Data analysis project 3:

Applying machine learning methods to movie ratings data

Mission command preamble: As in general, we won't tell you how to do something. That is up to you and your creative problem solving skills. However, we will tell you what we would like you to do. One exception: We do expect you to do this work yourself, so it reflects your intellectual contribution.

Purpose: In this project, you will demonstrate essential machine learning skills. We revisit the same dataset you already used in project 1 and project 2. This will highlight what machine learning methods can and cannot do for you, compared to Hypothesis testing and Prediction methods. Please write a report (1-3 pages, as needed) that answers all the questions below. Use figures as needed to make your case.

Dataset description: This dataset features ratings data of 400 movies from 1097 research participants.
1st row: Headers (Movie titles/questions) – note that the indexing in this list is from 1
Row 2-1098: Responses from individual participants
Columns 1-400: These columns contain the ratings for the 400 movies (0 to 4, and missing)
Columns 401-420: These columns contain self-assessments on sensation seeking behaviors (1-5)
Columns 421-464: These columns contain responses to personality questions (1-5)
Columns 465-474: These columns contain self-reported movie experience ratings (1-5)
Column 475: Gender identity (1 = female, 2 = male, 3 = self-described)
Column 476: Only child (1 = yes, 0 = no, -1 = no response)
Column 477: Movies are best enjoyed alone (1 = yes, 0 = no, -1 = no response)
Note that we did most of the data munging for you already (e.g. Python interprets commas in a csv file as separators, so we removed all commas from movie titles), but you still need to handle missing data.

What we would like you to do: (each question is worth 20% of the grade score):

1) Apply dimension reduction methods – specifically a PCA – to the data in columns 421-474. As laid out above, these columns contain self-report answers to personality and how these individuals experience movies, respectively. It is us to you whether you do one PCA each for personality and movie experience, or one overall, but regardless of that, we would like you to:
   a) Determine the number of factors (principal components) that you will interpret meaningfully (by a criterion of your choice – but make sure to name that criterion). Include a Scree plot in your answer.
   b) Semantically interpret what those factors represent (hint: Inspect the loadings matrix). Explicitly name the factors you found and decided to interpret meaningfully in 1a). Be creative.
2) Plot the data from columns 421-474 in the new coordinate system, where each dot represents a person, and the axes represent the factors you found in 1). Hint: If you identified more than 2 meaningful factors, it is a good idea to create several 2D (X vs. Y) subplots for better interpretability.
3) Identify clusters in this new space. Use a method of your choice (e.g. kMeans, DBScan, hierarchical clustering) to do so. Determine the optimal number of clusters and identify which cluster a given user is part of.
4) Use these principal components and/or clusters you identified to build a classification model of your choice (e.g. logistic regression, kNN, SVM, random forest), where you predict the movie ratings of all movies from the personality factors identified before. Make sure to use cross-validation methods to avoid overfitting and assess the accuracy of your model by stating its AUC.
5) Create a neural network model of your choice to predict movie ratings, using information from all 477 columns. Make sure to comment on the accuracy of this model.

Extra Credit: Use machine learning methods to tell us something interesting and true about the movies in this dataset that is not already covered by the questions above [for an additional 10% of the grade score].