ECEU601303 - Introduction to Data Science (Class A & B)
Final Project (35%)

Lecturer: Dr. Eng. Arie Wahyu Wijayanto, SST, MT

This is a group project. You need to form a group of 5 students. Once you have formed the group, please inform the class leader (Naila for Class A, Calista for Class B).

Data

In this final project, you will need to explore and analyze the house prices data.[1] You can download the data (and its description) from:
https://s.id/ProjectDatasetIDS

Task

1. Perform exploratory data analysis and clustering analysis on the data. Discuss and interpret the results.
2. Build a linear regression model to predict the house price.
    a. You need to show that you have to explore different models and perform model selection in order to obtain the most accurate model.
    b. Evaluate the prediction performance of the model.
    c. Discuss and interpret the results.
3. Build classification models to predict the class of house price. Note that you need to first transform the house prices into two (or more) distinct classes. For example, you can categorize all house prices below a certain threshold to be "cheap" and above the threshold to be "expensive".
    a. You need to use tree-based models.
    b. Perform and discuss methods to increase the model accuracy.
    c. Perform and discuss k-fold cross-validation / other necessary methods to prevent overfitting in the models.
    d. Evaluate the classification models. Remember to choose the appropriate evaluation metrics.
    e. Discuss and interpret the results.

Presentation (10%)

A verbal presentation for the final project will be conducted in the week of 9th and 10th. The presentation will be worth 10% points of the total evaluation mark.

You will have 10 minutes to present your project. A slide presentation needs to be submitted before the deadline of Tuesday, 5 Nov 2023 at 09.00PM.

[1] Source: Anna Montoya, DataCanary. (2016). House Prices - Advanced Regression Techniques. Kaggle. https://kaggle.com/competitions/house-prices-advanced-regression-techniques.

File name format must be: [class]_[groupnumber]_presentation.pdf, for example "A_1_presentation.pdf". The file type must be .pdf.

Send the report to arie.wahyu02@ui.ac.id (cc: ariewahyu@gmail.com). Email title should be: [class]_[groupnumber] Presentation, for example "A_1 Presentation".

Written report (25%)

A brief written report (maximum 15 pages including figures and tables) containing the justification of the approach, the presentation and the discussion of the results, and conclusions should be submitted before the deadline of Monday, 4 Dec 2023 at 09.00PM.

The written report will be worth 25% points of the total evaluation mark.

In your report, please properly describe the dataset and cite the data source(s). Do not forget to also cite all related works that you reference. The bibliography does not count towards the maximum 15 pages.

In the final page of your report, write a description on what each individual in your team has contributed into. Please also write the percentage of the contribution for each individual.

Write your group name on the heading of the first page. File name format must be: [class]_ [groupnumber]_finalproject.pdf, for example "A_1_finalproject.pdf".

Put the final report document and the *R* scripts into a single zip file. File name format must be: [class]_ [groupnumber]_finalproject.zip, for example "A_1_finalproject.zip".

Send the file to arie.wahyu02@ui.ac.id (cc: ariewahyu@gmail.com). Email title should be: [class]_ [groupnumber] Final Project, for example "A_1 Final Project".

Submitting without following the above guidelines will result in a penalty. Late submissions will also receive a penalty.

Marking

Both the presentation and the written report are marked based on the following criteria.

Components	*Weight*
Presentation	**25**
1. Clarity of presentations and explanations	15
2. Consistency of language and mathematical notation	10
Method	**40**
3. Justification of the choice of measures,	40

assumptions and approaches, validations	
Result	**35**
4. Scientific soundness	25
5. Originality	10