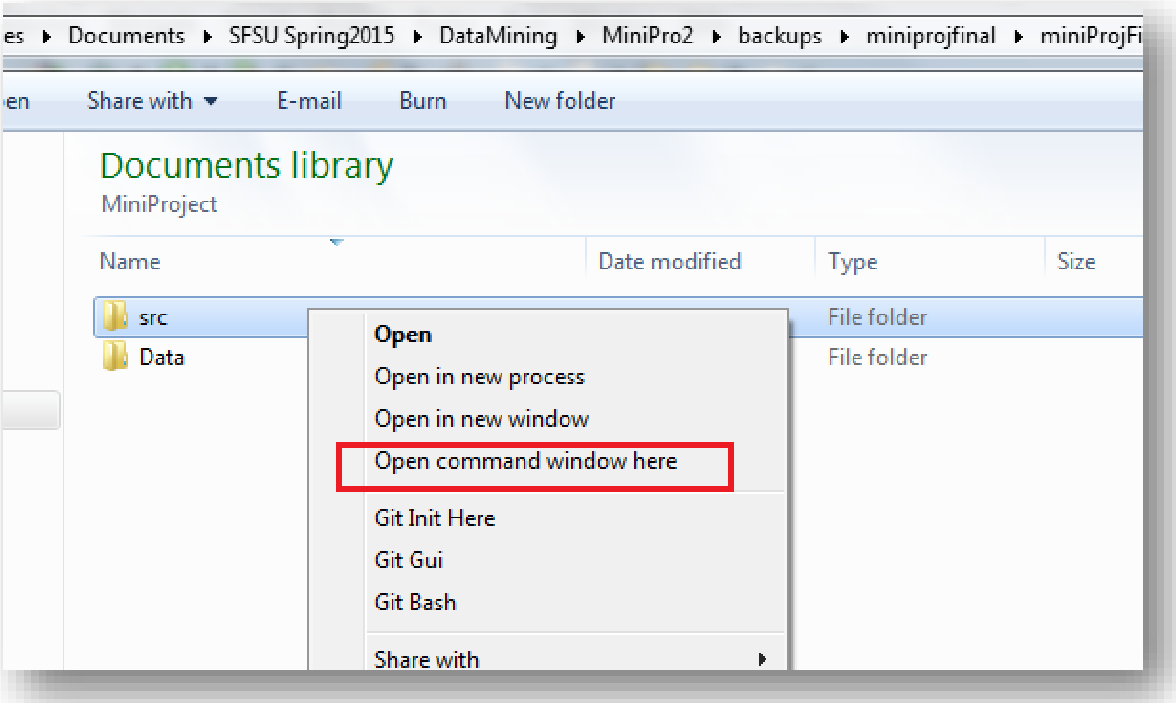
How to run the project and read output

Prerequisites:

- Numpy and Scipy library: http://www.scipy.org/scipylib/download.html
- Pandas library : http://pandas.pydata.org/pandas-docs/version/0.15.2/install.html

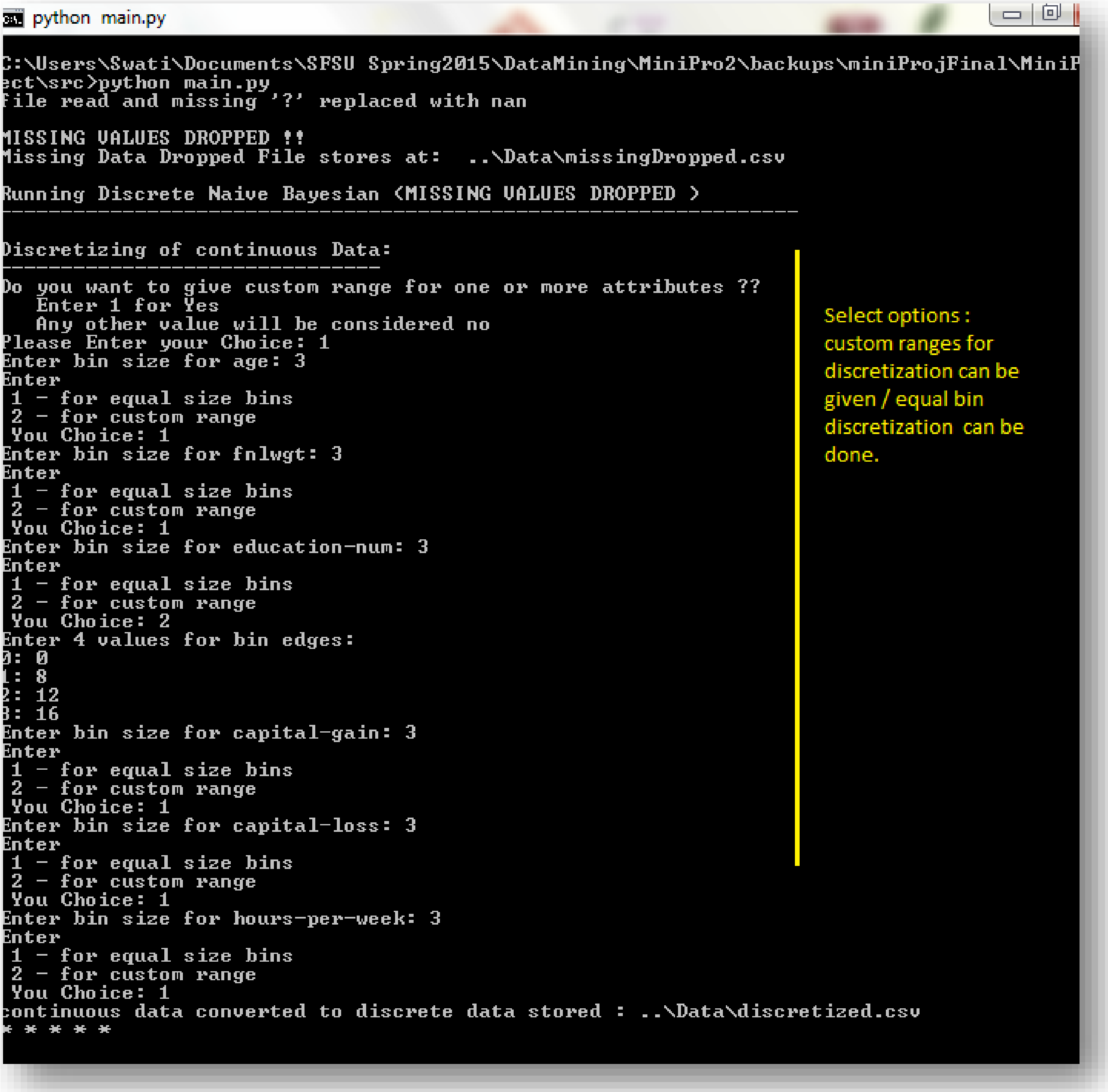
Step 1: Download the zip file and unzip the folder and open the command prompt in the src folder.



Step 2: run with command : python main.py

Program will ask for inputs for data discretization step . As shown below:

As the kfold evaluation progresses visual cues “*” will indicate the progress.

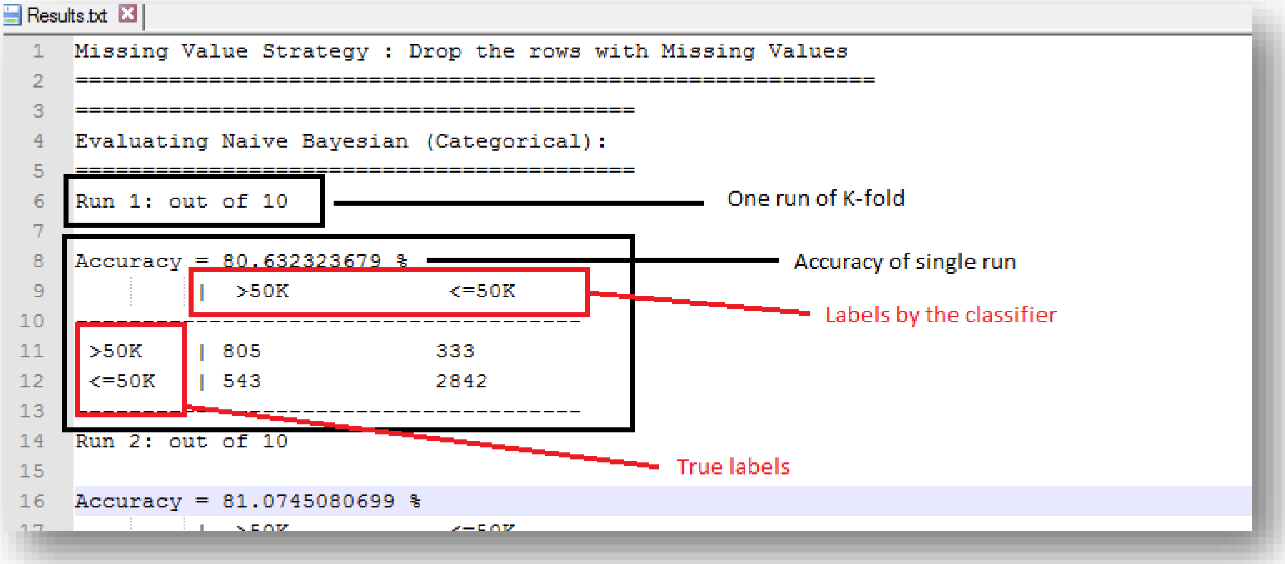


```
C:\Users\Swati\Documents\SFSU Spring2015\DataMining\MiniPro2\backups\miniProjFinal\MiniP
ect\src>python main.py
File read and missing '?' replaced with nan

MISSING VALUES DROPPED !!
Missing Data Dropped File stores at:  ..\Data\missingDropped.csv

Running Discrete Naive Bayesian (MISSING VALUES DROPPED )
-----------------------------------------------------------------

Discretizing of continuous Data:
-------------------------------
Do you want to give custom range for one or more attributes ??
   Enter 1 for Yes
   Any other value will be considered no
Please Enter your Choice: 1
Enter bin size for age: 3
Enter
 1 - for equal size bins
 2 - for custom range
 You Choice: 1
Enter bin size for fnlwgt: 3
Enter
 1 - for equal size bins
 2 - for custom range
 You Choice: 1
Enter bin size for education-num: 3
Enter
 1 - for equal size bins
 2 - for custom range
 You Choice: 2
Enter 4 values for bin edges:
0: 0
1: 8
2: 12
3: 16
Enter bin size for capital-gain: 3
Enter
 1 - for equal size bins
 2 - for custom range
 You Choice: 1
Enter bin size for capital-loss: 3
Enter
 1 - for equal size bins
 2 - for custom range
 You Choice: 1
Enter bin size for hours-per-week: 3
Enter
 1 - for equal size bins
 2 - for custom range
 You Choice: 1
continuous data converted to discrete data stored : ..\Data\discretized.csv
* * * * *
```

Select options : custom ranges for discretization can be given / equal bin discretization can be done.

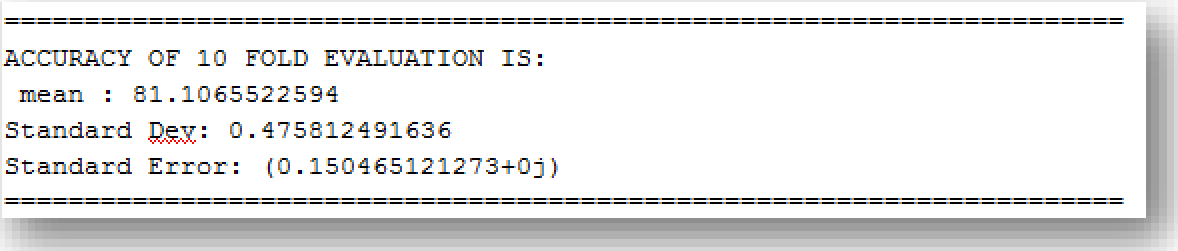
Step 3: Check Output in the Data Folder

- Result.txt : Stores the result of the single run of K fold as confusion matrix explained in figure below:



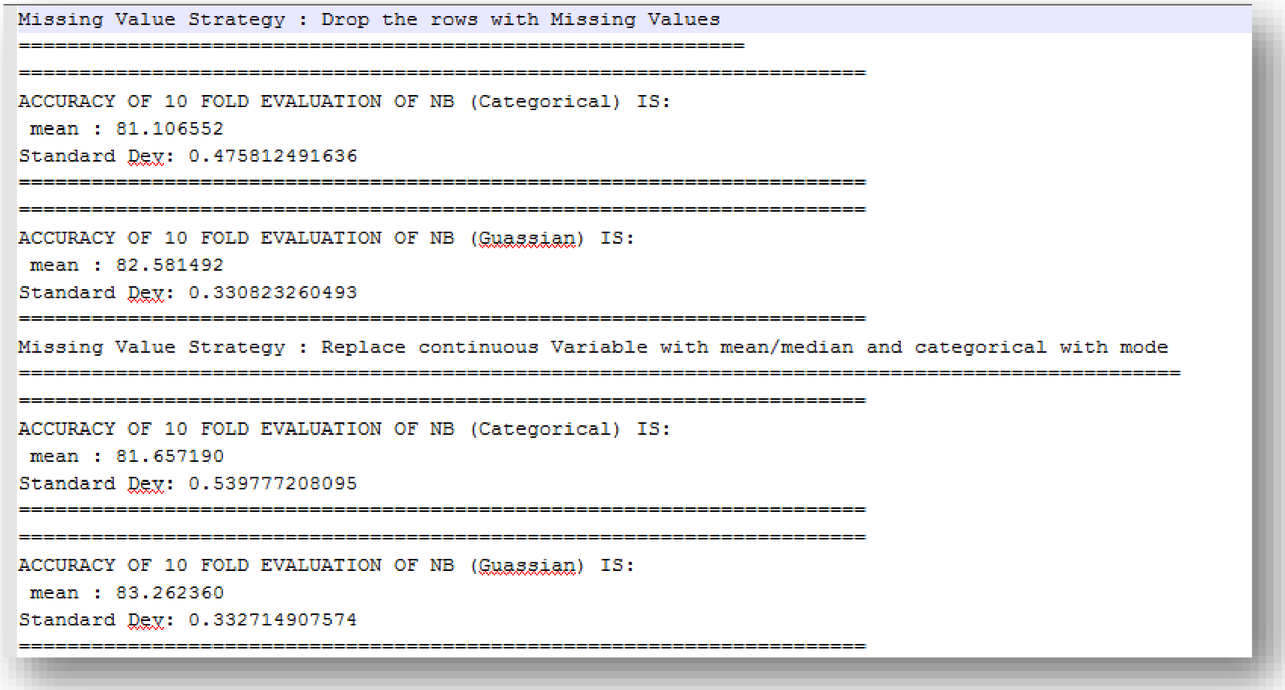
```
Missing Value Strategy : Drop the rows with Missing Values
============================================================
==========================================
Evaluating Naive Bayesian (Categorical):
==========================================
Run 1: out of 10

Accuracy = 80.632323679 %
        |  >50K             <=50K
-------------------------------------------------
 >50K   | 805               333
 <=50K  | 543               2842
-------------------------------------------------
Run 2: out of 10

Accuracy = 81.0745080699 %
        |  >50K             <=50K
```

Scroll down to find the average accuracy and standard deviation of 10 fold run:



```
======================================================================
ACCURACY OF 10 FOLD EVALUATION IS:
 mean : 81.1065522594
Standard Dev: 0.475812491636
Standard Error: (0.150465121273+0j)
======================================================================
```

- SummaryResult.txt : Stores the Summary Result of all four Evaluation , 2 different missing value handling * 2 different Naïve Bayesian implementation:



```
Missing Value Strategy : Drop the rows with Missing Values
============================================================
======================================================================
ACCURACY OF 10 FOLD EVALUATION OF NB (Categorical) IS:
 mean : 81.106552
Standard Dev: 0.475812491636
======================================================================
======================================================================
ACCURACY OF 10 FOLD EVALUATION OF NB (Guassian) IS:
 mean : 82.581492
Standard Dev: 0.330823260493
======================================================================
Missing Value Strategy : Replace continuous Variable with mean/median and categorical with mode
================================================================================================
======================================================================
ACCURACY OF 10 FOLD EVALUATION OF NB (Categorical) IS:
 mean : 81.657190
Standard Dev: 0.539777208095
======================================================================
======================================================================
ACCURACY OF 10 FOLD EVALUATION OF NB (Guassian) IS:
 mean : 83.262360
Standard Dev: 0.332714907574
======================================================================
```