SP2026 Statistical Learning Project III

- **Important:** This is not a group project. Please work independently.
- Data `proj3data.csv` The dataset is an extract from a survey of shopping mall customers in the San Francisco Bay area (1987). It consists of 14 demographic attributes. The dataset contains both categorical and continuous variables with a lot of missing data. The goal is to predict the Annual Income of Household from the other 13 demographics attributes.
- Training data: the first 4497 observations.
- Testing data: the last 4496 observations.
- Attribute Information

1 ANNUAL INCOME OF HOUSEHOLD (PERSONAL INCOME IF SINGLE)

1. Less than $10,000
2. $10,000 to $14,999
3. $15,000 to $19,999
4. $20,000 to $24,999
5. $25,000 to $29,999
6. $30,000 to $39,999
7. $40,000 to $49,999
8. $50,000 to $74,999
9. $75,000 or more

2 SEX: 1=Male, 2=Female

3 MARITAL STATUS

1. Married
2. Living together, not married
3. Divorced or separated
4. Widowed
5. Single, never married

4 AGE

1. 14 thru 17
2. 18 thru 24
3. 25 thru 34
4. 35 thru 44
5. 45 thru 54
6. 55 thru 64
7. 65 and Over

5 EDUCATION

1. Grade 8 or less
2. Grades 9 to 11
3. Graduated high school
4. 1 to 3 years of college
5. College graduate
6. Grad Study

6 OCCUPATION

1. Professional/Managerial
2. Sales Worker
3. Factory Worker/Laborer/Driver
4. Clerical/Service Worker
5. Homemaker
6. Student, HS or College
7. Military
8. Retired
9. Unemployed

7 HOW LONG HAVE YOU LIVED IN THE SAN FRAN./OAKLAND/SAN JOSE AREA?

1. Less than one year
2. One to three years
3. Four to six years
4. Seven to ten years
5. More than ten years

8 DUAL INCOMES (IF MARRIED)

1. Not Married
2. Yes
3. No

9 PERSONS IN YOUR HOUSEHOLD

1. One
2. Two
3. Three
4. Four
5. Five
6. Six
7. Seven
8. Eight
9. Nine or more

10 PERSONS IN HOUSEHOLD UNDER 18

0. None

1. One
2. Two
3. Three
4. Four
5. Five
6. Six
7. Seven
8. Eight
9. Nine or more

11 HOUSEHOLDER STATUS

1. Own
2. Rent
3. Live with Parents/Family

12 TYPE OF HOME

1. House
2. Condominium
3. Apartment
4. Mobile Home
5. Other

13 ETHNIC CLASSIFICATION

1. American Indian
2. Asian
3. Black
4. East Indian
5. Hispanic
6. Pacific Islander
7. White
8. Other

14 WHAT LANGUAGE IS SPOKEN MOST OFTEN IN YOUR HOME?

1. English
2. Spanish
3. Other

- Number of observations: 8993.

  These are obtained from the original dataset with 9409 instances, by removing those observations with the response (Annual Income) missing.

- The missing value is NA.

- Try all the methods that are applicable for the data set contained in proj3data.csv.

Project Report Format:

- Please typeset your report using R Markdown in RStudio to produce a PDF file. Your report should have the following
  - A title
  - Abstract
  - Introduction
  - Section(s) containing details of your data analysis, include only relevant R codes used for data analysis, graphs and explanation of the methods you used, etc. If random data were generated you must specified a random seed so that your results can be repeated.
  - Summary of Results and Discussion which include a table to summarize the test error rate of the models trained and give the predicting formula of your final model if possible.
- Limit your report within 20 pages. Only report necessary and relevant R code, graphs, and printout. Make your report neat, clean, readable, not like draft. Do not include the data in your report.
- **Important**: Using AI tools or anything like this is prohibited. Use only the R libraries and functions taught or mentioned in the lectures or the textbook or no credit!