ISTA 421: Introductory Machine Learning

Machine Learning Final Project

Instructor:
Salena Torres Ashton | salena@arizona.edu
448 Harvill Building, College of Information, University of Arizona

Class Hours: 11:00 am - 12:15 pm, Tuesdays & Thursdays
Class Location: 332B Harvill, in-person only
Office Hours: 448 Harvill Building, Tuesdays 9:00 - 10:15, or by appointment.

Machine Learning Final Project Rubric (220 Points Total)

A. Dataset Submission by 14 November 2024 (10 Points)

Detailed description of the dataset, including sources, qualitative and quantitative features, number of observations (at least 40), and why the dataset is suitable for your project. Homework #3 dataset is okay but you cannot use an ISLP dataset.

If I already cleared your homework #3 dataset, you do not need to have it cleared again. Otherwise, I need to clear your dataset by 14 November 2024. The reason you'll want the data set cleared is to be sure that the work you do is suitable for this particular course.

Do not email the dataset to me or push onto GitHub!

You must submit the actual dataset here and include the URL to the dataset.

B. Proposal Submitted by 14 November 2024 (10 Points)

- Problem and research question
- Significance (why we care)
- Method: what skills have you learned from this course to answer your research question? How will you answer your question? What's your hypothesis?
- 1 page max –informal– email is fine
- Do not use any work you have done for other courses
- Do not plagiarize others' work

C. Final Project Check In and Outline due 7 December 2024 at 12:00 pm Tucson Time (Outline consisting of 22 lines... see D2L for rubric... 50 Points)

D. Final Project due 17 December 2024 at 12:00 pm Tucson Time (Formal exam schedule for the university; 7 - 10 pages, 12 pt font, 1" margins on all sides, single-spaced lines. 200 Points)

1. Problem Formulation and Significance (25 Points)

- **Problem Statement (10 Points):** State a clear and concise description of the problem. Include the machine learning task and why it's important.
- **Significance (10 Points):** Explain why this problem matters in your chosen field and how solving it contributes to real-world applications.
- **Dataset Description (5 Points):** Detailed description of the dataset, including sources, qualitative and quantitative features, number of observations (at least 40), and why the dataset is suitable for your project.

2. Exploratory Data Analysis (15 Points)

- **Data Exploration (5 Points):** Summary statistics and visualizations that explore both qualitative and quantitative features. Discuss any relationships or patterns found.
- **Handling of Missing/Imbalanced Data (5 Points):** Explain how you handled missing, normalized, or imbalanced data.
- **Data Visualization (5 Points):** Appropriate visualizations (e.g., histograms, boxplots, scatterplots) and meaningful interpretation of key data patterns and trends.

3. Model Selection, Application, and Evaluation (45 Points)

- **Justify Model Choice (15 Points):** Explain the model(s) chosen with a clear rationale based on dataset characteristics and your research question/ problem statement. Model choice needs to reflect what was learned in class (linear models, tree-based methods, etc.).
- **Method Applications (15 Points):** Correctly implement your model(s) and include appropriate hyperparameter tuning. If you choose a non-parametric model, you must include the process of choosing your metrics (e.g., how you chose k for KNN or clustering).
  - You must write your own code! No tutorials or cut/paste from tutorials.
  - You are allowed to use generative AI under the following conditions:
    * You must include all discussions in your references and appendices. Be sure to label all code that is generative in your code comments, from start to finish for any and all sections.
    * You must share each discussion with me in an email. Include the link.
    * You are responsible for any mistakes that the generative AI creates. If it makes a mistake, I take points from your grade. Mistake areas include but are not limited to: computation, "valid arguments" like a philosopher or computer scientist would define, appropriate context, mathematically sound, correct "formal semantics" like a computational linguist or a computer scientist would define, and getting the correct answers for the correct reasons.
    * Again, if the generative AI makes mistakes, you are held accountable for it. It is worth your time to understand the formal meanings of "mistakes", as listed above. Even if your code is correct, if your written interpretation is too general, ambiguous, does not match your code, or has a mistake, as described above, you are held responsible and points will be taken off.
    * You must read a paper about generative AI and write a one-page response. This is due December 1, 2024. This paper will be printed and handed out in class. I will also post this paper on D2L. FAILURE TO GIVE ME YOUR REPONSE ON THIS PAPER (by December 1, 2024) REVOKES YOUR PERMISSION TO USE GENERATIVE AI ON THE FINAL PROJECT. NO EXCEPTIONS AT ALL.
  - You may use a package to verify your code, but your own model must be written from scratch.
  - You are allowed to cut and paste the code you wrote for your homework. You are not allowed to cut/paste from tutorials, demos, or generative AI.
- **Model Evaluation (15 Points):** Evaluate your model performance using relevant metrics (e.g., accuracy, precision, recall, AUC, RMSE). Compare results across models or hyperparameters and discuss the trade-offs involved.

4. Results, Conclusions, and Real-World Implications (30 Points)

- **Results Presentation (10 Points):** Present your results in a clear manner using tables, plots, and key statistics. Results should be tied back to the problem statement and dataset.
- **Conclusions (10 Points):** Interpret the results, discussing key takeaways and any limitations. Discussion should be thoughtful and relate to the problem objectives.
- **Workforce/Graduate School Preparation (10 Points):** Explain how your project developed or sharpened your skills that are relevant for future goals, whether for entering the workforce or applying to graduate school. Specific skills (e.g., data analysis, programming, optimization, focusing on the question at hand, communication, etc…) should be discussed.

5. Lightning Talk in Class (40 Points) 10 December 2024

- **3 Minutes Long (5 Points):** Fewer points if less than 2 minutes or more than 3 minutes.
- **Structure of Talk (20 Points):** Problem statement, significance, dataset, model selection, evaluation, results, conclusion, and how it will help you with your research/work goals.
- **Paying attention to Others' Lightning Talks (15 Points):** Not on devices, making appropriate eye contact, learning from another's research.

6. Code Implementation and Documentation (35 Points). All code must be pushed onto your Github repository.

- **Code Completeness and Accuracy (5 Points):** Code should run without errors and complete the project tasks. The project pipeline (data processing, model implementation, etc.) should be fully developed.
- **Code Documentation (7 Points):** Be sure that you have an introduction in your script. Document each of your functions. Add line comments to tricky code.
- **README.md (5 Points)**: Give instructions for running your code with a README.md file!
- **Reproducibility of Visuals (5 Points):** Any figures or tables that you present in your paper need to be generated by this code.
- **Dataset Submitted on D2L, not GitHub (5 Points)**: So I can run your code.
- **Code Readability and Organization (8 Points):** Code should be well-structured, with clear variable names and appropriate comments. Organization should be conducive to readability and understanding.
