Final Project Instructions

Course: Principles and Techniques of Data Science
Course Examiner: Hadi Fanaee

Step 1: Form a Group

- Form a team of two. If a single student remains ungrouped, they may join an existing pair to form a group of three. All other groups must consist of exactly two members.

Step 2: Dataset Selection

- Select a dataset from the UCI Machine Learning Repository.

Step 3: Define Your Research Questions

- Propose two data science questions:
    1. An **exploratory data science question** that uses a technique covered in lectures 1-8. Be sure to reference the lecture and slide number or relevant scientific paper if the method wasn't covered in class.
    2. A **predictive data science question**, where you'll build a predictive model (either regression or classification) using an iterative improvement process.

Note: Your questions should require a level of sophistication beyond basic statistics (e.g., they should not be answerable by calculating simple averages).

- **Submit your dataset choice, name, and URL on this Google Sheet.**
- **Deadline:** 21 Nov (Questions cannot be modified after this date.)

Step 4: Proposal Approval

- The course examiner will review and approve your proposal by 23 Nov.

Step 5: Project Execution

5-A: Data Preparation and Exploration

1. Clean the dataset, address missing values, and identify anomalies using unsupervised methods. If anomalies are identified, justify them using external sources when possible.
2. Summarize the dataset using descriptive statistics.
3. Apply appropriate dimension reduction and visualization techniques to reveal insights and trends.

5-B: Answer the Exploratory Question

- Select and apply the most suitable algorithm from lectures 1-8 to answer your question. Justify your choice, document any transformations or hyperparameter settings, and ensure the analysis generates meaningful and non-trivial insights. Provide a concrete recommendation based on your findings.

5-C: Answer the Predictive Question

1. Split your data into training (80%) and testing (20%) sets, with an additional 20% of the training data reserved for validation.
2. Build and improve your model through at least 5 iterations. Start with a simple model (e.g., linear or logistic regression), and improve it by modifying datasets, algorithms, or hyperparameters. Document each change.
3. Report model performance:
    - Regression: Mean Absolute Error (MAE).
    - Classification: Accuracy, and, for imbalanced classes, balanced accuracy or ROC AUC.

Avoid Data Leakage: Perform all transformations (e.g., normalization) only on the training set, then apply normalization parameters to the test set.

Example Iterations (for illustration):

- **Iteration 1:** Logistic Regression on original data
    - *Accuracy on Train:* 95.59%, *Test:* 93.63%
- **Iteration 2:** Logistic Regression with modified data (e.g., removing features, applying PCA, adding new features)
    - *Accuracy on Train:* 96.56%, *Test:* 94.51%
- **Further Iterations:** Document changes in algorithm, hyperparameter tuning, feature engineering, etc., and report performance improvements.

Document Use of LLMs (e.g., ChatGPT):

- Attach a screenshot of prompts and responses if using LLMs during your project. If not, submit a declaration signed by all group members stating no LLMs were used. Misrepresentation will be investigated very carefully and can be treated as cheating.

Note: Each iteration must involve substantial changes for performance improvement on the test set. Slight improvements over iterations (like very low numbers) is not acceptable.

Step 6: Presentation and Report

Report Requirements:

- Format: Max 10 pages, Calibri 11 pt font.
- Content: 7 pages for unsupervised learning (lectures 1-8) and 3 pages for supervised learning (lectures 9-13).
- Attach supplementary materials and references as necessary.

Presentation:

- 15 slides maximum. Summarize key findings of the written report.
- Presentation duration: 20 minutes (15 minutes for presentation, 5 minutes for Q&A).

Deadlines:

- **Dataset and Topic Selection:** 21 Nov, 23:59
- **Approval:** 23 Nov, 23:59
- **Final Topic Lock:** 1 Dec, 23:59
- **Report Submission:** 15 Dec, 23:59
- **Presentation Date:** 16 Dec, 08:00-12:00

Report Submission

Submit a single ZIP file including:

- Report (PDF, max 10 pages).
- Presentation slides (PDF).
- Screenshots from LLMs or signed declaration of non-use.
- Jupyter notebook demonstrating results, especially the 5 iterations of model improvement.

Evaluation Criteria:

1. **Data Cleaning and Preparation** - Clarity and thoroughness of tasks.
2. **Creativity and Relevance of Questions** - Creativity and importance of questions posed.
3. **Techniques and Tools** - Justification and alignment of techniques with course content.
4. **Insights from Exploratory Analysis** - Depth and significance of insights generated.
5. **Performance Improvement** - Creativity in improving predictive model performance.
6. **Presentation and Report Quality** - Transparency, justification of choices, adherence to guidelines.
7. **Completeness of Required Components** - Report on LLM use, Jupyter notebook, page limit, font size, slide number.

Good luck with your project!