**Project 1:**

In this project, you will apply three algorithms to two data sets. The data and the algorithms are provided by R. Please answer each question in the order they appear. Do not skip to later steps to answer earlier questions that ask you to predict outcomes based on your analysis of the data and understanding of the algorithms.

Submit your report in D2Lby midnight on the due date.

**The Data Sets**

Contact Lens data

<http://archive.ics.uci.edu/ml/datasets/Lenses>

Iris data

<http://archive.ics.uci.edu/ml/datasets/Iris>

**The Algorithms**

* K-Nearest Neighbors
* Decision Trees
  + classifiers/trees/Id3
  + classifiers/trees/J48
* Neural Networks
  + classifiers/functions/MultilayerPerceptron

**The Report**

Copy and paste the text below into a text editor or using LaTeX template of IEEE Templates for Transactions: <https://www.ieee.org/publications_standards/publications/authors/author_templates.html>

Submit only a PDF. **Do not** send MS Word document.

Be prepared to present your report in 3 slides.

**Data Sets**

What are the differences between the two data sets?

Which algorithm do you expect to perform best on the Contact Lens data data? Why?

Which algorithm do you expect to perform best on the Iris data? Why?

**KNN on the Contact Lens Data**

Run KNN on each data set with 1, 3, 5, 7 and 9 neighbors. Report the results for each run in a confusion matrix and comparisons in a table or graph.

Which K gives the best results? Why?

Holding K constant, try different distance functions on each data set. Which distance function(s) work best for each data set? Why?

**KNN on the Iris Data**

Run KNN on each data set with 1, 3, 5, 7 and 9 neighbors. Report the results for each run in a confusion matrix and comparisons in a table or graph.

Which K gives the best results? Why?

Holding K constant, try different distance functions on each data set. Which distance function(s) work best for each data set? Why?

**Decision Trees on the Contact Lens Data**

Based on R's vizualizations, which attribute do you expect to be chosen as the split attribute at the root node?

Run each decision tree on the data and report the results for each run in a confusion matrix and comparisons in a table or graph.

How do ID3 and J48 compare in terms of performance?

How does pruning affect test performance and generalization performance? What does that suggest about overfitting?

**Decision Trees on the Iris Data**

Based on R's vizualizations, which attribute do you expect to be chosen as the split attribute at the root node?

Run each decision tree on the data and report the results for each run in a confusion matrix and comparisons in a table or graph.

Why can't you run ID3 on the Iris data?

How does pruning affect test performance and generalization performance? What does that suggest about overfitting?

**Multilayer Perceptrons on the Contact Lens Data**

Experiment with different network structures (e.g. extra hidden layers, extra units). Report the results in graphs that show training time (epochs) versus error rate or accuracy.

Which network structures result in the most overfitting?

**Multilayer Perceptrons on the Iris Data**

Experiment with different network structures (e.g. extra hidden layers, extra units). Report the results in graphs that show training time (epochs) versus error rate or accuracy.

Which network structures result in the most overfitting?

**General**

Which algorithm performed best on each data set, for particular definitions of "best?"

Was the (comparative) performance of the algorithms as you expected? Why?

Which data set had the best performance in general across all of the algorithms? Why?