# Final Project Description

## 4780/6780 Fundamentals of Data Science

Kiril Kuzmin

September 19, 2025

In this project, you will take on the role of a real-world data scientist. Your goal is to identify meaningful questions, acquire appropriate datasets, and implement a data analytics solution in teams of 4 to 6 members (4 members is ideal size of the group). Below are the project phases and key details.

### Phase 1: Form a Group and Propose a Project

***Deadline: 11pm, Wednesday, October 15***

- ✓ Form a team of 4–6 members from your section. If you need help forming a team, let me know.
  - ✗ You cannot select members from another class.
  - ✓ Mixed teams of graduate and undergraduate members are encouraged.
- ✓ Choose a representative and select an analytics topic.
  - ✗ Avoid topics where the standard solution involves deep learning techniques.
  - ✓ Select a dataset (see suggestions below or find your own) and brainstorm a meaningful business problem based on it.
  - ✗ Avoid datasets that primarily consist of time series data.
- ✓ Suggested datasets:
  - UCI Machine Learning Datasets
  - DrivenData
  - FiveThirtyEight Datasets
  - Microsoft Research Open Data
  - Awesome Public Datasets
  - Harvard Dataverse

- GISAID
- data.world
- Kaggle Datasets
- AWS Open Data Registry
- OpenDataPhilly
- Zenodo
- Data.gov
- European Union Open Data Portal
- HealthData.gov
- NASA Climate Data
- World Bank Open Data

**Additional Search Options:**

- `https://datasetsearch.research.google.com/`
- `https://search.datacite.org/`
- `https://www.re3data.org/search`
- `https://guides.library.cmu.edu/machine-learning/datasets`
- `https://datahub.io/`

✓ The representative must email kkuzmin1@gsu.edu with:

- Subject: `4780/6780 TR Project Proposal:<Your group name>`
- CC all members.
- A brief (200–300 words) project summary outlining potential data sources and the problem.

✓ Project topics *must be approved by me* to ensure appropriate complexity.

## Phase 2: Business Understanding

✓ Define the business problem and select a suitable dataset.

✓ Your solution should involve a *supervised* learning model unless otherwise approved.

✓ Finalize the business problem, dataset, and supposed analytics solution.

### Phase 3: Data Understanding and Preparation

- ✓ Explore the dataset, handle missing values and outliers, and apply necessary transformations.
- ✓ Use at least two feature selection techniques to identify relevant features. Aim for 50 or fewer features, ideally 20 or fewer unless additional features improve performance significantly.
- ✓ For non-tabular data (such as time series or images), please consult me for guidance. However, it is best to avoid these cases, since handling such input goes well beyond the scope of our fundamentals class.

### Phase 4: Model Selection and Evaluation

- ✓ Train models using a train/test split and apply appropriate evaluation metrics.
- ✓ Implement custom evaluation metrics if necessary.
- ✓ Test a baseline model and at least three learning model categories:
  - **Information-based models:** Decision Trees
  - **Similarity-based models:** $k$-Nearest Neighbors
  - **Probability-based models:** Naive Bayes
  - **Error-based models:** Linear Regression, Logistic Regression, SVM, Random Forest
- ✓ You are encouraged to explore additional models not covered in class.
- ✓ Apply hyper-parameter optimization (e.g., grid search).
- ✓ Justify your recommended model based on performance, interpretability, efficiency, etc.

### Phase 5: Communicate Findings and Recommend Actions

- ✓ Analyze results, demonstrate relationships between features and target variable, and recommend actions *based on your findings.*
- ✓ Submit a final report and give a 12–15 minute group presentation.

## Deliverables

- ✓ **Processed datasets** (10 points): Submit raw and pre-processed datasets.
- ✓ **Project implementation** (35 points): Submit two Jupyter Notebooks for preprocessing (15 points) and modeling (20 points).
- ✓ **Presentation and discussion** (30 points): Present for 12–15 minutes and answer questions.
- ✓ **Final report** (25 points): Submit a detailed PDF report.

## Deadlines

- ✓ **11pm, October 15:** Submit group proposal.
- ✓ **November 21, December 2, December 4:** Presentations.
- ✓ **11pm, December 8:** Submit the final report.

**Suggestions**

- ✓ **Understand Your Data.** Analyze dataset behavior, trends, quality, and potential biases.
- ✓ **Start with Simple Model.** Begin with an explainable baseline model. Gradually experiment with features and more complex methods.
- ✓ **Prioritize Explainability.** Thorough validation improves results and insights.