Data Science Final Project

Capstone Analysis and Presentation

P.M.H

CS320/CS110: Fall 2025

Summary

This final project serves as a capstone for the entire course. You will be responsible for the complete data science lifecycle on a dataset **of your own choosing**. This includes sourcing and proposing a dataset, formulating a compelling hypothesis, performing all necessary data processing and exploratory analysis (EDA), building and evaluating appropriate predictive or analytical models, and presenting your findings in a formal PowerPoint and a live in-class presentation. This is a significant undertaking, estimated to require **approximately 25 hours of work**, and will test your cumulative skills in data manipulation, modeling, evaluation, and scientific communication.

Goals

The goals of this assignment are:

1. To identify a complex, real-world dataset and formulate a meaningful, testable hypothesis.
2. To demonstrate mastery in cleaning, managing, and preparing a dataset for advanced analysis.
3. To apply exploratory data analysis and visualization techniques to uncover deep insights and inform model development.
4. To select, implement, and evaluate appropriate machine learning models (e.g., classification, regression, clustering) to address your hypothesis.
5. To synthesize all semester-long concepts into a single, cohesive project.
6. To clearly and professionally communicate your methods, findings, and conclusions in both a written report (PowerPoint) and a live presentation.

Before You Begin: Comprehensive Review

Review Key Concepts

This project will draw on everything we have covered. Refresh your knowledge of:

- **Python Libraries:** Pandas (data manipulation), NumPy (numerical operations), Matplotlib/Seaborn (visualization), and **Scikit-learn** (modeling).
- **Data Sourcing:** Web scraping (e.g., BeautifulSoup) or API usage (e.g., requests) for data acquisition.

- **Data Cleaning:** Handling missing values, encoding categorical variables, feature scaling, and text processing.
- **Modeling Techniques:**
  - **Regression:** Linear, Polynomial, Ridge, Lasso.
  - **Classification:** Logistic Regression, k-NN, Decision Trees, Random Forest, SVM.
  - **Clustering:** K-Means, Hierarchical.
- **Model Evaluation:**
  - **Regression Metrics:** MAE, MSE, RMSE, R-squared.
  - **Classification Metrics:** Accuracy, Precision, Recall, F1-score, Confusion Matrix, ROC/AUC.
  - **Clustering Metrics:** Silhouette Score.
- **Advanced Topics:** Strategies for imbalanced data (SMOTE), feature engineering, and hyperparameter tuning (GridSearch).

Phase 1: Dataset Sourcing and Proposal

You must find your own dataset. The problem you choose to solve (regression, classification, etc.) will depend on the data you find.

- **Find a Dataset:** Look for a dataset that is interesting to you and non-trivial (i.e., not "iris" or "titanic"). It should be complex enough to require significant cleaning and analysis.
- **Good Sources:** Kaggle, UCI Machine Learning Repository, data.gov, Google Datasets, or any public-facing API.
- ***MANDATORY:*** You must submit a **1-paragraph proposal** via Canvas message by **[11/14/25]** outlining:
  1. A link to your dataset.
  2. The problem you intend to solve (e.g., "I will predict...").
  3. The type of analysis (e.g., "This is a classification problem.").
- I must approve your dataset and topic before you begin Phase 2.

Prepare Your Development Environment

- Set up your Python/Jupyter Notebook development environment (e.g., Anaconda, VS Code, Google Colab).
- Ensure all necessary libraries are installed, potentially including new ones for scraping or APIs.

Project Workflow: The Analysis Process

Phase 2: Hypothesis Formulation & Data Wrangling

- **Formulate a Hypothesis:** Based on your approved dataset, develop a clear, testable hypothesis. For example: *"Housing features (sq. ft., beds) will be more predictive of price than location-based features (zip code, school district)."* or *"User sentiment in text reviews is a significant predictor of a product's star rating."*

- **Data Cleaning and Integration:** Load the dataset(s). Perform all necessary cleaning (handling NaNs, correcting data types, encoding). If using multiple sources, integrate them.

Phase 3: Analysis and Modeling

- **Exploratory Data Analysis (EDA):** Generate descriptive statistics. Create meaningful visualizations to explore relationships between features and your target variable.
- **Model Building:** Perform the specific analysis required to test your hypothesis. This will involve splitting your data (train/test), building appropriate models, and evaluating their performance. appropriate, apply feature engineering and hyperparameter tuning.

Project Requirements

Your final submission will be a PowerPoint presentation **and** a live, in-class presentation.

Part 1: The Presentation (PowerPoint)

Your presentation must be structured to include the following slides/sections.

1. **Title Slide:** Project title, course, and your name.
2. **Introduction & Hypothesis:** Clearly state the problem you are addressing and why it is interesting. State your specific, testable hypothesis.
3. **Data Sourcing & Processing:** Describe where you got your data. Detail the steps you took to clean and prepare the data (missing values, encoding, feature engineering, etc.).
4. **Exploratory Data Analysis (EDA):** Showcase your most insightful EDA findings. Use visualizations (plots, graphs) to illustrate relationships between variables, especially their relationship with your target.
5. **Modeling & Analysis:**
   Implement **at least two different models** appropriate for your problem (e.g., two regression models, two classification models, etc.).
6. **Justify your model choices.** Why were they a good fit for your data and problem?
7. Evaluate **all models** using appropriate metrics (e.g., R-squared for regression, F1-score for classification).
8. Report and compare the performance of each model, discussing which performs best and why.
9. **Discussion & Conclusion:** Interpret your model results. Which factors were most important? Did the results support your initial hypothesis? Discuss the limitations of your analysis and suggest potential future work.

Part 2: The In-Class Presentation

- You will present your PowerPoint to the class.
- **Time Limit:** 10-12 minutes, followed by a 3-minute Q&A.

- **Goal:** This is a test of your ability to communicate complex technical findings clearly and concisely. Practice your timing. Do not just read your slides.

Important Notes:

- I am ***Looking for*** critical thinking, methodological justification, and a deep engagement with your chosen topic.
- This project is intended to be a 20-hour effort. Do not wait until the last minute.
- ***Extra Points - 10pts*** Extra points will be awarded for exceptional projects that demonstrate creativity and technical depth, such as:
  - Integrating a live API for your data.
  - Building a simple interactive dashboard (e.g., with Streamlit or Dash) to present your findings.
  - Applying a novel or highly complex modeling technique not covered in class (e.g., a neural network) and explaining it clearly.

Academic Integrity Policy

- **Cheating is strictly prohibited.** Copying directly from AI-generated content or any other published work without significant editing or proper citation will not be tolerated.
- Any instance of plagiarism or academic dishonesty will result in a score of zero for the assignment or project, and a report will be made to the school administration.
- Students are expected to submit their own original work.
- If you need help or clarification, seek assistance in a legitimate manner, and always ensure proper citation if you reference external sources.
- Remember, integrity is essential to your learning process and academic growth.
- Abiding by the rules helps maintain a fair and respectful academic environment for everyone.

Grading Rubric

Component	Expectations	Weight
Code & Data Processing	Dataset choice is non-trivial and approved. Clear, efficient, and well-commented code. Correctly loads, cleans, and prepares data for modeling.	25%
Analysis & Modeling	Appropriate and insightful use of EDA. Correct implementation and justification of at least two model types. Evaluation metrics are correct and well-explained.	30%
Discussion & Critical Thinking	Insightful interpretation of results. Thorough discussion of the findings in relation to the hypothesis. Acknowledges limitations and proposes future work.	25%
Presentation & Communication	PowerPoint slides are professional, well-organized, and clear. In-class presentation is well-paced, confident, and effectively communicates findings.	20%

Note: Points will be deducted for poor presentation quality (e.g., grammar, clarity, or structure issues) or for going significantly over/under the allotted presentation time.
Upload to Canvas the following:

- **Presentation File:** `LastName_FirstName_Final_CSXXX.pptx`
- **Code File:** `LastName_FirstName_Final_CSXXX.ipynb` (or `.py`)