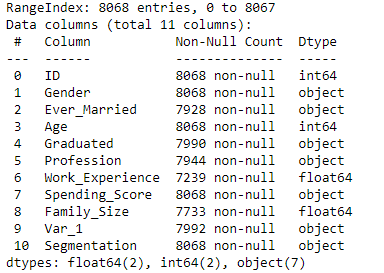
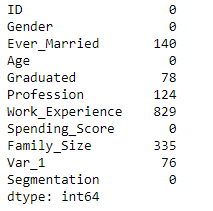
**Multiclass Classification**

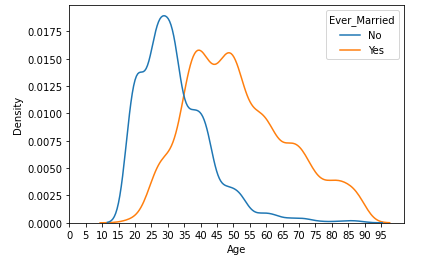
The dataset from Customer\_segmentation\_train.csv has 8068 rows and 11 columns. The column ‘Segmentation’ is the label class.



Above I have included information of the dataset. We can see that some data is missing. To see exactly how much, I have included below count for number of values missing for each column.



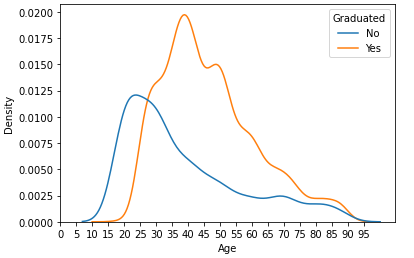
Plotting the Age distribution and separating based on Ever\_Married we get:



From this plot I decided to fill all null values for the Ever\_Married column in the following way:

* If Age <= 30, I will fill as No
* If 30 < Age < 40, I will drop the row
* If Age >= 40, I will fill as Yes

Plotting the Age distribution and separating based on Graduated we get:



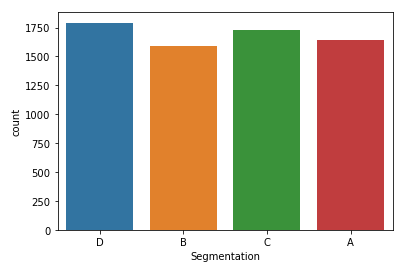
From this plot I decided to fill all null values for the Graduated column in the following way:

* If Age <= 25, I will fill as No
* If 25 < Age < 35 I will drop the row
* If Age >= 35, I will fill as Yes

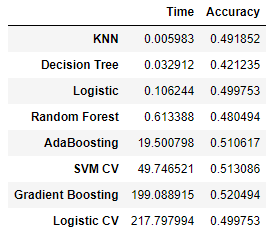
For Profession column I decided to fill the null values as ‘None’. Unable to find a pattern for the remaining columns, I decided to drop the rows which had null values in the columns Var\_1, Work\_Experience and Family\_Size.

I converted Spending\_Score column from object type to integer by mapping Low to 1, Average to 2 and High to 3. I dropped the ID column. I then used dummy variables to allow the machine learning models to work on the non-numeric data. The same process was done for Customer\_segmentation\_test.csv data as well.

We have 1639 examples of class A, 1588 examples of class B, 1730 examples of class C and 1792 examples of class D. We can say that the classes are balanced.



After training various classification models on the training data and comparing them with a validation set, we get:



From this we can see that SVM with cross validation has good accuracy without taking too long so we should use this model.

Checking feature importance using the decision tree we get that out of the 22 columns the ones with most importance are Age (0.346), Work\_Experience (0.151) and Family\_Size (0.115).