**Wide Bot Classification challenge**

**1-Preprocessing steps**

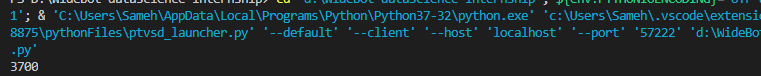
**2-Model selection steps**

**1-Preprocessing steps**

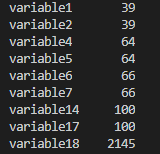
**-Removing null values**

The following facts was found when exploring the data:

1-the number of nulls in rows was 3700



2-the number of nulls in ‘variable18’ was very high 2145 values out of 3700



My decision was to remove ‘variable18’ column due to the large number of nulls in this row

3- after removing this column the number of nulls remaining in the rows was 178

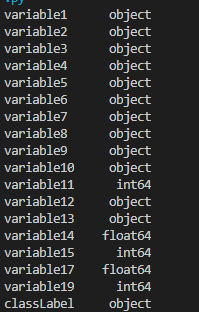


My decision was to remove the rows with the null values as it didn’t reach 5 percent of the total dataset

**-Converting the data to numeric values**

After removing the null values the following was found in the dataset

1-the datatypes of the columns was as follows:



To be able to perform classification data must be numeric so I converted all the columns with object datatype including the class label

2-after conversion the following columns could be normalized for efficient use of any classifier (the minimum and maximum value for each column):

Variable2:



Variable3:

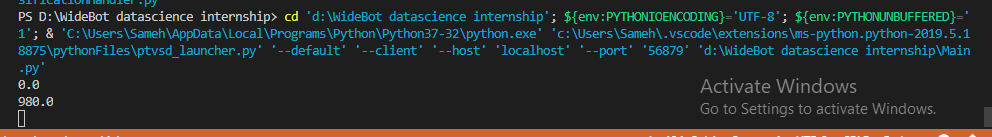


Variable 8:

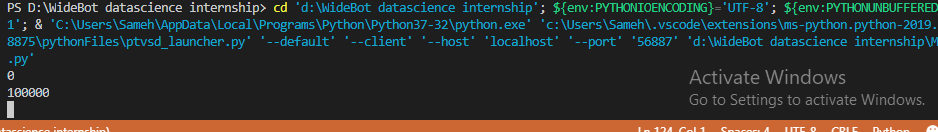
Variable 11:



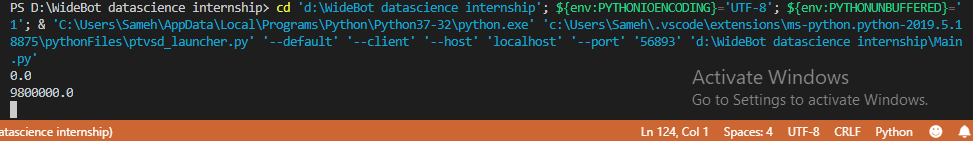
Variable 14:



Variable 15:



Variable 17:

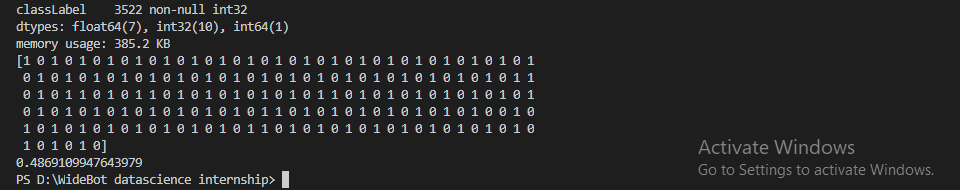


All the previous columns was normalized for efficient use of classifier

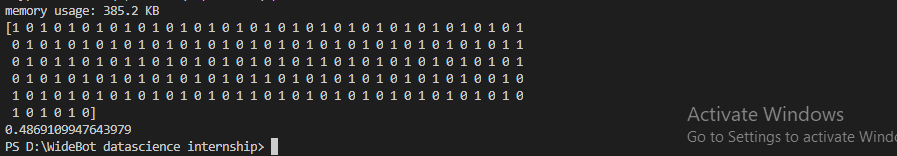
**2-Model processing steps**

1-Now that we have the data normalized and ready I have tested the following classifiers with their accuracy included:

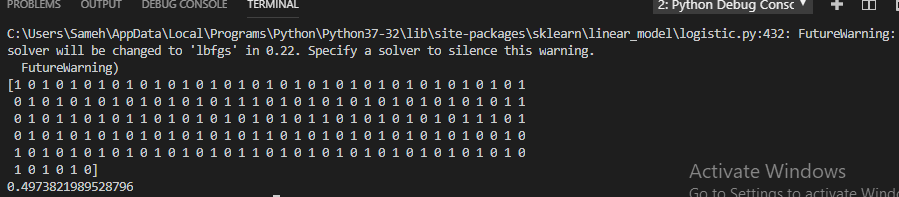
**1-Naief Bayesian: 48% accuracy**



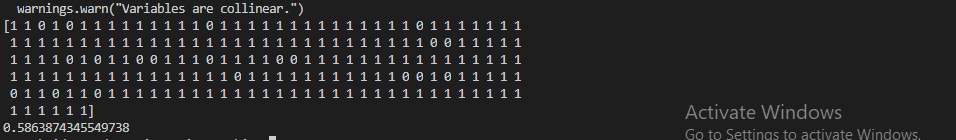
**2-Decision Tree: 48% accuracy**



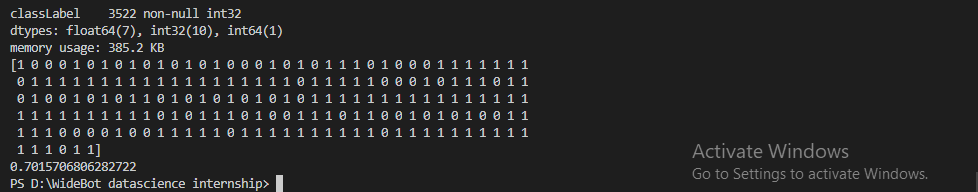
**3-logistic regression: 49% accuracy**



**4-Linear discrimination analysis: 58% accuracy**



**5-KNN: 70% accuracy**



From the previous finding my decision was to choose the knn classification model.