MAS456
Final Project Guideline

General Guideline

- The final project will test students' ability of data analysis to answer research questions in social epidemiology. Each student will analyze the provided data using statistical learning methods learned in the course, write a report and provide 15 min power point video presentation.
- **Deadline**: by December 17th (Sunday), 22:00:00 (KST).
- **Submission materials:** (1) Written Report (pdf file), (2) Power Point Video Presentation (mp4 file), (3) R or Python code file
  - ✓ All three items should be submitted in separate files through the KLMS.

General Guideline

- **Written Report and the Power Point Presentation should include the following sections** (No more than 10 pages for written report and No longer than 15 minutes for presentation).
  1. Title – project title, student's name and ID number
  2. Introduction – research background and objective
  3. Method – data description, details of statistical analysis
  4. Results
  5. Conclusions
- You may include as many figures or tables as you want in the Appendix (no page limit for the Appendix) if it is impossible to include all of them in the report given the page limit.

Data Description

- In this project, you will analyze a nationally representative longitudinal dataset obtained from the Korean Labor and Income Panel Study (KLIPS), launched in 1998. You will use the data from the 7th wave of the KLIPS (2004), where the participants' experiences of hiring discrimination were measured.

- The provided dataset includes 18 variables measured for **3,576 participants** who were waged workers at the time of the survey to ensure that all participants were eligible to answer either 'Yes' or 'No' to the question "Have you ever experienced discrimination in getting hired?" Among 3,576 participants, **3,479 people responded as either 'Yes' or 'No' but 97 people responded as 'Not Applicable'.**

Data Description

	Variable name	Description	Possible answers
1	disc_hire	Response to the question, "Have you ever experienced discrimination in getting hired?"	0:'No', 1:'Yes', NA:'Not Applicable'
2	Gender	Gender	0:male, 1:female
3	Age	Age	0:16–24, 1:25–34, 2:35–44, 3:45–54, 4:55–64, 5:65+ years old
4	Edu_cat	Education level	0:middle school graduate or less, 1:high school graduate, 2:college graduate or more
5	Marriage	Marital status	0:never married, 1:currently married, 2:previously married
6	Emp_fin	Employment status	0:permanent, 1:non-permanent
7	Income_quartile	Total household income divided by the square root of the number of household members	0:Q1, 1:Q2, 3:Q3, 4:Q4 (4 categories based on the quartiles)
8	Birth_region	Birth region	1:Jeolla-do, 0:other regions
9	Self-rated health	Response to the question, "How would you rate your health?"	0:'very good', 1:'good', 2:'poor', 3:'very poor'
10	Disability	Response to the question "Do you have any impairment or disability?"	0:'No', 1:'Yes'
11	Residence	Residential areas	1:Seoul, 2:Pusan, 3:Daegu, 4:Daejeon, 5:Incheon, 6:Gwangju, 7:Ulsan, 8:Kyunggi, 9:Kangwon, 10:Choongbuk, 11:Choongnam, 12:Jeonbuk, 13:Jeonnam, 14:Kyungbuk, 15:Kyungnam
12	disc_wage	Experience of discrimination in receiving income	0:'No', 1:'Yes', 2:'Not Applicable'
13	disc_jobedu	Experience of discrimination in training	0:'No', 1:'Yes', 2:'Not Applicable'
14	disc_promotion	Experience of discrimination in getting promoted	0:'No', 1:'Yes', 2:'Not Applicable'
15	disc_resign	Experience of discrimination in being fired	0:'No', 1:'Yes', 2:'Not Applicable'
16	disc_edu	Experience of discrimination in obtaining higher education	0:'No', 1:'Yes', 2:'Not Applicable'
17	disc_home	Experience of discrimination at home	0:'No', 1:'Yes', 2:'Not Applicable'
18	disc_social	Experience of discrimination at general social activities	0:'No', 1:'Yes', 2:'Not Applicable'

Research Questions to Address

- **Question 1**: What are the important variables that are associated with the experience of hiring discrimination? How are those variables related to the experience of hiring discrimination?
    - To answer this, you have to fit a logistic regression considering the ‘disc_hire’ as a response and others as explanatory variables. Use the data only for 3,479 participants whose value for ‘disc_hire’ is 0 or 1.

Research Questions to Address

- **Question 2:** What are the important principle components (PC) that explain a large portion of variation in the following 12 explanatory variables (gender, age, education level, employment status, income, self-rated health, disability, and experiences of discrimination in receiving income, training, getting promoted, being fired, and obtaining higher education)? How would you interpret those PCs?
  - To answer this, you have to conduct the principle component analysis for the 12 variables. You have to choose an appropriate number of important PCs based on some criteria. Though the 12 variables are discrete, you may just take them as continuous variables.

Research Questions to Address

- **Question 3**: Identify subgroups (clusters) based on the 12 variables you use for PCA to answer question 2. Also, identify subgroups based on the important PCs you find to answer question 2. Compare the clustering results.
    - To answer this, you have to conduct a cluster analysis. You have to decide an appropriate number of clusters based on some criteria. Once you determine the number of clusters, you need to provide description about the characteristics of each cluster.

Research Questions to Address

- **Question 4**: Is there a difference in under-reporting of hiring discrimination between males and females?
  - To answer this, you have to build a prediction model using the data for 3,479 participants to predict the response to the question about the experience of hiring discrimination for the 97 participants who responded as 'NA' even if they were eligible to respond as "Yes" or "No". Then, you have to predict the response for the 97 participants and compare the distribution of the predicted response between males and females. (See the next slide for specific objectives in each step of the analysis)

Research Questions to Address

- While building a prediction model, you have to consider two methods (Random Forest, Penalized logistic regression with Lasso penalty).
- You have to conduct a cross-validation to identify the optimal tuning parameter in each model.
- The predictive performance of each method should be compared by AUC of the ROC for the test data. You have to choose the best prediction model as the one with larger AUC.
- Once you choose the best prediction model, you have to determine the optimal threshold for the binary prediction based on the ROC curve such that the sum of sensitivity and specificity was maximized.

Research Questions to Address

- **Question 5:** Is there an association between the experience of hiring discrimination and health?
  - To answer this, you have to compare the distribution of self-rated health among the four groups of people whose response was (1) No', (2) 'Yes, (3) NA' but predicted as 'No', and (4) 'NA' but predicted as 'Yes'.
  - To compare the distribution, you may just take 'Self-rated health' as a continuous variable.
  - You have to conduct an overall test to see if there is a difference among the 4 groups and pairwise comparison applying an appropriate method for multiple testing adjustment.

Grading Guideline

- **Grading Criteria**

Item		%	Overall %
Final Project	Introduction • Background (Motivation) • Objective (Research Hypotheses)	5%	40%
	Methods • Data description • Statistical Method	20%	
	Results	10%	
	Conclusions/Discussions	5%	

Grading Guideline

Check points in each section

- Introduction
  - Research background and objective are well-described?
- Method
  - The analyzed data are well-described?
  - When performing the PCA, the important PCs are well-selected, well-reported, and well-interpreted?
  - When conducting a cluster analysis, the number of clusters is well-selected? The characteristics of each cluster are well-described?
  - When building a prediction model,
    - The two models (Random forest, penalized logistic regression with LASSO) are considered?
    - the tuning parameters are properly selected and well-reported for each model?
    - the considered models are well-compared and the model comparison is well-reported?
    - When predicting based on the best prediction model, the threshold for binary prediction was properly selected and well-reported?
- Results
  - All of the research questions are well-answered?
  - Tables and Figures are appropriately used to present the results effectively?
- Conclusion/Discussion
  - Are the conclusions well-made based on the results?
  - Are some discussions made for the limitation of the research and future research direction?