

# **CSI 5387: Data Mining and Concept Learning**

## Winter 2020

## **Assignment 2**

Submission Deadline: 4th March 2020 on Brightspace.

A sample of the dataset for women of the Pima Indian heritage living near Phoenix, Arizona that tested for diabetes according to World Health Organization is hereby provided. Develop a Feed-Forward Neural Network to predict the outcome of the test using either Weka, R or Python. Please note that there are some missing values in the data set.

The performance of the classifier should be evaluated by partitioning the dataset into a train dataset (75%) and test dataset (25%). Use the train dataset to build the Neural Network and the test dataset to evaluate how well the model generalizes to future results. Complete the following using a suitable version of the file provided for your platform.

#### A. Data Preprocessing (10 marks):

- 1. Some data points are not available, handle the missing data by applying central measure of tendency to derive the missing value.
- 2. Neural networks work best when the input data are scaled to a narrow range around zero. Rescale the data with a normalizing (e.g., *min\_max normalization*) or standardization (e.g., *z\_score standardization*) function.

### B. Model Development (60 marks):

Single Layer:

- 1. Train a simple multilayer feedforward network with only a single hidden node (not layer).
- 2. Plot the Neural Network. Your plot should depict the weights for each of the connections and the bias terms (if possible).
- 3. To estimate the model's performance, generate predictions on the test dataset.
- 4. Generate a confusion matrix for the model.

#### Multi-Layer:

- 1. Increase the number of hidden nodes to 5 and the number of hidden layers to 2. What impact does this have on the accuracy as compared to the single layer perceptron?
- 2. Try changing the activation function, varying the learning rate, epochs or removing the bias. What effects does any of these have on the result?

#### C. Model Comparison (10 marks):

1. Using the same data set partitioning method, evaluate the performance of a SVM classifier on the dataset.

#### D. Model Evaluation (20 marks):

- 1. Compare the results of the Multi-layer perceptron with the SVM model according to the following criteria: Accuracy, Sensitivity and Specificity.
- 2. Identify the model that performed best and worst according to each criterion.
- 3. Carry out a ROC analysis to compare the performance of the Multi-layer perceptron model with the SVM model. Plot the ROC graph of the models.