REPORT

STEP – 1: DATA VISUALIZATION

1. To create the pie charts of the categorical data, I used the MATLAB and created pie charts of the following categorical data: *(Saved the pie charts in .jpg format)*
   - Workclass
   - Education
   - Marital status
   - Occupation
   - Relationship
   - Race
   - Sex
   - Native country
2. To create histograms of the continuous data, I used MATLAB and created histograms with different bin widths for the following categorical data: *(Saved the histograms in .pdf format)*
   - Age
   - FNLWGT
   - Education-num
   - Capital-gain
   - Capital-loss
   - Hours-per-week
3. I did the secondary data visualization analysis (Stacked bar charts of attributes with respect to classes) in MS-Excel for both the categorical data and for continuous data. *(Saved the stacked bar charts in the .pdf format*
4. Recorded my insights in the **1.Data Visualization** folder.

STEP – 2: HANDLING MISSING VALUES

1. Ignoring or deleting attributes of no significant impact on the analysis.
   - After doing data visualization, it became clear that the attribute FNLWGT and Education-num will have no impact on the analysis that we are doing.
   - As education attribute is similar to education-num attribute. Keeping one of the two will help reduce unnecessary data.
2. I wrote a python script *handle-missing.py* which implements two strategies of handling the missing values, namely:
   - Removing the missing values:
     we lose a couple thousand rows.
   - Replacing the missing values:
     we replace the ' ?' with the most frequent value of the attribute.

STEP – 3: IMPLEMENTATION OF NAÏVE BAYESIAN CLASSIFIER

1. As mentioned in the pdf, we have to implement Naïve Bayesian classier using this strategy.
   - Conversion of continuous attributes to categorical attributes and using equal width binning method. I only considered: Age, Capital-gain, Capital-loss and Hours-per-week for this task.

2. I wrote a python script *contTOcat.py* to convert all the continuous attributes to categorical. The inputs are the output files from step 2. In the end, a csv is produced with all categorical attributes.
   - I divided the continuous attribute as follows:
     'AGE' -> 'CAT-AGE'
     Young (0-25), Middle-aged (26-45), Senior (46-65) and Old (66-95).

     'HOURS-PER-WEEK' -> 'CAT-HOURSpW'
     Part-time(0-25), Full-time (25-40), Over-time(40-60) and Too-much (60-100).

     'CAPITAL-GAIN' and 'CAPITAL-LOSS' -> 'CAT-CAPITAL-GAIN' and 'CAT-CAPITAL-LOSS'
     None (0-Median), Low (median-arbitrary) and High (max).

3. Implemented Naïve Bayesian classifier for all categorical data. Named the file *NaiveBayesian.py*

4. Calculated the accuracy of the classifier.

   CALCULATED ACCURACY: 76.337 %

STEP – 4: IMPLEMENTATION OF K-FOLD CROSS VALIDATION:

1. Using k=10, I wrote a script *k-foldCV.py* where we calculate the mean accuracy of our classifier.
   THE MEAN ACCURACY: 76.007 %