PROJECT FLOW AND REQUIREMENTS

- *Please prepare your project in R Markdown in the following format which is described in this document (1-8 and X-Z).*
- *"Clear comments before and after every little step" and "clear data visualizations with desriptions" will make good points.*
- *Don't ever put the whole data in the project report!! If you want to show your data, please just show the first rows (6-10); or use summaries- data visualization techniques instead of data itself.*
- *Write your names in the author part.*
- *Put your final docs (Project report as an html and a short video) in dys.*
- *You will be graded as: (80p following steps; 10p report clarity and quality, 10p video presentation)*

1. **(5p) Please find your original dataset or datasets; and describe your data in the first step.**

2. **(4p) Use "Exploratory data analysis". Write down your comments.**

3. **(4p) Use some "visualization techniques" and talk about your data further.**

4. **(4p) Check your data for multicollinearity, make your comments.**

5. **(7p) Apply PCA:**
   a. Use appropriate functions and arguments,
   b. Use visualization techniques for PCA, describe the result!
   c. Make your final comments clearly!

6. **(7 p) Apply Logistic Regression or Regression.**
   a. Use appropriate functions and arguments,
   b. Use visualization techniques for Regression, describe the result!
   c. Which performance scores you chose? What is the final result? Make your final comments clearly!

7. **(14 p) Apply at least 2 Clustering Techniques**
   a. Describe the reason you choose those 2 techniques.

      7.1 ……. Algorithm Application
         a. Use appropriate functions and arguments,
         b. Use visualization techniques. Describe the result!
         c. Make your final comments clearly.
      7.2 ……. Algorithm Application
         a. Use appropriate functions and arguments,
         b. Use visualization techniques. Describe the result!
         c. Make your final comments clearly.

b. Compare the results you have found in 7.1 and 7.2. Which performance scores you chose? What is your final decision? Make your comments!

8. **(14 p) Apply at least 2 Classification Techniques** (if you applied logistic regression to the same data in step 6, before, you can skip 8.2, and MENTION it (Ex: 8.2: I'll use the output of step 6 here, so I'm skipping 8.2 here)
   a. Describe the reason you choose those 2 techniques.

      8.1 ……. Algorithm Application
         a. Use appropriate functions and arguments,
         b. Use visualization techniques. Describe the result!
         c. Make your final comments clearly.
      8.2 ……. Algorithm Application
         a. Use appropriate functions and arguments,
         b. Use visualization techniques. Describe the result!
         c. Make your final comments clearly.

   b. Compare the results you have found in 8.1 and 8.2. Which performance scores you chose? What is your final decision? Make your comments!

X: (7p) Use the PCA results (principal components) you have found in "step 5" ; either for logistic regression or any other classification techniques: <u>Compare the "results with original data" and "results with components", make your comments!</u>

Use this part "at least once" under 6,7 OR 8.
If you really need to go on with components rather than original data, then you can use it through 6-8.

Your project headings for 6-8 might look something like:

6. (just to try PCA data)	7. (just to try PCA data)	7. (You really need to use PCA data)
a	a	a
b	7.1	7.1.X
c	a b c	a b c
6.X	7.2	7.2.X
	a b c	a b c
	b	b
	7.X (compare PCA data with 7.1 or 7.2, not both)	

Y: (7p) Missing Data imputation:

If you have your original data with missing values, do step "Y.a" after step 3 directly:

Y. a Use an imputation method to impute those NA value. Continue with complete data for the following steps.

Y. b In just one of your applications through 6-8:

- Apply the classification or clustering algorithm to "data with missing values" and "data with imputed values".
- Compare the "results with missing values" and "result with imputed values". Which performance scores you chose? What is your final decision?

If your original data is complete, do this step at the end, after step 8:

Y.

a. Delete about 20%-40% of your data set randomly. Make them NA values, as if they are missing. Describe what you did there.
b. Use an imputation method to impute those NA values.
c. Choose a classification or clustering algorithm (a new one or one of the techniques you used in 6-8).

- Apply the classification or clustering algorithm to "data with missing values" and "data with imputed values".
- Compare the "results with missing values" and "result with imputed values". (DON'T compare it with the original dataset!). Which performance scores you chose? What is your final decision?

Your project headings might look something like:

your original data has missing values	your original data is without missing values
… 3.	… 6.
Y.a	7.
4.	8.
5.	Y. a b c
6.	….
Y.b	
7.	
8. …	

Z: (7p) Imbalanced data set

If you have your original data as imbalanced, do step "Z.a" after step 3 directly:

Z. a Use oversampling, undersampling or both, to balance your data. Continue with the balanced data for the following steps.

Z. b In just one of your applications through 6-8:

- Apply the classification algorithm to "imbalanced data" and "balanced data".
- Compare the "results with imbalanced data" and "result with balanced data". Which performance scores you chose? What is your final decision?

If your original data is already balanced, do this step at the end of your project:

Z.

a. Make your data imbalanced, (in order to do it you should delete some part randomly). Describe what you did there.
b. Use oversampling, undersampling or both, to balance your data.
c. Choose a classification algorithm (a new one or one of the techniques you used in 6-8).
   - Apply the classification algorithm to "imbalanced data" and "balanced data".
   - Compare the "results with imbalanced data" and "result with balanced data". Which performance scores you chose? What is your final decision?

ZEYNEP FILIZ EREN