BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
Work Integrated Learning Programme
Data Mining IS ZC415

EC-1 (Lab Component) **Weightage 10%**

Due date of submission: 22/April/2017

This evaluation component comprises two lab activities. Use either scikit-learn or Weka to complete this assignment. In case you think more information is required to solve any problem, make a suitable assumption and proceed.

1. Please download lab.zip from the course page and extract it. There is a file **dataset.csv.** You will use it for this problem. First field of this file is binary label 1/0. Remaining 17 fields are features. Apply the following classification algorithms on the dataset.
   - Naïve Bayes
   - Random Forest
   - AdaBoost
   - SVM (linear)

   Use 10-fold cross validation and experiment with various parameters of the classifiers to achieve the best performance out of them.
   To submit: **Precision and Recall of all the classifiers on the given dataset**. [6]

2. There are files input_file_1 and input_file_2 in lab.zip. Run the following Unix/Linux command to create a single file out of these:
   cat input_file_1 input_file_2 > input_file

   You need to use **input_file** to complete the following problem. The file contains partly pre-processed social media messages in English over a period of last year's leap day (UCT).

   The messages in input_file are roughly sorted in time and there are one ore multiple messages every minute except in some cases. In such cases, use liner regression of the past 2 minutes of the missing minute and the next 2 minutes of received messages to fill the missing text of the minute. You need to sort the lines in time order and fill the following text as many times as the number you get from regression:
   "missing text here"
   For example:
   …
   Mon Feb 29 08:21:07 leonardo dicaprio deserved the award after many years of awesome movies and no oscar
   Mon Feb 29 08:22:45 leonardo dicaprio's performance has won him an oscar.
   Mon Feb 29 08:23:56 so, i can't compare myself of not getting girlfriend with leonardo dicaprio not get oscar.
   Mon Feb 29 08:24:19 ladies and gentlemen....simply and elegant leonardo dicaprio
   Mon Feb 29 08:26:02 the amount of leonardo dicaprio tweets is too high rn
   Mon Feb 29 08:26:02 leonardo di caprio wins an oscar. i can die happy now that everyone will shut up about him not winning one!
   Mon Feb 29 08:27:06 so happy for leonardo dicaprio. looks like his career is finally taking off.
   Mon Feb 29 08:27:11 finally leonardo dicaprio! the crowd were cheering for you.
   Mon Feb 29 08:27:27 and you finally received the award, congratulations leonardo!

Mon Feb 29 08:27:52 congratulations to leonardo for finally winning an oscar.
Mon Feb 29 08:28:09 so buzzing for leonardo di caprio!
Mon Feb 29 08:28:30 i liked a video best actor oscars 2016 leonardo dicaprio full speech hd
…

In the above snippet, messages for the minute Mon Feb 29 08:25 are missing, so you will use regression on minutes from Mon Feb 29 08:23 - Mon Feb 29 08:27. The updated text would now be:

…
Mon Feb 29 08:21:07 leonardo dicaprio deserved the award after many years of awesome movies and no oscar
Mon Feb 29 08:22:45 leonardo dicaprio's performance has won him an oscar.
Mon Feb 29 08:23:56 so, i can't compare myself of not getting girlfriend with leonardo dicaprio not get oscar.
Mon Feb 29 08:24:19 ladies and gentlemen....simply and elegant leonardo dicaprio
Mon Feb 29 08:24:19 ladies and gentlemen....simply and elegant leonardo dicaprio
Mon Feb 29 08:25:00 missing text here
Mon Feb 29 08:25:00 missing text here
Mon Feb 29 08:26:02 leonardo di caprio wins an oscar. i can die happy now that everyone will shut up about him not winning one!
Mon Feb 29 08:27:06 so happy for leonardo dicaprio. looks like his career is finally taking off.
Mon Feb 29 08:27:11 finally leonardo dicaprio! the crowd were cheering for you.
Mon Feb 29 08:27:27 and you finally received the award, congratulations leonardo!
Mon Feb 29 08:27:52 congratulations to leonardo for finally winning an oscar.
Mon Feb 29 08:28:09 so buzzing for leonardo di caprio!
Mon Feb 29 08:28:30 i liked a video best actor oscars 2016 leonardo dicaprio full speech hd
…

In the input file, there may be more than one continuously missing minutes. You need to fill them all. For simplicity, as has been done above, filled seconds can be put as **00.**
To submit: the **sorted file with filled missing values**. [4]

<u>You need to upload a zipped file having your ID and name as filename. It should contain all the solutions. The location to upload this file will be communicated later.</u>