IS 733 Homework 1, Due 2/16/2021 (by midnight through Blackboard)

In this homework task, you will practice the initial steps of data mining, including obtaining data and getting it in the right form for machine learning algorithms, preprocessing, and running an initial analysis using off-the-shelf tools. You can submit your homework to me on paper, just before class on the due date.

We will be using the `WEKA` software system for this homework. Start by downloading it from: `http://www.cs.waikato.ac.nz/ml/weka/` and installing it (use the current stable version). `WEKA` is java-based, and is hence cross-platform so you should be able to install it on any computer. Points

1. Download the Wine dataset from the UCI machine learning repository, at `http://archive.ics.uci.edu/ml/datasets/Wine` . This dataset is for a classification task where the goal is to classify wines according to their origin, based on their chemical properties.

   (a) Open the data file and its corresponding description in a text editor. How many instances are there? Which of the attributes in the file is the class? (The dataset description does not state this, but it should be straightforward to figure it out. Give your answer as an index, e.g. if the class were the 3rd column counting from the left, your answer could be "Attribute number 3.") (10)

   (b) Make a copy of `wine.data`, named `wine.arff`. Convert this data file to `WEKA`'s `ARFF` format, by manually adding the necessary header information using a text editor (see the slides for Lesson 2.) In the next step, `WEKA` may give you an error message if your file format was not correct, which can be useful for checking your work. List the header lines that you added to the file. (20)

2. Open `WEKA`, and open the `Explorer` interface, under `Applications`. Near the top-left of the `Explorer` window, click on `Open file`, and then browse to your `ARFF` file and open it. If you get an error message at this stage, or notice any other problems, fix your `ARFF` file and try again.

   (a) In the bottom-left of the `Explorer` window, in the `Preprocess` tab (which is open by default), there is a list of attributes. Clicking on each of these produces a histogram of its values, and statistics including the min, max, mean, and standard deviation for numeric attributes. Which class has the most instances? Which attribute has the largest standard deviation? Based on the histograms, name one attribute which is multimodal, i.e. has multiple peaks. (A peak, a.k.a. a "mode," can occur at the edge of the histogram. For this question, if there is a "valley" between two peaks we will count them as multiple modes, no matter how small the valley is.) (20)

   (b) Switch to the `Classify` tab. There is a drop-down menu to select the class attribute; use this to ensure that the correct attribute is selected as the class. Under `Classifier`, select `Choose`, then `weka -> classifiers -> trees -> J48`. This is a standard decision tree classification algorithm, also known as `C4.5`. Leaving the `Test options` as `Cross-validation (10 folds)`, click the `Start` button. The classifier will run on the dataset. Read the `Classifier output` panel. What percentage of test instances did the algorithm predict correctly? How much time did the classification algorithm take to build the model? (10)

3. Perform any two data preprocessing steps on this dataset. You may use any tool you like to perform these steps, including the `WEKA Explorer`'s `Preprocess` tab (under `Filter`), Excel, Python, etc.

   (a) Explain your rationale for each of these steps: why might they help improve performance or otherwise be beneficial (regardless of whether or not they actually turned out to help this time.) (20)

   (b) Report the percentage classification accuracy under the same experimental setup as in the previous question, for each of these preprocessing steps individually, and for the combination of both of

them. Did either of them improve performance versus the result in Question 2b? (It's absolutely OK if they didn't, this problem is relatively easy so it may be difficult to improve performance.) (20)

Total: 100