SCS3212 Assignment

24th April 2025

*Use the provide assignment reporting template.

*You will designate one person in your group to upload to the Googleclass.

Think through this assignment as you start to attempt it. Use the accompanying dataset.

PART 1

Use the Phone dataset

- Using WEKA use two algorithms:
    - k-Nearest Neighbour
    - Multi Layer Perceptron
    - Naïve Bayes
    - SVM
- Predict the length of call for the query data (second tab in the excel worksheet), with the requirements listed below.
    (a) Convert the data appropriately and adding value as discussed in class. You can use different categorizations and data conversion.
    (b) Use all the features and see what result you get, then attempt to discover the relevant attributes and see what results you get.
    (c) In all your model building use 10-cross fold validation
    (d) When using k-NN, determine the most appropriate k value

Phonedata data description

The data has been extracted from a local mobile phone service provider *(Equinox)*. It comprises of randomly selected subscribers and their calling data. Included too is some personal data.
The data attributes are:

- Call record ID
- Subscriber Age
- Subscriber Gender
- Subscriber Marital status
- Subscriber phone number
- Duration of call
- Time of call

PART 2

Use the Thyroid data set

Phonedata data description

- This is a real patient data from a hospital in Nairobi. It is about a cohort of patients who had their HIV status monitored before and after the use of ARV. The motivation of this reported data was to see if there is any effect on the Thyroid gland for patients on ARVs. Please use this dataset in confidence. Remember it is about people and we must be sensitive.
- Note that there are two datasets here.

- You are a machine learning engineer. You have been provided with this thyroid dataset. You will need to read up about Thyroid gland, ARVs, HIV conditions, etc. Using this information you can then prepare for your ML activities.
- Document your steps clearly.

Using WEKA use two algorithms:

- Decision Tree (J48 - an improvement of C4.5) decision tree algorithm)
    - Use the default settings
    - What rules do you get? What does this mean? (Note that the current dataset only has about ½ the original data. This data is before the treatment. The data was on two worksheets, only one was converted. The other worksheet has data after treatment)
    - Now separately set up the (after treatment) worksheet only and run j48 with the default settings. – What rules do you get? What does this mean?
    - Now combine the two worksheets and do the same as above. – What rules do you get? What does this mean?
- k-Nearest Neighbour (IBK)
    - Use a similar process as above (for Decision trees)
    - Use k values of 1 and 3 and describe what results you get

What have you discovered and what does it really mean?

PWWagacha