Midterm Examination

Datetime: 15:00, 13/04/2020 - 23:59 20/04/2020

**Open book;** Laptops/PCs/PDAs **are allowed**

|  |  |
| --- | --- |
| **SUBJECT: Introduction to Data Science (ITDS18IU11)** | |
| Approval by the School of CSE  Signature  Full name: | Lecturer:  Signature    Full name: Nguyen Thi Thanh Sang |
| Examiner  Signature  Full name: | Score |
| **STUDENT INFO** | | |
| **Student name:**  **Student ID:** | | |

# INSTRUCTIONS: the total of point is 100 (equivalent to 20% of the course)

1. *Purpose:*

* Test your knowledge about Data Science and the skill needed to be a data scientist
* Test your knowledge, skill in data exploration and classification

1. *Requirement:*

* Read carefully each question and answer it following the requirements.
* Describe your **STEP-BY-STEP** solutions clearly in the exam paper.
* Submit your answers via the Blackboard IU, name your file with your name and ID.

# QUESTIONS

Given the **glass identification data set**, which can be found at <http://archive.ics.uci.edu/ml/datasets/glass+identification>, including the following attributes:

% 1. Id number: 1 to 214

% 2. RI: refractive index

% 3. Na: Sodium (unit measurement: weight percent in corresponding oxide, as

% are attributes 4-10)

% 4. Mg: Magnesium

% 5. Al: Aluminum

% 6. Si: Silicon

% 7. K: Potassium

% 8. Ca: Calcium

% 9. Ba: Barium

% 10. Fe: Iron

% 11. Type of glass: (class attribute)

% -- 1 building\_windows\_float\_processed

% -- 2 building\_windows\_non\_float\_processed

% -- 3 vehicle\_windows\_float\_processed

% -- 4 vehicle\_windows\_non\_float\_processed (none in this database)

% -- 5 containers

% -- 6 tableware

% -- 7 headlamps

Attributes from 2 to 10 are numeric.

**Do the following tasks:**

**Q1. (40pts)** Select 20 sample records from the data set (file [glass.data](http://archive.ics.uci.edu/ml/machine-learning-databases/glass/glass.data)), draw boxplots for Attributes 2-10 using excel or R or Python. Note: please list the selected records and then present the boxplots.

**Q2. (30pts)** Compare and remark the five-number summaries from the above boxplots with the Summary Statistics of the corresponding attributes described in file [glass.names](http://archive.ics.uci.edu/ml/machine-learning-databases/glass/glass.names).

**Q3. (30pts)** Pick up one of the 20 sample records and find **TWO** nearest records using **Euclidean** and **Manhattan** distance metrics, using excel or R or Python. These comparisons are performed on the numeric attributes 2-10 only, exclude the *type* attribute. Note: please provide the selected records before listing the two nearest records, and explain how you get results.

**Q4. (10 bonus pts)** Remark the distribution of attribute values in the found records in Q.3, compared with the selected record.

*Note: if you use excel, please show the used formulae; if you use R/Python, please show your code.*

# YOUR ANSWERS

- END -