Data Science Fall 2024

Assignment 4: Data Science (written – 30%) and (presentation - 10%)

Due: BOTH written part of assignment and the presentation slides – 8th December 2024 by 8:00 pm ET by email (eleisa2@rpi.edu)

Submission method: email, please use the following file naming for electronic submission for documents: DataScience_A4_YOURGROUPNUMBER_written_part.xxx /PowerPoint/pdf..etc..

Late submission policy:

This assignment is due 12/08/2024 by 8:00 pm ET. If your group submission is more than 2 days late, it is likely that you will not have your grade for this assignment included in your final grade before they need to be submitted. 20% of score deducted each late day.

Note: Your report for this assignment should be the result of group work. All members of the group will receive the same grade. Take care to avoid plagiarism ("copying"), include all web resources, text, and class presentations. You are expected to work within a group setting contributing equally and with complementary skills and are encouraged to discuss your ideas and the tasks for this assignment with other student the group with other groups in the class.

General Assignment: Working with someone else's data related to the class project. The group chooses a data science investigation, finds, accesses, analyzes and presents/visualize the (more than one set/type of) data and manages the resulting products. The weighting score for each question is included below. Please use the question numbering below for your written assignment.

Written submission: Font size 12, New Times Roman, Double Spaced. Include each group members full name and email address at the top of the written document.

1. Abstract, Introduction, Literature Review and Project Workflow (5%):

a) Write an Abstract 300 – 500 words (1 %).

b) Introduction: write an introduction describing the project description, background and project goals (Minimum 1.5 page) (1.5%).

c) Literature Review: Include minimum 2-3 research papers that are relevant to your project work. Summarize the papers and discuss how they are relevant. (Minimum 1 page) (1.5%).

d) Project Workflow Diagram: Describe the project workflow by using a workflow diagram that shows the steps of the project. Explain each step of the workflow using 1-2 minimum sentences. (1 %)

2. Data Description and Methodology (5%):

Choose an investigation and identify pre-existing source of data that can address the data science project goal. Describe data description, data format(s) and any data cleaning activity conducted as a part of the project.

a) Choose, and state, the goal and reasons why the datasets were chosen and how they were found and managed, Min 8 sentences. (2.5%)

b) Document and discuss the data formats and any metadata standards/ conventions in use, and the method(s) of dataset discovery and access and how they helped or hindered the process, Min 5 sentences. (2.5%)

3. Data Analysis and Explore the statistical aspects of your datasets (10%)

a) Develop and state two particular questions/hypotheses related to the goal of the investigation and that can be answered using the datasets under consideration. Design an analysis study (preliminary, full and post) to answer the project questions and document the analysis design that helps to answer the questions/hypotheses, Min 3-4 sentences per question/hypothesis (5%)

b) Conduct Exploratory Data Analysis to show the distribution of data using histograms for quantitative data and barplots for qualitative (categorical) data. Use boxplot to show minimum, maximum, median, quartiles and any outliers that are in your data. Conduct any summary statistics for some chosen variables that are important in your data.

(5%)

4. Model Development and Application of models (5%):

a) Implement minimum two different types of models that support the two questions/hypotheses you mentioned in Q3. What types of models you used to describe the data (regression, classification, clustering, etc.), patterns/ trends you found, visual approaches that helped you choose models for the questions/hypotheses mentioned in Q3 and describe the variables (type/ number) in the models you decided to use, other parameter choices or settings for the models (e.g. distance metrics in clustering algorithms, or a type of kernels, etc.). (5%)

b) Apply the models to assess model performance (i.e. predict). Discuss the confidence in your results including any statistic measures relevant to the models you used (example:

Adjusted R-Squared value in Multivariate Regression or ideal number of clusters using the elbow method in KMeans clustering ect...).

Discuss howyou validated your models and performed any optimization (give details). (2%) Minimum 3 pages text + graphics for Question 4.

5. Conclusions and Discussion (5%):

Describe your conclusions; interpret the results, predictions you made using the models and project outcome and their characteristics, and a give summary of what changed as you went through the project (data, analysis, model choices, etc.), what you would do next, or do differently in a subsequent exploration. Provide a description of the choices of tools/methods used or a description of any code or scripts written during data format conversion (example: HDF5 to CSV etc.,) and describe how your results were stored and managed. Submit your code to course GitHub repository for evaluation. Include a

Reference section after the Conclusion.

Minimum 1 page text + graphics (optional).

References – websites, papers, packages, data refs, etc. You need to properly cite your references in the body of the document and should be included those properly cited reference at the end. You can choose any citation format (Chicago, MLA, etc.,) but be consistent when you cite your references (There is no specific citation format, just be consistent). Make sure to include the Figures/Graphs (Graphical Representations) with figure numbers in your written part and poster as well. The final written document should be a minimum of 10 pages (but can be more).

Include your code scripts! (e.g. codes in a zipped folder with the written part of the assignment and upload the zipped folder to LMS) and also include the GitHub URL that contains the code at the bottom of the conclusion section.

6. Oral Presentation (10%). Presentations will take place during the class. Please submit your slides by email using the same naming scheme mentioned above.

a). Title (followed by group members names) and Project name.

b). Problem area – what you wanted to explore/ solve/predict and why.

c). The data – where it came from, why it was applicable and the preliminary

assessments you made. Include the Questions/Hypotheses. Include some URLs for

data sources (or generally include the URL/website where you obtained the data).

d). Discuss your Exploratory Data Analytics (EDA) as well.

How you conducted your analysis: distribution, pattern/ relationship and model construction. What techniques did you use/ not use and why? What worked? What did not work?

e). What models you applied? How did you apply the model?

Explain any possible uncertainties in the data or models

f). Conclusion: What is the outcome/results. What did you predict and what decisions

(prescriptions) were possible. What was the outcome, conclusions?

THE END