DATA 221 Final Project

Introduction

The goal of the project is to explore a number of datasets that may be associated with political instability in the U.S. The data was taken from the Seshat Databank under Creative Commons Attribution Non-Commercial (CC By-NC SA) licensing. You are asked to follow the steps and/or answer the questions below. It is permitted to work on the project in pairs or individually (1-2 students per project). Throughout the project, use best practices.

Questions

1. **Wrangle** the data:
    - Read the (short) code book.
    - Numerical data need to be uploaded, interpolated, and properly saved. For the purpose of this project, **interpolate each variable such that you obtain one point per year** (within the range of available data).
    - Calculate (and then interpolate) the political instability index.
    - Display the DataFrame with all of the columns and the interpolated data for the years 1901-1910.
2. Perform some **exploratory data analysis**.
    - For your main findings, include appropriate visualizations.
    - Summarize your main findings (most interesting/insightful conclusions) in a short paragraph. You may include a figure if you find it helpful.
3. Find the **best regressor** that would predict the instability index from the various predictors.
    - To be clear, you are asked to compare a limited set of regressors of your choice – not to identify the theoretically optimal one.
    - Explain your modeling choices.
    - Interpret any evaluation metrics you use.
    - Summarize your conclusions in a short paragraph, i.e., the most interesting conclusion(s), the model that produced it, and how the model was chosen. You may include a figure if you find it helpful.
4. Find the **best dimensionality reduction for regression**.
    - You can restrict this part to reducing the data to two dimensions, to three dimensions, or explore both options.
    - You can test your variables using the best regressor found in the previous section or a small number of regressors (2-3 models at most).
    - Explain modeling choices and evaluation metrics.
    - Summarize your conclusions in a short paragraph, i.e., the most interesting conclusion(s), the model that produced it, and how the model was chosen. You may include a figure if you find it helpful.

5. Find the **best dimensionality reduction for unsupervised classification**.
    - Use only the predictor columns and not the outcome (instability) for classification.
    - You can restrict this part to reducing the data to two dimensions, to three dimensions, or explore both options.
    - You can test your variables using k-means or a small number of classifiers (2-3 models at most).
    - Explain modeling choices and evaluation metrics.
    - Summarize your conclusions in a short paragraph, i.e., the most interesting conclusion(s), the model that produced it, and how the model was chosen. You may include a figure if you find it helpful.
6. Briefly **explore the clusters of instability** scores.
    - Consider the cluster labels from the best clustering scheme from previous section or from clustering using all/most of the original features. Apply it to the corresponding records of the outcome column (instability).
    - Create a visualization of the results.
    - Summarize your conclusions. To be clear, the summary can be very short, and may be that the clusters do not exhibit any discernable or interpretable pattern. You may include a figure if you find it helpful.
7. Consider a real life modeling/prediction problem where you try a large number (100? 1000?) different models, examine their performances, and select the one that scores best on your performance metric of choice. Briefly discuss the potential disadvantage (or potential danger) of such and approach and how you might go about mitigating it.

Grading

- Best practices as discussed in class. These include concise and clear explanations of modeling choices and good visualizations (this is not an exhaustive list).
- Quality of reporting.
- Quality of the analyses and the results.
- How challenging the analyses are in terms of innovative thinking and exploration of methods and ideas.

Statements on the Use of AI Tools:
Students are only allowed to use AI tools, such as ChatGPT or Dall-E, when advance permission is given by the instructor. Students must submit a written request with an explanation of how they will use a particular tool in their assignment. Students are not permitted to use these tools until permission is granted in writing.
Unless given permission, each team is expected to complete the assignment without substantive assistance from others, including AI tools. If you are unclear if something is an AI tool, please check with your instructor. Unauthorized use of AI tools for any purposes in this assignment will violate the Universitys academic integrity policy.