

Project 2 (Report)

5/11/2023

68/70 Points

Attempt 1



In Progress

NEXT UP: Submit Assignment



View Feedback

Unlimited Attempts Allowed

▼ Details

Project 2

- Worth 15% of your grade
- Data set + group members may be decided at any time, **but you must first verify with me!** (Send an e-mail, discord message, or talk in person)
 - **Due date for presentation + writeup (3-5 pages):**
 - Thursday 5/11 2:30pm
- Grade will be assigned based on how well you handled the data set and utilized methods learned in the course, not based on your kaggle performance score.
- **Kaggle requires one submission, and only one student should submit.**
 - Display the submission page and results in your writeup.

Objective

- Pick a Kaggle data set that involves a specific **classification**, or **clustering** task.
 - Classification will involve a **submission** using an external test set
 - Clustering is more open-ended.

- Here is an excellent example to follow if you want to do clustering: <https://www.kaggle.com/code/micheldc55/mall-customer-segmentation-model-interpretation> (<https://www.kaggle.com/code/micheldc55/mall-customer-segmentation-model-interpretation>). (Note that PCA is used at the end for visualization in an excellent way!)

Timeline:

- **Friday 4/28:**
 - Groups + Project Decided
- **Tuesday 5/2:**
 - Project "Work Day" (no official class material)
- **Thursday 5/4:**
 - Presentations










* Writeup can be handed in later (end of week) as long as your presentation is ready for Thursday 5/4. If you receive feedback during your presentation that you want to implement into your writeup, you may get an extension (to be discussed during the presentation).

Data Set Instructions


Pick one of [classification](#), or [clustering](#).

Here are some recommended datasets for each. You may pick a different one, but verify with me first.

Classification:

- (>2 class) **Classification with a Tabular Vector Borne Disease Dataset**  (<https://www.kaggle.com/competitions/playground-series-s3e13/data?select=train.csv>) (based off of **Vector Borne Disease Prediction**  (<https://www.kaggle.com/datasets/richardbernat/vector-borne-disease-prediction>)) - uses **mean average precision**  (<https://www.kaggle.com/code/nandeshwar/mean-average-precision-map-k-metric-explained-code/notebook>)
- (Binary) **Binary Classification with a Tabular Kidney Stone Prediction Dataset**  (<https://www.kaggle.com/competitions/playground-series-s3e12>) (based off of **Kidney Stone Prediction based on Urine Analysis**  (<https://www.kaggle.com/datasets/vuppalaadithyasairam/kidney-stone-prediction-based-on-urine-analysis>)) - uses AUC
- (Binary) **Binary Classification with a Tabular Pulsar Dataset**  (<https://www.kaggle.com/competitions/playground-series-s3e10/data?select=test.csv>) (based off of **Pulsar Classification**  (<https://www.kaggle.com/datasets/brsdincer/pulsar-classification-for-class-prediction>)) - uses log-loss; see "Evaluation tab"
- (Binary) **Binary Classification with a Tabular Reservation Cancellation Dataset**  (<https://www.kaggle.com/competitions/playground-series-s3e7>) (based off of **Reservation Cancellation Prediction dataset**  (<https://www.kaggle.com/datasets/gauravduttakiit/reservation-cancellation-prediction>)) - uses AUC

Clustering:

- **Heart Disease Patients**  (<https://www.kaggle.com/datasets/kingabzpro/heart-disease-patients>)
- **Customer Clustering**  (<https://www.kaggle.com/datasets/dev0914sharma/customer-clustering?select=segmentation+data.csv>)
- **Bullying in Schools**  (<https://www.kaggle.com/datasets/leomartinelli/bullying-in-schools>) (see me for a cleaned version of this dataset)

Writeup Instructions

- Email me your written submission. (One member submits for the entire group)
- Each group member receives the same grade.
 - If you feel as though a group member did not contribute significantly, please let me know.
- Should be 3--6 pages in total, including figures, tables, etc.
 - Describe your data set
 - Describe model choices and modelling process
 - Supplement with plots and tables (e.g. comparing model metrics) as needed
 - Discuss final model choice (and also provide kaggle score)

Grade breakdown (Classification)

- **(10%) Overall report quality** (no typos, easy to read, etc.)
- **(15%) Thorough use of data visualization methods**
- **(15%) Model choices** (choices are informed by plots and data context, and thoroughly use methods discussed in class, such as whether it should be parametric or nonparametric (and why), avoiding overfitting but using enough relevant variables, feature engineering, justifying model choice based on model assumptions, etc.)
- **(15%) Model refinement** (showcasing initial models and how they can be improved after initial assessments)
- **(15%) Correct use of statistical methodology**

Grade breakdown (Clustering)

- **(10%) Overall report quality** (no typos, easy to read, etc.)
- **(15%) Thorough use of data visualization methods and exploratory data analysis**

- **(10%) Correctly utilizing data for clustering** (e.g. transforming when needed, potentially removing collinear features, potential feature engineering, handling categorical and interactions)
- **(10%) Choosing, assessment, and visualization of final clustering model** (choosing a reasonable and justified number of clusters and clustering methodology, visualizing final clusters e.g. through principle component analysis)
- **(15%) Interpretation of final clusters** (summary statistics and plots that are cluster specific, connections to data set and context, coherence with data set and prior knowledge)
- **(10%) Correct use of statistical methodology**

Presentation Instructions

- Present during Week 15 (Thu 5/4) 2:30pm-[until all presentations done]
- Aim for a 8--10 minute presentation
- Each group member should say something during the presentation.
 - You may be asked certain questions.
 - Points may be docked off if you say things incorrectly, either during the presentation or question answering.
- Grade breakdown (first and third bullets are graded **individually**, not as a group)
 - **(10%) Overall presentation quality** (data set explained well enough, modelling choices discussed and motivated by relevant plots, results slides, etc.)
 - **(10%) Thorough use of figures and plots**
 - **(10%) Correct use of statistical methodology**

File Name
