# Problem Set 5

Data Science 602: Data Analysis and Machine Learning

Spring 2022

1. **Logistic Regression** In the example given in class, using the Iris dataset, we predicted the species using only the petal length and width. Repeat the study adding the sepal width and length (in addition to the petal width and length) as predictors. Does classification accuracy improve with the additional predictors?

2. **Binary classification** In the example given in class, we used a logistic regression (`LogisticRegression`) classifier to classify digits. Continuing on the example presented in class that uses the MNIST 784 dataset, use the classifier to classify whether a digit is a 3. The classifier includes the hyperparameter $C$ where $10^{-4} \lessapprox C \lessapprox 10^4$. This parameter defines the regularization strength, the meaning of which we will discuss later in this course. Experiment with at least 4 different values of this hyperparameter across different orders of magnitude. What value of $C$ produces the highest $F_1$ scores?

   **Notes/Hints:**

   (a) While you can compute different values manually, the `GridSearchCV` class is recommended.

   (b) Training may require about an hour on the full dataset of 60,000 records using Colab, with training time increasing as $C$ increases. You can use a smaller subset of the training dataset if needed.

   (c) If you receive a message that the mode failed to converge due to maximum number of iterations being reached, increase the `max_iter` parameter.

3. **Precision-Recall curve** Generate and plot the precision-recall curve for the highest-performing model you identified in the previous question.