# CSE455/CSE552 – Machine Learning (Spring 2021)

# Term Project

**Handed out**: April 12, 2021.

**Due**: 11:55pm June 21, 2021.

**Hand-in Policy**: Via Teams. No late submissions will be accepted.
**Collaboration Policy**: No collaboration is permitted.
**Grading**: This project will be graded on the scale 100.

**Description**: The aim of this project is to implement a random decision forest algorithm using a specific decision tree algorithm. The algorithm should work for both classification (more than two classes) and regression. It should be tested on at least four different datasets (two for classification, two for regression) and compared to the baseline implementation provided by a library of your choice.

You should do the following:

1. Implement a decision tree classifier/regressor using trained neural networks as decisions at a given node. Such an algorithm may look like the following:
    a. Assume that all your features are real-valued.
    b. Decision tree training algorithm as discussed in class.
    c. At each node, the decision is made by training an MLP with one hidden layer over the data reaching the node.
    d. The number of nodes in the hidden layer is a hyperparameter.
    e. As the tree branches more and more (depth), there will be fewer and fewer data reaching those nodes. This will make training an MLP from scratch very unstable. You are expected to devise a strategy to cope with this. Possible strategies may include:
        i. Train a base model with the given training data and refine this model at each node with the available data there.
        ii. Refine the model from the parent node.
        iii. Randomly pick a trained node and refine.
        iv. Use all the data with different weights (increased weights to the data trickling to the current node).
        v. You should consider the fact that the number of nodes in MLPs may change.
    f. Make sure that your decision tree algorithm or algorithms can handle multiclass classification as well as multinomial regression problems.
2. [Graduate Students Only] Implement the random forest algorithm discussed in class.
    a. Use the decision tree algorithm above as the base classifier.
    b. You should include additional randomness by changing the number of nodes in the hidden layers of MLPs.
3. Report the performance of your implementations using an appropriate k-fold cross validation on four different datasets.
    a. Pick two datasets for classification with multiple classes.

b. Pick two multinomial regression datasets.
c. Establish a baseline performance using an RDF implementation of your choice.
d. Compare the baseline performance against your implementation.
e. Discuss the results with particular focus on the decision tree algorithm, ensemble performance (for graduate students) and overfitting/underfitting (model complexity). The discussion section should be at least one page long (letter size, 11pt, single spaced).

**What to hand in:** You are expected to hand in the following

- **Project_lastname_firstname_studentnumber_code.ipynb** (the Python notebook file containing the code and report output).

Your notebook should include something like the following:

**Implementations:**

**Tests and Results:**

**Discussions:**