Assignment 5

DSCI 222 – Data Science Workflow using Python
School of Mathematical and Data Sciences
West Virginia University

Instructions

This is a homework assignment to be solved in a small group (3-4). It is a great opportunity to show your teamwork skills. Enjoy it! All material needed for the assignment can be found in Github

- Create and share with me a single folder with all members of the group clearly listed.

  Files included in your deliverables folder

  - Deliverable 1 Video Recording
  - Deliverable 2 Colab Notebook
  - Deliverable 3 Colab Notebook
  - Deliverable 4 Report
  - Anything additional or supplementary that contributed to the assignment (this includes .csv files, .json files, etc.)

- Include your **full name** as a Python comment at the top of the notebook and at the top of your report.
- All data manipulation should occur within the Python script. No manipulation of the supporting file(s) should occur prior to importing the file(s) into your script. You may check your work by manually performing data analysis.
- Include your report in PDF format, written in LaTeX.
- Set share folder permissions so chood@mix.wvu.edu can access and run every file and notebook.
- Everything counts! Include as much as you want in your deliverables, even if the activity is not fully complete by the deadline. *Important:* Review the grading policy and course policies in the online syllabus.
- **Total: 100 points.**

Activity 1: Python Source Code and Video Recording (60 points)

Scenario

In this assignment we will be exploring some FitBit Tracking Data to see how well we can predict the calories burned by an individual using a variety of machine learning techniques. One type of machine learning model we have not learned about in class is decision trees. These are the basis of many more complex models, including random forests and gradient boosted trees (the most common base model for gradient boosted techniques). Just like many other models, decision trees can be used for both regression and classification.

For Activity 1, you will be teaching a lesson on what decision trees are, and how to program one in python. To help, I have provided a sample script that includes only the bare bones for how to produce a decision tree. This assignment has two components. The first will be a Google Colab used to teach about decision trees, and a video recording of the "lecture".

Deliverables 1 and 2 (combined):

Google Colab Notebook and Video Recording

- In the Assignment 5 folder on GitHub, you will find the file `Decision_tree.ipynb`.
- Convert this short script into a well-structured notebook suitable for teaching a lesson to DSCI students on what decision trees are, how to create them, and how to interpret them.
- Add substantial explanatory text, comments, spacing, data outputs, and visual outputs throughout the notebook to improve clarity and readability.
- The current script uses only a minimal number of parameters. Expand on the parameters and explain their purpose to improve the educational value of the notebook.
- You may add additional code to enhance the lesson. At a minimum, you must explain how to evaluate the validity of the decision tree model and how to use it to predict calorie counts.
- Once your notebook is complete, create a video recording of your group teaching the lesson on decision trees.
- For the video component, you may begin with your completed `.ipynb` file. The provided `Decision_tree.ipynb` file is simply a starting point and does not need to be referenced during your presentation. You may present the final code as if it is entirely your own work.
- In the video, you must clearly explain what each line of code does. Every part of the script should be addressed and nothing should be skipped.

- Be sure to clearly explain what a decision tree is and how the `DecisionTreeRegressor` in scikit-learn constructs one.
- Your video must be a screen recording of the code as you explain it so that viewers can follow along. You may appear on screen if you choose, but this is not required.
- You should include explanatory text, comments, and Markdown cells in the `.ipynb` file to support your explanation.
- All members of the group must speak during the video.

Activity 2: Python Source Code and Report (40 points)

Scenario

We are now going to look at a variety of ways to perform machine learning to make predictions on the total number of calories an individual wearing a FitBit may expend.

Deliverable 3:

Google Colab Notebook

Problems

Problem 1: Our data begins with 14 potential features to predict Calories. Perform some sort of analysis to determine which 5 of these features you would like to use for the remainder of the assignment (these will be justified in the report).

Problem 2: Create a table, where five methods of regression are selected and 4 regression evaluation metrics are selected to predict the calories burned.

Problem 3: Use five different methods of classification to predict whether a person is inactive (under 2000 calories), active (between 2000 and 2500 calories), and very active (over 3000 calories). These methods of classification should be sorted by model accuracy.

Problem 4: Use DBSCAN to detect any outliers in our data.

Problem 5: Perform PCA, t-SNE, and UMAP to reduce our data to 3 dimensions. Present the results visually.

- If you identify more than one interpretation of a questions listed above, make sure to explain why the chosen approach was selected over others.
- Before each block of code, include text explaining how that block works.
- Within each block of code, include comments explaining how each component of the code works and why it was used.
- Any libraries, functions, methods, etc., used that are not explicitly discussed in class must be clearly explained.

Deliverable 4:

Report Write a report on what conclusions you are able to draw from the problems addressed in deliverable 3.

- How did you decide which five features to include in your analysis?
- Which method of regression best fits the data? Provide justification for why this is the best model and explain how the model works. Be sure to explain what the evaluation metrics selected mean and how they can be interpreted within the context of the assignment. Artificially create a new row of data and use your model to predict the total calories.
- Which method of classification best fits the data? Provide justification for why this is the best model and explain how the model works. Artificially create a new row of data and use your model to predict the total calories for this new set of data.
- Based on the visuals that you have produced, which (if any) methods of dimensional reduction would be appropriate for our data? How are you able to tell?