BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
K. K. BIRLA GOA CAMPUS

First Semester 2018-19
Data Mining (CS F415)
Bonus Assignment
Malware Detection System

Maximum Marks: +10
If Assignment not submitted: 0 Marks
Minimum Marks: -10

Start Date: 11/11/2018
End Date: 21/11/2018 **End Time: 5pm**

Answer the following questions: (in Report.pdf)

1. Propose a model for Malware Detection using Machine Learning Algorithm(s) (you can use whatever is taught in the class including Classification / Clustering/combination etc. etc.).
   (Do design multiple models and compare them with proper justification)
2. Draw the KDD process diagram for the above model.
3. Discuss your insight about the data and the model etc.

Dataset Details:

Number of opcodes: 1808
Number of Benign files: 2709 (opcode_frequency_benign.csv)
Number of Malware files: 4060 (opcode_frequency_malware.csv)

Assignment Submission Format:

A zip file consisting of the followings only:

- Portable source code (jupyter notebook):
  - Must contain all required packages/libraries.
  - Path for any required file(s) should not be local to your machine
  - Instructor should be able to run your code after direct download.
- Source Code (jupyter notebook pdf version)
  - Should contain all the intermediate steps + results to reach the conclusion
- README.txt
  - Step by step instructions to run your code.
  - ~~Download package 1, download xyz.jar, install MySQL~~
- Report in PDF format (max 3 pages. 11pt. Times New Roman.)
  - Insights, inferences, results and conclusions drawn from the assignment.
  - No source code or figures in this PDF
  - Proper references to the source code and figures.
- Figures (depends on the type of the assignment)
  - Self-explanatory caption to the figures. ~~1.jpg, q1.jpg, abc.jpg~~
- DO NOT UPLOAD THE DATASET

Assignment Submission Policy:

- Submission accepted through **Photon only.**
- No assignment will be accepted by **email or after the deadline**.

Plagiarism: Plagiarism will be checked for every submission with Turnitin.

- The rule is very simple
- If **(Plagiarism % from Turnitin Report) > 30**
    - Will be awarded **"Component Maximum Marks * -1"**