Final project, Introduction to data science, Spring 2024

The final project is your opportunity to practice what you learned in the course, from stating the question to analyzing the data. The work on the final project consists of the following two steps:

1. Submitting a proposal for the project – until July 31st
2. Sending the final project report – until September 30th

The project replaces the exam for the course, so the amount of work I assume you will invest is ~3 days per person. You do not have to use everything you learned in the course, but you should use a significant portion of it. You are welcome to use new things (clustering algorithms we did not learn, for example), in which case you should explain the general idea of the tools you used and why you used them.

Project proposal

Your proposal should include the following information:

1. **Research question**: what exactly do you plan to investigate? The question should be well-stated according to what we learned in the course. Avoid vague questions (e.g., “Who is the best player in the NBA”), simple ones (“how many shows does Netflix release in each category”), etc.
2. **Dataset**: please provide a link to the dataset you will use, or if the dataset is not available online, send it with the proposal. Describe the dataset – how many data points, how many features, what are the relevant columns for answering your question, what is the target column (if relevant, e.g., in prediction), and anything important for your research.
3. **Methods**: What are the steps in answering your questions, and how will you do them? For example:
    a. Start with exploratory data analysis, in which I will look for missing values and columns that do not look right by plotting the data and comparing the min/max values to the mean and standard deviation.
    b. Look for correlations between different variables and the target variable.
    c. Split the data into training, testing, and validation sets (ratio: 50%-30%-20%, respectively)
    d. Create a linear regression model, consider its R2 and significance, and look for the most contributing variables
    e. Turn the target column into a categorical variable according to (…some criteria), apply random forest, test the model’s success and investigate the most contributing variables.

    Of course, your description should be relevant to your research question.

Things can (and most likely will) change once you implement the project! The research proposal is important to ensure that you are asking an answerable question and know how to start your research.

Final project

After you finish your project, you should submit a zipped file (through Moodle) that contains the following parts:

1. Project report
2. The code you implemented
3. Anything else needed to understand the project (figure, tables, etc.)

The final report should include all the information I need to know to understand what you did and how you reached your conclusions. This includes:

1. **Background**: a short description of the topic and all the relevant information
2. **Question**: state the question you answered. If you made changes to the research proposal, please explain what they were and why you did them.
3. **Methods**: a general description of the steps you took to answer the question. Specify packages, parameters, etc. you used
4. **Results**: coherently describe the results. Include figures and tables unless they are too large, in which case add them separately.
5. **Discussion**: discuss your results and how they answer the question. What did you find, and what are the implications (if any)? Was your approach appropriate, and what would you do differently? Are there still unknown things about the question, and how would you address them? If you could not answer the question (this can happen), please explain why and what else needs to be done to answer it (if possible). You are welcome to include other thoughts you may have about the topic.

Notes on using ChatGPT and other AI

- You can use ChatGPT to help you phrase your report, write code, propose analyses etc.
- You should state clearly in your report how ChatGPT was used and what prompt you used. I will consider code and text that are clearly not yours and that do not have prompt information as cheating.
- I expect that you understand and be able to explain everything that is written in your report and code
- I will reduce points aggressively to anyone copying and pasting text from ChatGPT whose aim is to fill space in your report. Your report must only include text that is relevant to your project.

Comments

- Your research proposal and final reports do not have to be wordy! Please make sure the relevant information is included so I can understand what you did and your reasoning.
- Avoid looking at code on the internet that analyses your data. Your code should be different from anything that others have done.