Readme File

Clear and unambiguous instructions on how to reproduce the predictions from start to finish including data pre-processing, feature extraction, model training and prediction generation. Point out the corresponding file names:

Instructions on how to reproduce the predictions:

1. Take the file **heart\_disease\_uci.csv** and put it into the same folder as the final\_project.pynb file to run any predictions. If you have a new data file that has the same structure and want to use the code to predict, DO NOT FORGET TO UPDATE THE FILE NAME IN LINE 4 where **heart\_disease\_uci.csv** is**.**
2. Run the following lines in this order:

Line 1: to load all libraries and packages

Line 3: Load functions needed for different tasks in the process

Line 4: Load the dataset you which to analyze. If you want to use ours do not change anything in this line.

Line 6: To get an idea of number of rows, mean, standard deviation per attribute and 5 numbers.

**Data Cleaning and Preprocessing**

Line 17: Fills missing data with mean for continues variables and mode for categorical variables

Lines 19 to 20: Data Transformation

**Create Training/Validation/Test Split and Standardization**

Lines 22 to 27: Creates the splits and standardized the data

**Creating the transform into PCA and Kernel PCA**

Lines 45 to 56: Creates PCA and KPCA

**Running Model Training and Prediction**

Lines 62 to 155: Trains and runs model on data in order of Model. Start at the beginning of each section and run through cells in order.

Environment details regarding how the model was developed and trained, including OS, memory (RAM), disk space, CPU/GPU used, and any required environment configurations required to execute the code:

We ran our notebook in Google Collab on Python 3 Google Compute Engine backend with 12.69 GB of RAM and 107.72 GB of Disk space. We did not create specific environments to run this code, only a requirements.txt file.

Data Files:

Our data file that is used is heart\_disease\_uci.csv. Keep in the folder with the notebook. You can also find the data file at <https://www.kaggle.com/redwankarimsony/heart-disease-data>.