Assignment#1

***Due Date (Hard!): September 20, 2017 11:55PM Total Marks: 10**

***Submit your assignment to the Course page at the CSE Eduserver**

***All submissions MUST be individual work and any form of copying, if found would lead to ZERO marks**

*** Every submission will be evaluated at a later period**

*** A ZIP file containing 3 PDF files may be uploaded with file name as Your_First_Name-RollNo**

1: [3 Marks] Download a data set of your interest from the UCI website (https://archive.ics.uci.edu/ml/datasets.html) or from the Kaggle website (https://www.kaggle.com/). Open the data set in *Open refine* and carry out the following operations which are relevant with the data set you have chosen. Try to choose data set with missing values:

- Cleaning up inconsistent spelling of terms (i.e. "USA", "U.S.A", "U.S.", etc).
- Converting values that are text descriptions of numeric values (i.e. $123 million) to actual numeric values (i.e. 123000000) which are usable for analysis.
- Extracting and cleaning values for dates
- Fill in the missing values by the various options supported by Open refine
- Removing duplicate rows
- Using a scatter plot to visualize relationships between values in different columns
- Exporting cleaned data to Excel

For each of the relevant operations for your data set, take screenshots of the various steps you have followed to obtain the final preprocessed data. Create a PDF file with suitable descriptions for the individual steps and include the screenshots for each relevant step. At the beginning of your PDF file, give a brief description about the data set with how many features, their types, how many rows in the data set, and details of class label attribute.

2: [3 Marks] Repeat the above problem with WEKA and compare their results where ever possible. Submit a separate PDF file with the required details including screen shots. Do we have any operations not supported by WEKA which is possible with *Open refine* and vice versa? Give details.

3: [4 Marks] The following link gives a data set extracted from a survey consisting of 9409 questionnaires containing 502 questions; those were filled out by shopping mall customers in the San Francisco Bay area (Source: Impact Resources, Inc., Columbus, OH (1987)). https://sci2s.ugr.es/keel/dataset/data/missing/marketing.zip. Download the full data set. A detailed description of the data set can be found here:

https://web.stanford.edu/~hastie/ElemStatLearn/datasets/marketing.info.txt

3.1 Preprocess missing values with the various options provided by WEKA

3.2 Demonstrate Attribute Filter option of WEKA

3.3 How Discretization can be done with WEKA for the above data.

3.4 Illustrate various Visualization techniques supported by WEKA for the above data

3.5 Show how data transformation is supported by WEKA.

3.6 Demonstrate *attribute selection* feature of WEKA for the above data.

Prepare your PDF file as mentioned above and submit the same.