كلية علوم الحاسب و المعلومات

College of Computer and Information Sciences

CS372 Selected Topics

Big data and Data analytics

Computer Science Department

2021-2022 Second Term

Project Description: Mini Project for Selected Topics -1 (Big Data analytics)

Objectives:

This mini project aims to give students more practice in conducting an analytic project to some extent spanning its different phases through:

- developing an analytic plan
- developing suitable hypotheses
- preparing data for analytics
- selecting and applying suitable method(s) to analyze data
- Presenting and interpreting results

Description:

In this project, we assume that, as part of the data discovery phase, data sources have been identified and data sets have been collected. These data sets are listed below. Students are asked to choose one of these data sets to work on and handle various tasks by answering the following questions:

1. What kind of data analytic problem can you address based on the data set you selected? Propose some hypotheses.
2. How the various roles on the working team could be fulfilled
3. Is it necessary to perform data cleaning, conditioning and normalization why and how should it be done?(provide R code and results)
4. Perform some basic statistics to explore data. Provide R code and summary results and related plots.(feedback on results is much appreciated)
5. Select the analytic method(s) suitable to address your problem. Justify your choice.
6. Apply the method. Provide the R code
7. Present results with interpretation and discussion
8. What could be the impact of your key findings?

Deliverables:

- Scope and sequence of the final mini-project report and presentation will be determined on an individual/group basis upon selection of project topic.
- The project demonstration/presentation should be approximately 5-10 minutes with a few minutes for questions and answers. The project report should not exceed 8 pages

maximum, double-spaced, 12 pt., with 1-inch margins. Students should organize their report according to the following plan:

1. **Project title and name of student(s)**
2. **Abstract:** (100 words maximum) : A concise statement of the problem, questions, methodology, and conclusions from your project results.

3. **Summary of Data Discovery phase:**

   What is the problem?

   What are you hoping to find? What is the goal?

   What are your research questions?

   Are there similar projects?

4. **Summary of data Preparation phase:**

5. **Summary of Model Planning phase:**

6. **Summary of Model Building phase:** Including why did you choose this particular approach?

7. **Summary of project results communication:**

- What is/are the outcome(s) of your study?
- What were the problems you encountered?
- Visual output – maps, images, graphs, etc.

8. **Conclusion**

A summary of the problem, findings, and implications in a short paragraph.

- How did you interpret the results?
- What are your recommendations for future action?
- What other approaches might be possible?

- R script for the whole project must be delivered as well

التعليمات:

- **يتم ملئ الجدول بأسماء الطالبات فى كل مجموعة، يجب ان يتراوح عدد الطالبات فى المجموعة ما بين 5 بحد ادنى الى 6 بحد اقصى.**

- **يحتوى الجدول على مثال لملئ الجدول ، مع لينك للمواقع التى يمكن استخدمها لاختيار داتا للمشرع، مع العلم ان هذة المواقع غير ملزمة، اى يمكن للطالبة العمل على ملفات داتا اخرى غير التى بالموقع .**

Data sets Links :

Kaggle multiple data set: https://www.kaggle.com/datasets
Open ML Data sets: https://www.openml.org/search?type=data
UK data set: https://ckan.publishing.service.gov.uk/dataset
https://archive.ics.uci.edu/ml/datasets.php

#	TASK	IDEA	DATASET	NO. OF STUDENTS & NAMES
1	classification	Determine whether the mushroom is safe or poison	https://www.kaggle.com/uciml/mushroom-classification	#of students = 5 -student name1 - student name2 -…… -…..
2	Prediction	-whether smokers has higher chance to stroke -do men smoke more than women	https://www.kaggle.com/ukveteran/smoking-deaths-among-doctors	#of students = 6 -student name1 - student name2 -…… -…..
3				3
4				3
5				3
6				3