# IE 496: Independent Studies
# Eduardo Lopez Tasks for Mar 21–April 14

Instructor: Anirudh Subramanyam

March 21, 2023

## 1 Tasks for Weeks 10–14

Complete the tasks in this order. If you are finished with these tasks, let the instructor know so we can chart out next steps.

1. [Reading] Read the introduction to Chapter 24 and Section 24.1 of the book [1].
   - Supporting video lecture in this link.
2. [Reading] `scikit-learn` is a popular Python library for machine learning. Read the first section titled 'Machine learning: the problem setting' here.
3. [Reading] Read Section 20.1.1 of the book [1] about linear regression.
4. [Reading + hands-on exercises] Read Chapter 17 of the book 'Learning Statistics with Python'.
   - Cover everything upto and including Section 17.5.1. You may skip the later sections.
   - As you are working your way through the chapter examples, you may want to directly experiment with data columns of the 'vehpub.csv' file, since analyzing this data will be part of your programming tasks described next.
   - Note that this book uses the term "predictor variable" (this is a term common among statisticians) to refer to what is also known as a "feature" (this term is common in machine learning) in [1]. See this link.
5. [Reading] Read the first paragraph and understand the first two code blocks here talking about training and testing datasets.
   - Your goal should be able to distinguish between training and testing data and how to split a given dataset into these two components using the function `train_test_split`.
6. [Programming] Now that you have filtered your dataset by removing negative or other missing entrees, do the following:
   (a) Split the 'vehpub.csv' dataset into training and testing sets.
   (b) Choose a subset of up to 10 different features that can meaningfully predict `GSTOTCST` and provide an intuitive explanation for your choice.

(c) Fit a linear regression model on the training set.

(d) Compute $R^2$ values of your model on the testing set.

It will be helpful to follow the workflow shown in this example.

7. [Presentation (Due on Mar 31)] Make a roughly 15-minute presentation summarizing your results from the above programming task. In addition, address the following points:

   (a) Explain using examples the difference between supervised and unsupervised learning.

   (b) Explain using examples the difference between regression and classification models.

   (c) Why should we not train machine learning models on the test data? In other words, what are the roles of the train and test datasets?

   (d) Provide an interpretation of the $R^2$ values you obtained in your programming task.

   (e) Based on your findings, would you recommend to keep, drop or add new features to the ones you selected in Step 2 of the programming task.

8. [Reading] Read Chapter 26 of the book [1].

   - You may skip Section 26.3.
   - Supporting video lecture in this link.

9. [Programming] This supplements the regression analysis from the previous programming task with classification.

   (a) Split the 'vehpub.csv' dataset into training and testing sets.

   (b) Choose a subset of up to 10 different features that can meaningfully predict the U.S. residential state of a household and provide an intuitive explanation for your choice.

   (c) Fit a logistic regression classifier on the training set.

   (d) Compute the accuracy and confusion matrices of your model on the testing set.

   It will be helpful to follow the workflow shown in this example. Note that this example uses a different type of classification model (called `SVC`) but everything else should be similar.

10. [Presentation (Due on Apr 14)] Make a roughly 15-minute presentation summarizing your results from the above programming task. In addition, address the following points:

    (a) What are some common metrics that are used to evaluate the performance of a classification model?

    (b) Provide an interpretation of the accuracy values and confusion matrices you obtained in your programming task.

    (c) Based on your findings, would you recommend to keep, drop or add new features to the ones you selected in Step 2 of the programming task.

## References

[1] John V. Guttag. *Introduction to Computation and Programming Using Python, third edition: With Application to Computational Modeling.* MIT Press, London, England, 2021.