CSCI 1070: Intro to computer science: Taming big data

Final project: training and evaluating models on real–world data sets
Group 2

Brief description

In this project you will use well–known data sets to train and evaluate kNN and simple linear regression. You will practise the skills of loading data, feature manipulation, exploratory data analysis, model training, evaluation, plotting and communicating your findings. Despite using just two of the models seen in class, this project shows many of the elements of the full process involved in solving data science problems.

Tasks and analysis

- Dataset: California Housing available here
- kNN Task: scale numerical features; apply kNN regression to predict median house value.
- Linear regression: simple linear regression predicting house value from a single key feature (e.g., average rooms).
- EDA and graphs: provide correlations,
- Output: data pre–processing, model performance metrics (RMSE, R^2), and interpretation.
- Presentation: summarise the process and your findings.

Deliverables

- A document with the following structure:
  1. Detailed analysis of the project objectives and problem statement. Here you should clearly explain what predictive or classification problem is being addressed, and how it relates to the dataset and scope. If any critical decisions have been made about the project that were not explicitly stated in the project description, include these as well.
  2. Detailed description of data handling and modelling components. Describe the datasets used, including sources and key attributes. Explain the pre–processing steps such as feature transformation, normalisation, and encoding necessary for kNN and linear regression models.
  3. Detailed description of exploratory data analysis and results. Present the exploratory data analysis process including key graphs (histograms, scatter plots, correlation matrices) and data insights relevant to modelling. Follow this with a clear explanation of model training, evaluation metrics (such as accuracy for classification and RMSE for regression), and comparison of model results. Include discussion on interpretation and implications.
  4. The full document must be well–structured, clear, and detailed, submitted as a PDF including all graphics and analysis.
- A copy of the presentation. Preliminary copies are OK, so long as they show an outline that will later be used. Please make sure the format is one of the following: Google Slides link, Power Point or PDF.

- The accompanying code. Make sure it is as readable as possible, provide comments and printouts of the results.

Evaluation

The project is assigned in lieu of the final exam, therefore it will count as 30% of the final mark for the course. The usual Banner rules apply to late submissions.

The score will be assigned as follows:

Data exploration and pre–processing	25%
Model performance and code organisation	35%
Analysis and reporting	20%
Presentation and communication	20%

Dates and deadlines

Deadline for the document: Sunday, 14 December 2025.

Deadline for the presentation file: Monday, 15 December 2025.

Presentations: Tuesday, 16 December 2025.

Important!

- As always, make sure to use the methods found in the `scratch` module rather than `scikit`, `pandas` or `numpy`. Feel free to reuse or recycle any Python code found on Canvas. If external sources are used, please provide references.
- You will use Canvas to submit both the document and the presentation file. **Compress them to a zip** file and submit them before the assigned deadlines.
- The file name for you submission **must** be as follows:
  CSCI1070-F25-FinalProject-Group##-Document.zip or CSCI1070-F25-FinalProject-Group##-Presentation.zip
  Example: CSCI1070-F25-FinalProject-Group1-Document.zip