ASSIGNMENT 2: DQR, MODEL BUILDING, AND EVALUATION

JOHN D. KELLEHER

This is a group assignment and each group is given a separate dataset to work on. Deadline for submission is 10 am on Friday the 27^{th} of April.

1. What you need to submit

Each group has to submit:

(1) a Data Quality Report for you dataset (tables, visualisations)
(2) code that creates at least 3 different prediction models from the sci-kit library and runs evaluation experiments on them
(3) a brief report document (structure explained below)
(4) a 10 minute presentation based on the report that you will present to the group on Friday (each team member must speak for an equal portion of the presentation).

Each of these components is described in more detail in the following sections of this document.

2. Code

You can use the same code that you used for Assignment 1 to generate the Data Quality Report for this assignment. You do not need to resubmit this code. What you do need to submit is a Python program that:

(1) imports the dataset from a file
(2) transforms the data into the format requred for sci-kit
(3) splits the data into a training and a test set. **Note, often the documentation that accompanies the datasets suggest a particular split of the data into training and test set. Follow this suggestion where it is mentioned. Otherwise, the size of the test set should be set to account for the overall dataset, the larger the dataset the larger you can make the test set.**
(4) creates 4 different prediction models from the sci-kit library; for example, a decision tree, nearest neighbor, naive bayes models, random forest.

Team	Name	URL
A	Covertype	`https://archive.ics.uci.edu/ml/datasets/Covertype`
B	Diabetes	`https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008`
C	Record Linkage Comparison	`https://archive.ics.uci.edu/ml/datasets/Record+Linkage+Comparison+Patterns`
D	PAMAP2 Physical Activity Monitoring	`https://archive.ics.uci.edu/ml/datasets/PAMAP2+Physical+Activity+Monitoring`
E	Cenuse-Income Dataset	`https://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29`
F	Bank Marketting	`https://archive.ics.uci.edu/ml/datasets/Bank+Marketing`
G	Bike Sharing	`https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset`

(5) for each model created your program should run a 5 fold-crossvalidation and output the accuracy score for each fold
(6) for each model created your program should then test the model on the test set and output the confusion matrix and the accuracy score on the test set

Remember to **ensure that your code is properly commented**

3. Report Document and Presentation

The report and presentation should have the following sections:

(1) **Dataset Description**: A description of the dataset, including: an explanation of what the prediction task is, information regarding how many instances are in the dataset and how may features the dataset includes.
(2) **Data Quality Report and Analysis**: The data quality report section should include the initial tabular analysis of the continuous and categorical features, the relevant visualisations for the features, and a brief report that identifies data quality issues (cardinality, outliers, and missing values) and explains how you intend to fix these problems before you begin modelling.
(3) **Data Handling**: The data handling section should describe what data cleaning you performed and explain and motivate how you split the data into training and test set, including the proportion of the dataset that went into each.
(4) **Modelling**: A section describing the prediction models you created, including:
    (a) a description of the models you created and an explanation of the exploration of the hyper-parameters that you undertook,
    (b) an explanation of which evaluation metric you used and motivate why you selected that metric for your dataset
    (c) for each model present the accuracy score for each fold of the 5 fold cross validation,
    (d) present the confusion matrix and accuracy score of each model on the test set.
(5) **Jazz**: I want you to take ownership of the project and to do things because you wanted to try them out and to explore possible alternatives. So, this section is the section where you describe things that you did because **you thought that they were useful/interesting** (and not because you were told to do them). Remember, it could be something that you tried but didn't work out. A negative finding can be just as useful as a positive one. For example, did you exclude any features from the modelling process and if so why? did you design any new features based on the raw data? did you find particular model parameters to help with the model accuracy? did you use a validation set or did you apply pruning?
(6) **Reflections:** what do you think worked or didn't work on the project? why do you think these things did/didn't work? have you a theory as to why some models worked on your dataset and others didn't? . . .