In this assignment, you are given a dataset of wine ratings, as assigned by human tasters, along with other pertinent characteristics of each wine. Your task is to build a regression model that can forecast the rating for novel, unseen wines. Begin by downloading the archive file provided on Moodle — this archive contains three files: two CSV files (you can just read them like plain text files, or open them in a spreadsheet program like Excel) with red wine and white wine data, and a third text file that describes the format of the data.

You should try both for performing the regression, and present results from both methods:

- Use an off-the-shelf regression solver — for example the `scikit-learn` library in Python. You can find excellent online tutorials on how to use this.
- Implement stochastic gradient descent yourself, as described in class.

The Write-Up

Here's the overarching writing rule for this course: *you need to be sufficiently precise with your writing and include enough detail that a competent reader could reproduce your results.* You will submit a PDF. A paper from AAAI, a notable artificial intelligence conference, is provided for guidance. You may use any word processing platform you prefer for your report. AAAI provides a LaTeXtemplate for those interested. Your writeup must also include an ethics section in line with the NeurIPS (another notable AI conference) guidelines: `https://neurips.cc/public/EthicsGuidelines`.

Here are some additional things to address in your report, in no particular order. This is *not* meant to be an exhaustive list.

- What preprocessing did you perform on the data? Did you perform any exploratory data analysis? Generate any plots or charts? Describe these, along with any relevant findings, in your report.
- What regression models did you build? How do they compare in terms of performance? What was the best performing model, and how did it do? Optionally, you can go beyond the methods we've seen in class and try other linear regression variants — if you go this route, you should describe how your chosen algorithm works (and don't forget to include citations!).
- What was your model-building and tuning regime? How did you address overfitting? How did you make hyperparameter choices?

Deliverables

Your deliverables for this assignment are

- a PDF write-up,
- the code you wrote for the project, and
- a PDF of your data exploration presentation.

Everything should be in your GitHub repository.

Deadlines

- **Monday, Feb. 14, in class:** Slide presentation on data exploration results and planned approach.

- **Monday, Feb. 21, 11:55pm:** Final paper and code on GitHub.

Timetable

Here's how to budget your time over the next couple of weeks as you work on this project.

- **Week 1:** Explore the dataset, think about feature engineering, choose your modeling tools (Python + `scikit-learn` vs. your own implementation), build your first model. Outline potential biases and inequities that can result from this project.
- **Week 1:** Run more thorough experiments (hyperparameter tuning, further feature engineering, etc.), analyze your results and iterate, search the literature for related work on the problem, write your *Introduction* and *Background* sections, optionally meet with Dr. Kuchera to get advice/feedback (both on technical issues and on writing)
- **Week 2:** Complete experiments, take a step back and think about your report's narrative, write drafts of your *Experiments* and *Results* section, consult with Dr. Kuchera as appropriate
- **Week 2:** Wrap-up any pending experiments, write the *Conclusions* section and the abstract, revise and proof-read the entire report and submit it for peer review

Rubric

- Data exploration (presentation): /10
- Model choice/exploration: /10
- Code: /10
- Report:
  - Objective description: /5
  - Method description: /10
  - Analysis: /10
  - Results: /10
  - Ethics: /5
  - Writing: Technical Correctness: /5
- **Total: /70**