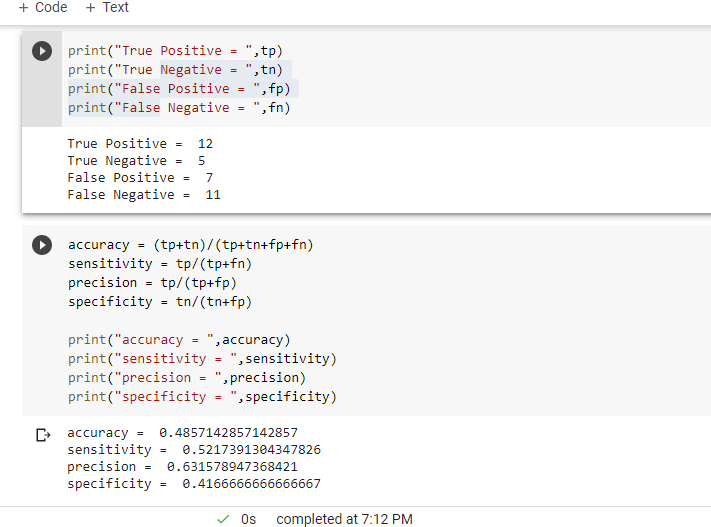
Hence I first tried to train the GDA algorithm over the raw dataset that was given to us. Splitting the data into 70-30 ratio in random for training and testing of the model, we obtained the confusion matrix and the multiple parameters that are used with binary classification problem as follows:



**Part 2**

In part 2, I implemented the Gaussian Discriminant Analysis algorithm on the Micorchip dataset, after performing any Feature Engineering as was given in the first question. Feature engineering is especially useful in case of labelled variables, where we create multiple binary independent variables from the labelled or class variables. When feature engineering is performed with data where columns have continuous data, and the model for which the data is being prepared requires the data to be IID, then often feature scaling may result in the model under-performing in some cases, because some the columns created by linear or exponential combinations of raw columns, may not strictly follow the IID requirement for the model.

Hence we addressed this question by creating three more columns by feature engineering where the newly created columns were the linear combinations of the raw columns. Upon splitting the dataset into training and testing and them training the GDA model and then testing over the test data, the following results were obtained:



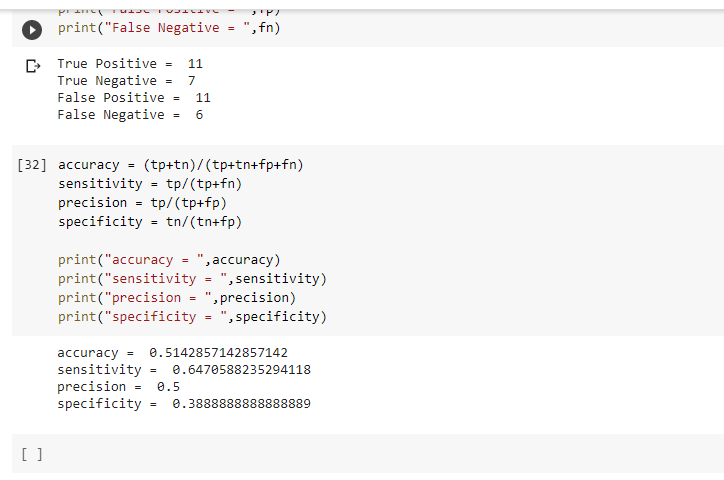
Here we can clearly see that adding columns with feature scaling resulted in loss of precision and overall cumulative performance of the GDA model.

**Part 3**

In part 3, I implemented the Gaussian Discriminant Analysis algorithm on the Micorchip dataset after having treated the given dataset with Box-Muller transformation. Basically the major task of Box-Muller transformation is to convert any independently identical distribution like say any uniform distribution to another independently identical distribution, preferably to a Normal or Gaussian Distribution.

Here I have also firstly applied MIN-MAX Normalization over the raw data, so that the raw data is under the range of [0,1]. Doing this would help us overcome any possible semantic error in transformation, like division by zero or taking log of any bad number.

Upon applying Box-Muller Transformation, training and testing over the splitted dataset, we received the following confusion matrics and the following output analysis parameters:



Here it should be noted that using Box-Muller transformation, did definitely help us in improvement of the overall performance of the model over the unseen data points, as it would result in one of the favourable condition where GDA model is preferred that the P(X|Y) is distributed according to a Multivariate Normal Distribution.